

2020

한국전과정평가학회·한국산업생태학회 ·한국환경경영학회
공동학술발표회



세브란스
SEVERANCE

텍스트 마이닝을 이용한 산업재해 원인의 탐색적 연구 :

논문과 언론보도 기사를 중심으로

An Exploratory Study of the Causes of Industrial Accidents Using Text Mining :
Based on Research Papers and Media Reports



연세의료원 김홍관

1. Introduction

- 산업재해의 원인이 되는 유해·위험요인을 충분히 제시되지 못하고 누락되고 있음. (Ref: jeong, 2018, Song, 2010)
- 위험성의 추정 및 결정방법의 적용을 통해 유해·위험요인별 현재 관리수준 Data Base 구축이 용이하고, 정량적이고 객관적인 평가로 데이터의 신뢰성 확보가 가능하다고 함. (Ref: Lee, 2019)
- 위험성평가 시 재해발생에 영향을 미칠 수 있는 다양한 요소가 적용되어야 하고, 더 많은 영향요소들이 추가될수록 재해발생에 대한 확률이 보다 정확해짐.
(Ref: Won, 2019)
- 빅데이터의 양적인 증가로 빅데이터의 생성, 가공, 저장 분석결과 및 시각화에 관심이 증가되는 추세에 있음
- 문서, 댓글 등의 데이터로 수집한 후 텍스트 간의 상관성(Correlation)에 기초해 분석함으로써 실증적인 관계에 대한 분석을 진행하여 새로운 통찰과 유용한 예측이 가능함. (Ref: Ah, 2017)

2. Objectives

- 위험성평가 관련 논문과 언론보도 기사를 대상으로 텍스트 네트워크 분석을 통한 연구동향 파악과 산업재해 원인 관련 텍스트의 분석으로 주요 키워드 출현빈도로 중요도 분석, 키워드 간의 관계성 분석
- 특징적인 패턴을 지닌 키워드들의 구조적 분석으로 지식구조를 파악



3. Methods

• 연구대상

- ✓ 국내 학술논문 585편과 언론보도 기사(한국언론진흥재단 제공 종합일간지 44개) 1,976건

• 연구설계

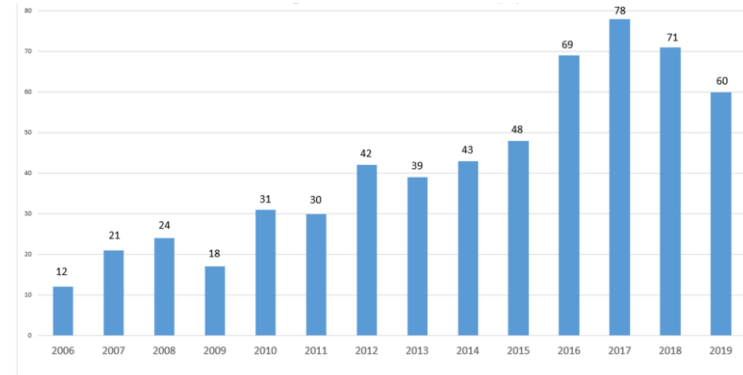
1 단 계	연구의 대상 선정		- 학술논문(학위논문 포함) : ‘위험성평가’ - 언론보도 기사 : ‘산업재해’ or ‘산재’ or ‘업무상재해’ and ‘원인’
2 단 계	분석 준비단계	데이터 수집	- 자료 수집 - 국내 논문 자료 수집 : RISS, NDSL, KCI - 국내 언론보도 기사 자료 수집 : BIGKINDS - 수집기간 : 2000년 1월 ~ 2019년 12월 - 수집대상 : 유해 · 위험요인 - 언 어 : 명사
		전처리 (데이터 정제)	- 프로그램 : Textom - 한글 형태소 분석 : 명사 추출 - 불용어, 유사어 등의 단어 표준화 데이터를 정형데이터로 변환
3 단 계	분석 모형 : 텍스트 마이닝 (Text-Mining)	네트워크 분석	- 프로그램 : Textom - 상위 40개 단어(핵심어 도출) - 빈도분석, TF-IDF 분석, 워드 클라우드
		CONCOR 분석	- 프로그램 : NetDraw - 분석과정 : 키워드 간의 상관성을 바탕으로 한 군집분석

