Journal of the Korean Society of Safety
Vol. 38, No. 3, pp. 69-76, June 2023
https://doi.org/10.14346/JKOSOS.2023.38.3.69

ISSN 1738-3803 (Print)
ISSN 2383-9953 (Online)
http://www.kosos.or.kr/jkosos

산업안전보건기준에 관한 규칙의 체계적 탐색과 이해를 위한 단어분포 지표와 Word2Vec 분석 방법

정재호[1] · 장성록[2] · 서용윤[3†]

Term Distribution Index and Word2Vec Methods for Systematic Exploring and Understanding of the Rule on Occupational Safety and Health Standards

Jae Ho Jeong[1] · Seong Rok Chang[2] · Yongyoon Suh[3†]

†Corresponding Author
Yongyoon Suh
Tel : +82-2-2260-3786
E-mail : ysuh@dgu.edu

Received : October 17, 2022
Revised : April 22, 2023
Accepted : June 11, 2023

Abstract : The purpose of the rules on the Occupational Safety and Health Standards (hereafter safety and health rules) is to regulate the safety and health measures stipulated in the Occupational Safety and Health Act and the specific instructions necessary for their implementation. However, the safety and health rules are extensive and complexly connected, making navigation difficult for users. In order for users to readily access safety and health rules, this study analyzed the frequency, distribution, and significance of terms included in the overall rules. First, the term distribution index was created based on the frequency and distribution of words extracted through text mining. The term distribution index derives from whether a word appears only in a specific chapter or across all rules. This allows users to effectively explore terms to be followed in a specific working environment and terms to be complied with in the overall working environment. Next, the related words of the previously derived terms were visualized through t-SNE and the Word2Vec algorithm. This can help prioritize the things that need to be managed first, focusing on key terms without checking the overall rules. Moreover, this study can help users explore safety and health rules by allowing them to understand the distribution of words and visualize related terms.

Key Words : rule on occupational safety and health standards, textmining, term distribution index, Word2Vec, typology of term

1. 서 론

산업안전보건기준에 관한 규칙은 산업안전보건법에서 규정하는 안전보건조치와 그 시행에 필요한 구체적인 지시사항을 규정함을 목적으로 하고 있다. 이에 따라 관리감독자 및 안전관리자 등 산업안전과 관련된 모든 이해관계자는 산업안전보건기준에 관한 규칙에 따라 작업장 안전보건조치를 수행하고 있다. 이를 위해 법령 사용자는 국가법령정보센터에서 제공하는 산업안전보건기준에 관한 규칙을 통해 검색 기능을 사용하며 법령을 탐색한다.

그러나 산업안전보건기준에 관한 규칙은 내용이 광범위하고 전문적이며, 조문 간의 관계가 복잡하게 연결되어 있어, 법령 사용자가 관리해야 할 주요 사항을 탐색하고 접근하는 데 어려움이 있다[1]. 또한, 법령을 처음 접하는 사용자의 경우 산업안전에 대한 전문성이 부족하여 어떠한 단어들이 규칙 내에 존재하는지 알지 못하는 경우가 많다.

[1]부경대학교 안전공학과 석사과정 (Department of Safety Engineering, Pukyong National University)
[2]부경대학교 안전공학과 교수 (Department of Chemical Engineering, Pukyong National University)
[3]동국대학교 산업시스템공학과 부교수 (Department of Industrial & Systems Engineering, Dongguk University(Seoul)

이로 인해 법령을 효과적으로 탐색하고, 일반인도 쉽게 이해할 수 있도록 관련 연구가 진행되고 있다. 판례에서 참조 법령을 추출하여 법령들을 임베딩하는 Law2Vec 모형을 구축하여[2], 법령 간 유사도 검색으로 연관 법령을 쉽게 검색하는 방법을 제시하거나, 법령 네트워크 시각화 분석을 통해 사용자가 연결 조항 또는 유사 조항을 검토하는 등의 활용방법을 제안한 연구가 있다[3]. 리스트화된 텍스트 위주의 법의 검색이나 다양한 시각화 기능 도구를 개발하여 일반 사용자의 사용성 측면을 강화하였다.

특히, 안전 분야의 경우 법령 체계 마련 및 정비 연구는 꾸준히 진행되고 있으나, 일반인을 위한 법령 내, 법령 간 탐색을 위한 실질적인 연구는 상당히 부족한 실정이다. 안전 법령은 사업장 책임자와 관리자들이 의무적으로 확인하고 지속해서 관리하기 위한 사항을 명령하면서도, 정작 그 법령을 확인하는 사업장 관계자에게는 이해하기 어려운 내용이 많다. 또한, 중소기업 안전 관계자는 무지로 인한 법령 리스크가 대기업보다도 크기 때문에, 법령의 이해를 돕기 위한 연구가 이루어질 필요가 있다.

