https://doi.org/10.7840/kics.2017.42.2.471

텍스트 분석 기술 및 활용 동향

김 남 규◆°, 이 동 훈*, 최 호 창*, William Xiu Shun Wong*

Investigations on Techniques and Applications of Text Analytics

Namgyu Kim◆°, Donghoon Lee*, Hochang Choi*, William Xiu Shun Wong*

요 약

최근 데이터의 양 자체가 해결해야 할 문제의 일부분이 되는 빅데이터(Big Data) 분석에 대한 수요와 관심이 급증하고 있다. 빅데이터는 기존의 정형 데이터 뿐 아니라 이미지, 동영상, 로그 등 다양한 형태의 비정형 데이터 또한 포함하는 개념으로 사용되고 있으며, 다양한 유형의 데이터 중 특히 정보의 표현 및 전달을 위한 대표적 수단인 텍스트(Text) 분석에 대한 연구가 활발하게 이루어지고 있다. 텍스트 분석은 일반적으로 문서 수집, 파싱(Parsing) 및 필터링(Filtering), 구조화, 빈도 분석 및 유사도 분석의 순서로 수행되며, 분석의 결과는 워드 클라우드(Word Cloud), 워드 네트워크(Word Network), 토픽 모델링(Topic Modeling), 문서 분류, 감성 분석 등의 형태로 나타나게 된다. 특히 최근 다양한 소셜미디어(Social Media)를 통해 급증하고 있는 텍스트 데이터로부터 주요 토픽을 파악하기 위한 수요가 증가함에 따라, 방대한 양의 비정형 텍스트 문서로부터 주요 토픽을 추출하고 각 토픽별 해당 문서를 묶어서 제공하는 토픽 모델링에 대한 연구 및 적용 사례가 다양한 분야에서 생성되고 있다. 이에 본 논문에서는 텍스트 분석 관련 주요 기술 및 연구 동향을 살펴보고, 토픽 모델링을 활용하여 다양한 분야의 문제를 해결한 연구 사례를 소개한다.

Key Words : Big Data, Data Mining, Text Analytics, Topic Modeling

ABSTRACT

The demand and interest in big data analytics are increasing rapidly. The concepts around big data include not only existing structured data, but also various kinds of unstructured data such as text, images, videos, and logs. Among the various types of unstructured data, text data have gained particular attention because it is the most representative method to describe and deliver information. Text analysis is generally performed in the following order: document collection, parsing and filtering, structuring, frequency analysis, and similarity analysis. The results of the analysis can be displayed through word cloud, word network, topic modeling, document classification, and semantic analysis. Notably, there is an increasing demand to identify trending topics from the rapidly increasing text data generated through various social media. Thus, research on and applications of topic modeling have been actively carried out in various fields since topic modeling is able to extract the core topics from a huge amount of unstructured text documents and provide the document groups for each different topic. In this paper, we review the major techniques and research trends of text analysis. Further, we also introduce some cases of applications that solve the problems in various fields by using topic modeling.

◆° First and Corresponding Author : Kookmin University School of MIS, ngkim@kookmin.ac.kr, 정회원
* Kookmin University The Graduate School of Business Information Technology, donghoonlee@kookmin.ac.kr, choi3684@kookmin.ac.kr, williamwong@kookmin.ac.kr
논문번호 : KICS2017-01-008, Received January 9, 2017; Revised February 7, 2017; Accepted February 16, 2017

Ⅰ. 서 론

최근 다양한 분야에서 빅데이터(Big Data), 빅데이터 분석, 그리고 빅데이터 활용에 대한 관심이 학계와 업계를 불문하고 급증하고 있다. 빅데이터는 데이터의 양이 너무 방대해서 기존의 방법이나 도구로는 수집, 저장, 검색, 분석, 그리고 시각화가 어려운 데이터를 의미하며, 빅데이터 분석의 경우 데이터의 양 자체가 해결해야 할 문제의 일부분이 된다. 빅데이터의 특징으로는 주로 데이터의 양(Volume), 데이터 입출력의 속도(Velocity), 그리고 데이터 종류의 다양성(Variety)을 나타내는 3V의 개념이 소개되고 있다. 이 가운데 속도는 단지 데이터의 생성 속도만을 의미하는 것은 아니며, 처리 및 분석 속도까지 포함하는 개념으로 이해하는 것이 바람직하다. 또한 다양성은 기존의 주요 분석 대상이던 정형 데이터뿐 아니라 이미지, 동영상, 각종 로그(Log) 및 센싱 데이터(Sensing Data) 등 다양한 형태의 비정형 데이터까지도 분석의 대상으로 포함함을 의미한다. 즉 기존에 비해 훨씬 다양한 형태의 데이터가 기존에 비해 훨씬 빠른 속도로 생성, 관리, 분석되기 때문에, 우리가 다루어야 할 데이터의 양이 급격하게 증가한 것으로 이해할 수 있다.

이러한 다양한 형태의 데이터 중 특히 비정형 텍스트 데이터에 대한 분석 수요가 사회 각 분야에서 급증하고 있다. 이처럼 텍스트 분석에 대한 수요가 증가하는 원인은 다음과 같은 측면에서 찾을 수 있다. 우선 텍스트는 인류가 정보를 표현하고 전달하는 데에 사용해 온 가장 대표적인 수단이므로, 상당히 많은 양의 정보가 이미 텍스트로 기록되어 있으며 향후에도 많은 양의 텍스트 기록물이 생성될 것이다. 또한 최근 다양한 소셜 미디어를 통해 유통되는 비정형 데이터의 양이 급증하고 있으므로, 이에 대한 분석을 통해 특정 제품 및 서비스 사용자의 인식뿐 아니라 일반 국민의 여론을 파악하기 위한 시도는 매우 의미 있다고 할 수 있다. 또한 텍스트 데이터의 경우 기존의 정형 데이터에 비해 상대적으로 대량 수집이 수월하다는 점도 텍스트 분석 수요 증가의 원인으로 이해될 수 있다. 정형 데이터의 경우 개별 데이터들이 민감한 정보를 포함하고 있기 때문에, 이들 데이터를 대량으로 수집하고 분석하여 의미 있는 결과를 도출하기가 쉽지 않았다. 예를 들면 금융 상품 대출 고객의 신상 정보, 온라인 쇼핑몰의 특정 고객의 구매 정보 등을 해당 기업의 업무와 관계되지 않은 목적으로 사용하는 것은 엄격히 제한되어 있다. 하지만 텍스트 분석의 주요 대상인 뉴스 기사, 블로그, SNS 게시물, 리뷰, 각종 댓글 등은 개별 데이터의 기밀성(Confidentiality)이 비교적 낮을 뿐 아니라 크롤링(Crawling)을 통해 비교적 수월하게 대량의 데이터를 수집할 수 있다는 특징이 있다.

요약하면, 분석 가능한 텍스트 데이터의 양이 증가함에 따라 텍스트 분석을 통해 새로운 지식을 창출하고자 하는 수요가 증가하고, 이러한 수요 증가로 인해 텍스트 분석에 대한 관심과 연구가 활발해지는 선순환 구조가 형성된 것으로 파악할 수 있다. 이로 인해 텍스트 분석과 관련된 기술 및 방법론을 연구하는 분야가 급격히 성장하고 있으며, 이러한 분야는 기존의 데이터 마이닝(Data Mining)과 관련되어 텍스트 마이닝(Text Mining)이라는 명칭의 새로운 분야로 자리매김하고 있다. 텍스트 마이닝은 기존의 데이터 마이닝과는 다른 새로운 분야로 인식되기도 하지만, 넓은 의미의 데이터 마이닝의 한 부분으로 이해하는 것이 일반적이다. 따라서 기존 데이터 마이닝에서 사용되던 빈도 분석, 군집화, 그리고 분류 등의 주요 기술이 텍스트 마이닝에서도 매우 중요하게 사용되고 있으며, 대부분의 상용 데이터 마이닝 도구의 최근 버전에서 텍스트 마이닝 기능을 제공하고 있다.

