텍스트 마이닝을 적용한 한국교통방송제보 비정형데이터의 분석

Analysis of the Unstructured Traffic Report from Traffic Broadcasting Network by Adapting the Text Mining Methodology

노 유 진* · 배 상 훈**

* 주저자 : 도로교통공단, 부경대학교 공간정보시스템공학과 박사과정 수료
** 교신저자 : 부경대학교 공간정보시스템공학과 교수

You Jin Roh* · Sang Hoon Bae**

* Koroad, Ph. D. Candidate, Pukyong National Univ.
** Professor, Pukyong National Univ.
† Corresponding author : Sang hoon Bae, sbae@pknu.ac.kr

Vol.17 No.3(2018)
June, 2018
pp.87~97

ISSN 1738-0774(Print)
ISSN 2384-1729(On-line)
https://doi.org/10.12815/kits.2018.17.3.87

Received 16 April 2018
Revised 16 May 2018
Accepted 31 May 2018

요 약

교통사고 관련 제보는 비정형 데이터로서 교통사고를 유발한 가해자나 피해자의 관점이 아닌, 교통사고 발생 지점과 구간, 시간대에 있었던 타 운전자의 관점에서 생성된 교통정보의 가치를 가지고 있다. 그러나, 비정형 데이터인 교통제보가 빅 데이터로서 교통사고 통계나 교통관련 연구에 활용되지 못하였으나, 텍스트 마이닝 기법을 활용한 본 연구를 통해 비정형의 빅 데이터를 시각화하고 해석하여, 기존의 정형 데이터에서 분석하지 못한 정보를 도출할 수 있었다. 그리고 교통사고 발생으로 인한 도로상 영향을 파악할 수 있었다. 이러한 분석으로 교통제보의 트랜드를 파악하고, 운전자가 제보하는 “도로명”, “지점명”, “시간대”를 추출하였으며, 교통사고 발생으로 다른 운전자에게 가장 많은 영향을 미치는 지점과 구간의 파악이 가능하였다. 향후 실제 교통사고 데이터와 결합하여 교통제보와의 상관성 분석 등을 통해 비정형 데이터의 활용방안을 모색할 계획이다.

핵심어 : 교통제보, 빅 데이터, 비정형 데이터, 텍스트 마이닝, 워드 클라우드

ABSTRACT

The traffic accident reports that are generated by the Traffic Broadcasting Networks(TBN) are unstructured data. It, however, has the value as some sort of real-time traffic information generated by the viewpoint of the drives and/or pedestrians that were on the roads, the time and spots, not the offender or the victim who caused the traffic accidents. However, the traffic accident reports, which are big data, were not applied to traffic accident analysis and traffic related research commonly.

This study adopting text-mining technique was able to provide a clue for utilizing it for the impacts of traffic accidents. Seven years of traffic reports were grasped by this analysis. By analyzing the reports, it was possible to identify the road names, accident spot names, time, and to identify factors that have the greatest influence on other drivers due to traffic accidents. Authors plan to combine unstructured accident data with traffic reports for further study.

Key words : Traffic Reports, Big data, Unstructured data, Text Mining, Word-cloud

Ⅰ. 서 론

1. 연구의 배경 및 목적

불특정 다수를 대상으로 하는 TBN 한국교통방송[1]은 1997년 부산, 광주에서 개국한 이후 운전자 대상 평일 채널 청취율은 1일 20시간 방송시간 가운데 전 시간대 청취율 1위를 차지하고 있으며, 운전시 선호도가 60.9%를 차지하며 교통정보의 신속성과 정확성에 대하여 응답자의 94.6%가 신속, 정확하다고 응답하고 있다.[2] 이러한 교통방송국의 교통 정보에 대한 운전자에 대한 선호도가 높아짐에 따라, 그 만큼 교통제보에 대하여 가치 있는 정보를 창출하기 위한 기술의 요구 또한 증가하고 있다. 그러나 대부분 교통제보가 비정형 형태로 구성되어 있는 텍스트 기반의 자료로서, 기존의 통계분석이나 데이터 마이닝(Date Mining) 기법을 적용하기에는 부적합하다. 따라서 본 연구에서는 비정형 자료 분석 기법 중 하나인 텍스트 마이닝(Text Mining) 기법으로 TBN 한국교통방송의 교통제보 자료를 분석하여 부산시내 교통사고와 연관된 단어들을 추출하여 사용빈도가 높은 단어들을 찾아내고 그 단어들을 이용하여 각 단어들 간의 특성을 파악하고자 하였다. 본 연구로 텍스트 형식으로 구성되어 있는 교통제보를 정보로서의 활용 방안을 찾아보고 교통사고 예방과 교통정체 해소에 도움을 줄 수 있는 방법을 모색하였다.

