J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

설비 오류 유형 구조화를 위한 인공신경망 기반 구절 네트워크 구축 방법☆

An Artificial Neural Network Based Phrase Network Construction Method for Structuring Facility Error Types

노 영 훈[1] 최 은 영[1] 최 예 림[2*]
Younghoon Roh Eunyoung Choi Yerim Choi

요 약

4차 산업혁명 시대의 도래와 함께 스마트 팩토리의 개념이 대두되면서 설비가동률과 생산성에 악영향을 미치는 설비 오류의 발생을 데이터 분석 기법을 통해 예측하고자 하는 노력이 이루어지고 있다. 데이터 분석 기법을 활용하여 설비 오류를 예측하기 위해서는 설비 오류가 발생한 상황과 설비 오류 유형을 명시한 데이터인 설비 오류 이력이 필요하다. 하지만 많은 제조 현장에서는 설비 오류 유형이 정확하게 정의/분류가 되지 않아 설비를 운영하는 작업자가 자신의 경험적 판단에 의거하여 정형화되지 않은 텍스트의 형태로 설비 오류 유형을 작성하고, 이에 따라 데이터 분석 기법의 적용이 어렵다. 따라서 본 논문에서는 수기로 작성된 설비 오류 이력을 활용하여 설비 오류 유형을 파악하고 구조화하기 위한 구절 네트워크 구축 방법을 제안하고자 한다. 구체적으로, 단어를 쓰임새에 따라 분류한 용도 딕셔너리를 활용하여 비정형의 텍스트 데이터로부터 설비 오류 유형을 의미하는 구절을 추출하고, 추출된 구절 간의 유사도를 계산하여 네트워크를 구축한다. 제안하는 방법의 성능을 실제 제조 기업의 설비 오류 이력 데이터를 활용하여 검증하였으며, 본 연구의 결과는 텍스트 데이터에 기반한 설비 오류 유형 구조화와 나아가서는 설비 오류 발생 예측에 이용할 수 있을 것을 기대한다.

☞ 주제어 : 스마트 팩토리, 설비 오류, 구절 네트워크, 텍스트 마이닝, 인공신경망, word2vec

ABSTRACT

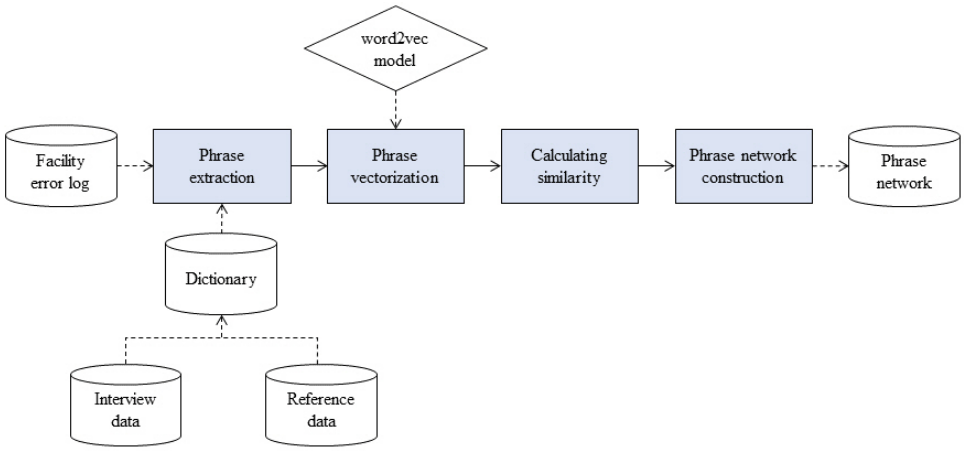
In the era of the 4-th industrial revolution, the concept of smart factory is emerging. There are efforts to predict the occurrences of facility errors which have negative effects on the utilization and productivity by using data analysis. Data composed of the situation of a facility error and the type of the error, called the facility error log, is required for the prediction. However, in many manufacturing companies, the types of facility error are not precisely defined and categorized. The worker who operates the facilities writes the type of facility error in the form with unstructured text based on his or her empirical judgement. That makes it impossible to analyze data. Therefore, this paper proposes a framework for constructing a phrase network to support the identification and classification of facility error types by using facility error logs written by operators. Specifically, phrase indicating the types are extracted from text data by using dictionary which classifies terms by their usage. Then, a phrase network is constructed by calculating the similarity between the extracted phrase. The performance of the proposed method was evaluated by using real-world facility error logs. It is expected that the proposed method will contribute to the accurate identification of error types and to the prediction of facility errors.

