ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 December: 29(4): 91~110 http://dx.doi.org/10.13088/jiis.2023.29.4.091

비정형 데이터를 이용한 화학물질 사고 대응 체계 정보속성 비교 분석 : 화학사고 예방, 대비 및 대응을 위한 OECD 지침서를 중심으로

김용진
한국화학연구원 연구기획실
(koine@krict.re.kr)

도충현
한국화학연구원 연구기획실
(dch125@krict.re.kr)

화학물질 사고는 신속한 대응 및 복구가 어렵고, 환경오염과 인명피해가 동반된다는 점에서 매뉴얼의 중요성이 점차 주목받고 있으며, OECD에서는 화학사고 예방, 대비 및 대응을 위한 OECD 지침서(이하 OECD 지침서)를 2023년 6월 개정하였다. 또한, 기존 연구에서는 화학사고에 대한 인식 제고를 통해 법규, 규정, 매뉴얼 등 시스템적 대응이 필요하다는 점을 강조하고 있으나. 매뉴얼에 대한 정보속성 비교연구는 찾아보기 힘들었다. 이에, 본 연구는 기존 OECD 지침서(2판)와 개정된 OECD 지침서(3판)을 비교분석하여 OECD 지침서별 정보속성을 파악하고 시사점을 발굴하는 것을 목표로 하였다. 세부적으로는 어떤 단어가 중요해졌는지 파악하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 분석을 적용하였으며, 유사하게 사용한 단어와 차별성있게 사용한 단어를 파악하기 위해 Word2Vec을 적용하였다. 최종적으로는 2X2 매트릭스를 제안하고, 각 사분면에 어떤 단어들이 있는지를 도출하여 OECD 지침서별 정보속성을 심층적으로 비교하였다. 본 연구는 연구자들이 정보속성을 파악하는데 도움이 되는 프레임워크를 제공하고자 하였으며, 실무적으로는 국내 화학관련 정부부처 및 기업의 표준메뉴얼 개정에 참고할 수 있을 것으로 보인다.

주제어 : 화학사고, 화학사고 예방, 비교분석, TF-IDF, Word2Vec

논문접수일 : 2023년 9월 5일 논문수정일 : 2023년 11월 26일 게재확정일 : 2023년 11월 28일
원고유형 : Regular Track 교신저자 : 김용진

1. 개요

화학물질은 우리 삶에 밀접하고 이용되고 있으며, 전 세계적으로 2014년 10만 종에서 2021년 약 20만 종의 화학물질이 유통되고 있을 만큼 급격하게 증가하고 있다 (환경부, 2022). 또한, 매년 3천여 종의 새로운 화학물질이 개발되고 있으며, 국내에서도 약 4만 4천종 이상의 화학물질이 유통되고 있다 (유지선, 정영진, 2014; 환경부, 2022). 그리고, 화학산업의 매출액에 있어 한국은 세계 5위권 국가이자, 화학과 같은 제조업 기반의 산업구조를 가지고 있다 (김민구 등, 2022; CEFIC, 2023). 이처럼 화학산업이 중요한 한국에서 화학물질 사고에 대한 대응과 의식은 그에 미치지 못하고 있다 (김용진 등, 2022; 박종서 등, 2012).

정책을 수립함에 있어, 사회적 중요성 및 관심사를 반영하여 다양한 이해관계자들이 공통의 목소리를 내는 과정이 중요하며 (김진솔 등, 2021), 환경부가 2012년 구미공단 불산 누출사고 이후 개정한 매뉴얼은 그 실마리가 될 수 있다. 매뉴얼은 사전적으로 내용이나 이유, 사용법을 설명한 글이며, 어떤 활동이나 기계의 조작 방법 등을 알기 쉽고

편리하도록 설명해 놓은 지침서이다 (국립국어원, 2023). 화학물질 사고는 갑자기 발생하며, 화학물질이 한번 퍼지고 난 이후에는 대응 및 복구가 어렵고, 환경오염과 인명피해가 동반된다는 점에서 매뉴얼의 중요성은 더욱 부각된다 (김용진 등, 2022).

2000년대 이후 높아진 화학물질 사고에 관한 관심과 더불어 대형 화학물질 사고 등으로 인해 관련 연구는 현재까지도 꾸준히 진행되고 있다. “화학물질 사고”라는 단어가 포함된 국내 논문은 2010년 2건, 2011년 3건에서 2012년 구미공단 불산 누출 사고를 계기로 12건으로 급증하였으며 2019년 51건으로 최고를 기록한 이후 서서히 감소하여 2022년에는 25건을 기록하였다 (한국학술지인용색인, 2023).

화학물질 사고 대응 체계와 관련한 주요 연구는 다음과 같다. 박종서 외(2012)에 따르면, 화학물질 사고 대응 매뉴얼을 개선하기 위해서는 정부 및 산업계와의 연계가 중요함을 강조한다. 유지선과 정영진(2015)은 화학물질 사고사례를 분석하여 문제점과 대처방안을 도출하였다. 이재석과 최돈묵(2015)은 화학물질 사고에 대한 국가적 시스템이 부족함을 지적하고, 체계적인 매뉴얼의 필요성을 강조한다. 김용진 등(2022)은 한국, 중국, 일본의 특정화학물질 관리체계를 분석하여 법령개선 등의 개선방안을 도출하였다. 기존 연구는 화학물질 사고에 대한 경각심을 일깨워 주면서 법령, 매뉴얼과 같은 시스템 관점의 대응이 필요함을 주장하였다. 하지만, 화학물질 사고 대응 체계에 핵심이라고 할 수 있는 매뉴얼 내, 정보속성을 분석하는 연구는 찾아보기 힘들었다.

화학물질 사고의 국민적 관심의 증대는 과학계와 산업계 그리고 정부가 현재 및 미래의 화학물질 사고를 해결해야 하는 필요에 대한 명확한 입장을 이끌었다. 그 결과, 2023년 6월, 화학사고 예방, 대비 및 대응을 위한 OECD 지침서(이하 OECD 지침서)가 개정되어 공개되었다. 개정된 OECD 지침서는 화학사고 예방과 함께 사고가 발생하였을 경우 적절한 조치를 취하는 방안을 포함하고 있으며, 위험설비의 안전한 계획 및 운영에 대한 지침을 제공한다 (OECD, 2023). 개정된 OECD 지침서가 공개된 지 얼마 되지 않아, 이에 대한 연구는 찾아볼 수 없었다. 또한, 한국에서 화학물질 사고와 관련하여 공개된 최신의 매뉴얼인 유해화학물질 유출사고 위기관리 표준매뉴얼(2012년 개정)은 OECD 지침서(2판)을 기준하고 있어 (OECD, 2003), 향후 국내 화학물질 사고 매뉴얼 수립에도 큰 영향을 미칠 것으로 보인다.

이에, 본 연구는 개정된 OECD 지침서 3판을 대상으로 정보속성을 기존 OECD 지침서인 2판과 비교·분석하였다. 정보속성을 비교함에 있어 본 연구는 단어를 사용하였는데, 이는 문서 내에 단어를 추출하는 것은 정보의 빠른 전달 등에 있어 중요하며, 문서 상호간 비교에 있어 중요 요인으로 활용되기 때문이다 (이동훈, 김관호, 2018). 또한, OECD 지침서 전체 내용을 대상으로 어떤 단어가 중요한지를 판별하고자 TF-IDF(Term Frequency-Inverse Document Frequency, 이하 TF-IDF) 분석을 적용하였다. 그리고, 2판과 3판에서 유사하게 사용한 단어와 차별성있게 사용한 단어를 파악하기 위해 Word2Vec를 적용하였다. 최종적으로는 2X2 매트릭스를 제안하고, 각 사분면에 어떤 단어들이 있는지를 도출하여 OECD 지침서별 정보속성을 심층적으로 비교하였다. 그 결과는 향후 연구자들이 정보속성을 파악하는 데 도움이 되는 프레임워크를 제공할 수 있을 것이며, 실무적으로는 국내 화학분야 정부 및 기업의 표준메뉴얼 개정에 참고할 수 있을 것으로 보인다.

본 연구는 총 5장으로 구성되어 있다. 2장은 화학물질 사고 현황 및 OECD 지침서에 대해 살펴본다. 3장 데이터 수집과 함께 본 연구의 분석

방법인 TF-IDF 및 Word2Vec를 설명한다. 4장 분석결과에서는 OECD 매뉴얼 2판과 3판의 핵심 단어 선정과 함께 의미가 유사하거나 변화한 단어를 도출한다. 5장에서는 분석결과를 바탕으로 금번 개정된 OECD 지침서의 정보속성을 파악하고, 시사점 및 연구의 한계점을 논의한다.

