**Analysis of Fire Accident Factors on Construction Sites**

**Using Web Crawling and Deep Learning Approach**

Jaehong Kim1, Sangpil Youm2, Yongwei Shan 3

1 Ph.D. Student, School of Civil & Environmental Engineering, Oklahoma State University, Stillwater, OK, [jaehong.kim@okstate.edu](mailto:jaehong.kim@okstate.edu)

2 Master Student, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, youms@iu.edu

3 Assistant Professor, School of Civil & Environmental Engineering, Oklahoma State University, Stillwater, OK, [yongwei.shan@okstate.edu](mailto:yongwei.shan@okstate.edu)

\*Corresponding author: [yongwei.shan@okstate.edu](mailto:yongwei.shan@okstate.edu)

**ABSTRACT**

The construction site is one of the industrial sites that can be exposed to fatal accidents. As the construction becomes complicated in recent years, on-site safety management has become very important. Fire accidents on construction sites are an important part of the site safety plan. However, because fire accidents have fewer frequencies than other types of accidents such as falls, plans related to fire safety at construction sites have been rarely studied. It has different characteristics compared to other types of accidents, such as the possibility of leading to further damage around fire accidents at construction sites. Without considering the characteristics of these fire accidents, it is unreasonable to evaluate the risk based only on the frequency of accidents. To fill the knowledge gap, this study was conducted to find factors related to accidents at the construction site and includes big data analysis. In this study, web-crawling was used to collect data related to accidents at construction sites in the past 20 years. Based on the collected data, the authors found the frequency of exposure of keywords related to accidents and provided similarity between related keywords through word embeddings and network analysis. In addition, it was visualized through the Uniform Manifold Approximation and Projection (UMAP) to show in multi-dimensional. The contribution of this study is important to reduce the risk by identifying factors related to construction site accidents, including fires. Through the visualization of keywords related to accidents, potential risks can be predicted in advance. Also, this study can play an important role in establishing the regulations necessary to increase safety in construction sites.

**KEYWORDS:** Construction sites, Safety, Fire accidents, Web crawling, Deep learning

**Introduction**

Recently, due to the development of various construction technologies, skyscrapers and large-scale construction projects are under construction. The development of construction technologies has shortened the construction period and improved the convenience of residents, but the safety of the construction site is developing slowly. The Bureau of Labor Statistics (BLS) recently released Census of Fatal Occupational Injuries (CFOI) in 2018, and the construction industry topped the list with 1,008 worker deaths. In addition, 5,250 fatal work injuries were recorded, a 2% increase from 2017. The fatal work injury rate was about 3.5 per 100,000 full-time equivalents (FTE) workers. As such, various types of accidents occur on construction sites, including fires. Fire accidents are greatly affected by the external environment, making it difficult to control and prevent them. This is especially dangerous because fire safety equipment has not been completed on construction sites.

Most construction accident-related research focuses on the frequency of accidents in the past. However, there are limitations when analyzing only the frequency of accidents. For example, fall accidents on construction sites are fatal but are unlikely to lead to additional accidents to the surroundings. The frequency of fire accidents is lower than that of fall accidents, but it can have a greater impact on the construction site. In other words, in order to improve the safety of the construction site, it is necessary to consider how the accident type affects the construction activity and the surrounding environment. The impact of each type of accident on the construction project and the surrounding environment involves various external conditions, and it is difficult to quantify these effects. For this reason, most construction accident studies focus on the frequency of accidents. To compensate for this limitation, this study investigated the frequency of media exposure. The media has characteristics that deal with social issues that have a great influence around them, and this is expressed in articles. Rather than minor accidents on construction sites, fatal accidents that can affect the surroundings are more likely to be exposed to the media. In addition, the articles provided by the media are organized in a similar format, which is efficient for many researchers to use as low data.

To enhance construction safety, including fires at construction sites, this study collected articles on construction site accidents reported in The New York Times over the past 20 years. The web-crawling method is used for efficient and accurate data collection. Using collected articles, we analyzed the frequency of keywords related to construction site accidents. In addition, similarity and relationships between related keywords were analyzed through word embedding and network analysis. To intuitively visualize words that have high relationship between fire accident and fall accident, the Uniform Manifold Approximation and Projection (UMAP) method is applied. In this study, the frequency of media exposure of construction site accidents was conducted to analyze accidents on construction sites and to present a new perspective to improve safety. In the case of fire accidents, the media frequency was higher than the actual frequency when compared to the fall accident. This shows the possibility that fire accidents may have a greater impact on the surroundings than the actual frequency. The results of this study can be used as data to establish new safety regulations for construction sites. In addition, it is possible to intuitively check the factors related to each accident type and help to institutionalize them.

**Background**

***Construction site safety***

The construction industry always considers safety, but the fatality rate at construction sites is always high(Abdullah and Wern 2011). The fatality rate of the construction industry was found to be the fourth highest after agriculture, mining, and transportation. According to the Occupational Safety & Health Administration (OSHA), 20.5% of fatal workplace accidents occurred on construction sites in 2014 (Zou and Zhang 2009, Hosseinian and Torghabeh 2012). Various studies have been conducted to analyze accidents on construction sites. The OSHA provided that the main reasons for the construction worker's fatal accidents are by falls, struck by an object, electrocutions, and caught-in/between. One study suggested that the seven causes of accidents are lack of training, deficient enforcement of safety, no provided safety equipment, unsafe methods, unsafe site condition, no use of safety equipment, safety ignorance, and isolated sudden deviation (O'Toole 2002). Other studies have shown that construction accidents have occurred due to improper safety management due to a lack of safety awareness by managers (Zhou, Fang et al. 2008). A study investigating the construction industry accident in Malaysia suggested that the accident on the construction site was due to the manager's fault and the attitude of the worker (Abdullah and Wern 2011). In another study, the cause of the accident was identified by analyzing 40 deaths at the construction site from 2003 to 2008(Ling, Liu et al. 2009). The findings indicated that unskilled workers and lack of safety training are the main causes.

