KIPS Tr. Software and Data Eng.
Vol.6, No.11 pp.521~526 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.11.521

Unstructured Data Processing Using Keyword-Based Topic-Oriented Analysis

Myung-Sook Ko†

ABSTRACT

Data format of Big data is diverse and vast, and its generation speed is very fast, requiring new management and analysis methods, not traditional data processing methods. Textual mining techniques can be used to extract useful information from unstructured text written in human language in online documents on social networks. Identifying trends in the message of politics, economy, and culture left behind in social media is a factor in understanding what topics they are interested in. In this study, text mining was performed on online news related to a given keyword using topic - oriented analysis technique. We use Latent Dirichiet Allocation (LDA) to extract information from web documents and analyze which subjects are interested in a given keyword, and which topics are related to which core values are related.

Keywords : Big Data, LDA, Tag Cloud, Correlation Network, Topic Modeling

키워드 기반 주제중심 분석을 이용한 비정형데이터 처리

고 명 숙†

요 약

빅데이터는 데이터 형식이 다양하고 방대할 뿐만 아니라 그 생성 속도가 매우 빨라 기존의 데이터 처리 방식이 아닌 새로운 관리 및 분석 방법이 요구된다. 소셜 네트워크 상의 온라인 문서에서 인간의 언어로 쓰여진 비정형 텍스트에서 Text Mining기법을 사용하여 유용한 정보를 추출할 수 있다. 소셜미디어에 남긴 정치, 경제, 문화에 대한 메시지에 대한 경향을 파악하는 것이 어떤 주제에 관심을 가지고 있는지를 파악할 수 있는 요소가 된다. 본 연구에서는 주제 중심 분석 기법을 이용하여 주어진 키워드에 관한 온라인 뉴스를 대상으로 텍스트 마이닝을 수행하였다. LDA(Latent Dirichiet Allocation)를 이용하여 웹문서로부터 정보를 추출하고 이로부터 사람들이 실제로 주어진 키워드에 대하여 어떤 주제에 관심이 있고 관련된 핵심 가치 중 어떤 주제를 중심으로 전파되고 있는지를 분석하였다.

키워드 : 빅데이터, LDA, 태그 클라우드, 상관관계 네트워크, 토픽 모델링

1. 서 론

빅데이터는 기존의 정형 데이터 뿐만 아니라 웹 로그에 존재하는 메타 데이터와 스키마를 포함하는 반정형 데이터와 텍스트 문서와 이미지, 동영상 또는 음성처럼 구조화되지 않은 비정형 데이터를 포함하는 것이다. 또한 시간이 흐르면서 데이터의 종류도 복잡해지고 다양화되고 있으며 특히, 비정형 데이터의 생성 속도가 매우 빠르게 증가하고 있는 추세이다. 이는 PC와 인터넷 같은 디지털 기기의 사용과 함께, 페이스북이나 트위터 등의 소셜 네트워크 서비스(SNS)가 모바일 폰의 사용과 결합되면서 엄청난 시너지 효과를 내고 있기 때문이다[1]. 또한 스마트 단말 사용자들이 거의 모든 시간대에 SNS로 자신들의 '감정' 데이터를 쏟아내고 있을 뿐만 아니라 스마트폰에 내장된 GPS, 가속센서, 거리센서 등도 시시각각 상황(context) 정보를 양산하는 등 모바일시대가 도래하면서 더 빠르게 진화하고 있다. 빅데이터는 데이터 형식이 다양하고 방대할 뿐만 아니라 그 생성 속도가 매우 빨라 기존의 처리 방식이 아닌 새로운 관리 및 분석 방법이 요구된다. 또한 소셜 빅데이터 분석은 훨씬 방대한 양의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인할 수 있기 때문에 보다 정확히 사회적 문제를 예측하고 현상에 대한 복잡한 연관 관계를 밝혀내며 정책을 개발할 수 있다[2].

소셜 네트워크 상에서 비정형 데이터로부터 정보를 뽑아내고 분석하는 방법은 온라인 문서에서 인간의 언어로 쓰여진 비정형 텍스트에서 자연어처리기술을 이용하여 유용한 정보를 추출하는 Text Mining, 소셜미디어의 문서에 담긴 텍스트 문장을 자연어처리 기술과 감정 분석 기술을 적용하여 사용자의 의견을 분석하는 Opinion Mining, 네트워크 연결구조와 연결 강도를 분석하여 어떤 메시지가 어떤 경로를

† 정 회 원 : 부천대학교 경영과 교수
Manuscript Received : September 18, 2017
Accepted : October 12, 2017
* Corresponding Author : Myung-Sook Ko(kms@bc.ac.kr)

통해 전파되는지를 파악하는 Network Analysis 등이 있다. 또한, 키워드간 상호관계를 예측하는 Data Mining과 시각화 등의 방법으로 빅데이터를 분석할 수 있다[3].