☞ keyword : Smart factory, Facility error, Phrase network, Text mining, Artificial neural network, word2vec

1. 서 론

대다수의 제조 현장에서 하나의 제품은 다양한 공정과 해당 공정을 구성하고 있는 여러 개의 설비를 거쳐 생산된다[1]. 실제로 자동차 산소 센서를 생산하는 국내 제조 기업에서 산소 센서는 센서 외주 공정, 센서 직선 라인 공정, 센서 조립 검사 공정, 마지막으로 센서 포장 공정의 총 네 가지 공정을 거쳐 생산된다. 이 네 가지 공정은 단순하게 하나의 설비로 구성되어 있지 않고 11가지 종류의 23개 설비로 구성되어 있다. 또한, 이 23개 설비는 독

1 Industrial and Management Engineering, Kyonggi University, Suwon, 16227, Republic of Korea
2 Assistant Professor, Industrial and Management Engineering, Kyonggi University, Suwon, 16227, Republic of Korea
* Corresponding author (yrchoi@kgu.ac.kr)
[Received 08 August 2018, Reviewed 17 September 2018, Accepted 30 October 2018]
☆ 본 연구는 경기도의 경기도 지역협력연구센터 사업의 일환으로 수행하였음.[GRRC 경기 2017-B01, 지능형 제조 빅데이터 분석 연구]
☆ 본 논문은 2018년도 한국인터넷정보학회 춘계학술발표대회 우수 논문 추천에 따라 확장 및 수정된 논문임.



(그림 1) 제안 방법 도식
(Figure 1) Framework of the proposed method

립적으로 운영되는 것이 아니라 서로 종속적인 관계로서 특정 설비의 오류가 발생할 경우 해당 설비뿐만 아니라 다른 설비의 운행에도 악영향을 미친다. 오류로 인해 설비 정지가 발생한다면 설비가동률은 감소할 것이며 공장의 생산성 또한 감소할 것이다.

이처럼 생산성에 큰 영향을 끼치는 설비 오류를 데이터 분석 기법을 통해 정확하게 예측하려는 연구가 다수 진행되어 왔다. 예를 들어, 설비의 진동 데이터와 같은 환경 데이터를 사용하여 설비의 잔여 수명을 계산함으로써 설비가 언제 고장 날지 예측하는 연구가 있었으며[2,3,19], 설비에 센서를 부착하여 설비의 물리적 데이터를 수집하는 IoT(Internet on things)를 활용한 설비 고장 예측 연구도 진행되었다[4]. 공정의 스케줄을 효율적으로 계획하기 위하여 사전에 설비 오류를 환경 데이터와 설비를 구성하는 부품의 수명 데이터를 사용하여 예측하는 연구 또한 진행되었다[20].

데이터 분석 기법을 사용하기 위해서는 학습 데이터의 수집이 선행되어야 하므로 다수의 제조 기업에서는 설비 오류의 유형과 설비 정지가 발생한 상황을 기록한 설비 오류 이력 데이터를 수집하고자 많은 노력을 기울이고 있다[2]. 하지만 설비 오류를 예측하기 위하여 필수적인 설비 오류의 유형을 정확하게 정의/분류하지 못하고 있는 제조 기업들이 다수 존재한다[14]. 이처럼 설비 오류 유형이 정확하게 분류되지 않은 상태에서는 오류가 발생했을 때 해당하는 오류 유형이 없거나 범주가 애매해 데이터를 정확히 입력하는 것이 불가능하다.

이와 같은 상황에서 작업자는 설비 오류 유형을 정확하게 기입하기 위해 본인의 경험적 판단에 의거하여 설비 오류 유형을 수기로 작성한다. 수기로 작성된 데이터는 통일되지 않은 단어와 규정된 형식 없이 자유롭게 구성되어 있어, 다수의 오·탈자를 포함하며 동일한 설비 오류 유형을 상이하게 표현하는 경우가 빈번하다. 예를 들어, 작업자에 따라 '니플'이라는 단어를 '니쁠'이라고 명시하는 경우가 있으며, '깨짐'이라는 단어를 '께짐'이라는 단어로 잘못 입력하는 경우가 빈번하게 발생한다.

비정형 텍스트 데이터를 다루는 다수의 기존 연구들에서는 일반적으로 텍스트 데이터를 단어 단위로 구분하여 분석을 진행하였다[5]. 하지만 제조 분야에서 사용되는 단어는 일반적으로 사용되는 단어들과 의미적인 차이점이 있으며 전문적인 의미를 내포하기 위해서는 여러 개의 단어가 묶음으로 사용되는 경우가 많다[6]. 즉, 하나의 단어로는 정확한 설비 오류 유형을 표현할 수가 없다. 예를 들면, 설비 오류 유형이 '실린더 마모'인 경우 하나의 단어의 '실린더' 또는 '마모'만을 사용하여 정확한 원인을 표현할 수 없다.

또한, 본 논문에서 해결하고자 하는 추가적인 문제점은 단어의 빈도수를 활용하여 설비 정지의 원인을 정확하게 파악할 수 없다는 것이다. 출현한 단어의 빈도수를 활용하여 설비 정지의 원인을 파악하고자 하는 연구는 다수 진행되었다[7,18,19]. 하지만 수기로 작성된 텍스트 데이터의 경우 동일한 의미지만 작업자에 따라 다른 형태로 표현된 단어가 있기 때문에 빈도수를 기반으로 분

석을 수행할 경우 의미론적으로 동일한 단어의 빈도수를 정확하게 표현하지 못한다. 그러므로 빈도수 기반의 분석을 수행하게 된다면 동일한 의미를 가진 설비 오류를 다른 유형으로 분류할 수 있다는 한계점을 지니고 있다.