In particular, it is important to prevent fire accidents on construction sites, as they are likely to cause secondary accidents such as collapse, burial, and explosion. According to a related study, fires on construction sites are mainly caused by the transfer of sparks to the surrounding insulation during welding (Lee 2012). In order to evaluate the fire hazards at construction sites, there are also studies evaluating fire hazard tracking systems and related training (Hui, Yongqing et al. 2012). In addition, several studies have been conducted on effective evacuation in case of fire at the construction site (Ingason, Lönnermark et al. 2010, De-Ching, Shen-Wen et al. 2011, Jeong, Lee et al. 2014). However, because these studies are based on case studies and specific projects, they are limited in application to general construction sites. In order to prevent these accidents, establishing relevant policies is one of the best ways. It is important to establish appropriate safety regulations for the construction, as well-established policies for construction safety can directly reduce accidents(Aires, Gámez et al. 2010).

***Web crawling***

Web crawling is a technique for systematically browsing the web for the purpose of web indexing (Paul, Mitra et al. 2017). It is often used for tracking web documents on the Internet to effectively collect the information the user needs. There are also studies to properly use online data for research purposes (Massimino 2016). This study provided guidelines on the skills and responsibilities required to collect online data. Because the data on the web is very huge, collecting web data manually can take a lot of time, and the accuracy can be reduced. However, web crawling technology automatically rotates the web server to repeatedly collect information that fits the purpose. These web crawling technologies are used in a variety of fields, especially in research involving decision models and prioritization (D’Haen, Van den Poel et al. 2016, Guy, Schwartz et al. 2019). Recently, research on safety and security through web crawling has been conducted (Morgan, Tietje et al. 2020). Web crawling technology has begun to be used not only for text but also for image analysis (Ali, Ali et al. 2018). In addition, research was conducted to utilize real-time data on the web rather than past data (Kim, Kim et al. 2019). To use web crawling in the research, researchers must set a clear target. Target is primarily a website, and researchers should make sure that web crawling technology is available on selected websites. This is very important because there are issues with data security. Next, the researcher determines the scope and frequency of data collection.

Traditionally, the main purpose of using web crawling in the construction field consists of two main parts. The first is construction material management and optimization. To improve the efficiency of construction material management, related researches used web crawling technology to collect relevant information and provide automated processes. (Yang, Wi et al. 2018, Hong, Lee et al. 2019). It also used web crawling to manage massive documents in construction projects. An example is a web crawling used to develop a system that collects text data with the latest information from the construction market and automatically assigns it to each applicable construction document (Moon, Shin et al. 2018). As above, the use of web crawling in the construction field was limited. Recently, this technology is used in various fields related to construction. An example is a study that collects a variety of geographic information on the web and provides a model to predict air emissions from heating (Lopez-Aparicio, Grythe et al. 2018). However, few studies have analyzed the factors related to the safety of a construction site using web crawling. In this study, web crawling technology was used to find factors related to site safety, which may suggest a new approach to improving construction site safety. The authors used the python language and libraries to implement web crawler.

***Word embedding and network analysis***

Word embedding is a technique that provides a way to express similar words with the same meaning through data analysis (Yin and Shen 2018). This is a new way to represent words and documents and is one of the key breakthroughs in deep learning (Lai, Liu et al. 2016). Each word has a unique vector value through analysis using an embedding layer or hidden layer (Yoshioka and Dozono 2019). This method of analysis was mainly used as a new method for analyzing text or documents (Shao, Zhang et al. 2017, Zhang, Huang et al. 2018). This methodology overcomes the limitations of a one-hot encoding analysis, which was widely used for text analysis (Rodríguez, Bautista et al. 2018). In the one-hot encoding method, since all words are composed independently, it is impossible to measure the similarity of each other. However, in the word embedding method, each word is expressed as a dependent relationship, and similarity can be measured (Rekabsaz, Lupu et al. 2017). The word embedding method is currently used mainly in text analysis and natural language processing (NLP). The main models using word embedding are Word2vec (Church 2017), GloVe (Hindocha, Yazhiny et al. 2019), and fastText (Choi and Lee 2020), and Word2vec was used in this study. Word2vec Model was created and published in 2013 by a research team of Google. This model has been subsequently analyzed and updated by other researchers. Word2vec has very efficient performance and accuracy (Ombabi, Lazzez et al. 2017). In addition, it has a lot of google news pre-trained data, so it is suitable for transferring learning based on it (Khatua, Khatua et al. 2019).

In this study, network analysis was used to analyze not only the vector value of each word, but also the relationship between linked words. Network analysis is one of the methodologies for finding the relationship between various types of low data, and basically consists of nodes and edges (Smith and Gorgoni 2018). Links between words can be expressed as graphs or structures, and relationships between words can be analyzed using them. Through the analysis of various centralities such as eigenvector centrality, degree centrality, and betweenness centrality, it is possible to check which nodes are important (Risselada, Verhoef et al. 2016). Also, the correlation between nodes can be analyzed by calculating the Jaccard coefficient that can measure the similarity of sample data (Bag, Kumar et al. 2019). This network analysis is widely used in recent safety issues, such as pandemic research. There are studies to effectively prevent pandemic by analyzing the paths and conditions in which the pandemic spreads using network analysis (Sandhu, Gill et al. 2016).

**Methodology**

***Web crawling***

This study implemented a web crawling method for data collection. The web crawler is the main operator to conduct web crawling. The web crawler usually traverses web pages by using a recursive algorithm and then it goes over a certain range which researcher confines. The crawler stores data in the format of the data structure (Mahto and Singh 2016). Researchers arrange data for using efficiently with their study. In this study, the author uses Python 3.7. Since Python has many libraries that can easily apply to the various areas, Python is able to be one of the most popular programming languages nowadays. Python also has an effective library for web crawling, and it is widely used for web crawling. To begin with, the author needs to set the scope in which crawlers should travel. This research collected data from The New York Times. The New York Times is the top three media company in terms of newspapers by circulation and thus it suffices representative of data that needed in this study. In addition, even though some media companies restrict crawling or limit the amount of data, The New York Times has generous terms for crawling the data. The New York Times provides a basic search environment that helps to narrow the range of search. This research utilizes the search term “construction accident” and retrieves the most relevant 1,010 data within 20 years. Among 1,010 data, we get rid of the irrelevant data that does not have an article format. Since different formats such as blogs and interactive documents consist of irregular structures, these formatted articles cannot be analyzed. Table 1 shows the scope of source data.

