Journal of the Korean Society of Safety, Vol. 33, No. 3, pp. 58-64, June 2018
 https://doi.org/10.14346/JKOSOS.2018.33.3.58

사고보고문서를 이용한 텍스트 기반 사고발생 유형 및 관계 분석

김범수 · 장성록 · 서용윤†

부경대학교 안전공학과

(2018. 4. 3. 접수 / 2018. 5. 22. 수정 / 2018. 6. 19. 채택)

Text Analytics for Classifying Types of Accident Occurrence Using Accident Report Documents

Beom Soo Kim · Seongrok Chang · Yongyoon Suh†

Department of Safety Engineering, Pukyong National University

(Received April 3, 2018 / Revised May 22, 2018 / Accepted June 19, 2018)

Abstract : Recently, a lot of accident report documents have accumulated in almost all of industries, including critical information of accidents. Accordingly, text data contained in accident report documents are considered useful information for understanding accident processes. However, there has been a lack of systematic approaches to analyzing accident report documents. In this respect, this paper aims at proposing text analytics approach to extracting critical information on accident processes. To be specific, major causes of the accident occurrence are classified based on text information contained in accident report documents by using both textmining and latent Dirichlet allocation (LDA) algorithms. The textmining algorithm is used to structure the document-term matrix and the LDA algorithm is applied to extract latent topics included in a lot of accident report documents. We extract ten topics of accidents as accident types and related keywords of accidents with respect to each accident type. The cause-and-effect diagram is then depicted as a tool for navigating processes of the accident occurrence by structuring causes extracted from LDA. Further, the trends of accidents are identified to explore patterns of accident occurrence in each of types. Three patterns of increasing to decreasing, decreasing to increasing, or only increasing are presented in the case of a chemical plant. The proposed approach helps safety managers systematically supervise the causes and processes of accidents through analysis of text information contained in accident report documents.

Key Words : accident analysis, text analytics, LDA(Latent Dirichlet Allocation), cause-and-effect diagram, chemical plant

1. 서 론

빈번히 발생하는 유사사고나 위험성이 큰 중대사고를 조사하고 예방하기 위하여 산업재해조사표의 작성이 법적 의무화되고 있다. 이에 따라 산업재해의 발생상황이나 원인 및 과정, 재발방지 계획 등의 사고보고문서 자료가 지속적으로 누적되고 관리되고 있다.

그러나 예방안전 및 사고조사를 위한 서술 기록이나 텍스트 자료가 산업재해조사표에 많이 포함되어 있음에도, 이 텍스트 자료를 이용한 체계적인 사고분석은 많이 이루어지고 있지 않다. 이로 인해 아직까지도 안전관리 실무에서의 재해정보 분석은 관리자가 직접 사고보고문서에 포함되어 있는 텍스트 정보를 개별적으로 확인하면서, 정성적으로 사고발생의 유형이나 원인을 해석하는 어려움이 있다[1-2]. 또한, 다양한 사고재해가 발생하면서 새로운 사고발생 원인이나 결과가 텍스트 형태로 작성되어 있음에도, 그 양이 많거나 작성체계가 명확하지 않아 사고발생 원인을 제대로 확인하거나 파악하지 못하는 문제점도 있다[3]. 결과적으로 사고보고문서의 정성적 분석은 안전관리자에게 시간적·업무적으로 버거우며, 실제 사고보고문서에 포함된 정보들을 제대로 파악하기 어렵다는 문제점이 발생하고 있다. 따라서 사고보고문서에 포함된 텍스트 정보를 정량적 · 과학적으로 분석하는 방법론을 적용하는 연구가 부족한 시점에서, 사고 텍스트 정보로부터 사고정보를 변환 및 도출하는 안전관리지원 연구가 필요하다[3-7].

산업재해조사표와 같은 사고보고문서의 텍스트 정보의 정량적 분석은 사고 프로세스를 이해하는데 도움

† Corresponding Author : Yongyoon Suh, Tel : +82-51-629-6467, E-mail : ysuh@pknu.ac.kr
Department of Safety Engineering, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 48513, Korea

을 줄 수 있다. 우선, 자연어 위주의 텍스트로 이루어진 정보를 기술적으로 처리하여, 재해통계에서 알 수 없는 사고요인들을 사고의 서술 혹은 텍스트를 통해 도출할 수 있다[4-6]. 또한, 텍스트 정보로 유의한 결과를 제시하기 위한 데이터 분석론이 활성화되면서, 사고보고문서를 보다 용이하게 분석이 가능하게 되었다. 데이터 과학의 발전으로 텍스트마이닝(textmining), 데이터마이닝(datamining), 확률모델(stochastic model) 등 다양한 데이터 처리 및 분석 알고리즘이 개발되면서, 과학적이고 체계적으로 분석하지 못했던 텍스트 데이터와 같은 비정형 데이터의 활용가치가 점점 높아지고 있다[7-8].

