# INFORMATION RETRIEVAL IN CONSTRUCTION HAZARD IDENTIFICATION

Hyunsoo Kim*, Hyun-Soo Lee, Moonseo Park, and Sungjoo Hwang

*Department of Architectural Engineering, Seoul National University, Seoul, Korea*
** Corresponding author (verserk13@naver.com)*

**ABSTRACT**: The repetitive occurrence of similar accidents is one of the most prevalent features of construction disasters. Similar accident cases provide direct information for determining the risks of scheduled activities and planning safety countermeasures. Researchers have developed many systems to retrieve and use past accident cases. However, this research has the limitations of performing too much retrieval to obtain results regarding construction site conditions, failing to reflect characteristics of safety planning steps, or both. To overcome these limitations, this study proposes an accident case retrieval system that can search similar accident cases. It also improves safety planning by using information retrieval and building information modeling. The retrieval system extracts BIM objects and composes a query set combining BIM objects with site information DB. DB compares a query set with past accident cases and seeks the most similar case. Results are then provided to safety managers. Based on the results of this study, safety managers can reduce excessive query generation. Furthermore, they can easily recognize construction site risks by obtaining coordination of objects where similar accidents occurred.

***Keywords***: *Safety Management, Accident Case, Information Retrieval, BIM*

## 1. INTRODUCTION

In spite of enormous safety efforts in the construction industry, it has had a poor record of preventing accidents. One of the noticeable characteristics of construction accidents is repetitive occurrence (Abudayyeh et al. 2003). Although field managers in construction sites understand the characteristics of accidents, the changeable conditions of a site make it hard for managers to apply proper accident cases to safety management. Accident cases are the strongest stimulation for planning safety management. Especially they which are directly related scheduled activities provide information to be able to predict risk and establish safety countermeasures (Ko et al. 2005).

Some approaches, such as case-based reasoning (CBR) and information retrieval systems based on accident cases, have attempted to put accident cases to practical use (Ahn 1994, Moon et al. 1997, Moon and Yang 1997, Ye 1998). The systems suggested by this research carry out information retrieval by translating construction site conditions into queries. However, these systems should separately retrieve information to find proper countermeasures for each of the various risk factors in construction sites. Although this research secures useful research results, it is limited by ineffective retrieval processes.

To overcome these limitations, this study proposes an accident case retrieval system that can search similar accident cases automatically. Also, the results provided by this system include remaining periods and coordination of influence factors.

## 2. PRELIMINARY STUDY

The progress of information technology has helped to address various challenges presented by construction projects (Paulson 1995), and it has been introduced to safety management. Many researchers use safety information by relating construction activities or retrieving accident cases applied to safety planning (Ahn 1994, Moon and Yang 1997, Moon et al. 1997, Ye 1998, Ko et al. 2005). Some research about safety information system is performed at the national level. Korea Occupational Safety and Health Agency (KOSHA) and Occupational Safety and Health Administration (OSHA) provide web-based search services to aid in the prevention of accidents. Recent research can be categorized into two groups. One combines activities and risk factors (Sacks et al. 2009), and another involves search systems based on CBR (Chua and Goh 2002, 2009, Goh and Chua 2010).

However, safety information retrieval systems developed from these researches use a Boolean retrieval model. The retrieval process is based on whether or not the documents contain the query terms. Although the Boolean retrieval model is simple and clear, too many or too few results can be retrieved by its strict matching rule. Another problem is that most research sets the weight of each index equally. This leads to a decrease in the effectiveness and usefulness of retrieval results.

A method overcoming these limitations of the Boolean retrieval model is to provide a frame that is able to be partly matched. It means that indices of queries or texts are allocated of non-binary weights (Manning et al. 2008). Nevertheless, by the suggested limitations, current researches improve safety management by retrieving accident cases. However, if managers want to identify countermeasures through retrieving accident cases, current safety retrieval systems should input as many accurate queries as the number of risk factors, as shown in Figure 1.

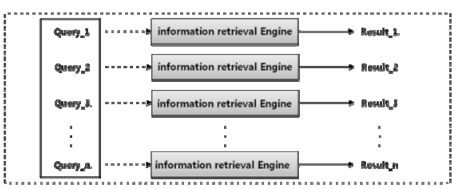


Fig. 1. Method of current retrieval systems

It is established through different levels of information and management. It is important to establish a proper safety management plan based on available information, because obtainable information is different for each step of safety management. As a construction project progresses, a preliminary plan, monthly plan, and weekly plan are established. At some specific phases, these may overlap. Therefore, they are different from the type and level of information that should be extracted from each step of the safety management level (An 2007), system ought to support these requirements.