텍스트 분석은 일반적으로 문서 수집, 파싱(Parsing) 및 필터링(Filtering), 구조화, 빈도 분석 및 유사도 분석의 순서로 수행되며, 텍스트 분석 관련 기술은 그림 1과 같이 크게 비정형 텍스트의 구조화까지의 단계와 구조화된 문서의 분석 및 활용의 두 단계로 구분하여 살펴볼 수 있다. 일부 연구에서는 논문, 특허, 연구보고서 등 분석 대상인 텍스트 문서가 데이터베이스 형태로 제공되기도 하지만, 대부분의 연구에서 텍스트 문서는 크롤링을 통해 직접 수집하는 경우가 많다. 크롤링을 통해 수집 가능한 텍스트는 인터넷 뉴스 기사, 블로그 게시글, 상품 리뷰, 댓글, SNS 메시지 등 매우 다양하며, 여러 관점의 분석을 수행하거나 검증하기 위해서는 텍스트 뿐 아니라 이와 연결된 정형 데이터(등록일, 평점, 조회수, 사용자 ID 등)도 함께 수집하는 것이 바람직하다.

이렇게 수집된 텍스트는 각 문서를 용어의 출현 빈도에 따라 벡터화한 벡터공간모델(Vector Space Model)로 구조화된다. 이 때 각 셀은 각 용어가 해당 문서에서 출현한 빈도를 나타낼 수도 있으며, 각 용어의 출현 여부를 '0'과 '1'로 나타낼 수도 있다. 또한 미출현 용어를 '0', 한 번 출현한 용어를 '1', 그리고 두 번 이상 출현한 용어를 '2'로 표기하는 방법도 사용될 수 있다. 하지만 학계 및 실무에서 문서의 용어 벡터화에 가장 널리 사용되는 방법은 TF-IDF(Term Frequency - Inverse Document Frequency)에 기반

을 둔 방법이다. 이는 각 용어의 각 문서에 대한 연관성을 TF와 IDF의 곱으로 나타내는 방식으로, TF는 각 용어가 각 문서에서 나타난 빈도를 근거로 도출되며 IDF는 해당 용어를 포함하고 있는 문서 수에 대한 전체 문서 수의 비율을 근거로 도출된다. 즉 임의의 용어 t가 다른 문서에서는 거의 출현하지 않고 문서 d에서 자주 출현한다면 용어 t의 문서 d에 대한 tf-idf 값은 매우 높게 나타나며, 이는 용어 t가 문서 d의 특징을 나타내는 매우 중요한 용어임을 의미하는 것으로 해석된다.

이러한 방식으로 비정형 텍스트 문서를 벡터의 형태로 구조화할 수 있지만, 벡터의 차원과 직결되는 용어의 수가 지나치게 많기 때문에 위 개념을 그대로 적용하는 것은 현실적으로 무리가 있다. 따라서 많은 수의 특질(Feature)을 처리 가능한 수의 차원으로 표현하기 위한 차원 축소 기법이 고안되어 왔으며, PCA(Principal Component Analysis), SVD(Singular Value Decomposition), NMF(Non-Negative Matrix Factorization) 등의 개념이 널리 사용되고 있다. 또한 이러한 차원 축소 개념과 함께 용어의 동시출현(Co-occurrence) 정보를 활용하여 용어 간 의미적 유사성을 표현하기 위한 시도가 활발하게 이루어지고 있으며, 대표적인 개념으로 LDA(Latent Dirichlet Allocation), LSA(Latent Semantic Analysis), pLSA(Probabilistic Latent Semantic Analysis), 그리고 최근 활용도가 높아지고 있는 Word2Vec을 들 수 있다.

이렇게 구조화된 텍스트 문서는 분석 목적에 따라 다양한 형태로 활용될 수 있으며, 대부분 전통적 데이터 마이닝의 대표적 기법인 빈도 분석(Frequency Analysis), 군집화(Clustering), 그리고 분류(Classification)의 주요 개념을 활용한다. 구체적으로 빈도 분석을 활용한 텍스트 마이닝의 대표적인 예로는 특정 문서 또는 문서 집합의 용어를 빈도에 따라 상이한 크기로 도식화하여 나타내는 워드 클라우드(Word Cloud) 분석, 용어 간 동시출현 관계를 도식화하여 나타내는 용어 망(Word Network) 분석, 그리고 빈출 어휘의 기간별 추이를 나타내는 추세(Trend) 분석 등이 있다. 또한 텍스트 마이닝 응용 중 군집화에 기반을 둔 가장 대표적인 분야로, 방대한 양의 문서로부터 주요 토픽을 추출하고 각 토픽에 대응되는 문서를 식별하여 제공하는 토픽 모델링(Topic Modeling)

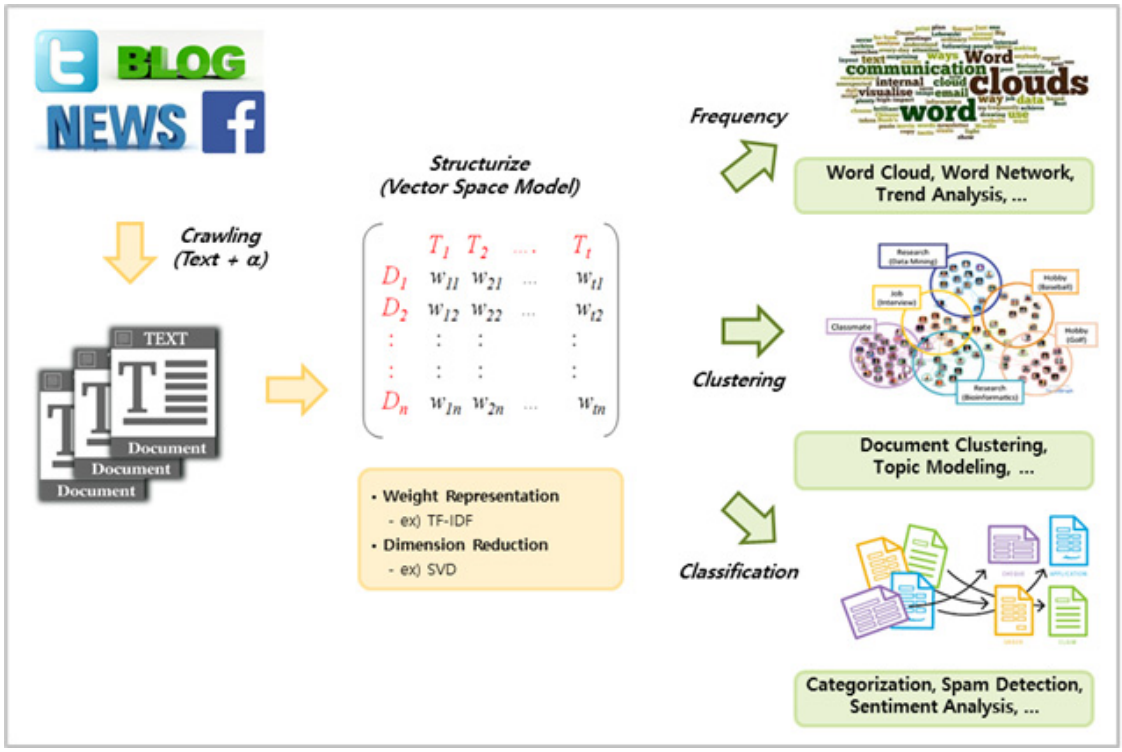


그림 1. 텍스트 분석 관련 기술 개요
Fig. 1. Overview of text analytics related techniques and applications

을 들 수 있다. 토픽 모델링은 방대한 양의 문서를 그 주제에 따라 묶음으로 군집화하여 제공한다는 측면에서 문서 군집화(Document Clustering)와 유사하지만, 전통적인 경성 군집화(Hard Clustering)와 달리 하나의 문서가 여러 토픽에 동시에 대응될 수 있다는 점에서 현실 세계의 모델링에 보다 적합한 것으로 평가받고 있다. 토픽 모델링은 그 자체로도 방대한 양의 문서에 대한 통찰(insight)를 제공한다는 점에서 의미가 있지만, 토픽 모델링 결과를 활용하여 다양한 형태의 분석을 수행할 수 있다는 점에서 더욱 활용가치가 있다. 가장 보편적인 토픽 모델링 활용 시나리오는 그림 2와 같다.

