공 학 박 사 학 위 논 문

텍스트 분석과 기계학습 접근법을 활용한 건설업 재해분석

2021년 8월

부 경 대 학 교 대 학 원

안 전 공 학 과

강 성 식

공 학 박 사 학 위 논 문

텍스트 분석과 기계학습 접근법을 활용한 건설업 재해분석

지도교수 장 성 록

이 논문을 공학박사 학위논문으로 제출함.

2021년 8월

부 경 대 학 교 대 학 원

안 전 공 학 과

강 성 식

강성식의 공학박사 학위논문을 인준함.

2021년 8월 27일

위 원 장 공학박사 이 의 주 (인)

위 원 공학박사 박 현 곤 (인)

위 원 공학박사 서 용 윤 (인)

위 원 공학박사 진 상 은 (인)

위 원 공학박사 장 성 록 (인)

목 차

제 1 장 서 론 ····· 1

1.1 연구배경 ····· 1

1.2 연구 필요성 및 목적 ····· 4

1.3 연구내용 및 범위 ····· 9

제 2 장 이론적 배경 ····· 12

2.1 산업재해통계 ····· 12

2.1.1 전체 재해 현황 ····· 12

2.1.2 건설업 재해 현황 ····· 13

2.2 산업재해조사표 ····· 15

2.3 텍스트데이터 전처리 ····· 20

2.3.1 텍스트 분석 ····· 20

2.3.2 차원 축소 ····· 25

2.3.3 데이터 분할 ····· 27

2.4 기계학습 방법론 ····· 29

2.4.1 비지도학습 방법론 ····· 31

2.4.2 지도학습 방법론 ····· 37

제 3 장 건설업 재해문서의 비지도학습 ····· 46

3.1 위험요인 연관성 분석을 위한 SOM 분석 방법 ····· 47

3.1.1 재해문서의 텍스트 데이터 전처리 ····· 49

3.1.2 SOM 분석 방법 ····· 52

3.2 재해문서의 군집화 결과 ····· 53

3.2.1 텍스트 분석 결과 ····· 53

3.2.2 SOM 분석 결과 ····· 54

3.2.3 위험요인지도의 해석 ····· 56

3.2.4 도출된 군집의 연도별 분석 ····· 63

3.3 소결 ····· 67

제 4 장 건설업 재해문서의 지도학습 ····· 70

4.1 위험성 키워드 도출을 위한 재해문서 분류분석 방법 ····· 71

4.1.1 위험성 키워드 도출을 위한 텍스트데이터 전처리방법 ····· 72

4.1.2 분류 모델별 분석방법 ····· 75

4.2 재해문서의 분류 결과 ····· 78

4.2.1 텍스트 분석 결과 ····· 78

4.2.2 PCA를 활용한 재해문서의 특성 선정 결과 ····· 79

4.2.3 k-fold cross validation을 활용한 데이터 분할 결과 ····· 81

4.2.4 모형별 분류 결과 ····· 82

4.2.5 분류 키워드 분석 결과 ····· 85

4.2.6 오분류된 재해문서의 키워드 분석 ····· 89
4.3 소결 ····· 93

제 5 장 결론 ····· 95
5.1 연구 기여점 ····· 97
5.2 연구의 제한점 및 추후 연구 ····· 100

참고문헌 ····· 102
부록 ····· 110
A. Keyword conversion table ····· 110
B. Analysis code ····· 111

Figures

Fig. 1 Flowchart for this study ····· 11
Fig. 2 Accident rate trend from 2001 to 2020 ····· 14
Fig. 3 Trend of the death rate from 2001 to 2020 ····· 14
Fig. 4 Number of collected industrial accident survey ····· 16
Fig. 5 Industrial accident survey table ····· 19
Fig. 6 Flowchart for text analysis process ····· 21
Fig. 7 Partition of data set. ····· 28
Fig. 8 Example for clustering data ····· 30
Fig. 9 Example for classifying data ····· 30
Fig. 10 Conceptual scheme of self-organizing map ····· 35
Fig. 11 Conceptual scheme of Logistic regression ····· 39
Fig. 12 example for decision tree analysis ····· 40
Fig. 13 Conceptual scheme of neural network analysis ····· 41
Fig. 14 Support vector of linear classifier ····· 42
Fig. 15 Flowchart for the methodology of unsupervised learning ····· 47
Fig. 16 Number of accident documents ····· 49
Fig. 17 Word cloud of risk factor keywords ····· 51
Fig. 18 Result of document-term matrix ····· 53
Fig. 19 SOM-based risk factor map ····· 55
Fig. 20 Major clusters of accident documents ····· 59
Fig. 21 Trends in the number of disasters in the cluster by year ····· 64
Fig. 22 Dynamic analysis for keywords change in each cluster ····· 66

Fig. 23 Flowchart for the methodology of supervised learning ····· 71
Fig. 24 Document-term matrix using TF-IDF ····· 78
Fig. 25 Variance for each PC ····· 79
Fig. 26 Document-PC matrix for PCA ····· 80
Fig. 27 Differentiation process of decision tree ····· 88

Tables

Table 1 Legal contents related to industrial accident survey table ····· 17
Table 2 Previous research on disaster documents using text mining ··· 24
Table 3 Previous research using unsupervised learning ····· 33
Table 4 Previous research using supervised learning ····· 38
Table 5 Confusion matrix ····· 44
Table 6 Analysis of large sells ····· 57
Table 7 Number of cells and accident cases each cluster ····· 58
Table 8 Analysis of major clusters ····· 62
Table 9 Number of training and test set ····· 81
Table 10 Confusion matrix and degree of accuracy for each methodology ····· 84
Table 11 Classification factors of logistic regression ····· 86
Table 12 Differentiation order and classification factors of decision tree · 88
Table 13 Definition of misclassification and implications of prediction and management ····· 90
Table 14 Contributions of this study ····· 99

Text mining and machine learning approach to analyzing industrial incidents in construction sites

Sungsik Kang

Department of Safety Engineering, Graduate School
Pukyong National University

Abstract

In the case of an occupational accident investigation, information on the accident should be prepared in an occupational accident survey table and reported to government agencies. Since 2001, more than 80,000 cases of industrial accident surveys have been collected every year, and by analyzing them, it is used to identify the current status of domestic accidents and to prepare industrial accident statistics to prevent accidents in the same type and similar industries. In particular, the disaster overview included in the Industrial Accident Survey Table includes the overall process of accidents, and since a large amount of data is accumulated, it is necessary to analyze keywords and analyze them through machine learning.

The purpose of this study is to derive keyword-type risk factors related to actual disasters using text analysis and machine learning in the disaster overview, identify the relationship between risk factors, and derive the risks of risk factors. To this end, first, unsupervised learning of high-risk fatal accidents is performed to search for risk factors affecting fatal accidents, and similar risk factors are grouped. Next, through supervised

learning of disaster documents, the risk of risk factors is identified by searching for keywords of risk factors that classify non-fatal injury and fatal injury.

This study was conducted under two main themes :

The first theme is a clustering study through unsupervised learning of the construction industry disaster overview. We collected 2,448 cases of fatal accidents in the construction industry, the industry where fatal accidents occur the most, and structured it into structured data in the form of a document-term matrix through text data preprocessing to enable machine learning. Using self-organizing map(SOM), an unsupervised learning methodology, structured data were clustered among disaster documents with similar characteristics and visualized as a risk factor map. Keyword analysis was performed based on the disaster documents clustered in each cell of the risk factor map, and clustering was additionally performed on adjacent cells to divide into 5 clusters with similar disaster characteristics. Five clusters were clustered into 1. material-oriented disaster, 2. high place moving-oriented disaster, 3. excavator and collapse-oriented disaster, 4. scaffold-oriented disaster, and 5. crane-oriented disaster. Risk factors with high relevance were derived by deriving keywords for risk factors included in each cluster and analyzing the keywords. In addition, dynamic analysis of risk factors was performed through keyword analysis by year according to cluster.

The second theme is a classification study of disaster documents through supervised learning of construction industry disaster overview. Among the collected documents, 2,853 fatal disasters and 24,133 non-fatal disaster documents in construction industry were analyzed. To perform supervised learning, text data was preprocessed using TF-IDF, and dimensionality reduction was performed through PCA. Through this, a PC-document matrix was created and then divided into a training set and a test set through k-fold validation, and a model that could classify disaster documents was created using four classification methodologies. The classification model used logistic regression, decision tree, neural network, and support vector machine. Confusion matrix was prepared for each classification model, and the accuracy of each model was evaluated

through precision, recall, and accuracy, which are accuracy indicators. Finally, keywords representing risk were derived by analyzing keywords according to the classification model, and risk factors affecting classification were derived by analyzing misclassified disaster documents in a neural network model with the highest accuracy. Misclassification can be divided into two types: Type I error and Type II error. Type I error means a high-risk injury accident, and the document that appears as an actual injury accident is incorrectly predicted as a fatal accident. One error resulted in a document containing risk factors highly related to fatal accidents. Type 2 error means accidental fatal disaster, and it is an error of misclassifying documents that appear to be actual fatal accidents as non-fatal accidents, and documents containing risk factors highly related to non-fatal accidents were derived.

The academic contribution was first presented on the usefulness of a text-based disaster summary. Through unsupervised learning of disaster overview, it is possible to more effectively analyze the correlation between risk factors that appear in the process of disaster occurrence, and to understand the risks of risk factors derived from the classification process of disaster documents. Second, a new method that can be used for safety management using machine learning methodology was suggested compared with the previous studies focused on frequency analysis that were analyzed using industrial accident statistics.

The practical contribution of this study is first, through the risk factor map created through unsupervised learning, risk factors are extracted and analyzed from text documents that are difficult for humans to interpret individually and visualized as an easy-to-understand map. Second, by identifying risk factors affecting the classification of documents derived through supervised learning, disasters occurring in the field can be analyzed and prevented in detail. It is expected to help safety managers and supervisors effectively manage safety by presenting detailed types and keywords for major factors of disasters and deriving factors related to risk factors that appear in actual sites.

제 1 장 서 론

1.1 연구배경

2020년 산업재해 현황분석에 따르면 국내 업무상 사고사망자 수는 882명, 업무상 사고 재해자 수는 92,383명으로 나타났다. 특히, 건설업의 경우, 가장 많은 사망재해가 발생하는 업종으로, 전체 업종 사망자 수의 51.9%(458명)가 발생하였다. 이와 같이, 산업재해분석은 유사, 동종업종의 사고를 예방하기 위한 목적으로 시행되고 있다. 하인리히(H. W. Heinrich)는 사망재해와 같은 대형사고는 우연히 발생하지 않으며, 사망재해 발생 이전에 수 차례의 경미한 사고가 발생함을 지적하였다. 이에 따라 사망재해를 예방하기 위하여 재해분석을 통해 경미한 사고에 대한 원인을 파악하고 개선하여 대형사고 및 인명피해를 예방할 수 있다는 이론을 제시하였다[1].

재해 발생 현황을 파악하고 유사재해를 예방하기 위하여 재해문서 자료를 지속적으로 작성 및 수집하는 것이 의무화되었다. 재해보고문서는 산업안전보건법 시행규칙 제73조(산업재해 발생 보고 등)에 따라 산업재해로 사망자가 발생하거나 3일 이상의 휴업이 필요한 부상을 입거나 질병에 걸린 사람이 발생한 경우에 산업재해가 발생한 날부터 1개월 이내에 산업재해조사표를 작성하여 관할 지방고용노동관서의 장에게 제출해야 한다. 이와 같은 재해보고문서는 재해 발생 시 현장 관리자가 재해 발생 개요를 보고하고, 재해조사를 수행하기 위해 사용되고 있다. 실제로 안전보건공단과 같은 공공기관에서는 재해조사보고서를 활용하여 규모별, 성별, 연령별, 근

속기간별, 재해 발생 시기별 등의 일반사업 현황과 재해의 발생형태별, 기인물별, 직접 원인별 등의 재해 발생 개요를 작성하고 원인분석을 실시하고 있다[2]. 이와 같이 고용노동부와 한국산업안전보건공단에서는 재해보고 문서를 수집하여 매년 산업재해통계를 작성하며, 이를 통한 재해분석으로 동종, 유사재해의 예방을 위해 사용하고 있다. 또한, 모든 재해문서는 지속적으로 축적되고 있으며, 이로 인한 데이터 분석가치가 증가하고 있다.

최근에는 축적되어있는 방대한 양의 데이터를 분석하기 위하여 기계학습을 활용한 연구가 증가하는 추세에 있다[3]. 특히, 재해문서와 같이 텍스트로 이루어진 문서를 효과적으로 분석하기 위하여 비정형데이터 분석을 활용한 기계학습 연구도 진행되고 있다. 텍스트 분석과 기계학습은 비정형데이터로 도출된 실험데이터를 통해 비지도학습(unsupervised learning)으로 기술분석을 하거나, 특허문서에 대한 자동 분류모델을 작성하는 연구, 진단 결과를 통해 병명을 예측하는 지도학습(supervised learning) 등 다양한 분야에서 사용되고 있다.

먼저, 기계학습을 활용한 주요 키워드를 도출하거나, 각 문서의 상관관계를 분석하는 연구는 비지도학습 기반의 기계학습 알고리즘이 활용되고 있다. 비지도학습은 레이블이 없는 데이터를 데이터의 특징만을 가지고 비슷한 데이터들끼리 군집화(clustering)하는 기계학습 방법론이다. 목표치가 주어지지 않고 입력값에 대한 기계학습을 수행하며, 유사한 특징을 가진 데이터끼리 군집화하여 데이터를 분석한다. 최근 산업계에서는 많은 데이터가 쏟아져 나오고 있지만, 이를 정제하고 레이블화하기 위해선 많은 비용이 소모되기 때문에 비지도 학습에 관한 연구가 활발히 이루어지고 있다.

다음으로, 기계학습을 활용한 문서분류와 관련된 연구는 지도학습 기반의 기계학습 알고리즘이 주로 적용되고 있다[4]. 지도학습은 목적변수와 입력데이터를 활용하여 생성한 훈련 데이터를 학습하고 학습된 결과를 바탕

으로 데이터를 분류하고 예측하기 위한 기계학습이며, 데이터에 목푯값이 개입되어 분석 정확도가 비지도 학습보다 높은 장점이 있다. 지도학습의 유형은 유추한 함수 중 연속적인 값을 출력하는 회귀분석(regression)과 주어진 입력 데이터가 어떤 종류의 값인지 나타내는 분류(classification)로 구분할 수 있다. 지도학습 기반의 기계학습 알고리즘은 예측 모델을 생성하여 문서분류뿐만 아니라 패턴인식과 주가 예측, 회귀분석 등 다양한 분야에서 사용되고 있다.

이와 같은 기계학습의 활용성에도 불구하고, 안전 분야에서는 재해예방을 위해 활용하고 있는 재해보고문서나 재해와 직접적인 관련이 있는 데이터들에 대하여 텍스트 분석이나 기계학습을 활용한 분석은 아직까지 미비한 실정이다.

1.2 연구 필요성 및 목적

현대 산업재해 예방의 기본이 되는 이론인 Heinrich's Law는 사고 발생에 대한 통계적 이론을 제시하고 있다. 5,000건의 사고를 분석한 결과, 대형사고 1건을 기준으로 소형사고 29건과 경미한 사고 300건의 비율을 가지는 것을 발견하였다. 이는 대형사고의 발생은 갑작스럽거나 우연히 발생하지 않으며, 반드시 사고 발생 이전에 관련 징후 및 경미한 사고가 반복되는 과정에서 초래되기 때문에 사소한 문제나 경미한 사고를 방치하지 않고 위험요인을 파악하고 개선하여 대형사고 및 인명피해를 예방할 수 있다는 안전관리 이론이다[1].

국내에서는 산업재해 예방을 위하여 재해 발생 시 의무적으로 작성 및 제출해야 하는 산업재해조사표를 수집하여 산업재해분석을 수행하고 있으며, 수집된 산업재해조사표를 바탕으로 산업재해현황과 산업재해통계를 작성 및 공시하고 있다. 그러나 현재 수행되고 있는 재해분석은 다음과 같은 한계점이 있다.

첫 번째로, 산업재해조사표의 활용방안에 대한 한계이다. 산업안전보건법 시행규칙 제4조(산업재해 발생 보고)에 따라서 산업재해로 사망자가 발생하거나 3일 이상의 휴업이 필요한 부상을 입거나 질병에 걸린 사람이 발생한 경우 사망재해는 즉시, 부상재해는 1개월 이내에 산업재해조사표를 의무적으로 작성하여 보고하여야 한다. 또한, 산업재해조사표는 재해의 전반적인 발생 과정인 재해 개요와 사업장 현황을 포함하고 있어 재해예방을 위해 필수적으로 활용되는 자료이다. 그러나 이와 같은 재해문서의 활용성에도 불구하고, 현재 재해문서분석은 산업재해조사표 수집과 관련한 관계부처에서만 분석되고 있으며, 재해 일자와 재해자 구분, 발생형태, 기인물, 업종명, 규모, 연령, 성별, 근속기간으로 유형화된 항목에 대해서만 분석이

진행되고 있다.

두 번째로, 재해문서를 활용하여 분석할 수 있는 연구 방법이 부족하다. 국내에서 발생하는 재해사례에 대하여 안전보건공단은 제조업과 건설업, 조선업, 서비스업, 직업병, 중대 산업사고, 질식·중독에 대해 공시하고 있다. 이와 같은 자료는 각 항목에 대하여 매년 10개 정도의 재해 개요 및 대책을 배포하고 있으며, 전체 재해 중 극히 일부에 대해서만 공개하고 있다. 이와 같은 이유로, 산업재해에 관한 연구는, 관계부처 내부에서 진행되거나 배포되고 있는 산업재해통계 보고서를 통한 연구가 주로 수행되고 있다. 재해문서의 분석은 텍스트로 이루어진 재해 발생 개요에서 재해 원인을 과학적으로 해석하는 데 어려움이 있으며, 텍스트 정보를 이용한 정량적인 분석은 이루어지지 않고 있다. 실제로 대부분 현장에서는, 기술통계 분석을 통하여 다양한 재해요인들의 빈도 또는 평균에 대한 분석을 시행하고 있을 뿐, 유사재해를 예방하기 위한 체계적인 분석은 이루어지지 않고 있다.

세 번째로, 산업재해의 위험요인에 대한 분석에 한계가 있다. 재해를 예방하기 위하여 필수적으로 수행되어야 하는 위험요인에 관한 연구는 중대재해를 바탕으로 분석되고 있으며, 중대 재해에서 도출된 위험요인을 통해 재해예방에 활용하고 있다. 또한 산업재해통계를 활용한 연구는 재해 형태와 기인물에 대한 위험요인 탐색만 가능하며, 위험요인 간의 연관성이나 위험성을 파악할 수 있는 실제 산업현장에서 필요한 연구는 미비한 실정이다.

이와 같은 한계점을 해결하기 위하여 다음과 같은 연구가 필요하다.

첫째, 지속적으로 수집되고 있는 재해발생개요를 활용한 연구가 필요하다. 재해 발생 개요는 재해 발생 시 현장의 작업자와 관리감독자, 안전관리자가 조사하여 서술형으로 작성되며, 수집기관의 담당자에 의해 관리되고

있다2). 또한 재해 발생 개요는 본질적으로 재해의 장소와 발생 시기, 재해 발생원인, 기인물 등 재해 발생 프로세스에 대해 분석할 수 있는 정보가 포함된 텍스트데이터이다. 또한 의무적으로 제출해야 하는 산업재해조사표의 특성으로 인해 재해문서는 지속적으로 축적되고 있어 데이터 분석 가치는 점차 증가하고 있다.

둘째, 대량의 텍스트데이터를 효과적으로 분석할 수 있는 방법론이 필요하다. 재해 개요는 사람이 직접적으로 분류하고 분석하기에 많은 시간과 인력이 필요하며, 정량적으로 분석하고 해석할 수 있는 연구가 필요하다. 컴퓨팅 시스템의 발전으로 문서 편집, 기계 제어와 같은 기능을 효율적으로 수행할 수 있을 뿐만 아니라 데이터 처리, 정보처리 등 데이터를 효과적으로 가공하여 사용할 수 있다. 기계학습은 기본적으로 컴퓨터가 어떤 작업을 하는 데 있어서 경험으로부터 학습하여 성능에 대한 측정을 향상시키는 것을 의미한다. 재해 발생 개요와 같이 많은 양의 데이터를 정량적으로 분석하기에 적합하며, 그 활용성이 점차 증가하고 있다5). 서술형 텍스트로 작성된 재해 발생 개요를 기계 학습하기 위해서는 텍스트 분석이 선행되어야 하며, 이를 통해 정량적인 데이터 분석이 가능하다. 이는 재해보고문서를 작성함에도 불구하고 안전관리자로 하여금 제대로 활용하지 못하게 하는 한계점은 물론, 대량의 데이터를 효과적으로 분석하여 관리하는 방법이 필요하다는 것을 의미한다.

앞서 언급하였듯이, 기존의 연구는 산업재해조사표에 포함되어있는 재해 개요의 활용방안이 부족하며, 산업재해의 위험요인에 대한 분석에 한계점이 있다. 기존의 산업재해분석의 한계점을 해결하기 위하여, 재해 개요의 활용방안과 새로운 재해분석 방법을 제시하고 산업현장에서 발생하는 위험요인을 도출하여 위험요인 간의 연관성을 파악하고 위험성에 따른 재해의 특성을 파악하는 것이 필요하다. 이를 위해 본 연구는 두 가지의 목적을

가지고 연구를 수행하였다. 첫째, 재해보고문서의 텍스트 정보에서 재해와 관련된 위험요인을 정량적으로 도출하고, 문서를 특성에 따라 군집할 수 있는 비지도학습 방법론을 활용하여 유사한 특성을 가진 위험요인을 도출하고 위험요인의 관계를 분석하고자 한다. 둘째, 위험요인의 위험성을 파악하기 위하여 사망재해와 부상재해를 문서의 특성에 따라 분류할 수 있는 지도학습을 활용하여 재해에 영향을 미치는 위험요인을 도출하고자 한다.

이와 같은 목적을 수행하기 위하여 첫 번째로, 위험도가 높은 사망재해에 대한 비지도학습을 수행하여 사망재해에 영향을 미치는 위험요인에 대한 탐색이 필요하다. 통계적 학습 방법인 기계학습 중 비지도학습을 활용하여 재해문서를 군집화할 수 있으며, 단편적으로 나타나는 재해의 위험요인에 대한 연관성을 분석하여 동시에 발생할 수 있는 재해의 위험요인을 찾을 수 있다. 이를 위해, 먼저 텍스트 분석을 활용하여 서술형으로 작성된 재해 발생 개요를 컴퓨터로 분석할 수 있도록 구조화한다. 다음으로, 구조화된 재해 데이터를 비지도학습을 활용하여 사망재해문서를 유사하게 발생한 재해끼리 군집화한다. 군집화된 문서 군에서 위험요인을 나타내는 키워드를 탐색하여 각 군집에 따른 키워드 분석을 수행하여 위험요인 간의 연관성을 도출한다. 이와 같은 연구를 진행하여, 건설업 사망재해문서가 포함하고 있는 위험요인을 탐색하고, 인접한 군집에 대한 키워드 분석으로 재해 발생 과정에서 유사하게 나타나는 위험요인을 도출하고자 한다.

두 번째로, 부상재해와 사망재해에 대한 지도학습을 활용하여 재해의 위험성을 분류하는 위험요인 키워드를 탐색하고자 한다. 본 절의 서두에서 설명한 Heinrich' law는 사망재해 예방을 위해, 사망재해에서 위험요인을 도출하는 것 뿐만 아니라, 그 전에 일어난 부상재해를 분석하고 사망재해와 비교하는 것이 필요하다는 것을 시사하고 있다. 또한, 하인리히가 주장하고 지금까지 재해관리의 중요한 이슈인 사망재해와 부상재해의 관계에

초점을 맞추어, 사망재해와 부상재해를 분류하고 두 재해 간의 위험성 및 위험요인을 비교·분석한 연구는 아직까지 미비한 실정이다. 재해는 기인물이나 재해 형태 등 다양한 요소들로 인해 발생하며, 사고의 결과는 사망 또는 부상으로 다르게 나타난다. 따라서, 두 재해 간에 위험요인의 차이를 파악하고 부상재해가 사망재해로 이어지지 않는 시사점을 파악해야 한다. 이를 위해, 재해보고문서에 포함된 재해 발생 개요의 내용에 대하여 텍스트 분석과 기계학습을 활용하여 사망재해와 부상재해를 분류하고, 재해의 위험성 즉, 각각 재해 결과에 영향을 미치는 위험요인 탐색과 키워드 분석이 요구된다. 먼저, 두 재해의 재해 발생 개요에 대해서 텍스트 분석을 수행하고, 부상재해문서와 사망재해문서를 분류하는 지도학습 모델을 작성한다. 다음으로, 지도학습에서 대표적으로 사용되는 방법론을 재해문서에 적용하여 분석모델의 정확도를 도출하여 모델을 검증한다, 정확도가 가장 높은 모델의 검증과정에서 나타난 오분류 문서를 통해 부상재해와 사망재해를 분류하는 주요 키워드를 탐색하여 위험요인 키워드의 위험성을 파악하고자 한다.

1.3 연구내용 및 범위

본 연구에서 중점을 둔 연구 분야는 산업재해조사표를 통해 수집한 재해 문서 중 재해 개요에 관한 내용을 효과적으로 분석할 수 있는 위험요인 키워드 분석에 관한 연구와 사망재해문서와 부상재해문서의 위험성을 분류할 수 있는 문서분류 방법론을 통해 위험요인을 분석하는 연구이다.

본 연구의 첫 번째 목적인 유사한 재해 특성을 가진 위험요인의 도출과 연관성 분석을 수행하기 위하여 기계학습 중 비지도학습을 활용하였다. 서술형 문장으로 작성되어있는 재해 개요를 기계학습을 수행하기 위하여 텍스트 분석을 활용하여 정형데이터로 변환하였다. 위험요인 키워드 도출과 연관성을 분석하기 위하여, 목적변수가 없이 키워드를 군집화하여 분석할 수 있는 비지도학습 방법론인 self-organizing map(SOM)을 활용하여 재해 문서에 포함되어있는 위험요인 키워드를 분석하였다. 또한, 두 번째 목적인 사망재해와 부상재해의 위험성에 따른 위험요인을 도출하기 위하여, 문서의 특성에 따라 분류할 수 있는 지도학습 중 일반적으로 사용되고 있는 네 개의 방법론(logistic regression, decision tree, neural network, support vector machine)을 활용하여 재해문서의 분류 방법을 제시하였으며, 분류된 문서에서 위험요인을 도출하여 재해문서에 포함된 위험성을 분석하였다.

