빅데이터 분석도구 R을 활용한 기상뉴스 데이터분석

김용수, 반재훈

고신대학교 IT경영학과

Analysis of Weather News using Big Data Analytics Tools R

YongSu Kim · ChaeHoon Ban

Dept. of IT Management, Kosin University

E-mail : ehyerisin@gmail.com · chban@kosin.ac.kr

요 약

정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅 데이터의 중요성이 강조되고 있으며 다양한 분야에서 이를 응용하고 있다. 빅 데이터 분석도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 R을 이용하여 기상뉴스에 나타난 기상관련 빅 데이터를 분석한다. 다양한 뉴스에서 기상 관련 데이터를 수집하고 어떠한 텍스트가 분포되어 있는지 빈도 조사를 수행한다.

키워드

Big Data, R, Text Mining, Transportation, Analysis

Ⅰ. 서론

정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 이를 응용하고 있다. 빅 데이터 분석 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 기상데이터의 활용으로 국가 주관의 행사와 자연재해에 대한 예측과 대응방안, 기업의 신제품 연구개발과 마케팅 등의 자원으로 활용됨으로 기상데이터가 중요하다고 볼 수 있다.

'기상'과 관련된 키워드를 중심으로 올해 현재까지를 기준으로 키워드의 빈도를 도출하고, 키워드 간 연관성과 올해 어떠한 자연재해가 발생하였는지 분석하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 '기상'과 빅데이터와 관련된 연구를 기술한다. 3장에서는 본 논문에서 구현한 워드 클라우드 형태의 그림을 표현하기 위해 R 프로그램을 활용한 데이터 분석 방법에 대해 기술한다. 4장에서는 워드 클라우드 형태의 그림으로 표현한 기상뉴스 데이터의 결과와 키워드를 설명하고, 마지막 5장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 관련연구

기존의 연구에서는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 소셜 마이닝 기법 등 다양한 기법을 통한 빅 데이터 분석연구가 있었다. 정보통신의 발달과 소셜 미디어의 급속한 확산으로 빅 데이터가 경제적으로 자산이 되고 있는 시대를 맞이하는 데 필요한 데이터 분석기법과 인프라 기술에 대해 알아보고, 한글 Text 데이터를 R 프로그램을 이용하여 usesejongdic() 이라는 옵션을 이용하여 명사만 추출하는 방법으로 비정형 데이터를 분석하였다.[1] 데이터 시각화 도구 통계 패키지인 R을 이용하여 대기오염의 자료를 여러 가지 방법의 데이터 시각화를 통하여 나타내었고, 데이터 시각화 방법별로 통계적인 방법을 활용한 분석과 연계하여 어떤 특징이 있는지를 나타냈다. 2차원의 히스토그램과 선점도, 상자그림, 3차원 산점도와 투시도 등 다양한 방법의 그래프를 구현하여 오존농도와 설명 변수들 간에 어떠한 관련성이 있는지를 분석했다.[2] 날씨는 사람의 심리와 행동에 다양한 영향을 미친다. 날씨가 계절상품의 매출에 미치는 영향에서, 국내 대형 할인점의 각 제품별 일일 판매데이터와 기상 정보 데이터의 분석을 통해 기상 요인들이 매출에 미치는 영향에 대한 분석연구가 있었다. 대형 할인점 중의 하나인 H사의 서울 K지점의 2년간 상품별 매출 데이터와 매장이 위치한 서울 지역의 기상과 날씨 데이터를 이용하여 상품 매출에 미치는 영향을 다변량 분산분석(MANOVA)을 통해 계절별, 주중/주말에 따른 매출액 차이와 다중회귀분석을 통해 기상 요인이 매출액에 미치는 영향을 분석하고 영향의 정도를 파악하였다.[3] 데이터 마이닝의 일부인 텍스트

마이닝의 기법을 이용하여 부산지역지인 국제신문과 부산일보의 기사들 중 제목에 '부산'과 '교통'을 동사에 포함한 기사의 기사 내용의 관계 또는 관련 있는 데이터에 내재되어 있는 의미 있는 패턴을 찾는 사회네트워크분석을 실시하여 정형화된 빅 데이터를 시각화하고 해석했다.[4] 빅데이터 분석도구인 R을 이용하여 성경의 텍스트 데이터를 성경전체, 구약성경, 신약성경, 모세오경, 사복음서 데이터 분석결과를 각각의 워드 클라우드 형태 그림으로 표현하여 성경데이터를 분석하여 성경을 읽는 독자에게 주는 메시지가 무엇인지에 대한 연구를 제시하였다.[5]

III. 데이터 분석 방법

빅데이터 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현한다. 기상뉴스 데이터는 검색 포털 사이트를 이용하여 '기상'키워드를 검색하여 본문내용을 중심으로 스크랩하여 텍스트 파일의 데이터를 수집했으며, 데이터를 분석하기 위해 언론사 일간지인 경향신문, 국민일보, 내일신문, 동아일보 외 9개의 언론사일간지의 지면기사를 기준으로 약 3천 건 이상의 기사를 분석하였다. 데이터 분석과정은 [그림 1]과 같다.

