빅데이터 분석 도구 R 언어를 이용한 비정형 데이터 시각화

남수태[1] · 진금회[2] · 신성윤[3] · 진찬용[2,*]

[1]부산대학교 · [2]원광대학교 · [3]군산대학교

Visualizing Unstructured Data using a Big Data Analytical Tool R Language

Soo-Tai Nam[1] · Jinhui Chen[2] · Seong-Yoon Shin[3] · Chan-Yong Jin[2,*]

[1]Pusan National University · [2]Wonkwang University · [3]Kunsan National University

E-mail : stnam@pusan.ac.kr / cjh564121294@gmail.com / s3397220@kunsan.ac.kr / jcy85366@wku.ac.kr

요 약

빅데이터 분석은 데이터 저장소에 저장된 대용량 데이터 속에서 의미 있는 새로운 상관관계, 패턴, 추세를 발견하여 새로운 가치를 창출하는 과정이다. 또한 대부분의 빅데이터 분석 기술 방법들은 기존 통계학과 전산학에서 사용되던 데이터 마이닝, 기계 학습, 자연 언어 처리, 패턴 인식 등이 이에 해당된다. 그리고 빅데이터 분석 도구인 R언어를 이용하여 전-처리된 텍스트 데이터를 이용하여 다양한 시각화 함수를 통해 분석결과를 표현할 수 있다. 본 연구에서 사용된 데이터는 한국정보통신학회 학회지 논문 중에서 2021년 3월호 논문 21편을 대상으로 분석을 하였다. 최종 분석결과는 가장 많이 언급된 키워드는 “데이터”가 305회로 1위를 차지하였다. 따라서 이러한 분석결과를 바탕으로 연구의 한계와 이론적 실무적 시사점을 제시하고자 한다.

ABSTRACT

Big data analysis is the process of discovering meaningful new correlations, patterns, and trends in large volumes of data stored in data stores and creating new value. Thus, most big data analysis technology methods include data mining, machine learning, natural language processing, and pattern recognition used in existing statistical computer science. Also, using the R language, a big data tool, we can express analysis results through various visualization functions using pre-processing text data. The data used in this study was analyzed for 21 papers in the March 2021 among the journals of the Korea Institute of Information and Communication Engineering. In the final analysis results, the most frequently mentioned keyword was “Data”, which ranked first 305 times. Therefore, based on the results of the analysis, the limitations of the study and theoretical implications are suggested.

키워드

네트워크 분석, 비정형 데이터, 빅데이터, 연관성 분석, 텍스트 마이닝

Ⅰ. 서 론

빅데이터 분석은 데이터베이스에 잘 정리된 정형 데이터뿐만 아니라 인터넷, 소셜 네트워크 서비스, 모바일 환경에서 생성되는 웹 문서, 이메일, 소셜 데이터 등 비정형 데이터를 효과적으로 분석하는 기술을 말한다. 대부분의 빅데이터 분석 기술 방법들은 기존 통계학과 전산학에서 사용되던 데이터 마이닝, 기계 학습, 자연 언어 처리, 패턴 인식 등이 이에 해당된다. 또한 정보통신기술의 발전

* corresponding author

은 우리생활에서 발생하는 대규모의 비정형 데이터를 수집하고 수집된 데이터를 이용하여 미래를 예측할 수 있는 빅데이터 기술의 중요성이 강조되고 있으며 다양한 산업에서 이를 활용되어 지고 있다. 텍스트 마이닝은 비정형 텍스트 데이터에서 새롭고 유용한 정보를 찾아내는 기술이라고 말할 수 있으며, 비정형 데이터를 자연어처리(natural language processing) 기술에 기반을 두고 데이터를 가공한다. 즉 전처리를 통해 비정형 데이터에서 정형화된 데이터로 바꾸어 특징을 추출하는 과정을 뜻한다. 감성분석(sentiment analysis)은 텍스트에 표현된 개체 및 속성에 대한 의견이나, 감성, 태도, 평가 등을 분석하여 텍스트에 나타난 감성을 분류하는 것을 말한다. 감성분석은 크게 2가지로 분류하는데 여기서 분석 데이터에 레이블(label)이 있는 경우와 없는 경우에 따라 지도학습(supervised learning) 및 비지도학습(unsupervised learning)으로 구분한다[1]. 오늘날 스마트 기기의 대중화는 시간과 공간을 초월한 인터넷 사용의 대중화를 가능하게 하였으며 이를 통해 사람과 사람은 물론 사람과 사회를 연결하는 매개역할을 하고 있다. 빅데이터 분석 기술은 매우 다양하게 존재한다. 그 중에서도 최근 가장 많이 사용하는 도구로는 R언어를 기반으로 한 통계기반의 데이터 분석을 가능하게 하는 언어의 환경이다.