일반적으로 빅데이터는 기존 데이터베이스 관리도구로 데이터를 수집, 저장, 관리, 분석하는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술로 정의하고 있다[3, 4]. 빅데이터는 기존의 데이터와 다른 측면인 데이터의 양(Volume), 데이터 유형과 소스 측면의 다양성(Variety), 데이터 수집과 처리 측면에서 속도(Velocity) 즉, 3V로 요약되며 기존의 작은 데이터 처리 분석으로는 얻을 수 없었던 통찰과 가치를 창출하는 새로운 방식으로 정의될 수 있다[2, 5].

빅데이터는 데이터 형식이 다양하고 방대할 뿐만 아니라 그 생성 속도가 매우 빨라 기존의 데이터 처리 방식이 아닌 새로운 관리 및 분석 방법이 요구되며 소셜미디어에 남긴 정치, 경제, 문화에 대한 메시지가 그 시대의 감성과 정서를 파악할 수 있는 원천으로 등장함에 따라 소셜미디어 상에서 이야기되고 있는 주제들의 경향을 파악하는 것이 중요하며 경향을 파악하는 것이 어떤 주제에 관심을 가지고 있는지를 파악할 수 있는 요소가 된다[6].

경향을 파악하기 위해서는 문서들의 주제를 분류해야 하는데, 본 연구에서는 이를 위하여 텍스트 속에서 주제를 자동으로 찾아주는 LDA(Latent Dirichiet Allocation)를 이용하여 웹문서로부터 주제를 추출하고 이로부터 사람들이 실제로 그 주제에 얼마만큼 관심을 가지고 있는지를 파악해 보고자 한다. 본 연구는 일정 기간 동안에 전체 뉴스 스크래핑을 수행하는 것이 아니라 특정 키워드를 기반으로 뉴스 스크래핑을 수행한 후 정보를 수집하고 그 결과로부터 주어진 키워드와 관련된 정책 또는 방향이 어떻게 진행 및 추진되었는지를 파악해해보고, 그 결과를 기반으로 향후 추진 방향 또는 미흡한 부분을 보완할 수 있는 기틀을 마련하는데 연구의 목적이 있다. 본 연구에서는 '정부3.0'의 효과적인 추진과 생애주기별 맞춤형 서비스 및 국민 행복 실현을 위하여 정부차원에서 빅데이터 추진 방안을 마련하는데 기반이 된 '정부3.0'과 관련하여 웹에 게시된 '정부3.0' 관련 뉴스인 비정형데이터에 대하여 의미있는 정보를 추출하기 위하여 웹스크래핑을 통하여 뉴스정보를 수집하고 텍스트마이닝 기법을 이용하여 데이터 셋을 얻은 후 LDA를 적용하여 정부3.0과 관련하여 사람들이 어떻게 생각하며 인식하고 있는지를 자동 생성된 주제(topic)를 기반으로 파악하고 핵심가치에 대한 내용이 얼마나 반영되었는지 분석하고자 한다.

본 연구에서는 웹뉴스를 기반으로 정부3.0에 대한 사람들의 인식과 관심도를 파악하고 핵심 가치 및 추진목표 등이 얼마나 전파되었는지를 웹스크래핑을 통하여 웹에 게시된 뉴스를 수집하고, LDA 기법을 사용하여 년도별로 정부3.0에 대한 전반적인 관심도 및 흐름을 파악해 보고자 한다. 제 2장에서는 관련 연구에 대해 다루고 제 3장에서는 텍스트 데이터 처리 방법 등 연구 방법에 대해 기술한다. 제 4장에서는 LDA를 적용한 연구 결과에 대해 분석하고 제 5장에서 결론 및 향후 연구 과제에 대하여 기술한다.

2. 관련 연구

2.1 웹 문서의 오피니언 마이닝

오피니언 마이닝(Opinion Mining)은 특정 주제에 대한 사람들의 주관적인 의견을 통계 및 수치화하여 객관적인 정보로 바꾸는 기술로서 문서의 주제보다 사건이나 인물에 대한 사람들의 의견 뿐만 아니라 긍정, 부정 및 중립 등의 감정과 태도로 분석하기 때문에 감성 분석(Sentiment Analysis)라고도 불린다[7].

