텍스트 마이닝을 이용한 한국정보통신학회 논문지의 주제 분석

우영운* · 조경원** · 이광의*

*동의대학교, **고신대학교

Topic Analysis of Papers of JKIICE Using Text Mining

Young Woon Woo* · Kyoung Won Cho** · KwangEui Lee*

*Dong-Eui University

**Kosin University

E-mail : ywwoo@deu.ac.kr

요 약

이 논문에서는 2007년부터 2016년까지 한국정보통신학회 논문지(JKIICE)에 게재된 3,668편의 논문들의 연구 주제 분야를 파악하기 위해 텍스트 마이닝 기법을 이용하여 논문들을 분석하였다. 자료 수집을 위하여 Python 기반의 웹 스크랩핑 프로그램을 사용하였으며, 자료 분석을 위해서는 R 언어로 구현된 LDA 알고리즘 기반의 토픽 모델링 기법들을 활용하였다. 연구 결과, 2016년까지 JKIICE의 투고 분야는 19개였으나 실제 최근 10년 동안 게재된 전체 논문들의 연구 주제는 크게 9가지로 대표됨을 알 수 있었다.

ABSTRACT

In this paper, we analyzed 3,668 papers of JKIICE from 2007 to 2016 using text mining methods for understanding research fields. We used web scraping programs of Python language for data collection, and utilized topic modeling methods based on LDA algorithm implemented by R language. In the results, we verified that representative research areas of JKIICE could be downsized to 9 areas only by the analysis though the submission areas were 19 areas by 2016.

키워드

텍스트 마이닝, 토픽 모델링, 한국정보통신학회 논문지, 연구 주제, 웹 스크래핑

Ⅰ. 서 론

최근 웹 크롤링의 발달과 그에 따라 수집되는 대규모 웹 문서들을 자동으로 분석하는 텍스트 마이닝 기법이 활발하게 사용되고 있다[1]. 특히 데이터 마이닝을 편리하게 활용할 수 있도록 해주는 R 또는 Python 등의 언어와, 그 언어에서 사용가능한 다양한 데이터 마이닝 관련 라이브러리 함수가 지원됨으로써 비교적 적은 분량의 코딩만으로 유용한 데이터 분석이 가능해지고 있다[2].

이 논문에서는 최근 한국정보통신학회 논문지(Journal of the Korea Institute of Information and Communication Engineering(JKIICE))에 게재된 논문들의 연구 주제가 어떤 연구 분야에 많이 포함되어 있는지, 그리고 연도별로 그 주제들이 비중이 어떻게 변화해 왔는지를 분석하였다. 2장에서는 자료 수집과 분석을 위한 방법을 설명하고 3장에서는 연구 결과를 제시하였다. 그리고 4장에서는 결과 분석에 대한 내용을 제시하고 5장에서 결론을 맺는다.

Ⅱ. 연구 방법

1. 자료 수집

이 논문에서는 한국정보통신학회 논문지의 연구 주제 변화를 분석하기 위하여 2007년부터 2016년까지 10년 동안 논문지에 게재된 논문들의 발간년월, 제목, 국문초록, 한글키워드를 논문별로 수집하였다. 이 정보들을 자동으로 수집하기 위하여 Python 웹 스크랩핑 프로그램[3]을 이용하였으며, 한국연구재단의 KCI 통합검색 사이트[4]에서 10년 동안 게재된 논문 3,668편의 발간년월, 제목, 국문초록, 한글키워드를 자동으로 추출하여

CSV 형태로 저장하였다.

2. 자료 분석

자료 분석을 위해서는 Anaconda Navigator 1.5에서 제공되는 Rstudio와 R 언어를 사용하였다. 자료 분석을 위해 가장 먼저 수집된 데이터의 전처리 과정이 필요하다.

이 논문에서는 데이터 전처리를 위하여 가장 먼저 제목, 국문초록, 한글키워드의 세가지 데이터를 한 단위로 처리할 수 있도록 한 문장으로 병합하였다. 그런 후 R에서 제공되는 NIADic을 사용하여 이 사전에 의해 형태소 분석을 수행한 후 보통명사와 영어단어로 추출되는 단어들만을 논문별로 저장하였다. 국문초록이라 하더라도 전문 용어인 경우에는 영어 단어를 그대로 사용하는 경우가 많기 때문에 연구 주제를 파악하기에는 영어로 된 전문 용어도 도움이 될 것으로 판단하여 함께 추출하였다. 이 때 추출된 보통명사와 영어 단어들 중 빈도수가 비교적 높으나 연구 주제를 파악하는데 관련이 없다고 판단되는 단어들을 모두 제거한 후 남은 단어들만을 논문별로 다시 저장하여 전처리 과정을 완료하였다.