2. 연구의 범위 및 방법

본 연구에서는 TBN 한국교통방송 부산본부에서 2007년부터 2012년까지 6년간 수집된 199,996건의 교통사고 관련 제보를 확보하여, 텍스트 마이닝 방법론으로 교통사고와 연관되는 주요 키워드를 도출하고 분석하였다. 교통방송 교통사고 관련 제보는 1997년부터 현재까지 20년 이상의 교통제보 데이터로서, TBN 한국교통방송 부산본부 편성제작국에 의뢰하여 수집되었으며 이 중에서 시간적 경제적 비용 부담으로 인해 TBN 한국교통방송 부산본부 개국 10년 후인 2007년부터 2012년까지의 교통사고 관련 교통제보를 추출하여 분석하였다. 초창기 교통제보는 교차로 및 가로명 등 지명의 통일 되지 않고 교통 용어도 정리되지 않아, 비교적 교통제보의 단어가 안정적인 방송국 개국 10년 후인 2007년을 시점으로 교통제보 데이터의 연구의 효율성과 경제성을 위하여, 균일성, 통일성이 확보되는 2012년까지 6년간의 데이터를 선택하였다. 이러한 교통 제보를 바탕으로 교통사고가 발생하는 도로명과 교차로명의 제보 키워드의 빈도 분석과 연관성 분석을 수행하여 교통사고가 발생하는 구간과 지점의 특성을 분석하고 운전자가 제보하는 교통사고 현황과 트렌드를 파악하였다. 본 연구를 통해 교통사고 분석 및 교통정보의 질적 제고에 도움을 주고 부산시 교통정책에 반영할 수 있는 유용한 정보를 도출하고자 하였다.

3. 기존 연구

국내에서 본 연구와 유사하게 텍스트 마이닝을 통해 비정형 교통정보를 분석한 사례는 찾아보기 어렵다. 다만, 부산지역 교통관련 기사를 이용한 비정형 빅 데이터의 정형화와 시각적 해석에서 지역신문 기사들 중에서 '교통'과 '부산'을 동시에 포함한 데이터의 패턴을 찾아 시각화한 것이 있다(Lee et al., 2014). 하지만,

1) TBN 한국교통방송 : 도로교통공단이 운영하는 교통정보전문 FM 방송채널이다. 1997년 부산과 광주에 개국하여 2017년 현재 전국 11개 네트워크를 가지고 교통정보와 지역정보를 제공하며 재난예방 방송을 하고 있다.
2) 도로교통공단 홈페이지(www.koroad.or.kr) 교통방송사업 사업소개

교통 분야에서 텍스트 마이닝 기법을 활용한 사례는, 화물자동차 사고데이터를 사용하여 경찰이 작성한 사고 개요를 텍스트 마이닝 기법을 통해 분석하여 정형 데이터에서 찾아내지 못한 교통사고 특징을 파악하였고(Kim et al., 2015), 교통사고 조사 서비스를 받은 민원인의 서술형 자료를 이용하여 사고조사서비스 만족도에 영향을 미치는 요인과 개선점을 분석한 (Jung et al., 2016) 연구 결과물도 있다. 교통제보와 유사한 비정형 데이터를 활용한 연구로는 기상학 분야에서 기상 콜센터의 전화 상담내용을 기록한 텍스트 자료를 이용하여 상담내용의 월별 특징과 시간대별 특징 등을 분석한 것이 있다(Lee et al., 2016).

텍스트 마이닝은 다양한 분야에서 적용되고 있으며, 의학 분야에서는 질병의 원인과 치료를 위한 분석을 위해 사용하여 왔고, 가장 텍스트 마이닝이 발전한 분야 중 하나이다(Chan et al., 2016). 금융 분야에서는 뉴스에 대한 텍스트 마이닝을 통해 뉴스가 주가 상승에 미치는 영향을 분석하는 모형을 제시하기도 했다(Ahn et al., 2010). 또한, 과거 부도가 발생한 기업의 뉴스 콘텐츠를 데이터로 확보하여 텍스트 마이닝 기법을 통한 기업 부도 예측의 가능성을 시도하였다(Choi et al., 2015). 그리고 기상청 기상연감 자료 분석을 통하여 이슈가 되었던 기상관련 소식과 기상현황, 그리고 기상청이 중점으로 하고 있는 업무현황의 트랜드를 파악하였다(Sun et al., 2017).

Ⅱ. 텍스트 마이닝

1. 텍스트 마이닝 방법론

텍스트 마이닝 방법론은 최근 발전된 정보처리 기술과 인프라를 활용하여 뉴스, 인터넷 등의 텍스트 문서로부터 정보를 획득, 키워드의 패턴을 분석하고 이를 토대로 예측을 수행하는 방법론으로서 최근 그 활용 영역을 확장해 나가고 있다. 텍스트 마이닝은 데이터 마이닝과 유사한 개념이지만, 기존의 데이터 마이닝이 관계형 데이터베이스나 XML과 같은 구조화된 데이터들만을 처리할 수 있는 반면, 텍스트 문서, e-메일, HTML 파일과 같은 비정형 또는 반 정형화된 데이터를 일정한 형식과 조건을 만족하는 자료로 가공하여 분석하는 방법론을 텍스트 마이닝으로 별도 구분하고 있다.

텍스트 마이닝 방법론에 대하여는 복수의 선행 연구자들이 정의를 내리고 있다. Choi et al.(2002)은 텍스트 마이닝을 구조화되지 않은 대규모의 텍스트 집단으로 부터 새로운 지식을 발견하는 과정을 의미하는 것이라고 정의하였다.

