**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. 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***

***One hot encoding***

***Hidden layer***

***Skip-gram vs Cbow***

***A close up of a map

Description automatically generated***

Figure 1. Skip-gram

***A close up of a map

Description automatically generated***

Figure 2. CBOW

* ***Setting of Word2Vec***

***Mincount : 200***

***Training Algorithm : CBOW***

***Dimensionality of the word vectors : 400***

* ***Cosine similarity***

1. ***UMAP***

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

**Results**

***Preliminary analysis***

* 웹크롤링을 통하여, 총 몇 개의 relevant articles을 찾아 냈습니다. relevant articles을 선별하는 과정에는 ~~~~~~~~어떠한 키워드/개발언어/라이브러리를 사용하였습니다. 이 라이브러리는 많은 연구에서 사용되고 있으며, 신뢰성이 있다고 판단됩니다.(참고문헌)
* The authors checked for unsuitable data, ~~~~~~~~ 우리가 총 수집한 데이터 중에서 어떠어떠한 이유로 몇개의 데이터를 삭제했다. 이유가 명확해야 되고, 데이터를 선별하는 작업에는 어떠한 언어 또는 소프트웨어를 사용하였으며, 결국 총 몇개의 article을 사용하게 되었다.
* 데이터 클리닝을 마친후, 저자는 articles 들을 분류 하였습니다. 일반적으로 기사들은 Title, date, body로 구성되기 때문에, 우리는 이 방법으로 나눴습니다. 이것을 나눌때는 어떠한 언어와 어떠한 방법을 사용했고, 데이터 포맷을 ~~ 입니다.
* 이 외에도 분석을 시작하기 전에, 했던 작업들이 있으면 추가

***Basic statistical analysis***



Figure 3. Frequency of keywords

* Preliminary analysis를 통해 얻어진 선별된 데이터를 바탕으로, Basic statistical analysis를 수행했습니다. 선별된 기사에 포함된 모든 단어들을 나열하여, 빈도 분석을 실시 하였습니다. 이 분석을 할 때, 필요없다고 판단되는 데이터를 우선 삭제하기 위해, ~~~ ~~~~~~~라이브러리를 이용하여 분석했습니다. (니가 저번에 영어 단어 중에 is,are 이런거 자동으로 빼준다 했던 그부분 설명/참고문헌 필요)
* 이 키워드 5개 말고, 모든 단어(is are이런거 빼고) 중에서 가장 많이 나오는 단어에 대한 설명이나 표가 추가 되면 좋을 듯
* 본 연구에서는 최소 빈도를 200회로 설정하여, text 분석을 했습니다. 우선적으로 분석을 마친 단어들 중, 본 연구와 관련이 높은 키워드 5개를 선정하였습니다. 5개의 키워드는 fire, fell, collapsed, building, people입니다. 이 선정 기준은 공사현장에서의 사고와 화재사고를 대표하는 키워드를 고려한 것이다. 많은 연구에서 빌딩화재의 3대 요소는 fire, building, people이라고 정의하고 있습니다. 또한 fell과 collapsed는 건설현장에서 가장 빈도가 높은 사고 유형입니다. 모든 키워드들은 충분한 빈도수를 가지고 있기 때문에, 이 키워드를 사용하는데 무리가 없습니다.
* 분석된 결과에 따르면, building과 people이 가장 많은 빈도를 보였습니다. 또한 fire, fell, collapsed는 비교적 비슷한 빈도를 보였습니다. Fell과 collapsed는 유사단어이기 때문에, 이를 합친 빈도를 고려할 수도 있습니다. 두 단어를 합쳐서 비교하면, Fire는 27% 빈도가 있으며, Fell과 collapsed는 73%의 빈도를 보입니다. 이것은 The Bureau of Labor Statistics (BLS) 의 사고 빈도 분석과 상당한 차이를 보입니다. BLS보고서에 따르면, 공사현장에서의 사고 중 Fell과 Collapsed와 관련된 사고는 전체 사고의 약 40%이며, Fire 사고는 2%입니다. 이를 비율로 환산하면, Fire사고는 Fell관련 사고에 5%의 빈도를 가지고 있습니다. 본 연구의 결과에서 나타난 27%와 BLS 보고서의 5%에는 큰 gap이 있습니다. 이 gap에 관한 가능성 있는 추론은 저자가 인트로에서 설명한 미디어의 특징으로 설명될 수 있습니다. 즉 Fell과 관련된 사고보다 Fire와 관련된 사고가 미디어에 노출이 더 많이 될 수 있으며, 이는 건설현장에서의 Fire사고가 Fell관련 사고에 비해 더 큰 영향을 미친다는 것을 설명할 수 있습니다.
* 이 외에 단어 빈도 분석을 통해 통계적이나, 데이터 쪽으로 더 설명 필요한 부분 있으면 추가 ~~~~~~~~~

