**Development of prediction model for construction site accident**

**through web crawling and machine learning**

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)

**이 논문은 크게 3개로 나눠서 생각하면 됨. 모든 챕터는 이거 3개를 기반으로**

**Web crawling / accident pattern / prediction model**

**Fire에 집중하는 것이 아니라, 전체적인 사고 유형을 포함**

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

**ABSTRACT**

Recently, construction is getting more complicated due to reflecting the diverse needs of society. As construction sites become more complex, the types of accidents on the site are becoming more diverse. Accidents on the construction site not only cause damage to human life but also increase the construction period and cause huge financial damage. However, plans to improve safety on construction sites are limited in responding to various types of accidents. It is very important to predict the risks of a construction site and prepare effectively but there are few systems or models that predict accidents on construction sites. To fill the knowledge gap, this study analyzed past accidents on the construction site and developed a model to predict on-site accidents. For this study, 5,132 construction site accident articles were collected through the web crawling. Through the text mining using the collected data, patterns by accident type were provided. In addition, an accident prediction model prototype was developed through time series analysis and machine learning. The results of this study can be used as useful data for providing a safety plans considering each schedule and situation on the construction site. Also, the accident prediction model can efficiently predict accidents and improve safety on the construction site.

**KEYWORDS:** Construction, Safety, Accidents, Web crawling, Machine learning

**Introduction**

Due to the continuous development of construction technology, various types of construction projects are in progress. As the construction projects diversify, activities on the construction site become more complicated. The complexity of these construction projects can lead to increased risk and accidents on the site. According to The Bureau of Labor Statistics (BLS), 5,250 fatal work injuries and 1,008 worker deaths on construction sites were recorded in 2018, this is a 2% increase from 2017. Accidents on construction sites cause significant financial damages as well as personal injury. In the National Fire Protection Association (NFPA) report, the frequency of fires related to construction among all building fires is reported to be about 1%, but direct property damage is reported to be about 2%. This shows that accidents on construction sites cause more financial damage than frequency. In addition, accidents on the construction site can lead to an extension of the construction period, which greatly affects the management of the construction project. Many related studies focus on construction site accidents with higher incidence rates. Accidents with lower incidence rates but the high impact may lose the opportunity for further study to those accidents with higher incident rates. In order to compensate for this limitation, this study presented a new analysis method using media. The media is always interested in relatively significant events, so there is very little chance of documenting minor accidents on the construction site. On the other hand, larger scale accidents that have a huge impact on the surroundings are likely to be reported in news articles. In addition, most media articles have a standard meta data format composed of a title, data information, and body, which is suitable for text-mining.

To effectively prepare for accidents on construction sites, this study collected articles related to accidents on construction sites. For accurate article collection, the web-crawling method was used. The web-crawling method is a technique used to collect various information on the web, and it can be converted into text data. Through text-mining using the collected data, patterns by accident type on the construction site were provided. This pattern was analyzed based on the type of accident on the site and time-related information. In addition, an accident prediction model prototype was developed through time series analysis and machine learning of the collected data. This model includes time-series data of the accident patterns and training through machine learning to improve accuracy. The pattern for each type of accident presented in this study makes it possible to intuitively check information about accidents on construction sites. This can be used as meaningful data to make safety regulations on construction sites. And the accident prediction model can predict the risk of the site at the stage of planning the construction project. This predictive model can efficiently establish a site safety plan during the construction project planning stage, and ultimately improve the safety of the construction site.