Table 1. Source data information

|  |  |
| --- | --- |
| **Category** | **Information** |
| Issued duration | 2000.01.01 – 2019.12.31 |
| News source | The New York Times |
| Total number of articles | 1010 |
| The number of relevant articles | 861 |
| Total number of words | 453,283 |

There are three vital libraries for web crawling: Selenium, HTMLParser, and Beautiful Soup. Selenium is the set of tools that assist the development of test automation for a web-based application (Mustika 2018). Selenium automatically traverses the web pages and stores data within the limit range. In this study, the author stores the URL of each article from the web. Next, HTMLParser is used for parsing HTML files in each article. With the broken HTML, Beautiful Soup scrapes data the researchers desire to utilize. Beautiful Soup is a Python package that analyzes HTML/XML, as it extracts and edits information in the web pages. It provides a developing friendly environment for building a prototype and mining the data (Zheng, He et al. 2015). In this research, the authors parse the data from broken HTML and sort out the part of the body, title, and date. The structure of the article provided by The New York Times has a tag for body and title. It also provides a meta part that represents the metadata of the article and the meta part gives date information. Collected data are used for word embedding in the next part of this study. This is the process that prepares suitable input for word embedding. Word embedding requires minimum changing from original data sources. Hence, the author minimizes data preprocessing, especially, in this study, we should proceed Natural Language Process, called NLP. NLTK library is used for conducting NLP. We exclude “English Stop word” that contains meaningless words such as “that”, “is”, and lower all letters in every article.

***Word Embedding (Word2Vec)***

Word2Vec is one of the powerful methods to implement word embedding. Since word2vec assigns a vector to each word, word2vec highly supports the analyzing words in diverse NLP tasks. Word2Vec resolves many problems that occur when researchers utilize words in their research. Before the presence of word embedding algorithms such as word2vec, researchers usually used one-hot encoding to analyze words. One hot encoding converts each word to specific (associated with) values(numbers) that differ from every word. Even though this helped them to conduct NLP, the researchers could not answer the relation of each word because every word in one-hot encoding is independent. Word2vec resolve this problem, as it adds the embedding layer into the model. Word2vec follows a deep learning approach that probabilistic prediction by using a hidden layer. It represented by as follows,

, where d is the embedding dimension and V is vocabulary.

Since Word2Vec conducts unsupervised learning that is trained on raw text data, it creates word embedding by figuring out the maximum likelihood of word prediction from their context (Ling, Dyer et al. 2015). There are two models for implementing Word2Vec: Continuous Bag-of-Words model (CBOW) and skip-gram model in figures below. Word2Vec algorithm compute cosine similarity to find out similarity or dissimilarity,

In this study, cosine similarity plays a significant role in analyzing words, which helps to compare the important words and construct relation among words. Continuous bag-of-words (CBOW) predicts word by considering context, which means target word is predicted from surrounded words of the target word and thus it can have one output vector.

***A close up of a map

Description automatically generated***

Figure 1. Continuous bag-of-words (CBOW)

The input vector is one-hot encoded. The weights between the input vector and the output vector are represented by V N matrix . The author formulates k-th row of **W** as **h.**

There is different weight matrix ; between embedding layer to the output layer. With these weights, this model computes a score for each word,

**h**

Through these formulae, the author can calculate the posterior distribution of words in the vocabulary as follows (Rong 2014),

The skip-gram model works oppositely to CBOW. In skip-gram, the model predicts context, C has given input words. That is, skip-gram computes the possibility of which prediction of the target words in several contexts. Skip-gram model also use weighted matrix, denoting the input vector of the word on the input layer, and it has the same definition of embedding layer outputs h,

At the output layer, unlike the result in CBOW, it has C multinomial distributions for the outputs, we can see its workflow as follows,

A close up of a map

Description automatically generated

Figure 2. Skip-gram

Each output is calculated by using the different weight matrix ; . The posterior distribution can be,

, where is the j-th word on c-th part of the output layer. is the true c-th word in the output context. In other words, is the input data of the j-th unit on the c-th part of the output.

Even though two models both are widely used for conducting word2vec, these models can be optimized depends on the goal and direction of the research. The skip-gram model can be effective when the research figures out fewer frequency words. On the other hand, CBOW is more suitable for studies that concentrate on high-frequency terms. In this study, since the authors focus on important keywords which can explain accident in the construction site and frequently appeared word, the authors implement Word2Vec with CBOW model. There are several parameters for conducting word2vec. The author restricts the minimum count of words, 200, that is, the author only considers words appear at least 200 and the dimension of the words vector is 400. Except for these two-parameter, word2vec is operated with the default setting.

***Network analysis***

Networks are a versatile method to show and analyze simple and complex interaction and thus they are used to study in widely diverse areas. The network representation is simple but rigid since many parts of a specific system are sorted out and concentrate on the interaction among its elements (Menczer, Fortunato et al. 2020). The network is represented with nodes and edges which connect nodes to each other. In this study, five keywords and similar words that have over 0.5 cosine similarity with the keywords. The author measures the Jaccard coefficient between each keyword. Jaccard coefficient is able to be described by the Jaccard Index. Jaccard Index is applied to compare similarity, dissimilarity and the formula to represent it as follows.

Jaccard coefficient calculates the result of the division between the number of features that can be seen to all divided by the total number of features (Niwattanakul, Singthongchai et al. 2013). The authors figure out the interaction over five keywords and examine the Jaccard coefficient between two keywords over ten combinations of keywords set.

***Uniform Manifold Approximation and Projection (UMAP)***

With the result of the Word2Vec, the authors generate 400-dimensional word vectors. As the author needs to analyze word in word vectors, visualizing in low dimension space can be a good way to figure out word vectors and thus UMAP is the state of the art technique for dimensional reduction. Dimension reduction creates low dimensional space without loss of structure in high dimensional space. According to the related UMAP study, its techniques widely used in various fields and larger sizes of data (McInnes, Healy et al. 2018). They describe UMAP in theoretical and computational view. In the theoretical view, Riemannian geometry and algebraic topology are theoretical ground that constructs UMAP. UMAP operates on weighted graphs, and it uses k-neighbors to cluster groups. UMAP has two phases with a graph learning algorithm. The weighted k-neighbor graph is created in the first phase, and a low dimensional layout of the graph is calculated in the second phase. UMAP is usually compared with t-SNE which widely used for dimensional reduction. UMAP performs faster and more efficiently. To be specific, performing with UMAP more correctly preserves global structure over performing with t-SNE, and UMAP performs faster than t-SNE. In this study, the authors conduct dimensional reduction with UMAP and visualize word vectors in two-dimensional space. Every word have their own vectors and they are reduced in two dimensions, and then we annotate every dots on the two-dimensional graph. Since the author focus on five keywords, annotation of keywords have a larger size of the font and similar words with five keywords represented in a different color. In other words, the author can easily examine keywords and similar words with keywords that are in the same cluster.