따라서 본 연구에서는 사고보고문서를 수집하여, 사고발생 원인과 과정을 도출하고 유형화하는 방법을 제시한다. 이를 위해 텍스트마이닝과 토픽 모델링(topic modeling) 기법 중 하나인 잠재 디리클레 할당(LDA: Latent Dirichlet Allocation) 알고리즘을 활용한다.

텍스트마이닝은 문서 형태의 정보를 컴퓨터에서 처리할 수 있는 데이터 형태(data format)로 변환하고, 문서에 포함된 텍스트 키워드 정보를 추출(keyword extraction)하기 위해 사용한다. 텍스트마이닝은 자연어 처리와 같은 기술적인 알고리즘 분야뿐만 아니라 특허분석, 의료정보 데이터 분석과 같은 응용분야에도 널리 활용되고 있다[6,9-10].

다음으로 LDA 알고리즘은 베이지안(Bayesian) 확률 알고리즘을 바탕으로, 데이터의 내용을 바탕으로 해당하는 주제 혹은 분야를 확률적으로 추론하는 기법이다[11]. 특히, 문서 데이터 분류에서 주로 활용되는 알고리즘으로, 문서-주제-단어 간의 관계에 따라 문서에 내포된 주제, 주제를 구성하는 단어들을 확률적으로 추론한다. 유사한 목적으로, 과거에 텍스트 기반 클러스터링을 주로 활용하였지만, 이는 데이터 간의 거리를 기준으로 유사성을 판단하기 때문에, 문서 내 같은 단어 혹은 키워드가 존재하지 않아도 동일한 클러스터에 분류되는 단점이 있다. 그러나 LDA는 문서 내에 포함된 단어의 포함 확률을 분석하기 때문에, 문서에 직접적으로 포함된 단어들의 조합으로 주제를 도출하게 된다. 이에 따라 주제의 의미를 파악하기 쉽기 때문에 문서 및 텍스트 분석에 널리 활용되고 있다[12-14].

이와 같은 장점에도 불구하고, 아직까지 사고조사보고문서를 분석하기 위해 LDA와 같은 문서에 특화된 방법론의 적용 연구는 많이 이루어지고 있지 않다[14]. 따라서 본 연구에서는 LDA 알고리즘을 통해, 사고보고문서 안에 포함되어 있는 단어들을 키워드로 추출하고, 현재 발생하고 있는 주요 사고발생유형을 제시한다. 또한, 사고발생 유형 및 키워드 간의 관계성을 나타내는 특성요인도(cause-and-effect diagram)를 작성하여, 전체적인 관계도를 작성할 것이다. 이는 안전관리자가 상황을 이해하기 쉬운 자연어 기반의 텍스트 정보를 추출했다는 점에서, 사고원인의 발생구조를 쉽게 파악할 수 있다. 마지막으로 시간에 따른 사고발생유형의 빈도수 변화에 따른 사고경향을 분석한다. 이 결과는 안전관리자로 하여금, 시간에 따라 관리해야할 주요사고유형을 결정하도록 지원하리라 기대된다.

2. 연구방법론

2.1. 데이터 수집 및 문제인식

사고 데이터는 울산 및 경기지역에 입지한 원료화학제품을 생산하는 외국계 기업의 화학 플랜트 1년 동안(2015.10.~2016.09.)의 사고보고문서를 수집하였다. 1년 동안 아차사고 및 공정사고를 포함한 사고보고 수는 총 127건이었으며, 각 사고보고문서는 사고 날짜, 내용, 유형 및 원인 분석과 대책 등을 포함하고 있다. 사고 발생 유형은 설비, 공정 및 휴먼에러 등으로 구분되어 있지만, 내용에 대한 구체적인 설명이나, 사고의 전반적인 주요원인들에 대한 체계적이고 정량적인 접근은 부족한 실정이다. 이를 위해 텍스트마이닝과 LDA를 활용하여 사고발생유형을 분류하고, 그 유형들의 관계를 토대로 특성요인도를 구조화한다. 또한, 각 유형들이 시간에 따라 변화하는 추세를 살펴보고, 유지보수의 필요성이 있거나 미래에 집중적으로 다루어야 할 사고발생유형을 제시한다.

2.2. 토픽 모델링 : 잠재디리클레할당(LDA)

토픽 모델링은 자연어 처리 기법 중 문서 분석에 특화된 방법으로, 많은 문서 안에서 나타나고 있는 잠재주제(latent topic)들을 도출하는 방법이다. 그 중에서도 가산자료의 이산다항-디리클레 분포를 이용하는 잠재디리클레할당(LDA) 알고리즘을 활용한다[11]. 디리클레 할당 분포는 각 확률변수의 값이 1이면서, 확률변수의 모양을 결정할 수 있는 형태모수를 가진 확률분포로, 단어나 문서 등 셀 수 있는 다양한 키워드 변수를 가산다항변수로 분석하는데 효과적으로 활용될 수 있다.