Bae et al.(2003)은 텍스트 마이닝을 문서 수집, 문서 전처리, 텍스트 분석, 그리고 결과 해석 및 정제 단계 등 4단계로 나누었는데, 전처리과정은 다시 필요 없는 단어 또는 기호를 정제하는 정제 과정과 문장의 정확한 의미 파악을 위해서 각 단어의 어간을 파악하고 동의어를 할당하는 정규화 과정으로 나누었다. 정규화 과정은 또 다시 한글 처리를 위해서 문장에서 최소의 의미단위를 추출해 내는 형태소 분석 단계와 통사구조를 파악하는 구문 구조 분석 단계, 의미 구조를 추출하는 의미 분석 단계, 그리고 문장들 사이의 관계를 분석하는 문맥 분석 단계로 나누었다. 텍스트 분석과정은 텍스트 군집화, 텍스트 분류, 그리고 텍스트 요약으로 나누어 설명하였다. 텍스트 군집화는 텍스트의 집단을 내용의 유사도에 따라 여러 개의 소집단으로 분할하는 과정으로서 데이터에 대한 지식 없이 분석 초기에 행하여 결과를 분석할 수 있다는 장점이 있으며, 중복 혹은 유사한 문서를 제거하고, 다른 문서의 주제와 다른 주제를 가진 문서를 구별하고, 대량의 문서집합의 개요를 획득하는데 적용할 수 있다고 한다. 텍스트 분류란 텍스트의 내용에 따라 미리 정해해 놓은 범주를 부여하는 과정인데, 군집과 같이 분류를 수행하기 위해서는 각 항목을 위한 학습데이터를 사용자가 선정하여 훈련시키는 과정이 필요하다고 정의하였다. 텍스트 요약은 문서의 전체 내용을 반영할 수 있는 일부 내용을

추출하는 과정으로 표면수준접근, 개체수준접근, 그리고 화법수준접근의 3가지 기법이 사용되는데 일반적으로 3가지 중 2가지 이상의 기법을 조합해서 이용한다.

또한, Kim et al.(2009)은 텍스트 마이닝을 다양한 정보원천으로부터 자동적으로 정보를 추출함으로써 이전에 알려지지 않았던 새로운 정보를 발견하는 정보기술이라고 정의하였다. 텍스트 마이닝 과정에 대하여는 전처리 과정과 텍스트 분석과정으로 나누어 설명하였다. 먼저 전처리 과정은 일반적인 텍스트 테이터들을 컴퓨터가 처리하기 쉽도록 변화하는 작업으로써, 특정 단어와 관련된 문서들을 신속하게 검색할 수 있도록 인덱스 파일을 만드는 것이라고 설명하고 있다. 그리고 인덱스를 만드는 방법으로 FB(Frequency-Based), IDF(Inverse Documnent Frequency), LSI(Latent Semantic Indexing) 등의 대표적인 방법을 열거하였다. FB는 문서 안에서 빈번히 나타나는 단어들을 그 문서를 대표하는 중요한 단어로 파악하고 가중치를 높게 주는 개념이며, IDF는 특정문서에서 중요한 단어가 무엇인지 뿐만 아니라 다른 문서와 구분을 해주는 단어가 무엇인지에 대한 정보를 포함하기 위한 계산을 한다는 것이다. LSI는 문서들이 공유하는 단어들을 파악하여 동일한 주제나 개념으로 인식함으로써 검색단어와 정확하게 일치하지 않더라도 개념이나 주제에 의하여 문서검색이 가능할 수 있게 한다는 것이다. 다음으로 텍스트 분석 과정은 전처리 과정을 거친 데이터들을 대상으로 정보 추출, 범주화, 문서요약과 같은 다양한 분석을 실시하는 것으로 설명하고 있다. 정보 추출은 특정 문서 안에서 유용한 정보 즉, 사람이름, 장소이름, 전화번호, 날짜, 화폐단위 등 문서내의 개체들 및 이들 사이의 연관성을 식별하여 검색하는 기술이라고 설명한다. 범주화는 수집된 문서들 중에서 유사한 내용의 문서들을 그룹화해서 분류하는 기술로써, 비구조적으로 모여 있는 문서들을 구조적으로 조직화하는 과정이라고 정의하였다.

2. 자료 전처리

본 연구에서 사용된 자료는 TBN 한국교통방송국 부산본부에서 2007년부터 2012년까지 6년간 수집된 199,996건의 교통사고 관련 제보이다. 부산본부가 1997년 개국한 이후 교통통신원 및 일반 운전자들에게 부산시내 교통상황을 제보 받아 교통정보로 활용하여 오고 있으며, 그 중 교통사고와 연관된 교통제보를 분석하였다. 교통제보는 사고 시간, 장소, 방향, 그리고 내용을 교통통신원이 제보 전화나 SNS로 실시간 전달하면 방송모니터요원이 접수받아 방송하는 구조이다.

교통사고 관련 제보 내용에는 구체적인 사고 지점 (중앙대로 서면교차로 등), 사고 관련 차량 종류(승용차 등), 사고유형(추돌사고 등), 그리고 사고영향(정체 등) 등이 기술되어 있다. 그러나 교통사고 제보 데이터를 텍스트로 교정하는 단계에서 도로명이나 방향의 미 표기, 오탈자 문제(승용차를 스용차, 승영차 등), 띄어쓰기 문제(개금 부산은행을 개금부산은행, 개금동 부산은행 등), 약자 표기(부경대학교를 부경대 등), 동일 장소를 다양하게 표현(충렬대로를 충렬로, 충렬길 등), 외국어 표기(센터를 센타로), 일반명사와 고유명사의 혼동(부산시청을 시청으로 표기), 과거 지명 사용(구만덕대로, 구 밀리오레앞 등), 불분명한 지명(삼성서비스센터 등) 등으로 신문기사나 사전 등과 같이 문법이 정확한 표기가 아닌 사례가 많은 것으로 분석되었다. 이러한 경우 형태소 인식기가 각 어절에 대해 품사 태깅(Tagging)[3]을 제대로 하지 못하여 자료 인식에 문제가 발생하는 경우가 많아 일반 텍스트 에디터의 정규표현식[4] 기술을 활용하여 교정하여야 한다. 또한, 일반 명사와 고유 명사의 혼동 문제는 형태소 분석기의 세종말뭉치를 교정하거나 예를 들어 시청을 부산시청으로 직접 교정하는 방법을 적용하였다. 본 연구에서는 상기와 같은 전처리 과정을 거쳐 텍스트 마이닝 방법론으로 교

