

Predictive Models from Accident Reports

Kamalesh Panthi, PhD and
Syed M. Ahmed, PhD
East Carolina University, Greenville NC

There has been an enormous amount of resources spent on collecting data related to construction accidents but there are very few researches done on analyzing the collected data beyond trend analysis. This research is based on the premise that construction accident reports can be exploited much more than they currently are to obtain valuable information. Analyzing texts in accident reports in addition to coded data provides far more information than merely analyzing coded data. By analyzing and understanding what is in the accident report database yields useful knowledge. Passing this knowledge onto others can improve an understanding of what went wrong with incidents from the past thereby greatly enabling the prevention of future accidents. This accident prediction model proposed in this paper relies on the vast amount of information available in the accident report summaries kept by OSHA, and text mining of such textual summaries will unveil the variables and their relationships that may not be evident through structured data.

Keywords: Data Mining, Text Mining, Accident Reports, BBN, OSHA

Introduction

Construction accidents reports consist of both structured and unstructured data. Structured data consists of nominal variables, categorical variables, and continuous variables. Data such as texts are considered unstructured data (Two Crows Corporation, 1999). While there has been good use of structured data in the form of statistical analysis to test different hypothesis as well as to analyse trends, unstructured data has been very sparsely utilized. There is an enormous amount of resources spent on collecting data related to construction accidents and if such data is not utilized to the full extent it does not make much sense to collect these data in the form of accident reports. The novel technical approach is to use data mining technique to uncover interesting patterns in both structured data as well as unstructured data from accident reports. When these reports provide predictive capabilities, i.e., predicting the likelihood of construction accidents in the presence/absence of certain variables (factors), a decision support system can be developed. Intervention strategies, aimed at reducing the likelihood and impact of construction accidents, may then be selected based on their effectiveness.

This research focuses on the added benefits of analyzing the textual data obtained from accident investigation reports. This research is based on the premise that analyzing text entries in addition to coded data provides far more information than by looking at coded data alone. On the same note, construction accident reports can be exploited to provide much more information than they currently are relied upon. This research aims to extract invaluable information from construction accident reports which otherwise would remain buried and unexplored.

There is a vast amount of construction accident data in anecdotal forms. While being able to capture and portray the sequence of events leading to accidents, these textual documents do little more than that. In light of this situation, how can one generate useful information from a vast amount of textual data? Text mining to explore association and causal relationship among the variables will provide the information that is required in unravelling the interrelationship between large number of factors and sub-factors in construction accidents. When these relationships are defined and captured well thorough the actual data collected over so many years, prediction of accident and injury occurrence in a construction site becomes easy and more reliable. This concept of using accident reports and predicting future accidents is not new in commercial air transport (Luxhoj, 2003). By integrating data mining and Bayesian belief network, the authors propose a conceptual model that is capable of predicting accident proneness of a construction site based on the state of safety factor conditions.

Literature Review

The safety performance of any construction activity is affected by a plethora of varied human, technical, environmental, and organizational factors. Although the apparent cause of any construction accident is generally attributed to a single trigger, very often either due to human error or due to technical fault, it is often the case that the accidents are the result of many latent factors and sub-factors contributing together to trigger an accident. Most models of the incidence of occupational accidents in the construction industry are composed of multiple factors. Although statistical techniques can be used to infer cause-and-effect relationships among these factors, the large number of factors involved and the complexity of the relationships among each other make it difficult for managers to identify potential hazards in construction projects to develop effective safety procedures (Cheng et. al, 2010). Statistical trends in the more recent past indicate that additional research and knowledge is needed on the causation and the preventive measures so that safety professionals can give counsel on how these injuries and fatalities can be further reduced (Manuele, 2008). While there have been research efforts in the past to identify factors affecting safety on construction sites, there is a paucity of analytical methods for analysing and interpreting the complex interactions of the various system risk factors. There has been a persistent need to develop advanced risk analytics that move beyond the essential identification of risk factors to enhanced system modelling and evaluation of complex causality as well as to assessing various combinations of risk mitigation strategies (Luxhøj, 2003).

A decision support system based on data mining (DM) and Bayesian belief networks (BBN), such as the one proposed in this study, enables potential cause-and-effect relationships to be identified thereby enabling the prediction of safety performance outcome. Although there has been a considerable research effort in identifying various factors causing construction injuries, review of literature shows that there has been very little done on assessing the probability of construction safety risk in advance based on the existing site conditions. Data mining and Bayesian belief network techniques are the major concept used in developing the conceptual model and therefore, are described in detail in the following paragraphs.