**Background**

***Construction accidents***

The construction industry is known as one of the most dangerous areas to work. According to the Occupational Safety and Health Administration (OSHA) report, 20% of all industrial worker deaths are included in the construction industry. The construction industry leads to non-fatal injuries, which cost the company millions of dollars per year. The “Fatal Four” from OSHA report leading causes of working fatalities in the construction industry are falls, being struck by an object, electrocution, and being caught in objects. These accident types are responsible for 59 percent of all construction worker deaths. According to the Centers for Disease Control and Prevention (CDC), of all industries, construction causes the most fatal fall accidents, accounting for 51% of all fall accidents in the United States. Also, one fatal injury at the construction sites costs an average of $991,027 in hospital costs. Many studies related to accidents on construction sites analyze causes by accident type. Using statistical data from the Bureau of Labor Statistics (BLS), researchers investigated the types of fatal construction accidents and found workers to die mainly from falls (Zhou, Fang et al. 2008). Hallowell and Gambatese (2009) used the Delphi process to identify key safety-critical activities in formwork construction work. According to this study, the most dangerous activity was exposure to harmful materials. In addition, several approaches have been made by introducing the concept of evaluating and assessment of safety on the construction site. For example, Yang, Chew et al. (2012) analyzed the all of construction accidents in the U.S. from 1995 to 2008 and designed a system to assess the accident possibility. They provided the designed system and confirmed that some construction site accidents can be predicted using the past statistical data. If workers are aware of the safety or risk issues on the construction site in advance, workers tend to behave safely (Mohamed, Ali et al. 2009). Thus, predicting construction site accidents and informing workers through training is a very important point to improve the safety of construction sites (Cooper and Phillips 2004, Hallowell and Gambatese 2009, Abdullah and Wern 2011). As such, in order to effectively notice to construction workers of the risks on the site, it is important to predict the risks on the construction site. One of the best ways to improve the safety of construction sites is to predict and prevent accidents. Providing a pattern based on past accidents and providing an accident prediction model can greatly improve the safety of the site. it can also provide economic benefits for construction sites. According to OSHA, construction companies can save $5 in indirect costs for every $1 invested indirect costs by spending to avoid risk at the working site. The construction industry is inherently dangerous, but the prediction of risks on the site using useful related data can mitigate some of these risks.

***Web crawling***

Web crawling technology is a new method to efficiently collect information by filtering out numerous data on the web (Paul, Mitra et al. 2017, Guy, Schwartz et al. 2019). It is also used for tracking web text documents such as articles and online books on the internet to collect the selected data the user needs (Kim and Ha 2016). Because the data on the web is very large, there is a limit to collecting web data manually. The web crawling technology automatically analyzes web servers and can repeatedly collect information that fits the purpose. Web crawling technology is widely used in research that collects huge data from the web and determines effective decision making and prioritization (D’Haen, Van den Poel et al. 2016, McClain, Aviña et al. 2016). In addition, research was conducted to improve safety by identifying risks through web crawling (Morgan, Tietje et al. 2020). At the beginning of the related research, the target of risk analysis was limited to internet-based such as web pages and e-commerce (Giordani 2018). Recently, it began to analyze risk by integrating web crawling technology into other fields. In the study, research was conducted to prevent driver's driving risk by combining web crawling with a driver monitoring system (Wu, Tsai et al. 2018). There are also studies on how to correctly collect and use online data for research purposes. This study provided guidance on the researchers' responsibilities and related techniques needed to collect and use online data (Massimino 2016). In order to use web crawling technology effectively in research, it is essential to set a clear target. It is necessary to investigate whether the selected targets meet the purpose of the study and whether the website allows web crawling technology. This preliminary investigation is a very important procedure to ensure proper data collection and compliance with security regulations. After completing this step, the researcher can determine the frequency and range of data collection. In the field of construction research, web crawling is a less widely used methodology but has been used limitedly. There are two main purposes of using web crawling in the construction field. The first is to efficiently manage the massive documents used in construction projects through web crawling technology. Most recent construction projects use software or apps to manage related documents. Web crawling can work efficiently in a space that stores construction-related documents. The related research developed a system that collects the information from the construction market and project through a web crawl and automatically assigns each document to the relevant department (Moon, Shin et al. 2018). It also used web crawling to optimize material management and productivity for construction projects. To improve the efficiency of material management in construction projects, related studies have used web crawling to collect material information and provide automated management processes (Yang, Wi et al. 2018, Hong, Lee et al. 2019). Traditionally, the use of web crawling in the construction sector has been limited, but the field of use in construction has been expanding in recent years. For example, there are studies that use the web crawling technology to collect various geographic information without on-site visits and to provide a model to predict the air emissions of each facility (Lopez-Aparicio, Grythe et al. 2018). However, few studies have used web crawling to analyze safety-related factors, such as accident patterns on construction sites. Web crawling is an efficient way to collect data, so research can use it to find accident patterns on construction sites. In this study, web crawling techniques have been used to find the frequency and patterns of accidents on construction sites, which can be a new approach to improving safety by predicting the risks of construction sites.