재해문서의 위험요인 분석을 위한 연구 절차는, 데이터 수집과 데이터 전처리, 데이터 분석, 데이터 해석의 순으로 Fig. 1과 같이 진행하였다. 먼저, 데이터는 사망재해가 가장 많이 발생하는 건설업 재해보고문서에서 재해 발생 개요를 수집하였으며, 재해 개요의 텍스트 분석을 통해 서술형으로 작성된 재해 발생 개요를 정형데이터로 변환하였다. 데이터 분석 방법은 기계학습 방법론인 비지도학습과 지도학습에 대하여 각각 진행하였으

며, 비지도학습은 재해문서의 군집화를 위하여 텍스트 분석을 통해 정형화된 데이터로 SOM을 활용하여 재해문서를 군집화하고 위험요인지도를 작성하였다. 다음으로, 지도학습은 수집된 재해 개요의 텍스트 분석과 데이터 분석의 효율을 높이기 위한 차원 축소과정을 거쳐 도출된 데이터로 재해문서의 분류를 수행하였다. 이를 통해 지속적으로 축적되고 있는 재해 개요에 포함되어있는 위험요인을 도출하고 연관성을 분석하였으며, 재해문서를 분류할 수 있는 특성을 나타내는 위험요인에 대해 분석하였다.

본 논문은 다음과 같이 구성되어 있다.

제2장 이론적 배경에서는 국내 산업재해 현황 및 산업재해조사표에 대한 법적 내용과 연구에서 활용한 텍스트데이터 전처리방법, 비지도학습, 지도학습에 대한 이론적인 설명이 이루어진다.

제3장 건설업 재해문서의 비지도학습에서는 건설업에서 발생한 재해문서를 통해 위험요인을 도출하고 연관성을 분석하기 위하여 강도가 높은 사망재해에 대한 문서를 수집하여 비지도학습 중 군집화방법인 SOM을 활용하여 위험요인 키워드 분석을 수행하였다.

제4장 건설업 재해문서의 지도학습에서는 건설업에서 발생한 사망재해와 부상재해를 분류하는 특성인 위험성에 따른 위험요인을 도출하기 위해 재해문서를 분류할 수 있는 지도학습 방법론인 logistic regression과 decision tree, neural network, support vector machine을 활용하여 재해문서를 분류하고 모델에 대한 정확도 평가를 수행하였다.

마지막으로 제5장 결론에서 연구에 관한 결과를 정리하고 연구의 기여점과 제한점, 추후 연구를 제안하였다.

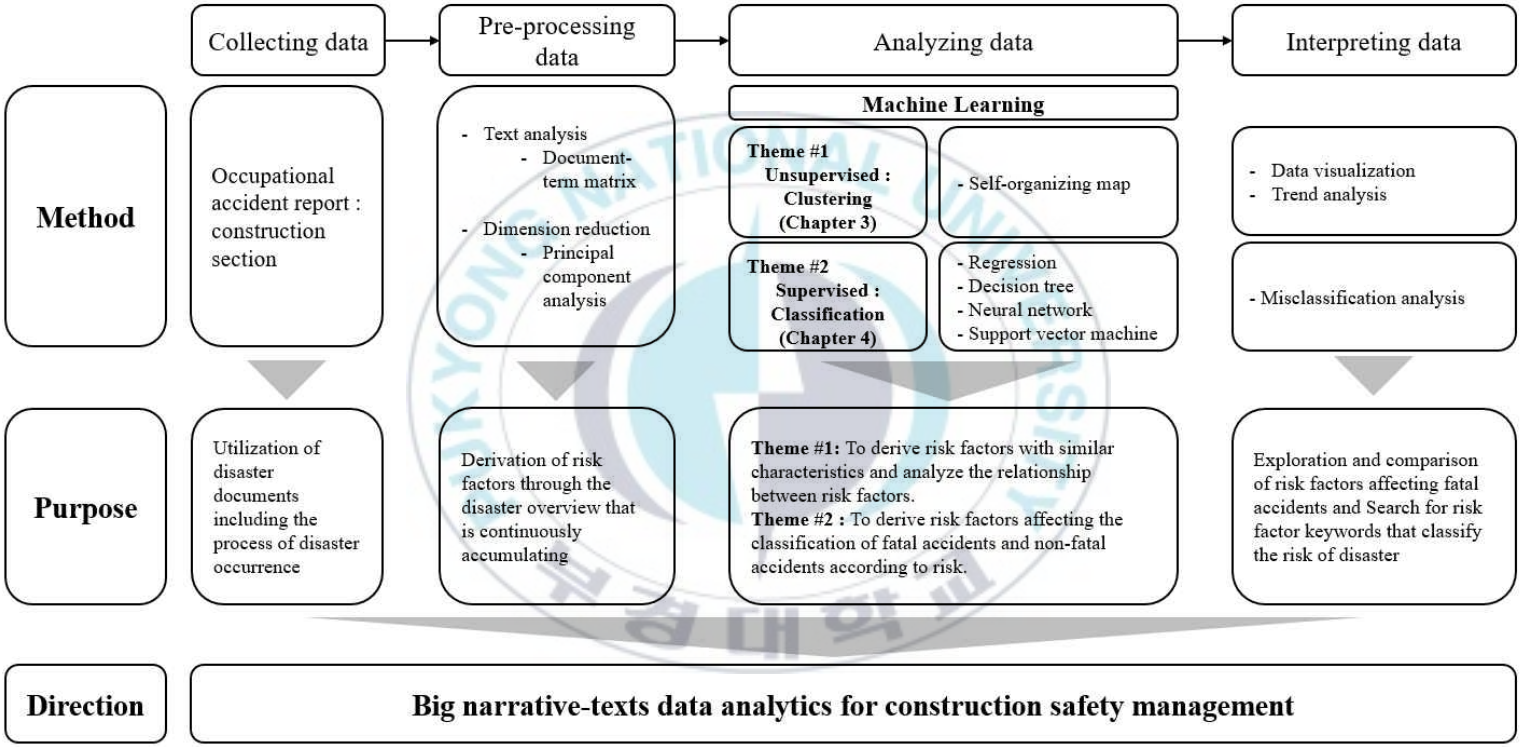

Fig. 1 Flowchart for this study

제 2 장 이론적 배경

본 장에서는 연구내용에 따른 배경 이론을 정리하였다. 먼저, 2.1절에서 국내 산업재해 현황을 살펴보고, 앞서 제시한 연구내용에 따라 수집한 텍스트데이터인 산업재해조사표에 대해 2.2절에서 설명한다. 다음으로 2.3절 텍스트데이터의 전처리 과정과 2.4절 기계학습에 대해 이론적 배경을 서술하였다.

2.1 산업재해통계

산업재해통계는 근로자가 업무와 관련하여 사망 또는 부상을 입거나 질병에 걸린 근로자를 집계한 통계이다. 재해자 수와 재해율, 사망자 수, 사망 만인율 등 여러 지표가 사용되며, 사업의 종류, 규모 및 유형별 산업재해 발생 현황을 파악하여 정부정책의 실효성 여부를 검증하고 향후 효과적인 산업재해 예방정책 수립에 활용한다.

2.1.1 전체 재해 현황

국내 산업재해 현황을 살펴보면, 2020년 재해율은 0.57%로 전년 동기 대비 0.01%p 감소하였으며, 업종별로는 기타의 사업이 37.4%(40,573명)에서 가장 많은 재해가 발생하였다. 규모별 재해율은 5인에서 49인 이하의 규모를 가진 사업장에서 43.4%(47,048명)로 가장 많이 발생하였으며, 연령별 재해율은 60세 이상 근로자가 30.6%(33,203명)로 가장 많은 재해자가 발생하였다. 2020년 사망 만인율은 1.09‰(근로자 1만 명당산업재해로 인한 사망자)로 전년 동기 대비 0.01‰p 증가하였으며, 건설업에서 51.9%(458명)로

가장 많은 사망재해자가 발생하였다. 또한, 5인에서 49인 이하의 사업장에서 45.6%(402명)의 사망재해자가 발생하였으며, 재해 발생형태로는 떨어짐으로 인한 37.2%(328명)로 사망재해가 가장 많이 발생하였다[6].

이와 같이 산업재해통계에서 나타나듯이 산업재해는 경제적인 손실뿐만 아니라, 근로자들에게도 영향을 미치며, 기업의 생산성과 사회, 국가적 측면에서도 중요한 문제로 나타나고 있다. 또한, 최근에는 산업구조 및 고용환경의 변화 등으로 인해 비정규직, 외국인, 고령, 여성 등 산재 취약계층 근로자의 증가와 소규모사업장에 대한 대기업의 하도급 증가 등 재해 유발요인은 지속적으로 증가할 전망이다. 그러므로 정부에서는 향후 정책방향으로 재해다발위험에 대한 집중 관리 등 산재 취약계층에 대한 실효성 있는 예방정책과 사업을 개발하는 것에 행정역량을 집중하고 있다.

2.1.2 건설업 재해 현황

사망재해가 가장 많이 발생하는 업종인 건설업의 2001년부터 2020년까지의 재해율과 사망 만인율의 추이는 Fig. 2와 Fig. 3과 같다. 먼저, 건설업 재해율은 2014년 0.73%로 나타났으며, 그 이후로 지속적으로 증가하고 있다. 또한, 전체 업종에서 발생한 재해율과 차이는 2020년 0.6%p로 가장 큰 폭으로 나타났다. 다음으로, 건설업 사망만인율을 살펴보면 2015년 1.47‰로 2001년 이후로 가장 낮은 수치를 나타내었으나, 2015년 이후로 지속적으로 증가하는 것을 알 수 있다. 또한, 전체 업종과 사망만인율을 비교하였을 때, 1.39‰p로 가장 많은 차이를 나타내고 있다. 이처럼 건설업은 높은 재해율과 사망만인율을 나타내고 있으며, 최근 구조물의 고층화, 대형화, 대단지화에 따라서 대형 건설장비 사용과 신재료 및 신공법 개발의 가속화 등으로 인해 중대 재해의 발생 가능성이 점차 커지고 있다[7].

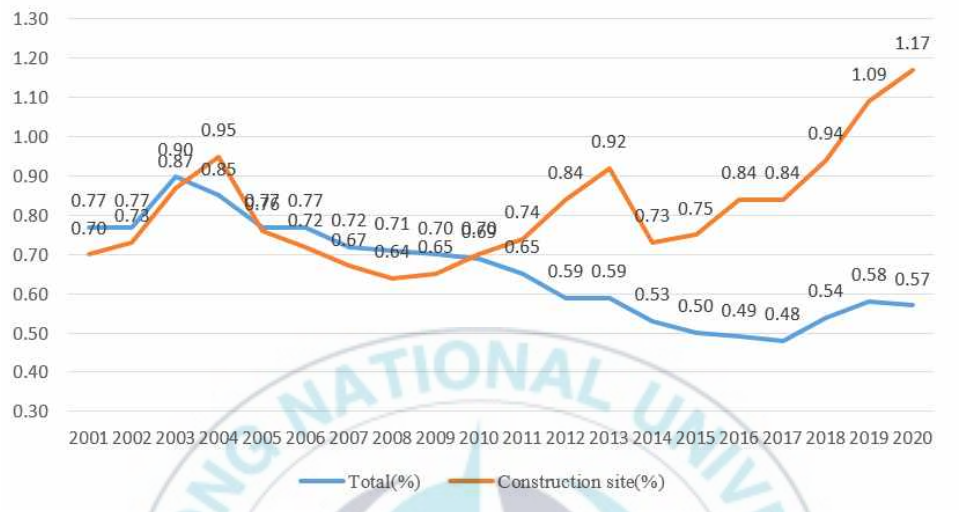

Fig. 2 Accident rate trend from 2001 to 2020

Fig. 3 Trend of the death rate from 2001 to 2020

2.2 산업재해조사표

산업재해조사표는 재해통계를 시작한 1964년부터 최근까지 지속적으로 수집되어 재해예방을 위해 연구 및 활용되고 있다. 2001년부터 2020년까지 20년 동안 1,863,723건의 산업재해조사표가 수집되었으며, 매년 수집된 산업재해조사표는 다음 Fig. 4와 같다. 2001년 이후, 매년 80,000건 이상의 산업재해조사표가 수집되었으며, 최근 3년간 더욱 증가하여 매년 100,000건 이상의 산업재해조사표가 수집되고 있다[8].

산업재해 발생 시 필수적으로 노동관서에 제출해야 하는 산업재해조사표는 다음 Table 1과 같이, 산업안전보건법 시행규칙 제4조(산업재해 발생 보고)에서 관계 법령을 규정하고 있다. 사업주는 사업재해로 근로자가 사망하거나 3일 이상의 휴업 재해를 입은 경우, 산업재해가 발생한 날로부터 1개월 이내에 산업재해조사표를 작성하여 관할 지방고용노동관서에 제출해야 한다. 또한, 재해근로자가 사업장이나 병원을 통해 근로복지공단에 제출하는 산재보험(요양, 휴업급여, 유족급여 등) 신청과는 별개로, 산업재해보험 처리가 승인되었더라도 반드시 제출해야 한다. 산업재해조사표를 미제출하거나 산업재해 발생일로부터 1개월을 초과하여 지연 제출할 경우 최고 1천만 원 이하의 과태료가 부과된다.

Fig. 4 Number of collected industrial accident survey

Table 1 Legal contents related to industrial accident survey table

산업안전보건법 시행규칙 제4조(산업재해 발생 보고) ① 사업주는 산업재해로 사망자가 발생하거나 3일 이상의 휴업이 필요한 부상을 입거나 질병에 걸린 사람이 발생한 경우에는 법 제10조제2항에 따라 해당 산업재해가 발생한 날부터 1개월 이내에 별지 제1호의2서식의 산업재해조사표를 작성하여 관할 지방고용노동관서의 장에게 제출(전자문서에 의한 제출을 포함한다)하여야 한다. ② 제1항에도 불구하고 다음 각 호의 모두에 해당하지 아니하는 사업주가 법률 제11882호 산업안전보건법 일부개정법률 제10조제2항의 개정규정의 시행일인 2014년 7월 1일 이후 처음 발생한 산업재해에 대하여 지방고용노동관서의 장으로부터 별지 제1호의2서식의 산업재해조사표를 작성하여 제출하도록 명령을 받은 경우 그 명령을 받은 날부터 15일 이내에 이를 이행한 때에는 제1항에 따른 보고를 한 것으로 본다. 제1항에 따른 보고기한이 지난 후에 자진하여 별지 제1호의2서식의 산업재해조사표를 작성·제출한 경우에도 또한 같다. 1. 법 제15조에 따른 안전관리자 또는 법 제16조에 따른 보건관리자를 두어야 하는 사업주 2. 법 제18조제1항에 따라 안전보건총괄책임자를 지정하여야 하는 사업주 3. 법 제30조의2제1항에 따라 재해예방 전문지도기관의 지도를 받아야 하는 사업주 4. 산업재해 발생사실을 은폐하려고 한 사업주 ③ 사업주는 제2조제1항제1호부터 제3호까지의 재해(이하 “중대재해”라 한다)가 발생한 사실을 알게 된 경우에는 법 제10조제2항에 따라 지체 없이 다음 각 호의 사항을 관할 지방고용노동관서의 장에게 전화·팩스, 또는 그 밖에 적절한 방법으로 보고하여야 한다. 다만, 천재지변 등 부득이한 사유가 발생한 경우에는 그 사유가 소멸된 때부터 지체 없이 보고하여야 한다. 1. 발생 개요 및 피해 상황 2. 조치 및 전망 3. 그 밖의 중요한 사항

산업재해조사표는 산업안전보건법 시행규칙 별지 제1호의2서식에서 Fig. 5와 같은 산업재해조사표 양식을 제시하고 있으며, 크게 사업장 정보와 재해정보, 재해 발생 개요 및 원인, 재발방지대책으로 구성되어 있다. 첫 번째로, 사업장 정보에는 산재 관리번호와 사업자등록번호, 사업장명, 근로자수, 업종 등이 포함되어있다. 또한, 재해자가 사내 수급인 소속인 경우와 파견근로자인 경우를 구분하여 작성하고 건설업의 경우, 공사종류와 공정률을 작성해야 한다. 두 번째로, 재해정보는 재해자에 대한 정보이며, 성명과 성별, 근속기간, 고용형태, 근무형태, 상해 종류, 상해 부위, 휴업 예상일수 등이 포함되어있다. 상해 종류는 재해로 발생한 신체적 특성 또는 상해형태를 뜻하며, 상해 부위는 재해로 피해가 발생한 신체 부위를 의미한다. 세 번째로, 재해 발생 개요 및 원인은 발생일시와 공정을 포함한 발생 장소, 재해 관련 작업유형, 재해 발생 당시 상황의 불안전한 행동과 불안전한 상황에 대하여 재해 원인의 상세한 분석이 가능하도록 작성하여야 한다. 마지막으로, 재발방지계획은 재해 발생원인을 토대로 작성한다.

특히, 재해 발생 개요는 발생 장소와 재해 관련 작업유형, 재해 발생 당시 상황에 대해 서술형으로 작성한다. 재해 관련 작업유형은 누가 어떤 기계·설비를 다루면서 무슨 작업을 하고 있었는지에 대해 상세 기술하며, 재해 발생 당시 상황의 경우는 재해 발생 당시 기계·설비·구조물이나 작업환경 등의 떨어짐, 무너짐 등과 같은 불안전한 상태와 재해자나 동료 근로자가 넘어짐, 까임 등 불안전한 행동을 했는지에 대해 상세히 작성해야 한다.

특히, 재해 발생 개요는 전반적인 재해 발생 과정을 포함하고 있기 때문에 데이터 분석의 적용이 가능하며, 텍스트 분석을 통해 서술형으로 작성된 재해 개요를 연구할 필요가 있다.

산업재해 조사표

※ 뒤쪽의 작성방법을 읽고 작성해 주시기 바라며, [ ]에는 해당하는 곳에 √ 표시를 합니다. (앞쪽)

Ⅰ. 사업장 정보	①산재관리번호 (사업개시번호)		사업자등록번호			
	②사업장명		③근로자 수			
	④업종		소재지	(-)		
	⑤재해자가 사내 수급인 소속인 경우(건설업 제외)	원도급인 사업장명	⑥재해자가 파견근로자인 경우	파견사업주 사업장명		
		사업장 산재관리번호 (사업개시번호)		사업장 산재관리번호 (사업개시번호)		
	건설업만 작성	발주자	[ ]민간 [ ]국가·지방자치단체 [ ]공공기관			
		⑦원수급 사업장명	공사현장 명			
		⑧원수급 사업장 산재관리번호(사업개시번호)				
		⑨공사종류	공정률	%	공사금액 백만원	

※ 아래 항목은 재해자별로 각각 작성하되, 같은 재해로 재해자가 여러 명이 발생한 경우에는 별도 서식에 추가로 적습니다.

Ⅱ. 재해 정보	성명		주민등록번호 (외국인등록번호)	성별	[ ]남 [ ]여
	국적	[ ]내국인 [ ]외국인 [국적:	⑩체류자격:]	⑪직업	
	입사일	년 월 일	⑫같은 종류업무 근속기간	년 월	
	⑬고용형태	[ ]상용 [ ]임시 [ ]일용 [ ]무급가족종사자 [ ]자영업자 [ ]그 밖의 사항 [ ]			
	⑭근무형태	[ ]정상 [ ]2교대 [ ]3교대 [ ]4교대 [ ]시간제 [ ]그 밖의 사항 [ ]			
	⑮상해종류 (질병명)		⑯상해부위 (질병부위)	⑰휴업예상 일수	휴업 [ ]일
				사망 여부	[ ] 사망

Ⅲ. 재해 발생 개요 및 원인	⑱ 재해 발생 개요	발생일시	[ ]년 [ ]월 [ ]일 [ ]요일 [ ]시 [ ]분
		발생장소	
		재해관련 작업유형	
		재해발생 당시 상황	
	⑲재해발생원인		
Ⅳ. ⑳재발 방지 계획			

Fig. 5 Industrial accident survey table

2.3 텍스트데이터 전처리

2.3.1 텍스트 분석

텍스트 분석은 재해 개요와 같은 문서 형태의 비정형 텍스트데이터에서 자연어처리기법(Natural Language Processing)에 기반하여 유용한 정보를 추출하기 위해 정보를 가공하는 방법론이다. 대량의 비정형데이터에서 의미 있는 지식을 도출하고, 각 정보에 대하여 통계적인 분석을 수행하기 위해 사용되고 있다. 텍스트 분석의 과정은 다음 Fig. 6과 같이 크게 3단계로 제시한다[7]. 먼저, 텍스트 데이터를 수집하고, 텍스트를 컴퓨터가 입력값으로 받아들일 수 있도록 말뭉치(Corpus) 단위로 변환하고 자연어처리기법을 활용하여 형태소 분류과정을 거쳐 분석에 필요한 형태소를 도출한다. 자연어처리기법은 인간의 언어분석에 대한 표현을 자동화하기 위한 기법으로, 기계번역과 대화체 질의응답 시스템, 정보검색, 말뭉치 구축, 딥러닝 등 인간의 언어정보처리에 대한 원리와 이해를 위한 언어학과 뇌인지 언어정보처리 분야까지 적용되는 언어처리 기법이다[9].

자연어처리기법의 분석과정은 형태소분석(morphological analysis)과 구문분석(syntactic analysis), 시멘틱분석(Semantic analysis), 실용분석(pragmatic analysis) 과정으로 진행된다. 먼저, 형태소분석은 입력된 문자열을 분석하여 형태소라는 최소 의미를 나타내는 단위로 분리한다[10]. 이는 분석하는 언어에 따라 분석난이도가 다르며, 영어는 한글이나 일본어보다 비교적 쉽게 분석할 수 있다. 다음으로, 구문분석은 문법을 이용하여 문장의 구조를 찾아내는 과정이며, 시멘틱분석은 통사에 대한 분석결과에 해석을 추가하여 문장이 가진 의미를 분석한다. 마지막으로 실용 분석은 분석된 문장과 실제 사용되는 의미의 연관 관계를 분석하며 지식과 상식의 표

현이 요구된다.

다음으로, 분석의 효율을 높이기 위하여 도출된 데이터의 Cleaning 과정을 수행한다. 이를 통해 부적절한 단어나 분석에 활용할 수 없는 불용어, 숫자, 기호 등을 제거한다. 마지막으로 각 문서에서 기록된 말뭉치의 키워드에 따라 행에서 문서를, 열에서 키워드를 나타내는 문서-키워드 행렬(DTM; document-term matrix)을 구축한다. 이를 통해 생성된 데이터 매트릭스를 input data로 활용하여 다양한 데이터 분석 알고리즘에 적용하여 새로운 지식을 도출한다. 이와 같은 텍스트 분석 알고리즘은 특허 분석과 고객 요구 분석, 감성 분석과 같은 사회과학 분야 등에서 텍스트 자료의 정량적인 분석을 위하여 활용되고 있다.

대표적으로 활용되는 정형데이터의 형태는 키워드의 도출 빈도를 추출하거나 TF-IDF(term frequency - inverse document frequency)와 같은 키워드의 가중치로 표현된다. TF-IDF는 정보검색과 같은 텍스트 분석에서 이용하며, 여러 문서로 이루어진 문서 군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치를 의미한다[11]. TF는 특정 문서에서 특정 단어의 빈도를 나타내며, 특정 단어가 도출된 문서의 수를 나누어 계산한다. TF-IDF는 문서의 벡터값을 해당 키워드의 빈도와 문서의 역수를 곱하여 계산하며, 각 문서에 포함되어있는 단어에 높은 가중치가 곱해지고, 문서 전체에 분포되어있는 단어에는 낮은 가중치가 곱해지는 방법론이다.

Fig. 6 Flowchart for text analysis process

재해문서의 텍스트 분석과 관련된 국내·외 기존 연구는 다음 Table 2와 같이 정리하였다[12-18]. 재해문서를 활용한 텍스트 분석에 관한 연구를 살펴보았으며, 저자와 논문 키워드, 분석 대상, 활용한 방법론, 연구의 목적을 정리하였다.

먼저, 재해문서를 활용한 텍스트 분석은 국내 연구는 2018년 이후로 관련 논문이 투고되기 시작하였으며, 재해문서에서 키워드를 도출하고, 다양한 방법론을 활용하기 위해 사용되고 있다. 박기창 외 1인(2021)의 연구는 안전보건공단에서 공시하고 있는 건설업 재해사례 2,026건을 수집하여 TF-IDF로 텍스트 분석을 수행하고, 네트워크 분석으로 건설업에서 발생한 재해의 빈도와 중심성 분석을 수행하였다. 다음으로, 김범수 외 2인(2018)은 사고 종류를 분석하기 위하여, 화학 플랜트 기업에서 발생한 사고보고서 127건을 텍스트 분석하여 토픽 모델링기법 중 하나인 latent dirichlet allocation(LDA)과 특성요인도를 활용하여 연구를 진행하였다. 또한 장우현 외 1인(2019)은 미국 OSHA에서 공시하고 있는 제조업 중 화학산업의 사고문서 519건을 바탕으로 텍스트 분석을 수행하였고, 이를 통해 재해의 이상치를 분석하였다. 마지막으로 이상규(2018)는 건설업에서 발생한 재해에 대한 뉴스기사 540건을 수집하여 텍스트 분석을 수행하였으며, LDA를 활용하여 사고 발생 동향과 관련 키워드를 도출하였다. 종합적으로, 재해문서의 텍스트 분석과 관련된 연구는 사고의 유형을 분석하거나, 재해문서에서 키워드를 도출하기 위해 사용되며, 문서 자체에서 영향력이 큰 키워드를 분석하기 위해 활용되거나 분석자가 원하는 결과를 얻기 위한 통계기법을 활용하기 위해 연구되고 있다. 이와 같이 재해문서의 텍스트 분석은 수행되고 있지만 아직까지 많은 양의 재해문서를 분석한 연구는 미비한 실정이다.

해외의 경우, 재해문서를 활용하여 다양한 연구를 진행하고 있으며, 건설업에 대한 분석뿐만 아니라 철도업이나, 운송업 등 다양한 분야의 재해문서를 활용한 텍스트 분석이 수행되고 있다. 특히, D. E. Brown(2016)은 5,469건의 철도 사고에 대하여 TF-IDF를 활용하여 재해와 연관된 키워드를 도출하였으며, LDA와 random forest 방법론으로 철도업에서 발생하는 위험요인을 분석하였다. 또한 N. XU 외 4인(2021)과 S. Shi 외 5인(2019)는 최근 중국에서 발생한 재해문서를 활용하여 키워드 분석을 수행하였으며, 건설업과 운송업에 대해서 발생한 사고를 분석하였다.

