Journal of the Korean Society of Safety, Vol. 32, No. 5, pp. 179-185, October 2017
 https://doi.org/10.14346/JKOSOS.2017.32.5.179

효과적인 산업재해 분석을 위한 텍스트마이닝 기반의 사고 분류 모형과 온톨로지 개발

안길승 · 서민지 · 허선†

한양대학교 산업경영공학과

(2017. 8. 9. 접수 / 2017. 9. 5. 수정 / 2017. 9. 6. 채택)

Development of Accident Classification Model and Ontology for Effective Industrial Accident Analysis based on Textmining

Gilseung Ahn · Minji Seo · Sun Hur†

Department of Industrial and Management Engineering, Hanyang University

(Received August 9, 2017 / Revised September 5, 2017 / Accepted September 6, 2017)

Abstract : Accident analysis is an essential process to make basic data for accident prevention. Most researches depend on survey data and accident statistics to analyze accidents, but these kinds of data are not sufficient for systematic and detailed analysis. We, in this paper, propose an accident classification model that extracts task type, original cause materials, accident type, and the number of deaths from accident reports. The classification model is a support vector machine (SVM) with word occurrence features, and these features are selected based on mutual information. Experiment shows that the proposed model can extract task type, original cause materials, accident type, and the number of deaths with almost 100% accuracy. We also develop an accident ontology to express the information extracted by the classification model. Finally, we illustrate how the proposed classification model and ontology effectively works for the accident analysis. The classification model and ontology are expected to effectively analyze various accidents.

Key Words : industrial accident analysis, text classification, ontology, support vector machine, mutual information

1. 서 론

산업재해는 근로자가 업무와 관련 있는 건설물, 설비, 원재료, 가스, 증기, 분진 등에 기인하여 발생하는 재해를 말한다. 우리나라의 산업재해자 수는 2016년 기준 82,780명으로, 매년 비슷한 수의 산업재해가 발생한다[1].

재해 분석은 과거에 발생한 재해의 원인과 특성을 분석하여 재해를 예방하기 위한 기초자료를 만드는 과정이다[2]. 즉, 재해 분석은 산업재해를 예방하는 데에 필수적이다. 이에 많은 연구자들이 다양한 형태의 재해를 분석하였다. 김정민 등은 2012년부터 2014년까지 국내 건설공사에서 발생한 재해를 작업 유형과 작업자의 인구 통계학적 특성을 기준으로 분석하였다[3]. 정옥남 등은 승강기와 에스컬레이터 관련 사고의 발생 원인을 이용자 과실, 관리 부실, 제조불량 등으로 구분하여, 원인별로 사고예방 대책을 수립하였다[4]. 윤유성과 권오헌은 제조업종에서 평균 재해율이 높은 재해를 분석하여, 자주 발생하는 재해 형태가 끼임(33.1%), 넘어짐(11.5%), 업무상질병(10.7%), 부딪힘(10.0%)이라는 사실을 파악하였다[5].

산업재해를 분석한 기존 연구 대다수는 설문 데이터와 사고 통계를 바탕으로 사망자 수, 사고 원인, 기인물 등을 단편적으로 분석하는 데에 그쳤다. 다시 말해, 기존 연구는 주로 어떤 유형의 재해가 많이 발생하는지, 사망자를 많이 내는 사고가 무엇인지 등을 분석하였다. 그러나 산업재해를 예방하려면 각각의 재해를 더욱 체계적이고 정밀하게 분석해야 하므로, 설문 데이터나 사고 통계는 분석 대상으로 부적합하다. 본 연구에서는 다방면에서 사고 상황을 분석해놓은 자료인

† Corresponding Author : Sun Hur, Tel : +82-31-400-5265, E-mail : hursun@hanyang.ac.kr
Department of Industrial and Management Engineering, Hanyang University, 55 Hanyang Daehak-ro, Sangrok-gu, Ansan 15588, Korea

※ 고용노동부, 한국산업안전보건공단에서 주최하고 한국안전학회에서 주관한 제1회 산업안전·보건 분야 대학(원)생 논문 공모전 수상작입니다.

사고 보고서를 바탕으로 재해를 예방하기 위한 기초자료를 생성한다. 하지만 설문 데이터나 사고 통계와는 다르게, 사고 보고서에서 원하는 정보를 획득하려면 보고서 내용 대부분을 직접 읽고 정리해야 한다. 더욱이, 보고서를 읽고 정리하는 과정에서 전문가의 주관이나 실수가 개입될 가능성이 있다.