이와 같은 문제를 해결하기 위해 본 연구에서는 산업안전보건기준에 관한 규칙 내 단어의 분포를 지표화하여 사용자가 탐색하고자 하는 단어의 유형을 나타내고자 한다. 단어분포 지표는 TF-IDF를 개량하여 개발하였고, 법령 내 단어의 유형을 파악함으로써 해당 단어가 특정 작업환경에서 준수해야 할 용어인지, 전반적인 작업환경에서 고려해야 할 용어인지 판단할 수 있다. 또한, 조문 내 단어의 연관성을 분석하기 위해 워드 임베딩(Word embedding) 기법의 하나인 Word2Vec을 활용하였다. 이후 훈련된 임베딩 모델을 시각화하여 사용자는 연관단어를 직관적으로 파악할 수 있다. 결과적으로 사용자는 단어의 유형과 연관단어를 효과적으로 파악함으로써 법령 탐색에 대한 접근성 향상에 본 연구의 목적이 있다.

2. 배경이론

2.1 TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency)는 특정 단어가 특정 문서 내에서 발생할 확률을 나타내는 지표로, 정보 수집 및 텍스트 마이닝 분야에서 주로 활용되고 있다[4]. TF(Term frequency)는 단어가 문서 내에 포함되어 있는지를 나타내는 값, 즉 빈도수를 의미한다. 빈도수가 클수록 문서 내에서 중요한 단어라고 생각할 수 있다. 그러나, 단어가 특정 문서 내에서 자주 사용되어 흔하게 등장했을 때는 중요도가 떨어질 수 있다. 단어가 출현한 문서의 수를 DF(Document frequency)라고 한다. TF와는 다르게 단어가 문서 내에서 여러 번 출현해도 한 번만 카운트한다. IDF(Inverse document frequency)는 DF의 역수이며, 단어가 문서 집합 전체에서 얼마나 공통으로 출현하는지를 나타내는 값이다. TF-IDF는 식 (1)과 같이 TF와 IDF의 곱으로 구성되며, tf(t,d)는 특정 문서 d에 포함된 t 단어의 개수, df(t)는 전체 문서에서 t 단어를 포함한 문서의 개수, N은 전체 문서의 개수를 나타낸다[5]. 따라서, 값이 클수록 문서 내에서 단어도 많이 포함되고, 그 단어를 포함한 문서도 전체 문서 내에 많이 없다는 의미이기 때문에, 문서를 대표하는 단어로 추론할 수 있다.

$$TF-IDF = tf(t,d) \times log\frac{N}{df(t)} \quad (1)$$

2.2 Word Embedding

워드 임베딩 기법은 단어를 벡터 공간에 표현하는 것을 말하며, 각각의 단어에 가중치 벡터를 부여한다. 비슷한 위치에서 등장하는 단어들은 비슷한 의미와 관계를 가진다는 것을 유추한다. One-hot vector로 표현하는 희소 표현(Sparse representation)과 달리 실수값으로 벡터의 차원을 표현하여 더 적은 차원으로 단어의 의미를 나타낸다. 이러한 표현을 밀집 표현(Dense representation)이라 한다. 밀집 표현은 희소 표현에서 일어나는 벡터 공간의 낭비를 줄여 계산량을 줄이는 장점이 있다.

워드 임베딩의 다양한 방법론 중 본 연구에서는 Word2Vec을 사용하여 단어 간의 연관성을 분석하였다. Word2Vec은 2013년 구글에서 발표한 CBOW와 Skip-gram 두 가지 모델이 존재한다[6,7]. CBOW는 주변 단어로 중심 단어를 예측하고, Skip-gram은 중심 단어로 주변 단어를 예측하는 방법이다. Skip-gram은 중심 단어로부터 여러 주변 단어를 예측 및 학습하기 때문에 CBOW보다 좋은 성능을 보이며[8], 따라서 본 연구에서는 Skip-gram을 사용하였다.

3. 연구설계

3.1 연구절차

본 연구는 다음 Fig. 1과 같이 진행하였다. 먼저, 국가법령정보센터에서 산업안전보건기준에 관한 규칙을 수집하였다. 다음으로, 텍스트 마이닝을 활용하여 전처리를 수행하고, 추출된 단어의 빈도수를 계산하였다.