**Results**

***Preliminary analysis***

Through web crawling, a total of 1,010 New York Times relevant articles were found. Basically, “construction accident” was used as a search term on the New York Times web page, and the top 1,010 relevant articles were retrieved. The author set the URL of the web page as the default URL and performed automated crawling using Selenium. Selenium is a set of powerful software tools working with many web browsers, programming languages, and web-based testing frameworks (Wang and Du 2012). This tool has the advantages of open source, easy access, and flexibility, and easy implementation. URL of each article obtained through this tool is implemented by parsing data through HTML parser and BeautifulSoup. HTMLParser is a library used as a basis for parsing text files formatted in HTML and XHTML. Beautiful Soup is a Python library that parses broken HTML. This creates a parse tree with a meaning almost similar to the original document (Vargiu and Urru 2013). These libraries are used in many studies and have sufficient reliability. The authors checked irrelevant data for data cleaning. Of all the data collected by web crawling, only document type articles were used for analysis, and 149 articles of the interactive document and blog type were excluded from the analysis. The article types excluded from the analysis are not valid for scraping because the structure is irregular. Therefore, a total of 861 articles were analyzed. After completing data cleaning, the authors classified articles according to composition. It was confirmed that the articles are generally composed of title, date, and body, and the text data included in the article was classified in consideration of this. The Beautiful Soup library was used to classify and scraped the parts needed for analysis.

***Basic statistical analysis***

Based on the selected data analyzed through preliminary analysis, the Natural Language Toolkit (NLTK) library was used to conduct the Natural Language Process. By using the NLTK library, the authors removed 'English stopwords' and changed all words to lower case. The definition of 'English stopwords' is unnecessary words that are filtered before and after the processing of natural language data. This is the data processing required for further analysis such as word2vec, and the minimum preprocessing is performed within the range that does not transform the original data as much as possible. As one of the basic statistics, frequency analysis was conducted for each keyword. The frequency of words appearing in each article was measured, and even if repeated twice or more in one article, it was measured once. Among the words satisfying this condition, 5 keywords related to this study were selected. The five keywords are fire, fell, collapsed, building, and people. The selection criteria considered the types of accidents on construction sites and keywords representing fire. In many studies, three major elements of building fire are defined as a fire, building, and people. In addition, fell and collapsed are the most frequent types of accidents at construction sites. All keywords have sufficient frequency to be used in this study.

According to the analyzed results, buildings and people were the most frequent. The fire, fell, and collapsed keywords related to the type of accident on the construction site showed relatively similar frequency. Since fell and collapsed are similar words, it may be reasonable to compare fire with the frequency of combining them. When comparing the combined numbers of fell and collapsed, the fire has a frequency of 27% and fell and collapsed have a frequency of 73%. This result differs significantly from the Bureau of Labor Statistics (BLS) analysis of construction site accident frequency. According to the BLS report, among the accidents at the construction site, fell and collapsed accidents account for about 40% of all accidents, and fire accidents are 2%. When converted to 100%, fire accidents have a frequency of 5% in fell-related accidents. There is a large gap in 27% of the results of this study and 5% of the BLS report. The possible reasoning about this difference can be explained by the characteristics of the media that the author described in the intro. Fire-related accidents may be more exposed to the media than fell-related accidents, which may explain that fire accident on construction sites have a greater effect than fell-related accidents.



Figure 3. Frequency of keywords

***Word embedding with Word2vec***

In this study, cosine similarity was used to provide a similarity between keywords. Cosine similarity is a concept mainly used to measure the similarity between words, and Word2vec also uses it. In order to express the distance between vectors in multidimensional space as a cosine value, this concept was used. A total of 10 combinations were made by classifying each of the 5 keywords selected above, and the similarity of each was calculated. In order to confirm the reliability of this study, cosine similarity among the most similar words among the analyzed words was first checked. The words chosen by the author are fell and collapsed. The dictionary meaning between these two words is very similar, so if the results of this study are reliable, the similarity between these words should be high. The similarity between the two words was 0.951, which was calculated much higher than the similarity between the other keywords. Through this, the author can confirm the reliability of this study.

In the results related to the fire keyword, the fire keyword was more similar to the building keyword than the people keyword. The similarity between fire and building was 0.525, and the similarity between fire and people was 0.331. It can be interpreted that fire showed a much higher degree of similarity to words related to the building than words related to people. For the fell and collapsed keywords, the similarity between each keyword and the people and building keywords was very similar. In the case of the fell, the similarity with the building keyword was 0.443, and the similarity with the people keyword was 0.486. This result contrasts with the fire keyword and fell has a higher similarity to the people keyword. Also, the similarity between people and buildings was calculated as a negative value, indicating that the correlation between the two keywords is not great. This table shows similarity between keywords.

Table 2. Similarity between keywords

|  |  |
| --- | --- |
| **Relationship between keywords** | **Similarity** |
| Fire and Building | 0.525 |
| Fire and People | 0.331 |
| Fire and Collapsed | 0.691 |
| Fire and Fell | 0.728 |
| People and Fell | 0.486 |
| People and Collapsed | 0.510 |
| People and Building | - 0.144 |
| Building and Collapsed | 0.554 |
| Building and Fell | 0.443 |
| Fell and Collapsed | 0.951 |

The table below shows 5 selected keywords and 20 words with high similarity. It was written based on the cosine similarity values, listed in order of similarity. In the results, the fell and collapsed keywords were calculated to be most similar to each other. This shows the reliability of this study, as well as the calculation of cosine similarity between keywords. In terms of similarity word results related to the type of accident on the construction site, such as fire, fell, and collapsed keywords, the words Monday and Friday are included in the list. This shows results consistent with research on the construction industry's 'The distribution of injuries'. In a related study, the day of the week where injuries were most likely to occur in the field was found to be Monday, which is consistent with the results of this study (Wigglesworth 2006).