그러므로 본 연구는 텍스트 마이닝을 바탕으로 사고 보고서의 개요에서 작업 유형, 기인물(original cause materials), 사고 유형, 사망자 수를 자동으로 획득하는 사고 분류 모형을 제안한다. 제안하는 모형은 사고 보고서의 개요가 단어 가방 형태(bag-of-words)로 입력되었을 때, 작업 유형, 기인물, 사고 유형, 사망자 수를 출력하는 서포트 벡터 머신(support vector machine, SVM)이다. 더 나아가, 사고 분류 모형을 바탕으로 획득한 작업 유형, 기인물, 사고 유형, 사망자 수를 효과적으로 저장하고 표현하는 온톨로지를 개발한다.

온톨로지란 사람이 현실 세계에서 인식하는 객체(entity) 간 관계를 명시하는 일종의 지식 명세서로[6], 다양한 분야에서 활용된다. 예를 들어, Yang 등은 제품 구성 지식(production configuration knowledge)을 표현하는 온톨로지를 개발하였다[7]. 이 온톨로지는 부품과 제품의 관계, 부품 간 관계, 부품의 특성 등을 포함한다. 다른 예로, García-Peñalvo 등은 소프트웨어 개발 프로젝트를 온톨로지로 모델링하여 소프트웨어 개발 프로젝트 이해 관계자 간 원활한 커뮤니케이션을 가능하게 하였다[8].

본 논문은 다음과 같이 구성된다. 제2장에서는 본 연구에서 개발한 사고 분류 온톨로지를 소개한다. 제3장에서는 산업재해 분석에 사용한 사고보고서 데이터와 그 활용 방안을 자세히 설명한다. 제4장에서는 사고 분류 온톨로지와 수집한 데이터를 바탕으로 사고 분류 모형을 수립하고 평가한다. 제5장에서는 개발한 온톨로지와 사고 분류 모형을 이용하여 산업재해를 분석한 결과를 제시한다. 마지막으로 제6장에서는 결론을 정리한다.

2. 사고 분류 온톨로지

사고 분류 온톨로지를 구성하기 위해, 안전보건공단에서 사용하는 사고 분류 기준을 참고하였다[1]. Table 1은 본 연구에서 개발한 계층 온톨로지에 포함되는 요소를 보여준다. 안전보건공단이 사고 보고서를 산업 분야 별로 구분하여 제공하기에, 산업 분야는 분류 모형을 사용하여 획득하는 것이 아니라 주어진다고 가정한다.

Table 1. Elements of the proposed ontology

Class	Elements
Industry fields (S)	Machinery industry (S_1) Electrical industry (S_2) Chemical industry (S_3) Construction industry (S_4) Shipbuilding industry (S_5)
Task type (T)	Handling, processing of an object (T_1) Connection/assembly/establishment/demolition of the object (T_2) Transportation, cargo handling and operating activities (T_3) Maintenance/installation of equipment and machines (T_4) Cleaning operations and additional work (T_5) Other work (T_6)
Original cause material (O)	Machine equipment (O_1) Components and materials, parts (O_2) Building structures and surfaces (O_3) Chemicals (O_4) Transport (O_5) Humans, animals and plants (O_6) Working environment, air conditions, natural phenomena (O_7) Others (O_8) Nit (O_9)
Type of accident (A)	Spill (A_1) Fall beneath (A_2) Bumped/striking (A_3) Hit (A_4) Trap (A_5) Collapse (A_6) Pressure, shock (A_7) Exposure for harmful and hazardous materials and environment (A_8) Explosion/Fire (A_9) Electrical shock (A_{10}) Others (A_{11})
The death toll (D)	No deaths (D_0) 1 death (D_1) 2 or more deaths (D_2)

Fig. 1은 Table 1에 제시된 요소를 바탕으로 개발한 사고 분류 온톨로지의 구조를 보여준다. Fig. 1에서 보듯이, 이 온톨로지는 사고가 일어난 산업 분야가 무엇인지, 사고자가 하고 있던 작업이 무엇이었는지, 사고를 일으킨 기인물이 무엇이었는지, 그 사고로 인해 몇 명이 사망하였는지를 저장한다. 개발한 온톨로지는 5장에서 사고를 분석하는 토대가 된다. 예를 들어, 기계 분야에서 물체의 가공, 취급 작업을 하다가 설비나 기계에 부딪혀 두 명 이상 사망한 사고가 몇 건 인지를 파악하는 데에 이 구조를 활용한다.

