차세대 에너지 관련 뉴스 빅데이터 분석

이예찬, 조해찬, 반재훈

고신대학교 IT경영학과

The Next Generation of Energy News Big Data Analytics

YeChan Lee, HaeChan Cho, ChaeHoon Ban

Dept. of IT Management, Kosin University

E-mail : ycl0188@naver.com · chc0451@naver.com · chban@kosin.ac.kr

요 약

대규모의 데이터가 생산되고 저장되는 정보화 시대에서 현재와 과거의 데이터를 바탕으로 미래를 추측하고 방향성을 알아갈 수 있는 빅데이터의 중요성이 강조되고 있다. 정형되지 못한 대규모 데이터를 빅데이터 분석 도구인 R을 통해 통계를 기초로 데이터의 정보분석과 정형화하도록 한다. 본 논문에서는 R을 이용하여 뉴스에서 나타나는 차세대 에너지 관련 빅데이터를 분석한다. 뉴스 기사에서 차세대 에너지 관련 데이터를 수집하고 수집된 키워드를 이용하여 근미래의 효율적인 차세대 에너지의 등장을 예측한다. 에너지 산업의 추진에 대한 흐름과 방향성을 제시하고 의사결정을 위한 기술적 과제를 도출함으로 탄력적인 경영과 의사결정에 도움을 주며 기술적 문제의 근원을 사전에 예측하고 방지할 수 있을 것으로 보여진다.

키워드

Big Data, R, Text Mining, Transportation, Analysis

Ⅰ. 서론

미디어의 발전과 확산으로 대규모의 비정형 데이터가 생산되는 정보화 시대에서 데이터를 수집하고 수집된 데이터를 이용하여 미래를 추측하고 방향성을 찾아가는 빅데이터 분석이 강조되고 있으며 다양한 산업에서 이를 활용하고 있다. 빅 데이터 수집 도구인 파이썬(python)은 데이터의 수집을 가능하게 하는 언어와 환경이다. 빅 데이터 분석 도구인 R은 통계기반의 정보 분석을 가능하게 하는 언어와 환경이다. 차세대 에너지의 데이터를 활용하여 미래의 변화를 예측함으로 차세대 국가 산업, 차세대 에너지 연구의 방향을 제시하고 미래 에너지 산업의 방향성을 제시함으로 차세대 에너지 데이터 분석이 중요하다고 볼 수 있다.

‘차세대 에너지’와 관련된 키워드를 중심으로 최근 5년간 데이터를 기준으로 주요 언론사(조선일보, 중앙일보, 동아일보)와 에너지관련 전문지에서 데이터를 수집하고, 키워드의 빈도를 도출하여 어떠한 차세대 에너지가 최근 5년간 노출되었는지 분석하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 ‘차세대 에너지’의 빅 데이터와 관련된 연구를 기술한다. 3 장에서는 본 논문에서 구현한 워드 클라우드 형태의 그림을 표현하기 위해 R 프로그램을 활용한 데이터 분석 방법에 대해 기술한다. 4장에서는 워드 클라우드 형태의 그림으로 표현한 각 신문사의 결과와 비교를 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] 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특

허검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행했다.[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 사회네트워크 분석을 실시하여 정형화된 빅 데이터를 시각화하고 해석했다.[4] 구글, 야후, 네이버 등 주요 포털의 지도에는 POI(Point of interest)가 서비스 되고 있다. 지도의 위치 데이터 즉, 현재 이용자가 위치한 장소는 인문학적인 스토리텔링의 시작점을 주목하여, POI는 카페, 레스토랑, 병원, 식당 등의 정보만이 서비스되는 한계점을 지적하고, 더 나아가 대안으로 POI 정보와 결합된 소위 '인문융합 지도 서비스'를 제안 했다.[5] 빅데이터 분석 도구인 R을 이용하여 미디어에서 나타난 부산교통관련 텍스트 데이터를 1년간 자료와 5년간 자료로 나누어 데이터 분석결과를 각각의 워드 클라우드 형태 그림으로 표현하여 부산 교통의 최근 1년간과 5년간의 변화를 통계/분석하여 부산교통의 변화를 찾고 부산교통의 발전 방향성을 제시하였다.[6]

Ⅲ. 데이터 분석 방법

빅데이터 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현한다. 신문기사의 데이터는 각 언론사 홈페이지를 이용하여 '차세대 에너지' 관련 키워드에 접속하여 본문내용을 중심으로 파이썬(python) 프로그램을 통해 데이터를 스크랩하여 텍스트 파일의 데이터를 수집하였으며, 데이터를 분석하기 위해 주요 언론사 조선일보, 중앙일보, 동아일보와 에너지관련 전문지 YTN사이언스, 에너지경제, 테크홀릭, 투데이에너지, 헬로디디, 가스신문, 동아사이언스, 동아오토, 전기신문에서 언론사 및 전문지의 지면기사를 기준으로 약 3500 건 이상의 기사를 분석하였다.

데이터 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 차세대 에너지 데이터를 명사로 변환하여 변환된 데이터를 확인 후 원하지 않는 데이터에 대한 'gsub'함수를 이용하여 데이터를 필터링 한다. 여기서는 2자리 이상의 명사만 추출하도록 프로그램을 구현하였다. 필터링 된 데이터를 텍스트 형식의 파일로 저장하여 테이블 형태로 변환하여 변수에 할당한다. 텍스트 형태로 각 명사에 대한 빈도수를 측정하여, 상위30위의 결과를 워드 클라우드 형태의 그래픽으로 출력한다. 출력 결과물을 이미지파일(JPGE, BMP, PNG 등)으로 저장한다. 데이터 분석과정은 [그림 1] 과 같다.