In addition, unlike the results of other keywords, the fire keyword result showed a high degree of similarity with the word “night”. Generally, the possibility of spread increases when a fire occurs at night. Because, most commercial buildings are less likely to stay after working hours, and even residential buildings have difficulty recognizing fire during sleep. In most fire accidents, the probability of the spread of a fire tends to increase as the initial detection of the fire is delayed. In particular, a construction site may have a small number of employees staying overnight for the monitoring or may be empty. The characteristics of these construction sites can lead to the rapid spread of fires in fire accidents. Also, during the construction phase of the building, it is more difficult to recognize fires because safety equipment such as fire and smoke alarms have not been completed (Hamid, Yusof et al. 2003). This was highlighted once again through the results of this study. There are many words that have a meaning of ‘administration’ or ‘inspection’ in words with high similarity to the building keyword. The top five words with the most similarity to the building keyword are department, inspectors, issued, commissioner, and city, and these words tend to have a common meaning. These results show that articles related to accidents on construction sites mainly deal with building inspection and management issues. In addition, words with high similarity to the people keyword have many words related to action or behavior.

Table 3. Top 20 list of similar words of keywords

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fire** | **Fell** | **Collapsed** | **Building** | **People** |
| police | collapsed | fell | department | died |
| investigators | side | debris | inspectors | injured |
| authorities | floor | floor | issued | killed |
| hospital | debris | ground | commissioner | men |
| injuries | ground | west | city | accident |
| chief | street | steel | cranes | authorities |
| night | Friday | street | contractor | debris |
| floor | avenue | death | violations | train |
| man | death | side | tower | injuries |
| dead | morning | site | crane | dead |
| Monday | worker | Friday | investigation | workers |
| Friday | injuries | march | office | ground |
| driver | Monday | wall | equipment | Monday |
| investigation | injured | injuries | site | crash |
| truck | dead | authorities | district | march |
| officials | wall | Monday | charges | residents |
| death | authorities | avenue | investigators | members |
| worker | steel | tower | company | Tuesday |
| fell | crash | Tuesday | officials | Thursday |
| working | died | injured | mayor | cars |

***Network analysis***

Network analysis is an analysis method that describes the relationship of data with nodes and edges. By using this network analysis, it is efficient to interpret the relationship between nodes of data. In this study, keywords are basically nodes, and words with a cosine similarity of 0.5 or higher are connected by edges. The text data of this study consists of a total of 136 nodes and 353 edges. Table 2 shows the number of nodes each keyword has, and the collapsed keyword has the most nodes. There was no significant difference in the number of nodes each keyword had.

Table 4. Network analysis information by keywords

|  |  |
| --- | --- |
| **Keywords** | **Connected Nodes** |
| Collapsed | 91 |
| Fell | 88 |
| Fire | 80 |
| Building | 50 |
| People | 50 |
| Total nodes: 136 / Total edges: 353 | |

In addition, the Jaccard coefficient between each keyword was calculated and provided. The Jaccard coefficient values between keywords are shown in Table 3. The Jaccard coefficient is a statistical value used to measure similarity and diversity of sample data. Through this, the network of each keyword can be expressed as one unified network. This network graph is visualized in Figure 4. This entire network has different sizes for nodes and annotations based on degree. This means that the more nodes are connected (the higher the degree), the larger the size of the node. Among the keywords in this study, collapsed has the largest degree. In addition to the five keywords, the word in which the node size is noticeably larger is'death'. This shows that the death word has a high degree except for the keyword in the network in Figure 4.

Table 5. Jaccard Coefficient between keywords

|  |  |
| --- | --- |
| **Relationship between keywords** | **Jaccard Coefficient** |
| Fire and Building | 0.262 |
| Fire and People | 0.262 |
| Fire and Collapsed | 0.513 |
| Fire and Fell | 0.541 |
| People and Fell | 0.484 |
| People and Collapsed | 0.424 |
| People and Building | 0.020 |
| Building and Collapsed | 0.205 |
| Building and Fell | 0.160 |
| Fell and Collapsed | 0.772 |

A close up of a logo

Description automatically generated

Figure 4. The network of keyword's similar words

***Visualizing with UMAP***

In this study, the results analyzed through Word2vec were visualized in Uniform Manifold Approximation and Projection (UMAP). UMAP is one of the novel manifold learning technique for dimension reduction. The result of this method is a practical scalable algorithm that applies to real text data (McInnes, Healy et al. 2018). If the working principle of UMAP is approached from the computational view, it can be divided into two main steps. In the first step, the graph is constructed, as in the k-neighbor graph based algorithms, and in the second step, the low-dimensional layout of this graph is calculated. Through the UMAP technique, it is possible to express the physical distance between two vectors in two-dimensional space. This allows intuitive visualization of vector values and is effective in grouping between vectors. Figure 5 shows the overall UMAP graph for this study. In order to increase the discernment of the graph, each keyword is expressed in a different color, and the range of each keyword is indicated by gradation. The gray dots on the UMAP are words with low similarity to keywords. As shown in the figure below, fell and collapsed were almost overlapped on the UMAP. This means that the similarity between the two keywords in UMAP is very high. In the case of the fire keyword, the range appears wider than other keywords, and there is an intersection with the building keyword. For the fell and collapsed keywords, the people keyword appeared to be closer than the building keyword.

A picture containing text

Description automatically generated

Figure 5. Word2vec result overview with UMAP

The figures below show the UMAP visualization limited to each keyword. Through this, it is possible to know how similar words with each keyword are expressed at the UMAP and the distance between each word. By using this, this study can provide information grouped according to the similarity of words.

