잠재 디리클레 할당 기반 토픽 모델링을 통한 건설재해 사례 분석

Analysis of Construction Accident Incident Using Latent Dirichlet Allocation-based Topic Modeling

김창재[1] · 김하림[2] · 이창수[3] · 조훈희[4*]
Kim, Changjae[1] · Kim, Harim[2] · Lee, Changsu[3] · Cho, Hunhee[4*]

Abstract

The construction industry has more safety accidents than other industries. Although there have been more attempts to reduce safety hazards in the industry such as the enforcement of the "Serious Accidents Punishment Act (SAPA)", construction accident has not been reduced enough. In this study, analysis of safety risk factors has been made through Latent Dirichlet Allocation (LDA)-based topic modeling. Risk analysis in construction site would be improved with natural language processing and topic modeling.

키 워 드 : 건설 안전, 건설재해, 잠재 디리클레 할당, 토픽 모델링
Keywords : construction safety, construction accident, latent dirichlet allocation, topic modeling

1. 서 론

1.1 연구의 목적

건설업은 타 산업에 비해 재해자 및 사망자 수가 많고, 전체 근로자에 대한 재해자의 비율 또한 많다[1]. 이에 정부에서는 안전 인식 개선과 중대재해처벌법 시행 등으로 산업 전반의 안전성 향상을 도모하고 학계에서는 건설 현장 안전성을 높이기 위한 연구가 계속되고 있으나, 건설재해의 저감이 크게 이루어지고 있지 않은 실정이다[2]. 건설재해는 공정이나 사용되는 자원에 따라 달라지기 때문에 설계, 시공, 해체 등 모든 단계에서의 위험 예방이 중요하다[3]. 따라서 본 연구에서는 미국 직업안전위생관리국(Occupational Safety and Health Administration, OSHA)에 기록된 사고 데이터 요약문을 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기반 토픽 모델링을 통해 자연어 기반 건설재해 데이터를 분류하고, 토픽(주제) 분포와 토픽별 재해 요소를 도출하여 발생하는 사고에 따른 공종, 신체 부상 부위에 대해 분석하였다.

2. LDA 기반 토픽 모델링을 활용한 건설재해 사례분석

2.1 LDA 기반 토픽 모델링

토픽 모델링은 문서들이 포함하는 단어들을 분석하여 문서의 토픽을 찾아내는 통계적 방법이다[4]. 토픽 모델링 기법 중 하나인 LDA는 문서들의 토픽과 단어의 확률 분포를 통해 토픽의 분포와 토픽별 단어의 분포를 추정하는 기법이다[5]. LDA 모델을 도식화하면 그림 1과 같은데, K개의 토픽이 문서 집합 D에 걸쳐 존재한다고 가정하면, LDA 모델은 문서의 단어 $W_{d,n}$에 대해 한 개의 토픽 $Z_{d,n}$을 무작위로 할당한다. 이를 전체 문서에 대해 반복하면 문서 집합의 모든 단어에 대해 토픽이 무작위로 할당된다. 이후 문서의 단어 중 특정 토픽에 해당하는 단어들의 비율과 특정 단어를 포함한 문서 중 특정 토픽에 할당된 비율을 기준으로 다시 할당하면 유사한 토픽을 구성하는 단어, 그리고 유사한 토픽을 가지는 문서가 수렴한다.

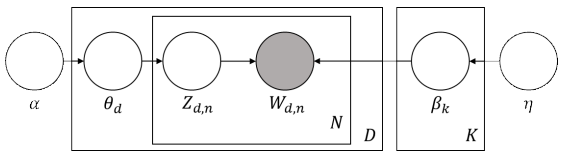

그림 1. LDA 모델

1) 고려대학교 건축사회환경공학과 석사과정
2) 고려대학교 건축사회환경공학과 박사과정
3) 고려대학교 건축사회환경공학과 박사수료
4) 고려대학교 건축사회환경공학부 교수, 교신저자(hhcho@korea.ac.kr)

2.2 건설재해 사례분석

본 연구에서는 2012년 3월부터 2021년 7월까지의 OSHA에서 수집된 건설재해 데이터를 활용하여 분석을 수행하였다. 해당 데이터는 요약, 경위, 키워드로 이루어진 데이터인데, 길이가 긴 사고 경위문의 경우 사고뿐 아닌 병원 이송 내용 등까지 포함하는 경우가 많고, 불필요한 정보가 포함되기 쉬울 것으로 판단되기 때문에 본 연구에서는 사고의 원인과 부위, 피해가 명료하게 작성된 사고 요약문 데이터를 활용하여 분석을 수행하였다. 본 연구에서는 영어 자연어 처리를 위한 파이썬 기반 라이브러리인 NLTK(Natural Language Toolkit)를 통해 OSHA의 사고 요약문 데이터를 전처리한 후, 토픽 모델링을 위한 패키지인 tomotopy를 통해 15개 토픽으로 이루어진 LDA 모델을 구성하여 분석하였다.