LDA는 베이지안 추론을 기반으로 하는 방법으로 확률적 계산 방식은 샘플링을 통한 사후확률의 최대화 과정을 통해 추정하게 된다. 이 때, 사후확률은 결국 특정 단어가 문서에 포함될 확률을 의미하며, 사후확률을 최대화하는 과정에서, 잠재변수인 주제들의 확률분포를 추정한다. Fig. 1을 참고하면, 문서-주제-단어

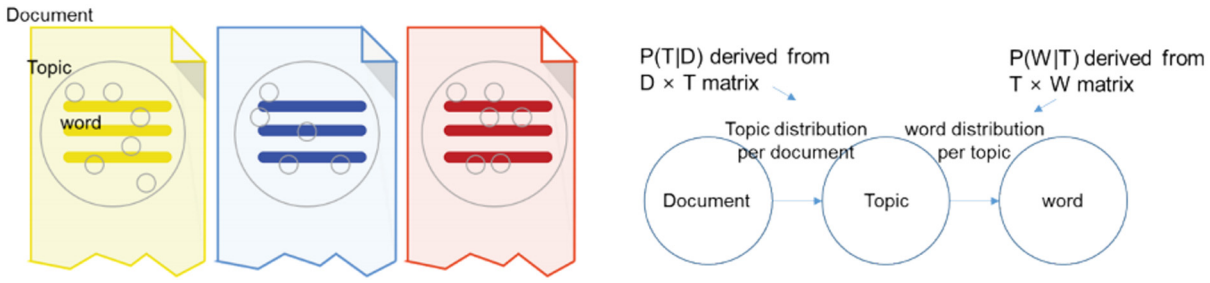

Fig. 1. Concept of LDA.

간의 포함관계와 베이지안 추정을 위한 사전-사후확률 관계를 확인할 수 있다. 문서 분석에서 활용하는 결과치는 문서 내 주제들의 확률분포와 주제를 구성하는 단어들의 확률분포들이다. 즉, 특정 주제에 포함될 단어들의 확률분포(P(W | T))와 문서에 포함될 주제들의 확률분포(P(T | D))의 결합확률과 조건부확률을 수렴 또는 최대화하여 추정하는 것이 LDA의 핵심이다.

LDA의 사후확률 추정식은 아래의 식 (1)과 같다[11].

$$p(\theta^{(d)},z,w|\alpha,\beta) = p(\theta^{(d)}|\alpha)*\sum_{n=1}^{N}p(z_n|\theta^{(d)})*p(w_n|z_n,\beta) \quad (1)$$

$\theta^{(d)}$: 문서 d에 포함된 주제들의 확률분포
z_n: 주제 n의 확률분포
w_n: 주제 n을 구성하는 단어집합의 확률분포
α: 문서 내에 존재하는 주제비율(사전확률)
β: 주제들을 구성하는 단어비율(사전확률)

세부적으로 분석해보면, 최종적으로 단어들로 이루어지는 주제들이 포함될 문서확률($p(\theta^{(d)},z,w|\alpha,\beta)$)은 문서 내의 주제들의 확률분포를 추정하는 식과 문서의 주제를 구성하는 단어의 비율을 추정하는 식으로 구분된다. 먼저, 수식의 앞 부분인 $p(\theta^{(d)}|\alpha)$는 주제의 사전확률 비율에 대해 실제 문서에 주제가 나타날 사후확률(이는 데이터를 통해 실제로 도출)을 추정하는 식을 나타낸다. 다음으로, 수식의 뒷 부분은 문서에 포함될 주제들의 조건부 확률($p(z_n|\theta^{(d)})$=p(T | D))을 나타내고, 이 주제들의 확률은 마지막에 있는 단어들의 사전확률에 대해 실제 주제들이 단어들로 구성될 사후확률(($P(w_n|z_n,\beta)$=P(W | T))로 추정하게 된다. 이 결과는 단순히 한 번만으로 추정되는 것이 아니고, 데이터 샘플링을 통해 특정단어의 탈락과 진입을 반복적으로 수행되면서, 모수인 α와 β를 학습해나가며 가장 가능성이 높은 결과(이를 Gibbs sampling 알고리즘이라 하며, 수렴 확률을 추정하여 결정[15])를 최종적으로 도출한다 (LDA의 이보다 구체적인 확률 알고리즘과 수식 전개는 Blei et al.[11]을 참고하는 것을 추천한다).

