**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***

* **데이터 콜렉션과 관계된 부분 작성**
* **파이썬과 같이 실제 사용한 개발 언어 및 라이브러리 관련도 작성**
* **이전 논문과 비슷하게 쓰되, 중복되면 안됨**

***Pattern analysis by accident type***

* **수집된 데이터를 어떻게 분석할 것인지 방법론 작성**
* **시계열 분석 등 사용할 방법에 관한 모든 것**

***Accident prediction model***

* **예측에 사용된 모든 방법론 정리**

**(Regression model, data mining, machine learning 등)**

* **논문의 핵심 부분이 될 수 있으므로, 상세히 정리하는게 좋을 듯**

**Results**

***Preliminary analysis***

* **웹크롤링을 통해서 수집한 기사 관련 내용**
* **이전 논문의 같은 파트와 비슷하게**
* **데이터 클리닝 방법 및 총 개수 등 상세히 설명 (표 첨부 필수)**

***Accident pattern***

* **그래프나 관련 표 등이 많이 나와야 할 부분 / 웹크롤링으로 수집된 데이터를 어떤식으로 보여줄지에 대해 고민 필요.**
* **일단 할 수 있는것부터 그래프랑 표 만들면서 추가 아이디어 있으면 반영**

1. **20년 (2000~2019) 각 연도별 사고관련 기사 총 개수 및 유형별 정리 필요**
2. **사고 유형별 노출 빈도 분석 (Fall, Collapse, Fire, Equipment(또는 Crane), Explosion 정도)**
3. **OSHA에서 제공해주는 4가지 주요 사고가 Falls/ Struck-By/ Caught-In.Between/ Electrocutions 인데 이 중에서 키워드로 분석가능한 것을 생각해봐야 될 듯.**
4. **사고 유형별 – 요일/월/계절/날씨~~~~~~~~ 등 시계열 분석**
5. **요일을 분석할 방법 찾기. 기사가 하루 이틀 지나고 나올 수 있기 때문에, 요일을 사용하려면, body에서 키워드 검색을 해야하고, 년/계절/월을 분석할때는 기사의 date로 뽑아내는게 나을 듯.**
6. **사고 유형별 어떠한 차이를 보이는지 그래프 등으로 분석 / 한눈에 볼 수 있는 그래프가 있으면 좋을 듯. 각각이 아닌 겹쳐지게.**

***Prediction model***

* **카톡으로 말한 부분 분석 순서 및 결과를 나열 해야함 (모델을 만들어서 80을 가지고서 트레이닝 시키고 그걸로 나머지 20의 데이터에 대해서 테스트)**
* **수식으로 모델 구현**
* **구현된 모델로 테스트 또는 Case study 형식으로 한번 해보면 좋을 듯**

**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.

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.

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.