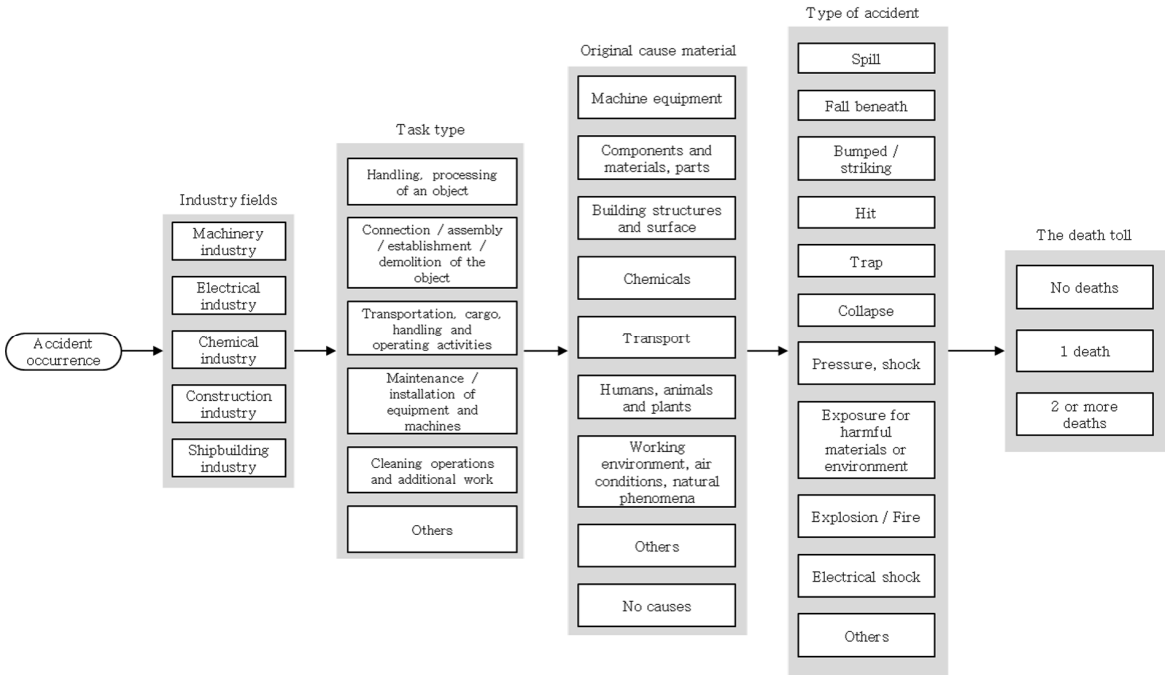


Fig. 1. Structure of the proposed accident ontology.

3. 사고 보고서 데이터

2007년 4월부터 2017년 4월까지 안전보건공단에 등록된 1,655건의 국내 재해사례 개요 문서를 수집하였다[9]. Table 2는 수집한 문서의 산업 분야 별 분포와 그 예시를 보여준다.

수집한 사고 보고서 데이터는 사고 분류 모형을 학습하고 평가하여, 산업재해를 분석하는 데에 사용한다. 데이터의 구체적인 활용 과정은 크게 네 단계로 구분한다. 첫 번째 단계에서는 1,655개 문서 집합을 각각 1 000개와 655개의 문서 집합으로 나눈다. 여기서 1,000개 문서 집합에 포함된 모든 문서는 모형 학습과 평가에 사용하기 위해, 여러 기술 문서를 참고하여 작업 유형, 사고 유형, 사망자 수를 수작업으로 라벨링(labeling)한다. Table 3은 라벨링한 예시 문서를 보여준다. 두 번째 단계에서는 라벨링한 1,000개 문서로 SVM 모형을 학습하고 평가한다.