재해문서에 관한 연구는 기존의 빈도분석뿐만 아니라 키워드를 도출하여 통계기법을 활용한 다양한 연구들이 진행되고 있다. 키워드 분석을 통하여 기존의 분석 방법으로는 탐색할 수 없었던 위험요인을 도출할 수 있으며, 이를 활용하여 사고의 예방과 안전관리에 사용되고 있다.

Table 2 Previous research on disaster documents using text mining

Author (Year)	Keywords	Analysis target (cases)	Methodology	Objective
K. Park and H. Kim (2021)	Construction accident, Hazard, text mining, frequency analysis, centrality analysis	Text data of accident cases in the construction industry (2,026 cases)	TF-IDF, Co-ocurrence network, Louvain algorithm	construction safety
BS. Kim, S. Chang, and Y. Suh (2018)	Accident analysis, text analytics, LDA, cause-and-effect diagram, chemical plant	Accident report in chemical plant (127 cases)	LDA, cause-and effect diagram	Accident type analysis
W. Jang and Y. Suh (2019)	Abnormal Accidents, Accident Report Documents, Anomaly Detection, Local Outlier Factor, Decision Tree	disaster report in manufacturing by OSHA (419 cases)	Local outlier factor	Discovery of anomalies in the chemical industry
S. Lee (2018)	Unstructured data, Construction Safety Accident, Topic Modeling, Latent Dirichlet Allocation (LDA), News Article, Time-Series	News related to construction safety accidents (540 cases)	LDA	Keyword classification and trend identification of accidents occurring in construction
D. E. Brown (2016)	Rail safety, safety engineering, latent dirichlet allocation, partial least squares, random forests.	rail accident report in USA (5469 cases)	LDA, random forests	Rail safety
N. XU, L. MA, Q. Liu, L. WANG, and Y. Deng (2021)	Construction safety, Automatic risk identification, Workplace accident, Text mining	metro construction accident reports in China (221 cases)	TF-IDF, information entropy weighted term frequency	Construction safety
S. Shi, D. Zhang, Y. Su, M. Zhang, M. Sun, and H. Yao (2019)	Inland ship collision, R language, Text mining, Risk factors analysis	inriver accidents' cause in china (419cases)	textmining	Risk factor analysis

2.3.2 차원 축소

데이터 분석에서 차원 축소란 다차원의 매트릭스로 이루어진 데이터를 차원을 축소하여 새로운 차원의 데이터로 생성하는 것을 의미한다. 특히, 차원의 저주(dimensionality curse) 문제를 해결하기 위한 대표적인 방법이다. 차원의 저주는 고차원의 매트릭스를 사용하여 데이터를 분석할 경우, 차원이 증가할수록 분석기법들의 성능이 급격히 떨어지는 현상을 의미한다[19]. 또한 데이터 분석의 효율을 높이기 위하여 고차원의 매트릭스를 저차원 매트릭스로 변환하는 방법이 사용되며, 효과적인 데이터 분석 기법은 분석 속도가 빠르면서 동시에 전체 차원에 대해 분석했을 때와 동일한 결과를 도출해야 한다[20]. 차원 축소는 기계학습에 활용되는 기존 데이터가 가지고 있는 특징을 최대한 보존하는 형태로 발전하고 있으며, 특히 텍스트데이터는 문서에 포함된 단어의 수가 증가할수록 매트릭스 크기의 증가폭이 커지기 때문에 차원 축소기법이 필요하다.

차원 축소에 관한 기존 연구에서는 고차원 공간에서 저차원 공간으로 변환하기 위해 다양한 수학적 변환 기법들이 활용되고 있다. 가장 일반적인 차원 축소기법으로는 주성분 분석(principal component analysis : PCA)을 이용한 기법이 있다. 또 다른 기법들로는 이산 코사인 변환(discrete cosine transform)과 이산 푸리에 변환(discrete fourier transform)을 이용한 기법 등이 있다[21]. 일반적으로 PCA나 이산 코사인 변환을 이용한 기법이 이산 푸리에 변환을 이용한 기법에 비하여 더 나은 성능을 보이는 것으로 알려져 있다.

PCA는 다변량 데이터를 효과적으로 분석하기 위한 대표적인 분석기법이며, 데이터 간의 연관성을 분석하여 연관성이 작은 고유 공간(eigen space)을 위한 고유 벡터(eigenvector)를 구하는 기법이다[22]. 또한 고윳값

(eigenvalue)을 기준으로 선택된 고유 벡터를 사용하여 연관성이 높은 고차원 벡터에서 평균 제곱의 오차를 최소화하여 저차원 벡터로 변환할 수 있다. 정형데이터에서 키워드의 특성을 나타내는 변수를 추출하거나 고차원 데이터를 저차원 데이터로 차원을 축소하기 위한 방법론으로 차원 축소와 시각화, 군집화, 압축 등 다양하게 활용된다. 또한, 데이터의 분산을 최대한 보존하면서 서로 직교하는 새 축을 찾아 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환하여 데이터 분포에서 분산력이 가장 큰 직선(혹은 평면)을 찾는 방법론이다[23]. PCA에서 사용되는 변수추출(feature extraction)은 기존 변수를 조합하여 새로운 변수를 만드는 기법으로, 기존의 변수를 선형결합(linear combination)하여 새로운 변수를 만들어내는 방법을 사용한다. PCA를 단계적으로 진행할 경우, 공분산 행렬을 통해 각 PC에 대한 eigenvalue와 eigenvector를 구할 수 있으며, 분산량에 해당하는 각각의 고윳값으로 해당 principal component(PC)가 차지하는 전체 고윳값의 비율을 구한다. 축소하는 차원의 수는 분석자가 설정하며, 주성분 벡터의 explained variance ratio를 계산하여 적절한 차원의 수를 선정한다. 각각의 주성분 벡터가 이루는 축에 투영한 결과의 분산에 대한 비율을 말하며 이를 explained variance ratio라고 하며, 주성분 벡터가 가지고 있는 eigenvalue의 비율과 같은 의미로 사용된다. 이를 활용하여 축소할 차원의 수에 따라 원 데이터에 비해 분석에서 제외되는 데이터의 분산을 도출할 수 있다. 이와 같은 과정을 통해 분석에 적합한 PC의 개수를 선정하여 차원을 설정하며, 유사한 키워드들의 조합으로 특성을 가진 축소된 차원을 선정한다.

2.3.3 데이터 분할

데이터 분할은 전체 데이터를 여러 부분으로 분할하는 것을 의미하며, 데이터 분석모델이 주어진 데이터에 대해서만 높은 성능을 보이는 문제를 방지하기 위하여 일부 데이터로 학습을 하고, 일부 데이터로 검증을 수행하는 것을 의미한다[24]. 분할 데이터의 구성은 학습용 데이터(training set)와 검증용 데이터(validation set), 평가용 데이터(test set)로 구성된다. 먼저, training set은 데이터를 학습하여 분석모델을 만드는데 직접적으로 활용되는 데이터이다. 다음으로, validation set은 과적합, 부적합 등 모형의 성능을 개선하기 위한 데이터이며, test set은 모델의 성능개선 및 적합성 검증하기 위해 사용되는 데이터를 의미한다.

데이터 분할 시 고려해야 할 사항은 training set과 test set이 전체 데이터에 대한 대표성을 가져야 하며, 과거 데이터로부터 미래 데이터를 예측하는 경우에는 데이터를 혼합하여 사용하지 못한다. 이 경우, training set에 있는 데이터보다 test set의 모든 데이터가 미래의 것으로 구성되어야 한다. 마지막으로 training set과 validation set, test set에서 데이터의 중복이 없도록 구성해야 한다[25].

전체 데이터를 training set과 test set으로 나누어 training set을 사용하여 모델을 학습시키고, test set으로 그 모델 성능을 평가한다. 그러나 전체 데이터의 일부인 test set을 사용해 모델 성능을 평가하는 것에 대한 문제는 test set이 모델에 학습되지 않기 때문에, test set의 선정에 따라 성능이 다르게 나타날 수 있어 test set에 대한 성능 평가의 신뢰성이 떨어지게 된다. 이와 같은 문제를 해결하기 위하여 k-fold cross validation을 사용하였다. 이는 모든 데이터를 최소 한 번은 test data set으로 사용하여 모델 성능을 평가하는 방법론이며, 추가적인 데이터를 수집하기 어렵거나 비용

을 많이 소모하는 경우 사용된다[26]. 또한 기존의 training set과 validation set, test set 세 개의 집단으로 분류하는 것보다 training set과 test set으로만 분류할 때 학습 data set을 더 많이 생성할 수 있기 때문에 더 정확한 분류모델을 만들기 위하여 사용된다. K-fold cross validation은 Fig. 7과 같이 전체 데이터를 k개의 그룹으로 나누어 (k-1) 개의 test fold와 1개의 validation fold로 지정하여 총 k 회의 교차검증을 통해 각 parameter를 지정하고 분할 정확도가 가장 높은 hyperparameter를 적용하여 분류모델을 평가한다. 또한, 전체 데이터에 대하여 민감도 분석을 통해 k값을 설정하여 fold의 개수에 따라 최소 및 최대 정확도를 도출할 수 있으며, 정확도가 가장 높은 k값으로 분석을 실시한다[27].

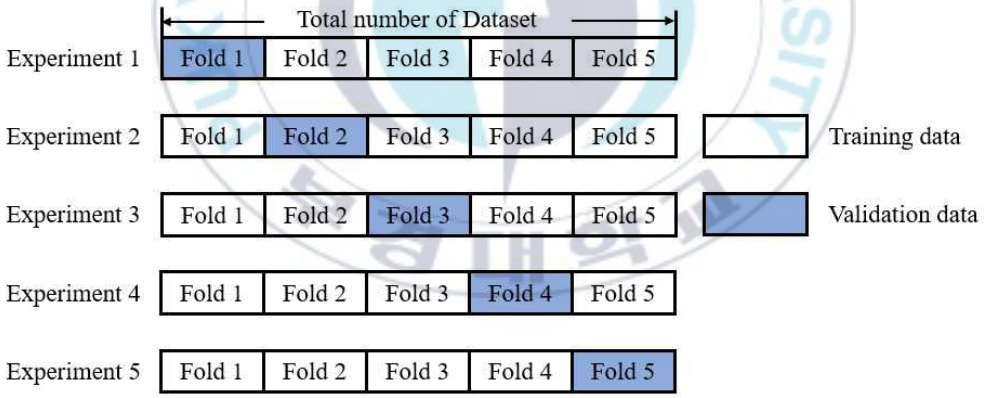

Fig. 7 Partition of data set.

2.4 기계학습 방법론

기계학습은 경험적인 데이터를 통해 자동으로 결괏값을 도출하는 컴퓨터 알고리즘이며 인공지능의 한 분야이다. 과거에는 학문에 대한 깊이는 있었으나 실생활과 다소 거리가 있는 분야였으나, 최근에는 기계학습 자체 기술과 데이터의 증가에 따른 활용방법, 하드웨어의 성능 등 관련 분야의 발전으로 실용성이 높은 방법론으로 사용되고 있다. 기계학습은 데이터의 평가를 나타내는 표현(representation)과 아직 입력되지 않은 데이터에 대한 처리를 나타내는 일반화(generalization)로 학습을 수행한다[28]. 훈련 데이터(training data)를 통해 학습된 데이터의 속성을 기반으로 앞으로의 입력될 데이터를 예측하는 것에 초점을 두고 많은 연구가 진행되고 있다. 훈련 데이터는 유한하며 그 결과가 불확실하여 알고리즘의 결과를 장담할 수 없으므로 통계적 추론과 유사함을 나타낸다[29].

기계학습에서 일반적으로 사용되는 분석 프로세스는 앞서 설명한 바와 같이 데이터 수집과 데이터 전처리를 수행하여 정형데이터를 도출한 후, 데이터 학습과 학습모델의 검증 및 평가를 수행한다. 기계학습 알고리즘을 활용하여 훈련 데이터로 학습하는 모델을 작성하고, 데이터 검증 및 평가를 위해 검증데이터를 사용하여 작성된 모델과 비교·검증한다[30].

기계학습은 학습 방법에 따라 크게 비지도학습과 지도학습, 2가지로 나누어진다. 먼저, 비지도학습(unsupervised learning)은 출력데이터에 대한 결괏값이 없을 경우, 입력 데이터들을 Fig. 8과 같이 cluster로 구분하는 학습방법이다. 이는 특정 집단을 그룹화하거나, 데이터 사이의 유사성을 확인하기 위해 사용된다. 다음으로, 지도학습(supervised learning)은 입력 데이터와 출력데이터를 통해 관계를 유추하여 올바른 해답을 제시하는 방법론이다. 지도학습을 통해 유추된 알고리즘을 통하여 연속적인 값을 예측하는

회귀분석(regression)과 어떤 종류의 값인지를 표시하는 분류(classification)를 수행할 수 있다. 예를 들어, Fig. 9는 지도학습 방법론 중 하나인 support vector machine을 도식화한 그래프이며, 출력데이터에 대한 결괏값을 가지고 있는 데이터들을 분류할 수 있는 회귀식을 찾아 각각의 데이터를 분류하는 방법론이다.

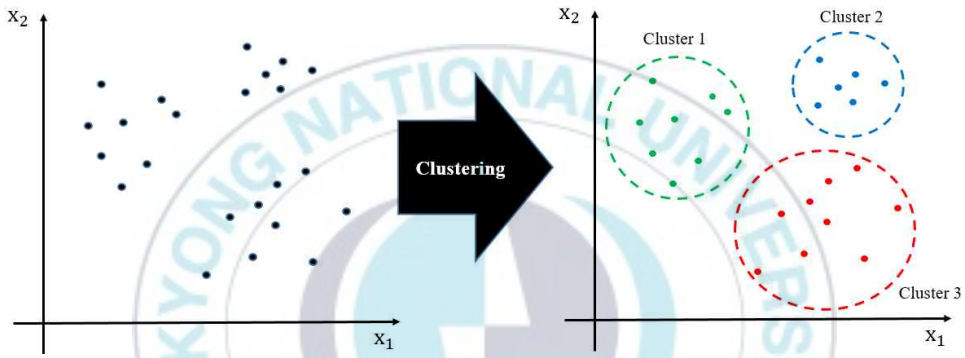

Fig. 8 Example for clustering data

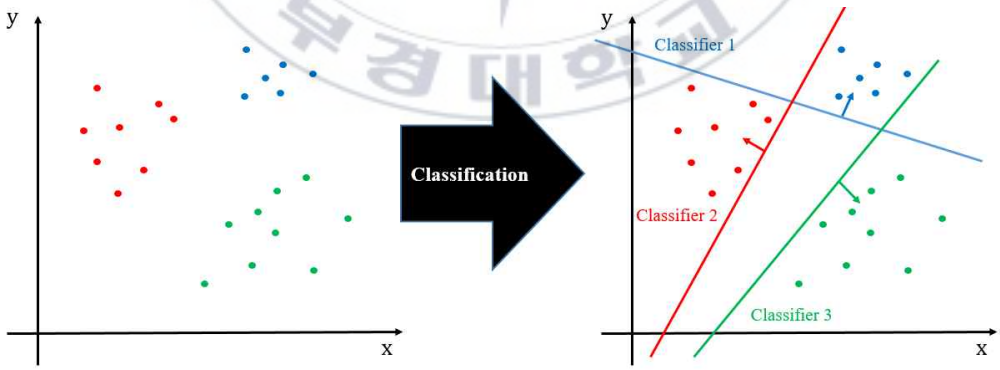

Fig. 9 Example for classifying data

2.4.1 비지도학습 방법론

비지도학습은 레이블이 없는 데이터를 데이터의 특징만을 가지고 비슷한 데이터들끼리 군집화(clustering)하는 기계학습 방법론이며, 데이터의 관계를 알지 못하는 경우에 유용하게 활용된다[31]. 목표치가 주어지지 않고 입력값에 대한 기계학습을 수행하며, 지도학습에 사용될 적절한 목표치를 찾기 위한 전처리방법으로도 사용된다. 최근 산업계에서는 많은 데이터가 쏟아져 나오고 있지만, 이를 정제하고 레이블화하기 위해선 많은 비용이 소모되기 때문에 비지도 학습에 관한 연구가 활발히 이루어지고 있다. 하지만 아직까지 비지도 학습은 레이블이 주어지는 지도학습보다 성능 면에서 떨어지는 편이며, 상용화된 대부분의 일반적인 기계학습 방법은 대부분 지도학습으로 이루어지고 있다[32],

텍스트데이터를 활용한 비지도학습은 문서 내에서 키워드를 도출하여, 유형화와 연관성 파악 등 다양한 분석에서 사용되고 있다. 다음 Table 3은 비지도학습을 통해 키워드를 군집화하여 안전관리에 활용하기 위해 진행된 연구이다[33-38]. 먼저, 국내에서 수행된 연구를 살펴보면, 손기영과 류한국(2019)의 연구는 건설업에서 발생한 추락과 관련된 1,618건의 재해정보를 수집하여 텍스트 분석을 수행하고, 비지도학습으로 재해의 연관규칙을 찾는 연구를 수행하였다. 다음으로 신명우와 서용윤(2019)의 연구는 OSHA에서 제공하고 있는 물질안전보건자료의 가이드 텍스트를 수집하여 군집화하였으며, 화학물질의 위험성에 대해 시각화 연구를 수행하였다. 국내에서 재해문서를 활용한 비지도학습 연구는 미비한 실정이며, 해외에서 진행된 기존연구는 다음과 같이 진행되었다. M. Shujiro 외 3인(2011)은 제조업 공장에서 발생한 품질 사고 데이터 54건을 텍스트 분석하고 비지도학습인 self-organizing map을 활용하여 시각화하였다. 이와 같은 연구를 통해 사

업장에서 발생하는 사고의 요인을 시각화하고, 위험관리를 위해 사용하는 방안을 모색하였다. 또한 F. Palamara 외 2인(2011)은 재해예방을 위해 재해로 이어지는 사건들을 발견하기 위해 비지도학습을 활용하여 재해분석을 수행하였다. 이탈리아의 목재산업에서 발생한 1,207건의 재해개요 데이터를 수집하였으며, self-organizing map과 k-means clustering을 활용하여 위험요인을 군집하였다. A. Chokor 외 3인(2016)의 연구에서는 OSHA의 건설업 재해문서 513건에 대하여 TF-IDF로 텍스트 분석을 수행하고 k-means clustering으로 문서를 군집화하였다. 이를 통해 안전 점검에 대한 지원과 사고 데이터베이스를 재구성하는데 있어서 재해문서의 텍스트 분석과 비지도학습의 강점을 평가하였다. 마지막으로 A. Vema와 J. Maiti(2018)의 연구는 철강제조회사에서 나타난 아차사고 8,877건을 분석하여 사고의 근본적인 원인을 파악하고자 하였다. 연구방법은 특성요인도와 TF-IDF, 계층적 군집화 방법론을 활용하였으며, 재해를 구성하는 근본적인 위험요인을 도출하였다.

이와 같이 비지도학습 방법론을 활용한 연구는 다양하게 수행되고 있으며, 특히 재해문서와 같은 텍스트데이터를 활용한 연구도 진행되고 있다. 하지만 국내 상황에 맞는 재해문서를 활용하여 위험요인을 도출하고, 이를 분석하는 연구는 미비한 실정이다.

Table 3 Previous research using unsupervised learning

Author (Year)	Keywords	Analysis target (cases)	Methodology	Objective
K. Son and H. Ryu (2019)	Construction safety, falling objects; data science, machine learning, association rules, hierarchical clustering	Disaster information caused by falling objects (1618 cases)	Association rules, Hierarchical clustering	Disaster association rule analysis
M. Shin, and Y. Suh (2019)	MSDS, MSDS map, textmining, SOM, visualization, chemicals	MSDS guide text by OSHA (539cases)	SOM	Visualization of hazardous substance information
M. Shujiro, O. Koji, S. Junko, and I. Hiroaki (2011)	Knowledge discovery, risk management, text data mining	Quality accident data in headquarters factory (54 cases)	SOM	Risk management
F. Palamara, F. Piglione, and N. Piccinini (2011)	Occupational accidents, Wood industry, Neural networks, Self-Organizing Map, Clustering algorithms	INAIL accident database (1207 cases)	SOM, k-means clustering	To discover the most common sequences of events leading to accidents for devising preventive actions.
A. Chokor, H. Naganathan, W. K.Chong, and M. E. Asmar (2016)	Accident, Construction, Data Analysis, Injury, Natural Language Processing, OSHA, Safety.	Construction accident report by OSHA (513 cases)	TF-IDF, k-means documents clustering	To Evaluate the strength of unsupervised machine learning and Natural Language Processing(NLP) in supporting safety inspections and reorganizing accidents database
A. Verma and J. Maiti (2018)	Cause and effect, root causes, text mining, clustering analysis, incident data	Steel plant incidents records (8877 cases)	Cause and effect diagram, TF-IDF, hierarchical clustering, expectation-maximization algorithm	to develop a text clustering-based cause and effect analysis methodology for incident data to unfold the root causes behind the incidents

이와 같이 선행연구에서 대표적으로 활용한 비지도학습 방법론으로는 SOM(self-organizing map)과 k-means clustering 등이 있다. SOM은 Fig. 10과 같이 투입되는 데이터들의 가중치를 조정하는 인공신경망에 기초한 기계학습의 유형으로, 지속적으로 업데이트되는 입력 데이터들을 비교사학습(unsupervised learning)으로 군집화하는 방법론이다[39]. SOM은 서로 연결된 노드들로 구성된 네트워크를 만들며, 연관성에 의해 주어진 데이터를 군집화함으로써 내부의 셀의 위치가 특정한 입력값에 정해지는 방향으로 결괏값을 변경하는 알고리즘이다. 이와 같은 네트워크는 2차원으로 배열되기 때문에, 고차원 자료의 차원을 감소시키는 동시에, 위상적 관계를 보존할 수 있는 시각화가 가능하다. 그러므로 SOM은 고차원 자료를 시각화하기에 적합하며, 특히 키워드를 하나의 벡터로 표현하여 분석하는 문서 데이터의 분석에 장점이 있다[40].

SOM의 네트워크는 입력층과 경쟁층으로 구분되며, 입력되는 자료의 수에 따라 경쟁층의 노드 수가 결정되고, 각 노드에는 임의의 가중치 값이 설정된다. 네트워크의 출력층이 활성화되기 위해 입력 데이터 간의 경쟁을 수행하고, 입력값에 맞는 하나의 출력값만을 활성화하기 위한 경쟁 학습을 기초로 한다. 입력된 데이터는 근삿값을 가진 각 노드로 배정되며, 앞의 과정을 반복하여 가중치의 변동이 나타나지 않으면 학습이 종료되어 군집 결과를 산출하게 된다[41].

SOM의 크기는 일반적으로 quantization error와 topographic error를 고려하여 선정한다. quantization error는 각 데이터 벡터와 가장 유사한 노드(BMU : best matching unit) 사이의 평균 거리를 나타내며 지도의 해상도에 따라 측정되는 오류값을 나타낸다. 그리고 topographic error는 지형적 오차로 첫 번째와 두 번째 BMU가 인접 단위가 아닌 모든 데이터 벡터의 비율을 나타내며 위상에 따라 측정되는 오류값을 의미한다[42]. 또한 샘플의

크기나 결과에 따라 유연하게 조정 가능하며, 분석자의 재량에 따라 지도의 크기를 다르게 분석하여 시각적으로 유효한 정보를 표현하는 매트릭스를 선정할 수 있다.

SOM은 다차원의 데이터를 이차원의 지도에 표현하는 것에 효과적이며, 군집 개수에 따른 검증 방법인 davies-bouldin index(DBI)로 검증하여 군집 개수의 객관적인 산출에 신뢰도가 높다[43]. 또한 유사한 데이터들이 군집화되어 가중치가 가장 큰 키워드를 가진 문서와 그 주변의 문서들의 유사성을 지도로 표현하여 거리에 따라 나타내며, 이를 통해 문서의 관계를 시각적으로 도출한다[44].

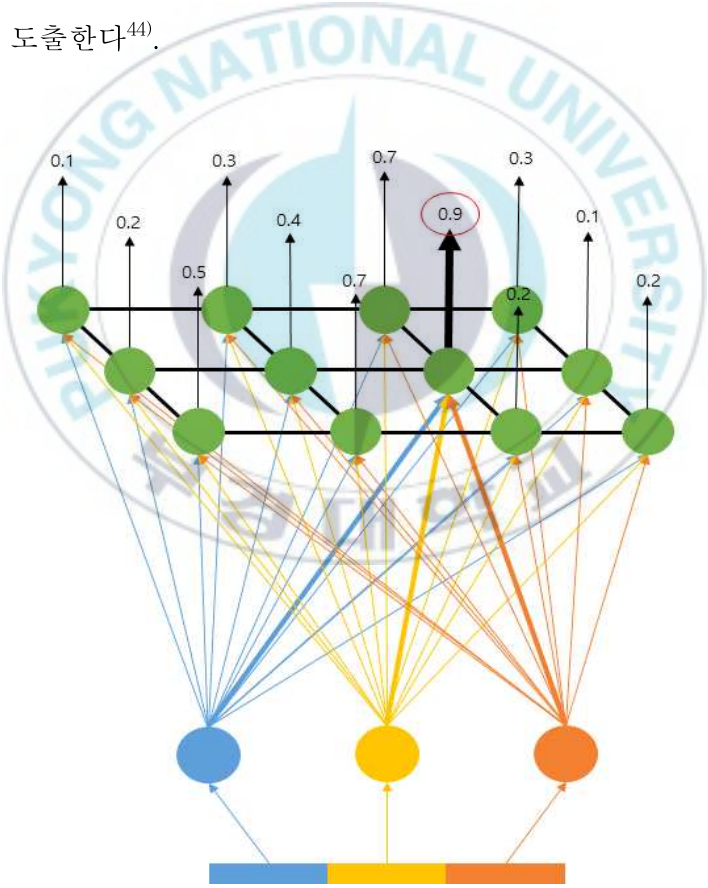

Fig. 10 Conceptual scheme of self-organizing map

다음으로 k-means clustering은 대표적인 비지도학습 기술인 군집화 알고리즘이며, 각 데이터로부터 그 데이터가 속한 클러스터의 중심까지의 평균 거리를 최소화하는 방법론이다. 각 군집은 하나의 중심을 갖고, 각 개체는 가장 가까운 중심에 할당되어 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성한다. 형성되는 군집의 수는 hyperparameter인 k를 분석자가 선정하여 군집 수를 정해 알고리즘을 실행한다. 군집 수는 데이터의 응집도로 선정하며, 각 데이터로부터 자신이 속한 군집의 중심까지의 거리를 의미하는 inertia 값으로 확인하며, 이 값이 낮을수록 군집화가 더 잘되었다고 볼 수 있다. k-means clustering은 먼저, k개의 임의의 중심점을 배치하고 각 데이터를 가장 가까운 중심점으로 할당한다. 다음으로, 군집으로 지정된 데이터들을 기반으로 해당 군집의 중심점을 결괏값이 수렴할 때까지 업데이트하여 군집화한다[45].

