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

장비점검 일지의 비정형 데이터분석을 통한 고장 대응 효율화 사례 연구☆

Unstructured Data Analysis using Equipment Check Ledger: A Case Study in Telecom Domain

주 연 진[1] 김 유 신[2] 정 승 렬[1*]
Yeonjin Ju Yoosin Kim Seung Ryul Jeong

요 약

비정형 데이터의 수집, 분석 그리고 활용에 대한 필요성이 대두되고 있지만 여전히 비정형 데이터를 효과적으로 활용하지 못하고 있는 실정이다. 본 연구에서는 국내 유수 이동통신 기업의 통신 시설장비 점검 시스템에 기록된 비정형데이터를 분석하여 장비고장 대응과 예방에 적극 활용할 수 있는 기반을 만들고자 하였고, 약 220만 건의 작업일지 데이터를 텍스트 마이닝을 통해 구조화/정형화 하였다. 이를 위해 장비 고장과 관련된 4가지 분석 프레임, 고장인지, 고장원인, 고장대상, 조치결과를 구성하였고 분석 결과로는 크게 3가지의 효율화 방안과 관련한 인사이트를 얻을 수 있었다. 첫 번째로는 신속한 조치를 통한 시간 단축을 도모하고, 두 번째로는 고장장비 Unit 수요를 예측하고, 마지막으로 현장 출동의 최소화를 지원할 수 있을 것으로 기대되었다. 결론적으로, 본 사례연구는 통신시설 장비 고장 대응을 위해 데이터 분석 대상을 정형 데이터뿐만 아니라 장비일지라는 비정형 빅데이터로도 범위를 확장했으며, 이를 분석에 활용하기 위해 처음으로 텍스트 마이닝을 시도를 했다는데 의의를 가진다. 또한 N사는 정형 데이터 뿐 만아니라 년 80만 건씩 축적되던 비정형 데이터의 활용 가치를 확인할 수 있던 기회를 가졌으며, 향후 비정형 데이터의 활용 방안에 대한 발전 방향 그리고 추후의 정형 데이터와의 연계 분석 방안 등에 대한 가이드를 확보할 수 있었다.

☞ 주제어 : 장비점검일지, 비정형분석, 통신시설, 고장예방

ABSTRACT

As the importance of the use and analysis of big data is emerging, there is a growing interest in natural language processing techniques for unstructured data such as news articles and comments. Particularly, as the collection of big data becomes possible, data mining techniques capable of pre-processing and analyzing data are emerging. In this case study with a telecom company, we propose a methodology how to formalize unstructured data using text mining. The domain is determined as equipment failure and the data is about 2.2 million equipment check ledger data. Data on equipment failures by 800,000 per year is accumulated in the equipment check ledger. The equipment check ledger coexist with both formal and unstructured data. Although formal data can be easily used for analysis, unstructured data is difficult to be used immediately for analysis. However, in unstructured data, there is a high possibility that important information. Because it can be contained that is not written in a formal. Therefore, in this study, we study to develop digital transformation method for unstructured data in equipment check ledger.

☞ keyword : Equipment Ledger, Unstructured Data, Telecom Domain, Anomaly Detection

1. 서 론

4차 산업혁명은 '데이터 혁명'이라해도 과언이 아닐만큼 데이터의 속도, 양 그리고 다양성이 지속적으로 확산되고 있다. 이러한 빅데이터는 IT 과학기술의 발전으로 수집 및 처리가 원활해졌으며, 그만큼 빅데이터 활용에 대한 관심이 높아지고 있지만, 데이터의 저장 및 분석 관련 기술에 비해 데이터가 증가 속도는 더 급격하다. 특히 인터넷 환경에서 파일, 이메일, 동영상 등과 같이 구조화되지 않은 비정형 정보들이 폭발적으로 증가하고 있는데 (김종현, 2013), 그 중에서도 비정형 텍스트 데이터는 기업환경에서도 매우 많은 비중으로 생성 활용되고 있지만, 정작 빅데이터 분석영역에서는 아직까지 초기 단계 머물러 있는 실정이다. 비정형 텍스트 데이터는 텍스트를 포함한 숫자나 날짜 형식의 데이터 모두를 포괄한다. 비정형 데이터 분석을 위해 정형화가 불가피한데, 질적내용

1 Business IT Graduate School, Kookmin University, Seoul, South Korea
2 Altimedia Chief Data Scientist, Alticast, Seoul, South Korea
* Corresponding author: (srjeong@kookmin.ac.kr)
[Received 10 September 2019, Reviewed 23 September 2019(R2 7 November 2019), Accepted 7 December 2019]
☆ 본 논문은 주연진의 석사학위논문을 바탕으로 작성되었음

분석이 방법론중 하나에 해당된다. 질적 내용 분석 이란 문서 자료 뿐아니라 그림, 상징적 기호, 의사소통과 같은 모든 종류의 자료를 활용하는 연구방법 중 하나로 발전해왔다(최성호 외, 2016). Weber(1985)는 "텍스트로부터 타당한 추론을 끌어내기 위해 일단의 절차를 사용하는 연구방법"을 내용분석으로 정의하며 질적연구의 타당성을 언급하였다.

내용분석은 질적분석과 양적분석이 적절하게 조합되어 이루어져야한다. 양적내용 분석이 키워드 빈도수 기반이라는 비판과 질적 내용분석이 체계적이지 못하고 추상적이라는 비판보다는 단점을 최소화하여 함께 활용할 수 있는 방안을 찾아야한다. 이를 보완하고자 한 연구에서는 분류된 토픽만 가지고 키워드를 분석하는 것은 한계가 있다고 제시하며 문서와 토픽, 키워드의 확률분포를 활용하여 각 주제에 대한 단어 및 빈도수 결과와 결합하여 분석하는 방식을 택했다(고명숙, 2017).