Bayesian Belief Network

Bayesian Belief Networks (BBNs), also referred to as belief networks, were first developed at Stanford University in the 1970s (McCabe et. al, 1998). BBNs describe cause-effect relationships among variables through graphical models. Belief networks consist of nodes, representing variables of the domain, and arcs, representing dependence relationships between the nodes (McCabe et. al, 1998).

The use of BBNs in construction has been mainly on the improvement of construction operations (McCabe et. al, 1998), diagnosing upsets in an anaerobic wastewater treatment system (Sahely and Bagley, 2001), estimating the false-work erection productivity (Tischer and Kuprenas, 2003), making inferences in highway construction costs (Attoh-Okine, 2002), analysing risk in construction contracts (Adams, 2006) and computing probability of schedule delay in construction projects (Luu et. al, 2009). A review of aforementioned studies indicates that Bayesian Belief Networks, which can be very useful tool in forecasting the probability of construction safety risk happening given the state of conditions of the construction site, has not been utilized in construction safety research.

A simple BBN is designed to illustrate its use as shown in *Figure 1*. In this case the network represents the influence of Site Conditions and Personal Protective Equipment on the occurrence of construction accident. This network consists of three nodes, with the following states: *Site Conditions* (good, poor), *Personal Protective Equipment* (adequate, inadequate) and *Accident Occurrence* (high, low). The nodes *Site Conditions* and *Personal Protective Equipment* are the parent nodes for the *Accident Occurrence* node, which is also called the child node. For example, it may be inferred, if there are no other variables, that if the site condition is good and use of personal protective equipment is adequate, the chance of a construction accident occurrence is low. This chance is denoted by probability. However, unlike the case shown in this simple BBN example, sites conditions are affected by many other factors (both tangible and intangible) and so will be the safety management. Therefore, it is essential to represent the network well with all critical factors and their relationships before such probability can be deduced. In order to deduce such critical factors and their inter-relationships, data mining has been found to be a useful technique provided that there is plenty of recorded data.

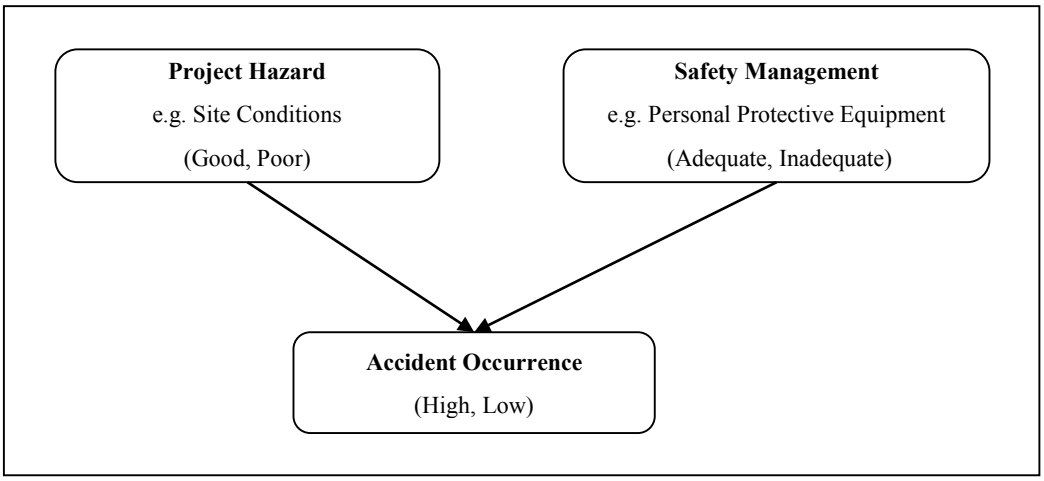

Figure 1: A Simplified Example of BBN

Data Mining

Data Mining can be defined as "the process of discovering interesting knowledge from large amounts of data stored either in databases, warehouses, and other information repositories" (Han & Kamber, 2001). Data mining has been an active analytical technique in many scientific areas for many years. These areas range from business, industry, medicine, and agriculture to engineering (Chang & Chen, 2005). However, the applications of data mining techniques to analyse the construction-related safety are relatively few. Chang & Chen (2005) have used data mining techniques to analyse freeway accident frequency. Cheng & Lin (2010) have used data mining to explore cause-effect relationship of construction related accidents in Taiwanese construction industry. Huang & Hsueh (2010) have utilized data mining approach to study the customer behaviour and decision making in construction refurbishment of buildings. Building from these literatures, it can be argued that there is a huge potential of using data mining techniques to analyse a vast pool of construction safety data that is recorded in the form of accident reports.