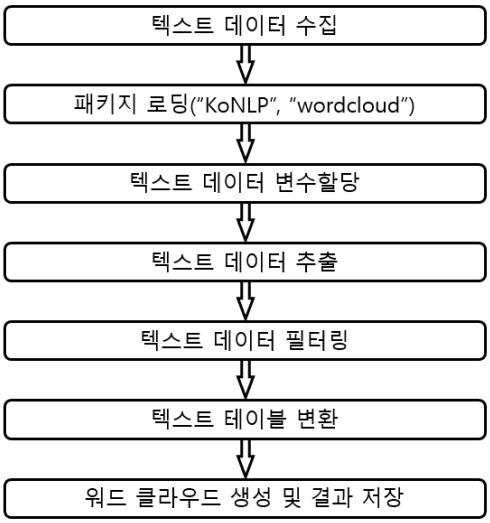

그림 1. 데이터 분석과정

빅데이터 분석도구인 R을 설치하고 한글 데이터 분석에 필요한 패키지("KoNLP"), 워드 클라우드 생성에 필요한 패키지("wordcloud")를 설치하고 R 소스에 로딩한다. 수집한 데이터를 경남신문, 국제신문, 부산일보, 중앙일보, 전체기사의 그룹으로 분류하여 각 그룹의 데이터를 변수를 할당하여 대입한다. 한글의 명사를 추출해주는 함수인 'extracNoun'함수를 사용함으로서 성경 데이터를 명사로 변환하여 변환된 데이터를 확인 후 원하지 않는 데이터에 대한 'gsub'함수를 이용하여 데이터를 필터링 한다. 여기서는 2자리 이상의 명사만 추출하도록 프로그램을 구현하였다. 필터링된 데이터를 텍스트 형식의 파일로 저장하여 테이블 형태로 변환하여 변수에 할당한다. 텍스트 형태로 각 명사에 대한 빈도수를 측정하여, 상위 30위의 결과를 워드 클라우드 형태의 그래픽으로 출력한다. 출력 결과물을 이미지파일(JPGE, BMP, PNG 등)으로 저장한다.

IV. 데이터 분석 결과 및 비교

본 논문에서는 '기상' 관련 데이터 분석의 결과를 워드 클라우드와 키워드 빈도 수에 대하여 표현하였다. 워드 클라우드는 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심단어를 시각적으로 돋보이게 하는 기법이다. 예를 들면 키워드의 빈도수가 높을수록 키워드를 크게 표현하고 각각 색상을 적용함으로 한 눈에 들어올 수 있게 하는 기법이 있으며, 표, 그래프와 같은 시각적인 표현도 가능하다.

그림 2. 전체 키워드

표 1. 전체 키워드 빈도수

기상	지진	기온	발생	지역
2939	1914	1704	1695	1260
전국	폭염	예보	최고	이상
1254	1006	962	854	832
오후	제주	규모	예상	전망
831	806	770	760	760
서울	영향	기록	영하	날씨
756	756	716	596	594
오전	시작	먼지	올해	미세
515	512	499	494	471
평년	지방	태풍	시간	관계자
471	464	459	399	395

[그림 2]는 2016년 1월1일 부터 9월30일까지의 각 언론사 일간지의 지면기사를 수집한 데이터의 전체 키워드를 그림으로 표현하였고, [표 1]은 전체 키워드 중 상위 30위에 해당하는 키워드 빈도를 나열하였다.

표1에서 검색어인 '기상'키워드를 제외하고 그

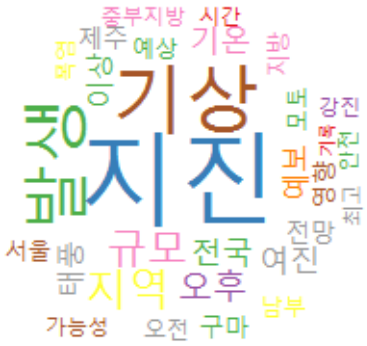

그림 3. 4월·7월·9월 키워드

다음으로 '지진' 1,914회, '폭염' 1,006회 이상 832회, 영향 756회, 기록 716회, 먼지499회, 미세471회, 시간399회를 놓고 보았을 때, 미세먼지에서 '미세'와 '먼지'의 키워드가 분리되어 나타났고, '지진', '폭염', '태풍' 등의 자연재해 현상과 관련된 단어들이 나타났다.