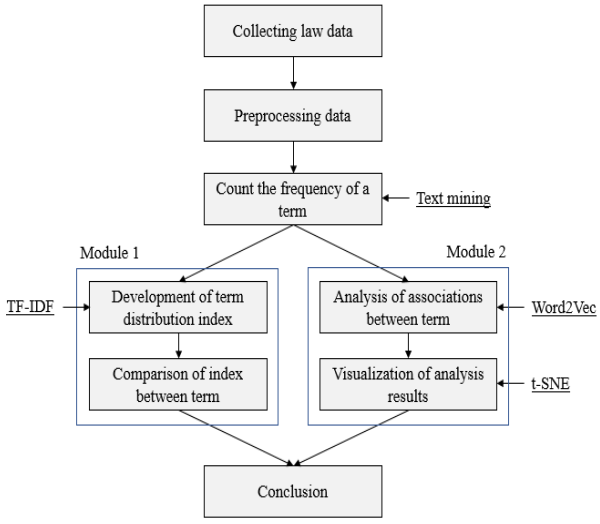

Fig. 1. Procedure of the study.

Module 1에서는 TF-IDF를 개량하여 단어분포 지표를 개발하였고, 단어 간의 지표를 비교하여 단어의 유형을 4가지로 매핑하였다. Module 2에서는 Word2Vec을 활용하여 추출된 단어의 연관성을 분석하였고, t-SNE를 통해 3차원으로 시각화하였다.

3.2 데이터 수집

본 연구의 분석 데이터는 국가법령정보센터에서 제공하는 산업안전보건기준에 관한 규칙을 사용하였다. 먼저, 국가법령정보센터에서 산업안전보건기준에 관한 규칙 680개의 조문을 수집하였다. 다음으로, 조문 데이터를 Python에 적용할 수 있도록 .txt 파일을 사용하여 680개의 행으로 List화 하였다.

3.3 데이터 전처리

수집된 데이터는 Python KoNLPy Package[9]의 Mecab 형태소 분석기를 사용하여 텍스트마이닝(textmining)을 수행하고, 전처리를 수행하였다. 조사, 수사, 관형사 등 분석에 필요하지 않은 단어는 제거하였다. Python 형태소 분석기 특성상 단어를 분해하여 추출하는데, 결합형 단어가 많은 규칙에서 이러한 추출은 적합하지 않다. 이를 해결하기 위해 한국산업안전보건공단에서 제공하는 안전보건용어사전에 명시된 단어를 형태소 분석기 사전(Dictionary)에 포함하여 분석에 필요한 단어를 추출하도록 하였다.

3.4 단어분포 지표 개발

먼저, 산업안전보건기준에 관한 규칙 내 단어의 분포를 확인하기 위해 단어가 출현하는 조항, 절, 장의 수를 계산하였다. 이는 단순히 전체 산업안전보건기준 규칙에 포함되는 단어가 아닌, 특정 절이나 장에 포함된 단어 비율을 탐색하기 위해서이다. 산업안전보건기준에 관한 규칙에는 기술용어가 많이 적시되어 있고, 또 많은 조항에 분포되어 있다. 따라서, 기술용어마다 분포도를 살펴보아, 특정 장이나 절을 대표할 수 있는 기술용어를 확인하는 것이 필요하다.

이를 위해, TF-IDF 지표를 개선(modification)하여, 특정 절이나 장에 포함된 단어들을 도출하고자 한다. 기존의 TF-IDF는 특정 문서를 대상으로 하였지만, 본 연구에서는 특정 조항(article), 특정 절(section), 특정 장(chapter)에 한정되어 사용된 용어들을 탐색하기 위해 아래와 같은 조정 TF-IDF 지표를 개발하였다.

$$\text{출현 조항 지수} = \frac{\text{출현 조항 수}}{\text{전체 조항 수}} \tag{2}$$

$$\text{출현 절 지수} = \frac{\text{출현 절 수}}{\text{전체 절 수}} \tag{3}$$

$$\text{출현 장 지수} = \frac{\text{출현 장 수}}{\text{전체 장 수}} \tag{4}$$

$$\text{역 출현 지수}(IFI) = \frac{1}{\text{출현 조항 지수} \times \text{출현 절 지수} \times \text{출현 장 지수}} \tag{5}$$

수식 (2)~(4)는 IDF에 해당하는 조정 부분으로, 최종적으로 (5)와 같은 역출현지수(IFI: Inversed Frequency Index)를 개발하였다. 이는 문서의 구조가 다단계로 이루어져 있어, 전체의 구조상 비율을 종합적으로 반영하기 위해서이다. (2)는 전체 조항 중 특정 단어가 들어간 비율, (3)은 전체 절에서 특정 단어가 포함된 절의 비율, (4)는 전체 장에서 특정 단어가 포함된 장의 비율로써, 최종적으로 각 지수의 역수의 곱으로, (5)와 같은 IFI를 개발하였다. 역 출현 지수가 낮을수록, 모든 조항, 절, 장에 포함됐다는 의미이기 때문에, 단어가 그만큼 규칙 내 넓게 분포함을 의미한다.