Table 2. Distribution and example of the collected documents

Industry Sector	Number of accidents	Example document (in Korean)	Example document (in English)
S_1	448	17년 2월 충북 음성군 소재 롤러 생산 작업장에서 범용 선반에 환봉을 장착하여 구동시킨 후, 사포의 양끝을 손으로 잡고 환봉의 표면을 연마 하던 중, 면장갑이 환봉 표면에 감기면서 선반의 척에 머리를 부딪쳐 사망	In Feb. 2017, a person mounted a round bar on a universal shelf and drove it in a roller production workshop located in Eumseong, Chungcheongbuk-do Province. After that, the victim was hit by the lathe chuck and dead, while polishing the surface of the round bar by holding both ends of the sandpaper by hand.
S_2	41	진공교반기를 고압세척기를 이용하여 세척하던 중 누전된 고압세척기 세척건에 접촉되어 감전 사망함	While the vacuum stirrer was being cleaned using a high-pressure washer, it was contacted with a short-circuited high-pressure washer-cleaning gun, resulting in an electric shock.
S_3	54	승합차 해체 작업장에서 피재자가 자동차에서 분리된 폐부품을 담을 용기를 만들기 위해 산소절단기로 폐드럼을 용단하던 중 드럼 내 잔류되어 있던 인화성물질이 폭발하여 사망	At the dismantling of the van, the victim was burning the waste drum with an oxygen cutter to make a container for the waste parts separated from the car. The remaining flammaable material in the drum exploded and the victim died.
S_4	911	건물 지붕 보수 작업 중, 지붕재로 사용된 선라이트(sunlight)가 파손되면서 지상 1층 콘크리트 바닥으로 추락(높이 약 7미터)하여 2명 사망, 1명 부상.	During the repair work on the building roof, the sunlight used as the roofing material was broken, and the building crashed to the first floor concrete floor (about 7 meters high), resulting in two deaths and one injury.
S_5	201	데크 상부에서 재해자가 탑재된 블록 용접용 작업발판을 설치하는 과정에서 몸의 균형을 잃고 11.7m 하부 바닥으로 떨어져 사망한 사고임	In the process of installing a work platform for block welding with a disaster on the top of the deck, the victim lost his or her balance and fell to the bottom of 11.7m and died.

Table 3. Example labeled documents

Example document (in Korean)	Example document (in English)	T-Label	O-Label	A-Label	D-Label
합성수지 제조업 공장에서 원재료 창고에 적재되어 있던 톤백이 무너지면서 아래에 있던 재해자가 깔려 사망	In the synthetic resin manufacturing factory, the back bag that was loaded in the raw material warehouse collapsed, and the victor under the back bag was crushed to death	T_1	O_2	A_3	D_1
건물 지붕 보수 작업 중, 지붕재로 사용된 선라이트(sunlight)가 파손되면서 지상 1층 콘크리트 바닥으로 추락(높이 약 7미터)하여 2명 사망, 1명 부상.	During the repair work on the building roof, the sunlight used as the roofing material was broken, and the building crashed to the first floor concrete floor (about 7 meters high), resulting in two deaths and one injury.	T_4	O_8	A_1	D_2

모형을 평가하기 위해, 5겹 교차 검증(5-fold cross validation)을 활용한다. 즉, 1,000개 문서 집합을 다섯 개 묶음으로 나누어 네 개 묶음으로 모형을 학습하고 나머지 한 개 묶음으로 모형을 평가하는 과정을 다섯 번 반복한다. 세 번째 단계에서는 학습된 모형을 바탕으로 나머지 655개 문서의 작업 유형, 사고 유형, 사망자 수를 추론한다. 마지막 단계에서는 작업 유형, 사고 유형, 사망자 수를 추론한 655개 문서와 미리 라벨링한 1,000개 문서를 합하여 사고 분석을 수행한다.

4. 사고 분류 모형 개발

제안하는 사고 분류 모형을 개발하는 과정은 크게 특징 추출, 특징 선택, SVM 학습이라는 세 단계로 구분한다. 특징 추출 단계에서는 사고 보고서에 출현한 모든 단어를 추출하여, 단어의 출현 여부를 특징으로 생성한다. 단어의 출현 여부는 텍스트를 분류하는 데에 효과적이고 계산량이 적다[10]. 사고 보고서에 등장한 단어를 추출하기 위해, 파이썬(python) 기반의 한국어 형태소 분석기인 KoNLPy[11]를 사용한다. 중복을 제외하고, 모든 사고 보고서에서 추출한 단어 개수는 1,856개다.

특징 추출 단계에서 추출한 모든 특징이 사고를 분류하는데 효과적이지는 않다. 심지어 몇몇 특징을 투입하면 분류 성능이 떨어지기도 한다. 따라서 추출한 특징 중 사고 분류에 효과적인 특징을 선택해야 한다.

특징 선택 단계에서는 상호 정보량(mutual information)을 이용하여 사고 분류에 효과적인 특징을 선택한다. 상호 정보량은 식 (1)과 같이 계산한다.

$$MI(w_i) = \sum_{k=1}^{l} \Pr(w_i, C_k) \times \log_2 \frac{\Pr(w_i, C_k)}{\Pr(w_i) \times \Pr(C_k)}, \quad (1)$$