본 논문에서는 이와 같은 한계점을 극복하여 설비 오류 유형을 정확하게 파악/분류하고자 설비 오류 유형을 나타내는 구절 추출 방법과 추출된 구절 간의 네트워크 구축 방법을 제안한다. 제안하는 방법은 크게 두 단계로 구성된다. 첫 번째 단계에서는 단어의 쓰임새를 구별해 놓은 용도 딕셔너리[17]를 사용하여 작업자가 작성한 설비 오류 이력으로부터 설비 오류 유형을 나타내는 구절을 추출한다. 두 번째 단계에서는 텍스트 데이터의 의미론적 속성을 표현하여 벡터화할 수 있는 인공신경망 기반의 word2vec 모델을 활용하여 구절을 표현하고, 벡터로 표현된 구절 간의 유사도를 계산하여 구절 네트워크를 구축한다.

본 논문은 다음과 같이 구성된다. 2장에서는 구절 추출 방법과 네트워크 구축 방법을 상세하게 설명하며, 3장에서는 이를 기반으로 실제 제조 회사의 데이터를 사용하여 제안 방법을 평가한다. 마지막으로 4장에서는 본 연구의 결론에 대해 논의한다.

2. 제안 방법

2.1 개 요

본 논문에서는 설비 오류 유형의 정의와 분류를 돕기 위해 작업자가 수기로 작성한 설비 오류 이력 데이터를 사용하여 설비 오류 유형을 정확히 표현할 수 있는 구절을 추출하고, 구절 사이의 유사도를 계산하여 구절 네트워크를 구축하는 방법을 제안한다. 그림 1은 제안 방법의 프레임워크를 나타낸다. 제안 방법은 (a) 구절 추출, (b) 구절 벡터화, (c) 구절 간의 유사도 계산 및 네트워크 구축의 총 세 단계로 구성된다.

가장 먼저, 정형화된 구조 없이 수기로 작성된 텍스트 데이터를 사용하여 설비 오류 유형을 정확히 표현할 수 있는 구절을 추출하고, 추출된 구절을 벡터로 표현하기 위하여 word2vec 알고리즘을 적용한다. 다음으로, 벡터를 활용하여 각기 다른 특성을 지닌 두 구절의 유사도를 계산한다. 마지막으로, 유사도별 계산 결과를 활용하여 추출된 구절 간의 네트워크를 구축하고, 설비 오류를 야기하는 유형을 분류하고 설비 오류 유형을 구축한다.

2.2 구절 추출

제안 방법에서는 텍스트 분석을 단어 단위가 아닌 구절 단위로 진행한다. 텍스트 데이터의 분석 방법은 다양하지만, 단어를 기본 단위로 한 분석 방법이 가장 일반적이다[3]. 하지만 단어 단위의 분석은 여러 개의 단어가 모여 전문적인 의미로 사용되는 설비 오류의 경우에 적합하지 않다. 따라서 본 논문에서는 설비 오류 유형을 정확하게 표현할 수 있는 최소한의 의미 단위인 구절 단위로 분석을 하고자 한다. 구절 추출을 위해 현장 전문가의 인터뷰와 참고문헌을 바탕으로 단어를 쓰임새에 따라 구별한 용도 딕셔너리를 사용한다.

본 논문에서는 단어의 쓰임새를 총 네 가지로 정의하였다. 첫 번째, 설비 오류 유형의 정확한 의미를 표현하기 위하여 필수적으로 필요한 단어, 두 번째, 설비 오류 유형의 의미를 표현하지 않아 불필요한 단어, 세 번째, 추출하는 구절을 다른 구절들과 구분함과 동시에 설비 오류 유형의 의미를 표현하는 단어, 마지막으로 다른 구절과 구절을 구분하지도 않으면서 구절에 표현할 필요가 없는 단어이다. 이러한 단어의 네 가지의 쓰임새를 기준으로 제조 용어들을 구분한 용도 딕셔너리를 구축한다.

구축된 용도 딕셔너리를 사용하여 작업자가 수기로 입력한 설비 오류 이력으로부터 오류 유형에 해당하는 구절을 추출한다. 예를 들어, ‘삽입부 계측오류로 인한 설비 정지 후 계측기 재측정 및 영점 셋팅 후 재가동’이 설비 오류 이력 중 작업자가 수기로 작성한 텍스트라고 할 때, 용도 딕셔너리를 이용하여 구절을 추출하면 ‘삽입부 계측 오류’, ‘설비 정지’, ‘계측기 재측정’과 같은 총 세 개의 구절을 얻을 수 있다.

2.3 word2vec을 활용한 구절 벡터화

추출된 구절과 같이 텍스트 형식의 데이터는 컴퓨터가 학습하거나 처리할 수 없으므로 벡터 공간에 할당하여 표현해야 한다[8]. 또한, 구절을 벡터로 표현함으로써 구절 사이의 유사성을 계산할 수 있으므로 벡터화는 필수적이다. 텍스트 데이터를 벡터 공간에 할당하기 위한 연구는 오랜 기간 동안 다수의 연구가 진행되어 왔으며 다양한 알고리즘들이 존재한다[9].