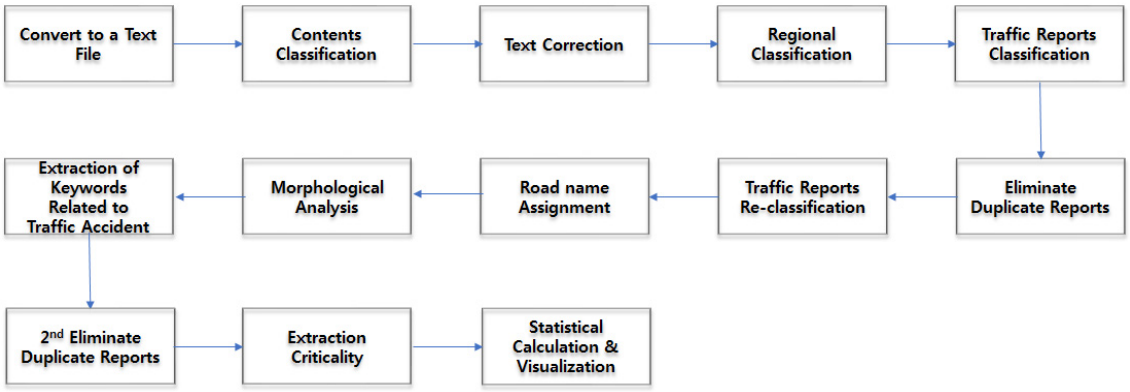
3) Tagging : 웹상에서 어떤 이미지나 파일 따위에 해당 내용을 대표하는 키워드를 다는 것
4) 정규표현식(regular expression) : 특정한 규칙을 가진 문자열의 집합을 표현하는데 사용하는 형식 언어

통사고와 연관되는 주요 키워드를 도출하고 분석하였다. 이를 위하여 교통사고가 발생하는 도로명과 교차로명의 제보 키워드의 빈도 분석과 연관성 분석을 수행하여 교통사고가 발생하는 구간과 지점의 특성을 분석하고 운전자가 제보하는 교통사고 현황과 트렌드를 파악하였다. 본 연구를 통해 교통사고 분석 및 교통정보의 효율화에 도움을 주고 부산시 교통정책에 반영할 수 있는 유용한 정보를 도출하고자 하였다.

3. 분석 수행 절차

본 연구에서는 TBN 한국교통방송국 부산본부 교통제보를 교통사고 발생 시간, 사고, 방향, 내용으로 나누어 각 항목에 대하여 다음 <Fig. 1> 분석 절차에 따라 텍스트 마이닝 기법을 각각 수행하였다. 텍스트 마이닝 기법을 이용하여 보다 더 정확한 분석을 위하여, 선택된 단어가 텍스트 내에서 정확하게 사용되는지 확인하는 과정이 필요하다. 이 과정에 따라 분석의 토대가 되는 단어 추출 결과가 달라지기 때문에 전처리 과정에서 사전 구축과 띄어쓰기 수정 및 용어 통일 과정이 필요하다. 사전을 구축하기 위해 오픈소스로 제공되는 세종 말뭉치를 기본으로 부산 지역의 지명이나 교통용어 등 세종 사전에 등록되지 않은 단어들을 추출해 텍스트 내 단어들을 사전에 추가하였다.

전처리 과정이후 통계 프로그램인 KNIME[5]에서 형태소 분석기를 활용하여 사전에 등록된 단어를 태깅된 품사에 따라 단어를 추출하였다. 이후 추출한 단어를 mecab[6] 패키지를 이용하여 말뭉치로 만들고 분석에 쓰이지 않을 단어를 제거하였으며, 변환된 결과를 원래의 자료와 비교하여 분석의 목적에 맞게 단어가 추출되었는지 확인하였다. 또한, 정제된 말뭉치를 기반으로 용어-문서 행렬을 만들어 빈도표를 작성하고 이를 이용하여 단어의 출현빈도를 파악하였다. 이 과정에서 분석 결과를 효과적으로 나타내기 위하여 단어 빈도표를 바탕으로 wordcloud[7]를 이용하여, 단어들을 구름모양으로 나타내어 빈도가 높고 핵심어일수록 큰 글씨로 중심부에 표현하여 분석 결과를 시각화하였다. 이후 구체적인 수치를 통해 단어의 흐름을 파악하고, 특정 단어가 출현한 횟수를 비교하기 위하여 단어가 출현한 절대 도수가 아닌 상대도수를 계산하여 분석하였다. 특정 단어의 출현 횟수와 총 단어의 출현횟수 합계를 이용하여 단어들의 상대도수를 구하고 상대도수 그래프를 그려서 분석결과를 나타내었다.