따라서 본 연구에서는 키워드기반의 빈도를 활용하는 방법이 아닌 문맥상의 의미를 파악하여 룰을 생성하고 이를 활용하는 구조화 방법을 사용하고자 한다. 문장 내에서의 키워드의 위치 또는 의미를 룰로 정의하여, 기존의 단순한 키워드 빈도를 세는 것이 아닌 룰을 활용한 정형화 분석을 진행한 사례연구를 제시하고자 한다. 본 사례연구를 통해 비정형데이터의 정형화 프로젝트의 현황을 공유하고 발생 가능한 이슈사항을 확인할 수 있다.

2. 관련연구

2.1 비정형 데이터 분석 관련

과학 기술의 발전으로 많은 양의 데이터를 의미하는 빅데이터의 수집, 처리가 가능해졌고 인터넷의 발전과 소셜네트워크서비스의 부상으로 여기서 발생하는 비정형 데이터의 활용도도 매우 높아지고 있다. 이러한 비정형 빅데이터 분석은 금융분야에서는 뉴스 콘텐츠를 분석하여 주가 예측에 활용하기도 하고(김유신 외 2016), 온라인에 게시된 영화 리뷰와 평점을 분석하여 영화의 흥행 가능성을 예측하기도 하였으며(Kim 외, 2018), SNS에 게시된 사회 갈등 이슈에 대한 찬반 의견을 분류하는 연구에 텍스트 마이닝이 이용되기도 하고(강선아 외, 2015) 고객서비스센터에 접수된 고객의 불만 텍스트를 분석하여 서비스 개선에 활용하기도 하였다(김유신 외 2016).

비정형 데이터를 분석하기 위해서는 데이터를 구조화하고 정형화해야한다. 배유진(2014)는 방대한 정보를 구조화시키는 것은 비정형의 콘텐츠를 정형화시키고 정형화된 데이터는 유형화하여 체계화시키는 것이라고 언급한 바 있다. 곧 전달하려는 내용을 정리하고 배열하여 분류하는 과정을 구조화라 할 수 있고, 정보의 구조화란 전달하려는 정보의 내용 요소들을 정리 및 배열하여 통일된 조직으로 만드는 과정을 말한다. 구조화를 활용하여 효율적으로 정보를 기억하거나 신속하게 정보 전달을 할 수 있다. 특히 많은 내용을 전달해야 하거나 내용 구성요소 들의 관계가 복잡하면 구조화의 효율성이 대두된다. 본 연구에서는 비정형 데이터의 활용을 위해 데이터의 정형화를 진행하고, 향후 연구과제로 데이터의 구조화를 진행하고자 한다.

2.2 장비 고장 분야 관련

빅데이터 분석은 다양한 분야에서 활용되고 있으며, 특히 고장 장비와 관련한 분야에서도 많은 연구가 진행되고 있다.

한 연구에서는 전자장비의 사용중 고장에 대비해 고려되어야 하는 예비 부품, 장비의 품질보증 등의 측면에서 고장률 예측의 필요성을 언급하며, 고장률을 예측하는 모델을 활용하여 연구를 진행했다(주철원 외, 1991). 또한, 부식 가스로 인한 전기 제어 손상은 시스템 운영에 위험을 주고, 예방비용보다 더 많은 비용을 수리비로 사용하게 된다고 언급하며, 장비 결함 예방의 중요성을 강조한 연구도 존재했다(이상윤, 2008). 그 외에도 수도권 지하철 역사 내에 설치된 공용중계기에서 발생되는 전원 불량과 같은 장애요인에 대한 연구를 진행하였고, 또한 장비와 시설에서 발생하는 불량 또는 지하철 터널 구간에서 발생되는 통신 서비스에 대한 문제점을 조사하는 등의 활발한 연구가 진행되고 있다(신지윤 외, 2003). 그리고 고장데이터와 A/S데이터를 기반으로 장비의 신뢰성을 예측하고 평가하는 연구(윤변동 외, 2016)를 진행하여 장비 데이터의 활용성에 대해 확인할 수 있었다.

연구들을 살펴보았을 때, 장비 고장과 관련하여 연구들이 활발하게 진행되고 있다. 장비 고장에 대한 데이터들이 활용성이 있으며, 장비 고장과 관련한 데이터 분석에 대한 니즈가 존재하고 중요하다는 것을 확인할 수 있었으며, 이를 어떻게 활용할 것인가에 대한 고찰이 필요하다. 그러나, 관련 연구들의 분석은 정형 데이터 위주였으며, 비정형 데이터 분석에 대한 내용은 확인할 수 없었다. 그렇기 때문에 본 연구에서는 비정형 데이터의 정형화를 통해 기존 연구와는 차별을 두어 비정형 데이터를

활용한 사례 연구를 진행하고자 한다.

2.3 분석 툴 소개

본 연구에서 사용될 툴인, IBM Watson Explorer Content Analytics는 정형 및 비정형 데이터를 탐구하고 분석하며 이해하고, 그 결과를 시각화 하여 목적 별로 신속하고 의미 있는 인사이트를 획득할 수 있도록 도와주는 툴에 해당되며, 자체적으로 형태소 분석을 진행할 수 있기 때문에 활용도가 높다. 또한, 문맥상의 의미 또는 규칙을 설정할 수 있는 룰을 생성할 수 있어 단순 키워드 빈도뿐만 아니라, 다른 방향의 활용도 가능하다는 장점을 가지고 있다. 본 연구에서는 IBM WEX(Watson Explorer) 11.0.2 버전을 사용하였다.