Limitations of current safety management systems and requirements of safety managers are as follows. First, they do not classify information for each step of safety management. Second, numerous queries are generated to obtain search results for whole risk factors in a construction site. Third, the results are not classified by the progress of a project and do not represent hazardous areas. Therefore, current systems have low efficiency because safety managers do not map and establish links between hazardous areas or develop appropriate countermeasures. To address the issues mentioned above, this study proposes a system that can generate queries automatically, perform retrieval, and provide search results with the timing and locations of risk factors.

## 3. RETRIEVAL ALGORITHM

### 3.1 Analysis of Retrieval Method

Generally, information retrieval can be divided into ad hoc systems and filtering systems. Ad hoc models input queries whenever a user requires new information. Filtering systems store user interests in a user profile, and their main objective is selecting new generated information and deleting useless information. As shown in Figure 2, the difference between these two systems is whether a query is short-term (one-off) or not (semi-permanent).

In this study, a routing model, one of the information filtering models, is adopted to achieve a safety manager-centric system. The most important part of a routing model is not ranking results but compiling a user profile (Manning et al. 2008). Before compiling a user profile, the information required by safety managers should be defined and converted to indices. This process is described in detail in the next section.

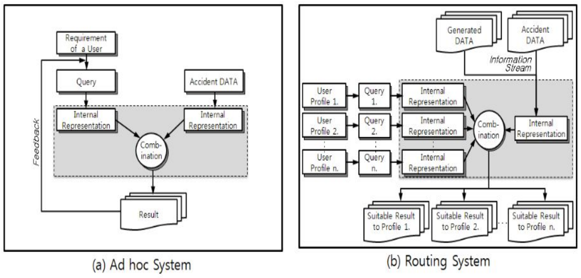


Fig. 2 Ad hoc system and routing system

3.2 Indices for Retrieval

(1) Extraction and Weight of Indices

Construction accidents are caused by the influence of various risk factors (Lee et al. 2009). Therefore, preventing accidents requires the consideration of various factors complexly. In this study, the risk factors suggested by Lee et al. (2009) are used, and the extraction process is as follows.

First, 27 factors were selected through a literature review. Then, the number of factors was reduced to 17 using measurable and statistical data. Finally, 10 factors were extracted by surveys conducted with 42 safety managers, each with more than 10 years of experience in safety management. The details of the extraction process are shown in Figure 3.

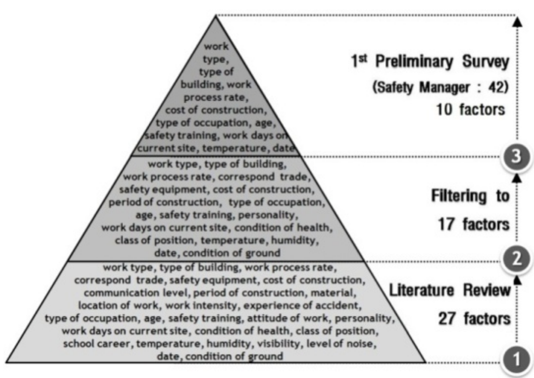


Fig. 3. Extraction of factors

Safety management plans are organized into three steps (preliminary, monthly, and daily). The uncertainty of construction projects makes it necessary to acquire information at each safety planning phase.. To remedy the fact that current accident retrieval systems do not consider the differences of each phase, ten factors are classified by the possibility of obtaining factors at each phase. The classification is arranged according to the degree of possibility of being affiliated with factors mentioned above (Lee et al. 2009), as shown in Table 1.

To weight the indices extracted in the former section, the analytic hierarchy process (AHP) developed by Saaty (1980) is used. AHP's capacity to convert qualitative to quantitative values is used to weight indices.

The questionnaire for calculating weights follows that of Lee et al. (2009). It was sent to 50 experts who had worked more than ten years as safety managers. Of the 50 questionnaires distributed, 43 were collected. The weights of indices were analyzed based on questionnaires with a contingency index below 0.1.

