https://doi.org/10.5392/JKCA.2021.21.11.785

웹 애플리케이션 기반의 텍스트 데이터 분석 모델

Text Data Analysis Model Based on Web Application

진고환
우송대학교 IT융합학부

Go-Whan Jin(gwjin@wsu.ac.kr)

요약

4차 산업혁명 이후 인공지능, 빅 데이터와 같은 기술들의 발전으로 사회 전반에 다양한 변화가 일어나고 있으며, 핵심적인 기술 적용 과정에서 수집할 수 있는 데이터의 양도 급속하게 증가하고 있는 추세이다. 특히 학계에서는 연구 동향을 파악하기 위하여 기존에 생성된 문헌 데이터에 대한 분석이 이루어지고 있으며, 이러한 문헌 분석은 연구의 흐름을 정리하고, 어떤 연구 방법론이나 주제, 또는 현재 학계에서 화두가 되고 있는 대상에 대한 파악을 통하여 향후 연구 방향 설정에 많은 기여를 하고 있는 상황이다. 그러나 문서 데이터의 분석을 위하여 데이터 수집이 필요하나, 일반적으로 프로그램에 대한 전문 지식이 없는 경우 접근하기 어렵다. 본 논문에서는 텍스트 마이닝 기반의 토픽 모델링 웹 애플리케이션 모델을 제안한다. 제안 모델을 통하여 데이터 분석 기법에 대한 전문적인 지식이 부족하더라도, 연구 논문의 수집, 저장, 텍스트 분석과 같은 다양한 작업을 진행할 수 있으며, 연구자들이 선행 연구 분석과 연구 동향을 파악하기 위하여 데이터 분석에 투입되는 시간 및 노력을 단축시킬 수 있을 것으로 기대된다.

■ 중심어 : | 텍스트 분석 | 텍스트 수집 | 토픽 모델링 | 잠재적 디리클레 할당 | 웹 크롤링 |

Abstract

Since the Fourth Industrial Revolution, various changes have occurred in society as a whole due to advance in technologies such as artificial intelligence and big data. The amount of data that can be collect in the process of applying important technologies tends to increase rapidly. Especially in academia, existing generated literature data is analyzed in order to grasp research trends, and analysis of these literature organizes the research flow and organizes some research methodologies and themes, or by grasping the subjects that are currently being talked about in academia, we are making a lot of contributions to setting the direction of future research. However, it is difficult to access whether data collection is necessary for the analysis of document data without the expertise of ordinary programs. In this paper, propose a text mining-based topic modeling Web application model. Even if you lack specialized knowledge about data analysis methods through the proposed model, you can perform various tasks such as collecting, storing, and text-analyzing research papers, and researchers can analyze previous research and research trends. It is expect that the time and effort required for data analysis can be reduce order to understand.

■ keyword : | Text Analysis | Text Collection | Topic Modeling | LDA((Latent Dirichlet Allocation) | Web Crawling |

접수일자 : 2021년 10월 04일
수정일자 : 2021년 11월 02일
심사완료일 : 2021년 11월 03일
교신저자 : 진고환, e-mail : gwjin@wsu.ac.kr