3. 사례연구

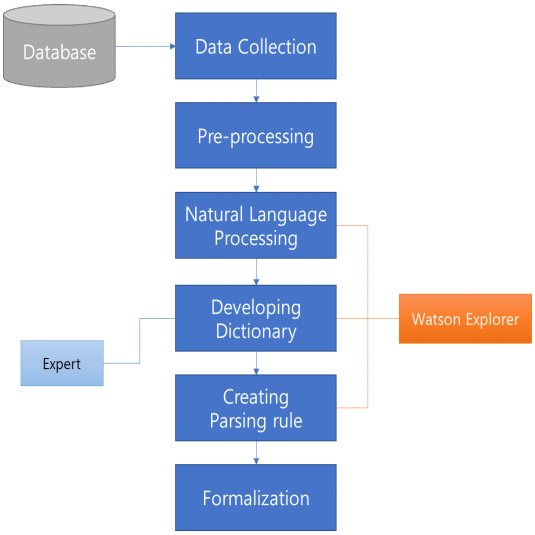
3.1 개요: 통신장비운용회사 N사의 비정형데이터 정형화 프로젝트 비정형 데이터 분석 관련

N사는 전국의 유무선 네트워크를 운용하는 네트워크 장비회사이다. 네트워크 장비의 고장이 발생하게 되면 N사의 현장작업자가 출동하게 된다. 이때 현장작업자는 관제 센터로부터 트러블 티켓(Trouble ticket)을 전달받은 후 출동한다. 보통 트러블 티켓에 적혀진 정보는 출동 장소(국소명)과 알람이다. 출동하는 횟수에 따라 비용이 발생하기 때문에, 한 번의 출동으로 문제를 해결하면 좋지만, 작업자들이 전달받는 트러블 티켓에 담긴 알람과 국소명만으로는 원인을 파악하기는 어렵다. 설령 문제점을 파악했다 하더라도 경험이 없는 작업자들이 한번에 문제 해결을 위한 조치방법을 예상하기에는 어려움이 따른다. 오래된 작업자의 경우, 노하우를 발휘할 수 있는 경력과 경험이 존재하지만, 이는 특정 작업자에 한정된다. 그렇기 때문에 현장 출동자 중 하위 50%는 망관리 센터의 가이드를 제시받고 있는 실정이다. 현장 작업자들은 출동 후에 작업일지를 작성한다. 작업일지에는 국소명, 업무구분, 작업자 등의 정형화된 필드가 존재하며, 비정형 필드로 세부항목이 존재한다. 세부항목에는 작업 내용 및 결과를 작성하도록 되어 있다. 이러한 작업일지는 년 80만 건씩 DB에 축적되고 있다.

N사에서 분석에 사용하고 있는 데이터는 정형에 한정되어 있으며, 대부분의 중요한 문제 발생 원인 및 해결방안은 정형이 아닌 비정형 필드로 관리되고 있기 때문에 비정형 데이터의 활용이 불가피함을 인지하고 이를 활용할 수 있도록 하고자 한다. 따라서 본 연구를 통해 그동안 데이터 분석에서 제외되어온 비정형데이터를 정형화하여 분석에 활용될 수 있도록 하고, 비정형 데이터의 정형화 분석은 고장 원인에 따른 조치사항을 확인하는 것뿐만 아니라 어느 국소에서 어떠한 부품 혹은 장비에서 빈번하게 문제가 발생하는지를 더 자세히 확인하고자 한다. 이는 고장 원인을 미리 예방하고자 하고, 향후 부품 및 제품 관련하여 투자 방향을 재설정할 수 있는데 도움이 될 것으로 기대된다.

3.1 비정형 빅데이터 정형화 방법론

본 연구는 비정형 빅데이터 정형화하기 위해 다음 그림 1의 과정을 통해 진행되었다.



(그림 1) 비정형 데이터의 정형화 프로세스
(Figure 1) Unstructured data processing

3.2.1 데이터 수집

N사의 현장 출동자는 출동 후 작업일지를 통해 작업내용 및 조치 결과 등을 기록해야한다. 작업일지에는 정형화 필드와 더불어 세부 작업 내용을 작성할 수 있는 비정형 필드가 존재한다.

작업일지는 데이터베이스에 년 80만 건씩 데이터가 축적되고 있었으며, 본 연구에서는 N사의 DB로부터 3년간의 데이터인 약 220만 건을 수집하였다. 데이터에는 총 24개의 컬럼이 존재했으며, 22개의 정형 필드와 2개의 비

정형 필드로 구성되어 있다. 본 연구에서는 정형필드인 "국소명", "본부", " 작업자", " 시설구분", "일자", "팀", "파트", "담당업무", "업무구분", "조치구분", "조치내역", "조치세부내역" 총 12개와 "제목", "세부내역" 비정형 필드 2개를 활용하였다.

3.2.2 전처리

수집된 데이터 작업일지 약 220만건의 비정형 필드는 정해진 규격이 존재하지 않아 작업자 마다 규칙적인 형식이 존재하지 않았으며, 작성 내용 또한 가지각색이였다. 또한 "세부내역"을 아예 작성하지 않은 작업자도 존재하였다.

우선, "제목" 또는 "세부내역"을 작성하지 않은 작업자들의 문서들을 정형화 분석 대상에서 제외하는 작업을 진행하였고, 작업자마다 사용한 구분자가 달라 분석에 어려움을 있음을 인지하고, 특수문자를 제거하는 작업을 진행하였다. 분석 대상 데이터를 정제하고, 더불어 툴에 임포트하기 위해 CSV로 파일형식을 변환하는 작업을 진행하였다.

3.2.3 자연어 처리(NLP)

본 연구에서 활용되고 있는 툴 WEX는 자연어를 처리할 수 있는 형태소 분석 기능이 포함되어 있다. 제공하는 24개의 Part of Speech 중 명사와 명사시퀀스를 활용하여, 대상 데이터에 포함된 핵심 단어와 복합명사 후보군(명사시퀀스)가 포함되어 있는지에 확인해보았다. 비정형 필드인 '제목'과 '세부내역'에서 고빈도 10,000개의 명사와 명사시퀀스를 추출하였고, 이를 통해 작업일지에서 어떤 유의미한 분석을 진행할 수 있을지에 대해 검토하였다. 대부분 작업일지의 '세부내역'에는 노후, 고온, 불량 등의 고장의 원인과 관련된 단어들과 장비명, Unit명과 같이 고장 위치에 대한 내용, 그리고 제어, 교체, 탈실장 등의 조치결과에 대한 내용이 존재했고, '제목'에는 알람 명이 기재되어있어 고장원인을 인지할 수 있었다. 이를 통해 장비 고장의 분석 프레임을 그림 2와 같이 정의하였다.