3. 분석결과

3.1. 사고발생유형분석 : 주제 도출

화학 플랜트에서 수집된 127건의 사고보고문서를 텍스트마이닝과 LDA를 수행한 결과, Table 1과 같이 10개의 주제와 관련 키워드 7개씩을 도출하였다. 텍스트마이닝과 LDA는 각각 *R*의 tm 패키지와 topicmodels 패키지를 활용하였다. 주제의 수(10개)는 전체적인 확률분포의 차이를 최대화하는 결과치로 결정하였다. LDA는 문서가 모든 주제에 포함될 확률을 도출하는 soft clustering 결과로, 일반적으로 한 주제에 가장 높은 확률을 가지는 결과로 수렴된다. 따라서 LDA는 가장 높은 확률을 가지고 있는 주제를 주요 군집으로 분석

Table 1. Topics of chemical accidents

Topic	Keywords	Frequency
1-Chemical storage	water, leak, waste, Gas 1, drum, Gas 2, pump	11(8.7%)
2-Material removal/ disconnection	shut, MSDS, Gas 3, fume, operation, remove, employee	10(7.9%)
3-Pressurization	tank, low, compressor, air, materials, mixture, activated	9(7.1%)
4-Cylinder handling	floor, cylinder, circulation, heater, storage, high, operation	13(10.2%)
5-Safety switch activation	switch, tank, back, NH3, mixture, safety, activated	15(11.8%)
6-Pipeline flow	pipe, flow, inside, small, synthesis, drum, employee	9(7.1%)
7-Process line	site, normal, closed, line, local, materials, operators	14(11.0%)
8-Gas detector	line, plant, check, gas, alarm, detector, pressure	20(15.7%)
9-Ventilation	cell, fan, alarm, electric, flow, detector, broken	14(11.0%)
10-Power supply	inspection, power, Gas 4, chemical, level, top, check	12(9.4%)
Sum		127(100%)

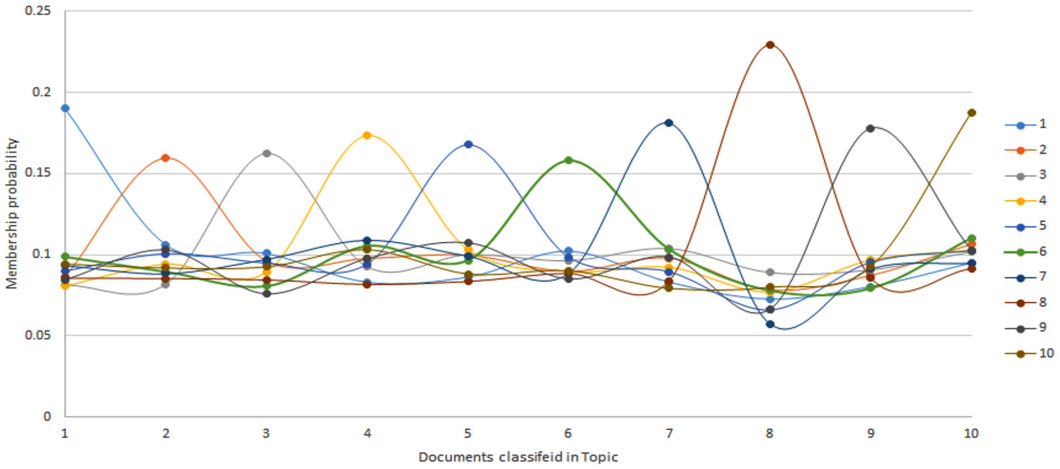

Fig. 2. Probability of classifying documents into each topic.

을 수행한다. Fig. 2에서 보듯이 문서들이 가장 높은 확률로 포함된 주제의 확률이 다른 주제에 할당된 확률보다 크게 차이가 있음을 확인할 수 있으므로, 앞서 언급했듯이, LDA는 가장 높은 확률에 할당된 주제들로 문서를 분류하게 된다. X축은 문서들을 분류한 각 주제번호를 나타내며, Y축은 그 주제에 포함된 문서들이 전체 각 주제들에 포함될 확률 그래프를 보여준다. 예를 들어, 1번 주제에 가장 높은 확률을 가지고 있는 문서들은 다른 주제에 대해 상대적으로 낮은 확률을 가지고 있음을 보여주고 있다. 또한 키워드 수는 분석자가 주제를 적절히 해석할 수 있는 수를 주관적으로 설정하였다. 본 연구에서는 정보를 기억하고 구별하기 용이한 개수로 알려진 7개로 설정하였다.