2.4.2 지도학습 방법론

지도학습 방법론은 정답을 알려주고, 그에 따라 학습시키는 방법론으로 특정한 label을 갖는 것이 특징이다. 정해진 카테고리에 따라 분류(classification)하거나 데이터들의 특징을 토대로 값을 예측하는 회귀분석(regression)을 수행하기 위해 사용된다.

특히, 텍스트 문서의 분류는 텍스트를 미리 정의된 카테고리 중 어느 하나에 속하도록 분류하는 것을 말하며, 사전에 분류의 기준이 되는 (문서, 카테고리) 형태의 훈련 데이터를 활용해 학습을 시킨다. 훈련 데이터의 학습은 문서를 분류할 수 있는 함수를 구하고, 이를 활용하여 새로운 문서가 어느 카테고리에 속할지를 예측한다. 문서의 분류과정에서 고려해야 할 사항으로 텍스트로 이루어진 문서를 특징 벡터를 갖는 매트릭스 형태로 나타내는 것이다[46].

지도학습 방법론을 활용한 재해문서의 분류에 대하여, 다음 Table 4와 같은 연구들이 진행되었다[47-52]. 국내에서는 김연철 외 2인(2017)이 건설업 산업재해통계 16,248건에 대하여 neural network를 활용하여 분석하였다. 이 연구는 산업재해조사표에서 수집되는 유형에 따라 데이터를 분류하고 예측모델을 생성하고자 하였다. 국외의 연구를 살펴보면, Y. M. Goh와 C.U. Ubeynarayana(2017), M. Cheng 외 2인(2020)의 연구는 OSHA에서 공시하고 있는 건설업 재해 개요를 수집하여 분석하였으며, 건설 분야의 안전관리를 위하여 분류모델을 작성하고 분석하였다. 분류모델은 support vector machine과 decision tree, linear regression, naive bayesian 등 다양한 분류기법을 활용하여 연구를 수행하였다. 이와 같이, 재해문서에 대한 분석이 다양하게 진행되고 있지만, 국내에서는 아직까지 재해 개요를 활용한 텍스트 분석과 지도학습에 관한 연구는 미비하다.

Table 4 Previous research using supervised learning

Author (Year)	Keywords	Analysis target (cases)	Methodology	Objective
Y. Kim, W. S. Yoo, and Y. Shin (2017)	Construction Safety Accident, Artificial Neural Network, Prediction	Construction Industry Accident Statistics (16,248 cases)	Neural network	Predicting possible accidents at construction sites
Y. M. Goh and C.U. Ubeynarayana (2017)	Accident classification, Construction safety, Data mining, Support vector machine, Text mining	Construction accident narratives by OSHA (1,000 cases)	SVM, linear regression, random forest, k-nearest neighbor, decision tree, Naïve Bayes, SVM	Construction safety
B. U. Ayhan and O. B. Tokdemir (2020)	Occupational health and safety in construction, Artificial neural networks (ANN), Latent class clustering analysis(LCCA)	Construction accident statistics (4,109 cases)	Latent class clustering analysis, Artificial neural network	Latent class clustering analysis, Artificial neural network
M. Cheng, D. Kusoemo, and R. A. Gosno (2020)	Construction project safety, Natural language processing, Gated recurrent unit, Symbiotic organisms search, Accidents cause classification	Construction accident narrative historical data, OSHA (16,323 cases)	DT, k-NN, Naïve baysian, Logistic regression, SVM, LSTM, GRU, SGRU	Classification results can be used to aid safety assessments of construction projects.
B. Zhong, X. Pan, P. E. D. Love, J. Sun,and C. Tao (2020)	Deep learning, Construction Hazards, Topic model, Text mining, Safety	Metro hazard report in Wuhan (3,000 cases)	LDA, Convolution Neural Network, Word Co-occurrence Network, Word Cloud	To provide the ability to analyse hazard records cutomatically
P. Sankarasubramanian (2020)	KNN, NLP, SQP, SVM	Accident report	TF-IDF, k-NN, Logical regression, naïve bayes, decision tree, SVM, SQP	Analyzing the historical accident reports identifies what went wrong in the past and this valuable knowledge prevents future accidents

문서분류에 관한 연구는 일반적으로 널리 사용되는 logistic regression과 decision tree, neural network, support vector machine 등이 있다.

Logistic regression은 이진 목적 데이터(binary target data)의 분류를 위한 가장 전통적인 방법이며, 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하기 위해 사용되는 통계기법이다. Logistic regression은 종속 변수와 독립변수 간의 관계를 독립변수의 선형 결합을 이용하여 사건의 발생 가능성에 대한 예측 모델에 사용한다. 선형 회귀분석과의 차이점으로는 Fig. 11과 같이 종속 변수가 범주형 데이터를 대상으로 하며 예측값의 범위가 0과 1사이의 확률값으로 제한되고, 종속 변수가 이진으로 나타나기 때문에 조건부 확률의 분포가 정규분포 대신 이항 분포를 따른다[53].

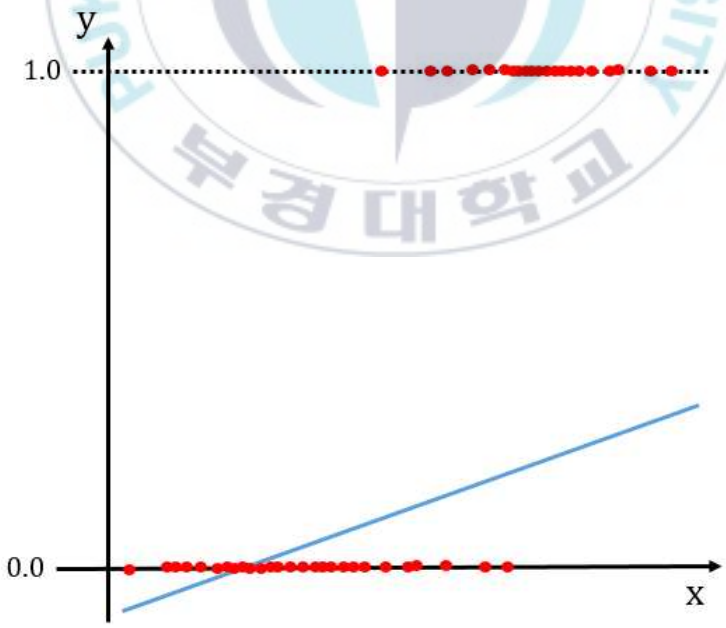

Fig. 11 Conceptual scheme of Logistic regression

Decision tree analysis는 나무구조를 통해 요인 간의 관계를 분화시켜 분류 및 예측을 수행하는 방법론이다. 이 방법론은 Fig. 12와 같이 가장 위에 있는 부모 마디로부터 가지의 조건에 따라서 자식 마디로 분화하며 데이터를 분류한다. 데이터의 분화는 지니 지수와 엔트로피, 카이제곱의 지표를 활용하여 데이터 분류의 순도를 계산하여 분류한다. 첫 번째로, 지니 지수는 사회과학과 경제학에서 사용되는 척도이며, 무작위로 두 개의 데이터를 뽑았을 때 그 둘이 서로 같은 클래스일 확률을 나타낸다. 활용한 데이터의 분화는 지니 지수를 최대화하는 방향으로 진행하며, classification and regression trees(CART) 알고리즘을 활용하여 분류모델을 형성한다. 두 번째로, 엔트로피는 정보이론에서 사용되며 메시지 당 평균 정보량을 측정하는 척도이다. 엔트로피는 불순도를 의미하며, 이를 최소화하는 방향으로 분화가 진행된다. 엔트로피를 활용한 decision tree 분석은 iterative dichotomiser(ID) 3과 C4.5, C5.0 알고리즘을 주로 활용한다. 마지막으로 카이제곱은 실제 관찰된 관찰도수와 이론적 빈도와 기대도수의 차이가 있는지를 검정하는 척도이며, 최초 데이터 수의 비율과 관련이 있다. 데이터의 분화는 카이제곱값을 최대화하는 방향으로 진행하며, chi-square automatic interaction detection(CHAID) 알고리즘을 활용하여 분류모델을 작성한다[54].

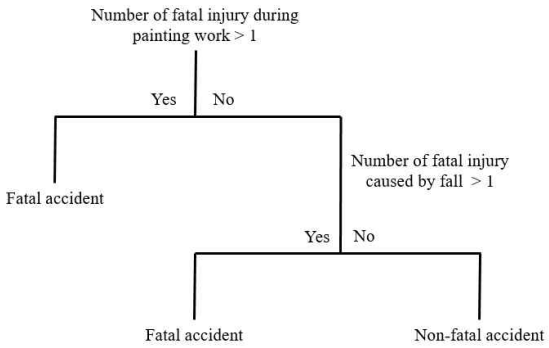

Fig. 12 example for decision tree analysis

Neural network analysis는 인간의 학습체계를 비유하여 생물학적 두뇌의 신경세포(neuron)를 수학적으로 모델링한 인공지능의 한 분야이다. 일반적으로 널리 사용되는 역전파 인공신경망 분석은 Fig. 13과 같이 입력층(input layer)과 은닉층(hidden layer), 출력층(output layer)으로 구성되고 각 층은 노드들로 구성된다[55]. 이와 같은 노드는 가중치와 바이어스 값의 조합으로 구성되며, 전파되는 과정에서 가중치와 바이어스 값을 조절하여 최적화된 노드를 갖게 된다. 인공신경망 분석은 전진파와 역전파로 데이터를 분석하며, 각 뉴런의 가중치를 입력값에 따라 업데이트하면서 최종 결과치에 가장 유사한 값을 도출하도록 학습하는 모델이다. 또한, 계산된 출력값과 실제 값의 차이를 계산하여 에러를 최소화하도록 역방향으로 가중치를 수정하여 모델의 정확도를 높인다[56].

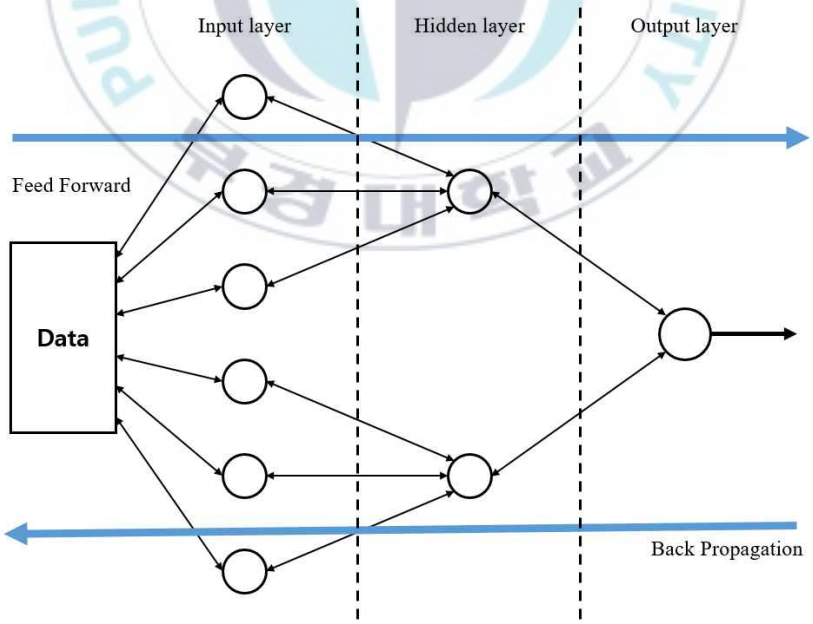

Fig. 13 Conceptual scheme of neural network analysis

Support vector machine(SVM)은 기계학습 중 하나로 패턴인식이나 자료 분석을 위해 사용되는 지도학습 모델이며, 기본적으로 두 범주를 갖는 데이터들을 분류하는 방법론이다. 기존의 분류 방법론 중 대부분은 경험적 위험(empirical risk)을 최소화하는 방식으로 기계학습이 수행되는 반면에, SVM은 구조적 위험 최소화(structural risk minimization)를 기반으로 하여 일반화 오류의 상한을 최소화할 수 있는 기계학습 방법론이다[57]. 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 카테고리에 속할지 판단하기 위해 사용되며, 비확률적 이진 선형 분류모델을 작성하여 데이터를 분류할 수 있는 가장 큰 폭을 가진 경계, 즉 최적의 회귀식을 찾기 위한 방법론이다. SVM은 Fig. 14와 같이 데이터가 분포되어있는 공간을 나누는 초평면(hyperplane)을 계산하여 데이터를 분류하며, 초평면과 최소의 거리를 갖는 선형조합으로 만들어지는 벡터인 support vector를 찾는다.

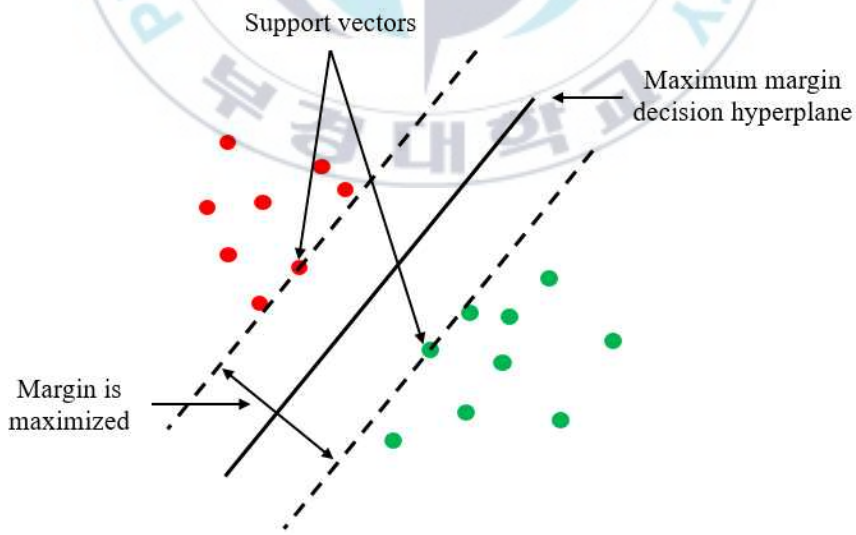

Fig. 14 Support vector of linear classifier

또한 SVM은 선형 분류 뿐만 아니라, 커널 함수(kernel function)를 활용하여 입력데이터의 고차원 공간상으로의 맵핑하여 비선형분류도 효율적으로 수행한다. 대표적인 커널 함수는 선형(linear) 커널과 다항(polynomial) 커널, 가우시안(gaussian) 커널 등이 있다. 선형커널은 문서분류 등에서 자주 발생하는 대량의 자료를 분석할 때 유용하며, 다항 커널은 이미지 처리에 자주 사용되며, 가우시안 커널의 경우 자료에 관한 사전정보가 없을 때 사용되는 일반적인 목적의 커널이다[58].

SVM은 범주 혹은 수치 예측 문제에 대해 사용할 수 있으며, 노이즈 데이터에 영향을 크게 받지 않고 과적합화가 드물게 일어나는 장점이 있다. 그러나 최적의 모델을 찾기 위해 커널과 모델에서 매개변수의 조합에 대한 테스트가 필요하며, 입력 데이터셋의 크기가 커질수록 다른 방법론들보다 훈련 속도가 크게 느려지는 단점이 있다[59].

분류모델을 평가하는 방법은 training set으로 분류모델을 작성하고 test data set으로 분류모델의 정확도를 평가한다. 분류방법론의 정확도는 통계학에 기반한 가설검정인 Table 5와 같은 confusion matrix를 통해 정확도 지표들을 계산하여 평가한다[60]. 모델을 평가하는 요소는 모델의 결괏값과 실제값의 관계로써 정의할 수 있다. 결괏값은 옳게 예측한 경우인 true와 예측이 틀렸을 경우인 false로 나누어져 있으며, 사망재해를 positive로, 부상재해를 negative로 이루어진 2*2 매트릭스로 표현할 수 있다. 또한 실제 발생한 부상사고를 귀무가설로 설정하여 분석을 수행하였다. Confusion matrix를 살펴보면 true positive(TP)는 실제 positive인 정답을 positive라고 예측하여 옳게 예측한 경우이다. 또한 false positive(FP)는 실제 negative인 정답을 positive라고 예측하여 오답으로 분석된다. 다음으로 false negative(FN)는 실제 positive인 정답을 negative라고 예측하여 오답으로 나타난 경우이다. 마지막으로 true negative(TN)는 실제 negative인 정답을 negative라고 예측하여 정답으로 나타난 경우이다. 이와 같은 confusion matrix를 통해 정확도 지표를 활용하여 오류를 정량화하여 평가한다[61]. 이는 통계의 1종 오류(Type Ⅰ error)와 2종 오류(Type Ⅱ error)와도 연결되는데, 가설은 일반적으로 다수 표본인 negative로 인식하여, false positive가 1종 오류, false negative가 2종 오류를 설명하게 된다.

Table 5 Confusion matrix

	Predicted : Positive	Predicted : Negative
Actual : Positive	True Positive	False Negative Type Ⅱ(Beta) Error
Actual : Negative (Null Hypothesis)	False Positive Type Ⅰ(Alpha) Error	True Negative

재해문서의 분류모델의 성능은 정밀도(precision)와 재현율(recall), 정확도(accuracy)를 정확도 지표를 활용하여 분류모델의 정확도를 평가할 수 있다[62]. 한 개의 재해문서가 사망재해와 부상재해에 속할 수 있는 레이블에 대해 독립적으로 확률을 예측한다. 정밀도(precision)란 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 자료로, 모델이 positive라고 분류한 것 중에서 실제 positive인 것의 비율로 수식 (1)과 같다. 재현율(recall)이란 실제 positive인 것 중에서 모델이 positive라고 예측한 것의 비율을 나타내며, 아래와 같은 수식 (2)로 계산한다. 정확도(accuracy)란 전체 데이터에서 올바르게 예측한 데이터의 비율을 의미하며, 수식 (3)으로 표현한다.

$$(\mathrm{Pr}ecision) = \frac{TP}{TP+FP} \quad \cdots (1)$$

$$(Recall) = \frac{TP}{TP+FN} \quad \cdots (2)$$

$$(Accuracy) = \frac{TN+TP}{TN+FP+FN+TP} \quad \cdots (3)$$

제 3 장 건설업 재해문서의 비지도학습

비지도학습은 레이블이 없는 데이터를 데이터의 특징만을 가지고 비슷한 데이터들끼리 군집화하는 기계학습 방법론이며, 데이터의 관계를 알지 못하는 경우에 유용하게 활용된다. 특히, 텍스트데이터를 활용한 비지도학습은 문서 내에서 키워드를 도출하여, 유형화와 연관성 파악 등 다양한 분석에서 사용되고 있다.

본 장에서는 비지도학습 중 데이터의 군집화 및 시각화를 위해 대표적으로 사용되는 self-organizing map(SOM)을 활용하여 사망재해문서로부터 주요 텍스트 키워드를 도출하고, 키워드 간의 관계를 분석 및 시각화하는 위험요인지도를 개발하고자 한다. 구체적으로, 텍스트마이닝을 통해 문서형태의 재해 발생 개요에서 정형화된 데이터로 변환하고, 문서에 포함된 주요 재해 원인 키워드를 도출한다. 다음으로 신경망에 기반한 데이터 시각화 방법인 SOM을 이용하여, 주요 키워드에 가중치를 부여하고 유사한 문서들을 군집화하는 과정을 통해 효율적인 위험요인지도로 표현(data representation or visualization)한다. 이처럼 지도 형태로 시각화된 위험요인지도는 각 재해문서의 주요 키워드들의 유사도에 따라 구성되기 때문에, 재해요인 간의 관계를 파악하기 위해 유용하게 사용된다. 본 연구는 사례분석으로, 국내 업종 중 가장 많은 사망재해가 나타난 건설업을 대상으로 위험요인지도를 개발하여 주요 재해 요인과 그 요인들의 관계를 파악하고자 한다. 이와 같이, 텍스트마이닝과 SOM을 이용하여 개발된 위험요인지도는 안전관리자와 관리감독자에게 재해의 주요 요인 관계를 시각적으로 분석·표현하고 이해할 수 있는 효과적인 도구를 제공할 수 있다.

3.1 위험요인 연관성 분석을 위한 SOM 분석 방법

재해문서에서 위험요인을 도출하여 연관성을 분석하기 위하여 비지도학습 중 대표적으로 사용되는 SOM을 활용하여 연구를 수행하였다. 연구 절차는 Fig. 15와 같이 재해보고문서에서 텍스트 정보를 추출하여 위험요인 지도를 개발하는 방향으로 진행된다. 먼저, 산업재해통계자료에서 재해발생 개요를 추출하고 분석에 적합하도록 목록화하였다. 텍스트 분석을 통해 재해문서가 가지고 있는 키워드 형태의 텍스트데이터를 추출하였으며, 분석에 필요하지 않은 데이터를 제거하는 전처리 과정을 수행하였다.

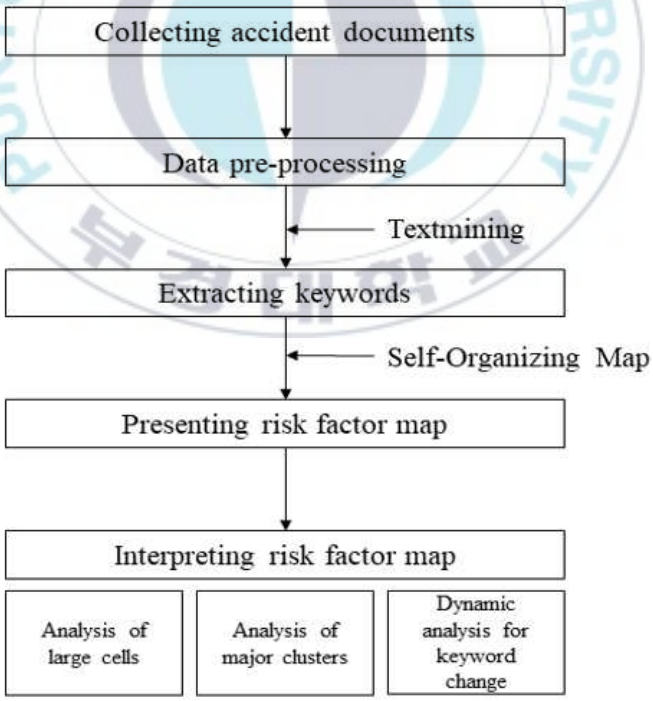

Fig. 15 Flowchart for the methodology of unsupervised learning

다음으로, 기계학습이 가능하도록 텍스트데이터를 문서-키워드 매트릭스의 형태로 구조화하였다. 구조화된 데이터를 SOM 방법론을 통해 재해문서에 포함되어있는 위험요인의 특징에 따라 군집화하여 재해문서가 군집된 셀을 생성하였으며, 지도 형태의 이미지로 시각화하여 위험요인지도를 작성하였다. 마지막으로 작성된 위험요인지도에서 셀 간의 유사도에 따라 셀을 추가적으로 군집화하여, 군집화된 문서군에 대한 키워드 분석을 수행하였다. 키워드 분석은 문서가 많이 군집된 셀에 대한 분석과, 군집된 문서군에 대한 분석, 문서군에 대한 연도별 분석을 수행하였다.

3.1.1 재해문서의 텍스트 데이터 전처리

본 연구를 위해 2009년 1월부터 2013년 12월까지 5년 동안의 사망재해 데이터를 안전관리를 담당하는 공공기관 및 협회를 통해 수집하였으며, 사망재해가 가장 많이 나타나는 업종인 건설업에 대하여 Fig. 16과 같이 2,448개의 사망재해문서를 수집하였다. 수집한 재해문서는 재해 일자와 발생형태, 업종, 연령, 재해 발생 개요 등이 포함되어있다. 본 연구에서는 재해 간의 유사성을 도출하기 위해 재해 상황이 기술되어있는 재해 발생 개요를 중심으로 연구를 수행하였다.

텍스트데이터의 전처리를 위하여 먼저, 재해문서를 텍스트 분석을 통해 키워드 벡터로 변환하는 구조화 단계를 수행하였다. 또한, 본격적인 분석에 앞서 재해문서를 분석 가능한 형태로 만드는 전처리 과정을 수행하였다. 수집된 재해문서는 통계적 분석이 가능한 형태로 변환되어야 하며, 문장을 단어의 형태로 분해해야 한다.

Fig. 16 Number of accident documents

키워드 분석은 전체 재해문서에 대한 키워드가 도출하기 위하여 전체 재해문서를 하나의 파일로 합치는 작업이 선행되어야 한다. 합쳐진 파일은 각 문서의 번호(label)를 다음과 같이 표기하여 작성하였다.

"11202" -> 11 : Years(2011), 202: Document number

모든 문서에 대하여 위와 같이 레이블링 후, 텍스트마이닝을 위한 프로그램 R의 tm package와 한글 자연어처리를 위한 KoNLP package로 텍스트 분석을 수행하였다. 명사와 동사 위주의 분석에 용이한 글자 수 2~4개의 명사 키워드를 추출하였으며, 분석의 효율을 높이기 위하여 '은/는', '와/과', '이'와 같은 약한 의미를 갖거나 분석에 필요하지 않은 불용어와 숫자, 기호 등을 제거하는 cleaning 과정을 수행하였다. 다음으로, 전체적인 재해문서의 키워드 동향을 살펴보기 위하여 Fig. 17과 같이 워드 클라우드로 나타내었다. 워드 클라우드는 프로그램 R의 wordcloud package를 사용하여 작성하였으며, 전체 문서에서 도출 빈도가 높은 키워드일수록 글자의 크기가 크게 표현되고, 중앙에 배치하도록 설정하였다. 또한, 너무 많은 단어의 표현을 방지하기 위하여, 출력된 단어들의 최소빈도는 5로 설정하여 워드 클라우드를 작성하였다. 그러나 워드 클라우드에서 나타나듯이, 알고리즘의 불확실성으로 인해 데이터 분석용으로 활용하기 어려운 불완전한 단어와 분석에 부적절한 단어, 조사, 기호 등과 같은 불용어(stopword)가 포함되어있다. 따라서 이를 제거하는 데이터 전처리를 추가적으로 수행하고, 정성적으로 도출된 키워드를 직접 살펴보면서 스크리닝하는 과정을 거쳤다. 다음으로 전처리된 비정형데이터에서 전체 재해문서를 대상으로 키워드 빈도 행렬을 도출하였다. 키워드 빈도 행렬은 각 행을 재해문서로, 각 열을 키워드로 하여, 해당 재해문서에서 키워드가 몇 번 포함되어있는지를

계산한다. 이를 통해 재해문서와 도출 키워드로 이루어진 정형데이터로서 DTM을 추출하였다. 이 DTM을 이용하여 위험요인지도를 구축하게 된다. 이를 통해 재해문서의 유사도를 분석하고, 유사재해를 분류 및 시각화하는 과정을 통해 위험요인지도를 개발하였다.