Methodology

A conceptual modelling approach is adopted for this study as shown in Figure 2. Causal factors are identified using database of construction accidents and injuries maintained by OSHA. Interactions among the causal factors are obtained from data mining to form a network of relationship diagrams known as Influence Diagrams. When conditional probability tables for these influence diagrams are obtained through either empirical data or through subject matter experts, Bayesian Belief Network (BBN) is constructed. Conditional probability represents the chance that one event will occur given that a second event has already occurred. For example, referring to the previous example, chance of an accident occurrence given that the site conditions are hazardous is represented by a conditional probability. With the aid of such a BBN, construction safety risk is quantified in a probabilistic form.

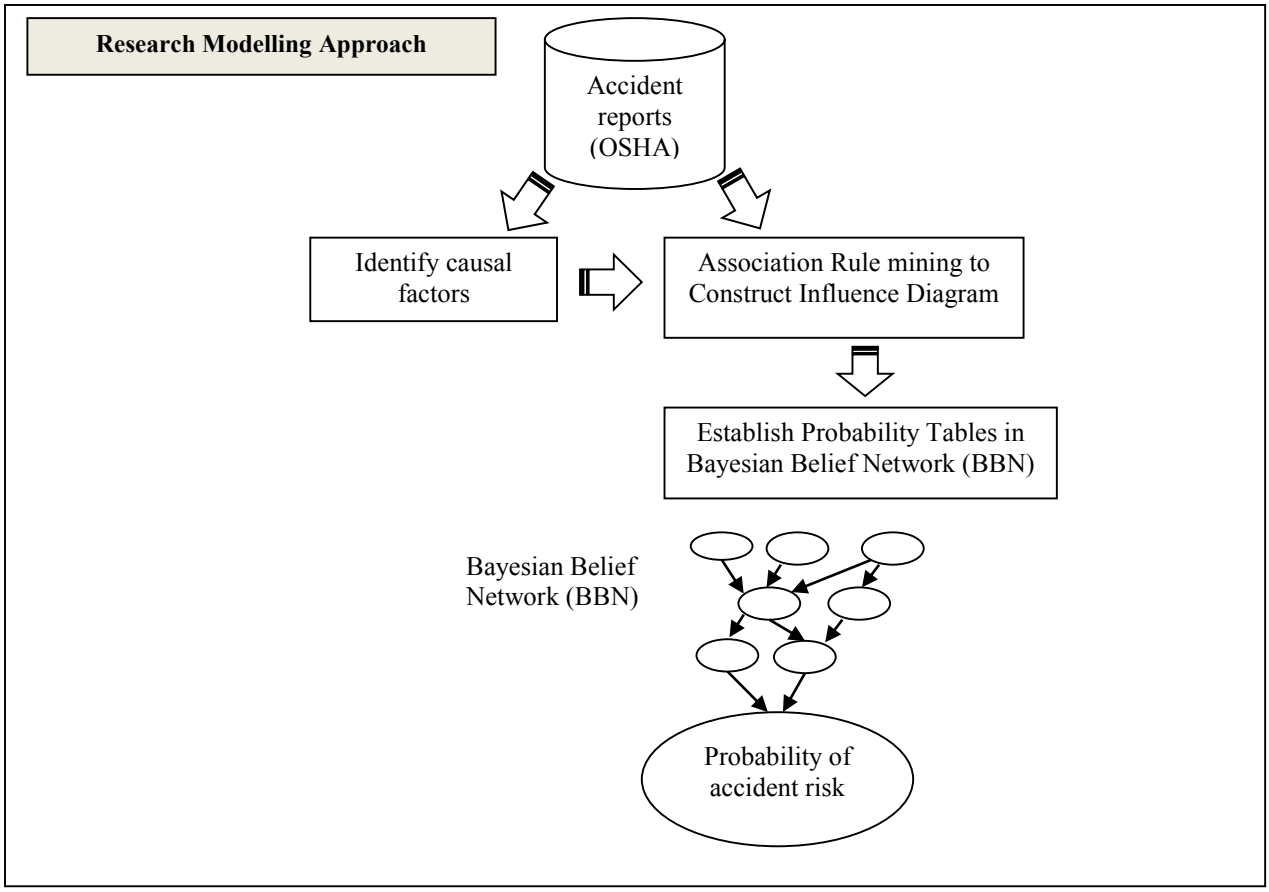

Figure 2: Conceptual Modeling Approach

Each of the major components of the conceptual model is explained in further detail as follows:

Accident Reports

The Occupational Safety and Health Administration, a division of the U.S. Departmental of Labor, is the federal agency charged with developing standards for workplace safety. When the agency, known as OSHA, conducts workplace inspections to determine compliance with its standards, it compiles reports of the results of these inspections. These reports are available to the public under the provisions of the Freedom of Information Act. Many of these reports are available online at the OSHA website. The following paragraph outlines how such data may be accessed for the purpose of the safety research such as this.