I. 서 론

4차 산업혁명은 정보통신기술과 다양한 산업분야의 융합을 통하여 사회 전반에 혁신적인 변화를 일으키고 있으며, 기존의 산업혁명과 비교하면 더 빠른 속도와 더 넓은 범위에서 많은 영향을 미치고 있다[1]. 정보통신기술의 핵심인 빅 데이터, 인공지능, 사물인터넷과 같은 기술은 시간적, 공간적인 제약을 없애고 현실세계와 가상세계를 융합하기 위한 4차 산업혁명의 핵심기술이라 할 수 있으며, 기술 적용 과정에서 수집할 수 있는 데이터의 양도 급속하게 증가하고 있는 추세이다[2]. 특히 빅 데이터를 중심으로 의미 있는 데이터를 수집하고 분석하기 위하여 산업체를 포함한 다양한 분야에서 많은 노력을 기울이고 있다. 이러한 변화는 학계의 연구자들 사이에서도 일어나고 있는 상황으로, 학계에서는 연구 동향을 파악하기 위하여 기존에 생성된 문헌에 대한 분석이 이루어지고 있으며, 이러한 문헌 분석은 연구의 흐름을 정리하고 어떤 연구 방법론이나 주제, 또는 현재 학계에서 화두가 되고 있는 대상에 대한 파악을 통하여 향후 연구 방향 설정에 많은 기여를 하고 있는 상황이다. 연구 동향의 파악은 그 시대의 연구 흐름을 파악할 수 있고, 새로운 분야를 발견하여 개척이 가능한 기준점이 될 수도 있다[3]. 최근 문헌 분석의 대표적인 텍스트 분석 기법인 토픽모델링 및 텍스트 마이닝 기법을 통하여 다양한 연구 주제와 방법에 대한 분석이 이루어지고 있다. 토픽 모델링은 비구조화로 이루어진 대량의 문서에서 주요한 토픽을 찾아내기 위한 알고리즘이며, 텍스트 마이닝은 대량의 텍스트를 키워드나 문장으로 요약하여 다양한 분석을 위한 전 단계인 정형 데이터로 구조화시키기 위한 방법론이다. 데이터 분석을 하는 가장 중요한 목적은 텍스트 분석의 개념 및 이벤트를 파악하고, 데이터의 특성을 파악하는 것이라 할 수 있으며, 데이터를 수집하기 위하여 인공지능 모델을 통한 데이터 수집이 선행되어야 한다. 웹에 존재하고 있는 데이터를 수집하기 위해서는, API를 제공하는 서비스를 기반으로 웹 크롤링을 수행하거나, 프로그래밍을 작성하여 수집하여야 한다. 그러나 일반적으로 API를 통하여 데이터를 수집할 경우 대량의 정보와 수집 범위에 한계가 있으며, 웹 크롤링을 통하여 정보를 수집하려면, 프로그램에 대한 전문 지식이 없는 경우에는 접근하기 어려운 단점을 가지고 있고, 간단한 절차를 통하여 논문을 수집할 수 있는 웹 서비스에 대한 연구는 매우 미미한 상황이다. 본 논문에서는 논문 수집의 과정을 자동화하여 웹 애플리케이션을 통한 서비스를 제공할 수 있는 모델을 제안한다. 제안 모델을 통하여 수많은 학계의 연구자들이 선행 연구 분석과 연구 동향을 파악하기 위하여 데이터 분석에 투입되는 노력과 시간을 단축시킬 수 있을 것으로 기대된다.

II. 관련 연구

1. 토픽 모델링

텍스트 분석 측면에서의 토픽 모델은 단어순서의 정보를 무시하고, 단어 모음에 의존하는 것으로서, 각각의 문서는 주어진 말뭉치를 통하여 단어 발생을 포함하는 히스토그램으로 표현될 수 있다. 또한 히스토그램은 특정한 수의 주제에 대하여 분포로 모델링할 수 있고, 각 주제들은 어휘의 단어에 대한 분포라 할 수 있다[4]. 고차원적이고 과다하게 분산되어 있는 단어 표현이 있는 문서를 분석하기 위한 일반적인 방법은 비지도 방식으로 주제를 추출하는 토픽 모델링을 수행하는 것이며, 다른 또 하나의 방법은 레이블과 문서를 공동으로 모델링하는 지도 학습을 수행하는 방법이 있다[5]. 문서를 분석하는 방법은 LSI(Latent Semantic Indexing)[6], LDA(Latent Dirichlet Allocation)[7], multimodal learning[8], document categorization[9], collaborative filtering[10] 같은 다양한 방법이 존재하고 있으며, 토픽 모델은 문서에서 추상적인 주제를 발견하는데 사용되는 통계 모델이라 할 수 있다[11]. 또한 토픽 모델링은 분석하고자 하는 문서들 내에서 주요한 주제가 어떤 것인지 분석하는 기법으로, 단어 사이의 특정한 관계를 군집화 하여 주제를 추출하고 해당하는 문서의 단어들이 어떠한 주제와 같이 나타나는지를 도출하는 것으로, 신속하게 문서의 맥락을 이해할 수 있고, 대량의 문서를 전부 파악하는데 소요되는 시간과 비용을 절약하기 위한 자동화 방식으로[12], LDA 기반의 토픽 모델링은 문서의 주제 파악과 각 주제에 대한 내용을 파악할 수