다음으로, 전체 문서에 포함된 단어의 빈도수(TF: Term Frequency)를 고려하여 식 (6)을 사용하여 단어분포지표(TDI: Term Distribution Index)를 계산하였다. 단어분포 지표 값이 클수록 규칙 내 단어가 많이 등장하며, 여러 조문에서 공통으로 나타나는 단어가 아닌 특징이나 의미가 있는 단어이다. 빈도수를 고려해준 이유는 어떤 단어가 규칙 내에서 범용적으로 사용되고 있는지를 알 수 있기 때문이다.

$$TDI = IFI \times TF \quad (6)$$

다음으로, 단어분포 지표 및 빈도수의 정규화를 식 (7)과 (8)을 사용하여 수행하였다.

$$\text{정규화단어분포지표}(NTDI) = \frac{(\text{해당 단어분포 지표} - \text{지표 최솟값})}{(\text{지표 최댓값} - \text{지표 최솟값})} \quad (7)$$

$$\text{정규화빈도}(NTF) = \frac{(\text{해당 단어 빈도수} - \text{빈도수 최솟값})}{(\text{빈도수 최댓값} - \text{빈도수 최솟값})} \quad (8)$$

정규화 과정을 수행한 후, 정규화 단어분포 지표(NTDI: Normalized Term Distribution Index)와 정규화 빈도(NTF: Normalized Term Frequency)를 통해 규칙 내 용어의 유형을 Fig. 2와 같이 4가지로 매핑하였다.

먼저, 핵심 용어(Key term)는 NTDI와 NTF가 모두 높은 조항으로, 특정 장에 분포하면서, 빈도수도 높은 단어를 나타낸다. 핵심 용어는 특정 작업환경에서 준수해야 하는 단어이면서, 빈도수도 높아 해당 단어가 포함되고 있는 장을 대표한다고 볼 수 있다.

전문 용어(Technical term)는 NTF는 낮고, NTDI가 높아, 특정 장에 분포하지만, 빈도수가 낮은 단어를 나타낸다. 전문 용어는 특정 작업환경에서 준수해야 하는 단어이지만, 빈도수가 낮아 해당 단어가 포함되고 있는 장을 대표한다고 볼 수 없다.

반면, 범용 용어(General term)는 NTF는 높으나 NTDI가 낮은 유형으로, 대부분 장에 고르게 분포하면서 빈도수가 높은 단어를 나타낸다. 범용 용어는 규칙 내에서 자주 등장하지만 고르게 분포되어 있어 특정 장을 대표한다고 볼 수 없으며 전반적인 작업환경을 고려해야 하는 단어라 볼 수 있다.

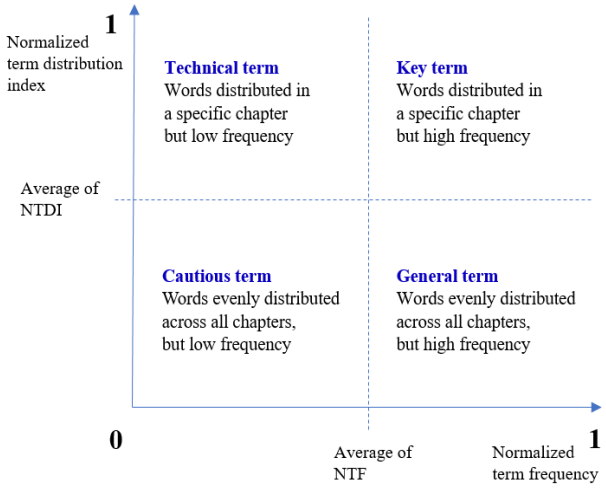

Fig. 2. Typology of term.

마지막으로, 주의 용어(Cautious term)는 대부분 장에 고르게 분포하지만, 빈도수가 낮은 단어를 나타낸다. 주의 용어는 규칙 내에서 자주 등장하지 않으나 고르게 분포되어 있다. 또한, 다른 용어에 비해 사용 문맥이 다를 수 있어, 규정된 장이나 절에 따라 주의 깊게 살펴봐야 하는 유형이다.

이와 같은 유형화는, 사용자가 스스로 규정해야 할 용어들에 대해 검색하고 살펴볼 때, 단순히 해당 절만 보면 되는지, 전체적으로 주의 깊게 살펴봐야 하는지를 유형에 따라 살펴볼 수 있다. 핵심 용어나 전문 용어는 비교적 그 용어가 처음 발견된 그 특정 부분만 살펴보면 문제가 없으나, 주의 용어의 경우에는 그 절뿐만 아니라 다른 절에서도 규정된 규칙을 검색해본 뒤 준수해야 한다. 즉, 사용자는 단어의 유형을 통해 해당 단어가 특정 작업환경에서 준수해야 할 용어인지, 전반적인 작업환경에서 준수해야 할 용어인지 구분하여 판단해야 한다.