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

A close up of a logo

Description automatically generated

Figure 4. The network of keyword's similar words

The size of nodes varies depend on degree.

* 키워드 이외에 death 가 큰 degree 를 갖음을 알 수 있다.

|  |  |
| --- | --- |
| Basic Information | |
| Number of Nodes | 136 |
| Number of Edges | 353 |

|  |  |
| --- | --- |
| Keywords | Connected Nodes |
| Fire | 80 |
| Fell | 88 |
| Collapsed | 91 |
| Building | 50 |
| People | 50 |

* 총 136개의 nodes와 353개의 edges 로 이루어 져있다
* 각 키워드의 연결된 node는 cosine similarity 0.5 이상의 단어를 연결된 노드로 갖는다 . Fire 는 80 개 , Fell 은 88개, Collapsed 91 개, Building 50개 , People 50 개를 갖는다 .

***Word embedding with Word2vec***

Retrieve similar words by cosine similarity

Table 2. Similarity between keywords

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

* 본 연구에서는 cosine similarity을 활용하여, 각 키워드간의 Similarity를 제공하였습니다.
* cosine similarity는 ~~~~~ 개념이며, 이를 구하는 방법은 ~~~~ 이며, 수식은 아래와 같습니다. (개념과 수식 넣어야함). cosine similarity의 범위를 알려줫으면 좋겠음. 예를 들어서 0.5?? 이상이면 유사도가 있다고 판단된다 등등 일반적인 범위 ~~~~~
* 앞서 선정한 5개 키워드들 2개씩 분류하여 총 10가지 조합을 만들었으며, 각각의 유사도를 계산했습니다. 본 연구의 신뢰성을 확인하기 위하여, 분석된 단어중 일반적으로 가장 유사한 단어간의 cosine similarity를 우선 확인하였습니다. 저자가 선택한 단어는 Fell과 collapsed 입니다. 이 두 단어간의 사전적 의미는 매우 유사하기 때문에, 본 연구의 결과가 신뢰성이 있다면, 이 단어간의 유사도가 높게 나타나야 합니다. 두 단어간의 유사도는 0.951로 나타났으며, 이는 다른 키워드들간의 유사도보다 매우 높게 계산되었습니다. 이를 통해, 본 연구의 신뢰도를 간접적으로 확인할 수 있습니다.
* 본 연구의 결과를 Fire 관점에서 확인해보면, fire는 people보다 building키워드와의 유사도가 높게 나타 났습니다. Fire와 building의 similarity는 0.525이며, fire와 people의 similarity는 0.331로 나타났습니다. 이는 fire는 people과 관련된 단어보다, building과 관련된 단어들과 훨씬 높은 유사도를 보였다고 해석할 수 있습니다. Fell과 Collapsed의 경우, people과 building 키워드간의 유사도는 매우 일정하게 나타났습니다. 또한 fell의 경우, building 키워드와의 similarity는 0.443이며, people 키워드와의 similarity는 0.486으로 나타났습니다. 이는 fire와는 반대로, fell은 people 키워드와의 유사도가 조금 더 높습니다.
* 또한 people과 building 간의 similarity는 negative value로 계산되었으며, 이는 두 키워드간의 관계성이 크지 않다는 것을 나타냅니다.