Table 1. Weight of Indices by safety planning phase

| Influence Factors | Preliminary Plan | Monthly Plan | Daily Plan |
|---|---|---|---|
| Work process rate | 0.241 | 0.167 | 0.099 |
| Cost of construction | 0.175 | 0.131 | 0.077 |
| Work type | 0.334 | 0.238 | 0.125 |
| Building type | 0.250 | 0.173 | 0.096 |
| Occupation type | - | 0.181 | 0.100 |
| Date | - | 0.109 | 0.059 |
| Age | - | - | 0.069 |
| workdays on current site | - | - | 0.118 |
| Safety training | - | - | 0.166 |
| Temperature | - | - | 0.089 |
| No. of distributed surveys | 50 | 50 | 50 |
| No. of collected surveys | 43 | 43 | 43 |
| No. of collected surveys with a contingency index below 0.1 | 43 | 31 | 23 |

To judge the degree of match between each index, it is important to determine the proper data format (Ye 1998). Also, determining weights of indices is necessary to heighten similarity measurement (Kolodner 1993).

The calculating similarity index method and data format are shown in Table 2. If data follows a numeric format, the distance between two values which correspond to construction site condition value and accident case is calculated using Formula 1.

$$SI = 1 - \frac{|A - B|}{A} \quad (1)$$

Where A = condition value of a current site; B = condition value of a accident cases

Table 2.Mehtod for calculating similarity index

| Influence Factors | Data Form | Calculating Method |
|---|---|---|
| Work process rate | Num | Formula 1. |
| Cost of construction | Num | Formula 1. |
| Work type | Char | If A=B, SI=1<br>If A≠B, SI=0 |
| Building type | Char | If A=B, SI=1<br>If A≠B, SI=0 |
| Age | Num | Formula 1. |
| Occupation type | Char | If A=B, SI=1<br>If A≠B, SI=0 |
| workdays on current site | Num | Formula 1. |
| Safety training | Num | Formula 1. |
| Date | Num | Formula 2. |
| Temperature | Num | Formula 2. |

Formula 1 cannot be directly applied to some indices. Therefore, in the cases of date and temperature, construction site condition value and accident case values are converted to calculate similarity. For example, if site condition (A) is January and accident case (B) is December, they have a very high degree of similarity in the view of continuity. However, degree of similarity is determined as zero if formula 1 is applied. To solve these problems, the weights of each month and each temperature are adopted to calculate similarity indices as suggested by Lee et al.(2009). The conversion method is as follows. First, the weights which conform to site condition value and case condition value are extracted from Figure 5. Then, they are applied to Formula 2.

$$SI = 1 - \frac{|A' - B'|}{A'} \qquad (2)$$

Where A' = converted condition value of a current site; B' = converted condition value of a accident cases

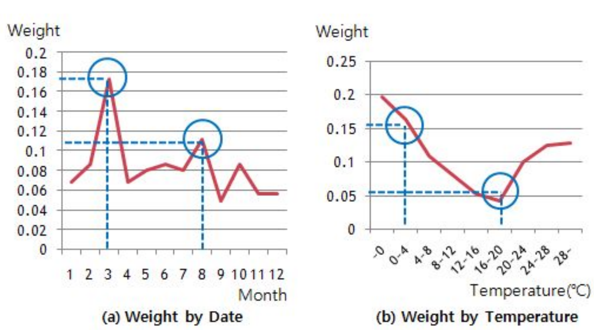


Figure 5. Weight graph of date and temperature

For example, if a site condition value is March and a case condition value is August, the weights of each case become 0.173 (A') and 0.111 (B'), respectively. After calculating the similarity index of each influence factor, the similarity score (SS) is determined. This value can be expressed as a sum of the multiplication of the similarity index (SI) and weight (M), using Formula 3.

$$SS = \sum_{i=1}^{n} M_i SI_i \qquad (3)$$

3.3 DATA Source for Developing Systems

Systems that can collect the aforementioned factors can be broadly divided into commercial programs and established project management information systems (PMIS). PMIS has abundant information for managing construction projects. It also has the benefit of minimizing additional data input by using a developed system and database. However, PMIS lacks information about extracted factors-related coordination and objects. If information about indices is extracted by an established PMIS, a data conversion process is necessary. Also, a mapping process linking data, coordination, and objects is necessary to satisfy the requirements of safety managers.