〈Fig. 1〉 Flow of Text Mining for Report data for TBN

5) KNIME : 머신러닝, 예측 알고리즘 모델링으로 빅데이터 SW 플랫폼으로 세종 말뭉치를 오픈 소스로 사용
6) MeCab : 오픈소스 형태소 분석엔진, 띄어쓰기에 크게 의존하지 않고, 사전을 참조하여 어휘를 구분
7) wordcloud : 문서의 키워드 등을 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법

Ⅲ. 분석 결과

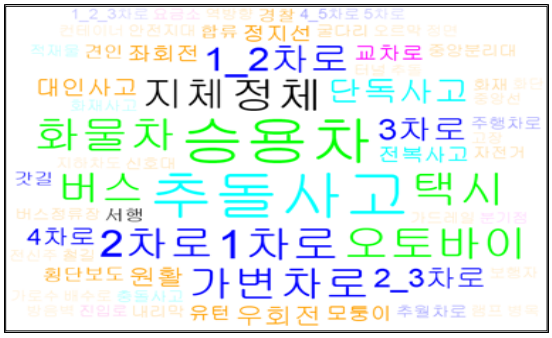
이러한 분석으로 TBN 한국교통방송국 부산본부에 제보되는 교통정보의 트랜드를 파악하고, 이를 통해 교통운전자가 제보하는 도로명, 교차로명, 시간대를 추출하였다. 그리고 그 내용들 간의 연관성을 파악하여 교통사고 발생으로 운전자에게 가장 많은 영향을 미치는 시간대와 지점과 구간 등을 도출하였다.

1. 교통관련 키워드 분석

TBN 한국교통방송 부산본부에 제보된 내용을 분석한 결과 “추돌사고”와 “승용차”의 빈도가 높게 나타났다. 빈도 합계를 살펴본 결과 “추돌사고”의 출현빈도가 74,034건으로 1위이며, 승용차가 72,831건으로 2위였다. 이외에도 “화물차”, “택시”, “정체”라는 단어가 상위를 차지하고 있다. 위의 <Fig. 2>와 <Fig. 3>과 같이 교통관련 키워드를 워드클라우드를 통해 시각화 하였다.

교통제보를 하는 운전자들이 가장 많이 도로상에서 영향을 미치는 것이 “승용차”와 “추돌사고”라는 단어임을 알 수 있다. 부산이라는 대도시에서 운행하는 대부분의 차량이 “승용차”이고, 도심 내에서 발생하는 사고가 “추돌사고”라는 것을 보여주고 있다.

〈Fig. 2〉 wordcloud 1 by key word

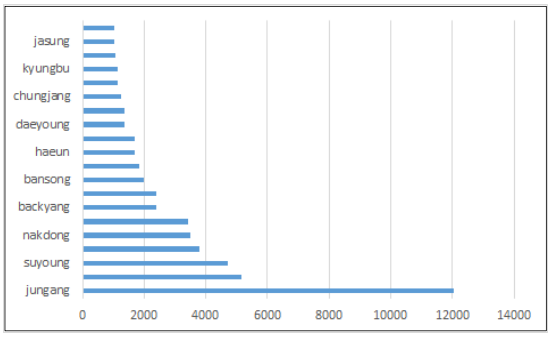


〈Fig. 3〉 wordcloud 2 by key word

2. 교통 제보 도로별 분석

교통제보 내용 중 도로별로 분석하여 보면, 상위 20개 도로가 68%를 차지하고 있으며 “중앙대로”가 12,031건 15.3%로 압도적으로 많은 것으로 나타났다. 이러한 현상은 중앙대로가 중구에서 금정구까지 남북을 연결하는 중심도로로서, 총 연장이 20km를 상회하며, 부산시내의 주 간선도로로 대부분의 통행량을 처리하는 것에 기인한 것으로 보인다. 다음으로는 “충렬대로”로 해운대구에서 동래구까지 동서로 연결하는 총 연장 5.1km 도로로서 출퇴근 시간대 정체가 많이 발생하는 구간이다. “수영로”와 “가야대로”등 출퇴근시간대 동서로 연결하는 도로망에서 많은 제보가 나타나는 것으로 분석되었다. 다음 <Fig. 4>는 도로별로 워드클라우드를 통해 시각화 하였으며, <Fig. 5>는 상대도수를 그래프로 표현하였다.

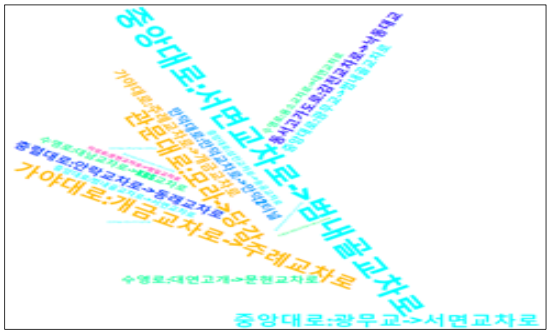
〈Fig. 4〉 wordcloud by road name



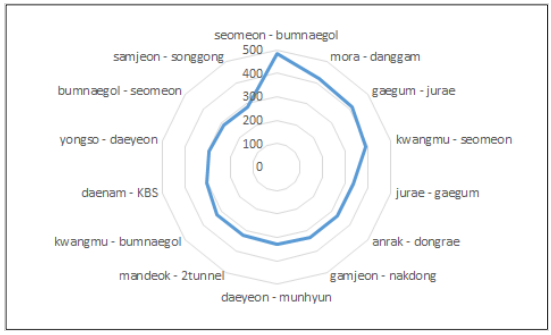
〈Fig. 5〉 statistical analysis by road name