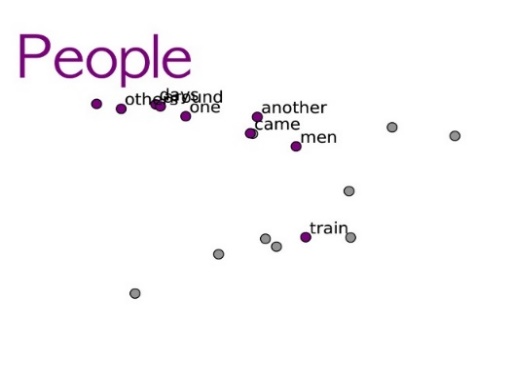
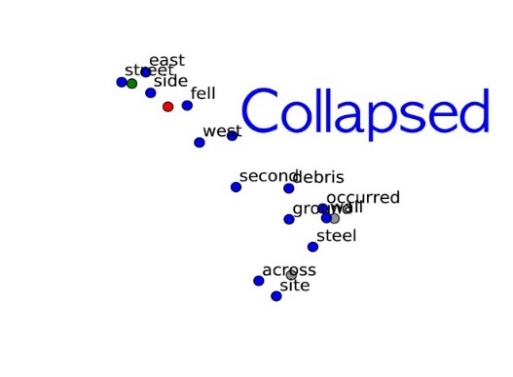
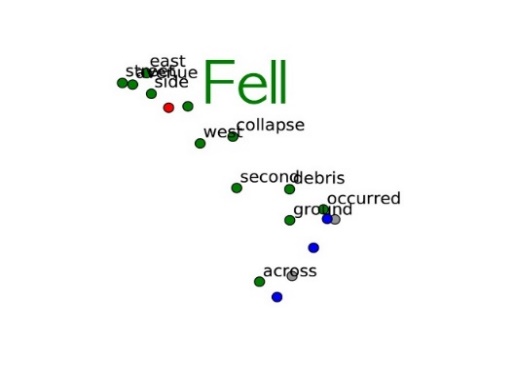
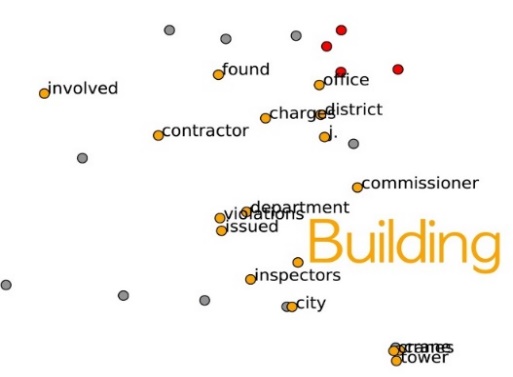
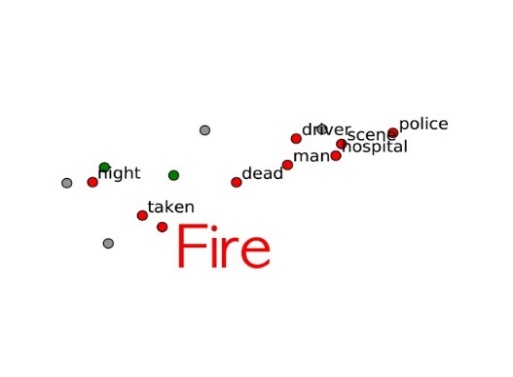


Figure 6. Subgraphs for keywords

**Discussion**

This study explored the relationship between factors related to fire accidents at construction sites through web crawling and deep learning. For data collection to be used in this study, web crawling was used to collect enough news articles related to accidents on construction sites. A total of 1,010 articles were collected using the keyword search method, and 861 articles were finally used after data cleaning. In addition, network analysis and word embedding techniques were used to investigate the relationship between factors related to accidents on construction sites. Through this analysis, the similarity and relationship between each keyword were defined and visualized by UMAP. To verify the reliability of this methodology, the relationship between similar words was analyzed.

First, keyword analysis was conducted with 861 articles collected by web crawling in this study. Since the authors set the search condition when using web crawling as a “construction accident,” most articles focus on accidents on construction sites. For the next step of data analysis, five keywords related to accidents and safety on the construction site were selected, which are fire, fell, collapsed, building, and people. Since the frequency of these keywords must be sufficient for further analysis of the collected data, the authors checked the frequency. The smallest frequency was fire, measured 192 times. The keyword with the highest frequency was building, which was measured 514 times. Of the 5 keywords, the minimum frequency was 192, which was sufficient for further analysis. The most interesting thing about frequency analysis is that the frequency of fire is relatively high. As mentioned earlier, the frequency of fire is 192 times, which is not much different from the frequency of fell and collapsed. The frequency of fire was about 10% smaller than fell and about 20% smaller than collapsed.  In the BLS report, the frequency of fire accidents is investigated within 2% of the total construction accidents, so it differs from the results in this study. The difference between the frequency of actual accidents and the frequency of media exposure can be used as evidence to confirm the great impact of fire accidents on construction sites. Of all the accidents on the construction site, fire accidents have a low frequency, and thus fire accidents are excluded from the list of major accidents on the construction site. OSHA announces "Fatal Four" on the construction site every year. The "Fatal Four" section of this report includes accident types related to falls, struck by object, electrocutions, and caught-in/between. As OSHA publishes reports and statistics related to 'Fatal Four' every year, many stakeholders on the construction site can check this and reflect it on the construction site. However, because the report is based on the frequency of accidents and deaths, there are limitations. Fire accidents are excluded from this “Fatal Four,” but the results of this study confirmed that the fire accident had a similar impact on the media as the main types of accidents on the construction site. Types of accidents in which the media frequency is higher than the actual accident frequency, such as fire accident, means that the impact of one accident frequency is greater than that of other accident types. This result is in line with the fire accident characteristics. Fire accidents are more likely to lead to additional accidents than other types of accidents on the construction site. Due to the spread of fire, it can affect the surrounding buildings and roads. Currently, most research and reports related to the safety of construction sites focus on the frequency of accidents. If researchers consider the characteristics of the accident types and their surrounding influence along with the frequency of the accident, this could be a new approach to improving construction site safety.

