빅데이터 분석 도구 R 언어를 이용한 논문 데이터 시각화

남수태[1] · 신성윤[2] · 진찬용[1,*]

[1]원광대학교 · [2]군산대학교

Visualizing Article Material using a Big Data Analytical Tool R Language

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

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

E-mail : stnam@wku.ac.kr / s3397220@kunsan.ac.kr / jcy85366@wku.ac.kr

요 약

최근 빅데이터 활용은 매우 다양한 산업 분야에서 광범위하게 관심을 가지고 있다. 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 특정 학회지 논문 중에서 29편을 대상으로 분석을 하였다. 최종 분석결과는 가장 많이 언급된 키워드는 “연구”가 743회로 1위를 차지하였다. 따라서 이러한 분석결과를 바탕으로 연구의 한계와 이론적 실무적 시사점을 제시하고자 한다.

ABSTRACT

Newly, big data utilization has been widely interested in a wide variety of industrial fields. 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 were analyzed for 29 papers in a specific journal. In the final analysis results, the most frequently mentioned keyword was “Research”, which ranked first 743 times. Therefore, based on the results of the analysis, the limitations of the study and theoretical implications are suggested.

키워드

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

Ⅰ. 서 론

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].

* corresponding author

II. 선행 연구

R을 이용한 빅 데이터 사례 분석[2]에서 정보통신의 발달과 소셜 미디어의 급속한 확산으로 생산된 빅 데이터를 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. 이외 R언어를 이용하여 다양한 비정형 데이터를 분석한 선행 연구를 찾아볼 수 있었다.

III. 빅데이터 분석

본 연구에서 사용된 데이터는 특정 논문 중에서 29편을 대상으로 하였다. 빅데이터 분석도구인 R언어를 이용하여 텍스트 데이터를 R Studio에서 제공하는 다양한 시각화 도구(함수)를 통해 나타내고자 한다.

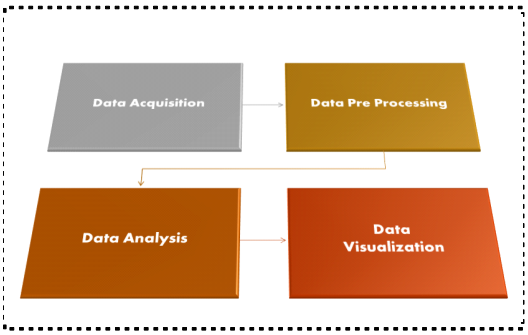

그림 1. 데이터 분석과정

먼저, 빈도분석에 해당하는 워드 클라우드 함수를 이용하여 그림으로 표현하고 데이터 사이의 관계를 분석하여 네트워크 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. 본 연구에서 여기까지 하고 차후 연구에서는 보다 정교한 데이터 분석을 위해 딕셔너리를 구축하여 해당되는 단어를 추출과 분석을 수행하고자 한다. 데이터의 분석과정은 그림 1과 같다.

IV. 분석결과 및 시각화

최종 분석결과를 살펴보면, 가장 많이 언급된 키워드는 "연구"가 743회로 1위에 위치하였다. 다음으로는 사용이 677회, 인지가 468회, 분석이 404회 각각 차지한 키워드인 것으로 나타났다. 아래 그림 2는 데이터 분석결과를 시각화한 소셜 네트워크 그래프로 표현한 그림이라고 말할 수 있다.

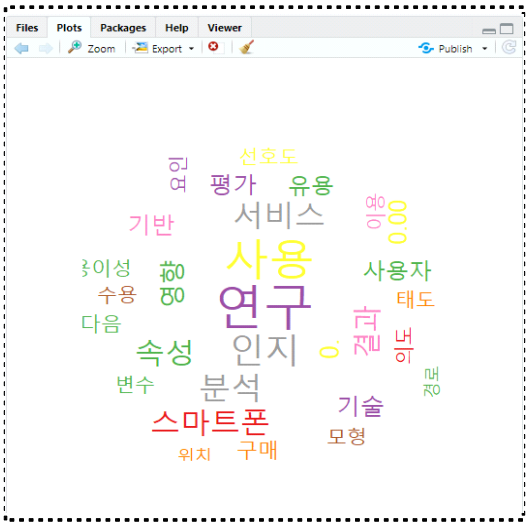

그림 2. 데이터 분석결과에 대한 시각화

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.