1단계	2단계	3단계	4단계
고장 인지 (Alarm)	고장 환경(원인)	고장 위치 (Unit, 장비)	조치 결과

(그림 2) 장비 고장의 분석 프레임
(Figure 2) Equipment Diagnose Process

3.2.4 사전 구축

'장비 고장의 분석 프레임'을 활용하여 장비 고장 분야의 사전을 구축하였다. 사전 구축을 위해 비정형 필드인 제목과 세부항목, 그리고 기존 분류 체계 등을 활용하였으며, 이를 통해 제목과 세부항목과 같은 비정형 데이터를 정형화할 수 있는 기반을 마련하였다. 사전은 고장인지를 위한 알람사전과 고장진단을 위한 원인사전 그리고 고장 부위 확인을 위한 장비사전·Unit사전·Unit이외사전이 속하며, 최종적으로 동의어를 포함하여 1,757단어로 8개의 사전을 구축하였다.

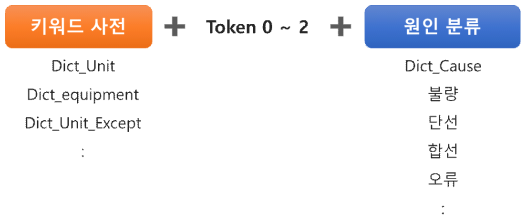
3.2.5 파싱룰 구축

본 연구에서 파싱룰을 구축하는데 있어 구축한 사전을 이용하였다. 문장 안에서 의미 있다고 여겨지는 문자열을 최소단위인 토큰으로 분류하고 이들을 구문 트리로 재구성하는 구문 분석 과정을 파싱이라 정의한다. 이를 활용하여 룰을 만드는 것을 파싱 룰이라 일컫는다. 즉 파싱룰이란, 문장 안에서 유의미한 정보를 추출하기 위해 패턴 또는 규칙성을 찾는 것이다.

파싱룰을 구축하여 활용하면 좋은 점은 크게 3가지로 제시할 수 있다. 첫째로 룰을 적용하여 문맥상의 의미 파악을 용이하게 할 수 있다는 점이다. 키워드의 빈도만 가지고 분석하게되면 문맥상의 의미를 파악할 수 없어 해석에 있어 오류를 범할 수도 있다. 예를 들어, "오늘은 비가오고, 내일은 눈이온다." 문장에서, 우리는 '오늘'의 날씨는 '비', '내일'의 날씨는 '눈'이라는 정보를 얻고자한다. 하지만, 빈도수로 카운트 하였을때, 이문장에서는 시점의 속성은 '오늘'과 '내일', 날씨의 속성은 '비'와 '눈'으로 추출된다. 이처럼 한 문장 안에서 추출하려는 속성 값이 각각 두 개씩 나왔을 경우, 이들 간의 관계를 설정하는 것이 쉽지만은 않다. 이를 룰을 활용하여 시점 속성과 날씨 속성이 가까이 존재하는 것을 관계설정하는 룰을 활용하여, "오늘-비", "내일-눈"을 추출할 수 있도록 조건을 설정할 수 있다. 두 번째로는, 룰을 설정하게 되면 한글어의 특성상 분석에 있어 어려움을 주는 문장의 조사 혹은 수식어로부터 조금은 벗어날 수 있다는 것이다. 가령 예를 들면, "오늘 비가 온다.", "오늘은 비옴.", "비오는 오늘" 세 문장에서 도출할 수 있는 정보는 시점 정보는 "오늘", 그리고 날씨 정보는 "비" 임을 알 수 있다. 결과적으로 파싱룰을 통해서 키워드 사이의 토큰 조건을 주고 형태가 다른 세 문장에서 동일한 정보인 "오늘-비"를 정의하여

추출하면, 이의 빈도가 3회로 측정될 수 있다. 통해 특정 케이스에 대한 예외처리를 할 수 있다. 사전에서 특정 단어만 제외하고 싶을 때, 혹은 다른 경우의 수가 존재할 때 룰을 통해 이를 조정할 수 있다. 예를 들어, 날씨 중 "태풍"을 제외하고 싶으면, 이를 제외하는 조건 룰을 만들어 적용 시킨다. 그러면, 문장 혹은 문서에 "태풍"라는 단어가 등장하더라도 이에 대한 키워드 추출 및 분석은 진행되지 않을 것이다.

본 연구에서는 총 95개의 룰을 만들어 활용했으며, 각각의 룰은 분야별로 활용될 수 있도록 종속적, 독립적으로 생성되었다. 예를 들면, 고장 원인과 관련된 룰의 경우, 룰 정의를 위해 원인의 키워드가 될 수 있는 '불량', "오류", "단선", "노후", "합선" 등의 키워드를 선정하였고, 그 앞에 올 수 있는 단어들의 후보를 2개와 3개 사이로 추출하여 검토하였다. 그 결과, UNIT명(장비명 또는 부품)이 대부분을 차지하였으며, 그 외의 기타 단어들을 하나의 분류(Unit_Except)로 추출할 수 있었다. 이를 활용해 다음 그림 3과 같은 룰을 생성하였다.