***Accident prediction model***

Research on the accident prediction model is one of the useful ways to improve the safety of the construction industry. Previous related studies have focused on fall accidents on construction sites. For example, one study analyzed OSHA's fall accident data to provide a risk model based on the height of the construction site. The study used a decision tree model and the results showed 75% fall accident prediction reliability (Chen and Luo 2016). Another study used data mining techniques to provide a decision tree for fall accidents. In this study, fatality chances increased as the distance of the fall increased, and safety education was found to be a way to reduce fatality chances (Mistikoglu, Gerek et al. 2015). There are also studies that reduce accidents on construction sites by predicting the behavior of construction workers. In one study, a survey of 215 construction workers in New Zealand was analyzed and a construction worker safety behavior model was developed and tested (Guo, Yiu et al. 2016). The results of the study were production pressure on construction sites was identified as a critical factor that has direct effects on safety motivation, safety knowledge, and safety compliance. And there is a study that provided a model to predict the working behavior of workers at construction sites using 10 safety climate constructs determined through literature review (Patel and Jha 2015). As a result, safety climate constructs such as manager supervision, work pressure, employee engagement, awareness of risk were significant relationships with worker safety behavior. Recently, there are studies that provide accident prediction models based on various technologies and theories of other fields. For example, there is a study that developed a model for predicting accidents on the construction sites using artificial intelligence technology after collecting data at a construction site using the Delphi method (Ayhan and Tokdemir 2019). The model was able to predict 84% of the accident results on the construction site. In addition, there is a study that developed a model that predicts the safety status of a construction site using real-time data rather than existing statistical data. There are studies that have developed systems that use Real-Time Location System (RTLS) on construction sites to detect in advance the risks of a site and alert workers (Li, Yang et al. 2016). It is based on the stochastic state sequence model and predicts site risk through this mathematical model. In this study, time series analysis was performed with data related to the accidents on the construction site. In addition, machine learning methods were applied to the collected data to develop an accident prediction model.

**Methodology**

***Data collection using the web crawling***

In this study, web crawling method was used for data collection. Web crawling operated with python 3.7 and mainly use Selenium for automated browsing and scarping data from web and Beautifulsoup for parsing article.

Selenium

Selenium provides API to write functional tests by using Selenium Web Driver, which is operated with the chosen browser, for example, Firefox, Chrome. Given that Chrome is relatively faster and widely used on labtop or desktop, researchers choose Chrome browser. Selenium is able to automatically roam all the sites assigned by multiple URL seed. (J.Peng, Y. Ma et al. 2019)

(J. Peng, Y. Ma, F. Zhou, S. Wang, Z. Zheng and J. Li, "Web Crawler of Power Grid Based on Selenium," 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 2019, pp. 114-118, doi: 10.1109/ICCWAMTIP47768.2019.9067730.)

It visits every pages and is able to get page source that indicates location of each articles in the page. With the page sources, we can prepare to use Beautifulsoup for paring HTML data.

Parsing and Structuring data (Beautifulsoup and pandas)

After we gather page sources with Selenium, Beautifulsoup begins to retrieve data we need. Since web pages in New York Times are formatted in HTML, we are able to use ‘html.parser’ method in Beautifusoup which analyze HTML data. The number of articles are over 5,000 in total and thus URLs of articles are separated in five different lists for effective executing. Then, irrelevant format of articles are removed since we are able to only analyze HTML formatted article with text. In other words, the article formatted in blog or interactive article such as animation cannot be analyzed.

Given that lists of articles URLs are prepared, we can parse the data into three parts, body, title and date. This process is also operated in separated lists, and then gather data into a singe list. Thus, we can get list of dates, list of bodies and list of titles.