결과적으로 10개의 주제들은 해당 화학 플랜트에서 주로 발생한 사고유형들로서, 해당 키워드들을 바탕으로 사고유형주제를 해석하였다. (1) 화학물질 저장 및 보관, (2) 물질 제거 및 설비 분리, (3) 공정 가압, (4) 실린더 취급, (5) 안전스위치 작동, (6) 배관 유체 흐름, (7) 공정 배관, (8) 가스 감지기, (9) 환기 시스템, (10) 전력 공급 계통들로 도출되었다. 주제들에 포함된 키워드들은 사고를 발생시킨 설비(machine), 미디어(media), 공정 및 운전조건(process)들을 주로 포함한다.

각각의 사고 관련 주제에 세부적으로 포함된 키워드들을 살펴보면, (1)번 주제의 경우, 화학물질 보관/저장 시 드럼이나 펌프 등에서 Gas 1, Gas 2 또는 물이나 폐수 등이 누출되어 벌어진 사고라는 것을 추정할 수 있다 (Gas는 특정제품명으로, 가스의 종류가 다르다는 의미만을 순번으로 표현하였다). 마찬가지로 (2)번 주제에서는 화학물질의 제거 또는 설비 분리 시에 Gas 3 이나 흄(fume)이 누출되고 공정이 Shutdown 된 사고로 해석할 수 있다. (3)번의 경우는 공정을 가압하는 과정에서 비정상 압력에 의한 물질의 누출, 이로 인한 인터락(inter-lock)의 작동 등으로 이해할 수 있고, (4)번의 경우는 실린더 저장 및 취급 과정에서 물리적 위험이나 연결/분리 과정에서 가스의 누출 등을 추정해 볼 수 있다. (5)번은 설비 등을 가동 시에 공정 불안전 상태로 인하여 탱크(tank)와 연동된 안전장치가 가동되고 이 과정에서 Gas 4 등이 누출된 것을 나타낸다. (6)번과 (7)번의 경우는 공정배관 및 배관 내의 유체에 의한 것으로 운전자(operator)의 조정 하에 흐름의 막힘이나 합성 관련된 사고로 볼 수 있고, (8)~(10)번은 가스감지기, 환기 장치, 전력 공급 장치 관련한 정상 감지 또는 감지 오류로 가스 및 가스의 압력 감지, 환기 장치 팬의 유량 감지, 전력 장비의 손상 감지 등의 문제로 판단된다.

전체적으로 1년 동안 가장 자주 발생한 사고발생유형은 (8)번 가스 감지기와 관련한 사항 (20건, 15.7%)으로 나타났으며, (5)번 안전스위치 작동 (15건, 11.8%), (7)번과 (9)번의 공정 배관 및 환기 시스템 (각 14건, 11.0%), (4)번 실린더 취급(13건, 10.2%) 순서로 나타났다. 이를 토대로 볼 때, 해당 사업장에서는 전체 사고 관련 주제의 1/4을 초과하는 공정 상태 (누출, 압력, 유량 등) 감지장치(detector)와 관련된 (8)번과 (9)번 (34건, 26.7%)을 특별히 관리할 필요성을 가진다.

3.2. 사고발생유형 관계분석 : 특성요인도

사고발생유형의 주제를 구성하는 키워드들은 크게 설비(machine), 미디어(media), 공정(process)과 관련된 상태 및 행위들을 나타낸다. 이를 각 주제 및 키워드들과의 연관관계를 추정하여 Fig. 3과 같이 특성요인도로 구조화하여 현장에서 발생하고 있는 사고의 상태를 진단한다. 전체 사고발생유형 관계에 대한 특성요인도의