II. 선행 연구

오늘날 빅데이터 분석 연구에는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 소셜 마이닝 등등 다양한 통계기법을 통한 빅데이터 분석 연구가 급속하게 증가함을 알 수 있다. 먼저, R을 이용한 빅 데이터 사례 분석[2]에서 정보통신의 발달과 소셜 미디어의 급속한 확산으로 생산된 빅데이터를 분석하는 기법 및 인프라 기술에 대해 기술하고 한글 텍스트 데이터를 R언어를 이용하여 usejongdic() 함수를 이용하여 명사만 추출하는 방법으로 비정형 데이터를 분석하였다. 그리고 R 소프트웨어를 이용한 대기오염 데이터의 시각화[3]에서는 대기오염의 자료를 여러 가지 방법의 데이터 시각화를 통해 히스토그램과 선점도, 상자그림, 3차원 산점도와 투시도 등 다양한 방법의 그래프를 구현하여 오존농도와 변수들 간에 미치는 영향을 분석하고 있다. 다음으로 빅데이터 분석 도구 R언어를 이용한 교육 자료 시각화[4]에서는 초중등 교과 내용을 포함하여 교육 자료를 시각화하여 그 특성을 파악하는데 빅데이터 분석 기술을 적용하고자 하였고, 마인드 맵 형태의 시각화 교육 자료를 통해 교수자와 학습자는 교육 내용에 대한 이해와 학습 제고를 가져올 것으로 기대 한다고 하였다. 다음으로 빅데이터 분석도구 R을 이용한 성경 데이터의 빈도와 소셜 네트워크 분석[5]에서는 성경 중에서 신약성경의 4복음서의 데이터를 분석하였으며, R을 이용하여 어떠한 텍스트가 분포되어 있는지를 빈도 조사를 수행하여 정확한 데이터의 분석을 위해 한 문장에서 나오는 단어들을 쌍으로 표현하고 단어 간의 관계성을 분석하는 소셜 네트워크 분석을 통해 성경을 분석한다. 이외 R언어를 이용하여 다양한 비정형 데이터를 분석한 선행 연구를 찾아볼 수 있었다.

III. 빅데이터 분석

본 연구에서 사용된 데이터는 한국정보통신학회 학회지 논문 중에서 2021년 3월호 논문 21편을 대상으로 하였다. 21편의 논문 중에서 영문 논문 1편을 제외하고 20편의 논문을 최종 분석에 사용되었다. 아래 표 1은 분석에 사용된 논문 리스트이다.

표 1. 분석에 사용된 논문 리스트

No	Title of Articles
1	GF(p) 상의 다중 체 크기를 지원하는 고성능 ECC 프로세서
2	TheImpactofBlockchainTechnologyonBanks' Conventional
3	YOLO 기반 차선검출 시스템
4	광 주입 파장 잠금 반도체 레이저를 이용한 광학 복소 신호 생성시
5	군집 비행 드론의 충돌 방지를 위한 UWB 레이다의 속도 감응형
6	기계학습을 이용한 동영상 서비스의 검색 편의성 향상
7	딥러닝기반객체인식을활용한퍼스널모빌리티안전보조
8	딥러닝 기반 자동 변조 인식 성능 분석
9	마이크로프로세서 기반의 얼굴 마스크 감지
10	명품 하울 유튜브 영상 댓글에 나타난 상대적 박탈감 여부와 특징
11	샴 네트워크 기반 객체 추적을 위한 표적 이미지 교환 모델
12	스마트 그리드를 위한 블록체인 기반 LoRa 멀티홉 네트워크 설계
13	스마트 조선소를 위한 사물인터넷 기반 용접작업장 센서네트워크
14	이동 장애물을 고려한D QN기반의 MaplessNavigation 및 학습
15	자율주행차용 우선순위 기반 다중 DNN 모델 스케줄링 프레임워크
16	적응적 가우시안 혼합 모델을 이용한 불법주정차 무인단속시스템
17	제품디자인의 감성공학 요소가 브랜드 선호도와 충성도에 미치는
18	증강현실기반의키즈콘텐츠제작을위한관찰조작형모델의
19	코로나 19와 서울 소상공인 상권의 상관관계 분석
20	팀 기반 전투형 게임에서 여성 게이머가 선호하는 여성 캐릭터
21	효과적인 디스플레이 제조를 위한 AIBIG DATA 기반 스마트 팩토리