4. 연구결과

4.1 Module 1 - 단어 간 지표 비교

수집된 조문 데이터의 전처리를 통해 총 2,545개의 단어가 추출되었다. 산업안전보건기준에 관한 규칙 내에서 가장 많이 등장하는 단어는 사업주(932회)이다. 그리고 작업(881회), 근로자(700회), 설치(643회) 등의 단어가 등장하였다. 단어의 빈도수를 기준으로, 단어분포 지표의 정규화 결과는 Table 1과 같이 요약되었으며, 단어들의 평균 사용량은 67회이며, NTF는 0.052, NTDI는 0.021이 평균으로 드러났다. 최대 사용빈도 단어나 장이나 절, 조항에 많이 분포된 단어들이 그만큼 높은 수치를 기록하고 있음을 보인다.

구체적으로 NTDI의 상위, 하위 값을 가지는 용어들을 살펴보면, Table 2와 같이, 특정 장, 절, 조항에서 언급된 정도가 가장 높은 용어는 로봇이었으며, 소음, 거푸집, 잠수작업, 동바리, 베릴륨 등이 순서로 나타났다. 반면, 여러 장, 절, 조항에서 거쳐 나오는 용어로는

Table 1. Result of NTF and NTDI

	Frequency	NTF	TDI	NTDI
Average	67	0.052	295,077	0.021
Maximum	932	1	13,812,050	1
Minimum	20	0	1,120	0

Table 2. Normalization term distribution index(top, bottom 10)

Term	NTDI	Term	NTDI
Robot	1.000	Business owner	
Noise	0.309	Revision	0
Mold	0.243	Worker	
Diving	0.232	Use	
Shore	0.214	Action	
Beryllium	0.143	Concern	~
Decompression	0.134	Prevention	
Diver	0.132	Occurrence	
Asbestos	0.131	Work	0.0001
Train	0.127	Installation	

Table 3. Comparison of term distribution index for shore and fall

Term	Frequency	NTF	NTDI
Shore	30	0.015	0.243
Fall	30	0.015	0.002

사업주, 개정, 근로자, 사용 등 규칙 내 조항에 전반적으로 사용되는 용어들이 도출되었다. 이는 지표 값을 통해 전문적인 용어와 일반적인 용어를 구별하는데 효과적임을 보여주고 있다.

예를 들어, 사망사고가 많이 나타나는 재해 발생형태인 추락과 함께, 같은 빈도수를 나타내는 동바리를 Table 3과 같이 비교하였다. 빈도수보다 정규화 단어분포 지표에서 큰 차이를 보이며, 실제로도 동바리(shore)는 규칙 내 1개의 장에 분포하고 있었으며, 추락(fall)은 10개의 장에 분포하고 있음을 알 수 있었다. 지표만으로 각 용어가 특정 장이나 절, 조항에만 나타나는지, 전체적으로 많이 나타나고 있는지를 확인할 수 있는 부분이다.

Table 4에서는 앞서 살펴본 용어의 유형에 해당하는 사례들을 도출한 결과이다. 먼저, 핵심 용어에는 '석면, 잠수작업자'가 있다. 석면은 실제로 제3편의 제2장 제6절에만 특수하게 규정되어 있으며, 용어의 활용도도 전체적으로 많음을 살펴볼 수 있다. 다음으로, 전문 용어에는 '로봇, 소음, 거푸집, 동바리, 베릴륨, 감압, 열차, 방사선, 안전밸브, 크레인' 등의 단어가 있다. 이 용어들은 특정 장이나 절에서만 사용되나, 그 언급의 빈도가 핵심 용어보다는 전체적으로 적은 경우이다. 범용 용어에는 '사업주, 작업, 근로자, 설치, 사용, 조치, 장소, 위험, 이상, 방지' 등의 단어가 있다. 즉, 일반적인 동사 위주로 규정되어 있다. 마지막으로, 주의 용어에는 '기준, 오염, 재료, 접촉, 노출, 제조, 운전, 조립, 추락, 국소배기장치' 등의 단어가 있으며, 언급된 건수도 적으면서, 모든 장에 분포된 경우를 의미한다. 이 용어들은 사용횟수는 적으나, 넓게 분포되어 있어 각 장 간의 규정 근거를 유의 깊게 살펴볼 필요가 있다.