있는 방식이며[13], 다양한 주제가 혼재되어 있는 텍스트 문서에서 어떤 토픽들이 같이 존재하는지 도출한다[14]. 이 기법은 주로 뉴스 분석이나 고객의 리뷰 및 기업의 사회적 책임에 대한 키워드 트랜드 분석[15]과 같은 비정형적인 데이터의 분석에 주로 활용되고 있으며, 토픽모델링과 머신러닝을 활용하여 이직률을 예측하는 연구가 이루어지고 있다[16].

2. LDA 분석

잠재적 디리클레 할당 모델이라고 하는 LDA는 문서의 토픽 비중과 토픽의 단어 비중이라는 2개 변수의 결합 확률 분포에 의하여 문서가 가지고 있는 토픽을 찾아내는 과정으로, 2개 변수 전부 양의 실수를 요소로 가지고 있으며, 디리클레 분포에 의하여 모든 요소를 더한 값이 1이 된다[17]. LDA를 통하여 가정하고 있는 문서의 생성과정이 합리적일 경우 문서의 토픽 비중과 단어의 비중을 결합한 확률이 가장 크다고 할 수 있고, 대상 문서는 이러한 확률을 가장 크게 만드는 토픽에 할당될 수 있으며, 다음의 (식 1)과같이 표현할 수 있다[18].

$$p(w|d) = \sum_{t=1}^{T} p(w|t)p(t|d) \qquad (1)$$

여기서, d는 문서이고, t는 토픽이며, $p(t|d)$는 문서에서 토픽에 대한 비중이다($\sum tp(t|d) = 1$). $p(w|t)$는 토픽에 등장하는 단어인 w의 비중을 의미하며($\sum p(w|t) = 1$), $p(w|d)$는 두 개의 확률을 결합한 문서에서 어떤 단어가 등장할 것인지에 대한 비중으로서, 높은 비중을 가지고 있는 단어들의 토픽이 해당하는 문서의 토픽으로 할당된다.

LDA를 기반으로 하는 토픽 모델은 자연어 처리, 텍스트 마이닝, 정보 검색, 소셜 미디어 분석 등에 적용하여 하용하고 있다. 소셜 미디어 분석을 위하여 사용하고 있는 토픽 모델링은 온라인 커뮤니티에서 당사자들의 반응과 대화를 이해하는데 많은 도움을 줄 수 있으며[19], 소셜 미디어 웹 사이트에서 공유되고 있는 내용 이외에서 상호 작용에서 유용한 패턴과 이해가 가능한 패턴을 추출하는데 사용되고 있다[20]. 또한 LDA는 분석하고자 하는 문서에서 토픽의 비중과 토픽의 단어 비중인 두 개 변수의 결합 확률 분포를 기반으로 해당 문서의 토픽을 찾아내는 과정으로써, 영상 및 음성 분석, 생물학, 텍스트 마이닝 분야 등 많은 데이터 분석에 활용되고 있다. 이러한 LDA 기법은 각 분야에서 활용하는 방법에 의하여 다양한 알고리즘이 존재하고 있으나, 가장 널리 쓰이고 있으며, 다양한 방식으로 확장되어 사용되고 있다[21]. 그러나 데이터 분석을 하고자 하는 문서의 생성된 주제 중에서 일부는 관련이 없는 토픽을 생성할 가능성도 존재하고 있어[22], 지식 기반의 토픽 모델이 제안되기도 하였다[23].

III. 제안 모델

1. 제안 모델 프로세스

다음의 [그림 1]은 제안 모델의 전체적인 프로세스로서 크게 4단계로 이루어져 있다. 첫째, 데이터 수집 단계로서, 웹 크롤링을 통하여 해당 사이트에서 분석하고자 하는 데이터를 수집한다. 둘째, 데이터 저장 단계로서, 엑셀 파일이나 csv 형태의 파일로 저장한다. 셋째, 탐색적 데이터 분석 단계로서, 논문의 저자, 년도, 발행기관 및 시각화를 진행한다. 넷째, 데이터 분석 단계로서 토픽 모델링을 수행하며, 워드 클라우드를 진행한다.