These data are formed in one data frame by using pandas library that is useful library to structure data into data frame. The data frame enables intuitive slicing data to form new index objects.(W.McKinney, 2011)

(Wes McKinney, “pandas: a Foundational Python Library for Data Analysis and Statistics”, 2011)

The data frame which has three columns, date, title, and body is the base for analyzing distribution of accident type frequency .

***Distribution of accident type***

***Construction accident pattern***

**NLTK**

Before analyzing statistical results, articles body contents need to be cleansed and each sentences should be broken down into word level. In this study, stop words provided from NLTK library are removed since most of these words are functional words and no need to be analyzed. All bodies in article are represented in words level, which is helpful for count the frequency of each words.

**Statistical Results – getting frequency of words**

This study addresses the distribution of keywords frequencies, and visualizes distributions with graph to analyze pattern of accident type. Matplotlib is used for drawing graphs. It is a portable 2D plotting and imaging package visualizing many types of data for example, engineer, scientific and financial data. ( P Barrett, J Hunter et al , 2005) (matplotlib – A portable Python Plotting Package, Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C., & Greenfield, P., 2005)

In this research, we basically examine frequency of keywords related to accident types such as fire, fell , collapse, struct by, explosion ,caught and electrocution. The frequencies of each keywords are counted only once even if they appear several times in a single article. Keywords frequencies are stored into the sets by using python dictionary. These distributions are analyzed by days, seasons and years, and then results are represented in bar chart.

With the yearly times series distribution, data from other resources are used for comparing the time series trend. To be specific, the number of fatal injuries, value of construction and employed persons in thousands are utilized. These four different time series graph are plotted in line graphs within one figure, depicting same x-axis and different y-axis depends on the data.

***Accident prediction model***

***Time series data***

In this part, the study aims at recognition of time series accident pattern. To analyze time series data, we need to generate timeseries data. Keywords frequencies are shown monthly in 20 years. Keywords are based on ‘Fatal Factor’ of construction accident provided by OSHA , it includes falls, struck by object, electrocutions, and caught-in. In addition, this research also examine fire.

***Recurrent Neural Networks***

Recurrent Neural Networks appears to resolve the shortcoming of neural network. Neural network is one of the most finest way to recognize pattern of data, and thus it is widely implemented for diverse areas such as classification, clustering and prediction modeling. It utilizes machine perceptron that helps to label the raw data. The networks are composed of three layers, input layer, hidden layer and output layer. Raw data are given as input set on input layer. The networks learns through input set and results , forming the probability-weighted layer between inputs and results. The weighted layer is hidden layer which performs the labeling input sets which helps to classifying or making prediction modeling. After input sets pass through hidden layer, it generates output sets.

The shortcoming of traditional neural networks is that the model cannot remember the information from other layers. That is, when there are multiple layers in hidden layers, each layers performs independently and it is operated only forward without recurrence. Unlike this feedforward neural networks, recurrent neural networks, RNNs are able to manage various lengths sequence input by building a recurrent hidden layers are dependent on that of the previous time. (Xingyou Wang et al, 2016)

(Xingyou Want, Weijie Jiang, Zhiyong Luo, Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts, 2016)

RNN is also called the Elman recurrent network. It utilizes the context layer which creates a copy of the hidden layers output in the previous time step. The dynamics of the change in neuron activations in the context layer is represented as follows:

where and represent the output of the context state and input neurons. and represent their corresponding weights. is a sigmoid function. (C. Rohitash and Z. mengjie, 2012)

(Rohitash Chandra, Mengjie Zhang, Cooperative coevolution of Elman recurrent neural networks for chaotic time series prediction, 2012)

***LSTM(Long short-term memory)***

LSTM is one of specific method of RNN. Unlike normal RNN, LSTM has four layer to compute hidden state. Four different layers are called forget gate layer, input gate layer, cell state update layer, and output gate layer. Forget gate is the process that decide to forget such information by considering hidden state and t-th input. Input gate layer comes after forget gate layer, deciding what information are stored by considering hidden state and t-th input. Cell state update layer compute what to be updated with tanh function based on results from forget gate layer and input gate layer. Output gate layer decide what parts of the cell should be returned as output and then this result is put on tanh function and multiplied by sigmoid gate. Thus, the output is the hidden state of such point. The process of LSTM can be depicted as follow,