우선 분석하고자 하는 주제를 선정하고, 방대한 문서 집합으로부터 해당 주제에 속하는 문서를 추출하기 위한 주제 정의식을 도출한다. 주제 정의식이 올바르지 않으면 분석 주제와 무관한 문서가 분석에 포함되거나, 반대로 분석 주제를 담고 있는 문서가 분석에서 대량 누락될 가능성이 있다. 따라서 올바른 주제 정의식의 도출은 목표 주제에 대한 정확한 분석을 위한 필수 조건이라 할 수 있다. 예를 들어 분석 주제가 "감염병 대응 기술"인 경우 주제 정의식은 "감염", "전염", "바이러스" 중 최소한 하나의 용어를 포함하면서, 이와 동시에 "방역", "면역", "격리", "대응", "역학조사" 중 최소한 하나의 용어를 포함하는 것으로 작성될 수 있다. 이렇게 추출된 분석 대상 문서에 대해 SAS Enterprise Miner와 같은 상용 소프트웨어 또는 R과 같은 오픈소스 소프트웨어를 사용하여 토픽 모델링을 수행함으로써 주요 토픽을 도출하게 된다. 대부분의 경우 토픽은 주요 키워드의 조합으로 기술되며, 용어 사전(Start List) 또는 불용어 사전(Stop List)을 보완하며 이 과정을 반복적으로 수행함으로써 양질의 토픽을 도출할 수 있다. 또한 각 토픽에 속한 문서 수의 기간별 추이를 분석함으로써 해당 토픽의 추세를 파악할 수 있으며, 특정 기간 특정 토픽에 대한 추가 분석을 통해 해당 토픽과 관련된 심층 내용을 파악할 수 있다. 예를 들어 그림 2에서 "환자, 검사, 차, 호흡기, 서울"로 기술된 Topic1의 문서 수가 기간 P4에 급증한 것으로 나타나는 흥미로운 양상을 보였다. 따라서 Topic1에 속하는 문서 중 기간 P4에 해당되는 문서에 대한 워드 클라우드 분석을 통해 이 토픽에 대해 어떤 내용이 언급되었는지 파악할 수 있으며, 기간 P4의 문서 중 Topic1과의 관련도가 가장 높은

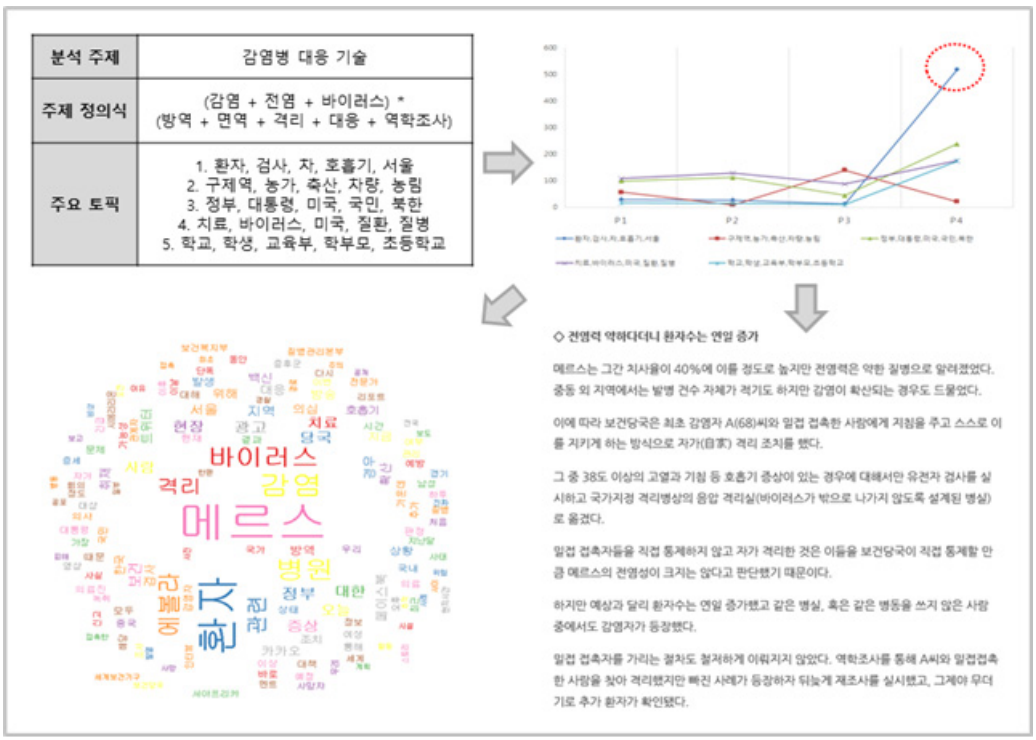


그림 2. 토픽 모델링 분석 주요 활용 시나리오
Fig. 2. Usual Scenario of Topic Modeling Applications

문서를 직접 조회함으로써 해당 기간에 해당 토픽에 대해 어떤 논의가 있었는지 더욱 정확하게 파악할 수 있다.

본 절에서는 빅데이터 분석의 다양한 분야 중 비정형 텍스트 분석을 다루는 텍스트 마이닝의 개념 및 관련 기술을 간략히 소개하였으며, 특히 최근 분석 수요가 급증하고 있는 토픽 모델링의 개념 및 분석 시나리오를 예시를 통해 소개하였다. 본 논문의 이후 구성은 다음과 같다. 우선 다음 절인 2절에서는 텍스트 마이닝 분야의 최근 연구 및 기술 동향을 소개하고, 3절에서는 토픽 모델링을 활용하여 다양한 분야의 문제를 해결하기 위한 최근 연구 사례를 몇 가지 요약한다. 마지막 절인 4절에서는 앞서 소개된 내용의 간략한 요약과 함께 본 연구의 의의를 제시한다.

Ⅱ. 텍스트 분석 관련 기술 및 연구 동향

텍스트 마이닝은 방대한 양의 텍스트 데이터에 포함된 수많은 정보를 사용자가 원하는 목적에 맞게 요약하는 기법 및 과정으로 정의할 수 있다. 많은 분야에서 다양한 텍스트 처리 기술을 활용하여 문서를 구조화하고 있으며, 이렇게 구조화된 문서를 분석하여 각 분야의 문제 해결을 위한 새로운 통찰을 얻고 있다. 본 장에서는 텍스트 마이닝에 활용되고 있는 주요 기법들을 설명하고, 최근 수행된 텍스트 마이닝 연구들의 동향을 살펴보고자 한다.

2.1 텍스트 분석 기술에 대한 연구

최근 분야를 막론하고 텍스트 분석 수요가 지속적으로 높아지고 있으며, 이에 따라 텍스트 처리 및 분석을 위한 관련 기술 또한 많은 관심을 받고 있다. 텍스트는 사용자의 전달 목적이나 관점, 심지어는 전달 매체에 따라 각각 상이한 형태로 기술된다는 비정형성을 갖기 때문에, 분석에 앞서 컴퓨터가 이해할 수 있는 형태로 변형시키는 구조화 과정이 반드시 필요하다.

벡터공간모델(Vector Space Model)은 텍스트의 구조화를 위한 대표적 기법으로, G. Salton과 그의 동료들이 SMART 정보 검색 시스템을 위해 처음 고안하였다[1-2]. 이 모델은 각각의 텍스트 문서들을 "수많은 용어의 집합"으로 간주하고, 이들을 해당 공간 내의 벡터로 표현함으로써 문서를 구조화한다. 이때, 각 문서가 갖는 용어의 수는 이 모델에서 형성되는 벡터공간의 수와 같으며, 각 용어가 출현하는 빈도 정보는 해당 벡터의 값으로 표현된다. 벡터공간모델에서 개별 벡터들은 사용자의 질의문(Query)을 기준으로 해당 공간 내에서 서로 다른 위치에 정렬되며, 벡터 간 거리는 의미적 유사성에 따라 표현된다. 즉, 의미상 유사성을 띄는 벡터들은 서로 가까운 위치에 존재하지만, 의미상 관련이 없는 벡터들은 멀리 떨어진 위치에 존재하게 된다. 벡터공간모델은 각 용어가 대상 문서에 출현하는 빈도를 용어 가중치(Term Weight)로 나타내며, 용어 가중치는 구조화 이후의 분석 과정에서 여러 용도로 활용된다.