In this study, the number of connected nodes of each keyword was found through network analysis. The three keywords related to accidents on the construction site (collapsed, fell, and fire) had a similar number of nodes. The fire had 80 connected nodes, which is about 12% less than the collapsed with the most nodes. Coefficient values between keywords were calculated by network analysis. The coefficients of fell and collapsed were the highest, and people and buildings had the lowest. The fire keyword has the same coefficient between buildings and people and fell and collapsed keywords have higher coefficients with people than building keyword.This study also conducted an analysis using Word2vec, one of the word-embedding models. Through this model, the similarity between each keyword was calculated, and the reliability of this analysis was verified. The reliability of the analysis method of this study was confirmed by calculating the cosine similarity of fell and collapsed, which have similar meanings. According to the results of this study, the fire keyword showed higher similarity to the building keyword than the people keyword. This can be explained by the fact that fires on construction sites have a higher relationship with building factors. In the case of fell keyword, the similarity with the people was higher than the building. This shows that in the case of a fell accident, the impact of people's activity or behavior may be greater than that of building factors. These results can be used as relevant data when establishing safety regulations on construction sites. For example, when trying to improve the fire safety of a construction site, it may be more efficient to provide regulations considering factors related to building. In addition, in the case of fell-related accidents, it may be effective to establish rules that affect people's activity or behavior. The list of words with high similarity to the three keywords related to accidents (fire, fell, collapsed) commonly includes Monday and Friday. This result is consistent with statistical data related to accidents on construction sites. According to a related study, workers' injuries on construction sites were the highest on Monday. Unlike other types of accidents on construction sites, the word 'night' has a high degree of similarity in the fire keyword. Construction sites tend to have few occupants after work hours compared to other building types. This may not be sufficient for the initial detection and response to the most important in extinguishing a fire. In the case of a construction site, safety facilities related to fire are not completed, so systems and regulations to compensate for this are necessary. Additional equipment or monitoring systems, such as fire and smoke alarms, are necessary for fire safety in construction sites. In addition, regulations must be supplemented so that such fire safety equipment can be installed effectively and compulsorily on construction sites. Also, words with a high similarity between the fire and building keywords have words related to ‘inspection’ in common. This shows that inspection is an important issue in construction site fire accidents and safety articles. The results of this study remind us once again that periodic and detailed inspection is essential to improve the safety of construction sites.

In this study, the results through Word2vec analysis were visualized and provided as UMAP. When the results are visualized in 2D space through UMAP, the results can be interpreted more intuitively. UMAP expresses words that show relatively high frequency among all the languages used in the article. In the whole UMAP graph with all keywords marked, the five keywords the authors selected are concentrated in similar locations. This shows that the five keywords are organically related and can affect each other. In detail, the fire keyword was expressed close to the building keyword in UMAP, and some ranges overlapped. The fell and collapsed keywords existed in almost the same range and were relatively closer to the people keyword than the building keyword. Also, by providing sub-graphs of each of the five keywords through UMAP, it shows words with similar characteristics. By providing UMAP visualization graphs, the results of the research can be expressed in graphs, not texts, which can enhance the understanding of related researchers.

**Conclusion**

Although many studies have been conducted on the safety of construction sites, related studies have limitations in analyzing based on the frequency of accidents. Because of this, the main focus of these studies is limited to accident types such as falls, which have a high frequency on construction sites. In the case of fire accidents on construction sites, it has a low frequency, but the impact of one fire accident may be greater than that of other accident types. When analyzing accidents on the construction sites, it is necessary to consider not only the frequency of accidents but also various factors such as the surrounding impact and probability of spread. In order to accurately analyze the influence of factors related to accidents on construction sites, it is necessary to consider not only using existing statistical data but also introducing and analyzing new methodologies. In this study, the frequency of media exposure as well as the frequency of accidents on construction sites were investigated to find the impact and factors of each type of accident. New approaches such as web crawling, network analysis, and word embedding using deep learning were introduced and used in this study. By extending this study, the authors provided the correlation and similarity of construction site accident types and related factors using network analysis and word embedding methodology. Furthermore, the results of word embedding analysis were visualized through UMAP. By expressing text data in a vector space, the results were visualized more intuitively. Through this study, it was possible to investigate the media exposure frequency of keywords related to accidents on construction sites and to identify factors affecting accidents on construction sites. However, it is necessary to analyze by adding various keywords as well as the keywords analyzed in this study. In addition, analysis of other media that can investigate the frequency of media exposure is required in addition to the news article. In this study, an analysis was conducted to find the relationship and similarity between representative keywords, but analysis of sub-words may be required based on sufficient data collection. Further studies may consider various types of accidents and factors of the construction site. Also, new methodologies in other fields, such as web crawling and deep learning, must be actively introduced to suggest ways to improve the safety of construction sites. By investigating the frequency of media exposure and exploring the relationships between elements for accidents at construction sites where various factors are compounded, it is possible to make meaningful data in developing regulations that consider the characteristics of construction sites.

**References**

Abdullah, D. and G. C. M. Wern (2011). An analysis of accidents statistics in Malaysian construction sector. International Conference on E-business, Management and Economics, IACSIT Press Honk Kong.

Aires, M. D. M., M. C. R. Gámez and A. Gibb (2010). "Prevention through design: The effect of European Directives on construction workplace accidents." Safety science **48**(2): 248-258.

Ali, R., A. Ali, A. M. Khatak and M. S. Aslam (2018). Large Scale Image Dataset Construction Using Distributed Crawling with Hadoop YARN. 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), IEEE.

Bag, S., S. K. Kumar and M. K. Tiwari (2019). "An efficient recommendation generation using relevant Jaccard similarity." Information Sciences **483**: 53-64.

Choi, J. and S.-W. Lee (2020). "Improving FastText with inverse document frequency of subwords." Pattern Recognition Letters.

Church, K. W. (2017). "Word2Vec." Natural Language Engineering **23**(1): 155-162.

D’Haen, J., D. Van den Poel, D. Thorleuchter and D. F. Benoit (2016). "Integrating expert knowledge and multilingual web crawling data in a lead qualification system." Decision Support Systems **82**: 69-78.

De-Ching, H., C. Shen-Wen, L. Chien-Hung, H. Po-Ta, S. Yi-Ting and S. Huei-Ru (2011). "A study for the evacuation of hospital on fire during construction." Procedia Engineering **11**: 139-146.

Guy, I., I. Schwartz and K. Radinsky (2019). Search system for providing web crawling query prioritization based on classification operation performance, Google Patents.

Hamid, A. R. A., W. Z. W. Yusof and B. S. B. J. Singh (2003). Hazards at construction sites, Universiti Teknologi Malaysia.

Hindocha, E., V. Yazhiny, A. Arunkumar and P. Boobalan (2019). "Short-text Semantic Similarity using GloVe word embedding."

Hong, S.-H., S.-K. Lee and J.-H. Yu (2019). "Automated management of green building material information using web crawling and ontology." Automation in Construction **102**: 230-244.

Hosseinian, S. S. and Z. J. Torghabeh (2012). "Major theories of construction accident causation models: A literature review." International Journal of Advances in Engineering & Technology **4**(2): 53.