(그림 3) 고장원인과 관련된 파싱룰 정의 예시
(Figure 3) Keyword Parsing Examples for Broken Cause

키워드 사전은 불량의 원인이 될 수 있는 3 가지 :Dict_Unit(유닛사전),Dict_equipment(장비사전), Dict_Unit_Except(유닛 이외 사전)을 활용하였다. 그 후에 나올 수 있는 최소 단위인 토큰의 개수를 최소 0에서 최대 2로 설정해두었고, 마지막에는 고장 원인의 분류에 해당하는 Dict_Cause(원인사전)을 활용하였다. 두 개의 사전 및 정규 표현식 그리고 토큰의 조건을 활용하여 바로 앞에 나온 키워드가 아니더라도 사전에 존재하는 단어라면 룰에 적용될 수 있도록 하였다. (그림 2)를 적용한 룰의 결과를 예시로 확인해 본다면 Unit사전명에 해당하는 "DAU"와 원인분류 사전의 "불량"이 같이 도출된 경우, 조건으로 설정한 토큰 개수 안에서 띄어쓰기 과 상관없이 같은 정형화 값을 도출할 수 있다. "DAU 불량"과 "DAU불량", 뿐만 아니라 "DAU가 불량으로 판정됨" 등이 "DAU 불량"으로 동일하게 정형화 될 수 있다.

3.2.6 정형화

본 연구를 통해 작업일지의 비정형 데이터를 정형화하는 연구를 진행하였다. 비정형 필드는 "제목"과 "세부항목"이 해당되였으며, 그 안에서 유의미한 항목 고장 알람, 원인, 대상 그리고 조치결과에 대한 4가지 분석 프레임을 가지고 정형화를 진행했다. 이 프레임을 가지고 사전을 구축하였으며, 구축된 사전을 활용하여 룰을 생성하였다. 룰을 통해 분석하고자하는 항목을 일컫는 패싯(Facet)을 생성하였으며, 각각의 패싯에 원하는 조건의 룰이 태깅될 수 있도록 설정하였다. 그 결과 분석 툴인 WEX Miner 화면에서 패싯을 통해 정형화된 항목들이 전체 문서에 대해 몇 건 정도 출현하였는지 빈도수를 확인할 수 있었다. 또한, 기존 정형 필드와 정형화된 필드를 함께 분석 할 수 있는 패싯쌍(Facet Pair) 기능을 통해 비정형 필드의 정형화 및 분석 활용 가능성을 확인할 수 있었다.

3.3 프로젝트 진행 이슈

3.3.1 데이터 이슈

본 연구에서는 데이터 품질과 관련한 이슈가 존재했다. 데이터 품질이란, 데이터 사용자에 의해 사용하기 적합한 데이터(Wang et al,1996), 사용자에 의해서 정의되는 것으로 비즈니스 목적에 대한 적합성의 정도로서 측정되는 것(Kelly,1997)이라 정의된다. 이러한 측면에 있어 본 연구에서 데이터로 활용된 작업일지는 데이터 품질에 적합하지 않았다. 비즈니스 목적에 대한 적합성의 정도를 측정할 수 있는 기준이 없었으며, 또한 어떤 내용이 작성되어야하는지에 대한 자세한 가이드가 제시되고 있지 않았다. 그렇기 때문에 작업자들마다 각자의 양식과 주관을 가지고 작성하였으며, 그들이 작성한 일지의 양과 질은 개개인마다 달랐다. 또한, 작업일지를 작성하는 현장 출동자는 자신들이 작성한 작업일지에 대한 활용성에 대해 인지하고 있지 못했기 때문에 내용을 부실하게 작성하거나 혹은 작성하지 않는 경우도 많았다. 혹여, 공문의 내용을 그대로 복사해 넣은 경우도 많이 존재했다.

본 연구를 통해 세부내역의 필드가 비정형임에도 불구하고 형식 및 표준양식이 존재하지 않아 발생된 과거 데이터 품질 이슈를 확인하였으며, 이를 해결하기 위해 세부항목 및 제목에 대한 가이드를 제공하는 안을 제시하였다. 향후 데이터의 품질을 높이기 위해 네트워크 운용회사 N사의 표준 양식 및 가이드를 제시하였다. 비정형 필드로 정해져 있는 제목과 세부내역에 대해 제목에는

알람, 세부내역에는 고장 원인, 고장 부위(대상) 그리고 조치결과에 대한 내용이 포함될 수 있도록 교육이 필요하다.

3.3.2 분석 툴 이슈

Watson Explorer는 비정형 데이터를 정형화하기위해 단지 키워드 빈도가 아닌 사전과 룰의 조합으로 원하는 패턴의 정형화를 진행할 수 있는 툴이다. 형태소 분석 기능을 통해 기존의 분석이 어려웠던 복합명사와 관련된 분석도 일부 진행할 수 있었다. 그러나, 본 연구에서 툴의 장점 뿐 아니라 단점도 확인할 수 있었다. 장비 고장 분야의 특성상 문서의 키워드들은 한글과 영어가 혼재되어 존재했고, 또한 장비 또는 Unit명이 영어와 한글로 모두 정의되는 경우가 존재했는데, 툴의 특성상 한글 사전과 영어 사전을 따로 만들어 사용해야하는 하는 이슈가 있었다. 처음에는 이 이슈를 인지하지 못해 하나의 사전을 한국어로 구축하였는데, 영어의 대소문자 구별 및 띄어쓰기가 되어있는 복합 단어에 대한 동의어 처리가 너무 많이 필요하다는 것을 인지하게 되었다. 그로 인해, 사전 구축에 있어 생각보다 많은 시간을 할애하는 등의 어려움을 겪었다.