Fig. 17 Word cloud of risk factor keywords

3.1.2 SOM 분석 방법

비지도학습 방법론 중에서 SOM 분석은 다변수 데이터를 군집화하고 시각화하는 것에 강점을 지니고 있다. SOM에 활용되는 입력데이터는 재해문서에서 도출한 키워드 빈도 행렬이며, 각 열은 도출된 키워드를 의미하기 때문에, 각 재해문서의 키워드를 비교 분석한다. SOM 분석은 상용 프로그램인 MATLAB을 사용하였으며, 핀란드의 helsinki university of technology의 laboratory of computer and information science에서 배포하고 있는 Kohonen SOM toolbox 2.0을 활용하여 부록 B에 작성한 같은 코드로 구현하였다. SOM toolbox는 SOM을 수행하고 결과를 시각화하는 등 다양한 기능을 제공하고 있으며, MATLAB에서 구조화된 배열로 정리된다. SOM을 통해, 각 재해문서가 지도에서 특정한 노드에 위치되면, 가까운 노드에는 문서의 내용이 유사한 특허들이 위치하게 된다.

또한, SOM은 셀의 개수가 적을 경우 군집화가 과도하게 발생하며, 너무 많을 경우 문서를 포함하지 않은 셀이 발생하는 문제점이 있다. 이를 위해 위험요인지도에 대한 quantization error와 topographic error를 계산하여 가장 낮은 수치의 오류값을 나타낸 위험요인지도를 도출한다.

그러나 작성된 지도에서 데이터의 배열은 각 셀에서 문서에 대한 구조를 이해하는 것에 도움이 되지만, 지도에서 나타난 셀에 대한 클러스터를 구분하는 것은 분석자의 판단으로 결정해야 한다. 클러스터를 구분하기 위하여, 각 셀 간의 거리에 대한 정보를 바탕으로 클러스터의 개수와 영역을 결정할 수 있다. SOM에서 셀 간의 거리는 색으로 정의하며, 파란색일수록 각 셀 간의 유사도가 높으며, 노란색일수록 유사도가 낮음을 나타낸다.

3.2 재해문서의 군집화 결과

3.2.1 텍스트 분석 결과

각 문서의 비정형데이터인 재해 발생 개요를 분석하기 위하여 전체 문서에 대하여 텍스트마이닝을 수행하여 문서와 키워드로 구성된 매트릭스로 구조화하였다. DTM은 명사와 동사 위주로 10개 이상의 키워드를 가진 문서를 채택하였으며, 이에 따라 781개의 위험요인(키워드)과 2,173개의 재해문서로 Fig. 18과 같은 매트릭스로 도출하였다. 매트릭스의 행은 레이블링한 문서의 번호를 나타내며, 열은 전체 재해문서에서 많이 나타난 키워드 순으로 위험요인을 나열하였다. 또한, 숫자로 각 문서에 포함된 키워드의 빈도를 표시하였다.

Term / Documents	추락	바닥	설치	신축공사	시공	…
9001	1	0	3	0	0	
9002	0	0	0	0	0	
9003	1	1	0	1	0	
9004	1	0	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	
13516	0	0	1	1	0	

Fig. 18 Result of document-term matrix

3.2.2 SOM 분석 결과

SOM의 본 연구에서는 분석에 적절한 셀의 개수를 도출하기 위해 5×5부터 행과 열의 크기를 하나씩 늘리면서 10×10까지 각각의 위험요인지도의 오류값을 계산하여 quantization error가 2.229, topographic error가 0.031로 가장 낮은 수치의 오류값을 나타낸 10×8의 위험요인지도를 Fig. 19와 같이 도출하였다. 표시된 문서번호는 각 셀에서 가중치가 가장 큰 키워드를 가진 문서를 나타내는 대표(vote)문서를 나타낸다. 셀 간의 거리는 가장 먼 셀이 0.965로, 가장 가까운 셀이 0.207로 나타났으며, 그 중앙값은 0.586으로 도출되었다.

먼저, 전체 80개 셀 중 79개의 셀에 문서가 군집되었으며, 각 셀에 포함된 문서는 유사한 키워드를 가진 문서로 군집됨을 알 수 있다. 다음으로, 각 셀 간의 관계를 살펴보면 가장 좌측 상단에 위치한 대표문서가 11202인 셀의 경우, 바로 우측에 인접해있는 대표문서가 10482인 셀보다 대표문서가 9384인 셀과의 거리가 더 가까움을 알 수 있으며, 이는 셀 간의 연관성이 더 크다는 것을 의미한다. 또한, 대표문서가 10020인 셀과 대표문서가 13281인 셀의 경우, 거리가 0.207로 가장 연관성이 큰 셀을 나타내며, 9316을 포함하고 있는 문서와 11471을 포함하고 있는 문서는 셀 간의 거리가 0.965로 가장 연관성이 없는 셀로 분석된다.

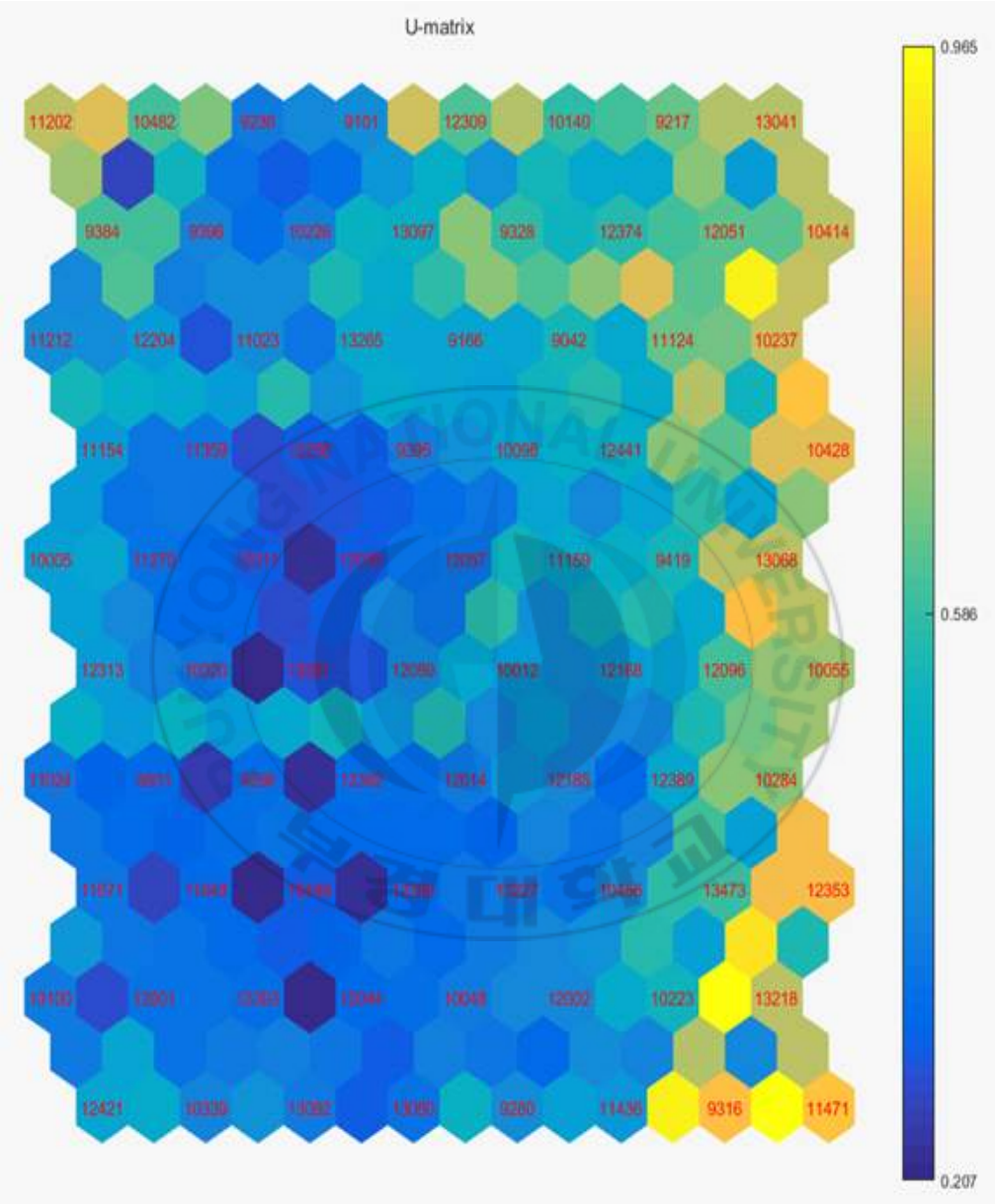

Fig. 19 SOM-based risk factor map

3.2.3 위험요인지도의 해석

작성된 위험요인지도의 해석을 위하여, 셀 단위로 군집된 재해문서를 살펴보고, 셀 간의 거리에 따라서 셀을 군집화하여 키워드 분석을 수행하였다. 키워드 분석은 셀에 포함된 문서번호에 해당되는 문서를 DTM을 통해 키워드 빈도를 도출하여 분석하였다. 하나의 셀에 군집된 재해문서는 위험요인에 대한 특징이 유사한 문서들로 나타나며, 인접한 셀에 대하여 추가적으로 군집화하여 유사재해에 대한 위험요인을 넓은 범위로 탐색할 수 있다.

먼저, 위험요인지도에서 재해문서가 가장 많이 군집된 셀에 대하여 다음 Table 6과 같이 분석하였다. 군집화한 문서 클러스터에서 나타나는 키워드 빈도에 따라서 도출된 상위 10개의 키워드를 추출하였다. 그리고 도출된 키워드를 재해의 발생 과정을 나타낼 수 있는 재해 원인 키워드와 작업 키워드, 재해 형태 키워드의 유형으로 분류하였다. 추가적으로, 한글로 작성된 재해문서를 분석하였기 때문에, 해당 키워드는 부록 A의 한-영 키워드 변환표에 따라 작성하였다.

가장 많은 재해문서를 포함하고 있는 주요 셀은 대푯값이 문서번호 12421인 셀로 90개의 문서가 군집화하였으며, 도출 키워드로는 굴착기(125회), 굴착(46회), 토사(26회), 운전(16회) 순으로 나타났다. 재해 원인 키워드로는 굴착기, 토사, 바다, 바퀴, 바지선이 나타났으며, 굴착작업과 운전이 작업분류 키워드로, 매몰, 붕괴, 전복이 재해형태 키워드로 도출되었다. 이와 같은 위험요인 키워드를 토대로, 해당 셀은 굴착작업 시 토사로 인해 발생한 재해 혹은 굴착기 운전 시 발생한 재해로 군집되어있는 것을 알 수 있다. 두 번째로 많은 재해문서를 포함하고 있는 셀은 대푯값이 문서번호 11048인 셀로 86개의 문서가 군집화하였으며, 도출 키워드로는 감전(9회),

전기(8회), 화재(8회), 외벽(7회)으로 나타났으며, 재해 원인 키워드로는 전기, 계단, 지하, 탱크가 나타났다. 또한, 작업분류 키워드로는 외벽과 교체, 철거, 도장이 나타났으며, 재해 형태 키워드는 감전과 화재로 나타났다. 이 셀의 경우, 전기로 인한 감전 및 화재가 발생한 재해문서를 포함하고 있으며, 전체적으로 키워드의 빈도가 낮은 문서들이 군집되었다. 세 번째로 많은 재해문서를 포함하고 있는 셀은 대푯값이 문서번호 11202인 셀로 81개의 문서가 도출되었으며 발생키워드로는 지붕(185회), 추락(70회), 바닥(49회), 설치(30회) 등으로 나타났다. 재해 원인 키워드로는 지붕, 바닥, 철근, 콘크리트, 판넬이 나타났으며, 작업분류 키워드로는 설치, 수리가 나타났다. 주요 재해 형태 키워드로는 추락이 도출되었다. 해당 셀은 고소작업 시 발생한 추락과 연관된 재해문서가 군집되어있으며, 설치 및 수리 작업도 포함되어있다.

Table 6 Analysis of large sells

Vote accident number	Cases of accident	Keywords (count)	Main cause of accident	Main work type	Main accident form
12421	89	excavator(125), excavation(46), soil(26), drive(16), bury(15), collapse(14), sea(13), rollover(13), tire(12), barge(12)	excavator, soil, sea, tire, barge	excavation, drive	bury, collapse, rollover
11048	86	electric(9), electricity(8), fire(8), external(7), stair(7), basement(7), tank(6), change(5), removal(4), paint(4)	electricity, stair, basement, tank,	external, change, removal, paint	electric, fire
11202	81	roof(185), fall(70), floor(49), install(30), steel(21), upper(17), balance(17), concrete(16), repair(15), panel(14)	roof, floor, steel, concrete, panel	install, repair	fall

다음으로, 셀 군집(cluster) 분석은 앞서 분석한 주요 셀을 포함하여 유사한 키워드를 가진 셀을 군집화한 전체 지도 분석과정이다. 앞서 언급한 바와 같이, SOM에서는 클러스터를 명확하게 구분하기 힘들며, 분석자가 직접 클러스터를 구분해야 한다, 본 연구에서는 셀 간의 거리가 중앙값인 0.586을 기준으로 Fig. 20과 같이 5개의 군집으로 분류하였으며, 각 군집별로 Table 7과 같이 11개, 20개, 27개, 14개, 7개의 셀이 군집되었으며, 각각 416개, 355개, 778개, 280개, 344개의 재해문서가 포함되었다.

각 군집을 살펴보면, 군집 1은 11개의 셀과 416개의 재해문서를 포함하고 있으며, 셀 개수에 비해 많은 재해문서가 군집되었다. 군집 2는 20개의 셀과 355개의 재해문서로 군집되었으며, 재해문서보다 셀의 개수가 더 많이 포함되어 다양한 위험요인이 포함되어있는 것을 알 수 있다. 군집 3은 가장 많은 셀과 재해문서를 포함하고 있으며, 셀은 27개, 재해문서는 778개로 군집되었다. 또한, 각 셀 간의 거리가 가장 가깝게 군집된 클러스터로 도출되었다. 군집 4는 14개의 셀과 280개의 문서로 구성되어 있으며, 셀 간의 거리가 멀어 유사도가 낮은 클러스터로 나타났다. 군집 5의 경우, 가장 적은 셀과 문서를 포함하고 있으며, 각 셀 간의 유사도가 낮은 셀들로 구성되어 있다.

Table 7 Number of cells and accident cases each cluster

Cluster number	Number of cells (rate)	Number of accident cases (rate)
1	11 (13.92%)	416 (19.14%)
2	20 (25.32%)	355 (16.33%)
3	27 (34.18%)	778 (35.80%)
4	14 (17.72%)	280 (12.89%)
5	7 (8.86%)	344 (15.83%)

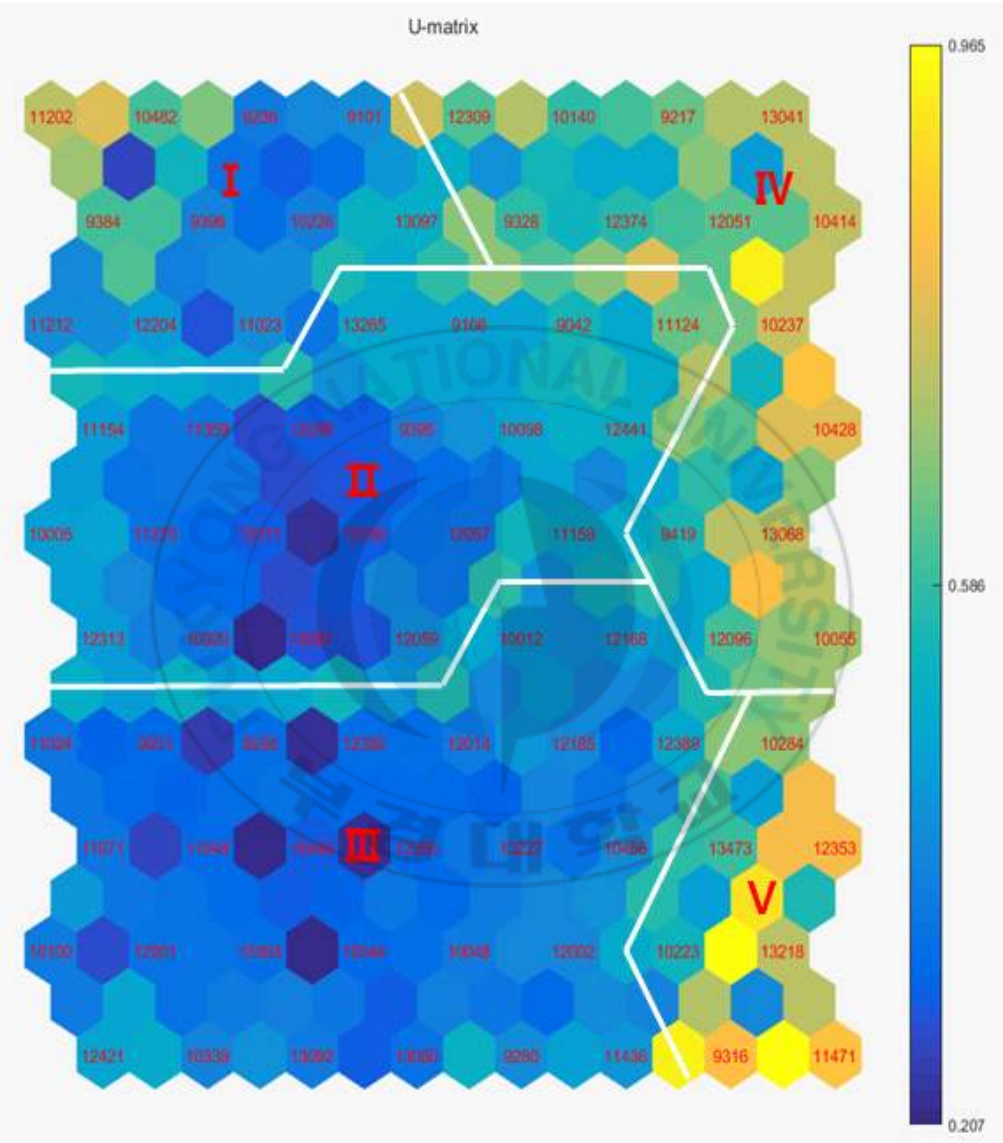

Fig. 20 Major clusters of accident documents

결과적으로 Table 8과 같이 각 군집의 특성을 정리하였다. 군집 3을 제외하고 나머지 4개의 군집은 추락으로 인해 발생한 재해인 것으로 나타났으며, 군집 3의 경우는 붕괴로 인해 발생한 재해로 군집화되었다. 각 클러스터를 살펴보면 다음과 같다.

먼저, 군집 1의 경우, 재해 수는 416건으로 나타났으며 바닥과 추락, 지붕, 콘크리트 등의 키워드가 도출되었고, 콘크리트나 철골로의 추락과 지붕에서의 고소작업과 설치작업, 작업 시 이동의 작업유형이 나타났다. 이를 통해, 군집 1에 포함되어있는 재해문서는 고소작업으로 인한 추락재해와 같은 재해들이 분포되어있음을 알 수 있다. 다음으로, 군집 2는 355건의 재해문서에서 추락, 사다리, 로프, 상부, 아파트 등의 키워드가 도출되었으며, 재해 원인 키워드는 사다리, 로프, 콘크리트로 나타났고, 작업 키워드는 외벽작업과 배관작업이, 재해 형태 키워드는 추락으로 나타났다. 이와 같은 위험요인들은 군집 2에 포함되어있는 재해문서들이 외벽작업 및 배관작업과 같은 높은 장소에서 이동 및 작업 시 발생할 수 있는 추락재해에 대한 문서들이 군집화되었다. 군집 3은 778건(35.80%)으로 가장 많은 재해문서가 할당되었으며, 굴착기와 차량, 머리, 굴착, 콘크리트 등의 키워드가 도출되었다. 그 중, 재해 원인 키워드는 굴착기와 차량, 토사로 나타났으며, 작업 키워드는 굴착작업, 운전이, 작업형태로는 붕괴와 매몰이 도출되었다. 이와 같은 위험요인 키워드를 통해 굴착작업 혹은 굴착기 운전 시 발생한 재해가 도출되는 것을 알 수 있다. 또한, 다른 군집들과는 다른 재해 형태인 붕괴로 발생한 재해를 나타내고 있다. 군집 4는 280건의 재해문서가 군집되었으며, 추락, 설치, 비계, 발판 등의 키워드가 도출되었고, 재해 원인 키워드는 비계와 발판, 작업 키워드는 설치작업, 해체작업, 이동이 나타났으며, 재해 형태 키워드는 추락이 도출되었다. 이를 토대로 비계 및 발판과 같은 가시설물의 설치·해체 작업에서 발생한 재해가 군집되어 있는 것을

알 수 있다. 마지막으로 군집 5는 344건의 재해문서에서 크레인, 설치, 추락, 거푸집, 작업대 등의 키워드가 나타났다. 재해 원인 키워드는 크레인, 거푸집, 작업대가, 작업 키워드는 설치·해체작업으로, 재해 형태 키워드는 추락으로 나타났다. 이 군집에서는 최근 중대 재해로 자주 발생한 크레인과 관련한 문서가 많이 포함되었으며, 이와 관련한 주요 요인들의 관계를 해석할 수 있다.

가장 많은 재해 형태인 추락이 위험요인지도에서 전체적으로 분포되어있으며, 군집 3의 경우에만 무너짐으로 인한 재해가 나타나는 차이점이 있었다. 또한, 군집 1과 2의 경우 유사한 재해 원인 키워드를 가지고 있으나 작업 키워드에서 설치작업과 이동작업, 외벽작업과 파이프 작업의 차이를 나타냈으며, 군집 4와 5의 경우 유사한 작업 키워드가 도출되었지만, 재해 원인 키워드에서 군집 4에서는 비계, 발판과 같은 가시설물과 군집 5에서 크레인, 거푸집, 작업대로 나타났다.

Table 8 Analysis of major clusters

Cluster number	Cases of accident (rate)	Keyword (count)	Main cause of accident	Main work type	Main accident form
1	416 (19.14%)	floor(415), fall(386), roof(251), concrete(117), steel(93), balance(82), install(63), rooftop(59), upper(54), move(40)	concrete, steel, roof	install, move	fall
2	355 (16.33%)	fall(352), ladder(116), rope(41), upper(35), apartment(33), balance(32), rooftop(32), external(30), piping(29), concrete(28)	ladder, rope, concrete	external, piping	fall
3	778 (35.80%)	excavator(163), vehicle(145), head(87), excavation(83), concrete(74), drive(74), soil(70), collapse(65), bury(65), reverse(60)	excavator, vehicle, soil	excavation, drive	collapse, bury
4	280 (12.89%)	fall(266), install(227), scaffold(205), foothold(193), floor(142), outside(53), move(65), balance(60), upper(53), disjoint(53)	scaffold, foothold	install, disjoint, move	fall
5	344 (15.83%)	crane(250), install(184), fall(147), mold(129), workbench(126), floor(92), upper(89), move(71), connect(67), disjoint(65)	crane, mold, workbench	install, disjoint	fall

3.2.4 도출된 군집의 연도별 분석

군집의 연도별 분석은 고소작업과 가시설물 설치·해체작업, 크레인과 굴착기와 같은 기계·기구를 사용하는 작업에 따라서 재해문서를 구분하고 도출되는 위험요인의 연도별 재해의 연관성을 파악할 수 있다. 전체 분석에서 군집화된 셀 군집에 대하여 재해 건수와 상위 도출 키워드, 재해 원인 키워드, 작업 키워드, 재해 형태 키워드로 2009년부터 2013년까지의 연도별 분석을 수행하였다.

먼저, 군집별 연도별 추이를 살펴보면 Fig. 21과 같이 나타났다. 군집 1의 경우, 2009년 82건의 재해에서, 2010년 85건, 2011년 100건, 2012년 56건, 2013년 93건의 재해문서가 나타났으며, 2011년까지 증가하는 추세를 보이다 2012년에 급격히 감소하는 것으로 나타났다. 군집 2의 경우, 2009년에 73건의 재해문서가 도출되었으며, 2010년 73건, 2011년 77건, 2012년 72건, 2013년 60건의 재해문서가 나타났으며, 2011년에 가장 많은 재해가 발생하였고 점차 감소하고 있다. 군집 3의 경우, 가장 많은 재해문서를 포함하고 있는 클러스터로 2009년에 155건, 2010년 152건, 2011년 146건, 2012년 152건, 2013년 173건으로 나타났으며, 2013년에 가장 많은 재해문서가 도출되었다. 군집 4의 경우, 2009년 55건, 2010년 56건, 2011년 59건, 2012년 55건, 2013년 49건의 재해문서가 도출되었으며, 2010년에 가시설물의 설치 및 이동작업에 대한 재해문서가 급격히 증가하였다. 군집 5의 경우, 재해문서가 2009년 69건, 2010년 62건, 2011년 60건, 2012년 76건, 2013년 77건으로 2011년까지 감소하는 추세를 보이다 점차 증가하는 것으로 나타났다.