3. 교통 제보 도로별, 방향별 분석

교통제보 내용을 방향별로 분석한 결과, 상위 20개 방향이 8%를 차지하여 방향별로 편중된 결과는 보이지 않았다. 빈도가 가장 높은 구간이 중앙대로의 서면교차로에서 범냇골 교차로 구간으로 486건의 교통제보가 있었다. 다만, 구간별로 중앙대로상 “광무교에서 범냇골교차로”, “광무교에서 서면교차로” 구간 등의 상위를 차지하고 있어 운전자들에게 가장 많은 교통사고 관련 제보 구간은 중앙대로의 “서면교차로에서 범냇골교차로”양 방향임을 알 수 있다.



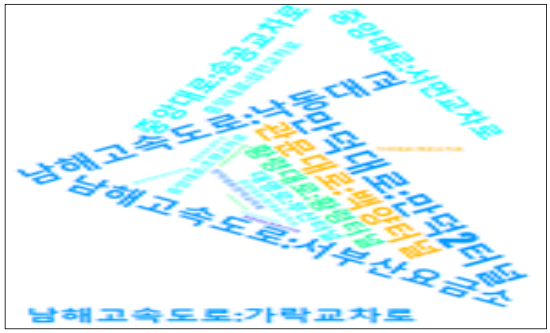
〈Fig. 6〉 wordcloud by direction name



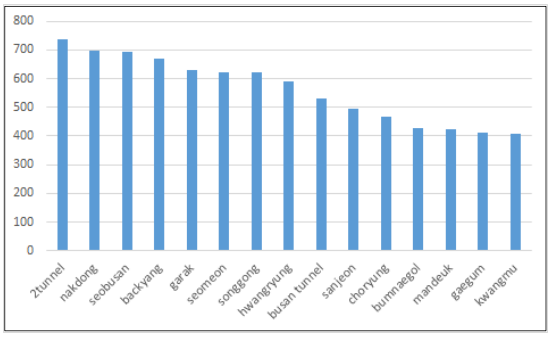
〈Fig. 7〉 statistical analysis by direction name

4. 교통 제보 도로별, 지점별 분석

교통제보 내용을 도로별, 지점별로 분석한 결과, 상위 20개 방향이 4%를 차지하여 편중된 결과는 보이지 않았다. 가장 빈도가 높은 지점은 만덕대로의 “만덕2터널”로 719건의 제보가 있었다. 다음 <Fig. 8>로 도로별, 지점별로 워드클라우드를 통해 시각화 하였다. 또한, <Fig. 9>에서 통계 분석한 결과를 그래프로 나타내었다.



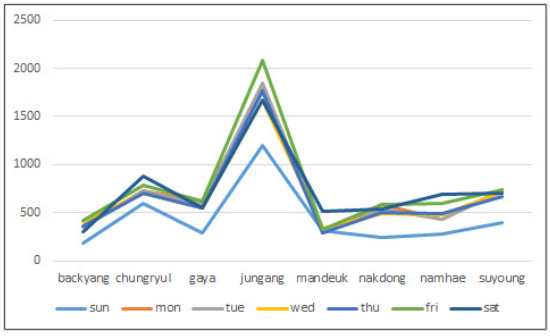
〈Fig. 8〉 wordcloud by road, spot name



〈Fig. 9〉 statistical analysis by road, spot name

5. 교통 제보 도로별, 요일별 분석

〈Fig.10〉 wordcloud by road, day



〈Fig.11〉 statistical analysis by road, day

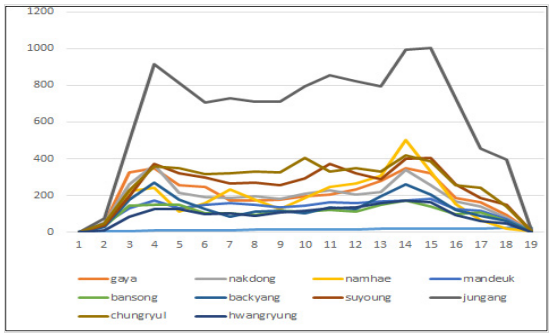
교통제보 내용 중 도로별, 요일별로 분석하면 부산시내 전 도로에서 금요일에 가장 많은 교통제보가 있으며, 일요일의 경우 가장 적은 편이다. 다만, “경부고속도로”의 경우에는 일요일에 교통제보가 가장 많고, 월요일과 화요일에 비해 높으며 금요일, 토요일, 일요일에 급격히 증가하는 것으로 나타났다. 이러한 분석결과는 일요일에 시 외곽에서 부산시내로 진입하는 차량들의 정체가 많은 구간에서 특히 많이 나타난다.

위의 <Fig. 10>은 도로별, 요일별로 워드클라우드를 통해 시각화 하였다. “중앙대로”는 모든 요일에서 교통제보가 높게 나타나고 있지만, 특히 금요일에 가장 많은 제보가 있었고 월요일이 가장 적은 것으로 나타났다. 또한, <Fig. 11>에서 통계 분석한 결과를 그래프로 나타내었는데 “충렬대로”의 경우에는 토요일과 화요일에 교통제보가 많고 월요일에는 가장 적게 나타난다. “수영로”의 경우에는 금요일, 토요일에 많으며, “만덕대로”의 경우에는 토요일과 수요일에 많은 것으로 나타났으며, 일요일의 교통제보가 다른 도로에 비해 상대적으로 높게 나타나고 있다.