3.3.3 프로젝트 협업 이슈

비정형 데이터를 정형화하는 본 연구에서는 분류 체계 구축 및 룰 생성을 위해 해당 분야의 전문적 지식이 필요했고, 이에 대해 전문가와의 협업이 절대적으로 필요했다. 하지만, 처음에는 전문가들의 협조를 얻고 설득시키는데 어려움이 있었다. 이유는 다음과 같이 2가지로 나뉘었다. 첫 번째 문제는 작업의 필요성에 대한 이해문제이다. 처음 사전을 구축하기위해 전체 비정형 데이터의 형태소분석(NLP)을 통해 도출된 명사와 복합명사 후보군들 중에서 유의미한 값들만 추출하여 분류체계 구축 및 사전 단어를 구성해야하는데, 이에 있어 공수가 많이 요구되었다. 이에 대한 업무를 수행할 수 있는 사람은 전문적인 지식을 가진 해당 업무의 담당자만이 가능했으며, 이들에게 왜 이 작업이 필요한지에 대한 이해를 시키는데 많은 시간이 걸렸다.

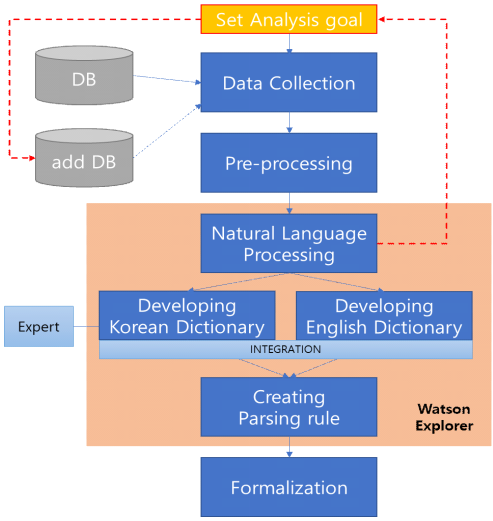
또한, 툴에 대한 교육을 통해 향후 이 작업들이 왜 필요한지 그리고 어떻게 쓰일지에 대해 설명하기 위해 노력했다. 실제적인 사전과 룰을 만드는 작업을 하는 것이 업무는 아니지만, 구축이 완료되었을 때 어떻게 사용해야 하는지 그리고 어떤 식으로 작동되는지에 대한 이해는 필요한 부분이다. 그렇기 때문에, 툴에 대한 설명 및 용도, 사용법을 교육함으로써 해당 전문가들도 자신들이 작업한 내용을 자신의 자리에서 확인하고 피드백을 줄 수 있도록 하였다.

3.3.4 프로젝트 범위 이슈

본 프로젝트는 작업일지 비정형 데이터의 정형화의 목적을 가지고 있었기 때문에 작업일지 DB만을 대상으로 분석이 진행되었다. 하지만, NLP 후 분류체계를 구축하고 분석프레임을 작성하면서 다른 데이터의 연계 필요성이 대두되었고, 그에 대한 고객의 니즈를 파악할 수 있었다. 부품정보 및 구매이력 등의 데이터가 담긴 DB 연계가 된다면 장비이력을 통해 어떤 부품이 가장 많이 교체되었는지 그리고 그 원인이 무엇인지를 더 상세히 파악할 수 있고 또한 부품을 구입했던 이력과 그 트렌드를 확인할 수 있을 것이라 기대된다. 해당 범위를 프로젝트 특성상 데이터 마트를 설계하거나의 작업 및 공수가 중간에 투입이 되기 힘든 상황이었기에 니즈에 대한 필요성만 인지한 채 프로젝트를 진행할 수밖에 없었다.

3.4 이슈 분석을 통해 도출된 정형화 모델

사례연구를 통해 제시된 이슈 중, 데이터 연계 문제 및 사전 구축과 관련하여 다음과 같은 추가적인 프로세스를 그림 4와 같이 제시한다.



(그림 4) 개선 프로세스

(Figure 4) Dictionary Development Proces

먼저, 분석 목표 설정이 우선적으로 진행된 상태에서 분석이 진행되어야하며, 만약 초기에 추가적인 분석 요건 혹은 분석 대상을 파악하지 못한 경우에는 NLP를 통해 분석 프레임을 설정할 때, 추가적인 분석 목표와 이에 해당하는 데이터 소스를 수집할 수 있도록 설계 되어야한다. 이때 한정적인 분석 결과 도출을 피하기 위해 추가적인 데이터 소스의 유입을 감안해야한다. 두 번째로, 같은 툴 안에서 한글사전과 영어사전을 따로 구축해야한다는 것을 인지하고 작업을 진행해야한다. 또한, 어떻게 통합하여 활용할 것인지 또는 다른 방법이 존재하는지에 방안 수립이 필요하다. 한글사전과 영어사전을 분리하되, 다음 depth인 대표어를 같은 단어로 설정하여 통합하는 방법을 이용한다면, 부득이하게 사전을 따로 구축하여도 통합이 용이할 것으로 예상된다. 또한, 이러한 사전 이슈인지는 사전을 활용하여 만드는 룰에도 영향을 줄 수 있기때문에, 사전이 통합에 대한 고려 없이 만들어진 경우, 사전 개수만큼 파싱룰을 생성해주어야 하는 불가피한 상황도 발생할 수 있다. 그렇기에 언어에 따라 달리 구축되어야하는 사전 이슈가 고려됨에 따라 작업의 중복성이 줄어들 것으로 예상된다.

3.5 프로젝트 결과 및 시사점

본 연구에서는 네트워크 운용 회사 N사가 현장작업자가 출동을 할 때, 한정된 정보를 활용하여 원인 추측 및 조치사항을 추천받을 수 있는 시스템을 구축하는데 필요한 비정형데이터의 정형화 프로젝트의 사례연구를 진행하였다. 약 220만 건의 작업일지 데이터를 이용하여 비정형데이터를 정형화하여 현재 분석되고 있는 정형데이터와 더불어 비정형 필드도 함께 분석될 수 있는 바탕을 만들고자 하였다. 비정형 데이터의 정형화를 위해 장비 고장과 관련된 4가지 분석 프레임, 고장인지, 고장원인, 고장대상, 조치결과를 구성하였다. 이러한 프레임 설정을 위해 WEX라는 툴을 활용했으며, 이 툴의 형태소 분석 기능을 활용하여 출현 빈도가 높은 명사와 명사 시퀀스를 추출하였다. 이 결과에 대해 유의미한지에 대한 정의를 전문가와 함께 진행하여 사전과 룰을 구축하고, 정형화를 진행하였다.