2. 배경조사

2.1 화학물질 사고 현황

2012년부터 2022년까지 한국에서 발생한 화학물질 사고는 총 831건에 달한다. 세부적으로 살펴보면 사고유형별로는 유출사고가 76.9%인 654건으로 가장 많았으며, 폭발사고 73건(8.5%), 화재사고 53건(5.7%), 기타 51건(6.9%) 순이었다. 또한, 2012년 9건에서 2013년 86건으로 급증한 이후, 2022년까지 화학물질 사고가 큰 변화없이 꾸준히 발생하고 있다(<표 1> 참조) (화학물질안전원, 2023).

〈표 1〉 화학물질 사고 형태별 발생 건수 (단위: 건)

구분	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	합계
전체	9	86	98	114	78	88	66	58	75	93	66	831
유출	4	65	75	78	58	75	60	48	63	78	50	654
폭발	2	11	10	11	11	7	0	2	4	9	6	73
화재	0	5	7	11	7	4	3	0	6	3	7	53
기타	3	5	6	14	2	2	3	8	2	3	3	51

화학물질 사고사례를 살펴보면, 첫 번째는 2012년 9월에 발생한 구미공단 불산 누출 사고를 들 수 있다. 이 사고는 불산(Hydrofluoric Acid, 산업분야에서 널리 쓰이는 비유기성 산의 일종으로 불산 자체는 불연성이나 금속과 반응, 수소를 발생하여 인화 및 폭발 위험과 무색에 자극적 냄새가 있는 발연성 액체로 강한 부식성과 물에 녹는 성질을 보유) 20톤을 적재한 탱크로리에서 공장 저장탱크로 옮기던 중 8톤의 가스가 누출되어 발생하였다. 구미공단 불산 누출사고는 국내 최대의 독극물 누출사고로 5명이 사망하고 주변 약 200헥타르(60만평)에 달하는 농경지에도 피해 발생하였다 (환경부, 2013). 두 번째로는 2016년 6월, 울산시에서 발생한 황산 누출사고를 들 수 있다. 이 사고는 황산제조시설 보수 작업 중, 시설 내부에 잔류한 황산(70%)이 누출되어 발생하였으며, 2명이 사망하고 4명이 다쳤다 (소방신문, 2023). 세 번째로는 2018년 11월 부산에서 발생한 황화수소 누출사고를 들 수 있다. 이 사고는 폐수를 집수조로 투입하는 과정에서 집수조내 잔류 폐수가 화학반응을 일으켜 황화수소가 누출되면서 발생되었다. 이 사고로 3명이 사망하고 7명이 다쳤다. 대표 사례에서 살펴본 바와 같이 화학물질 사고는 발생 건수 증가뿐만 아니라 피해 규모가 커지고 유형이 다양화되고 있다. 또한, 2022년에도 화학공장이 밀집한 울산, 전남 및 대구에서도 화학사고 및 인명피해가 지속되고 있다 (<표 2> 참조) (소방신문, 2023).

〈표 2〉 전국 주요 화학물질 사고사례(2022년)

지역	사고사례	인적피해
울산	- 톨루엔 저장탱크 청소작업 중 폭발 화재 - 폴리에틸렌 생산공장 내 사이클로 헥세인 유출 폭발 화재 등	- 사망 : 4명 - 부상 : 15명
전남	- 나프타 분해업체 열교환기 내부압력 시험 중 폭발 등	- 사망 : 4명 - 부상 : 4명
대구	- 정수사업소 저류조 내 황화수소 질식 등	- 사망 : 1명 - 부상 : 2명
경기	- 약품 생산공장 내 아세톤 취급 중 폭발 화재 등	- 사망 : 1명 - 부상 : 15명

2.2 OECD 지침서

OECD는 주요 환경문제를 해결하지 않는다면, 향후 몇십 년 내에 지속적인 경제 번영을 위한 환경적 기반이 크게 훼손될 것으로 경고하였다. OECD가 지적한 환경문제에는 기후변화, 생물다양성 감소 및 물부족 뿐만 아니라, 오염 및 유해화학물질도 주요한 요인으로 포함되었다 (김대환, 2010).

이에, OECD는 1992년에 화학사고 예방, 대비 및 대응에 관한 지침서 초판을 발간하여, 회원국들이 화학사고에 대한 표준화된 접근 방식을 공유하고 협력을 강화할 수 있도록 화학사고 예방, 대비 및 대응에 관한 지침을 수립하였다.

이후, 화학사고에 대한 경험이 축적됨에 따라 지침서에서 다루고자 하는 범위가 화학사고를 다루는 보건 인프라 개발, 중소기업의 가이드 원칙 규정 준수 등 7개로 확대되었다. 이에, 정부기관, 산업계, 환경단체 및 관련 업무를 담당하는 기타 국제기관들의 전문가들이 참여하여 개정작업을 하였다. 그 결과 2003년 2판이 발간되었으며, 중국어, 체코어, 불어, 독어, 헝가리어, 이태리어, 한국어로도 번역하여 전 세계적으로 널리 활용될 수 있도록 하였다.

기존 지침서와의 가장 큰 차이점은 행동원칙(Golden Roles)을 포함하였다는 점으로 화학사고 예방, 대비 및 대응과 관련된 이해관계자들의 책임에 관하여 상세하게 서술하고 있다. 또한, 사고 예방, 비상 준비 및 완화, 비상 대응, 사고의 사후처리, 특별 사항별 내용을 재구성하였다.

〈표 3〉 OECD 지침서 주요내용(2판)

구분	주요내용
제1장 사고예방	- 모든 단계 및 관련 부분에서 유해물질 사고예방의 중요성을 설명하고 있으며, 산업계, 정부기관 등의 역할과 책임에 관한 내용을 중점적으로 서술
제2장 비상조치 준비/완화	- 비상계획 및 사고완화 분야에서 산업계, 정부기관 등 이해관계자들의 역할과 책임 제시 - 발생 가능성이 있는 사고로 인한 건강, 환경, 물적 자산에 대한 피해를 최소화할 수 있는 비상대비계획, 프로그램을 포함
제3장 비상대응	- 화학사고 발생시 산업계, 정부기관 등 이해관계자들이 협력하여 건강, 환경, 물적 자산의 피해를 줄이기 위한 방안들을 제공
제4장 사고의 사후처리	- 사고 대응활동 이후에 취해지는 사항들로 향후 유사한 사고를 예방하기 위해 사고보고와 조사에 관한 내용을 제공
제5장 특별사항	- 국가간 협력 및 국제적 지원, 국제적 개발기술 이전사항을 포함한 국가간/국제적인 관심사 제공 - 수송 방법의 소유주/운영자, 하역작업을 하는 근로자 등 관련 이해관계자들의 역할과 책임에 관한 지침을 제시

2003년 개정판 발간 이후에도 화학물질 사고는 지속적으로 발생하였으며, 발생하는 화학물질 사고의 영향을 완화하기 위한 적절한 조치가 필요하게 되었다. 이러한 문제들을 보완하기 위해 OECD에서는 2023년에 세 번째 개정판을 발간하였으며, 기존 개정된 지침서와의 가장 큰 차이점은 2003년 이후 전 세계적으로 발생한 대형 화학사고를 통한 교훈과 최신 모범사례, 안전관리 성과를 제공한다는 점에 있다 (<표 4> 참조).

〈표 4〉 OECD 지침서 주요내용(3판)

구분	주요내용
제1장 사고예방	- 효과적인 사고예방을 위한 산업계 및 공공기관의 주요 역할 핵심 요소 및 화학사고 예방에 관한 일반 원칙을 제공
제2장 비상조치 준비/완화	- 사고현장 및 외부 비상계획을 위해 이해관계자 간의 비상 대비 및 완화에 대한 원칙을 제공
제3장 비상대응	- 화학사고 발생 시 산업계와 정부기관 등의 역할과 책임에 대한 원칙을 제공
제4장 사고의 사후처리	- 사고조사를 수행하는 방법과 사건 보고에 대한 원칙 및 데이터 수집, 결과 평가, 화학사고의 영향 분석을 위한 기준 등에 관한 내용 제공
제5장 특별사항	- 화학사고의 예방 및 완화에 기여하는 토지사용 계획의 개발 및 구현, 유해물질 운송 방법, 국제기구의 역할 등을 제공

3. 데이터 및 연구방법론

본 연구는 OECD 지침서(2판, 3판)를 비교분석하여 OECD 지침서별 정보속성을 파악하는 데 목적이 있다. 이 목적을 달성하기 위해 데이터를 수집하여 전처리하고, 분석방법론을 적용하였다.