**A screenshot of a video game

Description automatically generated**

Figure 1. Long Short-term Memory Cell

Figure 1 represents a single LSTM memory cell. is implemented by the following compound function(A.Graves, N. Jaitly et al, 2013)

(A. Graves, N. Jaitly and A. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, 2013, pp. 273-278, doi: 10.1109/ASRU.2013.6707742.) :

where, denote weight matrices, terms represent bias vector (e.g is input gate bias vector). is the logistic sigmoid function, and and are respectively the input gate, forget gate, output gate and cell state update layer, all of which are the same size as the hidden layer vector .

**Results**

***Preliminary analysis 표 삽입***

In this study, a total of 5,132 The New York Times articles were collected using a web crawling method. Basically, the author used 'construction accident' as a search condition and the author set the URL of the web page and performed automatic crawling using Selenium. Selenium is a powerful set of software tools that work in many web browsers, programming languages, and web-based frameworks (Wang and Du 2012). This tool has the advantages of easy access, open-source, and flexibility. The URL of each article is collected through this tool, and the collected data is parsed through HTML parser and BeautifulSoup. HTMLParser is a representative library used to parse various text files formatted in HTML and XHTML. The Beautiful Soup library is a Python library that parses broken or corrupted HTML. Using this library, a parse tree is generated where the damaged HTML is almost identical to the original document. This library is used in many studies, and sufficient reliability has been confirmed through various studies (Vargiu and Urru 2013). After these steps, the author classified the collected articles into three according to the composition. It was confirmed that the article generally consisted of the title, date, and body text. Accordingly, this study classified text data into three categories. The author used the Beautiful Soup library to classify text data. This step, as preliminary work to analyze the text data later, enables efficient analysis. The definition of 'English stopwords' used in this study is a set of unnecessary words that are filtered before and after processing natural language data (Moh and Bhagvat 2012). By removing unnecessary words, the author can improve the quality of the collected text data and perform analysis efficiently.

***Construction accident pattern***

In this study, frequency analysis was performed on the typical types of accidents on construction sites. Construction accident types include falls, struck by object, electrocutions, and caught-in in the 'Fatal Four' report provided by OSHA. Also, fire, explosion, and crane accident types have been added. Accidents related to fires and explosions have a low frequency of accidents on the construction site, but when an actual accident occurs, the surrounding influence may be great. The use of cranes at construction sites is increasing with the recent rise in skyscraper projects. As the use of crane increases, the risk of accidents may increase. Therefore, it was classified as one type of accident in this study, and the establishment of construction site regulations regarding the type of crane accident should be considered. According to the results of this study, the frequency of fell was the highest with 32.7%. The total frequency of fell keyword is 1,669 times. This result is the same as the result of a construction site accident type report provided by OSHA and BLS. In general, the most common type of accident on the construction site is a fall-related accident. These same results provide reliability in the data collection methodology of this study. The next most frequent type of accident was fire, with 24.3%. The frequencies of struck by object and caught in were 13.2% and 9.3%, respectively. In addition, explosion and crane were similarly measured at 7.6% and 7.4%, respectively, and electrocutions had the lowest frequency of 5.5%. In the case of the crane-related accident, it is not included in the OSHA fatal injuries category, but it has a relatively high frequency. This is due to an increasing number of accidents involving cranes in recent years and needs to be considered for adding cranes to new categories of related reports. Even for the minimum frequency of electrocutions, a sufficient amount of reliable data was measured with a total frequency of 317. The graph below shows the frequency of incident types of data collected through web crawling.