용어 가중치를 구성하는 요소는 크게 세 가지로 구분될 수 있다. 먼저 용어 빈도 요소(Term Frequency Factor)는 H. P. Luhn에 의해 처음 소개된 개념으로[3], 문서 내에서 등장하는 용어의 단순 발생 빈도를 가중치로 사용한다. 예를 들어, 대상 문서 내에서 해당 용어가 많이 발생하는 경우, 이 용어는 높은 가중치를 가지게 된다. 다음으로 수집 빈도 요소(Collection Frequency Factor)는 여러 문서 중 하나의 문서를 다른 문서와 구분하기 위해 사용되는 기법들을 총칭하며, 가장 대표적인 개념으로 역문서 빈도(Inverse Document Frequency)를 들 수 있다. 이는 K. S. Jones에 의해 처음 제안된 개념으로[4], 용어의 특수성이 용어가 등장하는 문서의 수에 반비례한다는 가정에 따라 각 용어의 가중치를 보정한다. 이에 따르면 많은 문서에서 공통으로 등장한 용어의 경우 낮은 가중치를 갖도록 조정되며, 매우 적은 문서에서 드물게 등장하는 용어의 경우 상대적으로 높은 가중치를 갖게 된다. 이처럼 역문서 빈도는 용어의 특수성을 가중치에 반영하는 역할을 수행하여, 여러 문서에서 자주 사용되는 일반적인 용어들이 문서의 핵심어로 나타나게 되는 부작용을 방지한다. 마지막으로 길이 정규화 요소(Length Normalization Factor)는 문서별 길이에 의한 차이를 보정하는 요소이다[5]. 각 문서는 서로 다른 길이의 텍스트로 구성되며, 문서의 길이가 길수록 당연히 많은 수의 용어를 포함하게 된다. 예를 들어, 트위터(Twitter)의 경우 140자 이내의 단문으로 구성되지만, 뉴스 기사(Article)의 경우 여러 개의 긴 문장으로 구성된다. 따라서 특정 용어를 포함하고 있는 문서를 검색할 경우, 트위터보다는 뉴스 기사가 결과 집합에 나타날 가능성이 크다. 길이 정규화 요소는 이러한 현상을 고려하여 분석 대상 문서들이 동일한 길이를 가지도록 변환하는 역할을 수행한다.

이상 세 가지 요소들은 문서가 포함하는 용어들의 가중치를 파악하기 위한 핵심요소로, 이를 응용하여 개발된 여러 개념이 많은 텍스트 마이닝 연구에서 활용되고 있다. 일반적으로 벡터공간모델의 용어 가중치

도출을 위해 TF-IDF(Term Frequency – Inverse Document Frequency)가 널리 사용된다[2]. TF-IDF는 하나의 용어가 특정 문서에서 가지는 출현 빈도와 전체 문서에서 가지는 출현 빈도의 역 비율을 곱하여 산출된다. 용어 가중치 도출을 목적으로 하는 다양한 기법들은 환경에 따라 성능의 차이를 보이기 때문에 상황에 따른 기법별 성능 비교에 대한 많은 연구가 수행되었으며, 그 결과 벡터공간모델에서는 용어 빈도와 역 문서 빈도, 길이 정규화의 세 가지 요소를 모두 반영하여 도출한 용어 가중치가 가장 우수한 성능을 나타내는 것으로 확인되었다[6]. 이에 따라, 최근 벡터공간모델을 활용하는 텍스트 마이닝 연구들은 TF-IDF와 길이 정규화 요소의 곱을 통해 최종적인 용어 가중치를 도출하고, 이를 각 분석 단계에 활용하고 있다.

앞서 설명된 벡터공간모델을 통해 구조화된 문서들은 벡터 행렬(Vector Matrix) 형태로 표현된다. 이때, 벡터 행렬을 이루는 각 행의 수는 분석 대상이 되는 문서의 수를 의미하며, 각 열의 수는 문서에서 사용된 용어의 수를 나타낸다. 하지만 이러한 방식으로 구조화 된 벡터 행렬은 지나치게 많은 차원을 사용하므로, 이를 그대로 표현하여 분석에 적용하기에는 무리가 있다. 따라서 엄청난 크기의 벡터 행렬을 실제로 처리 가능한 수의 차원으로 축소하는 과정이 반드시 필요하다.

차원 축소의 가장 고전적 기법의 하나인 PCA (Principal Component Analysis, 주성분 분석)는 역학에서 사용되는 주요 이론의 유사 이론으로 처음 고안되어[7], H. Hotelling에 의해 발전되고 명명되었다[8]. PCA는 개념적으로 변수들 간의 상호 연관성을 활용한다. 구체적으로, PCA는 수학적 투사법을 적용해 해당 데이터에서 핵심 정보를 추출하고, 이를 새로운 직교 변수 행렬의 도출에 사용함으로써 기존 데이터의 차원을 동일 또는 축소 변환시킨다. 또한 이 과정을 통해 새롭게 도출된 행렬은 관측치와 변수 사이의 유사성을 좌표상의 점으로 표현하는데[9-11], 이를 통해 패턴 및 추세와 이상치를 포함하는 데이터의 특성을 변환 이전에 비해 훨씬 쉽게 파악할 수 있다. 다음으로 SVD(Singular Value Decomposition, 특이값 분해)는 신호 처리와 통계 분야에서 자주 사용되는 기법으로, 많은 수학자들의 오랜 연구를 토대로 개발되었다. 학계에서는 SVD가 선형대수학을 이용해 행렬을 분해한 E. Beltrami(1835-1899), C. Jordan (1838-1921), J. J. Sylvester(1814-1897)의 연구와 적분 방정식을 적용해 행렬을 분해한 E. Schmidt (1876-1959), H. Weyl(1885-1955)의 연구에 의해 확립되었다고 인식하고 있으며[12], 이 외에도 수많은 학자들의 연구가 이 기법의 발전에 영향을 미쳤다. SVD는 원본 행렬을 양극단에 따라 세 행렬의 곱으로 분해하여, 이를 통해 차원을 재구성한다. 이렇게 분해된 행렬들은 일반적으로 원본 행렬보다 낮은 차원으로 구성되며, 앞서 언급된 주성분 분석의 수행에 활용된다. 또한 NMF(Non-negative Matrix Factorization, 음수 미포함 행렬 분해)는 계량분석화학 분야에서 연구되어 온 자기 모델링 그래프 분해(Self Modeling Curve Resolution) 알고리즘에서 발전된 개념으로, D. D. Lee와 H. S. Seung의 소개에 의해 널리 활용되기 시작하였다[13]. 이 기법은 내부 과정에서 뺄셈을 허용하지 않는 특성이 있으며, 하나의 비음수 데이터를 두 개의 행렬로 분해하는 역할을 수행한다. 예를 들어 원 데이터가 $m \times n$차원의 행렬로 구성되어 있을 때, NMF는 이를 $m \times r$차원과 $r \times n$차원의 행렬로 분해한다. 이때 $m \times r$차원의 행렬은 원 데이터에 비해 낮은 차원을 가지지만, 기존 데이터의 특징은 그대로 반영한다.

텍스트 마이닝은 앞서 소개한 구조화 및 차원 축소 기법과 함께, 용어들의 의미적 유사성을 파악하기 위한 다양한 기법들을 활용하고 있다. 이러한 기법들은 문서에서 도출되는 용어들 사이의 의미적 유사성의 정도를 수치로 나타내주는 역할을 하며, 해당 수치는 구조화 이후의 여러 분석 과정에서 사용된다. 일반적으로 벡터 간 유사도 측정에는 코사인 유사도(Cosine Similarity)가 널리 사용되며, 이는 "0"이 아닌 두 벡터 사이의 코사인 각도에 의해 산출된다[14]. 또한 텍스트 분석을 위한 많은 연구에서 전통적으로 LSI(Latent Semantic Indexing, 잠재 의미 색인)라고도 불리는 LSA(Latent Semantic Analysis, 잠재 의미 분석)를 통해 용어의 의미적 유사도를 파악하고 있다. 이 기법은 M. B. Koll의 연구에서 발전하여[15], S. Deerwester와 그의 동료들에 의해 확립되었다[16]. 이 기법에서는 의미가 유사한 용어들은 비슷한 특징을 가지는 문서에서 도출될 것이라 가정하고 있으며, 앞서 설명된 SVD를 사용해 대상 문서들을 분해하여 행렬로 나타낸다. 이때 대상 문서에서 도출된 용어와 구문들은 각각 분해된 행렬들의 행과 열로 구성되며, 일련의 과정을 통해 산출된 코사인 값을 통해 용어들 사이의 유사도가 도출된다. LSI는 비교적 단순한 알고리즘을 통해 용어들의 유사도를 산출할 수 있기 때문에 널리 사용되고 있지만, 분석 과정에서 대상 문서에 대한 정보의 손실이 있으며 잠재 정보의 해석을 위한

기준이 모호하다는 한계를 갖고 있다.

