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

**Using Web Crawling and Deep 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)

\*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 (큰 목차만 색깔 달리 한거다 보기 쉽게)**

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.

***Deep learning / Word Embedding / Word2Vec (대표 의미로 하나 선택해야함)***

* 논문 10개 정도 + 내용 한 두 페이지

**Methodology**

* 데이터 이용한 논문들 읽어보니, 방법론 하나하나 설명할때 수학적 설명 + 다른논문 레퍼런스를 꼭 넣더라.

1. ***Data collection procedures and sample (Web crawling)***

* ***Selenium***
* ***Beautiful Soup***
* ***Data facts***

|  |  |
| --- | --- |
| Issued duration | 2000.01.01 – 2020.04.22 |
| News Source | The New York Times |
| Number of Article | 861 |
| Number of words | 453,283 |

Table 1 : news data information

* 여기서 기간을 2019.12.31일로 해도 괜찮나? 뭔가 딱 20년이 나을듯
* 그리고 조금 필요없는 정보라도 data information을 추가하는게 좋을듯. 우리가 말하고 있는 기사의 계절별 뭐 요일별 등등 추가정보(같이 생각해보자)
* 그리고 이게 원래부터 861개가 아니자나. 처음에 추출된게 몇개였는데 블로그 형식 빼고 이렇게 해서 최종적으로 861개의 기사를 사용했다. 이런식의 설명이랑 표가 필요할듯
* ***Data preprocessing***

1. ***Network analysis***
2. ***Word Embedding - Word2Vec***

* ***Why Word2Vec***
* ***Theoretical grounds in Word2Vec***
* ***-neural network***
* ***skip-gram vs cbow***
* ***Setting of Word2Vec***
* ***cosine similarity***

1. ***UMAP***

* ***Why UMAP***
* ***Dimensional Reduction***

**Results**

**– 니가 데이터 흐름? 따라서 순서를 잡아야함. 내 생각에는 크롤링으로 빈도수 -> 네트워크 분석 -> word2vec -> UMAP 인데 의견도**

* 결론 부분에는 그래프, 표, 이미지 다수 포함. 소제목들은 일단 그래프 표 등 정리한 후에 결정

1. ***Basic statistical facts***Figure 1 : frequency of keywords

* 아직 어디에 제출할지 모르니, 모든 그래프는 흑백으로 작성하는게 좋을듯
* X축 y축 설명 넣어야하고, 각 그래프 위에 몇개인지 숫자를 적어주면 좋을듯
* 모든 그래프, 표에는 citation이 입력되니 이미지에 제목을 넣을 필요는 없음.

1. ***Word Embedding with word2vec***
   1. ***Retrieve similar words by cosine similarity***

***2.1.1 similarity between keywords***

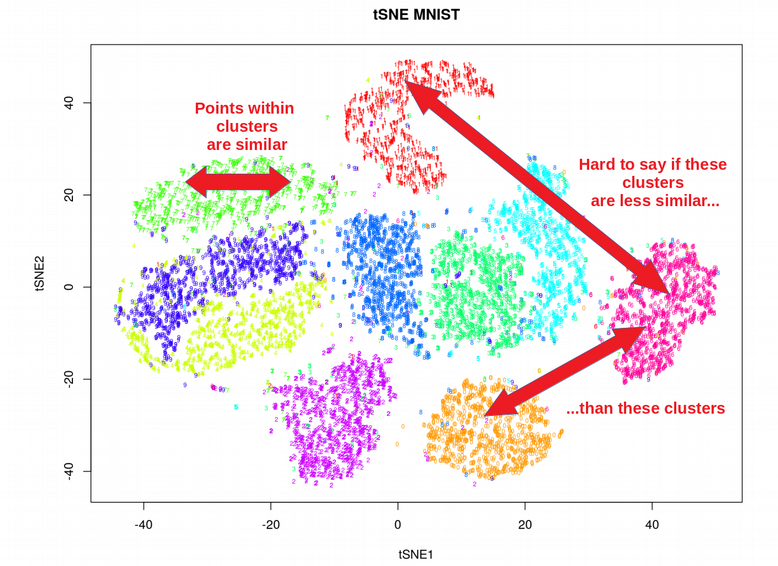
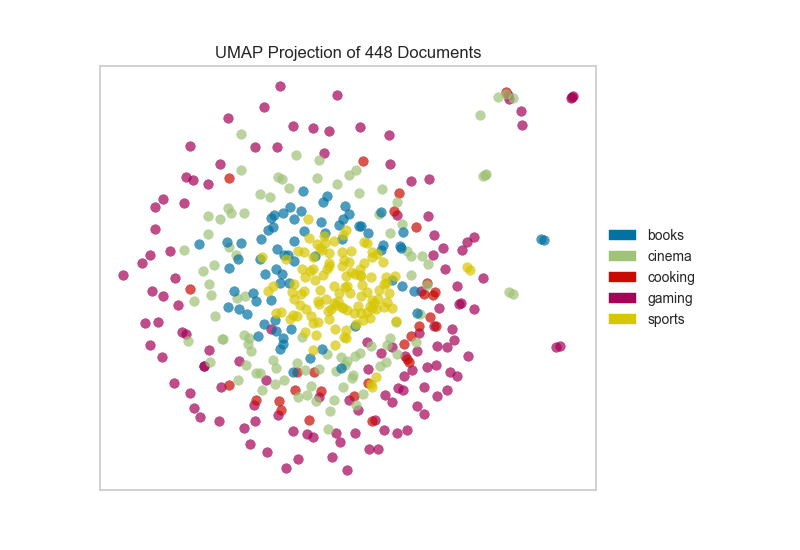
Table 2 : similarity between keywords