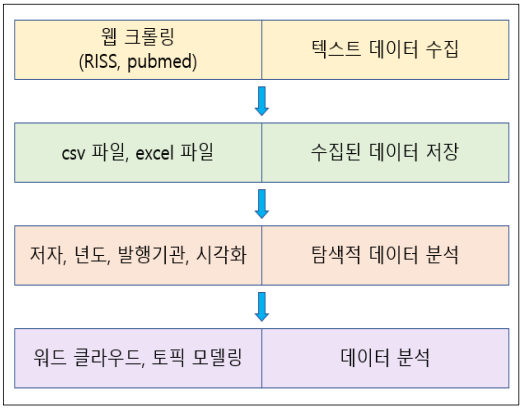

그림 1. 제안 모델 프로세스

2. 제안 모델의 구조

학계의 연구자들은 관심을 가지고 있는 연구 주제에 대하여 수시로 연구 동향을 파악하는 일은 매우 중요하다고 할 수 있다. 따라서 제안 모델은 키워드만 입력하면 연구자들이 원하고자 하는 결과물들을 차례로 확인할 수 있는 구조로 설계하였다. 다음의 [표 1]은 제안 모델의 구조를 나타낸 것이다. 국내 논문과 해외 논문을 분석할 수 있으며, 크게 3가지의 메인 메뉴로 구성하였다. 첫째, 논문 정보를 수집하는 기능으로, 키워드와 수집하고자 하는 논문의 개수를 입력하고, 웹 크롤링을 통하여 결과를 확인할 수 있다. 또한 데이터 탐색과 저자, 발행연도, 빈도 분석 등의 시각화를 제공하는 기능을 보유하고 있으며, 수집된 정보의 파일을 저장하도록 구성하였다. 둘째, 워드 클라우드를 제공할 수 있는 기능으로, 단어의 빈도수를 분석하고, 단어의 트리맵과 워드 클라우드를 생성할 수 있다. 셋째, 토픽 모델링을 생성할 수 있는 기능으로 토픽에 대한 시각화가 가능하다.

표 1. 웹 애플리케이션 결과물 구조

구분	메인메뉴	서브 메뉴	비고
국내논문, (riss) 해외논문 (pubmed)	논문 정보 수집	키워드 입력	분석 대상 키워드 입력
		논문개수 확인	1~1,000개
		크롤링 시작	RISS, pubmed
		결과 확인	DataFrame 형식
		데이터 탐색	칼럼, 데이터 수량 확인
		저자명, 발행기관, 발행연도, 빈도분석	각종 시각화 제공
		파일 저장	csv, excel 선택
	Word Cloud	단어 빈도수 분석	
		단어 트리맵	
		WordCloud 생성	사이드바 : 랜덤값 조정 (모양선택)
	Topic Modeling	토픽모델링	
		시각화	Interactive 방식

3. 데이터 수집

해당 분야의 연구 논문들을 수집하기 위하여 국내 논문뿐만 아니라 해외 논문도 수집할 수 있도록 RISS와 pubmed(https://pubmed.ncbi.nlm.nih.gov/)를 선택할 수 있도록 구성하였으며, 사용자들이 키워드만 입력하면 자동으로 웹 크롤링이 시작되어, 그 결과를 데이터 분석 패키지인 pandas의 DataFrame 형식으로 나타낸다. 제안 모델인 웹 애플리케이션에서 사용자는 인터랙티브한 방식으로 서비스를 이용할 수 있으며, 수집할 논문의 키워드를 입력하고, 수집할 논문의 수도 최대 1,000편까지 선택할 수 있다.

4. 제안 모델 구현

제안 모델을 구현하기 위하여 윈도우10(64bit) 운영체제 하에서 크롬 94.0의 브라우저를 사용하였다. 개발 언어는 Python 3.8, 개발 방식은 BeatifulSoup, Streamlit Library 방식으로 구현하였으며, 개발 환경은 Anaconda/Jupyter Notebook를 기반으로 하였다.

다음의 [그림 2]는 한글로 작성된 논문을 분석하기 위하여 riss에 접속하여 논문 정보를 수집하는 메뉴를 나타낸 것이다.

