**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. It is very important to predict the risks of a construction site and prepare effectively. However, plans and regulations to improve safety at construction sites are limited in responding to various variables. Also, 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 method. 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 of the data collected in this study. The results of this study can be used as useful data for providing a safety plan 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 sites, Safety, Accidents, Web crawling, Data mining

**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 the frequency of construction site accidents. However, if only the frequency of accidents collected is used for research, it is difficult to have a high accuracy of research. This is because the frequency analysis does not consider the size and impact of the accident. For example, when comparing a worker's fall at a construction site with a large fire on the construction site, the impact on the construction site is clearly different. However, both accidents are collected in one frequency. There is a limit to analyzing the impact of accidents at construction sites with only these frequencies. In order to compensate for this limitation, this study presented a new analysis method using media. The media is always interested in relatively large issues, so there is very little chance of documenting minor accidents on the construction site. On the other hand, large accidents that have a huge impact on the surroundings are likely to be reported as articles. In addition, most media articles are 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***

* **이전과 같이 각 항목별 논문 또는 관련책 10개씩 넣으면 될 듯**

***Web crawling***

* **이전과 같이 각 항목별 논문 또는 관련책 10개씩 넣으면 될 듯**

***Accident prediction model (Machine learning )***

* **이전과 같이 각 항목별 논문 또는 관련책 10개씩 넣으면 될 듯**

**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**