다음 단계로 전처리가 완료된 논문 데이터들을 모두 하나의 처리 단위로 하여 LDA(Latent Dirichlet Analysis) 분석 기법을 활용하여 토픽 모델링 분석을 수행하였다[3]. 이 논문에서는 적절한 토픽의 수를 결정하기 위하여 토픽의 수를 5개에서 30개까지 변화시켜 가면서 LDA 분석 알고리즘에서 사용되는 기법들 중 VEM기법을 이용하여 R 언어에서 제공되는 perplexity 함수 결과값과 토픽의 해석 가능성, 의미 유용성 등을 고려하여 17개로 결정하였다. 토픽수가 결정된 후에는 LDA 분석 알고리즘에서 Gibbs sampling 기법을 이용하여 17개의 토픽별 로 빈도수가 높은 상위 15개의 단어들을 최종적으로 추출하였다.

III. 결 과

분석 결과 추출된 17개의 토픽에 대한 intertopic distance map은 그림 1과 같다. 그림에서 알 수 있듯이 하나의 토픽으로 볼 수 있는 중복된 토픽들이 여러 개로 나뉘어 추출되었다. 중복된 토픽들과 그 토픽들에 대한 실제 추출된 단어들을 분석해 본 결과 토픽 1, 3, 5가 '인터넷/모바일 정보처리', 토픽 8, 12, 13이 '반도체(물성)', 토픽 9, 10이 '무선 통신', 토픽 2, 14가 '신호 처리'로 거의 유사한 토픽임을 확인할 수 있었다. 이 외에 추출된 토픽들의 상위 단어들을 분석한 결과 '컴퓨터 네트워크', '센서 시스템', '인공지능 및 지능시스템', '반도체(설계)', '센서 정보처리', '지능형 차량 정보처리', '데이터 통신'에 대한 토픽임을 알 수 있었다. 따라서 반도체 2가지를 묶고, '센서 시스템'과 '센서 정보처리'를 센서와 관련된 유사한 연구 주제라고 본다면 2007년부터 2016년까지 투고된 논문들의 연구 주제는 크게 9가지로 판단할 수 있었다.

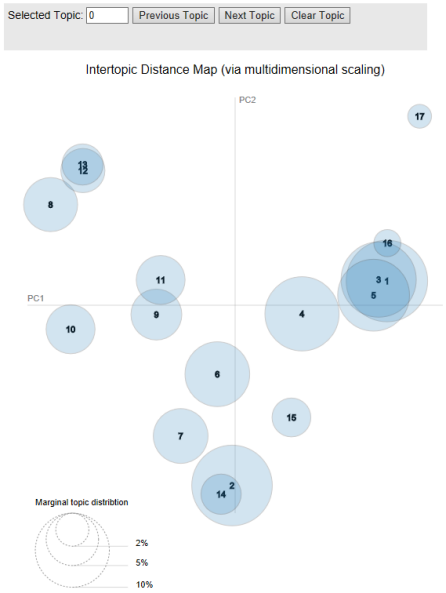

Fig. 1 Intertopic distance map for JKIICE papers

IV. 결 론

이 논문에서는 2007년부터 2016년사이에 JKIICE에 게재된 논문들을 대상으로 LDA 기반의 토픽 모델링 기법을 이용하여 대표 연구 주제들을 파악하였다. 2016년까지 논문 투고 분야를 19개로 운영하고 있었으나, 연구 결과 크게 9가지 정도의 연구 주제에 대한 논문이 가장 많이 게재되었음을 알 수 있었다. 이 결과가 향후 한국정보통신학회의 논문 투고 분야를 결정하는데 활용될 수 있을 것으로 기대한다.

참고문헌

[1] J. Y. Lee and Y. Bak, "Social Network Analysis of author's interest area in Journals about Computer," *Journal of the Korea Institute of Information and Communication Engineering*, vol.20, no.1, pp.193-199, Jan. 2016.

[2] J. Silge and D. Robinson, *Text Mining with R: A Tidy Approach*, Sebastopol, CA:O'Reilly Media, Inc., 2017.

[3] R. Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web*, 1st edition, Sebastopol, CA:O'Reilly Media, Inc., 2015.

[4] National Research Foundation of Korea. Korea Citation Index[Internet]. Available: https://www.kci.go.kr/kciportal/main.kci.