Table 4. Top 20 list of similar words of keywords

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fire** | **Fell** | **Collapsed** | **Building** | **People** |
| police | collapsed | fell | department | died |
| investigators | side | debris | inspectors | injured |
| authorities | floor | floor | issued | killed |
| hospital | debris | ground | commissioner | men |
| injuries | ground | west | city | accident |
| chief | street | steel | cranes | authorities |
| night | Friday | street | contractor | debris |
| floor | avenue | death | violations | train |
| man | death | side | tower | injuries |
| dead | morning | site | crane | dead |
| Monday | worker | Friday | investigation | workers |
| Friday | injuries | march | office | ground |
| driver | Monday | wall | equipment | Monday |
| investigation | injured | injuries | site | crash |
| truck | dead | authorities | district | march |
| officials | wall | Monday | charges | residents |
| death | authorities | avenue | investigators | members |
| worker | steel | tower | company | Tuesday |
| fell | crash | Tuesday | officials | Thursday |
| working | died | injured | mayor | cars |

* 아래의 표는 5개의 키워드들과 유사도가 높은 20개 단어를 보여줍니다. 이것은 ~~~~~~ 값(cosine similarity?)을 사용하여 계산한 후, 유사도가 높은 순서대로 나열했습니다.
* 결과에서, Fell과 collapsed 키워드는 서로 가장 유사도가 높다고 계산되었습니다. 이는 키워드간의 cosine similarity 확인과 함께, 본 연구의 신뢰도를 보여줍니다.
* Fire, Fell, Collapsed와 같이 건설 현장의 사고유형과 관련된 키워드에서, Monday와 Friday 단어의 유사도가 높게 나타났습니다. 이는 산업의 The Distribution of injuries관련 연구와 유사한 결과입니다. 관련 연구에서 현장에서 injuries가 가장 많이 일어난 요일이 월요일이라고 조사되었으며, 이는 본 연구의 결과와 유사하다고 볼 수 있습니다(Wigglesworth 2006).
* 그리고, fire 키워드에서, 다른 키워드과 달리 ‘night’의 단어와의 유사도가 높게 나타났습니다. 일반적으로, 화재가 밤에 발생할 때 확산 가능성이 더 커집니다. 왜냐하면, 일과시간 이외에는 건물에 거주자가 있을 확률이 적으며, residential building의 경우에도 사람들의 수면시간에 화재를 인식하기 힘듭니다. 특히 본 연구에서 다루고 있는 construction site의 경우에는 밤에는 관리를 위한 일부 직원이 머무르거나, 아에 비어 있을 가능성이 residential 건물보다 높습니다. 그리고 during construction phase 에서는 fire and smoke alarm과 같은 안전 설비가 완료되지 않았기 때문에 , 화재를 인식하기 더욱 힘듭니다(Hamid, Yusof et al. 2003). 본 연구의 결과가 이를 다시 한번 강조 시켰다고 볼 수 있습니다.
* Building과 유사도가 높은 단어에는 ‘administration’ or ‘inspection’의 의미를 가지고 있는 단어들이 많습니다. 가장 유사도가 높은 상위 5개의 단어가, department, inspectors, issued, commissioner, city로 비슷한 의미를 지니고 있습니다. 이는 건설현장에서의 사고와 관련된 기사에서 building의 inspection 및 management 이슈를 주로 다룬다는 것을 보여줍니다. 또한 people과 유사도가 높은 단어들은 act 또는 behavior과 관련된 단어들이 많습니다.