Commercial BIM programs, on the other hand, are designed to be able to add properties of an object. Two of the most widely used programs are Revit (AutoCAD) and ArchiCAD (Graphisoft). Geometric properties, such as coordination, volume, length, etc., can be extracted. However, developing tools provided in Revit or ArchiCAD, like GDL, are necessary to program objects which contain the various kinds of information suggested in this study. This function is not commonly used in the design process. Although these limitations have been overcome, data processing capacity declines noticeably when abundant information is inputted to object property value.

Therefore, in this study, DB incorporates the benefits of PMIS and commercial BIM programs. Kim et al. (2010) have demonstrated the possibility of combining and linking information between PMIS and influence factors. Lee et al. (2010) have suggested a model combining BIM with scheduling and safety information. Based on these suggested methods, DB is established on the outside of BIM system and saves values of influence factors and accident cases. Using the extraction function of BIM, geometric values are extracted and connected with DB. Through this kind of system composition, users can save, handle, and use information as much as they want. Also, property values can be fully used without decreasing the efficiency of commercial BIM programs.

## 3.4 Retrieval System Framework

The retrieval system framework suggested in this study is shown in Figure 6. The system consists of input, data processing, and output. The input has information about factors extracted from section 3.2, weight of each factor, and accident case DB. The number of influence factors differs with safety planning phases and construction site conditions. Weight of each factors DB has weight values acquired through AHP analysis. These values are used as indices of weight in an information retrieval module. Accident case DB is utilized to search similar accident cases by comparing queries based on construction site conditions.

Data processing comprises the BIM drawings module, extraction module, and information retrieval module. The BIM drawings module forms 3D oriented parametric models, has geometric property values of objects, and provides objects that add property values. The extraction module extracts objects from BIM drawings. The export function in commercial BIM programs is used as an extraction method. Finally, extracted objects and information corresponding with influence factors are combined. Information processed in the extraction module is sent to the information retrieval module and is used to find similar accident cases.

The information retrieval module determines the type of retrieval method by the number and degree of BIM property values. After that, AHP result values are loaded by the type of retrieval method. Also, queries are generated and are converted to internal representations. Similarity indices are calculated by comparing accident cases and internal representations based on construction site values. After this process, this module searches similar cases.

Output ranks similar accident cases through the similarity calculation process of the information retrieval module. The ranked cases are provided to safety managers. Similar accident cases classified by the number of inputted factors can be found. Each case provides coordination and remaining period of risk factors by combining objects extracted from BIM drawings.

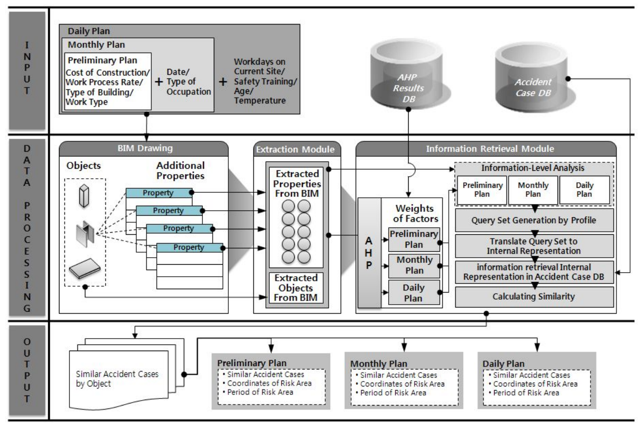


Figure 6. Retrieval system framework

## 3.5 System Development

Ye (1998) has used a two step search algorithm to achieve effective retrieval. However, since data processing speed has been improved by advanced technology, a simple retrieval algorithm may perform well. A comparison between a query and total contents of DB is adopted for retrieval.

Three types of DB are needed to perform retrieval: accident case DB, AHP result DB, and query set DB combining influence factors and objects from BIM.

An algorithm based on VBA in MS-Excel is used in order to ensure the usability of the system. Most construction sites manage scheduling and cost planning based on MS-Excel. When objects are extracted from a commercial BIM program, extracted data can be expressed as MS-Excel spreadsheets. These points can improve management efficiency. In addition, ArchiCAD is used as a modeling tool.

Each case is compared with every influence factor. After cells corresponding to each influence factor are identified, similarity scores are calculated by following the formulas in Table 2.