빅데이터 분석도구인 R언어를 이용하여 텍스트 데이터를 R Studio에서 제공하는 다양한 시각화 도구(함수)를 통해 나타내고자 한다. 먼저, 빈도분석에 해당하는 워드 클라우드 함수를 이용하여 그림으로 표현하고 데이터 사이의 관계를 분석하여 네

트워크 그래프를 생성하고자 한다. 다음으로 대상 논문 데이터에서 한글 단어를 추출하기 위해 R언어에서 제공하는 KoNLP 패키지 함수를 사용하여 데이터 분석을 하였다. KoNLP는 한국어 텍스트 기반 연구를 위한 형태소 분석과 형태 분석법을 제공하는 패키지이다. 본 패키지에서 제공하는 함수는 한글 명사를 추출하는 함수이며 “extracNoun” 함수를 사용하여 분석에 사용된 논문에서 명사 부분만을 추출하였다. 또한 시각화 분석에 불필요한 데이터는 필터링을 통해 정제 작업을 수행하였다. 그리고 2자리 이상의 명사만 추출하도록 코딩을 구현하였으며 필터링 된 데이터를 텍스트 형식의 파일로 저장하였다. 또한 상위 30위의 결과를 워드 클라우드 형태의 그래프로 출력하였고 단어를 노드로 표현하고 단어와 단어의 관계를 에지 형태로 표현하였으며 그 관계의 빈도를 노드의 크기로 표현한 소셜 네트워크 그래프를 생성하였다. 본 연구에서 여기까지 하고 차후 연구에서는 보다 정교한 데이터 분석을 위해 딕셔너리를 구축하여 해당되는 단어를 추출과 분석을 수행하고자 한다. 데이터의 분석과정은 다음 그림 1과 같다.

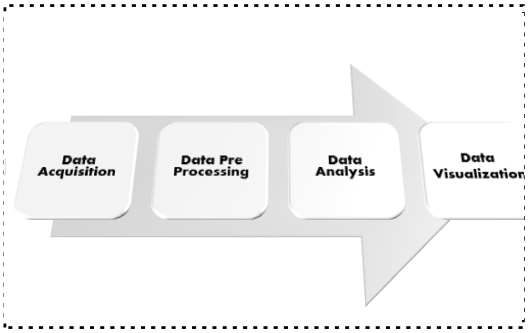

그림 1. 데이터 분석과정

Ⅳ. 분석결과 및 시각화

본 연구에서 사용된 데이터는 한국정보통신학회 학회지 논문 중에서 2021년 3월호 논문 21편을 대상으로 워드 클라우드 함수를 이용한 시각화 분석 결과는 다음 표 2와 같다. 워드 클라우드는 분석 대상 문서 내용의 키워드나 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 보여주게 하는 기법이다. 예를 들자면 많이 언급된 단어일수록 의미를 크가 부각시키는 방법으로 불 수 있으며, 그러한 단어를 크게 표현해 한눈에 직관적으로 보여주는 기법이다. 하지만 워드 클라우드는 단순히 단어의 반복됨을 파악하는 것이기 때문에 전체적인 단어 간의 관계, 관련성을 파악하기는 쉽지 않다. 최종 분석결과를 살펴보면, 가장 많이 언급된 키워드는 “데이터”가 305회로 1위에 위치하였다. 다음으로는 사용이 277회, 모델이 277회, 모델이 237회 각각 차지한 키워드인 것을 알 수 있었다. 다음 순으로는 그림이 198회, 학습이 167회, 기반이 155회 순으로 언급된 키워드 임을 확인시켜 주었다.

표 2. 워드 클라우드를 이용한 빈도분석

순서	키워드	언급수	순서	키워드	언급수
1	데이터	305	11	이용	129
2	사용	277	12	분석	125
3	모델	237	13	Fig	116
4	그림	198	14	영상	116
5	학습	167	15	네트워크	114
6	기반	155	16	제안	112
7	결과	149	17	하기	112
8	신호	149	18	적용	110
9	연산	144	19	시스템	109
10	경우	129	20	추출	109

다음 순으로는 결과가 149회이고 신호도 149회로 동일한 언급수를 기록하였으며, 그 다음으로는 연사이라는 키워드가 144회이었으며, 경우의 키워드가 129회 차지하여 10위로 언급된 회수로 기록되어 분석에 사용된 논문 중에 중간 정도의 의미로 부여할 수 있다는 것을 빅데이터 분석을 통해 확인 시켜 주었다.