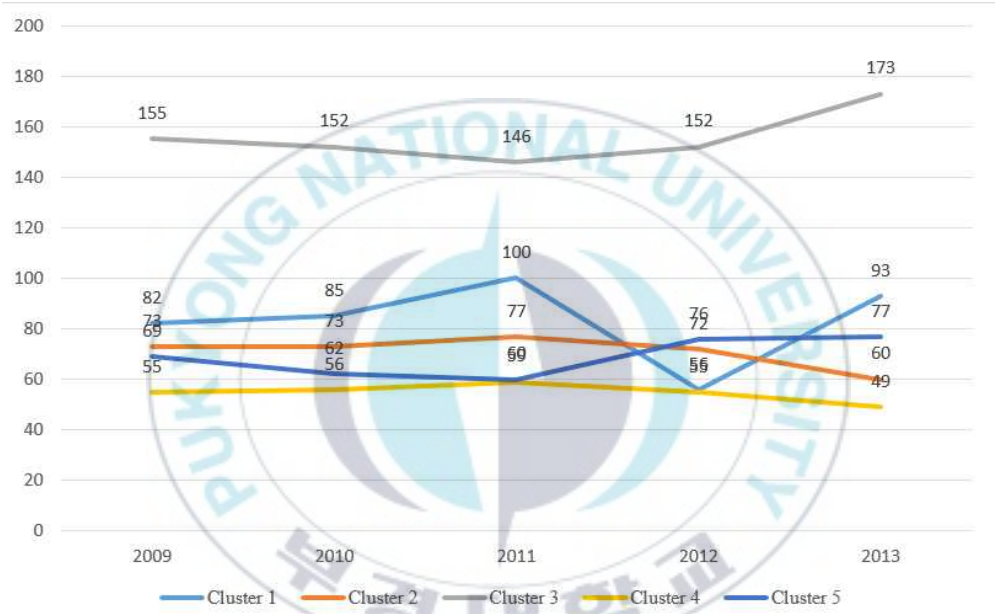

Fig. 21 Trends in the number of disasters in the cluster by year

군집의 연도별 키워드 분석을 정리하면, Fig. 22와 같이 상위 도출 키워드 5개에 대하여 연도별로 차이점을 제시할 수 있다. 먼저 군집 1의 경우, 콘크리트와 철근과 같은 재료와 관련된 재해가 군집되어 있으며, 2012년에 키워드 바닥과 추락이, 지붕과 콘크리트의 빈도순서가 변하였으며, 2013년에 키워드 철근과 상부가 도출되었다. 군집 2의 경우, 고소작업 중 외벽에서의 작업이나 높은 장소로 이동 중에 발생한 재해의 군집으로 위험요인은 추락과 사다리는 계속해서 나타났으며, 2009년에는 아파트와 레버, 옥상이, 2013년에는 상부와 외벽작업이 도출되어 다양한 재해 키워드가 나타났다. 군집 3의 경우, 굴착과 붕괴가 연관된 재해들이 군집되었으며 차량, 머리, 굴착기, 굴착, 토사의 순서였던 키워드가 도출되었다. 굴착기의 경우 시간이 지남에 따라 점차 증가하는 추세를 나타내고 있으며, 2010년, 2011년에 감소하였던 차량이 2012년, 2013년에 증가하였다. 군집 4는 비계와 발판과 같은 가시설물에 대한 재해문서가 군집되어, 설치작업과 추락, 비계, 발판이 계속해서 도출되었으며, 가시설물의 설치작업의 재해에 대하여 군집하였다. 또한, 군집 5의 경우, 크레인과 연관성이 큰 재해들이 군집되었으며, 크레인과 설치, 추락, 거푸집이 계속해서 도출되었으며 2010년에 키워드 작업대가 도출되어 2012년까지 감소하는 것을 볼 수 있으며 추락은 감소 추세를 보이다가 다시 증가하는 것을 볼 수 있다.

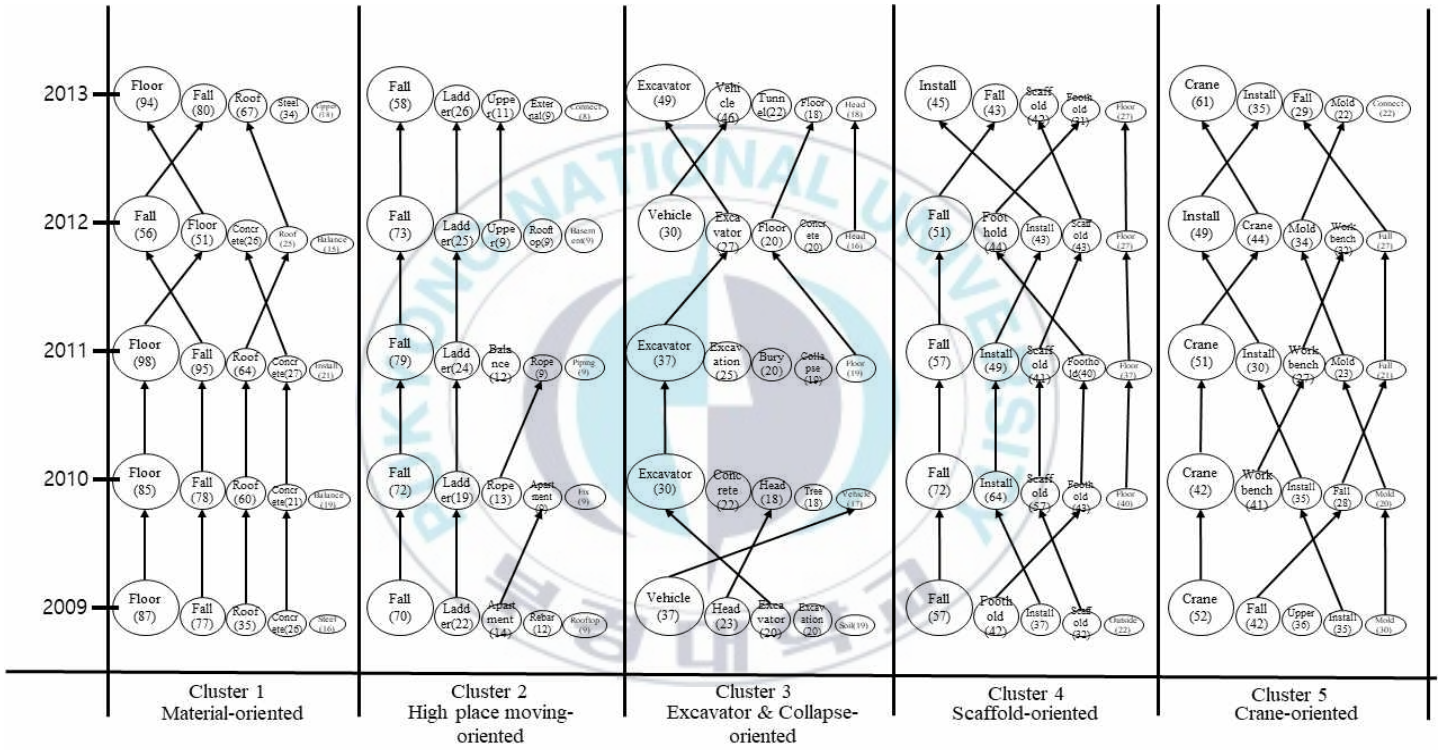

Fig. 22 Dynamic analysis for keywords change in each cluster

3.3 소결

재해문서에서 위험요인을 도출하고자 할 때, 연관된 재해문서를 군집화하거나 패키지화하려는 목적의 위험요인지도를 생성하는 것이 필요하다. 위험요인지도는 각각의 클러스터들이 공통의 위험요인을 보유하고 있는 유사한 재해들을 포함하는 방향으로 군집화하며, 그 결과를 시각적으로 표현한다. 또한, 위험요인지도는 현재 활용되고 있는 산업재해현황과 같은 재해통계에서 파악할 수 없는 재해문서 간의 관계를 토대로 위험요인을 군집화한다는 특징이 있다.

현재 작성되고 있는 재해 통계는 재해의 단편적인 특성이나 분류 카테고리를 가지고 분석을 수행하고 있으며, 특히, 사망재해가 가장 많이 발생하는 건설업에 대한 위험요인분석이 필요하다. 본 연구에서는 사망재해가 가장 많이 발생하는 건설업에 발생한 사망재해를 분석하여 위험요인지도를 작성하여 재해에 영향을 미치는 위험요인을 탐색하였다. 이를 위해, 서술형 문장으로 작성된 재해 개요를 텍스트마이닝 활용하여 재해 발생 개요에 포함된 주요 키워드를 도출하고 비지도학습 방법론인 SOM으로 군집화하였다. 이를 통해 재해의 주요 원인 키워드와 주요 작업분류 키워드, 주요 재해 형태 키워드를 도출하였다. 그 결과, 재해가 가장 많이 발생되는 추락재해에서 철근과 콘크리트와 같은 기인물로 인한 추락, 사다리와 로프로 인한 고소작업 시 추락, 비계와 발판 등 가시설물로 인한 추락으로 군집화되었으며, 굴착작업 시 발생하는 재해와 크레인으로 인한 재해로 크게 5가지로 군집화되었다. 또한 연도별 분석 결과, 굴착기로 인한 재해와 가시설물의 설치작업 시 발생하는 재해가 계속해서 증가하였고, 차량으로 인한 재해와 작업대와 추락으로 인한 재해가 감소 후 증가하는 추세를 나타내었다.

종합적으로, 군집된 5개의 셀 군집은 다음과 같은 위험요인을 포함하고 있다. 군집 1은 콘크리트와 철근과 같은 재료와 관련된 사고가 군집되었으며, 이동 시 혹은 설치작업의 작업 형태, 재해 형태는 추락을 위험요인으로 도출하였다. 군집 2는 고소작업 중 외벽에서의 작업이나 높은 장소로 이동 중에 발생한 재해가 군집되었으며, 외벽, 배관과 같은 작업 형태와 재해형태인 추락이 위험요인으로 도출되었다. 특히, 사다리와 로프가 지속적으로 고소작업 시 사용하는 도구에서 위험요인이 도출되었다. 군집 3은 굴착과 붕괴가 연관된 재해들이 군집되었으며, 굴삭기와 차량이 군집에서 연관성이 큰 위험요인으로 나타났으며, 다른 클러스터들과는 다르게 무너짐과 깔림이 재해 형태로 도출되었다. 군집 4는 비계와 발판과 같은 가시설물에 대한 재해가 군집되었으며, 가시설물의 설치·해체작업에서 발생한 추락이 주요 위험요인으로 도출되었다. 마지막으로 군집 5는 크레인과 연관성이 큰 재해들이 군집되어있으며, 크레인의 설치·해체작업과 추락이 주요 위험요인으로 도출되었다.

또한, 다른 셀 군집으로 군집화되었지만, 군집의 경계에 위치하고 있는 셀들은 거리에 따라 유사한 위험요인을 포함할 수 있기 때문에, 군집 간의 관계를 분석할 필요가 있다. 각 군집에 대하여 사망재해 형태와 기인물, 작업 형태와 같은 위험요인들을 비교해 볼 때, 연관성이 큰 위험요인을 가지고 있는 군집 1, 2, 4를 동시에 관리하고, 군집 3, 5에 대한 위험요인을 동시에 관리할 필요가 있다. 또한, 군집마다 유사하게 도출된 추락이 사망재해 형태라 하더라도, 기인물에 따라 자재(철근, 콘크리트), 안전 장비(사다리, 로프 등), 가시설물(비계, 발판)을 이용하는 경우의 차별화된 안전 지침을 고려해야 한다. 또한, 굴착 및 토사 작업 시와 크레인 작업 시에는 작업 기계를 설치할 때 붕괴와 관련된 사항을 필수적으로 점검하는 것이 요구됨을 도출할 수 있다.

이와 같은 결과는 기존의 일반적인 통계 방법으로 분석한 재해 통계와 상이함을 알 수 있다. 기존의 분석 방법의 경우, 발생형태와 기인물, 업종명 등 정해진 유형 내에서 전문가가 직접 분류하여 세부적인 재해의 키워드나 새로운 산업에서 발생하는 재해에 대한 분석에 어려움이 있다. 또한, 기술통계로 분석된 자료에서 얻고자 하는 재해의 정보를 찾아 비교·분석하는 어려움을 재해 개요의 텍스트 분석과 SOM을 통해 효율적이고 시각적으로 분석 가능하다고 사료된다.

제 4 장 건설업 재해문서의 지도학습

본 연구는 건설업 재해문서에서 사망재해와 부상재해를 분류하는 특성을 가진 위험요인을 도출하기 위하여 기계학습 알고리즘 중 지도학습인 분류 방법론을 활용하여 재해문서를 분류한 결과를 비교·분석하고, 분류된 재해를 결정짓는 주요 위험요인 키워드를 도출하고자 한다. 먼저, 기계학습 알고리즘을 적용하기 위하여 비정형데이터인 재해 발생 개요를 텍스트마이닝을 활용하여 정형데이터로 변환하는 전처리 과정을 수행한다. 키워드와 문서로 이루어진 정형데이터는 문서의 양이 늘어남에 따라 도출되는 키워드가 급격하게 증가하여 분석의 어려움이 있다. 이를 해결하기 위해 주성분분석(principal component analysis: PCA)을 통해 특성을 선정하여(feature selection) 도출된 키워드를 주요 특성으로 차원을 축소한다. 다음으로 지도학습 기반의 기계학습 알고리즘을 활용하여 특성에 따라 사망재해문서와 부상재해문서를 분류한다.

본 연구에서는 사망재해와 부상재해의 분류를 위하여 네 개의 대표적인 기계학습 알고리즘인 로지스틱 회귀분석(logistic regression), 의사결정 나무 분석(decision tree), 지지벡터 기계분석(support vector machine), 신경망 분석(neural network)을 사용하여 재해문서를 분류한다. 사망재해와 부상재해의 비교를 위해, 각 알고리즘의 문서분류 정확도를 도출하고 특히, 오분류된 문서를 통해 사망재해와 부상재해의 차이를 나타내는 핵심 위험요인을 도출한다. 이를 통해 부상재해에서 사망재해로 이어질 수 있는 위험요인을 도출하여, 사망재해를 감소시키고 이어 부상재해까지 관리할 수 있는 방안을 모색하고자 한다.

4.1 위험성 키워드 도출을 위한 재해문서 분류분석 방법

사망재해와 부상재해의 특성에 따른 위험성 키워드를 도출하기 위하여 지도학습 중 분류에 사용되는 네 개의 방법론을 활용하여 Fig. 23과 같이 분류연구를 수행하였다. 먼저, 산업재해통계에서 부상재해와 사망재해의 재해 발생 개요를 수집하였으며, TF-IDF를 활용하여 문서 내 키워드의 중요도에 따른 가중치를 도출하고 가중치를 기반으로 한 문서-키워드 매트릭스를 구조화하였다. 다음으로 차원 축소를 위해 PCA를 활용하여 재해의 특성을 선정하여 문서-PC 매트릭스를 작성하였다. 정형화된 매트릭스 데이터를 k-fold cross validation을 통해 training set과 test set으로 분할하였으며, 네 개의 분류방법론을 활용하여 재해문서를 분류할 수 있는 분류모델을 작성하였다. 분류모델은 일반적으로 많이 사용되는 logistic regression, decision tree, neural network, support vector machine을 활용하였다. 다음으로 각 모델의 confusion matrix를 작성하여 정확도 지표인 정밀도와 재현율, 정확도를 통해 각 모델의 정확도를 평가하였다. 마지막으로 분류모델에서 분류에 영향을 미치는 재해문서를 분석하기 위하여, 오분류된 재해문서를 추출하였다.

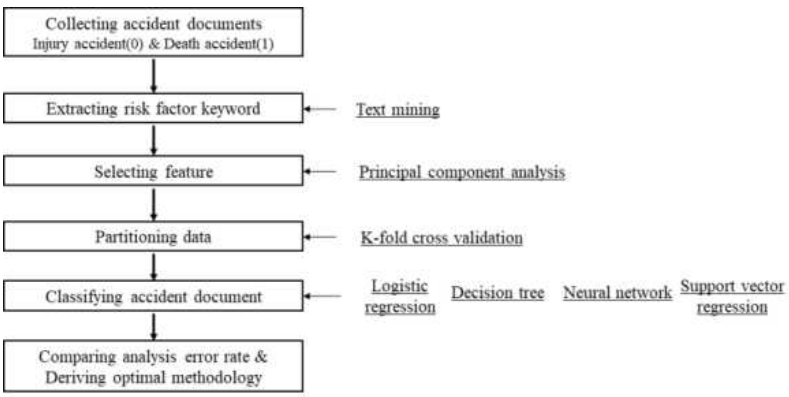

Fig. 23 Flowchart for the methodology of supervised learning

4.1.1 위험성 키워드 도출을 위한 텍스트데이터 전처리방법

재해문서에서 문서를 분류하는 특성을 갖고 있는 위험요인을 도출하기 위하여 재해율이 가장 높은 업종인 건설업에 대하여 2012년부터 2017년까지 발생한 부상재해와 사망재해의 발생 개요를 수집하였다. 그러나 사망재해문서와 부상재해문서의 불균형으로 인해 분류모델의 분석이 불가능하여, 2017년에 발생한 24,133건의 부상재해와 2012년부터 2017년까지 발생한 2,853건의 사망재해를 분석하였다. 산업재해조사표를 통해 수집된 데이터에서 도출할 수 있는 변수로는 연령, 업종, 사업장의 규모를 포함한 재해자의 정보와 재해 일자와 발생형태를 나타내는 재해의 전반적인 내용이 포함되어있다. 이와 같은 데이터에서 서술형으로 작성되어 재해의 전반적인 내용을 분석할 수 있는 텍스트데이터인 재해 발생 개요를 추출하여 엑셀을 활용하여 정리하였다. 결과적으로 26,986건의 재해문서를 수집하여 분석하였으며, 사망재해의 경우 1로, 부상재해의 경우 0으로 라벨링하여 분석을 수행하였다.

기계학습을 활용하여 재해문서를 분석하기 위하여, 데이터 전처리방법인 1) 텍스트 분석과 2) 차원 축소, 3) 데이터 분할을 수행하였다.

1) 텍스트 분석

수집된 사망재해문서와 부상재해문서의 재해 개요에 대한 기계학습을 수행하기 위해 필수적인 DTM 형태로 구조화하는 텍스트 분석을 수행하였다. 수집된 텍스트데이터에서 프로그램 R의 tm 패키지와 KoNLP 패키지에서 TF-IDF를 활용하여 DTM을 도출하였다. 텍스트 분석 코드는 부록 B에 작성하였다. 먼저, 위험요인의 효율적인 분석을 위하여 글자 수 2~5개로 이루어진 명사를 추출하였다. 다음으로 분석에 필요하지 않은 데이터를 전

처리하는 과정을 수행하였다. 이 과정에서 정확한 단어로 이루어져 있지 않은 불완전한 데이터와 분석에 부적절한 단어와 조사, 기호, 숫자와 같은 불용어를 stopwords 명령어를 통해 제거하였다. TF-IDF는 tm package에 내장된 가중치 명령어인 weightTfidf를 활용하여 재해문서와 위험요인 키워드로 이루어진 document-term matrix로 변환하였으며, excel을 이용하여 데이터를 추출하였다.

2) 차원 축소

데이터 분석의 효율성을 위하여 차원 축소를 수행하였으며, 차원 축소에 대표적으로 활용되는 PCA를 프로그램 R에 내장되어있는 prcomp() 함수를 활용하여 분석하였으며, 차원 축소에 활용한 코드는 부록 B에 작성하였다. prcomp() 함수는 숫자로 이루어진 데이터 프레임에 대하여 분석을 수행하기 전 변수가 단위 분산을 갖도록 하는 척도화 여부를 논리값으로 분석이 수행된다. PCA의 목적은 원 데이터의 분산을 최대한 보존하는 축(principal component: PC)을 찾아 투영하여 차원을 축소하는 것이다. 축소하는 차원의 수는 주성분 벡터의 explained variance ratio를 계산하여 누적 기여율(cumulative proportion)에 따라 적절한 차원의 수를 선정하여 문서-PC 매트릭스로 차원 축소를 진행하였다. 일반적으로, 각 PC에 대한 explained variance ratio를 PC와 variance로 이루어진 그래프에 작성하였을 때, elbow point에서 PC의 개수를 선택한다.

PCA를 활용하여 분류모델의 작성에 사용할 문서-PC 매트릭스와 전체 키워드를 축소하여 작성된 PC에 포함되어있는 키워드를 도출하기 위한 키워드-PC 매트릭스를 도출하였다.

3) 데이터 분할

본 연구는 한정된 데이터를 사용하여 분석을 진행하였기 때문에, 분류모델의 검증을 위하여 필요한 test set과 분류모델을 작성하기 위해 사용되는 training set으로 데이터 분할을 수행하였다. 데이터 분할은 k-fold cross validation을 사용하였으며, 상용 프로그램 R의 plyr package와 dplyr package를 활용하여 데이터 분할을 수행하였다. 데이터 분할에 활용한 프로그램 R의 코드는 부록 B에 작성하였다.

Fold를 나누는 파라미터인 k 값은 3부터 20까지 순차적으로 분석을 진행하여, 데이터의 분산과 편향이 적게 나타난 k 값을 도출하여 분석을 수행하였다. k가 낮을 경우, 모델의 평가에 대해 편중된 데이터가 나타날 수 있으며, k 값이 클 경우, 데이터의 분산이 크게 나타날 수 있다. 이와 같은 방법을 통해 총 5개의 fold로 교차검증을 수행하여 test set과 training set으로 데이터를 분할하였다.

4.1.2 분류 모델별 분석방법

재해문서를 통해 분류모델을 작성하기 위하여, 보편적으로 사용되고 있는 1) logistic regression과 2) decision tree analysis, 3) neural network analysis, 4) support vector machine으로 지도학습을 수행하였다. 각 분류모델의 분석 방법은 다음과 같으며, 분석을 수행하기 위한 파라미터를 도출하는 방법을 포함한 분석 방법을 서술하였다. 또한 분류모델을 작성하기 위해 사용한 코드는 부록 B에 작성하였다.

1) Logistic regression

Logistic regression은 회귀식을 작성하는 방법론으로 입력데이터를 분류할 수 있는 회귀식으로 사용할 수 있도록 기계학습을 수행한다. 분석 방법으로는 프로그램 R에 내장되어있는 generalized linear model(glm)과 anova, predict 명령어를 사용하여 지도학습을 수행한다. 먼저, 데이터 분할을 통해 작성된 training set으로 이진 분류를 위한 일반화 선형 모델을 작성한다. 작성된 모델을 카이제곱 검정을 수행하고, 작성된 모델에 test set으로 모델을 평가한다. test set을 분류기에 적용한 결과와 training set으로 작성한 분류모델의 데이터를 confusion matrix를 작성하여 모델의 정확도를 평가한다.

2) Decision tree analysis

Decision tree analysis는 특정 항목에 대한 의사 결정 규칙(decision rule)을 나무 형태로 분류해 나가는 분석기법으로, 분석과정을 관측할 수 있는 지도학습 모델이다. 분석 방법은 프로그램 R의 rpart package를 사용하였으며, rpart() 명령어를 통해 classification and regression

trees(CART) 방법론을 사용하여 training set으로 tree 모델을 작성하였다. CART 방법론의 경우 과적합화 문제가 발생할 수 있기 때문에, 이를 해결할 수 있는 가지치기(pruning) 과정을 수행하였다. 가지치기 과정은 교차검증을 통해 xerror가 가장 낮은 분화 개수를 선택하여 분석하였다. 작성한 decision tree 분류모델에 대해서 test set을 적용하였다. 마지막으로 Training set의 분류 데이터와 test set을 적용한 데이터를 confusion matrix를 통해 정확도를 도출하였다.

3) Neural network analysis

Neural network analysis는 생물학의 신경망에서 영감을 얻은 통계학적 기계학습 방법으로, 입력층과 은닉층, 출력층으로 구분되어있다. 특히, 입력데이터로부터 출력데이터까지 학습을 진행하는 feed forward와 출력층부터 입력층까지 데이터의 오류를 최소화하는 back propagation으로 모델의 오류를 최소화한다.

본 연구는 프로그램 R의 neuralnet package를 사용하여 모델을 작성하였다. 은닉층의 개수와 노드의 수는 tolerable error가 가장 작게 나타난 노드 수와 은닉층 수를 선정하였으며, 두 개의 은닉층으로 첫 번째 은닉층은 20개의 노드를 가지고 있으며, 두 번째 은닉층은 5개의 노드를 가지고 있는 분류모델을 작성하였다. 일반적으로 많은 은닉층이 존재할 경우 과적합 문제와 계산속도가 느려지는 문제가 발생한다.

4) Support vector machine

SVM은 분류 방법론 중에서 가장 정확도가 높고 다양한 분류 상황에서 좋은 성능을 보이며, 이상치의 영향도 적게 받는 방법론으로 알려져 있다. SVM 분석 방법은 상용 프로그램인 R의 e1071 package의 svm()함수를 활

용하여 재해문서의 분류를 수행하였다. e1071 package는 SVM을 수행할 수 있는 여러 package 중 가장 먼저 소개되었으며, 가장 직관적인 분석 모듈이다.

선행 연구데이터가 없을 경우 분류모델을 학습하기 위해 주로 사용되는 가우시안 함수를 커널 함수로 사용하였다. 선형커널을 제외한 모든 커널에서 요구되며 초평면의 기울기를 나타내는 파라미터인 gamma와 과적합에 대한 제약 위배의 비용을 의미하는 파라미터인 cost를 계산하였다. 최적의 파라미터값을 구하기 위하여, tune.svm() 함수를 이용하여 gamma는 2-5부터 1까지, cost는 1부터 16까지의 오류값을 계산하였다. 이를 통해, gamma는 0.03125, cost는 4로 SVM을 분석하는 것이 재해문서를 분류하기에 가장 적합한 것으로 나타났다.

4.2 재해문서의 분류 결과

4.2.1 텍스트 분석 결과

수집된 재해문서를 텍스트 분석을 통해 문서-키워드 매트릭스로 구조화하여 27,285개의 문서와 100개의 키워드로 이루어진 행렬을 작성하였으며, 도출 빈도가 높은 상위 키워드 100개와 키워드 가중치의 합계가 0인 문서를 제거하였다. 최종적으로 총 25,837개의 문서와 100개의 키워드에 대하여 분석을 수행하였다. TF-IDF 결과는 Fig. 24와 같이 문서번호와 키워드로 이루어진 매트릭스를 구조화하였으며, 재해문서에 포함되어있는 키워드의 가중치를 나타내어 작성하였다. 대표적으로 추락, 차량, 바닥, 설치 순으로 가중치의 합이 높은 키워드로 도출되었으며, 기존의 빈도 위주로 분석했던 DTM과는 다르게 재해와 직접적으로 연관된 키워드들 위주로 도출되었다.

Term / Documents	추락	차량	바닥	설치	높이	…
1	0	0.3538	0	0	0	
2	0.0463	0	0.0487	0	0	
3	0.2352	0	0	0.6686	0	
4	0	0.1061	0.2612	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	
25837	0	0	0	0.1955	0.1671	

Fig. 24 Document-term matrix using TF-IDF

4.2.2 PCA를 활용한 재해문서의 특성 선정 결과

키워드의 특성을 선정하고 행렬의 크기를 줄이기 위하여 principal component analysis를 활용하였다. 분석에 적절한 PC의 수를 선정하기 위하여, Fig. 25와 같이 explained variance ratio를 계산하여 누적 기여율이 90% 이상을 나타내는 88개의 PC를 선정하였다.