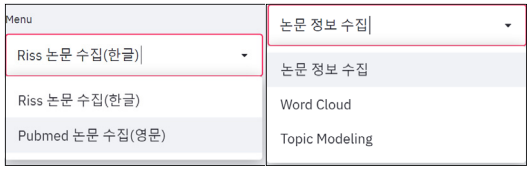

그림 2. 한글 논문 수집 메뉴 및 구조

논문 정보 수집 메뉴에서 키워드를 입력하고, 논문의 검색 수량을 입력한 후 크롤링 버튼을 클릭하여 크롤링을 시작한다. 다음의 [그림 3]은 '인공지능', '교육'과 같은 키워드를 입력하여, Pandas DataFrame으로 크롤링된 결과를 확인할 수 있는 화면이다.

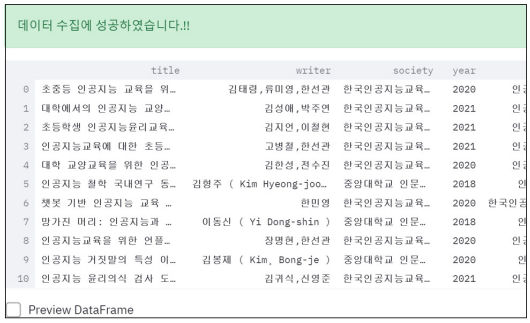

그림 3. 웹 크롤링 결과

다음의 [그림 4]는 영어 논문 데이터 수집을 나타내는 화면이다.

그림 4. 영어 논문 수집

앞의 [그림 3]의 웹 크롤링의 결과에서 데이터 탐색 기능도 보유하고 있다. 데이터 프레임을 통하여 처음 5개 데이터 및 마지막 5개 데이터를 설정하여 확인이 가능하며, 컬럼 이름과 데이터 수를 확인하는 항목을 통하여 화면에 표시할 수 있다. 데이터 분석을 위하여 수행한 정보는 발행 기관과 발행 연도이다. 크롤링하여 수집된 논문들 중에서 발행 기관과 발행 연도별로 논문의 빈도수를 분석하고 이를 시각화 하였다. 다음의 [그림 5]는 시각화한 결과를 나타낸 것이다.

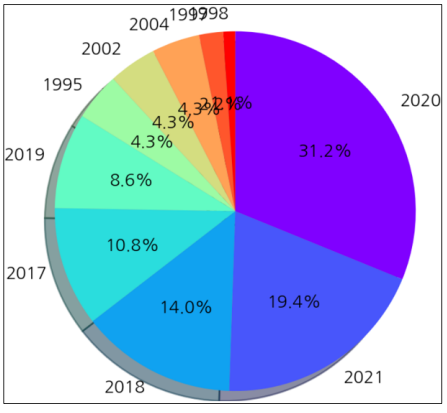

그림 5. 발행 연도별 시각화

시각화 결과를 살펴보면 인공지능이나 인공지능 교육과 관련된 논문들은 4차 산업혁명 이후인 2017년부터 2021년까지 전체 논문들 중에서 84%를 차지한 것으로 나타나고 있어, 인공지능에 대한 관심이 증대됨에 따라 많은 연구가 이루어지고 있음을 확인할 수 있다. 제안 모델은 사용자와의 인터랙티브한 방식으로 구현하였으며, 수집할 논문의 키워드 입력과 수집하고자 하는 논문의 수량까지 선택할 수 있고, 수집이 완료된 파일을 저장하기 위하여 엑셀 파일과, csv 형태의 파일로 저장할 지에 대한 선택이 가능하도록 하였으며, 로컬 PC에 저장된 파일을 업로드 할 수 있는 기능을 보유하고 있다. 저장된 파일을 통하여 워드 클라우드를 생성할 경우 랜덤 값을 사이드 바로 조정하면서 매번 다르게 생성되는 워드 클라우드를 확인하면서 선택할 수 있다. 다음의 [그림 6]은 워드 클라우드를 생성하기 위한 화면이다.