오피니언 마이닝은 데이터의 수집 및 분석을 위해 자연어 처리기법을 사용하며 감정 어휘나 문장의 통사구조를 분석하여 텍스트로부터의 의견을 추출하고 텍스트 마이닝 기법을 이용하여 의견의 극성을 분류하는 방식으로 기계 학습 기법을 이용하여 수행되며 의견 분석의 대상이 되는 단어는 개별 어휘나 구, 문장 등이다[8]. 즉, 내용을 작성한 사람의 감성을 추출해 내는 텍스트 마이닝(Text Mining)의 한 분류이며 문서의 주제보다 어떠한 감성을 가지고 있는지 판단하는 기술이며 사람들은 자신의 의사표현 도구로 블로그나 트위터와 같은 소셜 미디어를 많이 사용하는 경향이 있기 때문에 SNS상에 올려져 있는 주관적인 의견들은 오피니언 마이닝을 수행하기에 적절한 데이터라고 볼 수 있다.

2.2 웹문서의 토픽 모델링

토픽 모델링에 사용되는 LDA모델은 주어진 문서들이 잠재적으로 어떤 주제들이 존재하는지 추론해 내는 확률 모델로서 Unsupervised Generative Topic Model로도 불린다[9, 10]. Cho and Lee(2015)는 온라인 문서에 대하여 문서에는 나타나지 않지만 문서의 중요한 개념이나 내용을 함축하고 있어 문서 요약 및 정보 검색에 중요한 역할을 수행할 수 있는 잠재 키워드를 추출하기 위하여 주어진 문서와 유사한 문서의 키워드를 후보 키워드로 선택하고 후보 키워드를 구성하는 개별 단어들을 이용해 후보 키워드의 중요도를 평가하는 방법을 제안하였다[10]. Lee et al.(2015)는 사회적 이슈를 찾아내는 주제를 찾아내기 트위터 상의 정보를 클러스터링하기 위하여 LDA알고리즘을 이용하여 토픽 모델링을 수행하였다[11].

Jung et al.(2014)는 SNS인 트위터 토픽을 찾기 위해 토픽 모델링인 LDA모델을 이용하여 토픽들이 어떤 카테고리에 속하는지 파악하기 위하여 토픽 단어에 해당하는 내용들을 분석하고 토픽을 이루는 단어 그룹과 비교한 후 토픽 카테고리를 판단하는 방법을 제안하였다[12]. Jo et al.(2009)는 인터넷 상의 블로그 데이터를 네 가지 유형으로 분류하여 각 유형의 데이터가 가지는 특징과 그로 인해 나타나는 현상들을 살펴본 뒤에 혼잡도 개념을 도입하여 학습된 모델들의 성능을 비교하는 블로그 문서 처리 방법을 제안하였다[13]. 토픽 모델링 기법 중의 하나인 LDA는 단순하다는 특징과 함께 데이터의 차원을 축소하는데 유용하며, 의미적으로 일관성이 있는 주제들을 생산한다는 장점을 가지기 때문에 Text Mining에 유용하게 사용되어 왔다.

3. 연구 방법

본 논문에서는 topic-modeling 방법인 LDA모델을 사용하여 인터넷 상의 여러 개의 topic들이 섞여있는 각 문서(document)에 대하여 단어(word)들의 확률분포로 이루어진 topic들을 분류하였다. 분류된 topic만으로는 키워드의 핵심가치에 대한 분석이 어려우므로 분류된 topic들로부터 주제명을 추출함으로써 제시된 키워드와 관련된 핵심가치에 대한 사람들의 관심도를 분석해 보고자 한다. 즉 document, topic, word의 확률 분포를 이용하여 생성된 각각의 topic기반 주제명들이 키워드인 '정부3.0'의 핵심 가치인 '개방, 공유, 소통, 협력, 정보공개, 공공데이터 개방, 협업, 정보공유, 맞춤형서비스' 등에 얼마나 부합되는지 분석해 보고, 또한 각 주제에 대하여 (단어, 빈도수) 결과를 결합하여 관심도 및 인식 변화를 비교 분석해 보고자 한다.

다음 Fig. 1은 웹에 게시된 특정 키워드(정부3.0) 관련 온라인 뉴스 기사를 웹스크래핑을 통하여 데이터를 수집하고 수집된 데이터에 대하여 텍스트 데이터 처리 단계, LDA 모델링 등의 데이터분석 과정을 보여주는 다이어그램이다.