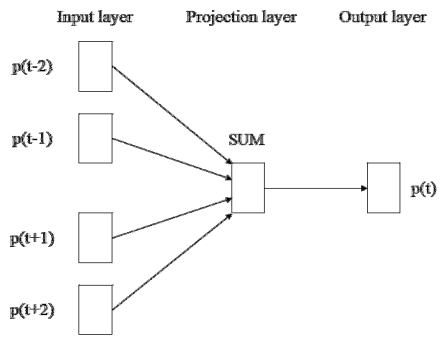
텍스트 데이터를 벡터화하기 위하여 가장 기본적인 방법은 텍스트의 발생 빈도를 활용하는 것이다[10]. 설비 오류 이력에서 추출된 구절은 통일되지 않은 단어와 규정된 형식 없이 작성된 텍스트 데이터이므로 동일한 의미

를 가지지만 다르게 표현된 구절이 다수 존재하여 빈도수 기반 분석은 적절하지 않다. 예를 들면 '실린더 노후'와 '실린더 내부 노후'는 다르게 표현되었지만 동일한 의미를 가진 구절이다. 이들을 빈도수 기반 방식을 사용해 분석하게 되면 서로 다른 원소에 빈도수 값을 갖는 벡터가 되어 이들 간의 의미적 유사도를 반영할 수 없다.

최근, 동일한 의미를 가진 구절을 근접한 벡터로 표현하기 위하여 word2vec[11] 알고리즘이 널리 활용되고 있다. word2vec 알고리즘은 신경망 분석 알고리즘의 한 유형으로 특정 단어의 앞/뒤에 위치한 단어의 분포를 활용함으로써 단어의 의미를 내포하며 벡터로 표현한다. 동일한 설비 오류가 발생하였지만 각기 다르게 작성된 텍스트로부터 추출된 구절은 word2vec 알고리즘을 활용하여 근접한 벡터로 표현이 가능하다.

word2vec의 대표적인 모델인 CBOW(continuous bag of words) 모델[11]을 사용하면 효과적으로 단어의 의미를 내포할 수 있다. CBOW 모델은 입력층(input layer), 투사층(projection layer), 출력층(output layer)의 총 세 개의 층으로 구성되어 있다. CBOW 모델을 적용하기 전 출현한 모든 단어를 활용하여 단어의 집합을 만든 후 텍스트 데이터를 벡터화하는 bag-of-words[21] 기법을 적용한다. 입력층에서 특정 단어 주변의 단어들을 입력받고 이를 이용하여 투사층을 지나 출력층에서 특정 단어에 대한 다차원의 벡터로 표현한다.

그림 2는 CBOW 모델을 도식화한 것이다. t 번째의 구절을 p(t)라 하고 구절의 순서가 p(t-2), p(t-1), p(t), p(t+1), p(t+2)라고 할 때, CBOW 모델은 p(t)를 벡터화하기 위하여 입력층에서 p(t) 주변 구절의 벡터값을 입력받는다. 구체적으로, p(t) 주변의 구절 p(t-2), p(t-1), p(t+1)과 p(t+2)을



(그림 2) CBOW 모델 도식
(Figure 2) Concept of CBOW model

활용한다. 입력 값들은 투사층을 거치며 모든 구절들의 출현 확률을 계산하여 가장 높은 확률의 구절이 출력층으로 출력된다.

2.4 구절 간의 유사도 계산 및 네트워크 구축

본 단계에서는 2.3장에서 벡터로 표현된 구절 사이의 유사도를 계산한다. 두 벡터의 유사도를 표현하기 위한 지표는 다양하다. 각 지표들의 유사도 계산 방식은 조금씩 상이하므로 지표에 따른 성능을 비교하고자 다양한 지표를 활용한다. 본 논문에서는 코사인(cosine) 유사도[12]와 피어슨(Pearson) 유사도[13]를 활용하여 구절 간 유사도를 도출하고 이 중 유사한 구절의 묶음을 하나의 설비 오류 유형으로 판단한다.

코사인 유사도는 두 벡터 간 각도의 코사인 값을 이용하여 측정된 벡터 간의 유사한 정도를 의미한다. 코사인 유사도를 사용할 경우 코사인 각도가 작은 구절 사이의 유사도를 높은 것으로 판단한다[12]. 코사인 유사도는 식(1)과 같이 계산한다.

$$Sim_{\cos(\vec{p_a},\vec{p_b})} = \frac{\vec{p_a} \bullet \vec{p_b}}{\sqrt{\vec{p_a} \bullet \vec{p_a}} \times \sqrt{\vec{p_b} \bullet \vec{p_b}}} \quad (1)$$

이때, $\vec{p}_a$는 구절 a의 벡터를, $\vec{p}_b$는 구절 b의 벡터를 나타내며, '•'은 벡터의 내적을 나타낸다.

피어슨 유사도는 두 벡터 사이의 상관계수를 의미하며, 두 구절 간의 선형 관계를 파악하기 위해 사용한다. 두 개의 연속적인 숫자열의 일대일 비교를 통해 상관성을 측정한다. -1과 1 사이의 값을 갖고, 양의 상관관계를 가지게 된다면 1에 가깝고, 음의 상관관계를 가지면 -1에 가까워진다. 상관성이 없을 경우 상관계수 값은 0에 가까워진다[13]. 피어슨 유사도는 식(2)와 같이 계산한다.

$$Sim_{Pearson(\vec{p_a},\vec{p_b})} = \frac{Cov(\vec{p_a},\vec{p_b})}{\sqrt{\vec{p_a} \bullet \vec{p_a}} \cdot \sqrt{\vec{p_b} \bullet \vec{p_b}}} \quad (2)$$

이때, $Cov(\vec{p_a},\vec{p_b})$는 벡터 $\vec{p}_a$와 벡터 $\vec{p}_b$의 공분산 값을 의미한다.