여기서 w_i는 단어 i를, $\Pr(w_i)$은 w_i가 임의의 문서에 출현할 확률을 나타낸다. C_k는 클래스 k를, l은 클래스 개수를 나타낸다. 상호 정보량이 클수록 클래스를 분류하는데 효과적이다. 따라서 상호 정보량이 높은 특징을 선택해야 한다. 그러나 개별 단어의 출현 여부만으로는 사고를 분류하기 어려운 경우가 잦다. 예를 들어, "옥상층 슬래브 콘크리트 타설 중 시스템동바리가 붕괴되면서 콘크리트 타설 작업자 4명이 추락하여 사망1명, 부상 3명." 이라는 사고 개요가 있다. 여기에서 사고 유형을 "붕괴"와 "추락"이라는 단어의 출현 여부만 가지고 분류한다고 하자. 주로 "붕괴"가 출현하면 사고 유형은 A_2(넘어짐/깔림)로, "추락"이 출현하면 A_1(떨어짐)으로 분류한다. 그러나 이와 같이 두 단어가 동시에 출현하면, "붕괴"가 출현할 확률, "추락"이 출현할 확률, "붕괴"가 출현했을 때 실제 사고 라벨이 A_2인 확률 등을 고려하여 사고 유형을 분류한다. 즉, 대부분 분류기는 이들 확률을 독립적으로 추정한다. 그러나 이들은 독립적이지 않다. 즉, "붕괴"가 출현했을 때 사고 라벨이 A_2일 확률은 높지만, "붕괴"와 "추락"이 동시 출현했을 때 사고 라벨이 A_2일 확률은 높지 않다. 다시 말해, "붕괴"만 등장한 사고 보고서는 주로 깔리는 사고와 관련이 있으나, "붕괴"와 "추락"이 동시에 등장한 사고 보고서는 다리나 지반 등이 붕괴하여 추락하는 사고와 관련이 있다.

따라서 본 연구에서는 한 단어의 출현 여부를 나타내는 특징 뿐 아니라 두 단어의 동시 출현 여부를 나타내는 특징도 고려한다. 이와 같은 방법으로 고려하는 클래스 별 특징 개수와 예시는 Table 4와 같다. Table 4의 예시에서, &로 엮인 두 단어는, 두 단어가 동시에 출현하는 지 여부를 고려하는 특징이다. 예를 들어, "덤프트럭&점검" 특징은 한 문서에 "덤프트럭"과 "점검"이 동시에 출현했는지를 나타내는 특징이다. 클래스 별로 고려하는 모든 후보 특징의 상호 정보량을 계산하여, 상호 정보량이 높은 순서대로 특징을 선택한다.

Table 4. The number of candidate features to classify each class

Class	Number of features	Examples
T	71840	dump truck&heck, drive
O	71944	vibration&furniture, timber&conveyor, shore
A	72655	explosion, accident&argon, chock&line
D	72812	one person&death, one person&fall

SVM 학습 단계에서는 앞서 선택한 특징을 바탕으로 안정적이고 우수한 예측력을 보인다고 알려진 SVM을 학습한다. 커널(kernel)은 텍스트 분류에 효과적인[12] 선형 커널(linear kernel)을 사용한다.

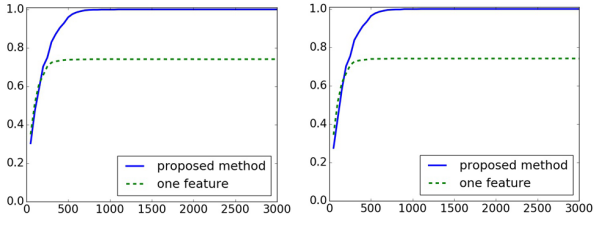
거시적 정확도(macro-accuracy)와 정밀도(macro-precision)는 본 연구에서 고려한 문제와 같이 클래스가 세 개 이상이고 클래스 간 데이터 불균형이 심하지 않을 때 적합한 성능 척도이다. 따라서 본 연구에서는 Fig. 2 - 5에 제시된 바와 같이 거시적 정확도와 정밀도를 이용하여 특징 개수에 따른 성능을 비교 평가하였다.

또한, 앞서 설명한 바와 같이, 특징은 상호 정보량이 높은 순서대로 선택하였다. 예를 들어, 작업 유형을 분류하는 데 500개의 특징을 사용하였다면, 전체 후보 특징 71840개 중 상호 정보량 기준 상위 500개를 사용하였다는 뜻이다. 거시적 정확도와 정밀도는 5-겹 교차검증을 했을 때의 평균을 사용하며, 거시적 분류 정확도와 정밀도는 각각 식 (2)와 (3)과 같이 계산한다.

$$M-acc = \frac{1}{l}\sum_{k=1}^{l}\frac{tp_k+tn_k}{tp_k+fp_k+tn_k+fn_k}, \quad (2)$$

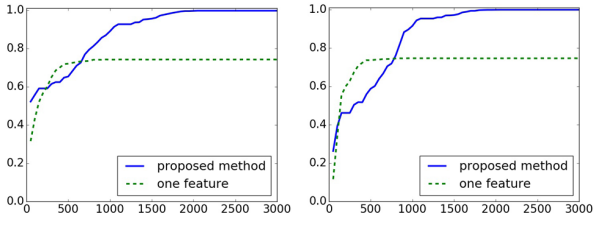
$$M-pre = \frac{1}{l}\sum_{k=1}^{l}\frac{tp_k}{tp_k+fp_k}, \quad (3)$$

여기서 tp_k, tn_k, fp_k, fn_k는 각각 클래스 k에 대한 참 긍정(true positive), 참 부정(true negative), 거짓 긍정(false positive), 거짓 부정(false negative)을 나타낸다.