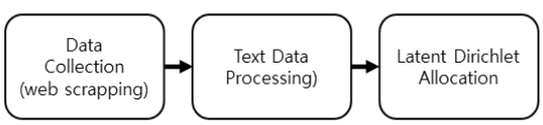

Fig. 1. Data Analysis Process

3.1 데이터 수집

다음 본 연구에서 사용한 데이터는 웹 스크래핑을 통하여 네이버(http://www.naver.com) 및 다음(http://www.daum.net)과 같은 포털싸이트에 '정부3.0' 키워드를 부여하고 개시되어 있는 뉴스를 수집하였다. 정부3.0 개념은 공공기관 중심으로 시작되었기 때문에 일반사람들에게 체감도와 인지도가 비교적 낮고 정부 체제가 바뀌는 시기를 거치면서 어떤 변화를 가져왔는지 살펴보고자 2개 년도(2016년, 2017년)에 대한 주제 중심 분석을 수행하게 되었다. 웹스크래핑 시 21~50개씩의 최신 뉴스를 스크래핑 하였으며, 2016년 6월~2017년 8월 까지 온라인 뉴스 기사 중 '정부3.0'이 등장하는 기사 총 1,785개를 스크래핑 하였다. 포털싸이트를 통하여 뉴스 데이터를 수집함으로써 특정 신문사에 편향되는 경향을 줄이면서 사람들의 관심도를 분석할 수 있다.

3.2 텍스트 데이터 처리

다음 본 연구에 데이터 처리 단계는 다음 Fig. 2와 같다. 모든 수집된 데이터는 웹 페이지 형식의 뉴스 기사이므로 html기호, 신문사 마크 등 불필요한 부분을 제거하여 내용(content)만을 추출하였다. 텍스트 데이터 처리는 아래와 같은 단계를 거치며 관련 프로세스는 아래와 같다.

1) 웹상의 뉴스를 스크래핑한다.
2) 공백을 구분자로 하여 내용을 추출한다.
3) 수집된 단어들로부터 명사만을 추출한다.
4) 뉴스별 명사별 출현 빈도로 이루어진 Term-Document Matrix를 만든다.
5) 한 글자로 이루어진 명사를 제거한다.
6) 명사-뉴스 간 matrix를 구성한다.
7) 단어-문서 출현빈도로 구성된 matrix를 구성한다.
8) 단어간 상관관계 계수 기반 matrix를 구성한다.
9) Latent Dirichlet Allocation 모델을 적용하여 뉴스의 topic을 구하고 각 topic에 속하는 단어(word)들에 대하여 (단어, 확률) 쌍으로 각 topic을 완성한다.

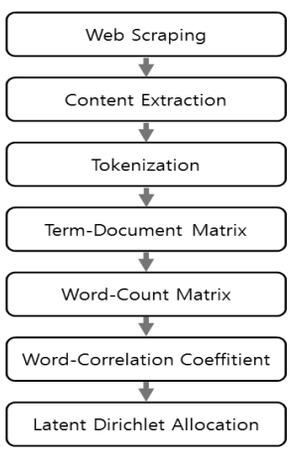

Fig. 2. Text Data Processing

3.3 데이터 분석

다음 본 연구에 사용된 LDA(Latent dirichlet Allocation) 모델은 unsupervised learning을 수행하는 확률그래프 모델로서 Dirichlet 분포를 이용하여 어떤 주제에 대해 단어들이 포함될 확률을 모델링하는 방법이다. 따라서 각 문서가 어떤 주제를 가지는지 알지 못하고 유일하게 관찰된 데이터인 단어들의 패턴만으로 학습이 이루어지는 방식이다[14].

LDA모델은 다음 Fig. 3과 같이 확률적 그래픽 모델로 표현된다. 파라미터 α와 β는 코퍼스(corpus) 레벨의 파라미터이며, α는 주제들이 어떤 패턴인지 나타내며, β는 단어들이 어떤 패턴인지 나타내는 파라미터이다. 변수 Θ는 문서 레벨의 변수로서 문서 하나에 대한 주제어 분포를 의미한다. z와 ω는 단어(word) 레벨의 변수이며 z는 해당 주제에 대한 단어 확률, ω는 해당 주제에 대해 실제 관찰된 단어를 나타낸다. LDA 모델은 사정 정보가 없는 상태에서 학습을 수행

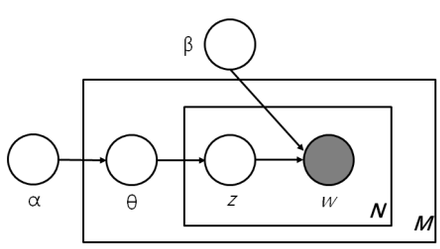

Fig. 3. Graphical Model of LDA

하므로 수집된 데이터의 단어들의 패턴만으로 학습한다. ω가 단어를 가리키고 z가 주제를 가리키며 단어ω가 어떤 주제 z를 포함할 확률을 찾는다[9, 10].