6. 교통 제보 도로별, 시간대별 분석

교통제보 내용 중 도로별, 시간대별 분석 결과 “중앙대로”에서 출퇴근시간대에 가장 교통제보가 많은 것으로 나타났다. 특히 “중앙대로”의 퇴근시간대인 19시대에 1,005건, 18시대에 999건의 교통제보가 있었다.

〈Fig.12〉 wordcloud by road, time



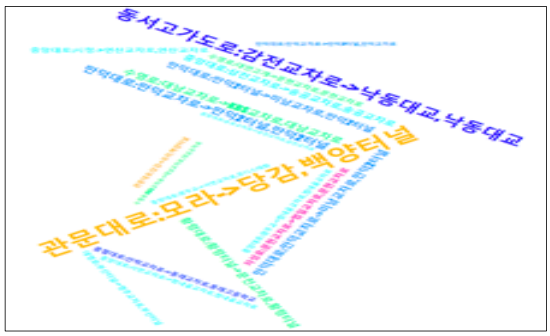
〈Fig.13〉 statistical analysis by road, time

"중앙대로"를 제외하면 "충렬대로"가 출퇴근시간대 교통제보가 집중되는 것으로 나타났다. 특히 9시 이후에는 타 도로에서는 교통제보가 줄어들지만, "충렬대로"에서는 전 시간대에서 교통제보가 이어져 교통정체 현상이 상시 발생하는 것으로 분석되며, 특히 "추돌사고"와 같은 잦은 교통사고가 정체로 이어지는 것으로 나타났다. 출근시간대 "수영로"의 경우에는 8시대에 집중적으로 교통제보가 있으며, 15시를 전후하여도 많은 교통제보가 발생하고 있다.

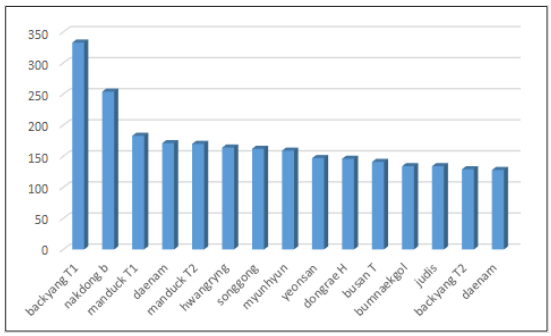
7. 교통 제보 도로별, 방향별, 지점별 분석

교통제보 내용을 도로별, 방향별, 지점별 교차 분석 결과, 상위 20개가 약 4% 비율을 차지하고 있다. 빈도가 가장 높은 지점이 관문대로 모라동에서 당감동 방향 "백양터널"이고, 다음이 동서고가도로의 감전교차로에서 낙동대교 방향의 "낙동대교"이다. 대부분 상위 지점이 교차로와 터널이지만, 터널에서는 "백양터널 양방향", "만덕2터널 양방향", 황령터널의 경우 "문전교차로 방향," 부산터널은 "영주교차로 방향"에서 많은 교통제보가 있었다.

교차로에서는 "송공교차로", "문현교차로", "연산교차로", "범냇골교차로"등 중앙대로 상의 대형 교차로가 교통제보 중 상위에 올랐으며, 중앙대로 상 "서면교차로"주변의 교차로에서 많은 교통제보가 있었다. 특히, 충렬대로에서는 "동래고등학교"가 많은 제보를 받았고, 중앙대로상의 "(구) 쥬디스태화"도 운전자들의 관심도가 높았다고 분석이 된다.



〈Fig.14〉 wordcloud by road, spot, direction



〈Fig.15〉 statistical analysis by spot, direction

Ⅳ. 결 론

TBN 한국교통방송 부산본부의 교통사고 관련 제보는 비정형 데이터로서 교통사고를 유발한 가해자나 피해자의 관점이 아닌, 교통사고 발생 지점과 구간, 시간대에 있었던 타 운전자의 관점에서 생성된 교통정보의 가치를 가지고 있다. 그러나 이러한 교통방송국의 교통제보 비정형 데이터가 교통사고 통계나 교통관련 연구에 활용되지 못하였으나, 텍스트 마이닝 기법을 활용한 본 연구를 통해 교통사고 발생으로 인한 도로상 영향을 파악할 수 있었다. 이러한 분석으로 TBN 한국교통방송 부산본부에 제보되는 교통제보의 트랜드를 파악하고, 이를 통해 운전자가 제보하는 “도로명”, “지점명”, “시간대”를 추출하였으며, 교통사고 발생으로 운전자에게 가장 많은 영향을 미치는 지점과 구간의 파악이 가능하였다.