3.1 데이터 수집 및 전처리

본 연구의 분석 데이터는 OECD 매뉴얼 내의 텍스트 데이터이다. 수집한 데이터는 분석할 수 있도록 다음과 같이 전처리 과정을 거쳤다. 첫 번째로는 토큰화(Tokenization) 과정으로 전체 문서를 문장 형태로 나누고, 문장은 단어로 나누는 과정이다. 대소문자를 소문자로 통일하고 특수문자를 제외하는 과정도 포함된다. 두 번째는 너무 긴 문자의 단어를 제거하는 과정이다. 본 연구에서는 OECD 지침서의 언어인 영어의 특성을 고려하여 1문자 이하의 단어는 제외하였다. 세 번째는 불용어와 숫자로만 된 단어를 제거하며, 단어들의 표제어를 추출한다. 마지막으로는 단어들의 어간을 추출한다. 데이터 전처리를 위해 본 연구는 파이썬 내의 라이브러리인 NLTK(Natural Language Toolkit) 및 Gensim을 사용했다 (Bird. 2006; Loper & Bird, 2002; Řehůřek & Sojka, 2011). <그림 1>은 불용어 제거 등의 전처리 전과 후의 데이터에 대한 예시를 보여준다.

- 1.1. The primary objective of safety-related programmes at hazardous installations is the prevention of accidents that result in harm to health, the environment or property (recognising that accidents involving hazardous substances may, nonetheless, occur).

전처리 전

⇩

- primary objective safety related programme hazardous installation prevention accident result harm health environment property recognising accident involving hazardous substance occur

전처리 후

〈그림 1〉 전처리 예시

3.2 분석방법론

3.2.1 TF-IDF

TF-IDF는 문서 내의 주요 단어를 추출하는 대표적인 기법으로서, TF와 IDF의 곱을 통해 계산된다 (김진솔 등, 2021). TF-IDF는 단어가 등장하는 빈도가 높을수록 값이 커지며, 해당 단어를 포함하는 문서가 많을수록 반비례하여 작아진다 (유은순 등, 2015; 이성직, 김한준, 2009). 또한 문서내 TF-IDF를 계산할 때 쓰는 가장 단순한 방법으로는 각 문서내 단어별 TF-IDF를 계산한 뒤 큰 순서대로 순위를 매기는 방식이 있다 (유은순 등, 2015; 이성직, 김한준, 2009). 이에, 본 연구에서는 TF-IDF 값과 함께 TF-IDF 순위를 동시에 고려

하여, 2판과 3판에 등장한 주요 단어를 추출하였다. 세부적으로는 특정 문서에서 단어의 빈도를 나타낸 TF 분석을 시행하고, 그다음 전체 문서에서 특정 단어가 포함된 문서의 수로 나눈 후 로그값을 취한 IDF 분석을 시행하여 TF-IDF를 제안하였다.

첫 번째, 전처리 된 단어를 대상으로 TF 분석을 실시하였다.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$n_{i,j}$는 문서 d_j에 등장한 단어 t_i의 횟수이며, $\sum_k n_{k,j}$는 문서 d_j에 등장한 모든 단어의 수이다 (유은순 등, 2015). 즉, 간단하게 설명하면, 단어가 문서 내에서 얼마나 등장했는지를 나타내는 값으로 문서 d_j에서 단어 t_i가 등장한 횟수를 문서 d_j에 등장한 모든 단어의 수로 나눈 값으로 표현한다.

<그림 2>는 <그림 1>의 전처리된 단어를 대상으로 hazardous, accident 및 harm에 대한 TF 계산을 보여준다. 전처리 후 단어는 총 20개이다. <그림 2>에서 볼 수 있듯이 'hazardous', 'accident'는 빈도수가 2이므로 TF값은 0.1이고, 'harm'은 빈도수가 1이므로 TF값은 0.05이다.

- primary objective safety related programme hazardous installation prevention accident result harm health environment property recognising accident involving hazardous substance occur

⇩

단어	빈도
hazardous	2
accident	2
harm	1
…	
전체	20

⇨

단어	TF
hazardous	0.1
accident	0.1
harm	0.05
전체	1

〈그림 2〉 TF 계산 예시

OECD 매뉴얼 2판과 3판에 등장한 단어는 총 3,238개 및 3,373개이다. 모든 단어를 대상으로 분석을 하는 것은 의미가 없으며 적절한 임계치를 설정할 필요가 있다 (Kim & Lee, 2020). 이에, 본 연구에서는 판별 TF 분포를 고려하여 상위 40개의 단어를 분석 대상으로 설정하였다.

단어의 중요도를 단어의 출현 빈도를 통해서만 판단하면 흔한 단어가 중요도가 높다고 판단할 수 있다. (박호연, 김경재, 2019). 예를 들자면, OECD 매뉴얼에서 'accident'라는 단어는 많이 등장하지만, 모든 문서(챕터)에 걸쳐서 등장하며 이 경우 'accident'라는 단어는 중요하다고 볼 수 없다. 그래서 각 단어가 몇 개의 문서에 나타났는지를 파악하는 것 역시 필요하다. 이는 DF(Document Frequency)라고 명명하면 다음과 같다.

$$DF_{i,j} = \frac{|d_j \in D : t_j \in d_j|}{|D|}$$

$|D|$는 문서의 전체수를 나타내며, $|d_j \in D : t_j \in d_j|$는 단어 t_j가 등장한 문서의 수를 나타낸다. <그림 3>은 <그림 2>의 문서를 문서 1로 가정하고, 가상의 문서 2, 3, 4에서 각각의 단어가 나온 횟수를 가정할 때, DF 값을 계산하는 예시를 나타낸 것이다.

단어	문서 1	문서 2	문서 3	문서 4
hazardous	2	2	2	2
accident	2	6	0	0
harm	1	0	1	0
…	…	…	…	…

⇨

단어	DF
hazardous	1
accident	0.5
harm	0.5
…	

〈그림 3〉 DF 계산 예시

DF가 높은 단어는 모든 문서에서 등장하므로 중요도가 높다고 볼 수 없다. 이에, Karen Spärck Jones는 “A statistical interpretation of term specificity and its application in retrieval”에서 DF의 역인 IDF(Inverse Document Frequency)를 최초로 고안했다. IDF는 총 문서가 커질수록 기하급수적으로 커지므로, log를 사용한다.

$$IDF_{i,j} = \log \frac{|D|}{|d_j \in D : t_j \in d_j|}$$

TF와 IDF를 이용하여, TF-IDF는 다음과 같이 계산한다.

$$TF-IDF = TF \times IDF$$

정보검색 등의 과정에서 가중치로 사용되는 TF-IDF는 주어진 문서 군의 특정 문서 내에 단어의 중요 정도를 나타내는 통계적 수치를 의미한다 (전병국, 안현철, 2015). 또한, 어떤 단어가 특정 문서에 얼마나 집중적으로 출현했는지를 파악함과 동시에, 흔한 단어는 상대적으로 중요도가 낮다는 점을 반영한 지표이다.

본 연구는 OECD 메뉴얼이라는 1개의 문서를 대상으로 하고 있으나, 매뉴얼 별 챕터가 독립적인 형태를 띠고 있으므로 하나의 챕터를 하나의 문서로 판단하여서 분석하였다. OECD 매뉴얼 내에 챕터는 5페이지 내외이며, 10페이지가 넘는 매뉴얼은 연구자의 판단에 따라 세부 챕터를 하나의 문서로 지정하였다.

3.2.2 Word2Vec

Mikolov et al. (2013)에 의해 최초로 제안된 방법인 Word2Vec는 분포 가설에 기반하여 의미상 유사한 단어를 동시 등장 정보 (Words of Co-occurrence)를 사용하여 근거리 내에 벡터화시켜 단어간 의미나 관계성을 추론하는 방법이다. Word2Vec분석은 비슷한 의미의 단어는 비슷한 문맥에서 발생한다는 점을 근거로 단어 벡터를 공간에 임베딩하여 실시한다.

Word2Vec은 크게 CBOW(Continuous Bag of Words)와 Skip-Gram으로 나뉠 수 있다. CBOW는 주변 단어들을 가지고 중심 단어를 예측하는 방식으로 학습하고, Skip-Gram은 반대로 중심 단어를 가지고 주변 단어를 예측하는 방식으로 학습한다. Skip-Gram은 단어 하나로 여러 단어를 예측해야 하므로 정확도가 낮아 보이지만, CBOW보다 중심단어의 학습기회가 많이 주어지기 때문에 상대적으로 더 좋은 결과를 보여주는 장점이 있다(Jang & Kim, 2019). 이에, 본 연구는 Skip-Gram 방식을 적용하였다.

벡터차원의 경우, 영어 문서를 적용한 기존 연구를 참고하여 100차원으로 적용하였으며, 빈도수가 6이하의 단어는 제외하였다 (Jang & Kim, 2019; Nawangsari et al., 2019).