LDA의 결과로 만들어진 주제 중심 모델은 각 topic들에 단어가 분류되는데 분류 기준이 알려져 있지 않기 때문에 구성된 모델을 보고 사람이 직접 주제를 부여하는 방식이다[10]. 즉, 정부3.0 관련 주제어를 부여하기 위해서는 관련된 핵심 가치 등에 대한 지식을 필요로 한다. 즉, topic을 구성하는 단어들을 보고 주어진 키워드의 핵심가치와 단어들의 의미적 연관성을 고려하여 topic 이름을 부여하는 것이다.

4. 연구 결과

4.1 연구 결과

국내 포털싸이트에 게시된 뉴스기사에 대한 웹 스크래핑을 이용하여 '정부3.0' 관련 키워드를 (단어, 빈도수)를 기반으로 연도별로 상위 50개씩을 추출하고 그 결과에 대하여 추출된 단어들에 대한 중요도를 직관적으로 파악할 수 있도록 상관관계 네트워크 및 태그 클라우드 기법을 적용하였다.

2016년도에 수집된 단어는 스크래핑한 단어들을 출현빈도수로 내림차순 정렬하면 각각 '예산, 요구, 억원, 정부, 부처, 올해, 내년, 증가, 분야, 스마트, 주택, 지출, 가장, 규모, 복지, 산업, 조원' 등이며, 2017년도에 출현한 단어는 출현빈도수 값으로 내림차순 정렬하면, '정부, 답변, 목록, 추천, 선택, 질문, 국민, 금융, 국민신문고, 사기, 위원회, 제공, 방송통신, 댓글, 배달, 서비스, 정보' 등과 같다. 2개 년도에서 TOP20에 공통적으로 등장한 단어는 '정부, 국민, 서비스, 정보'이다.

본 절에서는 생성된 각 주제들이 정부3.0의 핵심 가치 및 추진목표인 '개방, 공유, 소통, 협력, 정보공개, 공공데이터 개방, 협업, 정보공유, 맞춤형서비스, 국민중심, 정보보안' 등에 얼마나 부합되는지 분석해 보고, 또한 각 주제에 대하여 (단어,빈도수) 결과를 결합하여 최종 관심도 및 흐름 파악 결과도 비교 분석해 보고자 한다.

4.2 태그 클라우드와 상관관계 네트워크

태그 클라우드(tag cloud)는 워드 클라우드(word cloud)라고도하며 메타 데이터에서 명사 추출을 통하여 얻은 태그들을 분석하여 중요도나 인기도 등을 고려하여 핵심 개념들을 직관적으로 파악할 수 있도록 시각적으로 표현하는 기법이다[4]. 시각적 중요도를 강조하기 위해 각 태그들은 그 중요도에 따라 글자의 색상이나 굵기 등 형태가 변하며 또한 단어가 출현한 빈도수에 따라 글자 크기가 다르게 나타나므로 태그 클라우드를 통해 직관적인 파악이 가능하다.

태그 클라우드는 두 가지 의미를 지니는데, 첫째로 하나의 연결된 태그들이 얼마나 많으며 어떤 종류인지를 보여주는 것인데, 어떤 내용에 여러 사용자들에 의해 어떤 태그와 연결되어 있는지를 보여줄 수 있다. 두 번째는 각 태그들이 얼마나 인기도가 높은지를 보여주는 표시법으로 사용되는 경우이다. 이때 태그들의 글자 크기나 색상, 형태들이 인기도에 따라 변화되며 인기도는 사용자의 선택에 의해 자동적으로 갱신된다[4]. Correlation Network는 수집된 명사들 간의 상관관계 계수에 의해 만들어지는 네트워크 표현이다. word-word Correlation Matrix에서 weight값에 따라 연결성 및 강도(길이)가 다름을 알 수 있다.

추출된 명사들에 대한 상관관계 네트워크는 (답변, 국민, 방향, 선택, 정책, 댓글, 목록), (국민신문고, 위원장, 관련, 평가, 서비스, 맞춤, 선정, 분야) 등의 단어들이 밀접한 관계를 형성함을 볼 수 있으며 다음 Fig. 4와 같다. 또한, 2016년도 수집된 계획 중심의 추상적인 단어(예산, 요구, 억원, 정부, 내년, 증가, 규모)들에 비해서 2017년도에는 구체적인 단어(목록, 답변, 위원회, 댓글, 기관) 등으로 정부3.0에 대한 사람들의 관심도가 구체적으로 변해감을 알 수 있다.