유사도 지표를 이용하여 계산한 구절 간의 유사도를 이용하여 구절 네트워크를 구축한다. 네트워크 상에서 유사도 값이 큰 구절을 가깝게 위치하고 유사도 값이 작은 구절은 멀게 위치한다. 또한, 구절 간의 네트워크를 표현하기

위하여 유사성이 높은 구절을 선으로 연결하여 표현한다.

구절로부터 설비 오류 유형을 정의하기 위하여 구절 간 유사도를 기반으로 군집화를 시행하는 방법[16]을 적용한다. 동일한 군집에 할당된 구절들의 의미 판단을 통해 설비 오류 유형을 구축한다. 예를 들어, 특정 군집에 '리드선 마모', '실린더 마모', '배선 마모'와 같이 마모와 관련된 구절이 할당될 경우 해당 군집의 설비 오류 유형을 마모로 정의한다.

3. 실 험

3.1 실험 데이터

제안 방법의 구절 추출 성능과 네트워크 구축 결과를 평가하기 위해 실제 국내 제조 회사 '우진공업'의 산소 센서 공정의 데이터를 이용하여 실험을 수행하였다. 산소 센서 공정은 총 네 가지의 공정 구성되어 있으며, 데이터를 수집하지 않는 외주 공정을 제외하고 세 가지 공정의 23개 설비로 구성되어 있다. 산소 센서 공정은 독립적으로 구분되어 있지 않고, 연속적인 공정이므로 설비 오류 유형을 정확하게 파악하기 위하여 세 가지 공정의 데이터를 통합하여 사용하였다.

데이터의 수집 기간은 2014년 2월부터 2018년 1월까지이며, 수집된 데이터 중 부품 교환 또는 점검하기 위한 자주 보전과 같이 설비 가동을 위해 고의로 설비를 정지한 활동은 오류로 인한 설비 정지를 의미하지 않으므로 제거하였다. 또한, 설비 오류 유형을 기록하지 않아 분석에 사용할 수 없는 데이터도 제거하였다. 수집된 설비 오류 이력 데이터의 개수는 총 1,394개이다. 표 1은 수집된 설비 오류 이력 데이터의 예시를 나타낸다.

(표 1) 수집된 설비 오류 이력 데이터 예시
(Table 1) Example of the collected facility error logs

번호	설비 오류 이력
1	증상: 양품3번 실린더 앞 근접 센서 오동작 원인: 센서 고장 조치: 임시조치 센서 단선 시킴 임시 조치후 가동 확인중
2	삽입부 계측오류로 인한 설비정지 후 계측기 재측정 및 영점 셋팅 후 재가동
…	…
1394	절연불량 전후 실린더 에어밸브 잠금으로 인해 절연불량시 설비 정지로 인해 에어밸브 조정

단어 쓰임새에 따른 용도 딕셔너리를 이용하여 작업자가 작성한 텍스트 데이터에서 구절을 추출한 결과, 추출된 구절은 총 2,684개이다. 이때, 추출된 구절은 설비 오류 유형을 의미하는 구절과 정지로 인해 발생한 정지 현상을 나타내는 구절, 정지 발생 시 작업자가 어떠한 조치를 취했는지 기록한 조치 내역으로 구성된다. 설비 오류 유형과 현상은 명확하게 구별되지 않고 작업자에 의해서 혼용되어 사용되어왔다. 따라서 설비 오류 유형을 파악하고 분류하는 것이 본 연구의 목적이므로 이와 관련 없는 조치 내역 구절은 제외하여 총 2,446개의 구절을 분석에 사용하였다.

3.2 실험 환경

제안 방법에서는 추출된 구절의 의미론적인 측면을 반영하기 위하여 word2vec을 활용한 벡터화를 실시하였다. 구체적으로, 특정 구절의 주변 구절들을 같이 학습함으로써 해당 구절의 의미를 학습하였다. 이때, 다차원 공간에 텍스트를 할당하기 위하여 구절을 200차원의 벡터로 표현하였다.

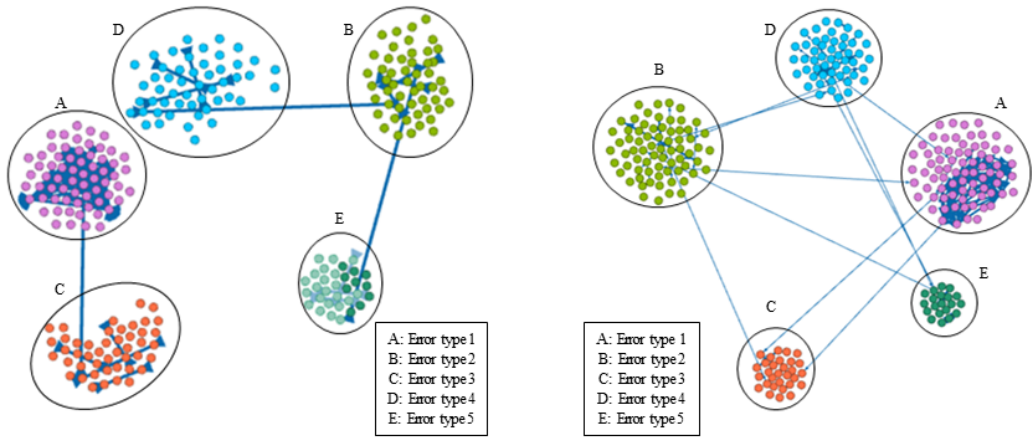
구절 간의 네트워크를 구축하는데 적합한 유사도를 판단하기 위해서 코사인 유사도와 피어슨 유사도의 두 가지 지표를 활용하여 비교하였다. 또한, 네트워크 시각화 분석 도구인 Gephi[15]를 활용하여 구절 네트워크를 가시화하였다.