다음으로 pLSA(Probabilistic Latent Semantic Analysis, 확률 잠재 의미 분석)는 pLSI(Probabilistic Latent Semantic Indexing, 확률 잠재 의미 색인)라고도 불리며, T. Hofmann에 의해 제안된 확률적 토픽 모델링 기법이다[17]. 이 기법은 개별 문서가 포함하는 내용에 수많은 토픽들이 잠재되어 있다고 간주하고, 이들이 해당 문서를 구성한다고 인식한다. 또한 잠재된 토픽들은 해당 문서에서 도출된 용어들의 분포에 의해 표현된다고 설명한다. pLSA의 내부 알고리즘은 개별 문서와 특정 토픽을 구성하는 임의의 용어, 그리고 기존 토픽으로부터 도출 가능한 잠재 토픽의 관계를 파악하여 이를 확률로 나타낸다. 이렇게 도출된 해당 문서의 잠재 토픽별 확률 분포는 개별 문서를 구성하는 토픽들을 나타낸다. 이 기법은 분석 대상에 새로운 문서가 추가될 경우 기존에 수행하였던 과정을 다시 반복한다는 한계점을 갖지만, 분석 과정을 통해 매우 의미 있는 결과를 도출할 수 있기 때문에 현재까지도 많은 연구에서 활용되고 있다. 또한 pLSA를 응용한 많은 알고리즘이 통계학 및 유전공학을 포함하는 여러 분야의 연구에 활용되고 있다. 텍스트 마이닝 연구에서 가장 빈번하게 사용되는 기법 중 하나인 LDA(Latent Dirichlet Allocation, 잠재 디리클레 할당)는 디리클레 분포(Dirichlet Distribution)를 적용하여 앞서 언급된 pLSA의 한계점을 보완한 기법으로, D. M. Blei와 그의 동료들에 의해 고안되었다[18]. LDA는 문서 내 용어들의 존재 여부만이 중요하다고 가정하며, 이들의 순서는 고려하지 않는 특징이 있다. 이 기법의 내부 알고리즘에서는 디리클레 분포의 매개변수를 사용하여, 개별 문서와 단어들이 특정 토픽에 포함될 확률과 전체 문서에서 도출된 개별 단어들이 특정 토픽에 포함될 확률을 산출한다. 이때, 해당 값은 전체 문서와 용어들이 포함되는 토픽 각각에 대한 정보를 나타낸다. LDA의 내부 구조는 모듈형으로 이루어져 있기 때문에 여러 형태로 쉽게 변환될 수 있으며, 다양한 분야에서 연구 목적에 따라 이 기법을 응용하고 있다. LDA의 대표적 응용기법으로는 계층적 잠재 디리클레 할당(Hierarchical LDA), 이중 잠재 디리클레 모형(LDA-dual model), 그리고 계층적 디리클레 프로세스(Hierarchical Dirichlet Process) 등이 있다.

이 외에도 단어의 의미적 유사성에 기반을 둔 다양한 기법들이 꾸준히 고안되고 있으며, Word2vec은 그 가운데 가장 많은 주목을 받고 있다. 이 기법은 구글의 연구원 T. Mikolov와 그의 동료들이 처음 고안하였으며[19], 이를 구성하고 있는 핵심 개념은 최근에도 여러 연구자들에 의해 재조명되고 있다[20]. Word2vec은 많은 문장으로 구성된 텍스트를 수백 차원의 벡터 공간으로 구조화하고, 텍스트에 포함된 용어들을 이 공간에 할당함으로써 용어 간 관련성을 표현한다. 이를 구체적으로 설명하면, 두 개의 신경망(Neural Network)을 통해 용어 간 전후 분포를 학습하고, 이를 토대로 각 용어와 관련된 용어를 도출한다. 이 기법의 용어 학습에는 CBOW(Continuous Bag of Words)와 Skip-gram 모델이 사용된다[19]. CBOW 모델은 주변 용어를 대상 용어의 예측 기준으로 사용하기 때문에 적은 문서 집합에 적합한 반면, Skip-gram 모델은 대상 용어로부터 그 주변 용어를 예측하기 때문에 큰 문서 집합에 적용되기에 바람직한 특성을 갖는다. Word2vec은 다른 기법들에 비해 비교적 간단한 알고리즘으로 구현됨에도 불구하고 비교적 빠른 시간 내에 매우 높은 정확도로 유사 용어를 도출하는 것으로 알려져 있다. 따라서 이 기법은 텍스트나 영상 또는 이미지를 포함하는 여러 분야의 연구에서 다양한 용도로 활용되고 있으며, 최근에는 이 기법을 기초로 구문이나 문서를 분석 대상으로 하는 Paragraph2vec이나 Doc2vec이 개발되어 다양한 분야에서 폭넓게 활용되고 있다.

본 절에서는 문서를 구조화하는 기본적 과정과 문서 특성을 도출하는 과정에 사용되는 대표적인 개념 및 기법을 소개하였다. 본 절에서 소개한 내용은 텍스트 분석 뿐 아니라 최근 그 관심이 급증하고 있는 딥러닝(Deep Learning)분야에서도 널리 활용되고 있으므로, 빅데이터 분석 전반에 대한 통찰을 위해 이들 개념 및 기법에 대한 보다 깊은 이해가 필요하다.

2.2 텍스트 분석 활용에 대한 연구

이전 절에서 소개한 과정에 따라 구조화된 텍스트는 빈도 분석(Frequency Analysis), 군집화(Clustering), 그리고 분류(Classification) 등 다양한 응용에 사용되며, 이러한 응용을 위한 대표적 기법들은 서론에서 간략히 소개한 바 있다. 본 절에서는 텍스트 마이닝을 활용한 국내외 연구 사례들을 소개하고, 이를 통해 텍스트 마이닝 분야의 최근 연구 동향을 파악한다.

2.2.1 빈도분석

텍스트 분석 응용 중 가장 직관적이면서도 널리 활용되는 분야로, 문서에 출현한 용어의 빈도를 다양한 관점에서 시각화하여 보여주는 빈도 분석(Frequency Analysis)을 들 수 있다. 워드 클라우드(Word Cloud)

기법은 특정 문서 집합에서 높은 빈도로 출현한 용어를 다른 용어에 비해 상대적으로 크게 표현함으로써 해당 용어를 강조하는 대표적 빈도 분석 기법이다. 워드 클라우드는 연구 프로젝트 보고서 분석을 통해 해당 연구를 심층 분석한 연구[J21], 중학 과정 중 가정과 소비생활 영역의 핵심 교육내용을 파악하기 위해 해당 교과서를 분석한 연구[J22], 주가 예측을 위해 뉴스 데이터를 분석한 연구[J23], 문서들의 개요 제공을 위해 뉴스 데이터를 분석한 연구[J24], 항생제 오용의 증거 도출을 위해 트위터 데이터를 분석한 연구[J25] 등 다양한 분야의 연구에 사용되고 있다.

다음으로 용어 망(Word Network) 분석은 용어 간 관계를 망 형태로 도식화하여 표현하는 기법으로, 일반적으로 용어 간 관계는 용어들의 동시출현 정보로부터 식별된다. 동시출현 단어 분석(Co-word Analysis)은 동시에 자주 출현하는 용어일수록 서로 관련성이 높을 것이라는 가정 하에 수행되며, 각 용어를 점으로, 용어 간 관계를 선으로 도식화하여 나타낸다. 이렇게 망 형태로 도식화된 정보는 최근 비약적인 발전을 보이고 있는 네트워크 분석의 다양한 기법을 활용하여 추가 분석이 가능하게 된다. 용어 망에 대한 네트워크 분석을 수행한 연구로는 국가 R&D 특허 데이터를 분석해 연구 키워드를 탐색하고 이들의 동향을 파악한 연구[J26], 의료 정보학(Medical Informatics) 분야의 논문을 분석하여 해당 분야의 학제적 특성을 확인한 연구[J27], 오픈 액세스 분야의 논문을 분석하여 해당 분야의 연구 경향이 반영된 지적구조를 도출한 연구[J28], 한국어교육학, 국어교육학, 국어학 논문을 분석하여 한국어 교육학의 정체성을 분석한 연구[J29] 등이 있다.

