**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**

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

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

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

**Background**

***Construction site safety***

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

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

***Web crawling***

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

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

***Word embedding and network analysis***

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

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

**Methodology**

* 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***

Table 1. Source data information

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

* 뉴스기사는 1010 개 까지의 most relevant article 을 추려냈다
* Non relevant article은 format이 블로그 형식 즉 글로만 이루어진article 형식이 아니다
* 그리고 조금 필요없는 정보라도 data information을 추가하는게 좋을듯. 우리가 말하고 있는 기사의 계절별 뭐 요일별 등등 추가정보(같이 생각해보자)
* ***Data preprocessing***

1. ***Network analysis***
2. ***Word Embedding (Word2Vec)***

* ***Why Word2Vec***
* ***Theoretical grounds in Word2Vec***
* ***Deep learning***
* ***Skip-gram vs cbow***
* ***Setting of Word2Vec***
* ***Cosine similarity***

1. ***UMAP***

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

**Results**

***Basic statistical analysis***

******

Figure 1. Frequency of keywords

***Network analysis (Threshold: cosine similarity 0.5)***



Figure 2. The network of keyword's similar words

The size of nodes varies depend on degree.

***Word embedding with Word2vec***

Retrieve similar words by cosine similarity

Table 2. Similarity between keywords

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

Table 3. Top 20 list of similar words of keywords (this is example / not a final version)



***Visualizing with UMAP***

A picture containing text

Description automatically generated

Figure 3. word2vec result overview with UMAP

Subgraphs for keywords / shows graph confining the scope of keywords.

A screenshot of a cell phone

Description automatically generated

A close up of text on a white background

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generated

A close up of a logo

Description automatically generated

Figure 4. subgraphs for keywords

**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. Bag, S., S. K. Kumar and M. K. Tiwari (2019). "An efficient recommendation generation using relevant Jaccard similarity." Information Sciences **483**: 53-64.
5. Choi, J. and S.-W. Lee (2020). "Improving FastText with inverse document frequency of subwords." Pattern Recognition Letters.
6. Church, K. W. (2017). "Word2Vec." Natural Language Engineering **23**(1): 155-162.
7. 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.
8. 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.
9. Guy, I., I. Schwartz and K. Radinsky (2019). Search system for providing web crawling query prioritization based on classification operation performance, Google Patents.
10. Hindocha, E., V. Yazhiny, A. Arunkumar and P. Boobalan (2019). "Short-text Semantic Similarity using GloVe word embedding."
11. 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.
12. 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.
13. 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.
14. Ingason, H., A. Lönnermark, H. Frantzich and M. Kumm (2010). Fire incidents during construction work of tunnels.
15. 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.
16. Khatua, A., A. Khatua and E. Cambria (2019). "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks." Information Processing & Management **56**(1): 247-257.
17. 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.
18. Lai, S., K. Liu, S. He and J. Zhao (2016). "How to generate a good word embedding." IEEE Intelligent Systems **31**(6): 5-14.
19. Lee, S.-R. (2012). "An Experimental Study on the Fire Risk at Welding· Cutting Process." Fire Science and Engineering **26**(3): 60-66.
20. Ling, F. Y. Y., M. Liu and Y. C. Woo (2009). "Construction fatalities in Singapore." International Journal of Project Management **27**(7): 717-726.
21. 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).
22. 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.
23. 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.
24. Morgan, J., R. Tietje, D. Wang and T. Pattabhi (2020). Web Threat Investigation Using Advanced Web Crawling, Google Patents.
25. O'Toole, T. (2002). "Construction site safety roles Journal of Construction Engineering and Management 128."
26. Ombabi, A. H., O. Lazzez, W. Ouarda and A. M. Alimi (2017). Deep learning framework based on Word2Vec and CNNfor users interests classification. 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT), IEEE.
27. 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.
28. Rekabsaz, N., M. Lupu and A. Hanbury (2017). Exploration of a threshold for similarity based on uncertainty in word embedding. European Conference on Information Retrieval, Springer.
29. Risselada, H., P. C. Verhoef and T. H. Bijmolt (2016). "Indicators of opinion leadership in customer networks: self-reports and degree centrality." Marketing Letters **27**(3): 449-460.
30. Rodríguez, P., M. A. Bautista, J. Gonzalez and S. Escalera (2018). "Beyond one-hot encoding: Lower dimensional target embedding." Image and Vision Computing **75**: 21-31.
31. Sandhu, R., H. K. Gill and S. K. Sood (2016). "Smart monitoring and controlling of Pandemic Influenza A (H1N1) using Social Network Analysis and cloud computing." Journal of Computational Science **12**: 11-22.
32. Shao, L., H. Zhang, M. Jia and J. Wang (2017). "Efficient and effective single-document summarizations and a word-embedding measurement of quality." arXiv preprint arXiv:1710.00284.
33. Smith, M. and S. Gorgoni (2018). "An introduction to network analysis." Networks of International Trade and Investment: Understanding globalisation through the lens of network analysis: 1.
34. 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.
35. Yin, Z. and Y. Shen (2018). On the dimensionality of word embedding. Advances in Neural Information Processing Systems.
36. Yoshioka, K. and H. Dozono (2019). The Classification of the Documents Based on Word Embedding and 2-layer Spherical Self Organizing Maps. Proceedings of the 2019 11th International Conference on Machine Learning and Computing.
37. Zhang, Z., Y. Huang, P. Zhu and H. Zhao (2018). Effective character-augmented word embedding for machine reading comprehension. CCF International Conference on Natural Language Processing and Chinese Computing, Springer.
38. 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.
39. 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.