Word2Vec로부터 OECD 매뉴얼간 유사하게 사용한 단어와 차별성있게 사용한 단어를 파악하기 위해서는 벡터공간 내 거리를 파악해야 한다. 본 연구에서는 벡터공간 내 거리와 관련한 지표 중, 대중적으로 사용되는 코사인 유사도(Cosine Similarity)를 적용하였다 (윤여일 등, 2019; 정예림 등, 2020; Levy & Dagan, 2015). 코사인 유사도는 두 벡터간 코사인 각도를 이용하며, 방향이 완전히 동일한 경우 1, 완전히 반대의 방향을 가지면 -1을 가지게 된다. 두 벡터 A와 B에 대해서 코사인 유사도는 식으로 표현하면 다음과 같다 (정예림 등, 2020).

$$similarity = \cos(\theta) = \frac{A \cdot B}{\| A \| \ \| B \|}$$
$$= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

본 연구에서는 OECD 매뉴얼 2판과 3판을 대상으로 Word2Vec 분석을 수행하여, 코사인 유사도가 높은 단어를 판별로 의미가 유사하게 사용한 단어로 정의하고, 코사인 유사도가 낮은 단어를 판별로 차별성 있게 사용한 단어로 정의한다.

3.2.3 매트릭스 분석

2판과 3판 사이에 특정 단어의 연속성을 고려할 때, TF-IDF 순위가 급상승하였다면, 그 단어는 2판에 비해 3판에서 대표성이 높아졌으며, 코사인 유사도가 낮다면 그 단어는 2판에 비해 3판에서 의미가 크게 변화하였다고 볼 수 있다. 우리는 이를 동시에 고려할 때 좀 더 유의미한 결과를 도출할 것으로 판단하였다. 즉, 의미의 변화가 크고, 단어의 대표성이 높은 단어를 지침서별 정보속성 파악에 핵심으로 판단하였다는 점이다. 이에, 2X2 매트릭스를 제안한다 (<표 5> 참조).

우리는 4장에서 각 사분면에 어떤 단어들이 있는지를 도출하고, 5장에서 1사분면에 있는 단어를 대상으로 코사인 유사도가 높은 단어를 도출해 2판과 3판에서 정보속성이 어떻게 다른지를 파악한 후 시사점을 도출한다.

〈표 5〉 매트릭스 분석

TF-IDF 순위변화 \ 코사인 유사도	높음	낮음
높음	(2사분면) TF-IDF 순위변화 및 코사인 유사도가 높음 → 대표성이 높아지지만, 의미는 크게 변화없는 단어 집단	(1사분면) TF-IDF 순위변화가 높고 코사인 유사도가 낮음 → 대표성이 높아지고 의미도 크게 변화한 단어 집단
낮음	(3사분면) TF-IDF 순위변화 낮고 코사인 유사도가 높음 → 대표성이 낮아지고 의미도 크게 변화없는 단어 집단	(4사분면) TF-IDF 순위변화 및 코사인 유사도가 낮음 → 대표성이 낮아지지만, 의미는 크게 변화한 단어 집단

4. 분석결과

4.1 대표 단어 추출

명사 및 형용사 형태만으로 정제하여 OECD 매뉴얼 2판은 총 3,238개의 단어 및 총 빈도수 36,297회, 3판은 총 3,373개의 단어 및 총 빈도수 34,141회를 추출하였다. 본 연구는 차별적 특성의 단어를 추출하기 위해, TF 및 IDF 분석을 개별적으로 실시한 다음, TF-IDF 값을 도출하였으며, 순위가 상승한 단어와 하락한 단어를 추가로 조사하였다.

첫 번째, OECD 매뉴얼 판별 TF값이 높은 단어를 정리하면 <표 6>과 같다. OECD 매뉴얼 2판에서 가장 많이 등장한 단어는 ‘accident(0.0173)’이고, ‘safety(0.016)’, ‘public(0.0143)’, ‘installation(0.0138)’, ‘information(0.0118)’ 등의 순이다. OECD 매뉴얼 3판에서 가장 많이 등장한 단어는 ‘accident(0.0191)’이었고, 그 다음으로는 ‘safety(0.0164)’, ‘public(0.012)’, ‘installation(0.0112)’, ‘chemical(0.0107)’ 등의 순이다.

〈표 6〉 OECD 매뉴얼 판별 TF 분석

2판		3판	
순위	단어(TF)	순위	단어(TF)
1	accident(0.0173)	1	accident(0.0191)
2	safety(0.016)	2	safety(0.0164)
3	public(0.0143)	3	public(0.012)
4	installation(0.0138)	4	installation(0.0112)
5	information(0.0118)	5	chemical(0.0107)
6	authority(0.0115)	6	authority(0.0099)
7	emergency(0.0088)	7	management(0.0098)
8	management(0.0084)	8	risk(0.0088)
9	substance(0.0083)	9	emergency(0.0086)
10	chemical(0.007)	10	information(0.0083)
11	appropriate(0.0067)	11	substance(0.0068)
12	response(0.0065)	12	response(0.0067)
12	risk(0.0065)	13	ensure(0.0062)
14	principle(0.0064)	14	system(0.0058)
15	enterprise(0.0063)	15	process(0.0057)
16	including(0.0062)	16	including(0.0057)
17	prevention(0.0056)	17	appropriate(0.0054)
18	country(0.0056)	18	plan(0.005)
19	ensure(0.0055)	19	prevention(0.005)
20	guiding(0.0051)	20	co-operation(0.0046)
21	planning(0.0049)	21	principle(0.0045)
22	related(0.0049)	22	hazard(0.0045)
23	employee(0.0048)	23	planning(0.0041)
24	industry(0.0048)	23	use(0.0041)
25	oecd(0.0047)	23	industry(0.0041)
26	system(0.0045)	26	procedure(0.004)
27	technology(0.0042)	27	change(0.004)
28	use(0.0041)	27	employee(0.004)
28	plan(0.0041)	29	level(0.0038)
30	example(0.004)	29	health(0.0038)
31	site(0.004)	31	enterprise(0.0035)
32	co-operation(0.0039)	32	activity(0.0035)
32	community(0.0039)	33	assessment(0.0035)
34	procedure(0.0039)	33	effect(0.0035)
35	relevant(0.0038)	35	event(0.0034)
36	process(0.0038)	36	example(0.0033)
37	responsibility(0.0036)	37	oecd(0.0033)
38	level(0.0036)	38	responsibility(0.0032)
38	health(0.0036)	38	international(0.0032)
40	event(0.0035)	40	facility(0.0032)

<표 7>은 OECD 매뉴얼 판별 IDF값을 내림차순으로 정리한 결과이다.

OECD 매뉴얼 2판에서 IDF 값이 가장 높은 단어는 ‘change(0.7655)’이고, 그 다음으로 ‘facility(0.6257)’, ‘assessment(0.5831)’, ‘plan(0.5423)’, ‘planning (0.5031), responsibility(0.5031) 등의 순이다. 반면, OECD 매뉴얼 3판에서 IDF 값이 가장 높은 단어는 ‘employee(0.499)’이고, 그 다음으로 ‘planning (0.4418)’, ‘change(0.4418)’, ‘procedure(0.3878), enterprise(0.3878), assessment(0.3878)’ 순이다. 또한 2판 5개의 단어인 ‘accident’, ‘chemical’, ‘prevention’, ‘principle’, ‘oecd’, 3판은 6개의 단어인 ‘accident’, ‘public’ ‘chemical’, ‘authority’, ‘prevention’, ‘principle’, ‘oecd’는 IDF값이 0로서 모든 문서에 등장했음을 의미한다.

〈표 7〉 OECD 매뉴얼 판별 IDF 분석

2 판		3 판	
순위	단어(IDF)	순위	단어(IDF)
1	change(0.7655)	1	employee(0.499)
2	facility(0.6257)	2	planning(0.4418)
3	assessment(0.5831)	2	change(0.4418)
4	plan(0.5423)	4	procedure(0.3878)
5	planning(0.5031)	4	enterprise(0.3878)
5	responsibility(0.5031)	4	assessment(0.3878)
5	international(0.5031)	7	plan(0.3365)
8	system(0.4654)	7	facility(0.3365)
8	hazard(0.4654)	9	use(0.2877)
8	employee(0.4654)	9	responsibility(0.2877)
11	industry(0.429)	9	international(0.2877)
11	effect(0.429)	12	emergency(0.2412)
13	procedure(0.36)	12	hazard(0.2412)
13	level(0.36)	12	level(0.2412)
13	enterprise(0.36)	12	activity(0.2412)
16	risk(0.2955)	16	system(0.1967)
16	use(0.2955)	16	effect(0.1967)
16	health(0.2955)	16	event(0.1967)
16	event(0.2955)	16	example(0.1967)