Visit the OSHA website home page and click on the tab “Data and Statistics” located at the top of the page. Data and Statistics tab will take you to the page where you can access the individual establishment inspection data. Once the "Data and Statistics" page loads, it offers you a series of choices. You will see links to various OSHA databases, such as commonly used statistics, workplace injury, illness and fatality statistics and establishment specific injury and illness data. Each presents OSHA statistics and information from a different point of view. The "Accident Investigation" search allows you to access the accident investigation summaries completed by OSHA investigators following an accident that resulted in a "fatality or catastrophe" reportable on OSHA Form 170. This search can be based on a keyword in the report, a phrase within the report, the date of the event or the industry code. The earliest summaries are those from 1984 and the newest are dated exactly one year earlier than the date you search the database. If for example, you perform a search on August 18, 2012, the newest reports you can retrieve are those dated August 18, 2011 (Charpentier, 2014). Figure 3 shows one of the search results among more than twelve thousand construction accident report summaries in OSHA’s database.

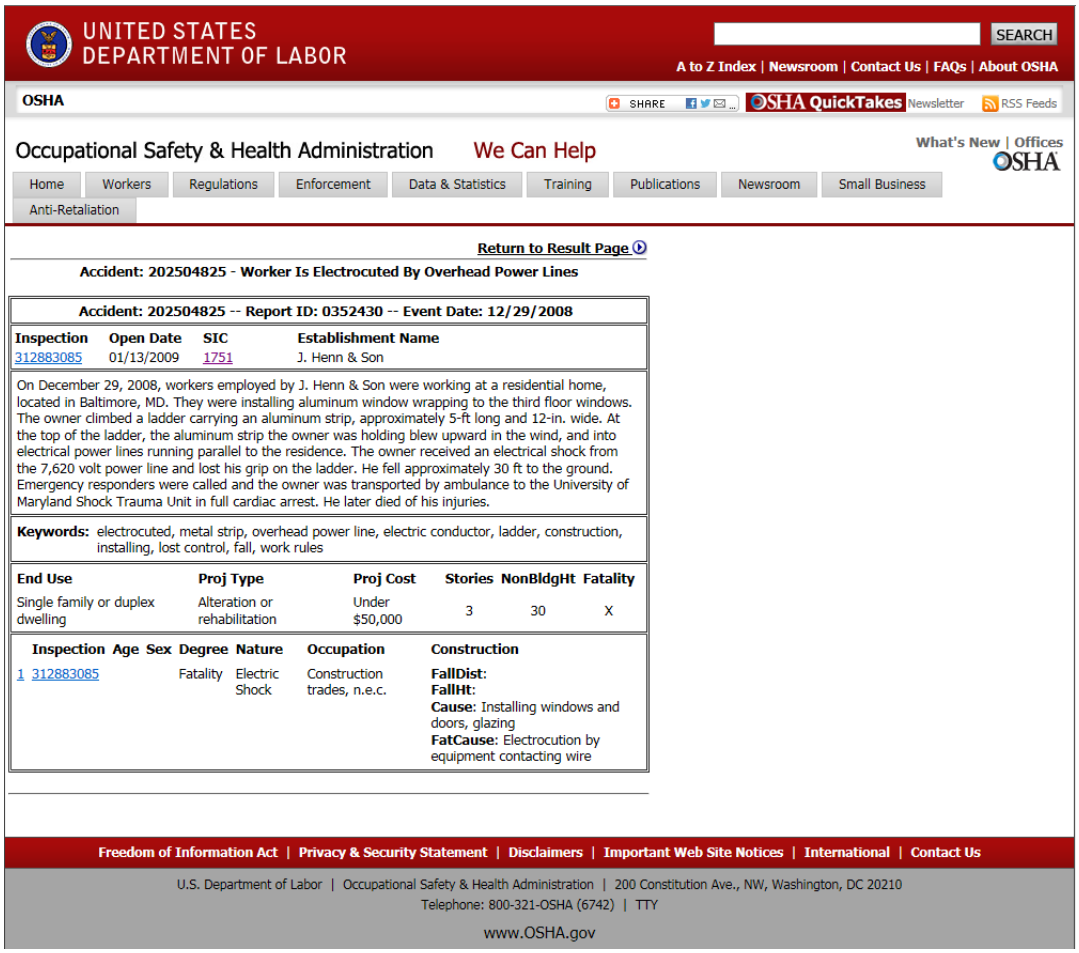

Figure 3: Search Result of Accident Summaries in OSHA's Database

Identify Causal Factors

In this component, causal factors of accidents applicable to construction projects are considered. Accident reports (OSHA database) will be used in developing a preliminary list of factors. Event tree analysis will be performed on the information obtained from the database.