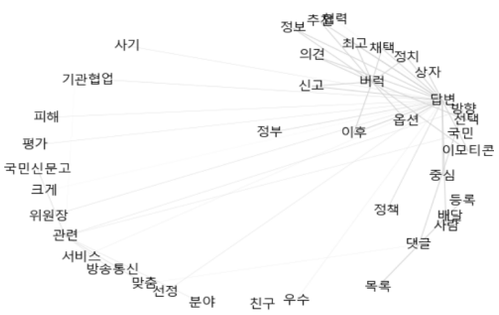

Fig. 4. Correlation Network by Extracted Nouns

아래 Fig. 5는 수집된 단어들로부터 출현빈도수를 기반으로 만들어진 태그 클라우드이다. 결과를 보면 '정부', '답변', '선택', '추천', '질문', '위원회' 등의 단어가 중요도가 비교적 높게 나옴을 알 수 있다. 태그 클라우드 결과로부터 2016년도에는 계획 및 준비하는 단계와 관련된 단어(예산, 요구, 억원 등)들이 많이 출현함을 볼 수 있다. 반면에 2017년도에는 실제적인 일을 추진하는 내용과 관련된 단어(정부, 답변, 목록 등)의 빈도수가 높았고 전체 결과를 보면 정부와 국민 관련 중요도가 비교적 높게 나타났음을 알 수 있다.

Fig. 5. Tag Cloud by Extracted Nouns

4.3 LDA 모델링 결과

LDA를 적용하여 각 topic별로 구성 word 벡터는 topic4 = (답변, 정부, 추천, 선택, 목록, 국민, 정보, 등록, 댓글, 채택), topic30 = (추천, 답변, 목록, 선택, 정부, 국민, 등록, 정보, 댓글, 신고), topic80=(금융, 사기, 피해, 정부, 위원회, 국민신문고, 협력, 목록, 위원장, 답변), topic91=(금융, 사기, 피해, 정

부, 위원회, 협력, 제공, 질문, 차단, 기관)이다. topic을 구성하는 단어의 교차 출현도를 살펴보면 (topic4, topic30)과 (topic80, topic91)이 주제를 구성하는 단어들의 분포가 비슷함을 볼 수 있다.

각 topic은 벡터로 표현되므로 두 topic간의 유사도를 수학적으로 측정하기 위해 두 주제를 이루는 단어들의 확률 분포 벡터의 코사인 유사도(Cosine Similarity)를 구하면 된다. topic을 구성하는 단어의 분포가 비슷한 (topic4, topic30)의 코사인 유사도는 0.0445이며 (topic80, topic91)의 코사인 유사도는 0.071임을 알 수 있다. 코사인 유사도 값은 0~1사이에 분포하며 1에 가까울수록 주제가 유사하다고 볼 수 있으므로 (topic4, topic30)에 대해서는 두 topic을 구성하는 단어에 대한 교차 출현 빈도가 높지만 코사인 유사도는 낮음을 알 수 있으며 (topic80, topic91)은 두 topic을 구성하는 단어 분포가 (topic4, topic30)보다 두 topic을 구성하는 단어에 대한 교차 출현 빈도가 낮지만 코사인 유사도는 (topic4, topic30)에 비해 높았다. 즉, 단어에 대한 교차 출현 빈도가 높다고 반드시 코사인유사도가 비례하지는 않는다는 것을 알 수 있다.

LDA를 적용한 결과 스크래핑한 웹 뉴스들로부터 99개의 주제를 추출하였고 각 주제별 별로 확률이 가장 높은 10개의 단어들을 (단어, 확률)의 쌍으로 나타낸 topic중 일부분은 다음 Table 1, Table 2와 같다. topic들 중에서 topic을 구성하는 단어의 확률의 합이 가장 높게 나온 주제는 topic2이며, topic2를 구성하는 단어는 (정보, 이후, 정치, 정책, 정부, 중심, 위원회, 의견, 차단, 이모티콘)이다. 또한 topic30, topic 96, topic2, topic95에서 보면 '목록'이라는 단어가 교차해서 나타나고 있으며, topic30과 topic2에서는 '국민'이라는 단어가 교차해서 나타남을 볼 수 있다. topic30, topic2, topic95에서는 '정부'라는 단어가 여러 topic에 높은 확률로 나타나는 것을 볼 수 있는데, 아는 하나의 단어가 다른 여러 topic에서 큰 의미를 부여한다는 것을 알 수 있다.

정부3.0의 핵심 가치 및 추진목표인 '개방, 공유, 소통, 협력, 정보공개, 공공데이터 개방, 협업, 정보공유, 맞춤형서비스, 국민중심, 정보보안' 등에 얼마나 부합되는지 분석해 보기 위하여, 정부3.0 관련 뉴스를 대상으로 주제 중심 모델링을 하여 각각 10개의 단어로 이루어진 주제들을 추출하였다. 주제 구성단어들은 주제 모델링의 수행 결과로 나온 것이며 주제명은 정부3.0의 핵심가치를 중심으로 직접 부여하였으며 결과는 다음 Table 3과 같다.