표 1. LDA를 통한 건설재해 사례 분석 결과

토픽번호	사고유형	단어(토픽 내 분포 확률 높은 순)
1	추락사고	ladder, roof, fall, injures, breaks, scaffold, leg, worker, back, head
2		fall, roof, injured, floor, during, construction, worker, opening, elevator, shaft
3		fall, injured, dies, scaffold, roof, ladder, employees, falling, collapses, workers
4	절단사고	finger, amputated, caught, amputates, worker, has, fingers, crushed, hand, machine
5		saw, operating, finger, amputates, using, back, power, amputated, fingers, kicks
6	복합사고	incurs, multiple, injuries, sustains, roof, worker, ladder, fall, scaffold, head
7	붕괴사고	trench, collapse, injured, collapses, worker, wall, excavation, crushed, employees, buried
8	깔림	struck, falling, worker, leg, injured, crushed, concrete, beam, steel, wall
9	건설기계관련사고	truck, struck, run, crushed, dump, loader, injured, leg, forklift, crane
10	교통사고	struck, vehicle, zone, highway, injured, road, motor, traffic, truck, construction
11	찔림	nail, gun, knee, punctures, eye, using, shoots, worker, operating, head
12	질병	dies, heat, worker, heart, attack, sickened, suffers, illness, heat-related, exhaustion
13	감전사고	power, line, electrocuted, shocked, electric, contacts, worker, shock, ladder, electrical
14	화상사고	burned, arc, flash, burns, worker, employees, electric, hot, electrical, water
15	기타	lift, scissor, aerial, carbon, employees, injured, monoxide, fall, basket, caught

LDA 모델에서 분포하는 토픽별 단어를 분석하여 표 1과 같이 사고유형으로 분류하였다. 토픽에 따른 단어 분포 분석 결과 추락사고, 절단사고, 복합사고, 붕괴사고, 깔림, 건설기계 관련사고, 교통사고, 찔림, 질병, 감전사고, 화상사고 등의 사고유형을 도출하였다. 각 토픽에서는 사고원인, 사고발생 공종, 신체 부상 부위 등의 단어들이 분포하는 것으로 나타났다. 가장 많은 토픽에 분포하는 것으로 나타난 추락사고의 경우에는 사다리, 지붕, 엘리베이터 샤프트 등의 부위에서 추락하여 작업자의 등, 다리, 머리 부위의 충돌로 인해 부상을 당하거나 사망하였다는 것을 확인할 수 있었다.

3. 결 론

OSHA는 유형화되지 않은 건설재해 데이터를 제공하나, LDA 기반 토픽 모델링을 통해 이를 분류할 수 있었다. 본 연구에서는 자연어로 구성된 건설재해 사례를 분석하여 사고사례의 토픽 분포와 토픽별 키워드 분포를 확인하여 발생하는 사고유형에 따른 사고의 상세 내용을 확인할 수 있었다. 향후 연구를 통해 전처리 과정에서 단어의 품사 태깅을 활용하여 사고를 유발하는 행위, 객체 및 피해 부위 등을 규명하고, 자연어 형태의 데이터 분류를 자동화할 수 있을 것으로 기대된다.

감사의 글

본 논문은 2022년 스마트건설기술개발사업(과제번호: 22SMIP-A158708-03)의 일환으로 수행된 연구임

참 고 문 헌

1. 고용노동부. 2021년 산업재해 발생현황. 2022.
2. 양성웅, 임형철. 건설 재해에 관한 연구동향의 의미연결망 분석. 대한건축학회논문집. 2021. p.231-236.
3. 김진원, 김요한, 김주형, 김재준. 건설재해의 유형분석을 통한 안전사고 저감방안에 관한 연구. 한국건축시공학회 학술, 기술논문발표회 논문집. 2010. p.137-140.
4. Blei, D. M. Probabilistic topic models. Communications of the ACM. 2012. p.77-84.
5. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. Journal of machine Learning research. 2003. p.993-1022.