마지막으로 추세(Trend) 분석은 시간의 추이에 따른 특정 객체의 변화 패턴을 분석하여, 이를 목적에 맞게 재분석하거나 이후 나타날 객체의 패턴을 예측하기 위해 활용된다. 최근 구글(Google)과 네이버(Naver)를 비롯한 많은 포털사이트에서 키워드 검색 통계량 기반의 추세 분석을 무료로 제공하기 시작하면서, 많은 연구에서 이러한 기초 정보를 적극적으로 활용하고 있다. 특히 추세 분석을 위한 많은 연구가 구글 트랜드(Google Trend)에서 제공하는 데이터를 활용하고 있으며, 미국의 소매 판매와 자동차 판매 등 다양한 판매정보를 예측한 연구[J30], 구글의 검색 통계와 영국의 소비 패턴 간 관계를 규명한 연구[J31], 독일과 이탈리아의 실업률을 예측한 연구[J32] 등을 그 예로 들 수 있다. 구글 트랜드 이외에도 다양한 문서들이 추세 분석에 활용되고 있으며, 그 예로 특허 문서를 대상으로 핫토픽을 검출하고 추세를 파악한 연구[J33], TF-IDF 기법을 활용하여 트위터 데이터에서 주요 이벤트를 식별하고 이들의 추세를 분석한 연구[J34], 그리고 IT분야의 추세를 분석한 연구[J35] 등을 들 수 있다.

2.2.2 군집화

비지도학습(Unsupervised Learning)을 활용하는 군집화는 텍스트 응용 분야의 고전적이면서도 대표적인 영역이다. 텍스트 분석 분야에서의 군집화는 크게 문서 군집화(Document Clustering)와 토픽 모델링(Topic Modeling)으로 나누어 이해할 수 있다. 문서 군집화는 문서 내 용어들의 문서 간 유사성을 통해 문서를 군집화하는 방식으로, 정형 데이터의 군집화와 마찬가지로 H. D. Steinhaus에 의해 처음 소개되고[36], J. MacQueen에 의해 명명된[37] K-means(K-평균) 알고리즘을 핵심 기법으로 사용한다. 문서 군집화를 활용한 연구의 예로는 k-means 기법을 활용하여 논문 검색을 군집화한 연구[J38], 리더-팔로우 기법을 활용하여 포털에서의 검색 결과를 군집화한 연구[J39], k-means 기법을 활용하여 법령 정보를 분석한 연구[40], 소셜 북마킹 서비스 "Delicious"에서 사용된 태그들을 군집화한 연구 등이 있다[41].

토픽 모델링은 각 문서를 임의의 이슈들의 집합으로 가정하고, 해당 문서를 구성하는 이슈 및 용어의 중요도를 확률적으로 제시하는 방법이다[42]. 이 기법은 개발과 동시에 텍스트 분석 연구자들의 뜨거운 관심을 받으며 활발하게 사용되었을 뿐 아니라, 여론 분석이나 사회적 이슈 동향 파악 등 실무적으로도 널리 활용되고 있다. 토픽 모델링은 사전에 설정한 토픽 수에 따라 그 결과가 상이하게 도출되기 때문에, 분석 목적에 맞춰 적절한 토픽의 수를 지정하는 것이 매우 중요하다. 따라서 적절한 수의 토픽을 탐색하기 위한 방안들이 연구되어 왔으며, 그 예로 토픽 수에 따른 정확도 비교를 통해 최적의 토픽 수를 계산한 연구[43], 확률 모델 최적화를 통해 적절한 수의 토픽을 추정한 연구[44], 많은 수의 토픽으로부터 유사한 토픽들을 결합한 후 이를 최종 결과의 도출에 활용한 연구[45-46] 등이 있다.

토픽 모델링은 이슈를 도출하고 이들을 추적(Tracking)하기 위한 연구에 널리 활용되고 있다. 이와 관련된 연구로는 뉴스 데이터를 분석하여 세부 기간별 주요 이슈를 도출하고, 이슈 흐름도를 통해 이슈의 동적 변이과정을 파악한 연구[47], 뉴스기사와 트위터에서 이슈를 도출한 후 그 변화를 추적한 연구[48]가 있다. 이와 더불어 토픽 모델링은 특정 분야의 연구동

향 파악을 위해서도 빈번히 활용되었는데, 그 예로 PNAS(Proceedings of the National Academy of Sciences of the United States of America, 미국국립과학원회보)의 초록 분석을 통해 최근 가장 활발히 연구되고 있는 주제(Hot Topics)와 점차 연구의 빈도가 줄어드는 주제(Cold Topics)를 파악한 연구[43], 문헌정보[49], 교통[50], 시뮬레이션[51]등의 연구 추세를 분석한 연구가 있다. 또한 소셜 네트워크 서비스(SNS, Social Network Service)의 이슈들을 추적하기 위한 연구들도 수행되었으며[52-53], 개인 이메일과 NIPS(Neural Information Processing Systems) 학회 프로시딩(Conference Proceeding), 200년 동안 작성된 미연방 대통령 연설문으로 구성된 데이터에 대한 분석을 통해 주요 토픽을 도출하고, 시간의 흐름에 따른 이들 토픽의 변화를 살피는 TOT(Topics Over Tirne) 모델을 소개한 연구[54]도 수행되었다.

이 외에도 토픽 모델링의 활용 분야는 실로 다양해서, 학술지의 학제성을 측정한 연구[55], 뉴스에서 도출된 이슈와 관련 트윗의 대응을 통해 정보의 가치를 측정한 연구[56], 아마존, 호텔 리뷰, 커뮤니티의 리뷰를 분석하여 사용자 경험(User Experience)을 파악한 연구[57-59], 뉴스로부터 사회문제를 분석하거나[60], 질병을 예측한 연구[61], 바이오분야의 연구 논문과 뉴스를 분석하여 에볼라와 관련된 키워드 및 토픽을 추출한 연구[62], Wikipedia에 등재된 객체들의 관계를 식별한 연구[63], 트위터와 New York Times 기사의 분석을 통해 트위터가 갖는 뉴스 매체적 특성을 확인한 연구[64] 등이 다양한 분야에서 토픽 모델링을 활용하여 활발하게 수행되었다.

2.2.3 분류

텍스트 마이닝 분야에서는 다양한 목적의 분류 분석이 수행되고 있다. 텍스트 분석 분야의 가장 대표적인 분류 응용인 문서 분류(Document Classification)는 각 문서의 내용에 따라 문서를 사전에 정의된 범주로 자동 할당한다[65-69]. 문서의 분류에는 전통적인 데이터 마이닝의 다양한 분류 기법이 적용될 수 있으며, 이 가운데 V. Vapnik에 의해 고안된 SVM(Support Vector Machine, 지지 벡터 기계)[70]의 사용이 특히 두드러지게 나타난다[71-73]. 이외에도 나이브 베이즈(Naïve Bayes) 기법을 활용해 문서를 분류한 연구[74], 나이브 베이즈 기법과 워드넷의 결합을 통해 문서를 분류한 연구[75], 그리고 이메일 분류[76] 및 신문 기사 분류[77] 등 다양한 문서에 대해 다양한 분류 기법이 적용된 바 있다.

한편 기계 학습을 통해 문서를 분류하는 대부분의 방식은 각 문서가 하나의 범주에만 속한다고 가정하기 때문에 복합 주제로 구성된 문서의 카테고리 식별에는 적용되기 어렵다. 따라서 이러한 한계를 극복하기 위해 카테고리와 카테고리, 특성과 특성, 그리고 카테고리와 특성으로부터 세 가지 상관성을 도출하고, 이를 통해 대상 문서에 다중 레이블을 적용할 수 있는 문서 분류기를 고안한 연구[78]가 수행되었다. 또한 개별 문서의 특징과 이들의 부분집합을 파악하고, 이를 통해 다중 분류기의 성능을 향상시킨 연구[79]도 진행되었다. 하지만 이러한 다중 분류 연구는 이미 다중 레이블이 부여된 문서의 학습을 통해 분류 규칙을 도출하기 때문에, 단일 레이블만 부여되어 있는 실생활의 대부분의 문서를 학습 데이터로 활용할 수 없다는 한계가 있다. 따라서 이를 극복하기 위해 카테고리, 토픽, 문서 간 관계를 분석하고, 이를 통해 단일 카테고리의 문서로부터 추가 주제를 발굴하여 이를 다중 카테고리로 자동 확장하는 방안을 제안한 연구[80]도 국내에서 최근 수행되었다.