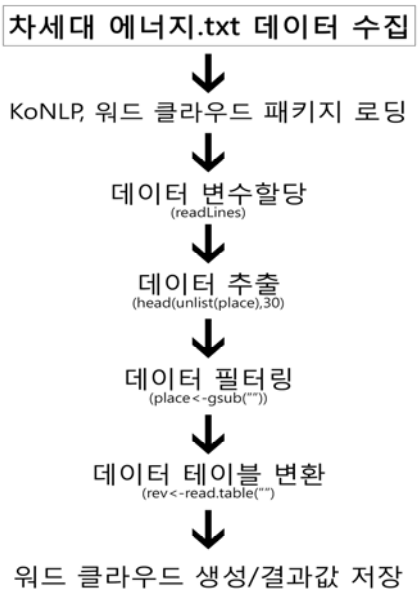

그림 1. 데이터 분석 과정

Ⅳ. 데이터 분석 결과 및 비교

본 논문에서는 '차세대 에너지' 관련 데이터 분석의 결과를 워드 클라우드와 키워드 빈도수에 대하여 표현하였다. 워드 클라우드란 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법이다. 예를 들면 표를 나타내어 키워드의 빈도수가 높을수록 상단에 그 키워드를 노출시키는 기법이 있으며, 키워드가 많이 언급될수록 단어를 크게 표현해 한눈에 들어올 수 있게 하는 기법 등이 있다.

표 1. 전체 키워드 빈도수

기술	개발	전기	시장	산업
3513	3025	1779	1626	1619
연구	발전	전력	세계	전지
1591	1532	1508	1401	1020
수소	자동차	생산	연료전지	제품
913	834	822	792	772
태양광	효율	경제	재생	스마트
745	719	717	696	632
성장	구축	소재	배터리	가스
604	555	554	534	492
태양전지	리튬	전자	연료	안전
487	473	454	427	424

[표 1]은 2011 년 10 월 4 일부터 2016 년 10 월 4일까지의 주요언론사 및 에너지관련 전문지

총 12곳의 신문사에서 최근 5년간 노출된 기사를 수집하여 상위 30개 단어의 키워드 빈도를 표로 나열하였다.

[표 1]에서 1, 2번째로 높은 빈도를 차지하는 '기술'의 빈도는 3513회, '개발'의 빈도는 3,025회로 3번째로 차지한 '전기'의 빈도 1,779회를 압도함으로 기술개발에 관한 기사들이 많이 시사되었음을 엿볼 수 있었다. 차세대 에너지의 전기분야인 '전지', '전력', '배터리', '태양전지'의 총 빈도가 2,621회로 '전기'의 빈도 1,779회를 밑돌아 차세대 에너지저장장치(ESS)의 투자를 확인할 수 있었다. ESS는 전기를 대용량으로 저장하였다가 원하는 시간에 방전할 수 있도록 함으로써, 전력산업의 패러다임을 바꾸는 에너지 신산업의 기반재이다. 전력주파수 평탄화를 통한 전력품질 제고, 풍력 등 신재생발전의 효율성 제고, 밤에 생산된 전력을 낮의 전력 시간에 사용함으로써 최대전력 수요 감소 및 전력의 절감효과를 기대 할 수 있다.

'효율'이 719회 언급되었고 '경제'가 717회 언급됨으로 차세대 에너지의 기술개발에서 효율성과 경제성을 높게 추구한다는 것을 알 수 있었으며 '스마트'의 빈도 또한 632회로 스마트 에너지원을 개발하려는 움직임도 확인되고 있다. '자동차'의 빈도가 834회로 자동차의 차세대 에너지원에 관한 신문기사가 시사되었음 또한 알 수 있었다.

'수소'는 913회 언급되었는데 수소는 차세대 무공해 에너지원으로서 중시되며 강력한 대체에너지로 부각되고 있다. 수소로 가는 자동차, 발전소, 수소 충전소 등이 점점 생기고 있고 기술이 개발되고 발전되고 있다. 미래에는 현시대의 기름을 사는 것과 같이, 수소를 편하게 사는 날이 오게 될 것이다.

그림 2. 전체 키워드

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 미디어 매체에 나타난 차세대 에너지관련 빅데이터를 워드 클라우드 형태의 그림으로 나타내어 신문기사에서 나타는 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 더욱 많은 빅데이터, 차세대 에너지관련 데이터를 수집하여 앞으로 차세대 에너지의 기술을 각 도시의 에너지 상황과 접목하여 효율적인 차세대 에너지의 개발과 관리에 도움이 될 수 있으며, 앞으로의 문제점과 해결방안에 대한 예측을 시도할 수 있고, 다양한 분야에 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] 장청윤, 장정환, 김석주, 이현군, 이창호, "빅데이터 분석 도구 R을 활용한 효율적인 특허검색에 관한 연구", 대한안전경영과학회지, 15(4), pp289-294, 2013.

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

[5] 이원태, 강장묵, "빅데이터 중 POI와 공간 메타포를 활용한 인문 융합 지도 연구", 한국인터넷방송통신학회. 15(3), pp43-50, 2015.

[6] 반재훈, 김용수, 이예찬, 정윤성, 정동민, 조해찬, "미디어에서 나타난 부산 교통 관련 빅데이터의 분석", 한국정보통신학회 2016 춘계종합학술대회, pp349-352, 2016.