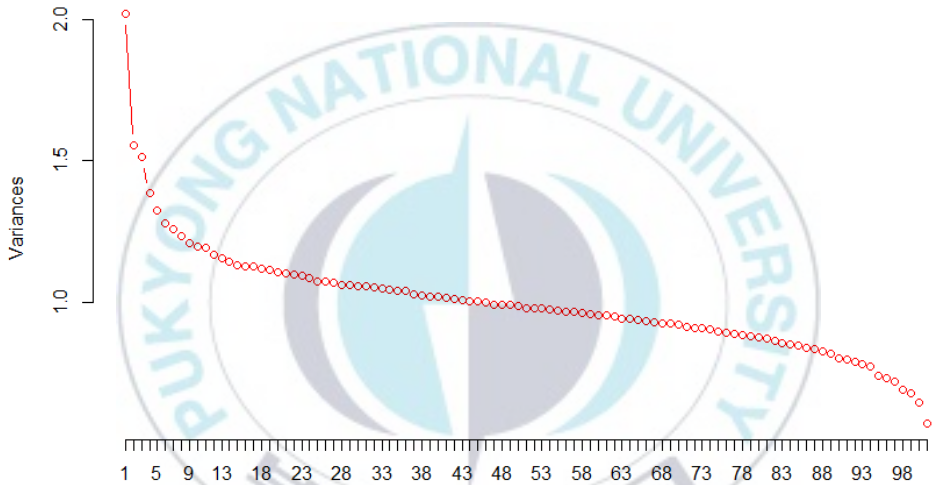

Fig. 25 Variance for each PC

Explained variance ratio가 높은 사고문서의 특성을 가장 많이 포함하고 있는 PC1의 경우, '절단'과 '손가락', '톱날', '엄지', '합판' 등의 키워드들이 분포되어있으며, 절단 작업 시 톱날로 인해 발생할 수 있는 위험요인들이 포함되어있다. PC2의 경우 '해체', '거푸집', '파이프', '비계', '형틀' 등의 키워드들이 분포되어있으며, 가시설물의 해체작업 시 발생할 수 있는 위험요인들이 포함되어있다.

PCA를 통해 도출된 문서-PC 매트릭스는 다음 Fig. 26과 같다. 가장 많은 가중치를 가지고 있는 PC부터 나열하였으며, 각 문서에 해당되는 특성을 PC의 조합으로 나타내었다. A열은 문서번호를 나타내며, B열은 사망재해는 1, 부상재해는 0으로 표기하여 사망재해와 부상재해를 라벨링하였다. 각 문서는 PC1부터 PC88까지의 조합으로 재해문서의 특성을 나타내며, 각 PC에 포함되어있는 위험요인을 분석하기 위해서 키워드와 PC로 구성된 매트릭스를 구조화하여 도출한다.

PC / Documents	Label	PC1	PC2	PC3	PC4	…	PC88
1	1	-1.5087	-0.5118	-1.0224	1.2419	…	-1.1859
2	1	-1.2843	1.3996	-0.0307	0.3879	…	0.0879
3	1	-1.3977	1.2075	0.2767	0.5976	…	0.0152
4	0	-0.3059	0.5403	0.2304	-1.3621	…	0.3487
⋮	⋮	⋮	⋮	⋮	⋮		⋮
25837	0	-1.3225	1.3657	0.3887	0.1845	…	0.1727

Fig. 26 Document-PC matrix for PCA

4.2.3 k-fold cross validation을 활용한 데이터 분할 결과

분류모델의 평가를 위하여 생성된 매트릭스를 k-fold cross validation을 사용하여 training set과 test set으로 분할하였다. 전체 데이터를 2에서 30까지의 k 값에 대하여 민감도 분석을 수행하여, 정확도가 가장 높게 나타난 5개의 fold로 나누어 교차검증을 수행하여 다음 Table 9와 같은 data set을 작성하였다. 그 결과, training set은 18,479건의 부상재해문서와 2,268건의 사망재해문서로 구성되었으며, test set은 4,505건의 부상재해문서와 585건의 사망재해문서로 분할하였다. 결과적으로, 20,747개의 training set으로 4개의 분류 방법론을 통해 각 모델을 학습시키고, 5,090개의 test set으로 모델을 평가하였다.

Table 9 Number of training and test set

	Training set	Test set
Non-fatal disaster (0)	18,479	4,505
Fatal disaster (1)	2,268	585
Total	20,747	5,090

4.2.4 모형별 분류 결과

분류모델별 정확도 분석 결과와 confusion matrix는 Table 10과 같으며, 각 모델의 작성에 필요한 파라미터는 4.1.2 분류모델별 분석 방법에서 서술한 방법으로 결정하여 연구를 수행하였다.

먼저, logistic regression의 모델은 R의 generalized linear model을 사용하여 작성하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 18개, 2종 오류가 26개로 나타났으며, 모델의 정확도 지표 결과는 정밀도는 0.9688, 재현율이 0.9538, 정확도가 0.9914로 나타났다.

다음으로, decision tree 모델은 R의 rpart package 중 지니지수로 순도를 계산하는 CART를 사용하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 108개, 2종 오류가 301개로 나타났으며, decision tree 모델의 정확도 지표 결과는 정밀도가 0.7244, 재현율이 0.4855, 정확도가 0.9196으로 나타났다.

Neural network 분류모델은 R의 neuralnet package를 사용하였으며, tolerable error가 가장 낮게 나타난 2개의 hidden layer로 총 node = (20,5)로 분석하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 14개, 2종 오류가 9개로 나타났으며, 정확도 지표 결과는 정밀도가 0.9763, 재현율이 0.9846, 정확도가 0.9955로 나타났으며, 4개의 모델 중 정확도가 가장 높게 나타났다.

마지막으로, SVM 모델은 R의 e1071 package를 사용하였다. 먼저, 2-5부터 1까지의 gamma값과 1부터 24까지의 cost값을 각각 대입하여 모델의 최적값을 계산하였다. 그 결과, gamma = 0.03125, cost = 4의 최적값을 도출하였으며, support vector의 수는 13,349개로 나타났다. 본 연구에서는 kernel function 중 RBF(radial basis function) - gaussian kernel을 사용

하여 training data set으로 분류모델을 작성하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 17개, 2종 오류가 163개로 나타났으며, SVM 모델의 정확도 지표 결과는 정밀도는 0.9613, 재현율은 0.7213, 정확도는 0.9646으로 나타났다.

모델별 정확도 지표를 살펴보면, neural network 분류모델이 전체적으로 가장 높은 정확도를 나타냈으며, 다음으로 logistic regression 분류 모델과 SVM 분류모델, decision tree 모델의 순서로 정확도가 높게 나타났다. Neural network와 logistic regression의 경우, 모든 정확도 지표에서 95% 이상의 정확도를 보이고 있으며, 가장 낮은 정확도를 보인 모델인 decision tree는 재현율에서 50% 미만의 정확도가 나타났다.

Table 10 Confusion matrix and degree of accuracy for each methodology

	Confusion matrix			Precision	Recall	Accuracy
Logistic regression		Predicted : Fatal	Predicted : Non-Fatal	0.9688	0.9538	0.9914
	Actual : Fatal	559	26			
	Actual : Non-Fatal	18	4487			
Decision Tree		Predicted : Fatal	Predicted : Non-Fatal	0.7244	0.4855	0.9196
	Actual : Fatal	284	301			
	Actual : Non-Fatal	108	4397			
Neural Network		Predicted : Fatal	Predicted : Non-Fatal	0.9763	0.9846	0.9955
	Actual : Fatal	576	9			
	Actual : Non-Fatal	14	4491			
Support Vector Machine		Predicted : Fatal	Predicted : Non-Fatal	0.9613	0.7213	0.9646
	Actual : Fatal	422	163			
	Actual : Non-Fatal	17	4488			

4.2.5 분류 키워드 분석 결과

사망사고와 부상사고를 분류하는 위험요인을 분석하기 위하여, 사고문서를 분류하는 PC를 파악하고, PC에서 위험요인 키워드를 도출하였다. 분류모델의 해석은 단순성과 투명성, 명료성을 기준으로 해석이 가능한 화이트박스(white box)와 분석과정의 해석이 불가능한 블랙박스(black box) 모델로 나누어진다. 먼저, 화이트박스 모델은 일반적으로 decision tree와 rule-lists, 회귀 알고리즘이 포함되며, 예측변수가 많지 않을 경우 이해하기 쉬운 모델을 의미한다. 다음으로 블랙박스 모델은 일반적으로 neural network와 랜덤 포레스트 등이 있으며 모델 내부에서 분석되는 과정을 시각화하거나 이해하기 어렵고, 목적변수와의 연관 관계를 파악하기 어렵다. 본 연구에서 작성한 분류모델 중 화이트박스 모델인 logistic regression과 decision tree에 대하여 재해문서의 분류에 대한 표준 추정값을 도출하였으며, 도출된 PC에 대하여 키워드-PC 매트릭스를 통해 PC가 포함하고 있는 키워드를 분석하였다.

먼저, 3개의 정확도 지표에서 95% 이상의 정확도를 나타낸 logistic regression의 경우, 사망사고와 부상사고에서 도출된 PC는 Table 11과 같으며, 사망사고에 대하여 표준 추정값이 큰 PC는 PC4과 PC6, PC23의 순서로 도출되었다. PC에 포함되어있는 위험요인 키워드를 살펴보면 PC4는 '해체', '거푸집', '비계', '설치', '추락'의 키워드를 포함하고 있으며, 가시설물의 설치·해체 시 발생하는 추락사고에 대한 재해특성을 갖고 있다. PC6은 '상부', '배관', '하부', '철골', '용접', '크레인'의 위험요인을 포함하는 PC로 나타났으며, 배관작업 및 용접 시 발생하는 사고의 특성을 나타낸다. PC23은 '목공', '형틀', '용접', '아파트', '철골', '기계' 등의 키워드를 포함하고 있으며, 목공 및 형틀 작업과 기계로 인해 발생한 사고의 특성이 도출되었다.

부상사고에 대해 표준 추정값이 큰 PC는 PC1과 PC9, PC5의 순서로 도출되었다. PC1에 포함되었는 키워드를 살펴보면, ‘절단’, ‘손가락’, ‘톱날’, ‘엄지’, ‘합판’, ‘전기톱’을 포함하고 있으며, 절단 작업 시 톱날로 인한 위험요인들이 도출되었다. PC9는 ‘철근’, ‘조립’, ‘골절’, ‘사다리’, ‘용접’, ‘설치’의 키워드가 도출되었으며, 용접 및 조립작업과 관련된 위험요인이 포함되어 있다. PC5는 ‘골절’, ‘해체’, ‘화장실’, ‘손목’, ‘배관’, ‘철거’의 키워드가 포함되어 있으며, 배관 및 철거작업으로 인한 골절이 위험요인으로 나타났다. 이와 같이 logistic regression을 통해 각 사고에 대해 분류 가중치가 큰 PC의 조합으로 사망사고와 부상사고를 분류하며, PC에서 도출한 키워드를 통해 사고의 위험성을 의미하는 위험요인을 도출하였다.

Table 11 Classification factors of logistic regression

	Number of PC	Overview of disaster characteristics
High weight for fatal accidents	PC4	Fall Disasters that occur during installation and dismantling of temporary facilities
	PC6	Disasters during piping work and welding
	PC23	Disasters caused by machinery during wood working and form working
High weight for non-fatal accidents	PC1	Hazards from saw blades when cutting
	PC9	Hazards associated with welding and assembly operations
	PC5	Fractures due to piping and demolition work

다음으로 decision tree 모델의 경우, 다음 Fig. 27과 같은 분화과정을 분석할 수 있으며, 분화 기준인 PC와 그에 따른 위험요인은 다음 Table 12와 같다. 먼저 PC1을 기준으로 분류를 수행하며, PC9와 PC25, PC4, PC78의 순서로 분화가 진행된다. PC1이 -1.111보다 크거나 같은 재해문서가 부상재해로 분류되며, PC1은 '철근', '손가락', '톱날', '엄지', '합판', '전기톱'의 위험요인이 포함되어있어 절단 작업 시 톱날로 인해 발생할 수 있는 재해가 분류된다. 다음으로, PC1에서 사망재해로 분류된 문서 중 PC9가 -0.08168보다 크거나 같은 재해문서는 사망재해로 분류되며, '조립', '골절', '사다리', '용접', '설치', '낙상'의 위험요인을 포함한 용접 및 설치작업 시 발생한 재해문서가 분류된다. PC9에서 사망재해로 분류된 문서 중 PC25가 -0.1674보다 크거나 같은 재해문서가 분류되며, '발목', '목공', '크레인', '형틀', '설치', '인양'의 위험요인 키워드가 포함되어있으며, 크레인의 설치·인양 작업 등으로 인한 재해문서가 분류된다. PC25가 부상재해로 분류한 재해문서 중 PC4가 0.2254보다 작은 문서에 대해 부상재해로 분류하며, PC4는 '해체', '거푸집', '비계', '설치', '추락', '외부'의 위험요인을 가지고 있으며, 가시설물 설치·해체 작업 시 발생하는 재해가 분류된다. 마지막으로 PC4에서 사망재해로 분류된 문서 중 PC78이 -0.2492보다 작은 문서에 대해 사망재해와 부상재해로 분류한다. PC78의 키워드로는 '기계', '얼굴', '주택', '엄지', '크레인', '제거'가 도출되었으며, 크레인 등 기계를 이용한 작업 시 발생한 재해가 포함되어있다.

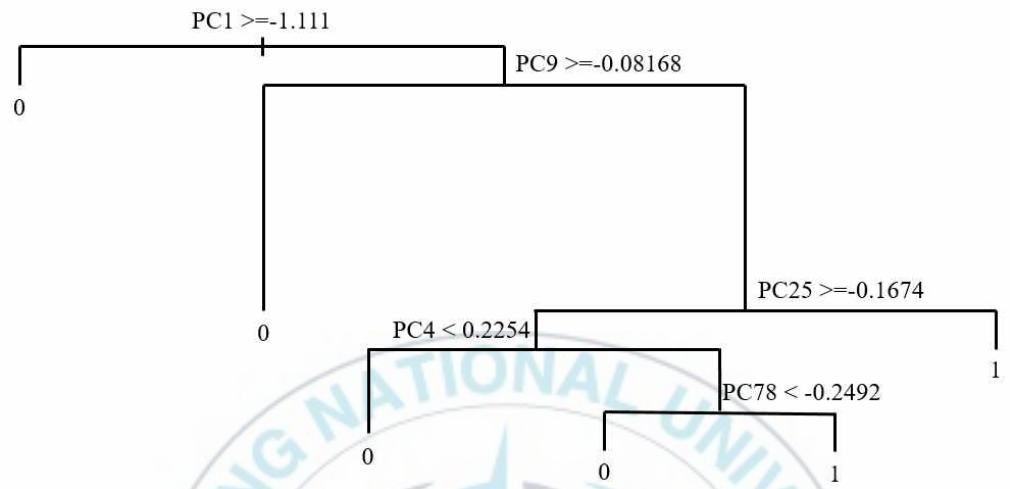

Fig. 27 Differentiation process of decision tree

Table 12 Differentiation order and classification factors of decision tree

Differentiation order	Number of PC	Overview of disaster characteristics
1	PC1	Disasters caused by saw blades when cutting
2	PC9	Disasters that occurred during welding and installation work
3	PC25	Disasters caused by crane installation and lifting work
4	PC4	Disasters that occur during installation and dismantling of temporary facilities
5	PC78	Disasters that occur when working with machines such as cranes

4.2.6 오분류된 재해문서의 키워드 분석

Heinrich는 부상재해에서 도출되는 위험요인을 파악하고 개선하여 부상재해뿐만 아니라 사망재해까지 예방할 수 있다는 이론을 제창하였다. 이와 같은 이론을 바탕으로 본 연구에서는 국내 건설업에서 발생한 사망재해와 부상재해의 관계를 파악하기 위하여 재해문서를 분류하여 비교·분석하였다. 재해문서의 분류는 위험요인 키워드를 포함하고 있는 PC에 따라 분류되며, 각 재해에서 많이 발생하는 위험요인이 문서에 포함되어있는가에 따라 문서를 분류한다. 재해문서의 오분류는 문서가 실제와는 다른 예측치가 나타났다는 점에서 유사한 위험요인이 포함되어있을 수 있다. 따라서 오분류된 재해문서를 분석하여, 사망재해와 부상재해의 경계에서 나타나는 위험요인 키워드와 재해 프로세스를 분석할 필요가 있다.

오분류는 2가지로 1종 오류(Type Ⅰ error)와 2종 오류(Type Ⅱ error)로 나눌 수 있으며, 그 정의와 예측 및 의미는 Table 13과 같다. 1종 오류는 실제 부상재해로 나타난 문서를 사망재해로 잘못 예측한 오류로, 사망재해와 연관성이 큰 위험요인을 포함하고 있는 문서가 도출되었다. 이는 고위험 부상재해를 의미하며, 사망재해로 이어질 뻔한 부상재해, 손실의 우연적인 결과로 인해 부상재해로 나타난 재해, 부상재해이지만 시정조치를 강하게 해야 할 재해, 고용노동부 또는 안전보건공단과 같은 기관에서 부상재해지만 사망재해의 정도로 인지하고 관리해야 할 재해 등으로 그 의미를 해석할 수 있다. 2종 오류는 실제 사망재해로 나타난 문서를 사망재해로 잘못 분류한 오류로, 부상재해와 연관성이 큰 위험요인을 포함하고 있는 문서가 도출되었다. 이는 우연성에 의한 사망재해를 의미하며, 부상재해에 가까울 뻔 했던 재해, 사망으로 이어진 특별하고 우연적인 이유를 분석해야 할 필요가 있는 재해를 의미한다.

Table 13 Definition of misclassification and implications of prediction and management

Error type	Definition	Implications of prediction and management
Type Ⅰ error	Error in predicting actual non-fatal accident as fatal accident	**High risk non-fatal injuries** • Non-fatal disaster that almost led to fatal accident • Disaster that resulted in non-fatal accident due to accidental loss • Non-fatal accident, but disaster that will strengthen corrective action • Disaster that is non-fatal injured, but needs to be recognized and managed as fatal accident by the Ministry of Employment and Labor or the Korea Occupational Safety and Health Agency
Type Ⅱ error	Error in predicting actual fatal accident as non-fatal accident	**Accidental fatal injuries** • Fatal disaster that was close to non-fatal accident • Disaster that resulted in fatal accident due to accidental loss • Disaster to analyze the special and contingent reasons that led to fatal accident

각각의 오류에 대한 위험요인 키워드 분석을 수행하기 위하여, 가장 높은 정확도를 나타낸 neural network 분류모델에 대해서 오분류된 문서를 탐색하였으며, 1종 오류에 해당하는 9개의 문서와 2종 오류에 해당하는 14개의 문서가 도출되었다. 그 중, 각 오류에 대한 재해문서를 도출하여 키워드 분석을 수행하였다.

1종 오류로 분류된 재해문서를 살펴보면, 2854번 문서의 경우, 위험요인 키워드는 '부라캣', '설치', '추락'으로 나타났다. 이 문서의 경우, 부라캣 설치작업 중 추락으로 인해 부상재해가 발생하였으며, 추락으로 인한 재해의 경우 사망재해로 이어질 수 있으며, 추락에 대한 안전 대책이 필요한 것을 알 수 있다. 다음으로 2882번 문서는 '리모델링', '철거작업', '벽체', '머리', '충격'의 단어로 위험요인 키워드가 도출되었으며, 벽체의 무너짐과 작업자의 보호구에 대한 대책이 필요하다. 또한, 2913번 문서의 위험요인 키워드는 '철근', '설치', '발판', '높이', '추락'과 같은 위험요인 키워드가 도출되었다. 1종 오류에 해당하는 문서의 경우, 사망재해에서 자주 도출되는 키워드들로 구성되어 있으며, 특히, 2882번 문서와 2913번 문서는 사망재해문서에서 가장 많이 도출되는 '철거작업'과 '머리', '설치', '발판', '높이', '추락'이 포함되어있다. 이처럼 부상재해에서 도출된 키워드임에도 불구하고 사망재해로 발생할 위험이 있는 것으로 나타난다. 이를 통해 1종 오류에 해당하는 문서들은 도출된 키워드의 위험성을 확인하고, 해당 키워드의 관계 분석을 통해 위험성을 탐색해야 할 필요가 있다.

2종 오류로 분류된 2712번 문서를 살펴보면, 위험요인 키워드로는 '단독주택', '신축공사장' '비계작업', '바닥', '응급실', '중환자실' 등으로 나타났다. 이와 같은 키워드들을 통해, 비계작업 시 안전관리가 필요하다고 볼 수 있다. 2793번 문서의 경우, '가스통', '토치', '스티로폼', '호스', '밸브', '응급처치', '구급차' 등이 주요 키워드로 도출되었으며, 키워드들을 통해 토치 사

용 시 발생한 화재로 호스나 밸브에 대한 관리가 필요하며, 화재가 발생할 수 있는 주변 환경에 대해서도 관리가 필요할 것으로 사료된다. 2종 오류로 분류된 2793번 문서의 경우, 가스통과 토치, 밸브 등의 부상재해와 관련성이 크게 나타난 키워드들이 분포되었지만 3~4초 동안 불에 노출되어 사망재해로 이어진 재해 발생 개요다. 특히, 2793번 문서의 위험요인 키워드를 살펴보면, 토치, 스티로폼, 호스 등 화상으로 이어지는 부상재해와 관련이 깊은 키워드들이 도출되었다. 이처럼 부상재해와 연관성이 큰 위험요인 키워드들도 사망재해로 이어질 수 있다.

4.3 소결

본 연구는 2017년 국내 건설업에서 발생한 부상재해와 사망재해를 비교·분석하여 두 재해 간에 위험요인의 차이를 파악하고 부상재해가 사망재해로 이어지지 않는 시사점을 파악하기 위하여 다음과 같이 연구를 진행하였다. 먼저, 재해 발생 개요를 수집하여 비정형데이터의 정형화와 특성 선정을 텍스트마이닝과 PCA를 활용하여 행렬 데이터로 작성하였다. 다음으로 k-fold cross validation을 활용하여 전체 데이터를 training set과 test set으로 분할하였으며, 분할된 training set으로 각 분류모델을 작성하고, test set으로 분류모델의 정확도를 평가하였다. 평가된 분류모델은 neural network와 logistic regression, support vector machine, decision tree의 순서로 정확도가 높게 나타났으며, 특히 neural network 분류모델은 사용된 3개의 정확도 지표 모두 95% 이상의 정확도를 나타내었다. Decision tree 분류모델은 데이터를 선형 분류하며, 단계적으로 분류를 진행하는 특성으로 재해문서의 분류에 정확도가 낮게 나온 것으로 사료된다.

또한 위험성이 큰 요인을 탐색하기 위하여 부상재해와 사망재해에 대한 분류모델을 구축하고 평가하였으며, 각 분류모델의 정확도를 계산하여 재해문서의 분석에 적합한 모델을 탐색하였다. 2017년 국내 건설업에서 발생한 재해의 경우 neural network 분석모델이 재해문서의 분류에 적합하다고 판단된다. 또한, 오분류된 문서들을 탐색하여 부상재해로 발생하였지만 비슷한 재해가 발생하였을 때 사망재해로 이어질 수 있는 위험요인 키워드들과 부상재해의 주요 키워드 중 사망재해로 발생할 수 있는 키워드를 탐색하였다.

사망재해와 부상재해를 분류하여 산업재해조사표를 수집하는 재해문서의 특성으로 실제로 발생한 재해문서에 대하여 분류모델을 수행할 필요가 없

다. 그러나 분류모델에서 도출된 사망재해와 부상재해를 분류하는 키워드를 분석하여 위험요인을 탐색할 수 있으며, 실제 재해와는 다르게 1종 오류와 2종 오류로 오분류된 재해문서를 분석하여 재해를 세부적으로 파악할 수 있다. 1종 오류에 해당하는 문서는 사망재해에서 나타나는 특성을 가지며 사망재해와 연관성이 높다고 볼 수 있다. 따라서, 고위험 부상재해를 의미하여 재해의 결과가 부상으로 나타났지만, 주의를 기울일 필요가 있는 문서로 판단된다. 또한, 2종 오류는 부상재해에서 주로 나타나는 키워드들로 구성된 문서가 분류되었으며, 이와 같은 오류는 재해 결과의 우연성에 의해 재해의 결과가 다르게 나타나는 우연성 사망재해로 볼 수 있다. 1종 오류와 같이 사망재해와 연관성이 높은 위험요인을 포함한 PC가 도출된 경우, 해당 위험요인뿐만 아니라 같은 PC에서 도출된 위험요인들에 대한 관리가 필요하다. 부상재해와 사망재해의 특성은 다르게 나타나며 사망재해에 대해 집중적으로 관리할 필요가 있지만, 2종 오류와 같이 우연성에 의한 사망재해로 분류되는 재해들에 대한 관리도 필요하다고 생각된다.

따라서 두 종류의 오분류는 업종과 작업내용, 재해의 기인물의 특성에 따라 비슷한 재해 프로세스에서 부상재해와 사망재해가 다르게 발생할 수 있으며, 재해의 분류뿐만 아니라 오분류에 대한 분석이 필요하다는 것을 시사한다. 사망재해가 가장 많이 발생하는 건설업과 관련된 재해문서의 경우는 고위험 부상재해인 1종 오류가 많이 발생할 수 있으며, 특히 고소작업의 경우, 낮은 높이에서 작업 중 추락으로 인한 재해의 경우, 실제로 부상재해인 재해 프로세스에서 사망재해가 많이 발생하는 '추락'과 같은 키워드로 인해 1종 오류가 발생할 수 있어 관리가 필요하다.

제 5 장 결론

현장뿐만 아니라 정책적으로 재해 예방에 대한 다양한 노력에도 불구하고, 건설업에서는 재해가 지속적으로 발생하고 있다. 서론에서 서술하였듯이 재해 예방을 위하여 기존에 발생한 재해에 대한 분석은 필수적이다. 이에 따라 본 논문에서는 축적된 재해보고문서에 대한 새로운 재해분석 방법을 사용하여, 재해와 관련된 키워드 형태의 위험요인을 도출하고 이를 활용할 수 있는 방안에 대해 제시하기 위해 비지도학습과 지도학습을 활용하여 연구를 진행하였다.

첫 번째로, 재해문서가 포함하고 있는 위험요인에 따라서 재해문서를 군집화하고 군집된 재해문서가 가지고 있는 위험요인과 각 위험요인의 연관성을 도출하였다. 데이터를 군집화하고 시각화하기 위해 사용되는 비지도학습 방법인 SOM을 활용하여 사망재해가 가장 많이 발생하는 업종인 건설업에 대해, 재해문서가 가지고 있는 특성에 따라 군집화하고 위험요인지도를 작성하였다. 위험요인지도는 기존의 텍스트로 이루어진 보고서보다 직관적으로 이해할 수 있으며, 군집화된 재해 유형에 따라 위험요인을 효과적으로 파악할 수 있다. 특히, 셀 군집 분석을 통해 도출된 대표적인 5개의 재해문서 군집을 도출하였으며, 해당 재해문서 군집에 포함되어있는 위험요인들의 연관성을 분석하여 재해분석 및 안전관리가 필요한 위험요인을 도출하였다. 또한, 군집화된 문서에 대하여 위험요인의 연도별 동적 분석을 통해 유형에 따른 추세를 살펴보았다.