본 연구를 통해 파악된 내용은 다음과 같다. 제보 빈도수가 가장 높은 도로는 주간선도로인 “중앙대로”이며, 중앙대로에서 교통사고가 발생할 경우 그 영향력이 매우 크다는 것을 알 수 있었다. 또한, 퇴근시간대 발생한 교통사고의 여파에 대한 관심사가 높다는 것을 분석을 통해 알 수 있었다. 특히, 부산의 경우 대부분 터널에서 사고가 발생할 경우 교통제보가 많으며, “백양터널”, “만덕2터널”, “황령터널” 순으로 교통사고 발생에 따른 여파가 큰 것으로 분석되었다. 교량에서는 동서고가도로와 남해지선고속도로를 연결하는 “낙동대교”의 관심 빈도수가 높은 것으로 나타나, 부산 시민들의 통행패턴에 대한 이해도 제고에도 도움이 될 수 있음을 알 수 있었다. 교통사고는 부산시내 어디에서나 발생하지만, 특히 교통사고로 인해 다른 운전자들이 반응하는 지점과 구간은 한정되어 있음을 알 수 있었다. 향후 실제 교통사고 발생 데이터, 당시의 기상 데이터 등 정형 데이터와 결합하여 교통사고 발생이 교통제보에 어떠한 영향을 끼치는지, 교통사고 결과가 교통제보로 전달되는지에 대한 추가적인 연구가 수행되어야 할 것이다.

이처럼 교통제보를 빅 데이터 분석을 통하여 새로운 정보를 습득할 수 있었으나, 아직까지 TBN 한국교통방송의 교통제보 방식이 무료 제보전화에 의존하고 있고, 이러한 무료 제보전화를 운영하는 방송모니터 요원들이 전문화되어 있지 않아 교통제보의 수집, 전달과정에서 정보의 왜곡이 심하게 발생하고 있었다. 따라서 이러한 문제점에 대한 보완이 필수적이라 사료된다. 또한, TBN 한국교통방송은 교통제보의 적극적인 활용방안모색을 통해 교통사고 예방 활동도 중요하지만, 교통사고의 여파가 도로이용자, 주변 상황에 어떠한 영향을 초래하는지에 분석을 통해 교통방송의 정보 수집, 전달 시스템의 개선도 고려되어야 한다.

TBN 한국교통방송 교통제보는 빅 데이터로서 무한한 잠재력을 가지고 있는 정보이다. 빅 데이터 분석은 다양한 형태로 축적되어 있는 대용량의 데이터로부터 잠재 되어 있는 가치를 찾아가는 과정이다. 교통정보는 현장에서 실시간으로 전달되고 그 여파가 시민의 실생활에 직접적인 연관을 가지고 있어 관계기관에서 여러 가지 형태로 저장되고 있다. 그러나 그 활용이 아직 다른 분야에서와 같이 활발하게 연구되지 않아 빅 데이터로서 교통사고 예방이나, 교통정체 해소 등 무한한 잠재력을 발휘하지 못하고 있는 실정이다. 본 연구에서는 비정형화된 빅 데이터를 시각화하고 해석하여 기존의 정형화된 통계분석에서 찾아내지 못했던 정보를 도출할 수 있었다. 앞으로 교통 분야에서는 이러한 빅 데이트를 통해 새로운 정보를 습득하기 위한 인공지능 머신러닝과 같은 기계학습 기법을 활용하는 것이 강조될 것으로 예상된다. 나아가 정형화된 도로교통공단의 TASS 자료와의 결합을 통해 추가적인 분석을 수행한다면, 고부가 가치의 정보 활용이 가능할 것으로 사료된다.

REFERENCES

Ahn S. and Cho S.(2010), “Stock prediction using news text mining and time series analysis,” In *2010*

Conference Proceedings of Korean Institute of Information Scientists and Engineers, 37, pp.364-369.

Bae S. and Park C.(2003), "A Study on the Application of Text Mining to the Analysis of Technical Information," *Korea Technology Innovation Society*, pp.79-83.

Chen P., Ponocko J., Milosevic N., Nenadic G. and Milosevic J.(2016), "Towards application of text mining for enhanced power network data analytics-part i; retrieval and ranking of textual data from the internet," *Mediterranean Conference on Power Generation, Transmission Distribution and Energy Conversion* (medpower 2016), pp.1-8.

Choi J., Han H., Lee M. and Ahn J.(2015), "The prediction of Corporate Bankruptcy Using text-mining Methodology," *Productivity Review*, vol. 29, no. 1, pp.203-206.

Choi Y. and Park S.(2002), "Interplay of Text Mining and Data Mining for Classifying Web Contents," *Korean Journal of cognitive science*, vol. 13, no. 3, pp.33-35.

Jung C. W.(2016), "A Study on Traffic Accident Investigation Satisfaction Factors," *Journal of Transport Research*, vol. 23, no. 4, pp.73-84.

Kim K. and Oh S.(2009), "Methodology for Applying Text Mining Techniques to Analyzing Online Customer Reviews for Market Segmentation," *The Journal of the Korea Contents Association*, vol. 9, no. 8, pp.272-284.

Kim Y., Heo J. and Kang K.(2015), "Overview of cargo accident using text mining," *2015 Conference of Korea Transportation Research Society*, pp.338-343.

Lee K., Roh Y., Yoon S. and Cho Y.(2014), "Structuring of unstructured big data and visual interpretation," *Journal of the Korean & Information Science Society*, vol. 25, no. 6, pp.1436-1437.

Lee Y., Lim C., Heo M. and Kim H.(2016), "Text-mining technique for Weather call center data analysis," In *2016 Spring Conference Proceedings of Korean Meteorological Society*, pp.153-154.

Sun H., Lim C. and Lee Y.(2017), "Analysis of the Yearbook from the korea Meteorological Administration Using a text-mining agorithm," *The Korean Journal of Applied Statistics*, vol. 30, no. 4, pp.603-613.