A picture containing accessory, umbrella

Description automatically generated

Figure 2. The total frequency of accident types

In the results of this study, the frequency of four types of accidents corresponding to 'Fatal Four' was measured similarly to the OSHA report. For the fall type, when comparing the frequency of the OSHA report and this study, it was measured very similarly with a difference of 0.8%. The struck by object type was also relatively similar with a difference of about 2%. In the case of electrocution and caught in type, there was a difference of about 3%, and the result of this study showed a higher frequency of caught in than electrocution, as opposed to the OSHA report. In the OSHA report and the results of this study, the total frequency of the four types of 'fatal four' was very similar, 58.6% and 60.7%, respectively. These four types of accidents account for about 60% of accidents on construction sites, and the results of this study also support this fact. There are types of incidents that differ significantly from OSHA reports in web crawling results. The biggest difference is the fire-related accident. In the OSHA report, the rate of fatal injuries in the fire and explosions category is very low at 1.6%. For this reason, fire and explosion accidents type is excluded from OSHA's 'fatal four' category. However, the frequency of fire and explosion was 31.9% in the results of this study. This is the second highest frequency after the fall accident. In detail, fire and explosion were 24.3% and 7.6%, respectively. This shows that for the type of fire and explosion accidents, unlike the other types, the actual frequency of accidents is less, but it is much more exposed to the media. For the four types of ‘fatal four', the actual frequency and the media frequency are similar, but in the case of fire and explosion, they are different. This can be explained that the fire and explosion types have a greater influence on the surroundings than other types of accidents, thus increasing the frequency of media access. The table below compares the results of this study with the OSHA report.

Table 1. Comparison of results of this study with OSHA report

|  |  |  |
| --- | --- | --- |
|  | OSHA report  (Fatal injuries in 2018) | Web crawling  (Frequency) |
| Falls | 33.5% | 32.7% |
| Struck by object | 11.1% | 13.2% |
| Electrocutions | 8.5% | 5.5% |
| Caught-in | 5.5% | 9.3% |
| Fires and explosions | 1.6% | 31.9% |

In addition, the data collected in this study were classified according to the season. The season with the highest frequency was summer, and the season with the lowest frequency was winter. The difference in frequency between summer and winter is 172, which is about a 4% difference. Also, the frequency of spring and summer and the frequency of fall and winter were similar. The result that the frequency of exposure to media related to accidents on construction sites was measured uniformly can explain that the type of accident on construction sites is not significantly affected by season. This shows that all seasons are exposed to construction site accidents, and there are no seasons where accidents are particularly reduced.

**A screenshot of a cell phone

Description automatically generated**

Figure 3. Frequency of construction accident articles by season

In this study, the graph below was provided to analyze the frequency of accidents on construction sites according to the day of the week. The results showed a high frequency of over 600 times on Friday, Sunday, and Monday. On the contrary, it had the least frequency on Saturday with 538. The gap between the day with the highest frequency and the day with the lowest frequency was about 3%, which was relatively small. The frequency of all accidents on the construction site did not show much difference in the days of the week as in the season. However, this graph is not classified by accident type, but has the limitation that it was analyzed with the total frequency of accidents collected. In order to compensate for this limitation, the frequency of each accident type was extracted and analyzed in the following chapter.

**A picture containing drawing

Description automatically generated**

Figure 4. Frequency of construction accident articles by the day of week

Through the web crawling method, the number of accident-related articles on construction sites by year was investigated. According to the results of the study, the year with the highest frequency of articles in the past 20 years is 2006. In 2006, a total of 476 articles related to construction accidents were written, which is about 3 times more than the minimum frequency of 2017. From 2000 to 2008, more than 200 construction accident articles have been reported for 9 years, and related articles have been decreasing since 2009. The authors compared this yearly data with the fatal injuries data of BLS to confirm the reliability of this study. Interestingly, the results of this study and the BLS report appeared very similar. The year with the highest number of fatal injuries in BLS was 2006, which was common to the results of this study. In addition, there were more than 1000 cases from 2000 to 2008, and the trend has been decreasing since 2009. Both the results of this study and the BLS report show the highest frequency in the past 20 years in 2006, and the frequency has decreased significantly since 2009. This result can be explained that the data collected through the web crawling method used in this study have a similar relationship, not an independent relationship with the statistical data of BLS. In particular, it is explained that the frequency data collected in this study is reliable because the trend of frequency in the past 20 years is very similar. The table provided below compares the frequency of results from this study and the BLS report.