정부3.0의 핵심 가치 및 추진목표인 '개방, 공유, 소통, 협력, 정보공개, 공공데이터 개방, 협업, 정보공유, 맞춤형서비스, 국민중심, 정보보안' 등과 관련하여 주제 모델링 수행 결과 정부3.0의 핵심가치와 관련하여 topic을 구성하는 단어들의 의미적 관련성 등을 고려하였을 때 '소통, 맞춤형서비스, 개방, 정보공유 및 국민 중심'이라는 주제명을 얻었다. 주제를 구성하는 단어에도 '정부'라는 단어가 교차해서 나타나는 것을 볼 수 있으며, Table 3의 결과로부터 키워드 관련 핵심 가치에 대한 전파가 '정보보안' 측면에서는 추진 및 관심도 확대 부분에 대한 보완이 필요함을 알 수 있다.

Table 1. Results of Topic30 and Topic96

Topic 30		Topic 96	
Word	Probability	Word	Probability
Recommendation	0.0986	Finance	0.1222
Answer	0.0960	Damage	0.0782
List	0.0770	Scam	0.0759
Select	0.0731	Government	0.0551
Government	0.0723	Cooperation	0.0439
Nation	0.0406	Committee	0.0436
Enrollment	0.0387	Agency	0.0374
Information	0.0359	National sinmungo	0.0337
Comment	0.0305	List	0.0326
Declaration	0.0285	Offer	0.0317

Table 2. Results of Topic2 and Topic95

Topic 2		Topic 95	
Word	Probability	Word	Probability
Broadcasting communication	0.0739	Government	0.0263
Committee	0.6959	Answer	0.0225
Government	0.0623	List	0.0222
Service	0.0511	Offer	0.0222
List	0.0465	Question	0.0219
Question	0.0446	Recommendation	0.0217
Excellence	0.0396	National sinmungo	0.0216
Evaluation	0.0378	Delivery	0.0213
Agency	0.0366	Service	0.0212
Nation	0.0359	Friend	0.0211

Table 3. Subject Results by topic modeling

Subject Name	Composition Word
Communication	Answer, Government, Recommendation, Select, List, Nation, Information, Enrollment, Comment, Selection
Customized service	Government, Broadcasting communication, Selection, Agency, Evaluation, Committee, National sinmungo, Answer, Service, Nation
Release	Government, Answer, List, Offer, Question, Recommendation, National sinmungo, Delivery, Service, Friend
Information-sharing	Broadcasting communication, Committee, Government, Service, List, Question, Excellence, Evaluation, Agency, Nation
Nation-oriented	Government, Service, List, Question, Answer, Recommendation, National sinmungo, Offer, Delivery, Relation

5. 결 론

본 연구에서는 특정 키워드를 기반으로 뉴스 스크래핑을 수행한 후 문서를 수집하고 추출된 정보로부터 주어진 키워드와 관련된 정책 또는 방향이 어떻게 진행 및 추진되었는지를 파악해해보고, 그 결과를 기반으로 향후 추진 방향 또는 미흡한 부분을 보완할 수 있는 기틀을 마련하기 위해 키워드 기반 주제 중심 분석을 수행 하였다. 웹스크래핑에 사용된 키워드는 '정부3.0'이며, 웹뉴스를 기반으로 정부3.0에 대한 사람들의 인식과 관심도를 파악하고 핵심 가치 및 추진목표 등이 얼마나 전파되었는지를 파악해 보기 위해 웹에 게시된 뉴스를 수집하고, LDA 기법을 사용하여 년도별로 정부3.0에 대한 주제 중심 분석을 수행하였다. 또한, 웹뉴스를 구성하는 단어들에 대한 출현 빈도수를 기반으로 태그 클라우드 및 상관관계도 네트워크를 통하여 핵심 단어들을 파악해 보았다. 주제 중심 모델링을 통하여 각각 10개의 단어로 이루어진 topic들을 추출하였다. topic은 벡터 형식이기 때문에 topic을 구성하는 단어에 대한 교차 출현 빈도가 높은 topic들이 반드시 코사인 유사도가 비례하여 높지는 않다는 결과를 얻을 수 있었다.