텍스트 분석 기술이 발전함에 따라, 텍스트 정보가 갖는 유용성은 더욱 높아졌다. 하지만 이와 더불어 텍스트 데이터의 왜곡을 통해 특정 목적을 달성하려는 시도 또한 증가하고 있으며, 이는 이른바 “스팸 데이터”(Spam Data)가 양산되는 현상으로 이어지고 있다. 이러한 스팸 데이터는 사용자의 효율적인 정보 검색을 저해할 뿐 아니라 정보 및 매체에 대한 신뢰를 저하시키기 때문에, 이를 사전에 방지하기 위한 스팸 탐지(Spam Detection)가 텍스트 분석 분야의 주요 이슈로 최근 각광받고 있다. 전통적인 스팸 탐지 연구는 대부분 이메일을 대상으로 수행되었으며, 그 예로 SVM 기법을 활용해 스팸 메일을 검출한 연구[81], SVM과 유의어 사전(Thesaurus Dictionary)을 결합하여 스팸 메일을 검출한 연구[82], 나이브 베이즈 기법을 적용하여 스팸 메일을 검출한 연구[83-85], N-gram 색인 후 나이브 베이즈와 SVM 기법을 활용해 스팸 메일을 검출한 연구[86] 등이 있다. 또한 이메일 외에도 스팸 블로그 식별을 위해 SVM 기법을 활용한 연구[87], 키워드와 태그를 분석하여 의미 있는 태그를 선별하는 연구[88] 등도 수행되었다.

또한 최근에 많은 주목을 받고 있는 감성 분석(Sentiment Analysis) 역시 분류 분석의 한 특수한 경우로 파악할 수 있다. 감성 분석은 제품이나 서비스 또는 조직이나 개인의 이슈, 사건, 토픽, 그리고 이들의 여러 속성에 대한 사람들의 의견, 평가, 태도, 감정 등을 파악하기 위한 분석기법이다[89]. 이 기법은 텍스

트를 구성하는 내용을 통해 사용자의 감성(Sentiment), 정서(Affect), 주관(Subjectivity), 감정(Emotion) 등 다양한 의견들을 식별하기 때문에 오피니언 마이닝(Opinion Mining)이라고도 불린다. 감성 분석은 각 문서의 최소 단위인 어휘들의 감성 극성(Sentiment Polarity)에 따라 이루어진다. 즉, 주요 어휘들이 가지는 감성 극성에 따라 사전에 정의된 감성 사전을 구축하고, 이에 따라 새로운 문서가 갖는 어휘들의 감성 극성을 분석하여 문서 전체의 감성을 도출한다. 일반적으로 이 기법을 다루는 연구들은 개별 문서가 하나의 감성을 표현한다는 가정 하에 분석을 수행하고 있지만[90-91], 다른 연구에서는 문서의 하위 개념인 구와 절 단위 또는 주관성 구분(Subjectivity Classification)에 의한 분석을 수행하기도 한다[92]. 더 나아가 최근에는 각 개체와 그 개체가 가지는 속성을 분석하기 위한 연구들도 시도되고 있다. 이러한 분석들은 개별 개체들의 속성에 대한 감성까지 파악할 수 있다는 장점을 가지지만, 매우 복잡한 과정을 통해 개체 인식이나 개체의 부분 요소 및 속성 파악이 수행되기 때문에 문서나 문장 단위의 연구에 비해 상대적으로 난이도가 높다[93].

살펴본 바와 같이 감성 분석의 수행에 있어서 감성 사전은 핵심 역할을 담당하기 때문에, 양질의 감성 사전의 확보는 성공적인 감성 분석을 위한 필수 요소라고 할 수 있다. 감성 사전은 크게 두 가지 방식으로 구축될 수 있다. 우선 사전기반 접근법에서는 WordNet을 포함하는 다양한 사전들로부터 도출된 시드 어휘를 기준으로, 어휘들 간의 유사성과 거리 관계 파악을 통해 다른 어휘들의 감성값을 도출한다[94-97]. 한편 말뭉치기반 접근법에서는 실제로 수집된 문장들에 대한 구분 분석을 통해 감성 사전을 구축한다[98-99]. 하지만 전통적 접근법을 통해 구축한 감성 사전들은 동일 어휘일지라도 상황이나 목적에 따라 상이한 감성 값을 가질 수 있다는 현상을 반영하지 못하는 한계를 갖고 있다. 따라서 이를 극복하기 위해 특정 목적에 맞는 감성사전을 구축하려는 시도도 다수 이루어졌다[100-102].

다양한 매체의 데이터들이 감성 분석 연구의 대상이 되어 왔다. 대표적 SNS인 트위터 데이터를 분석한 연구로는 이모티콘을 학습한 후 이를 감성 분석에 활용한 연구[103], 해시태그(Hash Tag)와 스마일 레이블을 동시에 활용해 감성 분석을 수행한 연구[104], 키워드로 해시태그를 활용한 준지도학습(Semi-supervised Learning)을 통해 감정의 극성을 분석한 연구[105], 특정 감정을 통해 다우 존스 평균 주가의 예측을 시도한 연구[106], 동일한 데이터에 SVM 기법과 나이브 베이즈 기법을 적용한 후 그 결과를 비교한 연구[107] 등이 있다. 트위터 외에도 감성 분석은 블로그 데이터[108-109], 뉴스 데이터[110], 인스타그램[111], 페이스북[112] 그리고 영화 리뷰[91, 113-114] 등 다양한 데이터에 대해 수행되고 있다. 또한 둘 이상의 매체에 대한 감성 분석도 활발하게 이루어지고 있으며, 그 예로 신문기사와 영화 및 상품 리뷰를 대상으로 문서별 긍부정 분류를 시도한 연구[115], 뉴스, 블로그, 트위터 데이터를 대상으로 의미적 요소의 결합을 시도한 연구[116], 그리고 뉴스, 블로그, 트위터 데이터에 감성 분석을 통해 선호 농식품 리스트를 추천한 연구[117]를 들 수 있다.

Ⅲ. 토픽 모델링 활용 방법론

토픽 모델링은 다량의 문서로부터 핵심 이슈를 식별하고, 시간의 추이에 따른 이슈의 변화를 파악하기 위해 주로 사용된다. 뿐만 아니라 토픽 모델링의 중간 산출물 및 최종 산출물은 다양한 형태의 가공을 통해 추가 분석에 사용될 수 있으며, 토픽 모델링을 활용하여 보다 의미 있는 지식을 창출하기 위한 방법론이 꾸준히 개발되고 있다. 이에 본 장에서는 토픽 모델링 활용 방법론을 제안한 최근 국내 연구 중 참신성 및 기여도 측면에서 주목할 만한 연구를 선정하여 각 절에서 소개하고자 한다.

3.1 텍스트 분석을 활용한 과학기술이슈 여론 분석 방법론[118]

3.1.1 연구 개요

이 연구는 과학기술 관련 주요 사회이슈를 발굴하고, 이들 이슈의 추이 및 특정 이슈에 대한 상세 여론을 분석하는 것을 목적으로 수행되었다. 이 연구에서 제안한 방법론이 이전의 유사 연구에 비해 갖는 차별성은 다음의 측면에서 찾을 수 있다. 우선 이 연구는 과학기술과 관련된 국민의 여론이 과학기술 자체에 대해 형성되는 것이 아니라, 각 과학기술이 특정 사회문제에 적용될 때 형성된다는 인식을 분석에 반영하였다. 또한 기존의 유사 분석에서 한계로 지적되어 온 키워드 선정의 문제, 즉 분석 대상이 되어야 할 문서가 분석에서 빠지거나 주제와 직접적인 관련이 없는 문서가 분석에 포함되는 부작용을 최소화하기 위한 장치를 고안하였다. 마지막으로 어휘가 과도하게 정제되는 경우 분석 결과가 상투적으로 나타나고 반대로 정제 수준이 지나치게 낮은 경우 다듬어지지 않은 어

휘가 그대로 결과에 포함되어 분석의 신뢰성을 하락시키는 딜레마를 극복하기 위해, 용어 사전과 불용어 사전을 순차적으로 적용하는 방안을 제시하였다.