Table 4. Example of representative terms in each type

Types	Term
Key terms	Asbestos, Diver
Technical terms	Robot, Noise, Mold, Shore, Beryllium, Decompression, Train, Radiation, Safety valve, Crain
General terms	Business owner, Work, Worker, Installation, Use, Action, Place, Danger, Abnormality, Prevention
Cautious terms	Standard, Pollution, Material, Contact, Exposure, Manufacture, Driving, Assembly, Fall, Local exhaust system

4.2 Module 2 - 단어 연관성 및 시각화

4.2.1 단어 연관성 분석

본 연구에서 단어 간 연관성 분석은 Word2Vec을 활용하여 수행하였다. Word2Vec은 도출한 단어에 임베딩 벡터를 부여해 단어 간의 유사도를 알 수 있다. 문장에서 주변 단어 또는 중심 단어(Target word)를 예측할 때 사용되며, Fig. 3과 같이 Sliding window 방식으로

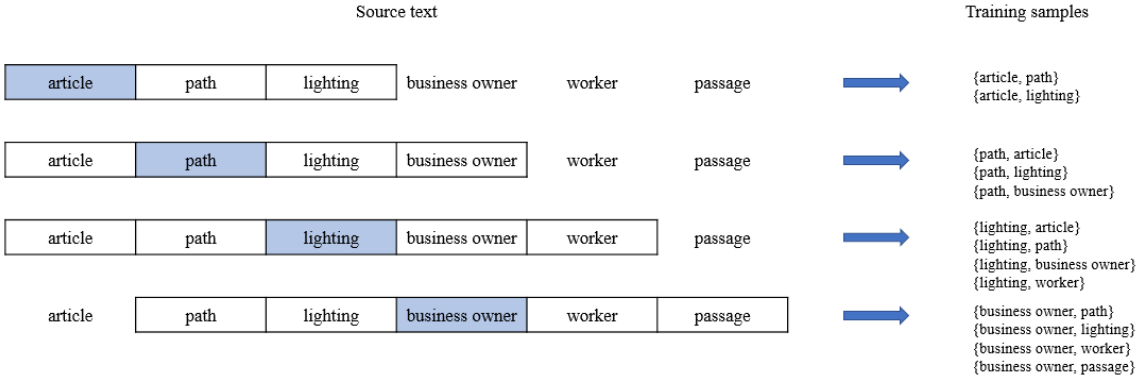

Fig. 3. Sliding window process (window=2).

Table 5. Embedding vector using Word2Vec

Term	Dimension			
	V1	V2	V3	V4
Business owner	0.073394	0.140322	-0.13751	0.22179
Worker	0.054613	0.051329	-0.11062	0.191942
Workshop	0.14326	0.276784	-0.37953	0.493998
Danger	0.121434	0.068634	-0.06925	0.329303
Floor	0.097688	0.514935	-1.11199	0.300977
Product	0.048245	-0.9661	-0.09166	0.727609

학습한다[10]. Window는 중심 단어 주위에 있는 단어이며, 값이 클수록 훈련에 고려하는 단어가 증가하여 단어의 의미적인 정확도가 증가한다.

또한, Fig. 3에서 조항(article)이란 단어는 분석과정에서 필요치 않은 공통사항이므로 이전의 전처리 절차에서 제거하였다. 한글의 경우 조사, 수사, 관형사 등을 제거하고 임베딩 벡터를 부여해야 한다.

Python Gensim package[11]의 Word2Vec 클래스를 사용하여 앞서 추출하였던 모든 단어에 임베딩 벡터를 부여하였다. 설정값은 100차원, Window는 5로 설정하였다. 이러한 설정은 관련 연구를 참고하여[12,13] 차원의 크기는 100~300차원, Window는 5일 때 가장 좋은 성능을 보여, 본 연구에서도 이를 적용하였다. 임베딩 벡터로 구성된 2,545×100 크기를 가진 데이터 모델을 Table 5에서 간략히 나타내었다. 2,545는 단어의 총 개수, 100은 차원의 크기를 의미한다. 이러한 임베딩 벡터 모델을 통해 단어 간의 유사도를 파악할 수 있다.

4.2.2 분석 결과 시각화

임베딩 벡터로 구성된 데이터 모델은 고차원이기 때문에 t-SNE 방법론을 사용하여 3차원으로 시각화하였다. 시각화는 구글에서 지원하는 Embedding projector[14]를 사용하여 진행하였다.

앞서 Module 1에서 지표 간 비교를 다루었던 추락은 주의 용어로써, 사용자가 유의 깊게 살펴봐야 하는 용어이기도 하며, 사고사망자가 가장 많이 발생하는 재해형태[15]이므로 Module 2에서도 다루어 보았다.

추락과 연관된 단어를 시각화한 결과는 Fig. 4와 같다. 코사인 유사도를 기준으로 한 추락의 연관단어로는 '안전난간, 안전대, 설치요건, 탑승설비, 안전망' 등이 있다. 또한, 추락과 단어분포 지표 비교를 수행한 동바리의 연관단어는 Fig. 5와 같다. 동바리의 연관단어로는 '거푸집, 연결재, 파이프 서포트, 깔목, 교차가새' 등이 있다.