본 연구를 통해 크게 3가지의 인사이트를 얻을 수 있었다. 첫 번째로는 신속한 조치를 통한 시간 단축할 수 있게 되었고, 두 번째로는 Unit 수요를 예측할 수 있게 되었고, 마지막으로 현장 출동의 최소화가 이에 해당한다.

3.5.1 고장원인 예측으로 시간 단축

지금까지 현장 작업자는 트러블 티켓을 받게 되면, 기본 부품(Unit)을 챙겨서 출동을 한 뒤,‘ 정확한 원인 및 예상 조치 결과를 추측하여 다시 출동하는 등의 프로세스를 거쳤다. 그러나, 본 연구를 통해 정형화된 Entity간의 상관분석을 통해 현장 추천 또는 고장예방에 활용할 수 있다는 인사이트를 얻었다. 예를 들면, E알람의 원인은 F가 가장 많았으며, 또한 E알람의 조치결과는 H가 가장 많았다는 결과를 통해 출동 이전에 예상원인, 그에 따른 조치사항을 확인하고 출동하여 전보다 빠르게 조치할 수 있게 되었다.

3.5.2 수요 예측을 통한 Unit 재고 관리

비정형 데이터 분석을 통해 알람과 원인에 따른 조치사항 추천으로 보다 신속하게 조치사항을 예측할 수 있다. 이는 곧, 교체 혹은 수급이 필요한 Unit에 대한 재고관리 및 수급 요청 또한 개선될 수 있음을 의미한다. 각 장비마다 필요한 Unit의 종류는 다양하고 이에 대한 모든 수량을 재고로 모두 보유하고 있기는 어렵다. 본 연구를 통해 최근 혹은 특정 기간에 많이 언급된 Unit에 대한 수요 예측 및 재고 관리를 진행 할 수 있으리라 기대된다.

3.5.3 현장 출동 최소화

기존 정형정보와 연계시, 리소스 투입과 관련하여 인사이트를 도출할 수 있다. 예를 들면, 지역별로 고장이 많이 발생하는 다발고장 비교분석을 통해 앞으로의 운용 방향을 설정하는데 도움을 얻을 수 있다. 특히, 주요환경 원인을 분석함으로써 장비의 취약점 및 국소환경 개선으로 "출동 중 자동복귀"등의 알람에 대해서는 현장 출동의 최소화 할 수 있다.

하지만, 본 연구는 한정적인 업무 분야인 장비의 고장에 국한되어 진행되었기 때문에 다른 업무분야 적용에 있어서는 보다 검토가 필요하며, 업무분야별 사전 및 분류 체계 구축에 대한 고려가 필요하다. 이에 따른 전문가들의 작업도 불가피할 것이다. 하지만, 방법론적인 접근은 유사할 것이며, 이러한 연구를 통해 비정형데이터에서만 도출할 수 있는 의미 있는 데이터를 확보할 수 있다면, 이는 인사이트를 얻기위한 분석 결과의 질을 향상시킬 수 있는 하나의 핵심요소가 될 수 있다.

본 연구는 비정형 데이터의 정형화를 키워드 빈도수 기반 뿐만이 아니라 사전과 룰을 활용하였다는 점에서

의미가 있다. 하지만, 여전히 비정형데이터를 정형화에 있어 이슈는 존재한다. 현재 분석되고 있는 문서들에 대한 사전은 본 연구를 통해 구축하였다. 그러나, 이는 절대적인 사전이 아니며, 지속적으로 업데이트되고 관리가 필요한 부분이다. 앞으로 새로 생겨나거나 기존 사전에 존재하지 않은 키워드가 출현한다면 이를 기존 사전에 추가하는 작업 등이 이루어져야한다. 그렇기 때문에 주기적인 사전 고도화에 대한 우려가 존재한다. 그렇기에 이를 보완할 수 있는 방법론 혹은 대안이 필요하다.

결론적으로, 데이터의 품질 문제와 여러 가지 연계 데이터 등의 이슈 등으로 완벽한 결과를 얻을 수는 없었다. 하지만, 본 프로젝트는 N사가 분석 대상을 정형뿐 아니라 비정형 빅데이터로 범위를 확장했으며, 이를 분석에 활용하기 위해 처음으로 정형화를 시도를 했다는데 의의가 있다. 이번 프로젝트로 N사는 정형 데이터 뿐 만아니라 년 80만 건씩 축적되던 비정형 데이터의 활용 가치를 확인할 수 있던 기회를 가졌으며, 향후 비정형 데이터의 활용 방안에 대한 발전 방향 그리고 추후의 정형 데이터와의 연계 분석 방안 등에 대한 첫 발걸음을 내딛은 것이다.

4. 결 론

과학 기술의 발전으로 빅데이터 수집 및 저장이 가능해지면서, 방대한 양의 데이터, 특히 비정형데이터를 처리하고 분석하는데 필요성이 증대되고 있다. 비정형데이터를 분석하지 못해 지금까지 얻지 못했던 유의미한 정보들을 현재 정형으로만 이루어져있는 분석에 함께 활용할 수 있다면, 보다 분석의 질이 향상될 것이며, 활용성도 높아질 것이다.