Construct Influence Diagrams

Association rule mining finds interesting associations and/or correlations among set of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely used example of association rule mining is *Market Basket Analysis*. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. These relationships can be used to create influence diagrams (Liu & Chen, 2011).

Establish Probability Tables

When large number of data is available such as the one in the OSHA database, it becomes possible to create conditional probability tables for constructing Bayesian belief networks (BBN) from the Influence diagrams. When conditional probability tables are added to the influence diagrams, a Bayesian belief network is formed.

Estimate Probability of Accidents

The developed Bayesian belief network model is able to predict the probability of accidents happening when the states of conditions (good, poor, etc.) of a construction project are input. Therefore, such a model becomes a predictive model. Based on these predictions, preventive measures may be taken to reduce the likelihood of an injury/accident happening in a construction site.

Conclusion

With the conceptual modeling approach outlined in the paper, a strong case is made to use data mining technique together with a Bayesian belief network to predict the occurrence of accidents and injuries in a construction site. There is clearly an enormous value in adopting such an approach as preventive action or safety intervention can be undertaken to minimize the likelihood of such accidents. Further, textual data in the form of accident summaries have hardly been utilized in construction safety analysis owing to the lack of understanding of converting unstructured data into useful information that is hidden in such data. While data mining has been largely utilized in businesses and manufacturing industries, its full utilization has yet to be achieved in construction, especially in the area of construction safety. While only conceptual model has been presented in the paper, it is the intention of the authors to use the proposed conceptual model with data from OSHA in the next stage of the research.

References

Attoh-Okine, N.O. (2002). "Probabilistic analysis of factors affecting highway construction costs: a belief network approach," *Canadian Journal of Civil Engineering*, 29(3):369–74.

Chang, L., & Chen, W. (2005). Data mining of tree-based models to analyse freeway accident frequency. *Journal of Safety Research, 36*(4), 365-375. doi:http://dx.doi.org/10.1016/j.jsr.2005.06.013

Charpentier, W. (2014). How to obtain OSHA safety reports? http://work.chron.com/obtain-osha-safety-reports-10511.html. (Visited on 10/23/2014)

Cheng, C., Lin, C., & Leu, S. (2010). Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. *Safety Science, 48*(4), 436-444.

Han, J., Kamber, M. (2001). Data Mining: Concept and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, USA 2001, pp. 1-7, 429-433.

Huang, C., & Hsueh, S. (2010). Customer behavior and decision making in the refurbishment industry-a data mining approach. *Journal of Civil Engineering and Management, 16*(1), 75-84. doi:10.3846/jcem.2010.07

Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery, Third ed., http://www.twocrows.com/intro-dm.pdf, (Visited on 9/17/2014)

Liu, K. F., & Chen, J.S.. (2011). Prediction and assessment of student learning outcomes in calculus a decision support of integrating data mining and Bayesian belief networks. pp. 299-303. doi:10.1109/ICCRD.2011.5764024

Luu, V.T., Kim, S.Y., Tuan, N.V., Ogunlana, S.O. (2009). "Quantifying schedule risk in construction projects using Bayesian belief networks." International Journal of Project Management, 27, pp 39-50.

Luxhøj, J.T. (2003). "Probabilistic Causal Analysis for System Safety Risk Assessments in Commercial Air Transport," *Proceedings of the Workshop on Investigating and Reporting of Incidents and Accident.*

Manuele, Fred A. (2008). *Professional Safety*, Vol. 53, Iss. 12; p.32.

McCabe, B., AbouRizk, S.M., and Goebel, R.(1998). "Belief networks for construction performance diagnostics," *Journal of Computing in Civil Engineering, ASCE* 12(2): 93–100.

Sahely B., and Bagley D.M. (2001). "Diagnosing upsets in anaerobic wastewater treatment using Bayesian belief networks," *Journal of Environmental Engineering, ASCE*, 127(4):302–10.

Tischer, T.E, and Kuprenas, J.(2003). "Bridge falsework productivity-measurement and influences," *Journal of Construction Engineering and Management, ASCE*, 129(3): 243–50.