* 표는 이미지로 만들지 말고, 엑셀이나 작업한곳에서 그대로 복사해서 가져오는게 나중에 편집할때 좋음
* 이것도 그림으로 나타낼수 있을지에 대한 고민 필요할듯
* 모든 그래프, 표에는 citation이 입력되니 이미지에 제목을 넣을 필요는 없음.

***2.1.2 Visualizing with UMAP***



Figure 2 : word2vec result overview with umap

* 이게 올바른 예인지는 모르겠는데 이런식으로 뭔가 선명한게 논문에는 좋다. 이 그래프는 컬러가 들어가도 괜찮을듯.
* 키워드 말고 다른 단어들은 저런식으로 굵은 점으로 표현해도 좋을것 같다. 각 키워드에 종속되는 단어들 색깔을 달리해서 점으로 표현하고 키워드는 글자 크게 보여주면 될듯
* 어차피 이거 다음에 각각 키워드를 보여줄거라 여기서는 큰 그림이 어떻게 생긴지만 선명하게 보여주면 될듯
* 이 그래프가 논문의 하이라이트로 보일 수 있게 눈에 띄어야함 ㅋㅋㅋㅋㅋㅋㅋ

Subgraphs for keywords

A close up of text on a white background

Description automatically generated

Figure 3 : subgraphs for keywords

Figure 3 shows graph confining the scope of keywords.

* 일단 나중에 합치더라도, 처음엔 5개 키워드 이미지를 각각 크게 만드는게 좋을듯
* 최대한 각 5개 그래프가 각각 선명하게 보이는게 좋음
* 이 그래프에서는 근처의 단어들도 명확하게 보이는 것이 좋음
* 중간중간 색깔이 다른 단어들이 있는데, 이 그래프에서는 키워드 단어를 제외하고는 검정으로 표현하는 것이 나을듯.

***2.1.3 Similar words of keywords***

***- network analysis (Threshold : cosine similarity 0.5)***



Figure 4 : network of keyword's similar words

The size of nodes varies depend on degree.

* 각 키워드 옆에 몇 단어가 링크되어 있는지 표시
* 링크가 많을수록 진해진다고 했는데, 진하기 차이를 더 많이 주면 좋을듯
* 각 키워드별로 나눠서 보여줘야할듯 – 각 키워드별로 보여줄때는 각 링크된 단어도 식별이 가능하면 좋을듯
* 네트워크 분석을 UMAP 앞에 넣을지 Umap 뒤에 넣을지 결정. 나는 마지막에 유맵을 보여줄 생각이었는데, 분석쪽인 니가 전문가니까 의견을 도
* ***Top 20 list of similar words of keywords***



Table 3 : top 20 similar words

* 표는 이미지로 만들지 말고, 엑셀이나 작업한곳에서 그대로 복사해서 가져오는게 나중에 편집할때 좋음
* 이거는 내가 정리한걸로 버릴꺼 제외한 Top 20을 보여주는게 나을듯
* 유사 키워드나 우리가 정한 5개 키워드 색깔 표시 이런건 더 고민해보자. 엑셀에서 바로 붙여놓으면 수정하기 그나마 편함
* 모든 그래프, 표에는 citation이 입력되니 이미지에 제목을 넣을 필요는 없음.

**Discussion**

**Conclusion**

**References**

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Guy, I., I. Schwartz and K. Radinsky (2019). Search system for providing web crawling query prioritization based on classification operation performance, Google Patents.
7. 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.
8. 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.
9. 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.
10. Ingason, H., A. Lönnermark, H. Frantzich and M. Kumm (2010). Fire incidents during construction work of tunnels.
11. 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.
12. 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.
13. Lee, S.-R. (2012). "An Experimental Study on the Fire Risk at Welding· Cutting Process." Fire Science and Engineering **26**(3): 60-66.
14. Ling, F. Y. Y., M. Liu and Y. C. Woo (2009). "Construction fatalities in Singapore." International Journal of Project Management **27**(7): 717-726.
15. 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).
16. 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.
17. 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.
18. Morgan, J., R. Tietje, D. Wang and T. Pattabhi (2020). Web Threat Investigation Using Advanced Web Crawling, Google Patents.
19. O'Toole, T. (2002). "Construction site safety roles Journal of Construction Engineering and Management 128."
20. 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.
21. 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.
22. 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.
23. 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.