4. Results

• 위험성평가 관련 연구현황

- ✓ Frequency trend of research papers by year

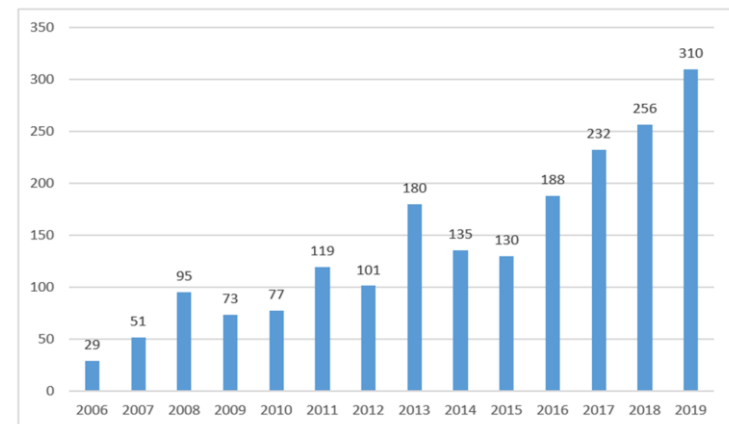
Year	Number of research papers	Rate (%)
2006	12	2.05
2007	21	3.59
2008	24	4.10
2009	17	2.91
2010	31	5.30
2011	30	5.13
2012	42	7.18
2013	39	6.67
2014	43	7.35
2015	48	8.21
2016	69	11.79
2017	78	13.33
2018	71	12.14
2019	60	10.26
Total	585	100.00



• 산업재해 원인에 대한 언론보도 기사 현황

- ✓ Frequency trend of media reports by year

Year	Before filtering	After filtering
	N	N
2006	214	29
2007	263	51
2008	378	95
2009	474	73
2010	491	77
2011	605	119
2012	495	101
2013	676	180
2014	822	135
2015	680	130
2016	632	188
2017	728	232
2018	774	256
2019	802	310
Total	8,034	1,976



4. Results

- 키워드 빈도분석 Top 10 keywords by appearance frequency : 2006~2019

No	Research papers			Media reports		
	Keyword	N(%)	TF-IDF	Keyword	N(%)	TF-IDF
1	Chemical substances	1,129(16.66)	1,389.40	Outsourcing	738(12.20)	1,108.03
2	Fire	743(10.77)	879.65	Semiconductor	479(7.92)	1,009.33
3	Gas	608(8.81)	900.31	Chemical substances	384(6.35)	870.80
4	Explosion	491(7.11)	794.89	Fall	379(6.26)	809.65
5	Leakage	398(5.76)	698.38	Explosion	328(5.42)	745.66
6	Machine equipment	325(4.71)	582.32	Fire	222(3.67)	590.97
7	Storage tank	210(3.04)	567.37	Leakage	180(2.98)	503.20
8	Dust	163(2.36)	548.36	Crane	178(2.94)	521.13
9	Hazardous material	162(2.35)	425.50	Stress	162(2.68)	455.99
10	Exposure	145(2.10)	359.40	Overwork	151(2.50)	443.73



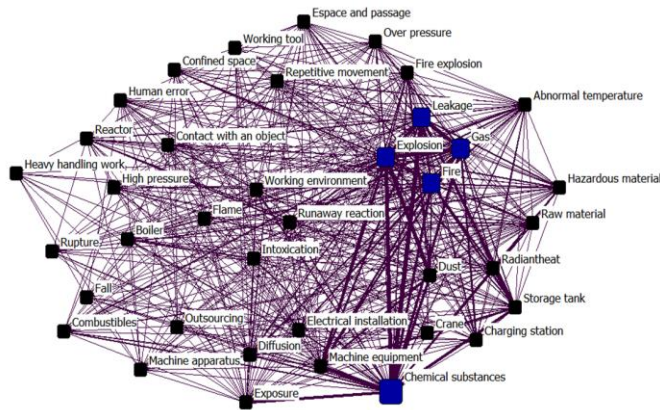
[Word cloud of keywords : 2006~2019 research papers]



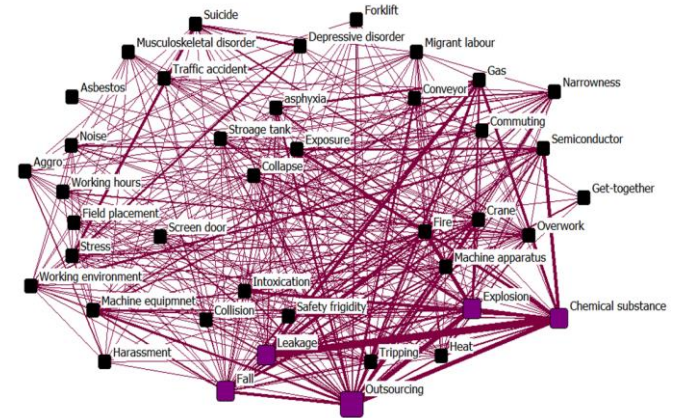
[Word cloud of keywords : 2006~2019 media reports]