2 판		3 판	
순위	단어(IDF)	순위	단어(IDF)
20	emergency(0.2647)	20	safety(0.1542)
20	process(0.2647)	20	risk(0.1542)
20	operation(0.2647)	20	response(0.1542)
23	substance(0.2348)	20	appropriate(0.1542)
23	ensure(0.2348)	24	installation(0.1133)
25	safety(0.2059)	24	substance(0.1133)
25	management(0.2059)	24	operation(0.1133)
27	response(0.1777)	24	industry(0.1133)
28	activity(0.1503)	28	management(0.0741)
28	example(0.1503)	28	ensure(0.0741)
30	public(0.1236)	28	including(0.0741)
30	appropriate(0.1236)	28	health(0.0741)
32	installation(0.0976)	32	information(0.0364)
32	authority(0.0976)	32	process(0.0364)
32	information(0.0976)	34	accident(0)
32	including(0.0976)	34	public(0)
36	accident(0)	34	chemical(0)
36	chemical(0)	34	authority(0)
36	prevention(0)	34	prevention(0)
36	principle(0)	34	principle(0)
36	oecd(0)	34	oecd(0)

TF 및 IDF 분석결과를 바탕으로 TF-IDF 값을 판별로 구분하여 <표 8>에 나타냈다. 빈도값이 소수인 단어의 TF-IDF가 높게 나오는 현상을 방지하고자, 3판의 TF값이 높은 40개의 단어를 기준으로 하였다. 또한, TF-IDF 값이 0인 경우, 순위가 다르더라도 동일한 순위로 판단하여 순위가 유지된 것으로 계산하였다.

2판의 결과를 살펴보면, ‘safety’, ‘planning’, ‘emergency’, ‘enterprise’, ‘employee’, ‘plan’, ‘system’, ‘industry’, ‘assessment’, ‘substance’ 등의 단어가 높은 중요도가 보임을 알 수 있다. 다음 3판의 결과를 살펴보면, ‘safety’, ‘emergency’, ‘employee’, ‘planning’, ‘change’, ‘plan’, ‘procedure’, ‘enterprise’, ‘risk’, ‘assessment’ 등의 단어가 높은 중요도가 보임을 알 수 있다.

세부적인 분석결과는 다음과 같다. 2판에서는 ‘safety’의 TF-IDF 값이 119.6으로서 가장 높았으며, 그다음 순위의 단어인 ‘planning(89.6)’에 비해 33.5% 크다. 이는 ‘safety’가 ‘planning’에 비해 문서내 출현 빈도(TF)와 몇 개의 문서에 나타났는지를 역순한 값(IDF)의 곱이 33.5% 크다는 것을 의미한다. 그 다음으로는 TF-IDF값이 80대인 단어는 ‘planning(89.6)’, ‘emergency(84.2)’, ‘enterprise(81.7)’ ‘employee(81.4)’, ‘plan(80.3)’이며, 70대 4개, 60대 4개, 50대 2개 등이다. 3판에서는 ‘safety’가 의 TF-IDF 값이 86.5로 가장 크며, 70대인 단어는 1개이며, 60대 2개, 50대 3개 등이다.

3판에서 TF가 높은 40개의 단어를 대상으로 2판에 비해 TF-IDF 값과 순위가 어떻게 변화했는지 살펴보았다. 본 연구에서는 순위를 우선적으로 비교하였으며, TF-IDF 값을 보완적으로 적용하여 체계적으로 비교하고자 하였다.

순위 관점에서 살펴본 결과 순위가 변동없는 단어는 8개, 순위가 상승한 단어는 16개, 순위가 하락한 단어는 16개이다.

순위 변동이 없으며, TF-IDF 값이 0이 아닌 단어는 ‘safety(1위)’, ‘plan(6위)’, ‘facility(15위)’이며, ‘safety’는 TF가 2판 및 3판 모두 2번째로 높은 단어로 큰 의미변화가 없이 여전히 중요한 단어로 판단된다. ‘safety’가 2판과 3판 모두 가장 높은 값이었으나, 2판에서는 119.6으로 3판의 86.5에 비해 38.3% 크다. 이는 2판에서 ‘safety’가 3판에 비해 문서내 출현 빈도(TF)와 몇 개의 문서에 나타났는지를 역순한 값(IDF)의 곱이 38.3% 크다는 것을 의미한다. ‘plan’과 ‘facility’는 TF에서는 3판이 높았지만, IDF는 2판이 높아서 순위가 동일한 것으로 보인다.

〈표 8〉 OECD 매뉴얼 판별 TF-IDF 분석

2판		3판		순위
순위	단어(TF-IDF)	순위	단어(TF-IDF)	
1	safety(119.6)	1	safety(86.5)	유지
3	emergency(84.2)	2	emergency(70.4)	상승
5	employee(81.4)	3	employee(67.4)	상승
2	planning(89.6)	4	planning(61.9)	하락
25	change(39.0)	5	change(59.6)	상승
6	plan(80.3)	6	plan(57.9)	유지
16	procedure(50.8)	7	procedure(53.5)	상승
4	enterprise(81.7)	8	enterprise(46.9)	하락
11	risk(69.7)	9	risk(46.4)	상승
9	assessment(72.3)	10	assessment(45.8)	하락
17	installation(48.8)	11	installation(43.2)	상승
21	use(43.7)	12	use(40.3)	상승
7	system(75.9)	13	system(38.8)	하락
31	hazard(32.1)	14	hazard(36.9)	상승
15	facility(56.9)	15	facility(36.3)	유지
22	response(41.9)	16	response(35.3)	상승
12	responsibility(66.4)	17	responsibility(31.6)	하락
26	international(38.7)	17	international(31.6)	상승
18	level(47.2)	19	level(31.6)	하락
35	activity(17.9)	20	activity(28.9)	상승
32	appropriate(30.2)	21	appropriate(28.4)	상승
10	substance(70.5)	22	substance(26.3)	하락
14	management(63)	23	management(24.8)	하락
20	effect(45.5)	24	effect(23.2)	하락
29	event(37.2)	25	event(22.6)	상승
34	example(21.9)	26	example(22.2)	상승
28	operation(37.6)	27	operation(17.9)	상승
8	industry(74.2)	28	industry(15.9)	하락
19	ensure(46.7)	29	ensure(15.8)	하락
33	including(22.1)	30	including(14.5)	상승
23	information(41.8)	31	information(10.3)	하락
27	health(38.7)	32	health(9.7)	하락
30	process(36.3)	33	process(7.1)	하락
13	public(64)	34	public(0)	하락
24	authority(40.6)	34	authority(0)	하락
36	accident(0)	34	accident(0)	유지
36	chemical(0)	34	chemical(0)	유지
36	principle(0)	34	principle(0)	유지
36	prevention(0)	34	prevention(0)	유지
36	oecd(0)	34	oecd(0)	유지
순위가 변동 없는 단어(8개) 순위가 상승한 단어(16개) 순위가 하락한 단어(16개)				

OECD 매뉴얼 3판이 2판과 비교하여 의미있게 변화된 점을 파악하기 위해서는 순위가 하락 및 상승한 단어를 자세히 살펴볼 필요가 있다. 이에 순위가 상승한 단어 16개, 하락한 단어 16개를 대상으로 순위가 어느정도 상승 및 하락했는지를 추가로 분석하였다 (<표 9> 참조). 또한 TF-IDF 값의 변화를 보완적으로 적용하여 체계적으로 비교하고자 하였다.

〈표 9〉 단어별 TF-IDF 순위 및 값 변화

상승한 단어			하락한 단어		
단어	상승폭		단어	하락폭	
	순위(단계)	값		순위(단계)	값
change	20	20.6	public	21	-64.0
hazard	17	4.8	industry	20	-58.4
activity	15	11.1	substance	12	-44.2
appropriate	11	-1.8	ensure	10	-40.6
procedure	9	-7.1	authority	10	-30.9
use	9	2.8	management	9	-38.2
international	9	-3.5	information	8	-31.5
example	8	0.3	system	6	-37.1
installation	6	-5.6	responsibility	5	-29.0
response	6	-6.6	health	5	-34.8
event	4	-14.6	enterprise	4	-22.3
including	3	-7.6	effect	4	-34.8
employee	2	-14.1	process	3	-29.1
risk	2	-23.3	planning	2	-27.7
emergency	1	-13.8	assessment	1	-26.6
operation	1	-19.7	level	1	-15.6

순위가 가장 많이 상승한 단어는 ‘change’로 총 20단계가 상승하였으며, TF-IDF 값도 39.0(2판)에서 59.6(3판)으로 증가하였다. 그리고 화학물질과 관련한 단어인 ‘hazard’이 두 번째로 높은 상승을 하였으며, 절차와 관련한 단어인 ‘activity’,

‘appropriate’, ‘procedure’, ‘use’ 및 국가적인 관점의 ‘international’이 순위가 상승함을 파악하였다. 하지만 TF-IDF 값에 있어서는 ‘change’, ‘activity’, ‘hazard’, ‘use’, ‘emample’의 5개 단어만 증가하였으며, 값의 변화와 순위의 변화 유사성을 찾기 힘들었다. 반면, ‘public’과 ‘industry’는 순위가 20단계 이상 하락하였으며, ‘substance’, ‘ensure’, ‘authority’ 등도 순위가 10단계 이상 하락하였다. 또한 TF-IDF 값에도 20개 단어 모두가 감소하였으며, 상대적으로 순위변화와 유사한 형태를 띄고 있다.