그림 6. 워드 클라우드 생성 선택 화면

수집된 논문들의 텍스트 분석을 위하여 초록과 제목을 저장한 후 한국어 형태소 분석기 패키지인 Konlp를 활용하여 명사만을 추출하고, 명사의 빈도수를 계산하여 워드 클라우드로 시각화 한 것으로, 다음의 [그림 7]은 워드 클라우드를 시행한 결과이다.

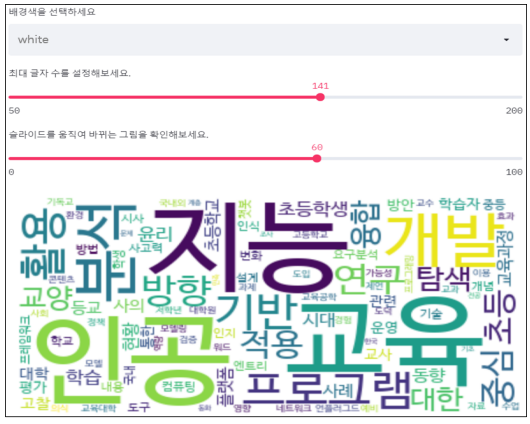

그림 7. 인터랙티브 방식의 워드 클라우드 생성 예

워드 클라우드는 인터랙티브 방식으로 생성한 것으로, 인공지능, 교육과 관련된 결과 화면의 예로서 여러 형태로 조정이 가능하다. 인공지능과 교육으로 검색된 논문들 이므로 대부분의 논문 제목과 초록에 인공지능, 교육이라는 단어가 포함되어 있다.

다음은 토픽 모델링의 대표적인 모델인 LDA를 기반으로 분석을 수행하였다. LDA는 분석하고자 하는 문서들이 혼합으로 토픽이 구성되어 있으며, 토픽들은 확률 분포를 기반으로 단어들을 생성한다고 가정하고, 문서가 생성되는 과정을 역추적하면서 토픽을 생성한다. 다음의 [그림 8]은 IDM(Intertopic Distance Map) 기법을 적용하여 html 파일 형식으로 시각화한 것으로, 이를 통하여 하나의 토픽이 다른 토픽과 어느 정도의 연관이 있는 지를 파악할 수 있고, 토픽들 간의 유사도를 알 수 있다. 전체적으로 5번과 6번 토픽이 겹치는 경우를 제외하면, 다른 토픽들은 서로 겹치는 영역이 거의 없기에 토픽 분류가 잘된 것으로 판단할 수 있다.

그림 8. IDM에 의한 시각화 결과

다음의 [그림 9]는 수집된 논문 중에서 토픽과 가장 관련이 있는 용어에서 상위 30개 용어를 시각화하여 나타낸 것으로 교육이 가장 많았고, 인공지능, 공학, 시스템의 순서로 토픽이 이루어진 것을 확인할 수 있다.

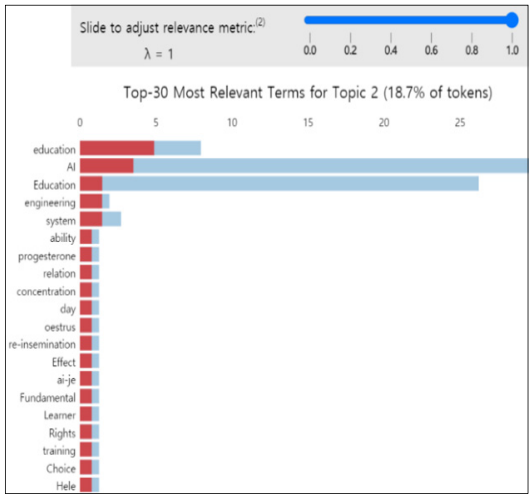

그림 9. 상위 30개 용어 시각화 결과

IV. 제안 모델 평가

제안 모델을 평가하기 위하여 학계 및 연구원의 연구자들과 일반 사용자 80명을 대상으로 30일간의 베타 테스트를 진행하였으며, 진행 사항의 결과를 설문조사를 통하여 만족도 평가를 실시하였다. 베타 테스트 참가자들의 연령은 설문 대상자 연령은 40대에서 60대를 대상으로 진행하였으며, 설문조사 결과를 다음의 [표 2]와 같이 정리하였다.