추출된 태그 클라우드 결과로부터 2016년도에는 계획 및 준비하는 단계와 관련된 단어들이 많이 출현함을 볼 수 있다. 반면에 2017년도에는 실제적인 일을 추진하는 내용과 관련된 단어의 빈도수가 높았고 전체 결과를 보면 정부와 국민 관련 중요도가 비교적 높게 나타났음을 알 수 있었다. topic을 구성하는 단어들은 주제 모델링의 수행 결과로 나온 것이며 '정부3.0'의 핵심가치와 관련하여 '소통, 맞춤형서비스, 개방, 정보공유 및 국민 중심'이라는 핵심가치와 연관된 주제명을 얻을 수 있었다. 이로부터 빅데이터와 관련된 개방, 정보공유 등의 주제 뿐만 아니라 앞으로는 정보 공개에 따른 '정보보안(데이터 침해, 개인정보보호)' 등과 관련하여서도 많은 이슈들이 제시되어야하고 제도적 관심이 요구된다고 할 수 있겠다.

소셜네트워크 상의 데이터를 처리함에 있어서 블로그나, 트위터 등은 사람들 개개인의 생각들이 많이 담겨져 있는 반면에 뉴스 기사와 같은 웹문서의 경우는 정해진 주제에 대하여 관련 내용들이 어떤 방향으로 다루어졌는지 파악해 볼 수 있는 좋은 소스로 사용될 수 있다. 또한 주어진 키워드에 대한 목표(핵심 가치 실현)가 있으므로 핵심가치에 대한 topic 구성 단어로 유도된 주제를 분석함으로써 방향 설정을 변경하거나 또는 부족한 부분을 파악하고 보강하는 등의 조치를 취할 수 있는 근거를 마련해 줌으로써 분석 결과가 목표에 더 가깝게 도달할 수 있도록 방향키 역할을 할 수 있을 것으로 판단된다. 향후 연구 방향은 키워드에 대한 긍정, 부정의 가설을 귀무가설과 대립가설의 형태로 세운 후 웹스크래핑 결과를 통하여 가설의 기각 또는 채택 과정을 통하여 주어진 특정 키워드에 대한 인식을 긍정적으로 또는 부정적으로 평가할 수 있도록 하는 실험 및 결과에 대하여 정량적으로 평가할 수 있는 방법에 대한 연구 및 적용이다.

References

[1] J. P. Woo, "Big Data Analysis will ahead," Maekyung Pub. pp.236-251, 2017.

[2] "The Guide for Advanced Data Analytics Professional," Korea Database Agency, 2014.

[3] T. M. Song and J. Y. Song, "Social Big Data Research Methodology with R," Hannarae Pub., ch. 1, pp.16-39, 2016.

[4] Tag Cloud [Internet], https://ko.wikipedia.org/wiki/

[5] K. T. Kim, J. G. Ahn, and D. H. Kim, "Big Data Weapering 1," AgePerson Pub., ch. 1, pp.4-12, 2017.

[6] Y. D. Yun, J. H. Jo, and H. S. Lim, "A Comparative Analysis of Cognitive Change about Big Data Using Social Media Data Analysis," in *KIPS Tr. Software and Data Eng.*, Vol.6, No.7, pp.371-378, 2017.

[7] Opinion Mining [Internet], https://www.facebook.com/kubigdata/posts/504333396427600

[8] B. I. Kang, M. Song, and W. S. Jho, "A Study on Opinion Mining of Newspaper Texts based on Topic Modeling," *Journal of Korean Society for Library and Information Science*, Vol.47, No.4, pp.315-334, 2013.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3(Jan.), pp.993-1022, 2003.

[10] Taemin Cho and Jee-Hyong Lee, "Latent Keyphrase Extraction using LDA Model," *Journal of Korean Institute of Intelligent Systems*, Vol.25, No.2, pp.180-185, 2015.

[11] R. D. Lee, J. M. Kim, and J. S. Lee, "Analysis of twitter topic using LDA," *Journal of Korean Institute of Intelligent Systems*, Vol.25, No.2, pp.180-185, 2015.

[12] B. M. Jeong, T. H. Kim, J. Lee, and J. S. Kim, "Twitter Topic Extraction and Topic Category Decision using LDA Model," *Proceedings of KISSE Winter Conference*, pp.787-788, Dec., 2014.

[13] Yohan Jo, Dongwoo Kim, Il-Chul Moon, and Haeyun Oh [Internet], http://seslab.kaist.ac.kr/xe2/?module=file&act=procFileDownload&file_srl=5591, 2009.

[14] S. Y. Bong and K. B. Hwang, "Applying Labeled LDA to Author Keywirds Recommendation," *Proceedings of KIISE Spring Conference*, Vol.37, No.1(C), pp.385-389, 2010.

고 명 숙

e-mail : kms@bc.ac.kr

1989년 이화여자대학교(이학사)

1993년 고려대학교 컴퓨터학과(이학석사)

1998년 고려대학교 컴퓨터학과(이학박사)

2001년~현재 부천대학교 경영과 교수

관심분야 : 유전자알고리즘, 빅데이터경영