4. Results

- 텍스트 네트워크 분석

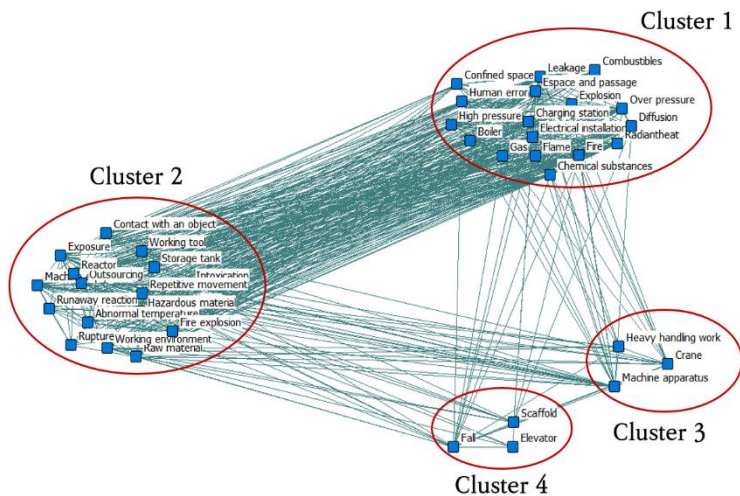


[Text network analysis : 2006~2019 research papers]

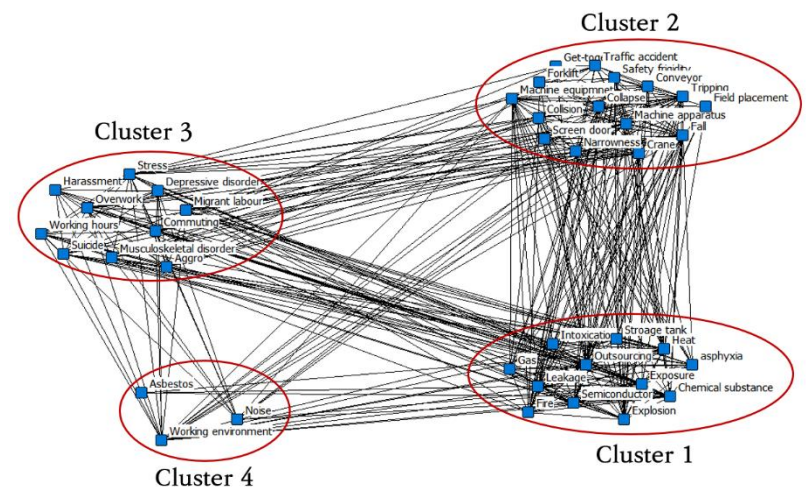


[Text network analysis : 2006~2019 media reports]

- 의미연결망 분석 CONCOR Analysis



[Visualization with CONCOR analysis : 2006~2019 research papers]



[Visualization with CONCOR analysis : 2006~2019 media reports]

5. Conclusion

- 키워드 빈도분석 결과, 공통적으로 '화학물질, 화재, 폭발, 누출' 등의 키워드가 높은 비중을 차지함
- 구조적 등위 관계분석으로 CONCOR 분석을 수렴수준 3으로 하여 연구논문과 언론보도 기사 모두 유사성을 가지는 그룹으로 4개의 클러스터(군집)를 형성함
 - ✓ 일반적인 재해의 유형은 어느 한가지 요인에 의한 것이 아니라 여러가지 복합적인 인과관계에서 발생하는 것으로 비슷한 유형에 대한 각각 일정한 군집으로 나누어 특성을 분석함으로써 유사성을 확인할 수 있다.
 - ✓ 이러한 유사성을 고려하여 위험성평가 시 유해·위험요인의 파악, 평가 내용의 적절성 분석, 위험성 결정에 도움이 될 것이다.
- 오랜 기간동안 축적된 데이터를 계량적인 분석을 통해 텍스트 네트워크 분석으로 텍스트를 추출하여 네트워크 간 유의미한 관계를 확인하였다.
 - ✓ 산업재해 통계에서 떨어짐(추락), 넘어짐(전도), 끼임(협착), 업무상 질병이 산업재해 원인의 가장 많은 비중을 차지하였으나, 연구논문과 언론보도 기사에서 화학물질, 화재, 폭발 등을 중요도 있게 다루고 있음을 알 수 있다.