Table 2. Frequency comparison by year

|  |  |  |
| --- | --- | --- |
| Year | Fatal injuries (BLS) | Web crawling |
| 2000 | 1,151 | 223 |
| 2001 | 1,226 | 256 |
| 2002 | 1,125 | 231 |
| 2003 | 1,171 | 223 |
| 2004 | 1,278 | 291 |
| 2005 | 1,243 | 228 |
| 2006 | 1,297 | 476 |
| 2007 | 1,239 | 307 |
| 2008 | 1,016 | 375 |
| 2009 | 879 | 200 |
| 2010 | 802 | 185 |
| 2011 | 781 | 194 |
| 2012 | 849 | 152 |
| 2013 | 856 | 173 |
| 2014 | 933 | 164 |
| 2015 | 985 | 195 |
| 2016 | 1,034 | 164 |
| 2017 | 1,013 | 151 |
| 2018 | 1,038 | 167 |

In addition, this study compared the yearly data collected through web crawling with various data published in the United States. To compare the data, the authors included the fatal injuries of BLS mentioned above and added two other data. First, the added data is ‘Construction spending in the United States by years’ provided by The United States Census Bureau. This data includes all amounts used in construction projects in the United States and is provided in the value of the construction category. Another added data is the ‘Number of construction workers by years’ released annually by The Bureau of Economic Analysis (BEA). The probability of an accident increases as more workers participate in the construction site. If the relevant department can predict the trend of workers participating in the construction site every year, this can predict and respond to the probability of an accident. Since the data compared with the results of this study is published by the US government every year, it has the advantage of continuously comparing and researching in the future. The four data sets, including the results of this study, had very similar trend lines. As shown in the graph below, all data has been increasing since the beginning of 2000 and is commonly peaked in 2006. It has been decreasing from 2006 to 2011, and most data sets are at their lowest point in 2011. Since 2011, the frequency has shifted to an upward trend, which continues until 2018. As described above, the similarity of the frequency of articles related to construction accidents, the frequency of fatal injuries on construction sites, the number of construction workers, and the value of construction spending means that these data can be used to predict accidents on construction sites. If all data sets show independent flows, the reliability of providing predictive models for accidents on construction sites using these data sets may be less reliable. However, since all the data used in this study show a similar flow, providing a predictive model using the collected data can help reduce accidents on the construction site. The graph below shows the annual trend for each data and is expressed by matching the color of the graph with the color of the x-axis.

**A close up of a map

Description automatically generated**

Figure 5. The trend of four sources data by year

In order to find out the monthly trend of frequency according to the type of accident, the collected data was classified into monthly categories. The graph below shows the monthly trend of frequency according to the type of accident on the construction site. In the case of the fall accident type, it has the highest frequency in every month because it has many occurrence frequencies. The fire accident type showed the second highest frequency, followed by the struck by object accident type. All three of these types of accidents have something in common with the highest frequency in May. The types of construction site accidents related to crane, explosion, caught in, and electrocution have relatively low frequency and showed different trends. The difference in monthly frequency for each accident type is because each accident type has different characteristics. In other words, the difference in the frequency of accidents on construction sites by month means that the difference in external conditions, such as climate, may affect the frequency of accidents.

**A close up of a map

Description automatically generated**

Figure 6. Accident frequency trends by month

The table below shows the maximum frequency, minimum frequency, and difference for each accident type. The highest frequency was measured in May from the sum of all types of accidents, and the minimum frequency was in January. In addition, the difference was calculated by comparing these maximum and minimum values. It can be explained that the type of accident with a large difference in monthly occurrence frequency is influenced by external conditions such as climate. The type of accident that showed the highest difference in this study was a crane, which showed a difference of up to 58% per month. The month with the highest frequency was March, and the month with the lowest frequency was December. Because the climates in the two months are not similar, one of the reasons for this difference can be determined as climate. This is related to the climate data that March has the highest wind speed in the United States. This difference is understandable because the risk of accidents involving cranes can vary depending on the wind strength. In addition, the type of electrocution also showed a high difference of 54%. This type had the highest frequency in July and the lowest frequency in January. The type of accident with the lowest monthly frequency difference was the fell type, with a difference of 18%. This result means that the external impact, which varies from month to month, has less impact on the type of fell accident.