3.3 실험 결과

표 2는 설비 오류 이력의 비정형 텍스트 데이터와 용도 딕셔너리를 활용해 추출된 구절을 나타낸다. 용도 딕

(표 2) 설비 오류 이력 텍스트로부터 추출된 구절
(Table 2) Example of the extracted phrases from facility error logs

번호	설비 오류 이력	추출된 구절
1	증상: 양품3번 실린더 앞 근접 센서 오동작 원인: 센서 고장 조치: 임시조치 센서 단선 시킴 임시 조치후 가동 확인중	[양품 3번 실린더 앞 근접 센서 오동작], [센서 고장], [임시조치]
2	삽입부 계측오류로 인한 설비정지 후 계측기 재측정 및 영점 셋팅 후 재가동	[삽입부 계측 오류], [설비 정지], [계측기 재측정]
…	…	…
1394	절연불량 전후 실린더 에어밸브 잠금으로 인해 절연불량시 설비 정지로 인해 에어밸브 조정	[절연불량 전후 실린더 에어 밸브 잠금], [절연불량 설비 정지]



(a) 코사인 유사도 시각화 결과 (b) 피어슨 유사도 시각화 결과

(그림 3) 유사도 지표에 따른 구절 네트워크 시각화 결과

(Figure 3) Comparison of the phrase networks using two similarity measures, (a) cosine and (b) Pearson.

셔너리 적용 결과, 통일되지 않은 단어와 형식으로 이루어진 비정형 텍스트 데이터로부터 설비 오류 유형을 표현할 수 있는 구절이 추출되는 것을 확인할 수 있었다. 이를 통해 용도 딕셔너리를 이용한 구절 추출 결과가 유의s미한 것으로 판단할 수 있다.

구절 네트워크 구축 단계에서는 벡터로 표현된 구절 간의 코사인 유사도와 피어슨 유사도를 계산하였다. 계산된 유사도를 활용하여 구절의 군집화를 실행한 결과 코사인 유사도와 피어슨 유사도를 활용한 경우 모두 5개의 군집으로 구절이 분리되었다. 표 3과 4는 각각 코사인 유사도와 피어슨 유사도를 이용한 구절 네트워크에서 구절 간 군집화를 수행한 결과의 예시를 나타낸다.

(표 3) 코사인 유사도 기반 오류 유형

(Table 3) The type of error based on cosine similarity

오류 유형	추출된 구절
오류 유형 1	설비 정지, 금형 미작동, 금형 뒤로 후진 안됨, 스위치 비정상 동작 …
오류 유형 2	센터 어긋남, 센서 위치 틀어짐, 팁 위치 이상, 실린더 고정 안됨 …
오류 유형 3	실린더 오동작, 척크 실린더 오동작, 실린더 동작않음, 센서 오동작 …
오류 유형 4	배선 단선, 포트 단선, 히터 단선, 리드선 단선, 끊어짐 …
오류 유형 5	팁 마모, 척크 마모, 척크 발 마모, 베벨기어 마모, 구동 베벨기어 마모 …

표 3의 다섯 개 오류 유형 중 첫 번째 오류 유형을 예로 들면, 코사인 유사도의 군집 결과는 '설비 정지', '금형 미작동', '금형 뒤로 후진 안됨'과 같이 부품이 정상적으로 작동되지 않는 유형이 동일한 군집에 할당되는 것을 확인할 수 있었다.

피어슨의 군집 결과 또한 코사인 유사도의 군집 결과와 동일하게 다섯 개의 오류 유형이 형성되었지만, 표 4의 두 번째 오류 유형에서 '고정 볼트 풀림', '척크 풀림', '감지부 센서 이상', '접촉 불량', '실린더 노후'와 같이 의미론적으로 유사하지 않은 구절이 동일한 군집에 할당되

(표 4) 피어슨 유사도 기반 오류 유형

(Table 4) The type of error based on Pearson similarity

오류 유형	추출된 구절
오류 유형 1	레이저 용접기 판 이상, 공연비 이상, 광센서 이상, 스프링 파손 …
오류 유형 2	고정 볼트 풀림, 척크 풀림, 감지부 센서 이상, 접촉 불량, 실린더 노후 …
오류 유형 3	베벨기어 마모, 레이저 용접 변색, 브레이크 실린더 마모, 센서 오류 …
오류 유형 4	히터 단선, 과다 불량 발생, 히터 불량 발생, 실린더 척크발 이상, …
오류 유형 5	실린더 빠짐, 에어 누수, 배선 단선, 스프링 파손, 절연 단선 …

는 것을 확인할 수 있었다.