이와 같이 탐색하고자 하는 용어의 연관단어를 직관적으로 제공함으로써, 산업안전에 대한 전문성이 부족한 사용자의 경우 연관된 단어 위주로 우선으로 검색해야 할 조문을 파악하는 것에 도움을 줄 수 있다. 또한, 안전관리자의 경우 찾고자 하는 단어의 연관단어를 파악함으로써 추가적인 안전보건조치를 수행할 때 고려해야 할 사항에 대한 정보를 얻을 수 있다. 즉, 추락과 관련하여 난간이나 안전대, 설치요건, 탑승설비, 안전망을 하나의 관리집합(set)으로 확정할 수 있으며, 동바리 역시 거푸집, 연결재, 파이프 서포트, 깔목, 교차가새 등을 관리집합으로 통합하여, 규칙 통제의 효과성을 높일 수 있다.

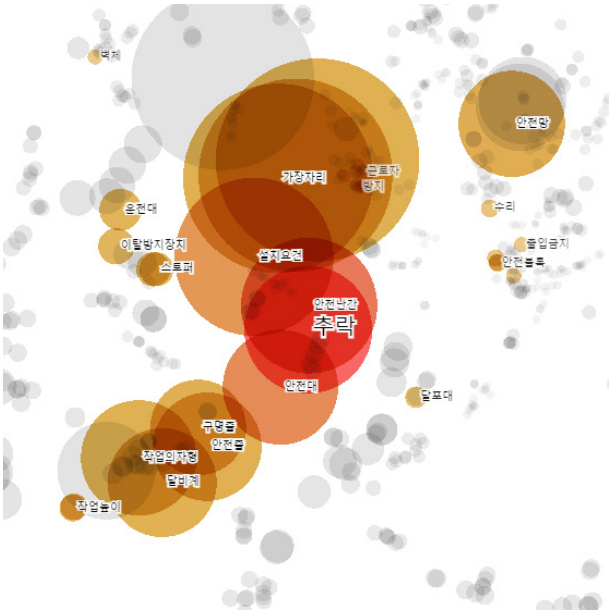

Fig. 4. Associated words for fall.

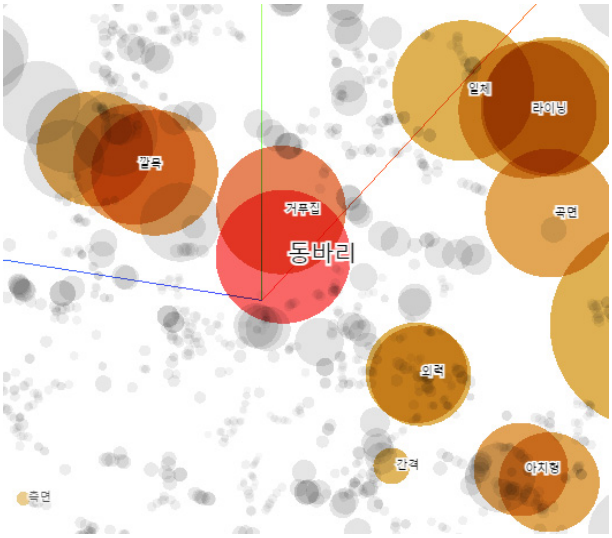

Fig. 5. Associated words for shore.

5. 결 론

본 연구는 산업안전보건기준에 관한 규칙 탐색의 접근성 향상을 위해 다음과 같이 연구를 진행하였다. 먼저, 국가법령정보센터에서 산업안전보건기준에 관한 규칙을 수집하고 텍스트 마이닝을 활용하여 규칙 내 단어를 추출하였다. 다음으로 TF-IDF를 개량한 단어분포 지표를 개발하여 용어의 유형을 4가지 형태로 제시하였다. 제시된 형태는 핵심 용어, 전문 용어, 주의 용어, 범용 용어로 나타내었다. 이후, 추락과 관련한 용어들의 주변 단어들을, Word2Vec을 활용하여 단어의 연관성을 분석하고 t-SNE 방법론으로 시각화하였다.

Module 1에서는 단어분포 지표를 통해 전문적인 용어와 일반적인 용어를 구별하는데 효과적임을 보여주고 있다. 또한, 용어의 유형을 4가지 형태로 제시하여 사용자가 스스로 규정해야 할 용어들에 대해 검색하고 살펴볼 때, 단순히 해당 절만 살펴보면 되는지, 전체적으로 주의 깊게 살펴봐야 하는지를 유형에 따라 고려할 수 있다. 즉, 법령을 탐색하는 사용자가 특정 작업환경에서 준수해야 할 용어와 전반적인 작업환경에서 준수해야 할 용어를 구분하여 판단할 수 있게 하였다.