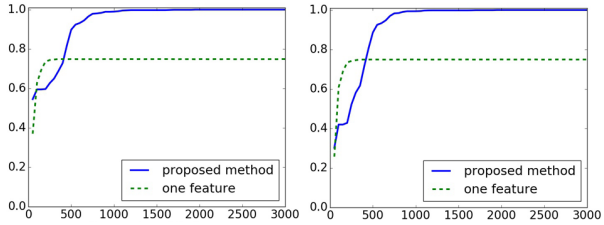
(a) Macro-accuracy (b) Macro-precision

Fig. 2. Macro-accuracy and precision of SVM for work type (T) classification according to the number of features.



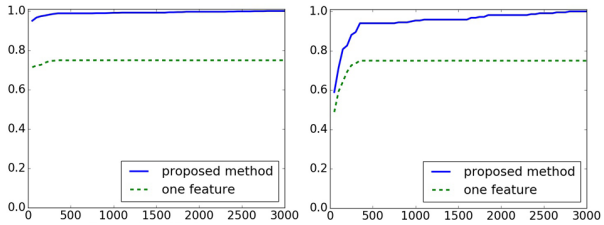
(a) Macro-accuracy (b) Macro-precision

Fig. 3. Macro-accuracy and precision of SVM for original cause materials (O) classification according to the number of features.



(a) Macro-accuracy (b) Macro-precision

Fig. 4. Macro-accuracy and precision of SVM for accident type (A) classification according to the number of features.



(a) Macro-accuracy (b) Macro-precision

Fig. 5. Macro-accuracy and precision of SVM for the number of deaths (D) classification according to the number of features.

위 그림에서 x축은 특징 개수를, y축은 정확도(정밀도)를 나타낸다. 또한, 점선은 한 단어의 출현 여부를 나타내는 후보 특징 집합에서 상호 정보량이 높은 특징을 선택하여 분류했을 때 성능을 나타낸다. 모든 그림에서 두 단어가 동시 출현하는 특징을 고려한 방법이 한 단어의 출현 여부만 고려하는 기존 방법보다 사고 분류를 할 때 우수함을 보였다. 특히, 제안하는 방법은 특징 개수가 증가함에 따라 정확도와 정밀도 모두 1에 수렴하였다. 이는 제안하는 사고 분류 모형이 정확한 산업재해 분석을 할 수 있다는 사실을 보여준다.

5. 사고 분류 모형을 이용한 산업재해 분석

본 장에서는 제안하는 사고 분류 모형을 이용하여 산업재해를 분석한다. 우선, 제안하는 사고 분류 모형을 바탕으로 라벨링하지 않은 655개 문서를 라벨링한다. 앞서 설명한 바와 같이, 제안하는 모형은 특징 개수가 일정 수를 넘어서면 사고 개요를 정확히 라벨링한다. 또한, 특

Table 5. The number of features when the accuracy is 1 for the first time

Class	Number of features
T	1087
O	2872
A	2801
D	1945

징 개수가 많아지면 사고 개요를 분류하는 시간 역시 증가한다. 따라서 655개 문서를 정확히 라벨링하면서 분류 시간을 최소화하기 위해, Fig. 2 - 5에서 분류 정확도가 처음으로 1이 되는 특징 개수를 사용한다. 클래스 별로 분류 정확도가 처음으로 1이 되는 특징 개수는 Table 5와 같다.

개발한 사고 분류 모형으로 655개 문서를 라벨링한 결과는 Table 6과 같다.

Fig. 7은 본 연구에서 제안한 분류 모형과 사고 온톨로지를 바탕으로 1,655건의 사고를 분석한 결과의 일부를 도식화한 것이다. 그림은 다음과 같이 해석한다. 전체 사고 중 911건이 건설업(S_4)에서 발생했고, 그 중 207건이 물체를 연결/조립/설치/해체(T_2)하던 중에 발생했다. 다시 207건 중 97건의 기인물이 건축물/구조물 및 표면이며, 97건 중 31건이 넘어지거나 깔리는 사고였다. 마지막으로 31건 중 28건에서 한 명이 사망했고, 나머지 3건에서 두 명 이상이 사망했다. 분석 경로에 포함되는데도 불구하고 화살표로 이어지지 않는 것은 그 사례가 없다는 것을 의미한다. 예를 들어, A_2에서 D_0로 가는 화살표가 연결되지 않았는데, 이는 31건의 사고 중 사망자가 없는 사고가 없음을 의미한다.