코사인 유사도 기반으로 구절들과 구절의 군집을 하나의 오류 유형으로 시각화한 결과는 그림 3의 (a)와 같다. 구절 네트워크 구축 결과, 설비 정지로 인한 오류를 의미하는 오류 유형 1과 설비의 오작동으로 인한 오류를 의미하는 오류 유형 3이 연결된 것을 확인할 수 있었다. 또한, 설비를 구성하는 부품의 부적절한 위치에 의해 야기된 오류 유형을 의미하는 오류 유형 2는 부품의 단선을 의미하는 오류 유형 4와 부품의 마모를 의미하는 오류 유형 5와 연관이 있는 것을 확인하였다.

그림 3의 (b)는 피어슨 유사도를 기반으로 한 군집화 결과를 시각화한 것이다. (a)와 달리 모든 유형이 연결된 것을 확인하였다. 이것은 표 4의 결과와 같이 유사하지 않은 구절이 동일한 오류 유형에 할당되어 있는 것을 의미한다. 이를 통해, 코사인 유사도 기반의 군집화 결과가 피어슨 기반의 군집화 결과보다 하나의 오류 유형 안에서 구절들이 더욱 긴밀히 연결된 것을 발견할 수 있다. 이것은 유사한 구절이 동일한 오류 유형에 할당되어 있는 것을 의미하고, 코사인 유사도를 기반으로 한 군집화 결과가 피어슨 유사도 군집화 결과보다 오류 유형 구축을 위해서 더 적합하다는 것을 의미한다.

4. 결 론

본 연구는 설비 오류 유형을 정확하게 분류하지 못한 실제 제조 기업의 문제점에 주목하였다. 설비 오류 유형을 정확하게 파악하기 위하여 단어의 쓰임새를 정의하고 구분한 용도 딕셔너리를 활용한 구절 추출 방법과 추출된 구절 사이의 유사도를 계산하여 네트워크를 구축하는 방법을 제안하였다. 실제 수집된 데이터를 활용하여 제안 방법의 유의성을 확인한 결과, 설비 오류 유형을 정확히 표현할 수 있는 구절이 추출되는 것을 확인하였으며, 유사도를 기반으로 구절 사이의 네트워크를 구축하여 설비 오류 유형을 파악한 결과 의미론적으로 유사한 구절들이 하나의 오류 유형으로 묶이는 것을 확인할 수 있었다. 특히, 유사한 구절을 하나의 설비 오류 유형으로 군집하는 결과에서 코사인 유사도가 피어슨 유사도보다 우수한 결과를 도출하는 것을 정성적으로 확인하였다.

본 논문에서 제안된 비정형 텍스트 데이터로부터 유의미한 구절 추출 방법과 유사도 기반 네트워크 구축 방법을 활용하여 설비 오류의 정확한 원인을 파악하고 설비 오류 유형을 구조화 할 수 있을 것을 기대한다.

참고문헌(Reference)

[1] Seifi M, Salem A, Beuth J, Harrysson O, and Lewandowski J. J, “Overview of Materials Qualification Needs for Metal Additive Manufacturing.“ The Journal of The Minerals, Metals & Materials Society, Vol. 68, No. 3, pp. 747-764, 2016. https://doi.org/10.1007/s11837-015-1810-0

[2] Seong Jun Kim, Byung Hak Choe, and Woo sik Kim, “Prognostics for Industry 4.0 and Its Application to Fitness-for-Service Assessment of Corroded Gas Pipelines.” Journal of the Korean Society for Quality Management, Vol. 45, No. 4, pp. 649-664, 2017. https://doi.org/10.7469/JKSQM.2017.45.4.649.

[3] Gee Wook Song, Woo Sung Choi, Wanjae Kim, and Nam Gun Jung, “Damage Analysis for Last-Stage Blade of Low-Pressure Turbine.” Transactions of the Korean Society of Mechanical Engineers B, Vol. 37, No. 12, pp. 1153-1157, 2013. http://dx.doi.org/10.3795/KSME-B.2013.37.12.1153

[4] Wang, Chen, Hoang Tam Vo, and Peng Ni, ″An IoT Application for Fault Diagnosis and Prediction.″ In Proceedings of the IEEE International Conference on Data Science and Data Intensive Systems, pp. 726-731, 2015. https://doi.org/10.1109/DSDIS.2015.97

[5] Ju Seop Park, Soon Goo Hong, and Na Rang Kim, “A Development Plan for Co-creation-based Smart City through the Trend Analysis of Internet of Things.” Journal of the Korea Industrial Information Systems Research, Vol. 21, No. 4, pp. 67-78, 2016. http://dx.doi.org./10.9723/jksiis.2016.21.4.067

[6] Bok Hee Lee, Kang Hee Lee, Tae Ki Kim, Han Soo Kim, “A Study on the Present State and Consistent use of Terminologies Concerning Grounding.” Journal of the Korean Institute of Illuminating and Electrical Installation Engineers, Vol. 27, No. 4, pp. 81-87, 2013. https://doi.org/10.5207/JIEIE.2013.27.4.081

[7] Ur-Rahman, Nadeem, and Jennifer A. Harding, “Textual Data Mining for Industrial Knowledge Management and Text Classification: a Business Oriented Approach.” Expert Systems with Applications, Vol. 39, No. 5, pp.