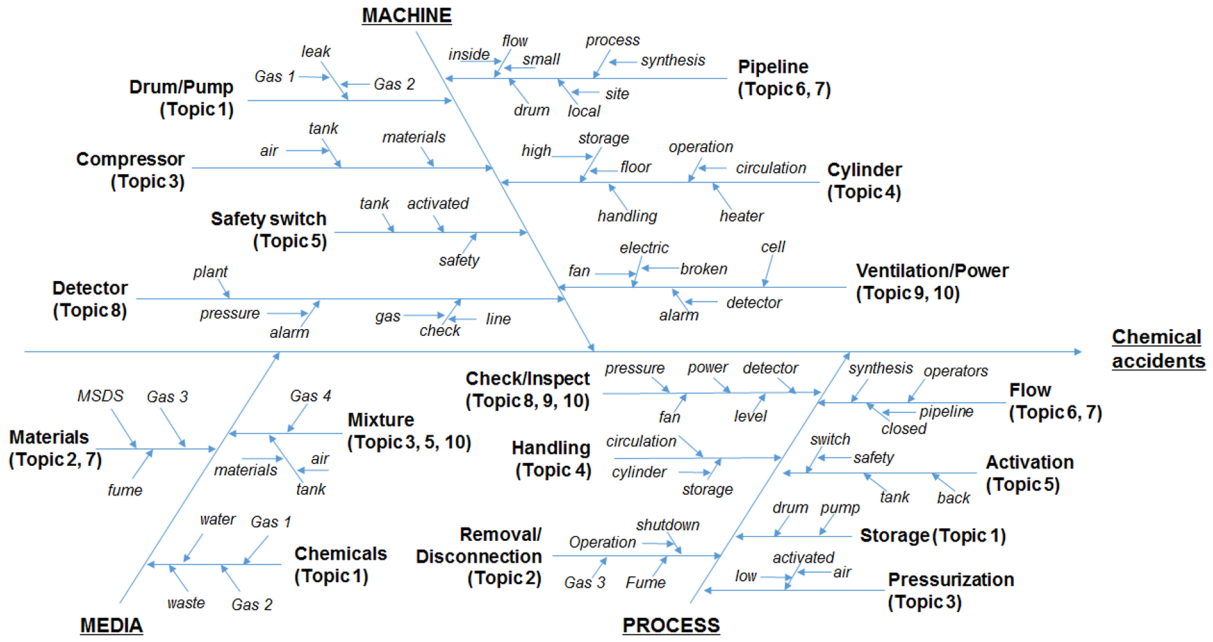

Fig. 3. LDA-driven cause-and-effect diagram of chemical accidents.

작성은 안전관리자로 하여금 설비, 미디어, 공정 측면에서 전체 문제점을 종합적으로 연계하여 이해하고, 안전진단을 체계적으로 수행하도록 지원한다.

먼저, 설비와 관련된 사고발생 키워드는 Drum/pump (주제 1), Compressor (주제 3), Cylinder (주제 4), Safety switch (주제 5), Pipeline (주제 6, 7), Detector (주제 8), Ventilation/Power (주제 9, 10)로 도출되었다. 좀 더 구체적으로 살펴보면, Drum/pump는 Gas 1, Gas 2의 누출(leak)에 따른 것이고, Cylinder는 취급 및 저장 시 오류와 연관되며, Pipeline은 공정유체의 흐름상 문제가 발생한 것임을 보여준다. Detector는 가스 감지 및 공정 상태(특히 압력) 변동에 따른 알람발생 사고이고, Ventilation/Power는 전기적 문제 등으로 팬의 가동이 중단되고 이에 따른 자동 감지 및 알람으로 이어진 사고임을 나타낸다.

둘째, 미디어와 관련된 키워드는 Chemicals (주제 1), Materials (주제 2, 7), Mixture (주제 3, 5, 10)에서 주로 도출되었다. 각각의 항목에 포함되어 있는 물질정보를 보면, Chemicals에서는 Gas 1, Gas 2, 물(water) 및 폐기물(waste)이 도출되었으며, Materials에서는 Gas 3과 흄의 관련성이 높게 구조화되었다. 마지막으로 Mixture는 탱크에 포함된 Gas 4 혼합물일 가능성이 높으며, 탱크 저장 및 운반 과정에서의 사고로 도출되었다.

마지막으로 공정 및 운전조건에서는 원료 및 제품의 보관/저장 방법 (주제 1), 설비의 연결 및 분리 방법 (주제 2), 공정 조건(압력)의 변경 (주제 3), 용기 및 기기의 취급 요령 (주제 4), 설비의 가동 (주제 5), 공정 설비 (배관) 및 운전상태(유량)의 불안정 (주제 6, 7), 설비 및 공정조건의 점검 및 감지 (주제 8, 9, 10)시의 오류로 인한 사고들이 도출되었다. 이 주제들은 공정 설비와 원료/제품의 투입 및 취급 과정과 밀접하며, 보다 현실적으로는 근로자의 작업환경과도 연결될 수 있는 정보라는 점에서 중요하다. 무엇보다 다양한 주제 (주제 8, 9, 10)의 키워드를 포함하는 Check/Inspect 사고요인에 대해서는, 제반 공정의 전반적인 상태를 확인하고 관리하는 운전원 및 안전관리자의 역할을 보다 부각시키는 결과를 보여준다.

3.3. 사고발생유형 변화분석 : 모니터링

사고의 추세와 관리 정도를 확인하기 위해 주제별로 도출된 사고발생유형들을 시간에 따라 변화되는 양상을 분기에 따라 분석하였다. Table 2는 3개월마다의 각 주제별 사고의 변화 추세를 나타낸다. 이 결과는 수집한 전체적인 사고 수가 많지 않아 통계적으로 유의미한 변화를 해석하기보다는, 단순히 증감추세와 비율변화 정도만을 확인할 수밖에 없는 한계점이 있다. 그럼에도 불구하고, 사고 수나 재해비율이 기간별로 증가 혹은 감소의 경향성을 가지는 결과 자체는 검토해볼 의의가 있다.