표 2. 설문 조사 결과

구분	사용자 만족도	제안 모델 효용성	제안 모델 효과성
매우 좋다	24	22	25
좋다	3	2	3
보통	2	3	1
나쁨	1	2	1
아주 나쁨	0	0	0

제안 모델 사용 테스트를 실시한 평가 결과를 종합하면, '좋다', '매우 좋다'라고 응답한 참가자들의 의견이 86.2%로 제안 모델에 대한 사용자 만족은 기대했던 것 이상으로 평가되었으며, 제안 모델을 처음 사용하는 경우 익숙하지 않은 관계로 접근 및 결과에 대한 해석이

쉽지 않았으나, 신속하게 접근하여 사용하였으며, 립모션을 처음 사용하는 경우 익숙하지 않아 접근이 쉽지 않았으나 빠르게 접근하였으며, 논문 분석에서 더욱 다양한 시각화 및 논문의 전체적인 텍스트 요약이 필요하다는 의견이 도출되었다.

V. 결론 및 향후 연구 방향

인공지능과 빅 데이터와 같은 기술의 발전은 시간적, 공간적인 제약을 없애고 현실세계와 가상세계를 융합하기 위한 4차 산업혁명의 핵심기술이라 할 수 있으며, 기술 적용 과정에서 수집할 수 있는 데이터의 양도 급속하게 증가하고 있는 추세이다. 빅 데이터를 중심으로 의미 있는 데이터를 수집하고 분석하기 위하여 산업체를 포함한 다양한 분야에서 많은 노력을 기울이고 있다. 이러한 변화는 학계의 연구자들 사이에서도 일어나고 있는 상황으로, 학계에서는 연구 동향을 파악하기 위하여 기존에 생성된 문헌에 대한 분석이 이루어지고 있으며, 이러한 문헌 분석은 연구의 흐름을 정리하고 어떤 연구 방법론이나 주제, 또는 현재 학계에서 화두가 되고 있는 대상에 대한 파악을 통하여 향후 연구 방향 설정에 많은 기여를 하고 있기에, 텍스트 데이터의 분석은 학술 연구에 매우 중요한 분야 중 하나라 할 수 있다. 본 논문에서는 학술 논문을 자동으로 수집하고, 이를 통하여 텍스트 데이터 분석이 가능한 LDA 기반의 웹 애플리케이션 모델을 제안한다. 제안 모델을 통하여 학계의 많은 연구자들이 연구 동향 파악과 선행 연구 분석에 투입되는 시간과 노력을 단축시킬 수 있어 더 많은 학문적 연구를 할 수 있을 것이며, 애플리케이션의 상용화를 통하여 다양한 토픽모델링의 서비스가 가능할 것으로 판단된다. 제안 모델의 베타 테스트를 실시한 결과 시스템 사용자들의 만족도 및 시스템의 효과성, 효율성에서 '좋음' 이상이 86.2%로 분석되어 기대치 이상으로 나타났으며, 신속하고 쉽게 크롤링 하고 그 결과를 시각적이고 인터랙티브하게 확인할 수 있다는 점에서 높은 관심을 보이고 있어 효용성 가치 또한 크다고 할 수 있다. 그러나 제안 모델은 한글 논문과 영문 논문의 크롤링 사이트를 1개씩만 지정하여, 연구한 것으로 더 많은 논문 사이트의 연결이 필요할 것으로 판단되며, 시스템의 확장성과 안정성을 위해 보다 안정적인 서버 구축과 사용자 편의성을 고려한 디자인 개선 또한 필요한 부분이라 할 수 있다. 향후 연구에서는 데이터 사이언스 교육 프로그램을 운영할 경우 애플리케이션 제작 과정을 교육과정으로 도입할 수 있는 커리큘럼에 대한 연구가 계속되어야 할 것이며, 또한 딥러닝 기술, 특히 자연어처리(NLP) 분야의 발전 속도가 매우 빠르고 상용화 분야도 확장되고 있는 상황으로, 논문의 내용을 자동으로 요약해주는 '텍스트 요약'과 같은 메뉴를 지속적으로 추가하면 사용자들의 만족도가 향상될 것으로 기대된다.