Hui, L., W. Yongqing, S. Shimei and S. Baotie (2012). "Study on safety assessment of fire hazard for the construction site." Procedia Engineering **43**: 369-373.

Ingason, H., A. Lönnermark, H. Frantzich and M. Kumm (2010). Fire incidents during construction work of tunnels.

Jeong, M.-J., M.-G. Lee and E.-G. Ham (2014). "Assessment of Fire Evacuation Safety for Building Construction." Journal of the Korean Society of Safety **29**(6): 119-124.

Khatua, A., A. Khatua and E. Cambria (2019). "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks." Information Processing & Management **56**(1): 247-257.

Kim, Y.-A., G.-H. Kim, H.-J. Kim and C.-G. Kim (2019). "Design and Implemention of Real-time web Crawling distributed monitoring system." Journal of Convergence for Information Technology **9**(1): 45-53.

Lai, S., K. Liu, S. He and J. Zhao (2016). "How to generate a good word embedding." IEEE Intelligent Systems **31**(6): 5-14.

Lee, S.-R. (2012). "An Experimental Study on the Fire Risk at Welding· Cutting Process." Fire Science and Engineering **26**(3): 60-66.

Ling, F. Y. Y., M. Liu and Y. C. Woo (2009). "Construction fatalities in Singapore." International Journal of Project Management **27**(7): 717-726.

Ling, W., C. Dyer, A. W. Black and I. Trancoso (2015). Two/too simple adaptations of word2vec for syntax problems. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Lopez-Aparicio, S., H. Grythe, M. Vogt, M. Pierce and I. Vallejo (2018). "Webcrawling and machine learning as a new approach for the spatial distribution of atmospheric emissions." PloS one **13**(7).

Mahto, D. K. and L. Singh (2016). A dive into Web Scraper world. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE.

Massimino, B. (2016). "Accessing online data: Web‐crawling and information‐scraping techniques to automate the assembly of research data." Journal of Business Logistics **37**(1): 34-42.

McInnes, L., J. Healy and J. Melville (2018). "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426.

Menczer, F., S. Fortunato and C. A. Davis (2020). A First Course in Network Science, Cambridge University Press.

Moon, S., Y. Shin, B.-G. Hwang and S. Chi (2018). "Document management system using text mining for information acquisition of international construction." KSCE Journal of Civil Engineering **22**(12): 4791-4798.

Morgan, J., R. Tietje, D. Wang and T. Pattabhi (2020). Web Threat Investigation Using Advanced Web Crawling, Google Patents.

Mustika, N. R. (2018). "Automated Black Box Testing using Selenium Python." International Journal of Computer Science and Software Engineering **7**(9): 201-204.

Niwattanakul, S., J. Singthongchai, E. Naenudorn and S. Wanapu (2013). Using of Jaccard coefficient for keywords similarity. Proceedings of the international multiconference of engineers and computer scientists.

O'Toole, T. (2002). "Construction site safety roles Journal of Construction Engineering and Management 128."

Ombabi, A. H., O. Lazzez, W. Ouarda and A. M. Alimi (2017). Deep learning framework based on Word2Vec and CNNfor users interests classification. 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT), IEEE.

Paul, S., A. Mitra and S. Dey (2017). Issues and challenges in web crawling for information extraction. Bio-Inspired Computing for Information Retrieval Applications, IGI Global**:** 93-121.

Rekabsaz, N., M. Lupu and A. Hanbury (2017). Exploration of a threshold for similarity based on uncertainty in word embedding. European Conference on Information Retrieval, Springer.

Risselada, H., P. C. Verhoef and T. H. Bijmolt (2016). "Indicators of opinion leadership in customer networks: self-reports and degree centrality." Marketing Letters **27**(3): 449-460.

Rodríguez, P., M. A. Bautista, J. Gonzalez and S. Escalera (2018). "Beyond one-hot encoding: Lower dimensional target embedding." Image and Vision Computing **75**: 21-31.

Rong, X. (2014). "word2vec parameter learning explained." arXiv preprint arXiv:1411.2738.

Sandhu, R., H. K. Gill and S. K. Sood (2016). "Smart monitoring and controlling of Pandemic Influenza A (H1N1) using Social Network Analysis and cloud computing." Journal of Computational Science **12**: 11-22.

Shao, L., H. Zhang, M. Jia and J. Wang (2017). "Efficient and effective single-document summarizations and a word-embedding measurement of quality." arXiv preprint arXiv:1710.00284.

Smith, M. and S. Gorgoni (2018). "An introduction to network analysis." Networks of International Trade and Investment: Understanding globalisation through the lens of network analysis: 1.

Vargiu, E. and M. Urru (2013). "Exploiting web scraping in a collaborative filtering-based approach to web advertising." Artif. Intell. Research **2**(1): 44-54.

Wang, F. and W. Du (2012). A test automation framework based on WEB. 2012 IEEE/ACIS 11th International Conference on Computer and Information Science, IEEE.

Wigglesworth, E. (2006). "Occupational injuries by hour of day and day of week: a 20‐year study." Australian and New Zealand journal of public health **30**(6): 505-508.

Yang, S., S. Wi and S. Kim (2018). "Development Methodology of Web Crawling Based on Physical Properties DB of Building Materials for the Efficiency of Building Energy Simulation." 한국생활환경학회지 **25**(4): 467-475.

Yin, Z. and Y. Shen (2018). On the dimensionality of word embedding. Advances in Neural Information Processing Systems.

Yoshioka, K. and H. Dozono (2019). The Classification of the Documents Based on Word Embedding and 2-layer Spherical Self Organizing Maps. Proceedings of the 2019 11th International Conference on Machine Learning and Computing.

Zhang, Z., Y. Huang, P. Zhu and H. Zhao (2018). Effective character-augmented word embedding for machine reading comprehension. CCF International Conference on Natural Language Processing and Chinese Computing, Springer.

Zheng, C., G. He and Z. Peng (2015). "A Study of Web Information Extraction Technology Based on Beautiful Soup." JCP **10**(6): 381-387.

Zhou, Q., D. Fang and X. Wang (2008). "A method to identify strategies for the improvement of human safety behavior by considering safety climate and personal experience." Safety Science **46**(10): 1406-1419.

Zou, P. X. and G. Zhang (2009). "Comparative study on the perception of construction safety risks in China and Australia." Journal of construction engineering and management **135**(7): 620-627.