우선, 주제별 사고의 경향은 {증가→감소} 추세와 {감소→증가} 두 가지 유형이 주로 나타났다. 첫 번째로, 사고 수와 비율 두 가지 모두 {증가→감소} 추세를 보이는 사고발생유형은 (1)번, (3)번, (6)번으로 나타났

Table 2. Trends of accidents

Topic	2015 10-12	2016 01-03	2016 04-06	2016 07-09	Trends
1	3	4	3	1	I* → D**
	11.1%	13.3%	9.7%	2.6%	
2	2	2	3	3	I
	7.4%	6.7%	9.7%	7.7%	D → I → D
3	0	2	5	2	I → D
	0.0%	6.7%	16.1%	5.1%	
4	2	3	4	4	I
	7.4%	10.0%	12.9%	10.3%	I → D
5	5	3	2	5	D → I
	18.5%	10.0%	6.5%	12.8%	
6	0	2	4	3	I → D
	0.0%	6.7%	12.9%	7.7%	
7	3	4	2	5	I → D → I
	11.1%	13.3%	6.5%	12.8%	
8	**3**	**4**	**5**	**8**	**I**
	11.1%	**13.3%**	**16.1%**	**20.5%**	
9	5	4	1	4	D → I
	18.5%	13.3%	3.2%	10.3%	
10	4	2	2	4	D → I
	14.8%	6.7%	6.5%	10.3%	
Sum	27	30	31	39	I

* I: Increasing; D: Decreasing
** I→D means Increasing and then, decreasing

다. 이는 화학물질 저장/보관, 공정 가압, 공정배관 내 유체 흐름이 관리 체계에 들어가, 사고가 줄어든 경우로 고려할 수 있다. 비율 측면에서는 (4)번의 실린더 취급 관련 사고도 해당 기간의 전체 사고 대비해서는 감소 추세에 있음을 확인할 수 있다. 두 번째로, {감소→증가} 추세를 보이는 사고발생유형으로는 (5)번, (9)번, (10)번으로 나타났다. 이는 안전스위치의 오작동 개선이나 환기장치, 전력공급 계통 설비의 유지보수 기간을 보다 짧게 두어 관리할 필요성을 나타낸다.

이보다 더 중요한 추세로는 지속적으로 증가하고 있는 사고 유형이다. (2)번과 (4)번은 증가추세지만, 그 증가 수가 미비하고, 해당 기간의 전체 사고 대비해서는 감소 추세에 있다. 그러나 (8)번 사고발생유형의 경우 3회 발생하던 사고가 마지막 3개월에 8회로 증가하였으며, 전체 사고의 20%를 초과할 정도로 높은 비율을 차지할 만큼 변화하였다. 이와 관련한 사고 문서들을 살펴보면 가스 감지기의 비정상 작동보다는 정상 작동했음에도 발생한 사고보고가 많이 나타났다. 이는 감지기의 과도한 감지 능력으로 미약한 외부 환경에도 작동하거나 잦은 공정 누출로 인한 것으로, 감지기의 적정한 감지 능력 설정과 실제 누출 사고를 구분할 수 있는 방안들이 시급히 수립될 필요가 있다.

4. 결 론

본 연구는 지속적으로 누적되고 있는 사고보고문서에 포함된 사고발생유형의 텍스트 분석방법을 제시하였다. 또한, 분석결과를 안전관리자가 이해하기 쉽게 사고발생 원인과 유형의 관계를 특성요인도로 시각화하였다. 마지막으로 시간에 따른 사고발생 유형의 변화와 경향을 정리하였다. 이는 기존의 단순히 문서를 직접 검토하던 작업에서 벗어나, 사고보고문서에 포함된 텍스트 데이터를 분석하면, 현재 발생하고 있는 주요 사고발생 원인과, 유형, 관계를 도출하고, 안전관리자가 특성요인도를 효과적으로 작성할 수 있도록 지원한다.

이와 같이 사고보고문서를 이용한 텍스트 기반의 분석은 새로운 안전관리 시스템을 설계하기 위해 학술적・실무적으로 중요한 역할을 제공하리라 기대된다. 우선 학술적으로 안전관리 분야에 경험적이고 수치적인 연구를 벗어나, 과학적이고 서술적인 연구를 새롭게 시도할 수 있다. 또한, 실무적으로도 현상을 이해하기 쉬운 텍스트 위주의 정보를 최종결과로 제시함으로써, 안전관리자가 기구축한 특성요인도의 상태를 지속적으로 개선할 수 있는 시사점을 제공한다. 따라서 안전관리자가 LDA 알고리즘 분석을 통해, 사고보고문서 안에 포함되어 있는 주요 사고원인을 분석하여, 작업자와 경영자에게 제시하는 업무를 지원한다는 기여점이 있다.