Table 6. Classification result for 655 accident documents

Class	Classification result
T	T_1(116), T_2(136), T_3(113), T_4(136), T_5(49), T_6(105)
O	O_1(310), O_2(45), O_3(206), O_4(6), O_5(29), O_6(6), O_7(2) O_8(50), O_9(1)
A	A_1(288), A_2(88), A_3(30), A_4(93), A_5(73), A_6(4), A_7(2), A_8(7), A_9(19), A_{10}(24), A_{11}(22)
D	D_0(4), D_1(609), D_2(42)

6. 결 론

본 연구에서는 체계적이고 정밀한 재해분석을 하기 위한 방법론으로 텍스트마이닝 기반의 사고 분류 모형과 온톨로지를 개발하였다. 사고 분류 모형은 사고 보고서의 개요를 단어 가방 형태로 입력받아, 작업 유형, 기인물, 사고 유형, 사망자 수를 판단한다. 특징으로 사용된 단어는 상호 정보량을 바탕으로 선택하며, 다른 연구와 차별되는 점은 사고 보고서라는 특성을 고려하여 두 단어의 조합까지 특징으로 고려하였다는 점이다.

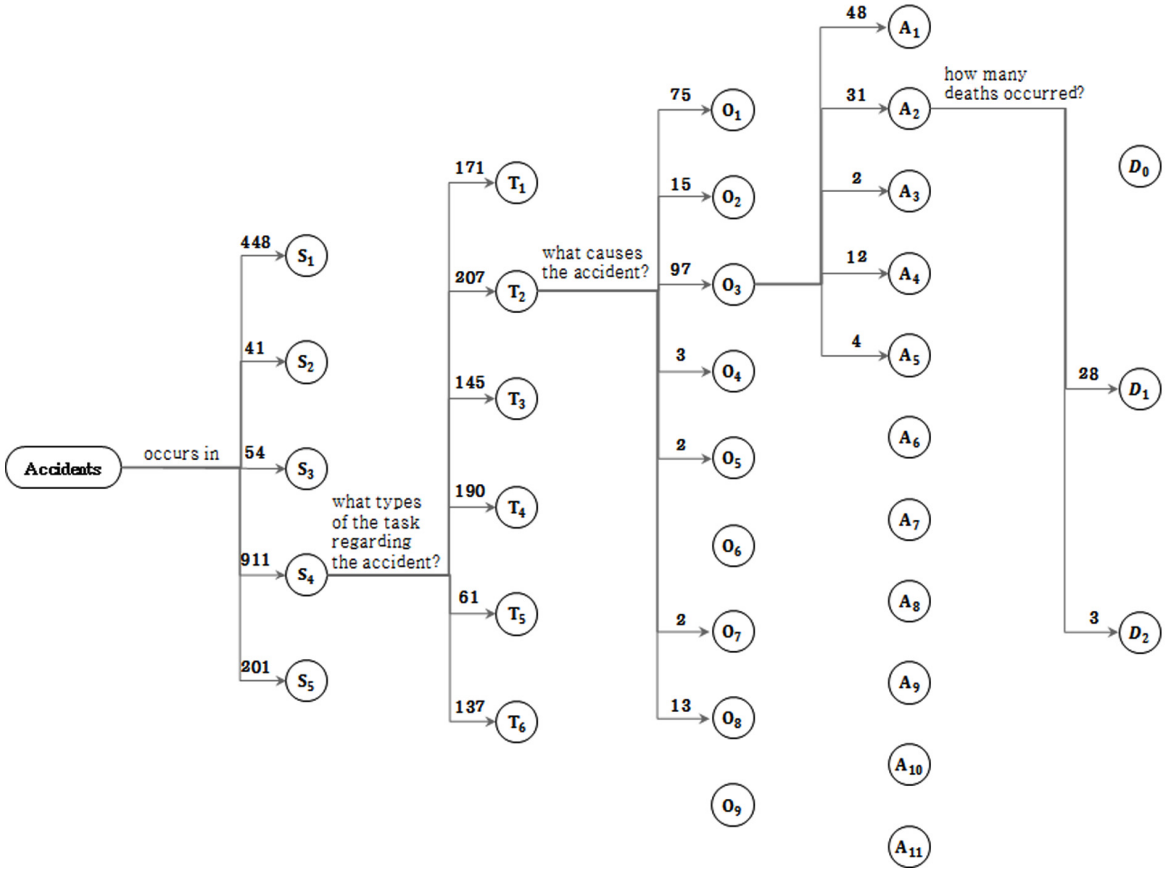


Fig. 7. Analysis result of the 1,655 accidents with the classification model and the ontology.