Table 3. Maximum, minimum and difference by accident type

|  |  |  |  |
| --- | --- | --- | --- |
|  | Maximum | Minimum | Difference (%) |
| Fell | 165 (May) | 115 (Feb) | 18% |
| Struck by object | 61 (May) | 34 (Jan) | 28% |
| Electrocution | 23 (Jul) | 7 (Jan) | 54% |
| Caught in | 42 (Aug) | 25 (Jul) | 26% |
| Fire | 108 (May) | 69 (Jan) | 22% |
| Explosion | 37 (Feb) | 17 (Jan) | 38% |
| Crane | 50 (Mar) | 13 (Dec) | 58% |

***Prediction model***

* **Training / Test (80%/ 20%)**
* **Error Table**
* **Time series plot (fell, caught, fire, electrical, struck)**
* **Prediction for keywords appearance**

**Discussion**

**Conclusion**

**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.

Ayhan, B. U. and O. B. Tokdemir (2019). "Predicting the outcome of construction incidents." Safety science **113**: 91-104.

Chen, H. and X. Luo (2016). "Severity prediction models of falling risk for workers at height." Procedia engineering **164**: 439-445.

Cooper, M. D. and R. A. Phillips (2004). "Exploratory analysis of the safety climate and safety behavior relationship." Journal of safety research **35**(5): 497-512.

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.

Giordani, A. (2018). "Artificial Intelligence in Customs Risk Management for e-Commerce: Design of a Web-crawling Architecture for the Dutch Customs Administration."

Guo, B. H., T. W. Yiu and V. A. González (2016). "Predicting safety behavior in the construction industry: Development and test of an integrative model." Safety science **84**: 1-11.

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

Hallowell, M. R. and J. A. Gambatese (2009). "Activity-based safety risk quantification for concrete formwork construction." Journal of Construction Engineering and Management **135**(10): 990-998.

Hallowell, M. R. and J. A. Gambatese (2009). "Construction safety risk mitigation." Journal of Construction Engineering and Management **135**(12): 1316-1323.

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.

Kim, S.-m. and Y.-g. Ha (2016). Automated discovery of small business domain knowledge using web crawling and data mining. 2016 International Conference on Big Data and Smart Computing (BigComp), IEEE.

Li, H., X. Yang, F. Wang, T. Rose, G. Chan and S. Dong (2016). "Stochastic state sequence model to predict construction site safety states through Real-Time Location Systems." Safety science **84**: 78-87.

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).

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.

McClain, J. T., G. E. Aviña, D. Trumbo and R. Kittinger (2016). Improving Analysis and Decision-Making Through Intelligent Web Crawling. International Conference on Augmented Cognition, Springer.

Mistikoglu, G., I. H. Gerek, E. Erdis, P. M. Usmen, H. Cakan and E. E. Kazan (2015). "Decision tree analysis of construction fall accidents involving roofers." Expert Systems with Applications **42**(4): 2256-2263.

Moh, T.-S. and S. Bhagvat (2012). Clustering of technology tweets and the impact of stop words on clusters. Proceedings of the 50th Annual Southeast Regional Conference.

Mohamed, S., T. H. Ali and W. Tam (2009). "National culture and safe work behaviour of construction workers in Pakistan." Safety science **47**(1): 29-35.

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.

Patel, D. and K. Jha (2015). "Neural network model for the prediction of safe work behavior in construction projects." Journal of Construction Engineering and Management **141**(1): 04014066.

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.

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.

Wu, Y.-L., H.-Y. Tsai, Y.-C. Huang and B.-H. Chen (2018). Accurate Emotion Recognition for Driving Risk Prevention in Driver Monitoring System. 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), IEEE.

Yang, H., D. A. Chew, W. Wu, Z. Zhou and Q. Li (2012). "Design and implementation of an identification system in construction site safety for proactive accident prevention." Accident Analysis & Prevention **48**: 193-203.

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.

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.