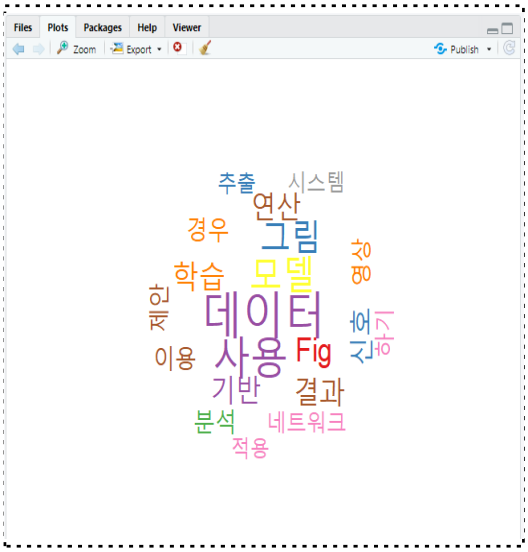

그림 2. 데이터 분석결과에 대한 시각화

앞에서 말한 것처럼 무조건 언급된 수가 높다고 하여 의미 있는 키워드라고는 말할 수는 없지만 자주, 많이 언급된다는 의미는 때에 따라서는 중요한 의미를 가질 수 있다는 것으로도 해석될 수 있다. 다음으로 위에 표 2는 워드 클라우드 함수를 이용

하여 빈도분석 수행한 결과를 나타내 다음으로 위에 표 2는 워드 클라우드 함수를 이용하여 빈도분석 수행한 결과를 나타내 것이라고 한다면 아래 그림 2는 데이터 분석결과를 시각화한 소셜 네트워크 그래프로 표현한 그림이라고 말할 수 있다.

Ⅴ. 결 론

본 연구에서는 R언어를 제공하는 워드 클라우드 함수에서 제공하는 시각화 도구인 소셜 네트워크 그래프를 이용하여 한국정보통신학회 학술지 2021년 3월호 모두에 해당하는 21편의 논문을 기초 데이터를 사용하여 빅데이터 분석 도구인 R언어를 이용하여 최종 분석을 수행하였다. 결론적으로 소셜 네트워크 그래프를 살펴보면 다음과 같이 결론을 내릴 수 있다. 먼저 워드 클라우드 분석에 결과에 따라 가장 많이 언급된 키워드가 그래프의 가장 중심에 그리고 가장 큰 글씨 크기로 표시하고 있다는 것을 우리는 확인할 수 있다. 여기에서 말하는 중심은 언급이 가장 많이된 “데이터”의 키워드가 전체의 추세나 트랜드를 대변 한다고 해도 무방하다고 할 수 있겠다. 다음으로 많이 언급된 키워드는 “사용”과 “모델”이 전체의 세나 조망을 보조하여 설명 한다고 볼 수 있다. 그런데 여기에서 적게 언급된 키워드라고 할지라도 중요하지 않다는 의미로 해석의 의미는 아니다. 본 분석결과를 통해 알 수 있는 것은 한국정보통신학회 학회지의 가장 중요한 핵심 키워드는 “데이터”이라는 것을 실증연구를 통해 증명되었다.

References

[1] H. Kim, S. Kim and H. Kim, “Crisis Prediction of Regional Industry Ecosystem based on Text Sentiment Analysis Using News Data - Focused on the Automobile Industry in Gwangju -,” *International JOURNAL OF CONTENTS*, Vol. 20, No. 8, pp. 1 - 9, Aug. 2020.

[2] H. Kim, *Big Data Case Study by Using R, M. S. thesis*, Hoseo University, Asan, Korea, 2014.

[3] Y. Oh, and E. Park, “Data visualization of air quality data using R software,” J*ournal of the Korea Data & Information Science Society*, Vol. 26, No. 2, pp. 39 - 408, Feb. 2015.

[4] Y. Kang, M. Kim, C. Hong S. Kim and S. Kwon, “Visualizing Educational Material using a Big Data Analytical Tool R Language,” *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol. 8, No. 3, pp. 915 - 924, Mar. 2018.

[5] J. Ban, J. Ha and D. Kim, “Frequency and Social Network Analysis of the Bible Data using Big Data Analytics Tools R,” *Journal of information and communication convergence engineering*, Vol. 24, No. 2, pp. 166p - 171p, Feb. 2010.