Module 2에서는 단어마다 연관단어를 직관적으로 제공함으로써 사용자가 산업안전에 대한 전문성이 부족할지라도 연관된 단어 위주로 우선으로 검색해야 할 조문을 파악하는 것에 도움을 줄 수 있다. 또한, 안전관리자는 찾고자 하는 단어의 연관단어를 파악함으로써 추가적인 안전보건조치를 수행할 때 고려해야 할 사항에 대한 정보를 얻을 수 있다.

그러나, 본 연구에서 사용자가 단어분포 지표를 통해 법령을 탐색하였을 때의 시간적, 이해적 이점을 파악하지는 못하였다. 관련 연구를 지속하기 위해서는 사용자의 직접적인 사용성 실험을 통해 단어분포 지표의 효율성을 파악하는 것이 필요하다.

중소규모 사업장은 전담 안전인력이 배치되어 있는 경우가 많지 않고, 특히 소규모 사업장에서는 안전업무를 다른 업무와 겸직시키는 경우가 많다. 중대재해처벌법이 시행된 이후로 50인 미만 중소규모 사업장에서는 법령 이해와 대응이 필요해진 시점에서, 본 연구는 법령에 대한 기본검색과 구조, 규칙 간 연관조치사항을 살펴볼 수 있도록 기여하였다. 또한, 본 연구에서 얻은 결과를 통해 안전관리자뿐만 아니라 안전 분야에 종사하는 비전공자나 일반인 등의 사용자가 효율적으로 법령의 구조를 파악하고 탐색하여, 교육훈련과 더불어 자체적인 학습을 하는데 이바지할 수 있으리라 생각한다. 한국산업안전공단에서 배포한 '안전보건법령 스마트검색'과 병행하여 사용한다면 더욱 효과적인 정보 획득이 가능할 것이다. 마지막으로, 산업안전보건기준에 관한 규칙뿐만 아니라 타 법령에도 적용하여 법령 탐색과 관련한 분석에 기초 자료로 사용되기를 기대한다.

Acknowledgement: This work was supported by the National Research Foundation of Korea(No. NRF-2020 R1C1C1007302). 이 성과는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1C1C1007302).

※ 본 연구는 부경대학교 정재호의 석사 논문을 기반으로 작성하였다.

References

1) S. Wang, "Analysis of Penalties Imposed in Violation of Industrial Safety and Health Act", Occupational Safety and Health Research Institute, 2018.
2) N. Kim and H. Kim, "A study on the Law2Vec Model for Searching Related Law", Journal of Digital Contents Society, Vol. 18, No. 7, pp. 1419-1425, 2017.
3) Y. Suh, "A study on Visualizing Law for Universal Understanding of Occupational Safety and Health Act", Occupational Safety and Health Research Institute, 2021.
4) A. Aizawa, "An Information-theoretic Perspective of Tf-idf Measures", Information Processing and Management, Vol. 39, No. 1, pp. 45-65, 2003.
5) S. Lee and H. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF", The Journal of Society for e-Business Studies, Vol. 14, No. 4, pp. 59-73, 2009.
6) T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv preprint arXiv:1301.3781, 2013.
7) T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", Advances in Neural Information Processing Systems 26, pp: 3111-3119, 2013.
8) L. Ma and Y. Zhang, "Using Word2Vec to process big text data", IEEE International Conference on Big Data, 2015.
9) E. Park and S. Cho, "KoNLPy: Korean natural language processing in Python", Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology, pp. 133-136, 2014.

10) S. Kang, S. Chang, J. Lee and Y. Suh, “Structuring Risk Factors of Industrial Incidents Using Natural Language Process”, J. Korean Soc. Saf., Vol. 36, No. 1, pp. 56-63, 2021.
11) R. Rehurek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora”, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta Malta, pp. 45-50, 2010.
12) S. Choi, J. Seol and S. Lee, “On Word Embedding Models and Parameters Optimized for Korean”, Korean Language information Science Society, pp. 252-256, 2016.
13) H. Kang and J. Yang, “Optimization of Word2vec Models for Korean Word Embeddings”, Journal of Digital Contents Society, Vol. 20, No. 4, pp. 825-833, 2019.
14) D. Smilkov, N. Thorat, C. Nicholson, E. Reif, Fernanda B. Viégas and M. Wattenberg, “Embedding Projector: Interactive Visualization and Interpretation of Embeddings”, arXiv:1611.05469, 2016.
15) Ministry of Employment and Labor, “Statistical Survey and Analysis of Industrial Disasters”, 2021.