4.2 차별적 특성의 단어 추출

3판에서 TF가 높은 40개의 단어를 대상으로 코사인 유사도 값을 구하였다 (<표 10> 참조). 코사인 유사도가 낮을수록 판별 단어가 차별적으로 사용되었으므로 코사인 유사도 값을 오름차순으로 정렬하였다.

2판과3판에서 동일하게 쓰인 단어 중, ‘information’의 코사인 유사도가 －0.2764로 가장 낮았으며, 이는 2판과 3판의 ‘information’이라는 단어가 차별적으로 사용되었음을 의미한다.

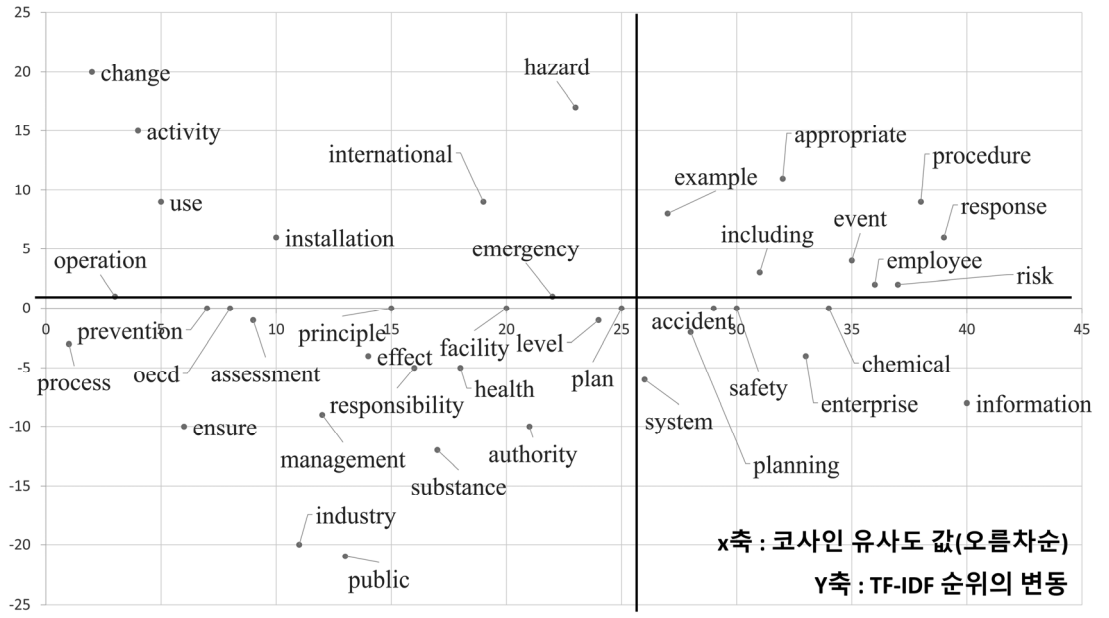
그 다음은‘response(-0.1955)’, ‘procedure(-0.1595)’, ‘risk(-0.1451)’, ‘employee(-0.1378)’, ‘event(-0.1375)’ 및 chemical(-0.1354) 순이다. 이 단어들은 2판과 3판에서 그 쓰임이 다르다는 것을 의미한다.

반면 ‘process’는 코사인 유사도가 0.2962로 49개의 단어 중 가장 높은 코사인 유사도를 기록하였으며, 이는 2판과 3판의 ‘process’ 라는 단어가 유사하게 사용되었음을 의미한다.

그 다음은 ‘change(0.2354)’, ‘operation(0.2298)’, ‘activity(0.1657)’, ‘use(0.1056)’, ‘ensure(0.0985)’ 순이다. 이 단어들은 2판과 3판에서 그 쓰임이 유사하다는 것을 의미한다.

〈표 10〉 판별 단어 유사성(Word2Vec)

순위	단어		코사인 유사도
	2판	3판	
1	information	information	-0.2764
2	response	response	-0.1955
3	procedure	procedure	-0.1595
4	risk	risk	-0.1451
5	employee	employee	-0.1378
6	event	event	-0.1375
7	chemical	chemical	-0.1354
8	enterprise	enterprise	-0.0907
9	appropriate	appropriate	-0.0844
10	including	including	-0.0810
11	safety	safety	-0.0784
12	accident	accident	-0.0540
13	planning	planning	-0.0447
14	example	example	-0.0447
15	system	system	-0.0170
16	plan	plan	-0.0092
17	level	level	-0.0078
18	hazard	hazard	-0.0021
19	emergency	emergency	0.0122
20	authority	authority	0.0201
21	facility	facility	0.0202
22	international	international	0.0214
23	health	health	0.0289
24	substance	substance	0.0347
25	responsibility	responsibility	0.0399
26	principle	principle	0.0439
27	effect	effect	0.0528
28	public	public	0.0546
29	management	management	0.0571
30	industry	industry	0.0618
31	installation	installation	0.0651
32	assessment	assessment	0.0658
33	oecd	oecd	0.0773
34	prevention	prevention	0.0867
35	ensure	ensure	0.0985
36	use	use	0.1056
37	activity	activity	0.1657
38	operation	operation	0.2298
39	change	change	0.2354
40	process	process	0.2962



〈그림 4〉 매트릭스 분석 결과

4.3 주요 정보 추출

본 연구는 2판과 3판의 OECD 지침서 내, 추출된 단어를 대상으로 어떤 단어가 중요 정보인지 파악하는 연구이다. 이를 위해서 TF-IDF를 통해 대표 단어를 추출하였으며, 어떤 단어가 판별로 중요한지를 <표 9>을 통해 도출하였다. 또한, <표 10>의 Word2Vec 분석을 통해 단어별로 의미가 유사하거나 차별화된 단어를 도출하였다. 우리는 <표 9>와 <표 10>의 결과를 <그림 4>와 같이 매트릭스화 하였다. X축은 코사인 유사도 순위이며, Y축은 TF-IDF 순위 변화이다. 해석을 쉽게 하도록, X축은 오름차순을 기준으로 정렬하였다. 즉, 코사인 유사도가 가장 낮은 단어인 'process'를 순위 1위로, 가장 높은 단어인 'information'을 40위로 설정하였다는 뜻이다. 또한, 기준선은 X축은 25, Y축은 0으로 설정하였다.

단어의 대표성이 3판에서 증가함과 동시에 의미도 크게 변화한 정보와 관련한 단어는 1사분면에 있으며 총 8개가 관찰되었다. 'appropriate'는 TF-IDF 순위가 11단계 상승하여 4번째로 높은 상승을 기록하였으며, 코사인 유사도값은 -0.0844로 9번째로 낮은 값을 기록하였다. 그 외에도 'response', 'procedure', 'example' 등이 TF-IDF 순위가 높아졌으며, 코사인 유사도 값은 낮았다.

단어의 대표성은 3판에서 증가하였지만, 의미 변화가 크지 않은 정보와 관련한 단어는 2사분면에 있으며, 총 8개가 관찰되었다. 'change'는 TF-IDF 순위가 20단계 상승하여 가장 상승폭이 컸지만, 코사인 유사도는 0.2354로 39위를 기록하였다. 그 외에도 'hazard', 'activity', 'international', 'use', 'installation' 등이 TF-IDF 순위가 높아졌으며, 코사인 유사도 값도 높았다.

단어의 대표성이 3판에서 유지되거나 감소하였

으나, 의미가 크게 변화한 정보와 관련한 단어는 4사분면에 있으며, 총 7개가 관찰되었다. ‘information’은 TF-IDF가 8단계 하락하여, 2판에 비해 3판에서 대표성이 감소하였다. 그 외에도 ‘safety’, ‘chemical’, ‘enterprise’ 등이 TF-IDF 순위가 유지 또는 감소하였으나, 코사인 유사도 값은 낮았다.

단어의 대표성이 3판에서 유지 혹은 감소하였고, 의미변화도 크지 않은 정보와 관련한 단어는 17개가 관찰되었다. ‘industry’는 TF-IDF 순위가 20단계 하락하여 두번째로 높은 하락세를 기록하였으며, 코사인 유사도값도 0.0618로 30번째로 유사도가 낮았다. 그 외에도 ‘public’, ‘substance’, ‘management’, ‘ensure’, ‘responsibility’ 등이 TF-IDF 순위가 유지 또는 감소하였으며, 코사인 유사도 값은 높았다.