참 고 문 헌

[1] 김상범, 김효관, "4차산업 혁명과 빅데이터 기반 기술," 대한전자공학회지, 제46권, 제11호, pp.17-25, 2019.

[2] 채호근, 이기현, 이주연, "토픽모델링 분석 기법을 활용한 국내외 금융보안 분야 연구동향 분석," 한국산업정보학회논문지, 제26권, 제1호, pp.83-95, 2021.

[3] 김민성, 정형록, 김미옥, 박진하, "한국 관리회계 사례 논문의 연구동향," 한국관리회계학회논문지, 제15권, 제1호, pp.71-112, 2015.

[4] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," International Journal of Advanced Computer Science and Applications (IJACSA), Vol.6, No.1, pp.147-153, 2015.

[5] H. Zhang, B. Chen, Y. Cong, D. Guo, H. Liu, and M. Zhou, "Deep autoencoding topic model with scalable hybrid Bayesian inference," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1-17, 2020.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, Vol.41, No.6, pp.391-407, 1990.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," The Journal of machine

Learning research, Vol.3, pp.993-1022, 2003.
[8] C. Wang, B. Chen, and M. Zhou, "Multimodal Poisson gamma belief network," In Proceedings of the AAAI Conference on Artificial Intelligence, Vol.32, No.1, pp.2492-2499, 2018.
[9] J. Zhu, A. Ahmed, and E. P. Xing, "MedLDA: maximum margin supervised topic models," The Journal of machine Learning research, Vol.13, No.1, pp.2237-2278, 2012.
[10] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.448-456, 2011.
[11] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," The annals of applied statistics, Vol.1, No.1, pp.17-35, 2007.
[12] 권혜진, 전재균, "텍스트 마이닝 기법을 통한 숙박공유서비스 에어비앤비 (Airbnb)에 대한 탐색적 연구," 한국조리학회논문지, 제26권, 제8권, pp.143-153, 2020.
[13] 박영욱, 정규엽, "DMR (Dirichlet Multinomial Regression) 토픽모델링을 이용한 온라인 리뷰 빅데이터 기반 고객감성 분석에 관한 연구: 국내 5 성급 호텔의 외국인 이용객 리뷰를 중심으로," 호텔경영학연구논문지, 제30권, 제2호, pp.1-20, 2021.
[14] 김원식, 김두산, "LDA 토픽모델링을 활용한 호텔등급별 서비스품질 요인도출," 한국무역연구원논문지, 제16권, 제5호, pp.779-791, 2020.
[15] 윤지혜, 이종화, "토픽모델링을 활용한 CSR 키워드 트렌드 분석," 한국인터넷전자상거래학회논문지, 제21권, 제5호, pp.73-91, 2021.
[16] 최진욱, 신동원, 이한준, "IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측: 토픽모델링과 머신러닝을 활용하여," 한국데이터정보과학회지, 제32권, 제5호, pp.1035-1047, 2021.
[17] 문성현, 정세환, 지석호, "Latent Dirichlet Allocation 기법을 활용한 해외건설시장 뉴스기사의 토픽 모델링 (Topic Modeling)," 대한토목학회논문지, 제38권, 제4호, pp.595-599, 2018.
[18] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, "Analyzing entities and topics in news articles using statistical topic models," International Conference on Intelligence and Security Informatics, Springer, Berlin, Heidelberg, pp.93-104, 2006.
[19] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, Vol.78, No.11, pp.15169-15211, 2019.
[20] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," Information fusion, Vol.36, pp.10-25, 2017.
[21] D. M. Blei, "Probabilistic topic models," Communications of the ACM, Vol.55, No.4, pp.77-84, 2012.
[22] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp.262-272, 2011.
[23] Z. Chen, A. Mukherjee, and B. Liu, "Aspect extraction with automated prior knowledge learning," In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol.1, pp.347-358, 2014.

저 자 소 개

진 고 환(Go-Whan Jin) 정회원

- 1990년 2월 : 한국과학기술원 산업공학과(공학석사)
- 1999년 2월 : 한국과학기술원 테크노경영대학원(공학박사)
- 2000년 3월 ~ 현재 : 우송대학교 IT융합학부 교수

〈관심분야〉 : 빅데이터, 토픽모델링, 기술경영, 인공지능