'지진' 키워드는 4월 577회, 7월 520회, 9월 698회로 [그림 3]과 같이 워드 클라우드가 나타났고, 각 월별 상위30위 중 1위에 해당하는 키워드로 나타났다. 키워드 유추결과 4월 14일 21시경 일본 구마모토 지진, 7월 5일 20시경 울산 동쪽 52km 부근 규모5.0의 지진, 9월 12일 20시경 경주 남남서쪽 8km 부근 5.8의 지진 발생을 알 수 있고, 4월·7월·9월의 키워드 빈도수는 [표 2]와 같다.

표 2. 4월·7월·9월 키워드 빈도수

지진	기상	발생	규모	지역
1795	1280	1147	692	590
오후	전국	이상	예보	여진
458	440	392	335	334
기온	태풍	영향	오전	예상
331	327	309	296	285
모토	구마	전망	지방	제주
270	268	252	246	240
폭염	남부	최고	서울	안전
231	224	199	198	195
시간	중부지방	기록	가능성	강진
192	191	187	180	180

'폭염' 키워드는 5월 74회, 6월 57회, 7월 222회 8월 619회로 여름의 막바지에 이를수록 그 빈도수가 증가하는 것으로 나타났으며, 키워드 유추결과 5월부터 8월까지에서 '열대야' 143회, '더위' 308회. '온열' 100회, 질환' 94회로 폭염과 연관되는 키워드와 질병 등이 나타났고, '태풍' 키워드는 7월 195회, 8월 96회, 9월 130회의 빈도로 나타났으며, 키워드 유추결과 7월부터 9월까지에서 '장마전선' 159회, '호우' 110회 등 태풍과 연관되는 키워드가 나타났음을 알 수 있었다.

이 밖에도 [표 1]에 나타나지 않은 '경보' 326회, '주의보' 279회, '특보' 234회 등의 안전조치와 경고 등의 뜻으로 전달되는 키워드와 '중국' 284회, '세계' 236회, '미국' 231회 등 인접국가에 대한 기상 및 자연재해에 대한 보도자료와 내용이 포함되어 있었고, '황사' 337회, '폭우' 135회,'장맛비' 128회, '엘니뇨' 178회 등의 자연재해도 나타나 있었다. '우리나라', '정부', '관리', '대책' 키워드를 나타냄으로 기상현상과 자연재해 대책과 관리에 대해 우리나라 정부는 물론 각 기관에서도 많은 관심을 가지고 있음을 추측할 수 있었다.

V. 결론 및 향후 연구

본 논문에서는 정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 응용하고 있다. 빅데이터 분석도구 R을 이용하여 기상뉴스에 나타난 '기상'관련 빅데이터를 워드 클라우드 형태의 그림으로 나타내어 언론사 일간지 지면자료에 나타난 키워드를 워드 클라우드로 시각화함으로써 빈도수에 따른 키워드를 쉽게 알아 볼 수 있었다. R프로그램을 이용함으로써 누구나 쉽게 접근하여 다양한 데이터를 워드 클라우드 형태의 그림으로 시각화 표현을 구현할 수 있다고 본다.

향후 연구 방향으로서 기상관련 빅데이터를 수집하여, 앞으로 대기업 마케팅의 전략으로서 기상마케팅과 슈퍼컴퓨터 활용으로 기상예측 정확성을 높이는 연구, 농업기술에 기상 빅데이터를 접목하여 스마트 농업기술의 발전에 도움이 될 수 있으며 앞으로의 기상현상에 도움이 될 수 있고 다양한 분야에 이를 응용할 수 있을 것이다.

참고문헌

[1] 김현근, "R을 이용한 빅 데이터 사례 분석", 호서대학교 일반대학원 정보통계학과 석사학위논문, 2014.

[2] 오영창, 박은식, "R 소프트웨어를 이용한 대기오염 데이터의 시각화", 한국데이터정보과학회지, 26(2), pp399-408, 2015.

[3] 홍진환, 이현정, 나준희, "기상요인이 할인점의 계절상품 매출에 미치는 영향", 유통경영학회지, 15(6), pp5-15, 2012.

[4] 이경준, 노윤환, 윤상경, 조영석, "부산지역 교통관련 기사를 이용한 비정형 빅데이터의 정형화와 시각적 해석", 한국데이터정보과학회지, 25(6), pp1431-1438, 2014.

[5] 김용수, 반재훈, "성경 데이터를 활용한 빅데이터 분석", 한국정보통신학회 2015 추계종합학술대회, pp349-352, 2015.