5. 토론 및 결론

화학물질의 양과 종류가 급속하게 증가함에 따라 화학물질 사고는 일부 기업만의 문제가 아닌 국가차원의 고민이 되었다. 2023년 6월 개정된 OECD 지침서는 이를 염두에 두고 수행되었다. 우리는 OECD 지침서를 대상으로 중요한 정보가 무엇인지 파악하기 위해, 이론적 조사와 함께 정량적 분석인 TF-IDF 분석과 Word2Vec 분석을 수행하고 이를 매트릭스화 하여 체계화하였다.

매트릭스 분석 결과, 1사분면에는 8개의 단어가 관찰되었다. 1사분면에 있는 단어는 코사인 유사도 순위가 낮아졌으며, TF-IDF 순위가 증가하였다. X축인 코사인 유사도 순위가 낮다는 점은 2판과 3판 내의 단어의 의미의 변화가 크다는 점을 의미한다. 즉, 같은 단어임에도 불구하고, 문맥상 다른 의미로 사용되었다고 볼 수 있다. 또한, Y축인 TF-IDF 순위가 증가하였다는 점은 TF-IDF 값을 동시에 고려하지 않는 한계가 있기는 하지만, 일반적으로 해당 단어가 더 많이 노출되거나, 더 적은 문서에 집중적으로 나타났다고 볼 수 있다. 코사인 유사도 순위가 낮고, TF-IDF 순위가 증가한 단어는 2판과 3판의 정보속성 분석에 있어 주요한 의미가 있다고 보았다. 반면, 2사분면에 있는 8개의 단어는 Y축인 TF-IDF 순위는 증가하였으나, 코사인 유사도 순위가 증가하였으며, 4사분면에 있는 7개의 단어는 코사인 유사도 순위는 낮아졌지만, TF-IDF 순위는 하락하였다. 이 단어들은 2판과 3판의 정보속성 분석에 있어 보완적인 의미를 지닌다고 볼 수 있다.

이에, 3판에서 코사인 유사도 순위가 낮아 의미의 변화가 크며, TF-IDF 순위가 증가하여 대표성이 높아진 8개의 단어를 대상으로 코사인 유사도 값이 큰 근접 단어를 추출하였다 (<표 11> 참조). 연구자가 판단하여 해석에 불필요한 단어 (예: sure, whose 등)는 제외하였으며, 2판과 3판에 동시에 나온 단어는 가능하면 제외하였다.

〈표 11〉 주요 단어별 코사인 유사도

example		including	
2판	3판	2판	3판
cost (0.7276)	disposal (0.7815)	environment (0.7016)	photograph (0.7134)
hospital (0.7253)	production (0.7697)	initial (0.6461)	picture (0.6749)
overwork (0.714)	building (0.7562)	owner (0.615)	manager (0.649)
event		appropriate	
2판	3판	2판	3판
property (0.7363)	workforce (0.848)	revising (0.7724)	action (0.7474)
security (0.7271)	potentially (0.8272)	measure (0.767)	adopt (0.7354)
formed (0.7069)	step (0.7208)	statutory (0.7529)	measure (0.7184)

procedure		response	
2판	3판	2판	3판
verification (0.9561)	manufacturing (0.8736)	preparedness (0.9408)	preparedness (0.9138)
certification (0.9194)	permanent (0.8479)	prevention (0.8752)	requested (0.8514)
conceptual (0.9054)	temporary (0.8227)	accident (0.7838)	hidden (0.8387)
employee		risk	
2판	3판	2판	3판
refusing (0.8622)	behalf (0.7847)	estimated (0.8271)	influenced (0.7225)
complain (0.8574)	representative (0.7792)	identification (0.7937)	uncertainty (0.7179)
reason (0.846)	motivation (0.7011)	hazard (0.7723)	conflict (0.672)

‘example’과 관련하여 2판에서 코사인 유사도가 높은 단어는 ‘cost’, ‘hospital’ 및 ‘overwork’이며, 3판에서는 ‘disposal’, ‘production’ 및 ‘building’가 높았다. 즉, 2판에서는 화학물질 사고로 인한 비용과 병원치료 및 그 원인이 되는 과로에 관한 내용이 사례로써 주로 언급되었다면, 3판에서는 화학물질 생산과 생산이 이루어지는 건물에 관한 사례가 많았으며, 화학물질 사고이후 처리에 관한 내용이 많았다고 볼 수 있다.

‘including’과 관련하여 2판에서 코사인 유사도가 높은 단어는 ‘environment’, ‘initial’ 및 ‘owner’이며, 3판에서는 ‘photograph’, ‘picture’ 및 ‘manager’가 높았다. 2판에서는 환경과 관련한 개념적인 내용과 초기 단계 예방, 대응 등에 관한 내용을 많이 포함하고 있지만, 3판에서는 사진과 같이 증거기반의 내용을 포함하고 있었다. 그리고 2판에서는 소유주, 3판에서는 관리자가 유사도가 높다는 것은 3판으로 오면서 실무적인 내용에 관한 관심과 대응이 강조됨을 추측할 수 있다.

‘event’와 관련하여 코사인 유사도가 높은 단어는 2판에서는 ‘property’, ‘security’ 및 ‘formed’이고, 3판에서는 ‘workforce’, ‘potentially’ 및 ‘step’이다. 이는 2판에서는 화학물질 사고로 인한 재산손실과 안전강화와 관련한 이벤트가 많이 언급된 반면, 3판에서는 화학물질 사고 예방과 대응을 위한 노동자의 역할을 단계별 이벤트로 제시하는 경우가 많음을 알 수 있다.

‘appropriate’에 대응하여 관련성이 높은 단어는 2판에서는 ‘revising’, ‘measure’ 및 ‘statutory’이고, 3판에서는 ‘action’, ‘adopt’ 및 ‘measure’이다. 이는 2판에서는 화학물질 사고시 적절한 측정방법을 제안하고, 기존내용을 수정하거나 법제화하는 방안이 주인 반면, 3판에서는 실제 행동으로 옮기고 사업장 등에서 적용하는 데 집중함을 확인할 수 있다.

‘procedure’와 관련성이 높은 단어는 2판에서는 ‘verification’, ‘certification’ 및 ‘conceptual’이고, 3판에서는 ‘manufacturing’, ‘permanant’ 및 ‘temporary’이다. 이는 2판에서는 개념적으로 화학물질 사고에 대해 정의하고, 증명하는 과정에 집중하였다면, 3판에서는 실제 제조과정에서 일시적으로 해야 할 내용과 영구적으로 해야 하는 내용을 담았다는 점에서 차이가 있다.

‘response’에 대응하는 단어로는 2판에서는 ‘preparedness’, ‘prevention’ 및 ‘accident’이고, 3판에서는 ‘preparedness’, ‘requested’ 및 ‘hidden’이다. 이는 2판에서는 사고를 방지하는 데 있어 대응책을 마련하는 데 집중하였다면, 3판에서는 기존에는 존재하지 않았던 숨겨진 요구에 대응하는 방안에 대해서 주로 작성되었음을 확인할 수 있다.

‘employee’와 관련성이 높은 단어는 2판에서는 ‘refusing’, ‘complain’ 및 ‘reason’를 들 수 있고, 3판에서는 ‘behalf’, ‘representative’ 및 ‘motivation’를 들 수 있다. 이는 2판에서는 종업원들의 불만 사항의 원인을 파악하는 데 집중하였다면, 3판에서는 조직의 대표만큼 능동적으로 대처하도록 동기

부여 방안에 관한 내용이 주로 작성되었음을 파악할 수 있다.

‘risk’와 관련 높은 단어는 2판에서는 ‘estimated’, ‘identification’ 및 ‘hazard’이고, 3판에서는 ‘influenced’, ‘uncertainty’ 및 ‘conflict’이다. 이는 2판에서는 위험물질에 대한 측정과 확인에 대한 리스크가 주로 작성되었으며, 3판에서는 사회적 영향과 불확실성 그리고 갈등에 관한 리스크가 주로 논의되었음을 알 수 있다.