Each similarity index of influence factors and AHP weights are multiplied. Then, all these values are summed to calculate the similarity score of each accident case. Safety managers receive accident cases that occurred under similar conditions of a current site. At the same time, safety managers also do remaining periods and coordination of risk factors, geometric information from BIM objects. All of this information can help managers take safety measures. These results are shown in Figure 6.

| | Similar Accident Case | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| No. Dis(1) | No. Dis(2) | No. Dis(3) | No. Dis(4) | No. Dis(5) | No. Dis(6) | No. Dis(7) | No. Dis(8) | No. Dis(9) | No. Dis(10) |
| SS | SS | SS | SS | SS | SS | SS | SS | SS | SS |
| A0037 | A0217 | A0501 | A0052 | A0321 | A0039 | A0037 | A0528 | A0287 | A0662 |
| A0583 | A0163 | A0224 | A0486 | A0015 | A0551 | A0324 | A0095 | A0354 | A0354 |
| A0172 | A0022 | A0386 | A0301 | A0286 | A0383 | A0274 | A0252 | A0320 | A0049 |
| A0036 | A0531 | A0099 | A0130 | A0402 | A0061 | A0128 | A0121 | A0359 | A0292 |
| A0384 | A0314 | A0145 | A0037 | A0449 | A0109 | A0072 | A0224 | A0147 | A0131 |
| A0202 | A0076 | A0324 | A348 | A420 | A0156 | A0164 | A019 | A0406 | A0090 |

Figure 6. Retrieval Result

## 4. CONCLUSION

Because current safety management uses experiential and uniform knowledge, it cannot reflect changeable construction site characteristics. To solve this problem, some research has been performed by retrieving accident cases based on the repetitive occurrence of accidents. They improve safety management by extracting accident cases. However, they do not include safety planning steps, identify coordination of risk factors, or fully define users.

Therefore, this study proposes an accident cases retrieval system that can automatically generate queries based on construction site conditions. The results include time and geometric information as well.

The suggested system helps safety managers prepare safety countermeasures as safety planning steps. Automatic retrieval of accident cases in the system increases the efficiency of identifying risk factors. Results from the system consist of accident cases corresponding to risk factors, their coordination, and a remaining period. Moreover, this study shows that BIM can be applied to scheduling and cost planning and maintenance as well as safety management.

Despite these advantages, this study has some limitations. If there is no abundant data source such as PMIS, enormous effort is required to input data. Also, accumulated accident cases are not sufficient to be applied to safety management. Future research will eventually remedy these limitations and help to improve safety management.

**Acknowledgment**

This research was supported by a grant from Super-Tall Building R&D Project (VC-10) funded by the Ministry of Land, Transport and Maritime Affairs of Korean government.

## REFERENCES

[1]Ahn, H.(1994). “A Knowledge Model for Enhancing Construction Hazard Communication.” Ph.D. dissertation, Seoul National Univ., Seoul, Korea.

[2]An, S. (2007). "Integrated safety risk management for construction projects." Ph.D. dissertation, Seoul National Univ., Seoul, Korea.

[3]Chua, D. K. H. and Goh, T. M. (2002). "Application of case based reasoning in construction safety planning." Proceedings of International Workshop on IT in Civil Engineering 2002, pp. 298-307.

[4]Chua, D. K. H. and Goh, T. M. (2009). "Case-based reasoning for construction hazard identification: case representation and retrieval" J. Constr. Eng. Manage., 135(11), pp. 1181-1189.

[5]Goh, T. M. and Chua, D. K. H. (2010). "Case-based reasoning approach to construction safety hazard identification: adaptation and utilization." J. Constr. Eng. Manage., 136(2), pp. 170-178.

[6]Kolodner, J. L. (1993). Case-Based Reasoning, Morgan Kaufmann

[7]Lee, H., Kim, H., Park, M., Lee, K., and Lee, S. (2009). “Construction Risk Assessment Methodology Using Site Risk Influence Factors”, KICEM, 10(6), pp. 117-126

[8]Manning, C. D., Raghavan, P. and Schutze, H. (2008). Introduction to information retrieval, Cambridge University Press, New York, USA

[9]Moon, M., Kim, E., and Yang, K. (1997). “A Study on the Implementation of the Accident Information Management System in Constructions”, AIK, 13(8), pp.429-440

[10]Ye, T. (1998). “Reasoning Model of the Case-Based Construction Safety Management System” Master Course. dissertation, Seoul National Univ., Seoul, Korea.