4729-4739, 2012.
https://doi.org/10.1016/j.eswa.2011.09.124

[8] Dino Isa, Lam Hong Lee, V.P. Kallimani, and R. Rajkumar, “Text Document Preprocessing With the Bayes Formula for Classification Using the Support Vector Machine.” IEEE Transactions on Knowledge and Data engineering, Vol. 20, No. 9, p.1264-1272, 2008.
https://doi.org/10.1109/TKDE.2008.76

[9] Rossant C, Goodman D. F, Platkiewicz J, and Brette R, “Automatic Fitting of Spiking Neuron Models to Electrophysiological Recordings.” Frontiers in Neuroinformatics, Vol. 4, No. 2, pp. 1-10, 2010.
https://doi.org/10.3389/neuro.11.002.2010

[10] Dhillon, Inderjit S, and Dharmendra S. Modha., “Concept Decompositions for Large Sparse Text Data using Clustering.” Machine Learning, Vol. 42, No. 1-2, pp. 143-175, 2001.
https://doi.org/10.1023/A:1007612920971

[11] Nam gyu Kim, Dong hoon Lee, Ho chang Choi, and Wong William Xiu Shun, “Investigations on Techniques and Applications of Text Analytics.” The Journal of Korean Institute of Communications and Information Sciences, Vol. 42, No. 2, pp. 471-492, 2017.
https://doi.org/10.7840/kics.2017.42.2.471

[12] Yung Shen Lin, Jung Yi Jiang, and Shie Jue Lee, “A Similarity Measure for Text Classification and Clustering.” IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 7, pp. 1575-1590, 2014.
https://doi.org/10.1109/TKDE.2013.19

[13] Monedero I, Biscarri F, León C, Guerrero J. I, Biscarri J, and Millán R, “Detection of Frauds and Other Non-Technical Losses in a Power Utility using Pearson Coefficient, Bayesian Networks and Decision Trees.” International Journal of Electrical Power & Energy Systems, Vol. 34, No. 1, pp. 90-98, 2012.
https://doi.org/10.1016/j.ijepes.2011.09.009

[14] 우진공업, http://www.ngkntk.co.kr

[15] Gephi, https://gephi.org

[16] Blondel V. D, Guillaume J. L, Lambiotte R, and febvre E, “Fast Unfolding of Communities in Large Networks.” Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008, No. 10, pp. 10008, 2008.
https://doi.org/10.1088/1742-5468/2008/10/P10008

[17] Yoo sin Kim, Sung Gwan Hong, Hee Joo Kang, and Seung Ryul Jeong, ″Electronic-Composit Consumer Sentiment Index(CCSI) development by Social Bigdata Analysis.″ Journal of Internet Computing and Services, Vol. 18, No. 4, pp. 121-131, 2017.
https://doi.org/10.7472/jksii.2017.18.4.121

[18] Tae Soo Park and Ok Ran Jeong, “Event Detection System Using Twitter Data.” Journal of Internet Computing and Services, Vol. 17, No. 6, pp. 153-158, 2016.
https://doi.org/10.7472/jksii.2016.17.6.153.

[19] Yong Woong Lee, Se Han Kim, Kyo Hun Son, In Hwan Lee, and Chang Sun Shin, “Implementation of Failure-Diagnostic Context-awareness Middleware for Support Highly Reliable USN Application Service.” Journal of Internet Computing and Services, Vol. 12, No. 3, pp. 1-16, 2011.
https://doi.org/10.7472/jksii.2015.16.4.71

[20] Wei Ji and Lihui Wang, “Big Data Analytics Based Fault Prediction for Shop Floor Scheduling.” Journal of Manufacturing Systems, Vol. 43, No. 1, pp. 187-194, 2017.
https://doi.org/10.1016/j.jmsy.2017.03.008

[21] Huma Lodhi, Craig Saunders, John Shawe- Taylor, Nello Cristianini, and Chris Watkins, ″Text Classification Using String Kernels.″ Journal of Machine Learning Research, Vol. 2, pp. 419-444, 2002.
https://doi.org/10.1162/153244302760200687

◐ 저 자 소 개 ◑

노 영 훈(Younghoon Roh)
2012년~현재 경기대학교 산업경영공학과
2018년~현재 (주)워드바이스 마케팅 매니저
관심분야: 스마트 팩토리, 머신러닝
E-mail: yesrohyh@gmail.com

최 은 영(Eunyoung Choi)
2015년~현재 경기대학교 산업경영공학과
관심분야: 스마트 팩토리, 텍스트 마이닝
E-mail: eychoi506@gmail.com

최 예 림(Yerim Choi)
2010년 서울대학교 산업공학과(공학사)
2016년 서울대학교 산업공학과(공학박사)
2016년~2017년 네이버랩스 Data Scientist
2017년~현재 경기대학교 산업경영공학과 조교수
관심분야: 인공지능/머신러닝, 빅데이터 기반의 인간 모델링
E-mail: yrchoi@kgu.ac.kr