본 연구는 다음의 세 가지를 한계점으로 제시한다. 첫 번째는 OECD 메뉴얼에 대한 정성적 분석이 부족했다는 점이다. 본 연구에서는 배경조사에서 OECD 매뉴얼에 대해 설명하였으나, 정책적으로 어떤 의미를 지니고 있는지에 대한 심층분석이 부족했다고 판단한다. 텍스트 분석을 통해 얻어지는 결과는 단어 간의 관계만을 단순히 보여지는 것으로 단어간 연결과 빈도 등에 대한 배경조사가 필요할 것으로 판단된다. 이에, 향후 연구에서는 전문가 인터뷰 등을 병행한다면 더욱 의미있는 결과가 도출될 것으로 보인다. 두 번째는 단어별 관련성 있는 단어에 대한 해석에 있어 전체 메뉴얼을 전수조사하지 않았다는 점이다. 특정 단어와 관련성이 높은 단어는 코사인 유사도로 도출할 수 있었으나, 이 단어들이 문장내에서 실제로 어떻게 사용하는지는 표본조사로 도출하였다. 메뉴얼 전체에 대한 리뷰를 향후 수행한다면 좀 더 의미있는 결과를 도출할 것으로 기대된다. 마지막으로는 코사인 유사도 및 TF-IDF의 순위와 값 모두가 의미있음에도 불구하고, 매트릭스 분석에 있어서 코사인 유사도 및 TF-IDF 순위를 중심으로 하였다는 점이다. 추후 연구에서는 코사인 유사도 및 TF-IDF 순위와 값을 동시에 고려할 수 있는 방법론을 적용한다면 좀 더 정확한 결과가 나올 것으로 기대된다.

참고문헌(References)

[국내 문헌]

국립국어원, (n.d.). 국립국어원 표준국어대사전. 국립국어원. Retrieved January 30, 2023, from https://stdict.korean.go.kr/main/main.do

김대환. (2010). OECD DAC 가입과 KOICA 의 환경 및 기후변화 ODA 추진전략. *Journal of International Development Cooperation, 5(2)*, 10-31.

김민구, 김용우, 정태현, 김영민. (2022). Organic Light-Emitting Diodes 디스플레이 기술의 특허 동향과 기술적 가치에 관한 탐색적 연구. *지능정보연구, 28(4)*, 135-155.

김용진, 정주미, 최호철. (2022). 한국의 화학무기용 특정화학물질 관리체계 개선방안 연구: 한·중·일 3 국간 비교를 중심으로. *한국정책과학학회보, 26(2)*, 65-86.

김진솔, 신동훈, 김희웅. (2021). 비정형 빅데이터를 이용한 COVID-19 주요 이슈 분석. *지식경영연구, 22(2)*, 145-165.

박종서, 정성봉, 안찬기, 김연응. (2012). 위험물 표준화를 통한 운송 사고대응메뉴얼 개발. 한국산업인력공단, 서울.

박호연, 김경재. (2019). CNN-LSTM 조합모델을 이용한 영화리뷰 감성분석. 지능정보연구, 25(4), 141-154.

소방신문. (2023), 소방청, 2022년 국내 화학사고 218건…소방청, 희귀 화학물질 사고대응 능력 높인다. 소방신문. Retrieved January 30, 2023, from https://www.sobangnews.kr/news/articleView.html?idxno=19086

유은순, 최건희,김승훈. (2015). TF-IDF 와 소설 텍스트의 구조를 이용한 주제어 추출 연구. *한국컴퓨터정보학회논문지, 20(2)*, 121-129.

유지선, 정영진. (2014). 유해화학물질 유출의 사례

분석. *한국화재소방학회 논문지, 28(6),* 90-98.

윤여일, 고은정, 김남규. (2019). 주제 균형 지능형 텍스트 요약 기법. *지능정보연구, 25(2),* 141-166.

이동훈, & 김관호. (2018). Word2Vec 기반의 의미적 유사도를 고려한 웹사이트 키워드 선택 기법. *한국전자거래학회지, 23(2),* 83-96.

이성직, 김한준. (2009). TF-IDF 의 변형을 이용한 전자뉴스에서의 키워드 추출 기법. *한국전자거래학회지, 14(4)*, 59-73.

이재석, 최돈묵. (2015). 국가재난관리체제 관점의 화학사고 대응체계 개선방안에 관한 연구. *한국화재소방학회 논문지, 29(5)*, 73-78.

전병국, 안현철. (2015). 사용자 리뷰 마이닝을 결합한 협업 필터링 시스템: 스마트폰 앱 추천에의 응용. *지능정보연구, 21(2)*, 1-18.

정예림, 김지희, 유형선. (2020). Word2Vec 을 활용한 제품군별 시장규모 추정 방법에 관한 연구. *지능정보연구, 26(1)*, 1-21.

한국학술지인용색인. (n.d.). KCI 통합검색. 한국학술지인용색인. Retrieved January 31, 2023, from https://www.kci.go.kr

화학물질안전원, (n.d.). 화학물질종합정보시스템. Retrieved January 31, 2023, 화학물질안전원. from https://icis.me.go.kr

환경부 화학물질과/화학물질안전 TF. (2013, September 25). 구미 불화수소 누출사고, 화학안전의 교훈. 환경부. Retrieved January 31, 2023, from https://www.korea.kr/briefing/pressReleaseView.do?newsId=155919163

환경부 물관리정책실 수질관리과. (2022, March 24). 미규제 미량오염물질 촘촘한 조사, 먹는물 안전관리 강화. 환경부. Retrieved January 31, 2023, from https://www.korea.kr/briefing/pressReleaseView.do?newsId=156500636

[국외 문헌]

Bird, S. (2006, July). NLTK: the natural language toolkit. *In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (pp. 69-72).

CEFIC. (2023, February 27). Facts and Figures 2023. CEFIC. Retrieved August 31, 2023, from https://cefic.org/a-pillar-of-the-european-economy/facts-and-figures-of-the-european-chemical-industry/

Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one, 14(8)*, e0220976.

Kim, Y. J., & Lee, D. H. (2020). Technology convergence networks for flexible display application: A comparative analysis of latecomers and leaders. *Japan and the World Economy, 55,* 101025.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics, 3*, 211-225.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028.*

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Nawangsari, R. P., Kusumaningrum, R., & Wibowo, A. (2019). Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study. *Procedia Computer Science, 157,* 360-366.

OECD. (2003) *OECD Guiding Principles for Chemical Accident Prevention, Preparedness and Response - Second Edition.* Paris: OECD Publishing.

OECD. (2023) *OECD Guiding Principles for Chemical Accident Prevention, Preparedness and Response - Third Edition.* Paris: OECD Publishing.

Řehůřek, R., & Sojka, P. (2011). Gensim－statistical semantics in python. *Retrieved from genism. org.*

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28(1)*, 11-21.

Abstract

Comparative analysis of information attributes in chemical accident response systems through Unstructured Data: Spotlighting on the OECD Guidelines for Chemical Accident Prevention, Preparedness, and Response

YongJin Kim* · Chunghyun Do**

The importance of manuals is emphasized because chemical accidents require swift response and recovery, and often result in environmental pollution and casualties. In this regard, the OECD revised OECD Guidelines for the Prevention, Preparedness, and Response to Chemical Accidents (referred to as the OECD Guidelines), in June 2023. Moreover, while existing research primarily raises awareness about chemical accidents, highlighting the need for a system-wide response including laws, regulations, and manuals, it was difficult to find comparative research on the attributes of manuals. So, this paper aims to compare and analyze the second and third editions of the OECD Guidelines, in order to uncover the information attributes and implications of the revised version. Specifically, TF-IDF (Term Frequency-Inverse Document Frequency) was applied to understand which keywords have become more important, and Word2Vec was applied to identify keywords that were used similarly and those that were differentiated. Lastly, a 2X2 matrix was proposed, identifying the topics within each quadrant to provide a deeper comparison of the information attributes of the OECD Guidelines. This study offers a framework to help researchers understand information attributes. From a practical perspective, it appears valuable for the revision of standard manuals by domestic government agencies and corporations related to chemistry.

Key Words : Chemical accident, Chemical accident prevention, Comparative analysis, TF-IDF analysis, Word2Vec

Received : September 5, 2023 Revised : November 26, 2023 Accepted : November 28, 2023
Corresponding Author : Yongjin Kim

* Corresponding Author: Yongjin Kim
Office of R&D Planning, Korea Research Institute of Chemical Technology,
141 Gajeongro, Yuseong-gu, Daejeon 34114, Republic of Korea
Tel: +82-42-860-7988, Fax: +82-42-860-7289, E-mail: koine@krict.re.kr
** Office of R&D Planning, Korea Research Institute of Chemical Technology

저 자 소 개

김용진
현재 한국화학연구원 연구전략본부 연구기획실 실장(책임연구원)으로 재직 중이다. 한국과학기술원 기술경영전문대학원에서 박사학위를 취득하였다. 주요 관심분야로는 경영정보시스템, 사회 네트워크 분석, 빅데이터 분석 등이며, Scientometrics, Technology Analysis & Strategic Management, International Journal of Technology Management, Japan and the World Economy 등 다수의 국제학술지에 관련 논문을 게재하였다.

도충현
현재 한국화학연구원 연구전략본부 연구기획실 선임연구원으로 재직 중이다. 충남대학교 행정대학원에서 석사학위를 취득하였으며, 관심 연구 분야는 화학분야 정보조사 및 분석, 화학산업 R&D 전략연구, 산학연 R&D 협력 등이다.