그러나 이와 같은 기여점과 시사점에도 불구하고, 실질적인 측면의 문제점을 개선하고, 실무에서 활용하기 위한 향후 연구가 필요하다. 첫째, 사고보고문서 데이터의 질(quality)의 문제이다. 사고보고문서의 사고개요와 장소, 위험성 등과 관련한 서술 내용이 비체계적으로 작성되고 있으며, 그 내용도 상세하지 못하다는 점이다. 자연어 분석 및 텍스트 분석은 불확정한 서술 내용임에도 유효한 분석결과를 제공하지만, 사고보고문서의 내용이 더욱 체계적이고 상세하다면 보다 유의한 결과를 제공할 것이다. 둘째, 사고보고문서 데이터의 양(quantity)의 문제이다. 현재 연구는 기업 단위의 사고보고문서 분석을 수행하여, 많은 데이터를 확보하지 못하였다. 그러나 안전보건공단 등 공공과 산업 차원에서 대량의 데이터 분석이 가능한 사고보고문서를 수집한다면, 기업이 아닌 건설, 화학, 제조, 조선 등 산업별 사고원인유형과 경향을 효과적으로 제시할 수 있

을 것으로 기대된다. 마지막으로, 사고보고문서 수집의 기간의 한계점이 있다. 사고보고문서를 보다 장기간 수집하였다면, 사고발생유형의 변화여부를 통계적으로 검정하여 보다 유의한 결과를 제시하리라 기대된다. 이는 사고발생유형 동적변화분석의 향후 연구로 수행될 필요가 있다.

감사의 글: 이 논문은 부경대학교의 지원(2016년 후기 신임교수 연구력강화 지원사업(공과대학))에 의해 연구되었음.

References

1) H. S. Lee and J. P. Yim, "A Study on Prevention Measure Establishment through Cause Analysis of Chemical-Accidents", J. Korean Soc. Saf., Vol. 32, No. 3, pp. 21-27, 2017.
2) D. H. Choi, J. W. Choi and W. C. Shin, "Accident Analysis and Research on Risk of the Actual Conditions", J. Korean Soc. Saf., Vol. 27, No. 5, pp. 111-116, 2012.
3) G. H. Choi, "Cause Analysis of Accidents Associated with Industrial Machines and Devices", J. Korean Soc. Saf., Vol. 33, No. 1, pp. 16-21, 2018.
4) F. Abdat, S. Leclercq, X. Cuny and C. Tissot, Extracting Recurrent Scenarios from Narrative Texts Using a Bayesian Network: Application to Serious Occupational Aaccidents with Movement Disturbance, Accident Analysis and Prevention, Vol. 70, pp. 155-166. 2014.
5) R. Moura, M. Beer, E. Patelli and J. Lewis, "Learning from Major Accidents: Graphical Representation and Analysis of Multi-attribute Events to Enhance Risk Communication", Safety Science, Vol. 99, pp. 58-70, 2017.
6) G. Ahn, M. Seo and S. Hur, "Development of Accident Classification Model and Ontology for Effective Industrial Accident Analysis based on Textmining", J. Korean Soc. Saf., Vol. 32, No. 5, pp. 179-185, 2017.
7) Y. Suh, "Data Analytics for Social Risk Forecasting and Assessment of New Technology, J. Korean Soc. Saf., Vol. 32, No. 3, pp. 83-89, 2017.
8) Q. Fang, H. Li, X. Luo, L. Ding, T. M. Rose, W. An and Y. Yu, "A Deep Learning-Based Method for Detecting Non-certified Work on Construction Sites", Advanced Engineering Informatics, Vol. 35, pp. 56-68, 2018.
9) F. Palamara, F. Piglione and N. Piccinini, "Self-Organizing Map and Clustering Algorithms for the Analysis of Occupational Accident Databases", Safety Science, Vol. 49, pp. 1215-1230, 2011.
10) H. Kwon, J. Kim and Y. Park, "Applying LSA Text Mining Technique in Envisioning Social Impacts of Emerging Technologies: The Case of Drone Technology", Technovation, Vol. 60-61, pp. 15-28, 2017.
11) D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation". Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
12) S. Moro, P. Cortez and P. Rita, "Business Intelligence in Banking: A Literature Analysis from 2002 to 2013 Using Text Mining and Latent Dirichlet Allocation", Expert Systems with Applications, Vol. 42, No. 3, pp. 1314-1324, 2015.
13) S. Tirunillai and G. J. Tellis, "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation". Journal of Marketing Research, Vol. 51, No. 4, pp. 463-479, 2014.
14) D. E. Brown, "Text Mining the Contributors to Rail Accidents", IEEE Transactions on Intelligent Transportation Systems, Vol. 17, No. 2, pp. 346-355, 2016.
15) Wei, Xing, and W. Bruce Croft., "LDA-Based Document Models for Ad-hoc Retrieval", Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178-185, 2006.