사고 분류 모형을 1,000개의 실제 사고 보고서 개요를 이용하여 학습한 결과, 정확도와 정밀도 모두 100%에 가깝다는 사실을 확인하였다. 온톨로지는 사고 분류 모형을 통해 나온 예측 결과를 효과적으로 표현한다. 제안한 사고 분류 모형과 온톨로지는 효과적으로 재해를 분석하는 데에 도움이 되리라 기대된다.

본 연구의 한계는 세 가지로 요약할 수 있다. 첫째, 전문가의 도움 없이 여러 기술 문서만 참고하여 라벨링하고 데이터를 구성하였기에 학습 데이터에 오류가 있을 수 있다. 그뿐 아니라, 본 연구에서는 하나의 사고에 대해 한 종류의 라벨만 부착하였다. 즉, 한 사고에 대해서는 하나의 작업 유형, 기인물, 사고 유형, 사망자 수만 라벨링 하였다. 그러나 실제로는 두 개 이상의 작업 유형, 사고 유형, 기인물이 존재할 수 있다. 따라서 현실적인 모형 수립을 위해서는 두 개 이상의 라벨을 부착해야 한다. 둘째, 5겹 교차 검증을 바탕으로 모형을 학습하였지만, 많은 특징이 사용되었기에 과적합(overfitting)이 발생했을 수 있어, 특징 선택 결과를 해석할 때 주의가 필요하다.

마지막으로 안전보건공단에서 제공한 데이터와 사고 분류 체계에 의존하여 분류 모형과 온톨로지를 구성하였으므로, 다양한 사고 보고서에 적용 가능한 지 확신할 수 없다. 따라서 추후 연구에서 특정 분야(예: 건설업)에 최적화된 분류 모형과 온톨로지를 개발하고 다양한 출처의 사고 보고서를 분석하도록 한다.

References

1) http://laborstat.molab.go.kr/
2) T. Kontogiannis, V. Leopoulous and N. Marmaras, “A Comparison of Accident Analysis Techniques for Safety-critical Man-machine Systems”, International Journal of Industrial Ergonomics, Vol. 25, No. 4, pp. 327-347, 2000.
3) J. Kim, J. B. Lee, and S. R. Chang, “A Study on the Accident Analysis of Architectural Work”, Journal of the Korean Society of Safety, Vol. 31, No. 3, pp. 96-101, 2016.
4) O. N. Jeong, Y. S. Yun and O. H. Kwon, “Accident Prevention for the Elevator and Escalator by the Accident Type Analysis”, Journal of the Korean Society of Safety, Vol. 31, No. 4, pp. 15-21, 2016.
5) Y. S. Yun and O. H. Kwon, “A Study on Industrial Injury Analysis for Manufacturing Industry”, Journal of the Korean Society of Safety, Vol. 28, No. 8, pp. 13-18, 2013.
6) M. Gruninger and J. Lee, “Ontology”, Communications of the ACM, Vol. 45, No. 2, p. 39, 2002.
7) D. Yang, R. Miao, H. Wu and Y. Zhou, “Product Configuration Knowledge Modeling using Ontology Web Language”, Expert Systems with Applications, Vol. 36, No. 3, pp. 4399-4411, 2009.
8) F. J. García-Peñalvo, R. Colomo-Palacios, J. García and R. Therón, “Towards an Ontology Modeling Tool, A Validation in Software Engineering Scenarios”, Expert Systems with Applications, Vol. 39, No. 13, pp. 11468-11478, 2012.
9) https://www.kosha.or.kr/board.do?menuId=541
10) K. N. Junejo, A. Karim, M. T. Hassan and M. Jeon, “Term-based Discriminative Information Space for Robust Text Classification”, Information Science, Vol. 372, pp. 518-538, 2016.
11) http://konlpy-ko.readthedocs.io/ko/v0.4.3/
12) H. Drucker, D. Wu and V. N. Vapnik, “Support Vector Machines for Spam Catecorization”, IEEE Transactions on Neural Networks, Vol. 10, No. 5, pp. 1048-1054, 1999.
13) Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization”, Information Retrieval, Vol. 1, No. 1, pp. 69-90, 1999.