본 연구는 비정형 데이터의 동향 및 중요성에 대해 언급하고, 이를 활용하기 위한 방안으로 정형화에 대해 다루고 있다. 사례연구를 통해 한 네트워크 운용 회사의 비정형데이터 정형화 프로젝트를 살펴보고, 그 안에서의 작업 과정과 이슈사항들에 대해 알아보았다. 이는 비정형데이터 프로젝트에서의 한계점 혹은 고려해야할 점 등을 미리 인지할 수 있는 기회를 마련했다고 볼 수 있다. 본 연구에서 사례연구로 제시한 네트워크 분야의 고장장비뿐 아니라, 다른 분야의 고장 장비에 대한 효율적인 조치사항을 위해서도 활용될 수 있을 것이라 예상된다. 다만, 도메인에 따른 추가적인 분석과 수집대상에 해당되는 데이터에 대해 고려를 통한 적용이 필요할 것으로 보인다.

연구를 진행하며 비정형 데이터를 정형화하는데 데이터 수집부터 정형화까지 총 6단계의 과정을 거쳤고, 사례연구를 통해 추가적인 프로세스를 제시하였다.

요즘 화두에 오르고 있는 인공지능, 챗봇과 같은 IT기술들로 인해 이 기술들을 뒷받침 해 줄 수 있는 텍스트마이닝, 빅데이터, 자연어 처리 등의 기술 등이 관심을 받고 있다. 본 연구에서 다룬 내용은 비정형 데이터를 정형화하여 정형데이터와의 분석을 진행하고 최종적으로 이를 이용하여 시스템을 구축하는데 목적을 두고 있다. 하지만, 시스템 구축에 도움을 주는것과 더불어 현재 각광받는 새로운 기술과의 융합도 가능할 것으로 예상된다. 예를 들어, 인공지능으로 화두에 오르고 있는 챗봇과의 융합 등이 이에 해당된다. 장비 고장과 관련된 알람명 혹은 원인을 검색하거나 혹은 챗봇에 질문했을 때, 검색 결과 혹은 답변을 얻을 수 있는 활용방안 등이 기대된다.

참고문헌(Reference)

[1] Sun-A Kang, Yoo Sin Kim, Sang Hyun Choi, "Study on the social issue sentiment classification using text mining", Journal of the Korean Data & Information Science Society, Vol.26, No.5, pp. 1167-1173, 2015. http://dx.doi.org/10.7465/jkdi.2015.26.5.1167

[2] Myung-Sook Ko, "Unstructured Data Processing Using Keyword-Based Topic-Oriented Analysis", KIPS Transactions on Software and Data Engineering, Vol.6, No.11 pp. 521-526, 2017. http://dx.doi.org/10.3745/KTSDE.2017.6.11.521

[3] Yoosin Kim, Seung Ryul Jeong, "Intelligent VOC Analyzing System Using Opinion Mining", Journal of Intelligence and Information Systems, Vol.19, No.3, pp. 113-125, 2013. http://dx.doi.org/10.13088/jiis.2013.19.3.113

[4] 김종현, "국내외 금융권 빅데이터 활용사례 및 도입 활성화를 위한 선결과제", 우리금융경영연구소, 금융경제분석 금융연구 2013-03. http://www.wfri.re.kr/

[5] BAE, Youjin, "Structuralization of legal contents with based on electronic documents : The practical use of online administrative appeal", Department of Law, The Graduate School of Ewha Womans Universi, 2014. http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=b8dd44f3afa09973f

fe0bdc3ef48d419

[6] Ji-Yun Shim, Duk-Kyu Park, "Failure Analysis on the Equipment of PCS common - repeater at the Metropolitan Subway", The Journal of the Korea Contents Association, Vol.3, No.2, pp. 78-86,2003. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE00684926

[7] 이상윤, "전자 장비 결함 예방을 위한 부식 가스 제어", Vol.66, pp.11-18, 2008. http://www.ndsl.kr/ndsl/search/detail/article/articleSearchResultDetail.do?cn=JAKO200851935718391

[8] 윤병동, 김근수, 황태완, 김수지, 전병주, "Equipment Reliability Estimation Based on Field Failure Data and Warranty Data", 대한기계학회 춘계학술대회, Vol.20, No.4, pp. 58, 2016. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06671765#

[9] 주철원, 이상복, 김성민, 김경수, "패키지형태에 따른 반도체소자의 고장률예측", Electronics and Telecommunications Trends, Vol.6, No.3, pp. 3-12, 1991. http://dx.doi.org/10.22648/ETRI.1991.J.060301

[10] Robert Philip Weber, "Basic content analysis", Sage Publications, p95,1985

[11] Yoosin Kim, Mingon Kang, Seung. R. Jeong, "Text Mining and Sentiment Analysis for Predicting Box Office Success", KSII Trans. Internet Inf. Syst., Vol.12, No.8, pp. 4090-4102, 2018. http://dx.doi.org/10.5392/JKCA.2014.14.10.527

◐ 저 자 소 개 ◑

주 연 진
2016년 명지대학교 컴퓨터공학전공 졸업(학사)
2018년 국민대학교 비즈니스IT전문대학원 졸업(석사)
관심분야 : 텍스트 마이닝, 빅데이터 분석 및 활용 etc
E-mail : juy3625@kookmin.ac.kr

김 유 신
2000년 국민대학교 정보관리학과 졸업(학사)
2009년 국민대학교 경영정보학과 졸업(석사)
2013년 국민대학교 경영정보학과 졸업(박사) & 미국 텍사스 주립대 Post-doctoral Research Fellow
현재 알티미디어 데이터사이언스부문장(Chief Data Scientist)
관심분야 : 모빌리티 & 금융 빅데이터 분석 및 AI 서비스 개발 etc.
E-mail : yoosin@alticast.com

정 승 렬
1985년 서강대학교 경제학과 졸업(학사)
1989년 미국 위스컨신 대학교 대학원 경영정보학과 졸업(석사)
1995년 미국 사우스케롤라이나 대학교 대학원 경영정보학과 졸업(박사)
1997년~현재 국민대학교 비즈니스IT전문대학원 교수
관심분야 : 텍스트 마이닝, 오피니언 마이닝, 빅데이터 분석, 정보시스템 구현, 프로세스 관리 etc.
E-mail : srjeong@kookmin.ac.kr