두 번째로, 사망재해와 부상재해의 분류에 영향을 미치는 위험요인을 도출하였다. 건설업에서 발생한 재해에 대하여 재해문서를 분류하기 적합한 지도학습 모델을 분석하고 오분류된 문서를 통해 키워드를 도출하였다.

분류에 사용한 neural network와 logistic regression, support vector machine, decision tree에 대하여 정확도 평가 및 검증을 수행하고, 특히 오분류된 문서들을 탐색하여 부상재해로 발생하였지만 비슷한 재해가 발생하였을 때 사망재해로 이어질 수 있는 위험요인 키워드들과 부상재해의 주요 키워드 중 사망재해로 발생할 수 있는 키워드를 탐색하였다. 이를 통해 고위험 부상재해와 우연성 사망재해에 대해 분석하였으며, 재해 발생 과정에서 나타나는 위험요인을 위험성에 따라 관리할 수 있을 것으로 사료된다.

5.1 연구 기여점

최근 20년 동안 평균적으로 매년 90,000건 이상의 산업재해조사표가 수집되고 있다. 산업재해조사표의 수집은 최근에는 더욱 증가하고 있으며, 현재 산업재해조사표에 관한 분석 방법은 데이터를 수집하고 재해 발생 개요에 포함되어있는 위험요인을 업종별, 재해 발생형태, 사업장 규모와 같은 정해진 카테고리에 따라 분류·분석하여 산업재해통계로 공시하고 있다. 즉, 지속적으로 축적되고 있어 데이터의 활용 가치는 증가하고 있다. 그러나 아직까지 유형화된 카테고리에 대해서 기술통계로 재해 현황을 분석하고 있는 실정이며, 산업재해조사표의 분류 및 분석에 많은 시간과 인력이 사용되고 있다. 산업재해통계는 재해의 빈도만을 분석하여 카테고리에 따라 작성되며, 정해진 카테고리 안에서만 분석이 가능할 뿐, 기계학습과 같은 방법론을 활용하여 분석하기에 어려움이 있다. 최근에는 안전보건공단에서 공시하고 있는 재해사례와 발생한 재해에 대한 뉴스기사 등과 같은 재해문서의 텍스트 분석과 기계학습을 활용한 연구가 진행되고 있지만 아직까지 미비한 실정이다.

본 연구의 기여점은 Table 14로 정리하였으며, 학술적 기여점은 첫 번째로, 텍스트로 이루어진 재해 개요의 활용성에 대해 제시하였다. 재해 개요의 비지도학습을 통해 재해 발생 과정에서 나타나는 위험요인 간의 연관성을 기존의 재해분석보다 효과적으로 분석할 수 있으며, 위험요인지도를 통한 셀 군집에서 도출된 재해 유형을 토대로 재해의 특성이 유사하게 나타난 위험요인들을 분석할 수 있다. 또한 산업재해조사표 수집 시 새로운 재해 발생 개요 및 원인을 기존의 분석모델에 적용하여 더욱 빠른 분석이 가능하다.

두 번째로, 재해문서의 텍스트 분석과 기계학습을 활용하여 분석모델을

작성하였고, 재해문서의 데이터 분석방법론에 대한 방향성을 제시하였다. 산업재해통계를 활용하여 분석했던 빈도분석 위주의 기존 연구들과 비교하여, 기계학습 방법론을 활용하여 안전관리를 위해 사용할 수 있는 새로운 방안을 제시하였다. 재해 개요의 텍스트 분석을 통하여 기인물과 재해형태와 같은 기존의 카테고리가 아닌 작업 형태, 재해 부위와 같은 새로운 유형의 위험요인 카테고리를 탐색할 수 있으며, 지도학습을 통해 연구자가 분석에 필요한 카테고리를 직접 선정하여 원하는 정보를 도출할 수 있을 것으로 사료된다. 또한, 위험성이 높은 부상재해와 우연하게 발생한 사망재해에 대한 분석을 통해 도출된 위험요인을 다양하게 분석하여 실질적인 재해분석에 도움이 될 것이라고 생각된다.

본 연구의 실무적 기여점은 첫 번째로, 비지도학습을 통해 작성한 위험요인지도를 통해 개별적으로 사람이 해석하기 힘든 텍스트로 이루어진 문서에서 위험요인을 추출·분석하여 이해하기 쉬운 지도로 시각화하여 위험요인 간의 관계를 제시하였다. 산업현장에서는 동종업종의 재해 현황을 파악하고 그에 따른 유사재해를 방지하기 위한 대책을 수립하기 위하여 산업재해통계분석을 활용하고 있다. 본 연구를 통해 동종업종이 아니어도 유사한 위험요인을 가지고 있는 사업장에 대하여 그에 따른 대책을 마련하는 것이 가능하다.

두 번째로, 지도학습을 통해 도출한 문서의 분류에 영향을 미치는 위험요인을 파악하여 현장에서 발생한 재해를 세부적으로 분석하고 예방할 수 있다. 세부적인 재해의 주요 요인의 유형과 키워드를 제시하여 실제 현장에서 나타나는 위험요인과 연관이 있는 요인을 파악하여 안전관리자와 관리감독자들의 효과적인 안전관리에 도움이 되리라 기대된다.

Table 14 Contributions of this study

		Previous	Current
Academic contribution	Use of data	Descriptive statistical analysis only for categorized categories despite the use value of data	Presenting the usefulness of the disaster summary made up of text - Extract risk factor keywords from disaster documents directly
	Methodology	Insufficient research on methodologies using disaster documents	Suggestion of direction for analysis methodology of disaster documents - **(Unsupervised learning)** Presenting a methodology to explore the relationship of risk factors - **(Supervised learning)** Presenting a methodology to identify the risk of risk factors
Practical contribution	Industrial site	Identification of similar disasters in the same industry by using disaster statistical analysis prepared in the form of a report	Provide tools for effective safety management for on-site safety managers and supervisors - **(Unsupervised learning)** Exploration of new risk factors related to discovered risk factors - **(Supervised learning)** Prepare for a high-risk disaster through safety management even though it appears to be an injury disaster

5.2 연구의 제한점 및 추후 연구

본 연구는 서술형으로 작성된 재해 발생 개요를 기계학습으로 분석하는 방법론을 제시하였다. 이는 기계학습에서 자주 사용되는 기초적인 방법론을 활용하였으며, 건설업에서 발생한 한정적인 재해문서로 연구를 수행하였다. 효과적인 안전관리 방안을 제시하기 위해선 향후 더 개선되어야 할 필요가 있다.

먼저, 본 논문에서 수행한 텍스트 분석과 기계학습은 건설업과 사망재해에 대한 분석을 한정된 데이터로 수행하였다. 분석하는 데이터를 추가하여, 건설업뿐만 아니라 전체 업종에 대한 재해 키워드 분석을 통해 다양한 업종에서의 발생하는 위험요인을 탐색할 수 있으며, 데이터의 불균형에서 오는 기계학습의 오류를 줄일 수 있을 것으로 사료된다.

다음으로, 비지도학습 방법인 SOM을 활용하여 작성한 위험요인지도의 크기와 클러스터의 구분에 대한 검증을 위해 다른 군집화 방법이 동시에 수행되어야 할 필요가 있다. 대표적으로 k-means 클러스터링과 같은 데이터마이닝 군집화 방법론을 활용하여 이를 보완하는 과정이 요구된다.

또한, 재해 요인 키워드의 유형화에 대한 구체적인 유형 설명이 부족하였다. 이는 재해문서 내에서 키워드들을 구체적으로 유형화하지 않은 근본적인 문제에서 기인하며, 실제로 수집되는 산업재해조사표에 포함되어있는 재해 개요를 살펴보면 작성자에 따라서 분석하기에 부적합하게 작성되고 있다. 현재의 산업재해조사표는 작성자의 단어 선택과 작성 방법에 따라서 다르게 분석될 수 있다. 따라서, 근본적인 해결방안으로 재해개요의 작성법을 구체적으로 제시하여 서술형으로 작성되는 재해개요를 유형화할 필요가 있다. 이를 위하여, 산업현장에서 발생하는 위험요인 키워드를 유형화한 키워드 사전(dictionary)을 구축하여, 텍스트 분석에 활용할 수 있는 방안이

필요하다. 또한, 현재 재해개요 데이터에 대한 신뢰성을 확보하기 위하여, 산업안전보건공단에서 직접 조사하여 공시하고 있는 중대재해분석를 수집하여 본 연구와 비교·분석할 필요가 있다.

지도학습의 경우는, 새롭게 발생하는 재해 데이터를 추가적으로 분류하기 위해서 2017년 이후의 데이터에 대한 분석이 필요할 것으로 사료되며, 다른 업종의 위험요인을 탐색하기 위해서 다양한 업종의 재해 데이터를 분석할 필요가 있다. 또한, 재해문서 자체의 편향으로 인해 데이터의 불균형이 발생하였으며, 데이터 수집 단계와 데이터 분할단계에서 이를 해결할 수 있는 방법에 대한 연구가 필요하다.

참 고 문 헌

1) H. W. Heinrich, “Industrial Accident Prevention: A Scientific Approach”, McGraw-Hill Book Company, 1931.

2) Korea Occupational Safety & Health Agency, “Statistical Survey and Analysis of Industrial Disasters”, 2019.

3) H. Park, “Trend Analysis of Korea Papers in the Fields of ‘Artificail Intelligence’, ‘Machine Learning’ and ‘Deep Learning’”, Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 13, No. 4, pp. 283-292, 2020.

4) C. Park, D. Seong, and K. Lee, “Automatic IPC Classification for Patent Documents using Machine Learning”, The Journal of Korean Institute of Information Technology, Vol. 10, No. 4, pp. 119-128, 2012.

5) S. Jun, “A Big Data Preprocessing using Statistical Text Mining”, Journal of Korean Institute of Intelligent Systems, Vol. 25, No. 5, pp. 470-476, 2015.

6) Ministry of Employment and Labor, “Classification of Causes of Deaths in Industrial Accidents”, 2020.

7) Korean Statistical Information Service, “Industrial Accident Status by Industry”, http://www.index.go.kr.

8) G. Salton and M. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, NY, 1983.

9) R. Feldman and I. Dagan, “Knowledge Discovery in Textual

Databases (KDT)", Proceedings of the First International Conference on Knowledge Discovery and Data Mining, pp. 112-117, 1995.

10) C. D. Manning and H. Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999.

11) A. Aizawa, "An Information-theoretic Perspective of Tf-idf Measures", Information Processing and Management, Vol. 39, No. 1, pp. 45-65, 2003.

12) BS. Kim, S. Chang, and Y. Suh, "Text Analytics for Classifying Types of Accident Occurrence Using Accident Report Documents", Journal of the Korean Society of Safety, Vol. 33, No. 3, pp. 58-64, 2018.

13) K. Park and H. Kim, "Analysis of Seasonal Importance of Construction Hazards Using Text Mining", Journal of Civil and Environmental Engineering Research, Vol. 41, No. 3, pp. 305-316, 2021.

14) W. Jang and Y. Suh, "Identifying Abnormal Accidents Using Local Outlier Factor and Decision Tree Algorithms", Journal of the Korean Institute of Industrial Engineers, Vol. 45, No. 4, pp. 329-340, 2019.

15) S. Lee, "A Study on the Trends of Construction Safety Accident in Unstructured Text Using Topic Modeling", Journal of the Korea Academia-Industrial cooperation Society, Vol. 19, No. 10, pp. 176-182, 2018.

16) D. E. Brown, "Text Mining the Contributors to Rail Accidents", IEEE Transactions on Intelligent Transportation System, Vol. 17, No. 2, pp. 346-355, 2016.

17) N. XU, L. MA, Q. Liu, L. WANG, and Y. Deng, “An improved text mining approach to extract safety risk factors from construction accident reports”, Safety Science, Vol. 138, 105216, 2021.

18) S. Shi, D. Zhang, Y. Su, M. Zhang, M. Sun, ans H. Yao, “Risk Factors Analysis Modeling for Ship Collision Accident in Inland River Based on Text Mining”, The 5th International Conference on Transportation Information and Safety, 2019.

19) R. Weber, H. J. Schek, and S. Blott, "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces", International Conference on Very Large Data Bases, pp. 194-205, 1998.

20) O. Egecioglu, "Parametric Approximation Algorithms for High-Dimensional Euclidean Similarity", European Conference on Principles of Data Mining and Knowledge Discovery, pp. 79-90, 2001.

21) S. Jeong, S. Kim, and B. Choi, “An Effective Method for Dimensionality Reduction in High-Dimensional Space”, The Institute of Electronics Engineers of Korea, Vol. 43, No. 4, pp. 88-102, 2006.

22) S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis”, Chemometrics and Intelligent Laboratory Systems, Vol. 2, No. 1-3, pp. 37-52, 1987.

23) T. K. Moon and W. C. Stirling, “Mathematical Methods and Algorithms for Signal Processing”, Prentice-Hall, 2000.

24) K. V. R. Kanth, D. Agrawal, and A. Singh, "Dimensionality Reduction for Similarity Searching in Dynamic Databases"

International Conference on Management of Data, pp. 166-176, 1998.

25) R. C. Sharma, K. Hara, and H. Hirayama, “A Machine Learning and Cross-Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multitemporal Data“, Hindawi Scientifica, Vol. 2017, Article ID 9806479, pp. 8, 2017.

26) C. A. Ramezan, T. A. Warner, and A. E. Maxwell, “Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification”, Remote Sensing, Vol. 11, No. 2, pp. 185, 2019.

27) F. Y. M. Ahmed, Y. H. Ali, and S. M. Shamsuddin, “Using K-Fold Cross Validation Proposed Models for Spikeprop Learning Enhancements”, International Journal of Engineering & Technology, Vol. 7, No. 4, pp. 145-151, 2018.

28) M. Leblanc and R. Tibshirani, “Combining Estimates in Regression and Classification”, Journal of the American Statistical Association, Vol. 91, No. 436, pp. 1641-1650, 1994.

29) Y. Kodratoff, R. S. Michaiski, R. S. Michaiski, J. G. Carbonell, and T. M. Mitchell, “Machine Learning: An Artificial Intelligence Approach”, Morgan Kaufmann, Vol. 2, 1986.

30) B. Jang, “Next-Generation Machine Learning Technology”, Communications of the Korean Institute of Information Scientists and Engineers, Vol. 25, No. 3, pp. 96-107, 2007.

31) S. Mun, S. Jang, J. Lee, and J. Lee, “Technology Trends in Machine Learning and Deep Learning”, Information and

Communications Magazine, Vol. 33, No. 10, pp. 49-56, 2016.

32) H. Trevor, R. Tibshirani, and J. Friedman, "Unsupervised learning. The elements of statistical learning“, Springer, New York, pp. 485-585, 2009.

33) K. Son and H. Ryu, “Association Rules Analysis of Safe Accidents Caused by Falling Objects”, Journal of the Korea Institute of Building Construction, Vol. 19, No. 4, pp. 341-350, 2019.

34) M. Shin and Y. Suh, “Development of MSDS Map for Visual Safety Management of Hazardous and Chemical Materials”, Journal of the Korean Society of Safety, Vol. 34, No. 2, pp. 48-55, 2019.

35) M. Shujiro, O. Koji, S. Junko, and I. Hiroaki, “Data Mining for Hazard Elimination through Text Information in Accident Report”, Asia Pacific Management Review, Vol. 16, No. 1, pp. 65-81, 2011.

36) F. Palamara, F. Piglione, and N. Piccinini, “Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases”, Safety Science, No. 49, No. 8, pp. 1215-1230, 2011

37) A. Chokor, H. Naganathan, W. K.Chong, and M. E. Asmar, “Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning”, Procedia Engineering, Vol. 145, pp. 1588-1593, 2016.

38) A. Verma and J. Maiti, “Text-document clustering-based cause and effect analysis methodology for steel plant incident data”, International Journal of Injury Control and Safety Promotion , Vol. 25, No. 4, pp. 416-426, 2018.

39) Thomas Hofmann, “Unsupervised learning by probabilistic latent

semantic analysis." Machine learning, Vol. 42, pp. 177-196, 2001.

40) X. Lin, D. Soergel and G. Marchionini, "A Self-Organising Semantic Map for Information Retrieval", Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Evelopment in Information Retrieval, pp. 262-269, 1991.

41) T. Kohonen, "The Self-Organizing Map", Neurocomputing, Vol. 21, No. 1-3, pp. 1-6, 1998.

42) T. Kohonen, "Self-Organization and Associative Memory", Springer-Verlag, Berlin, Heidelberg, 2012.

43) T. Kohonen, "Essentials of the Self-Organizing Map", Neural Networks, Vol. 37, pp. 52-65, 2013.

44) J. Vesanto and E. Alhoniemi, "Clustering of the Self-Organizing Map", IEEE Transactions on Neural Network, Vol.11, No.3, pp. 586-600, 2000.

45) A. Likas, N. Vlassis, and J. J. Verbeek, "The Global k-means Clustering Algorithm", Pattern Recognition, Vol. 36, No. 2, pp. 451-461, 2003.

46) J. Kim and J. Sung, "A Text Classification System based on a Supervised Learning Algorithm", International Conference Digital Library & Knowledge, pp. 421-430, 1998.

47) Y. Kim, W. S. Yoo, and Y. Shin, "Application of Artificial Neural Networks to Prediction of Construction Safety Accidents", J. Korean Soc. Hazard Mitig., Vol. 17, No. 1, pp. 7-14, 2017.

48) B. U. Ayhan and O. B. Tokdemir, "Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial

Neural Networks", Journal of Construction Engineering and Management, Vol. 146. No. 3, 2020.

49) Y. M. Goh and C. U. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques", Accident Analysis & Prevention, Vol. 108, pp. 122-130, 2017.

50) M. Cheng, D. Kusoemo, and R. Antoni Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning", Automation in Construction, Vol. 118, 103265, 2020.

51) B. Zhong, X. Pan, P. E.D. Love, J. Sun, and C. Tao, "Hazard analysis: A deep learning and text mining framework for accident prevention", Advanced Engineering Informatics, Vol. 46, 101152, 2020.

52) P. Sankarasubramanian, "Industrial Accident Report Analysis Using Natural Language Processing", International Journal of Scientific & Technology Research, Vol. 9, No. 6, pp. 470-475, 2020.

53) S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models", Journal of Biomedical Informatics, Vol. 35, pp. 352-359, 2002.

54) Y. Song and Y. LU, "Decision tree methods: applications for classification and prediction", Shanghai Arch Psychiatry, Vol. 25, No. 2, pp. 130-135, 2015.

55) J. P. Bigus, "Data Mining with Neural Networks", Macgraw-Hill, pp. 61-97, 1996.

56) R. Feraud and F Clerot, "A methodology to explain neural network classification", Neural Networks, Vol. 15, No. 2, pp. 237-246, 2002.

57) A. Mathur and G. M. Foody, “Multiclass and Binary SVM Classification: Implications for Training and Classification Users”, IEEE Geoscience and Remote sensing Letters, Vol. 5, No. 2, pp. 241-245, 2008.

58) A. Karatzoglou, D. Meyer, and K. Hornik, “Support Vector Machines in R”, Journal of Statistical Software, Vol. 15, No. 9, pp. 1-9, 2006.

59) C. Cortes and V. Vapnik, “Support-Vector Networks”, Machine Learning, Vol. 20, pp. 273-297. 1995.

60) J. Han, J. Pei, and M. Kamber, “Data Mining: Concepts and Techniques”, Elsevier, 2011.

61) S. Visa, B. Ramsay, A. Ralescu, and E. van der Knaap, “Confusion Matrix-based Feature Selection”, Proceedings of the Twenty-second Midwest Artificial Intelligence and Cognitive Science Conference, pp. 120-127, 2011.

62) M. Junker, R. Hoch, and A. Dengel, “On the Evaluation of Document analysis Components by Recall, Precision, and Accuracy”, International Conference on Document Analysis and Recognition, 1999.

부록

A. Keyword conversion table

영어-한글 키워드 변환표

영문	한글	영문	한글
apartment	아파트	move	이동
balance	중심	narrow	협착
basement	지하	outside	외부
board	탑승	paint	도장
bury	매몰	panel	판넬
check	확인	pillar	기둥
collapse	붕괴	pipe	파이프
concrete	콘크리트	piping	배관
connect	연결	rebar	철근
crain	크레인	removal	철거
disjoint	해체	repair	보수
drive	운전	reverse	후진
electric	감전	rollover	전복
equipment	장비	roof	지붕
excavation	굴착	rooftop	옥상
excavator	굴삭기	rope	로프
external	외벽	salvage	인양
fall	추락	scaffold	비계
fix	고정	soil	토사
floor	바닥	steel	철골
foothold	발판	tower	타워
forkcrane	포크레인	transport	운반
fracture	골절	tree	나무
glass	유리	tunnel	터널
head	머리	upper	상부
height	고소	vehicle	차량
install	설치	volt	볼트
ladder	사다리	weld	용접
materials	자재	wire	와이어
miss	실족	workbench	작업대
mold	거푸집	see	바다

B. Analysis code

Analysis code for SOM using MATLAB

```
   sM   = som_set('som_map');
   sD   = som_set('som_data');
   sTo  = som_set('som_topol');
   sTr  = som_set('som_train');
   sN   = som_set('som_norm');

D=som_read_data('cons_kosha.txt')       % fix 'D-labels'

msize = [10, 8]

sMap=som_make(D,'msize',msize)

%sMap = som_randinit(D);
%sMap = som_lininit(D);
%sTopol = som_topol_struct('data',D);
%sTrain = som_train_struct('phase','train','data',D,'map',sMap);

sMap=som_autolabel(sMap,D,'vote');
%sMap=som_autolabel(sMap,D);

sS = som_show(sMap,'umat','all');
%som_show(sMap1,'comp','all');
som_show_add('label',sMap.labels,'TextSize',11,'TextColor','r')
```

Analysis code for TF-IDF using program R

```
#-----------------------TF-IDF----------------------------#

Narr <- read.csv("Narr.csv", header = TRUE)

Corpus.Narr <- Corpus(DataframeSource(Narr))

data <- unlist(Corpus.Narr)
data.Noun <- extractNoun(data)

Corpus.data <- VCorpus(VectorSource(data.Noun))

dtm <- DocumentTermMatrix(Corpus.data,
                          control=list (tokenize=words,
                                        removeNumbers=T,
                                        removePunctuation=T,
                                        wordLengths=c(2,10),
                                        stopwords=c(stop_K),
                                        weighting=weightTfIdf)
)

matrix.tfidf <- as.matrix(dtm)

mt.count <- colSums(matrix.tfidf)
mt.order <- order(mt.count, decreasing=T)

freq.mt <- matrix.tfidf[,mt.order[1:100]]

write.csv(freq.mt,"TFIDF.csv")   #need preprocessing data (+ empty data)#
```

Analysis code for principal component analysis using program R

```
#------------PCA ---------------#

pca <- read.csv("TFIDF.csv", header=TRUE)

pca.tfidf <- prcomp(pca,center=T, scale=TRUE)


screeplot(pca.tfidf, main = "", col = "red", type = "lines", pch = 1, npcs = length(pca.tfidf$sdev))

summary.model <- summary(pca.tfidf)
summary.model


biplot.model <- biplot(pca.tfidf)
biplot.model

pca.key <- pca.tfidf$rotation
pca.matrix <- as.matrix(pca.tfidf$x)

write.csv(pca.matrix, "pca.matrix.csv")
write.csv(pca.key, "pca.key.csv")
```

Analysis code for k-fold cross validation using program R

```
#--------Partition / k-fold cross-validation------ #

data.pca <- read.csv("pca.matrix.csv")

k=5

data.pca$id <- sample(1:k, nrow(data.pca), replace = TRUE)
list <- 1:k

prediction <- data.frame()
testsetCopy <- data.frame()

progress.bar <- create_progress_bar("text")
progress.bar$init(k)

for(i in 1:k){ trainingset <- subset(data.pca, id %in% list[-i])
testset <- subset(data.pca, id %in% c(i))
}

write.csv(trainingset, "trainingset.csv")
write.csv(testset, "testset.csv")
```

Analysis code for logistic regression using program R

```
#-----Logistic Regression ------------#


r.train <- read.csv("trainingset.csv")
r.test <- read.csv("testset.csv")

glm.fit <- glm(CS~., family = "binomial", data = r.train)
sum.glm <- summary(glm.fit)
summary(glm.fit)

r.anova <- anova(glm.fit, test="Chisq")
r.anova

glm.test <- predict(glm.fit, newdata = r.test, type = "response")

round(glm.test)

r.table <- table(round(glm.test),r.test$CS)

confusionMatrix(table(round(glm.test), r.test$CS))

plot(glm.test, main = "Regression")
R.model <- write.csv(glm.fit$model, "r.model.csv")

write.csv(glm.test, "reg_form.csv")
```

Analysis code for decision tree analysis using program R

```
#------------------------------DT --------------#

train <- read.csv("trainingset2.csv")
str(train)
df.train <- train[,-1]
table(df.train$CS)

test <- read.csv("testset.csv")
str(test)
df.test <- test[,-1]
table(df.test$CS)

tree_model <- rpart(CS~.,df.train, method="class")
tree_model$cptable

plot(tree_model)


opt.cp <- tree_model$cptable[which.min(tree_model$cptable[,"xerror"]),"CP"]
tree_model_prune <- prune(tree_model, opt.cp)
plot(tree_model_prune)
text(tree_model_prune)

r_predict <- predict(tree_model_prune, df.test, type="class", na.action = na.pass)
write.csv(r_predict,"dt_predict.csv")

confusionMatrix(table(r_predict, test$CS))
```

Analysis code for neural network analysis using program R

```
#----------------------Neural network --

nn.model <- neuralnet(CS~., data = r.train, hidden = c(20,5))

plot(nn.model)

nn.predict <- predict(nn.model, r.test)

confusionMatrix(table(round(nn.predict),r.test$CS))
```

Analysis code for support vector machine using program R

```
#----------------------Support vection machine ---------------------------------

result  <- tune.svm(CS~., data=r.train, gamma = 2^(-5:0), cost = 2^(0:4), kernel="radial")

result$best.parameters


scm <- svm(CS~., data = r.train, gamma=0.03125, cost=4)
summary(scm)

svm.test <- predict(scm,r.test)
plot(x=r.test$CS, y=svm.test, main ="SVM")


mse <- mean((svm.test - r.test$CS)^2)
mse

confusionMatrix(table(round(svm.test), r.test$CS))

write.csv(svm.test,"svm_predict.csv")
```