***Visualizing with UMAP***

* 본 연구에서는 Word2vec을 통해 분석한 결과를 UMAP에 가시화 하였습니다. UMAP은 다차원 공간의 vector를 2차원 공간으로 차원을 축소하여 가시화 함으로써, 보다 직관적으로 키워드간의 관계를 볼 수 있습니다. UMAP으로 가시화하기umap library 를 사용하고 , Parameters : number of neighbors = 5, min\_dist = 0.1, n\_components = 2.(참고문헌 필요)
* UMAP의 간단한 절차 및 의미하는 바를 ~~~~~~~~설명
* As with other k-neighbour graph based algorithms, UMAP can be described in two phases. Un the first phase a particular weighted k-neighbor graph is constructed. In the second phase a low dimensional layout of this graph is computed.( Leland Mcinnes et al, 2018)\*\*
* 각 키워드를 각각의 색깔로 표현하였으며, 각 키워드의 범위를 gradation으로 나타냈습니다. UMAP 상의 회색 점들은 키워들과의 유사도가 낮은 단어들입니다.
* 아래의 figure과 같이, fell과 collapsed는 UMAP 상에서 거의 overlap되게 나타났습니다. 이는 UMAP 안에서도 두 키워드간의 유사도는 매우 높다는 것을 의미합니다. Fire 키워드의 경우, 다른 키워드들보다 범위가 넓게 나타나며, building 키워드와의 교집합이 존재합니다. Fell과 Collapsed의 경우, people 키워드가 building 키워드 보다 더 가깝네 나타났습니다.

A picture containing text

Description automatically generated

Figure 5. Word2vec result overview with UMAP

* 아래의 figures은 각 키워드에 한정된 UMAP 가시화를 보여줍니다. 이것을 통해, 각 키워드들과의 유사단어들이 UMAP 을 통해 표현되며, 단어들간의 유사도가 거리로 표현 되어있습니다.
* 또한 단어의 유사도에 따라서 그룹화 할 수 있습니다.

A close up of a logo

Description automatically generated

A close up of a device

Description automatically generated

A close up of a device

Description automatically generatedA close up of a piece of paper

Description automatically generated

A close up of text on a white background

Description automatically generated

Figure 6. Subgraphs for keywords

**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.

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.

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.

Bag, S., S. K. Kumar and M. K. Tiwari (2019). "An efficient recommendation generation using relevant Jaccard similarity." Information Sciences **483**: 53-64.

Choi, J. and S.-W. Lee (2020). "Improving FastText with inverse document frequency of subwords." Pattern Recognition Letters.

Church, K. W. (2017). "Word2Vec." Natural Language Engineering **23**(1): 155-162.

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.

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.

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

Hamid, A. R. A., W. Z. W. Yusof and B. S. B. J. Singh (2003). Hazards at construction sites, Universiti Teknologi Malaysia.

Hindocha, E., V. Yazhiny, A. Arunkumar and P. Boobalan (2019). "Short-text Semantic Similarity using GloVe word embedding."

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.

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.

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.

Ingason, H., A. Lönnermark, H. Frantzich and M. Kumm (2010). Fire incidents during construction work of tunnels.

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.

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.

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.

Lai, S., K. Liu, S. He and J. Zhao (2016). "How to generate a good word embedding." IEEE Intelligent Systems **31**(6): 5-14.

Lee, S.-R. (2012). "An Experimental Study on the Fire Risk at Welding· Cutting Process." Fire Science and Engineering **26**(3): 60-66.

Ling, F. Y. Y., M. Liu and Y. C. Woo (2009). "Construction fatalities in Singapore." International Journal of Project Management **27**(7): 717-726.

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.

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.

O'Toole, T. (2002). "Construction site safety roles Journal of Construction Engineering and Management 128."

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.

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.

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.

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.

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.

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.

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.

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.

Wigglesworth, E. (2006). "Occupational injuries by hour of day and day of week: a 20‐year study." Australian and New Zealand journal of public health **30**(6): 505-508.

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.

Yin, Z. and Y. Shen (2018). On the dimensionality of word embedding. Advances in Neural Information Processing Systems.

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.

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.

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.

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.