3.1.2 연구 동기

이 연구의 주요 동기는 과학기술에 대한 여론이 과학기술 자체에 대해 형성되는 것이 아니라 각 과학기술이 특정 사회 이슈에 적용될 때 형성되는 현상을 분석에 반영하는 것이다. 예를 들어 그림 3에서 "지능형 무인 비행체"에 대한 여론은 해당 기술이 "감시 정찰" 이슈에 적용될 때와 "무인 택배" 이슈에 적용될 때 서로 상이하게 나타남을 알 수 있다. 따라서 특정 기술과 특정 이슈가 교차하는 접점에서의 여론을 분석함으로써, 보다 현실적이고 활용 가능한 과학기술 여론 분석을 수행할 수 있을 것이다.

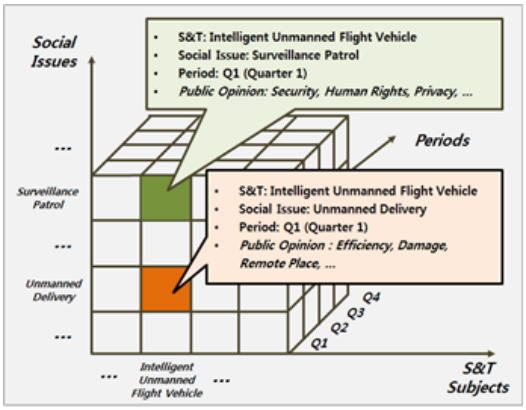


그림 3. 연구 동기 - 과학기술이슈 여론 분석
Fig. 3. Research motivation - Analysis of public opinion about science and technology issues

3.1.3 분석 데이터 및 실험 결과

이 연구는 트위터, 네이버 블로그, 다음 아고라, KBS 뉴스, 그리고 연합뉴스로부터 수집한 총 1,700,886건의 데이터를 분석에 활용하였다. 전체 데이터에 대해 논리 필터를 적용하여 "감염병 대응 기술"에 대한 문서 3,404건과 "지능형 무인 비행체 기술"에 대한 문서 450건을 추출하였으며, 이에 대한 토픽 모델링 및 이슈 트래킹을 수행하였다.

그림 4는 "감염병 대응 기술"과 관련된 이슈의 추이를 나타낸다. 그래프에서 가로축은 기간을 나타내며 2014년 6월 11일부터 2015년 6월 10일까지 1년의 기간이 Period 1 ~ Period 4로 구분되어 있다. 또한 세로축은 해당 기간 각 이슈에 속하는 문서의 수를 나타낸다. 해당 주제에 대한 이슈 트래킹 결과 "Issue 1"의

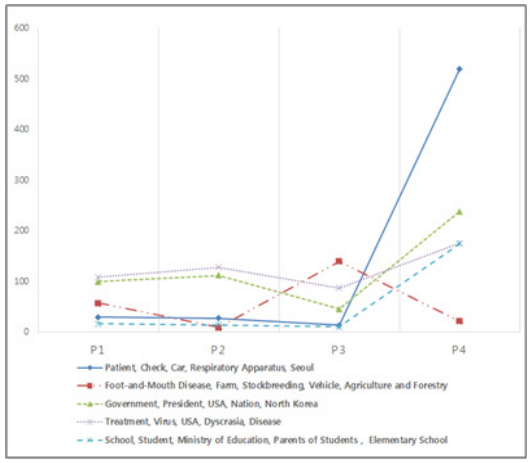


그림 4. "감염병 대응 기술"과 관련된 이슈의 추이
Fig. 4. A result of issue tracking for "Infectious disease control"

경우 Period 4의 기간에 다른 이슈에 비해 관심이 급격히 증가한 것을 알 수 있다. Issue 1은 "환자", "검사", "차", "호흡기", "서울"로 구성되어 있으며, Period 1 ~ Period 3의 기간 동안 거의 주목을 받지 못하다가 Period 4에 관련 문서의 수가 급증한 것으로 나타났다. 따라서 이 연구에서는 해당 기간 해당 이슈에 대한 관심이 급증한 원인과 여론을 분석하기 위해 워드 클라우드 및 용어 망 분석을 사용하여 이 이슈를 심층 분석 하였다.

3.2 비정형 텍스트 분석을 활용한 이슈와 동적 변이과정 고찰[47]

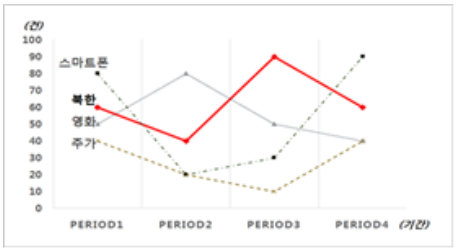
3.2.1 연구 개요

전통적 이슈 트래킹은 토픽 모델링을 활용해 특정 기간의 주요 이슈를 발굴하고, 해당 이슈를 구성하는 문서 수의 세부 기간별 분포를 분석하는 방식으로 수행되었다. 하지만 이 방식은 각 이슈를 구성하는 내용이 전체 기간에 걸쳐 변화 없이 유지된다는 가정에 따라 수행되기 때문에, 하나의 이슈를 구성하는 다양한 세부 이슈가 서로 영향을 주며 생성, 병합, 분화, 소멸하는 이슈의 동적 변이과정을 나타내지 못한다. 또한 전체 기간에 걸쳐 지속적으로 출현하는 키워드만이 이슈의 키워드로 도출되기 때문에, 세부 기간의 분석에서는 매우 다른 맥락으로 파악되는 구체적 이슈가 오랜 기간 동안의 분석에서는 큰 범주의 이슈에 함몰되어 가려지는 현상이 발생할 수 있다. 이러한 한계의 극복을 위해 이 연구에서는 특정 기간의 문서에 대한 독립적인 분석에 따라 세부 기간별 주요 이슈를 도출

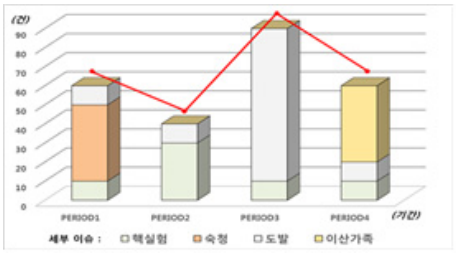
하고, 개별 이슈의 유사도에 기초한 이슈 흐름도를 도출하였다. 이러한 이슈 흐름도를 통해 각 이슈의 동적 변이과정과 특정 이슈의 선행 및 후행 이슈를 파악하고, 장기간 지속하는 일반적 이슈의 기간별 특성을 설명할 수 있는 방안을 제시하였다. 또한 각 문서의 카테고리 정보를 활용하여 카테고리 간 이슈 전이 패턴을 분석하였으며, 이를 통해 특정 카테고리 간에 단방향 전이와 양방향 전이의 흥미로운 패턴이 존재함을 확인하였다.

3.2.2 연구 동기

이 연구의 주요 동기는 하나의 이슈를 구성하는 다양한 세부 이슈가 서로 영향을 주며 생성, 병합, 분화, 소멸하는 이슈의 동적 변이과정을 규명하는 것이며, 동기의 일부가 그림 5에 소개되어 있다. 그림 5(a)는 전통적인 이슈 트래킹의 예를 보여주며, 가로 축은 기간을, 세로 축은 해당 기간의 각 이슈별 문서 수를 나타낸다. 이러한 분석은 각 이슈는 고정되어 있고, 이에 대한 반응만 기간에 따라 변화하는 것으로 가정하고 있다. 하지만 각 이슈는 기간에 따라 성격이 변화하게 되며, 이는 해당 이슈를 구성하고 있는 세부 이슈의 조합의 변화로 파악할 수 있다. 예를 들어 그림 5(b)는 "북한" 이라는 장기 이슈를 구성하는 세부 이슈가 Period1 ~ Period4를 거치면서 변화하는 경우를 보이고 있다. 이처럼 전체 기간의 이슈가 세부 기간별로 동적으로 변화하는 양상을 고찰함으로써, 사회적 이슈에 대해 보다 본질적인 이해의 폭을 넓힐 수 있을 것이다.



(a) Ecample of teaditional issue tracking



(b) Detailed issue of "North Korea"

그림 5. 이슈의 동적 변이과정 고찰을 위한 방법론의 개요
Fig. 5. Research overview - Investigating dynamic mutation process of issues

3.2.3 분석 데이터 및 실험 결과

이 연구는 2012년 7월부터 2013년 6월까지 1년의 기간 동안 국내의 한 뉴스 포털 사이트에 게재된 문서 중 53,739건의 뉴스 기사를 표본으로 추출하여 실험을 수행하였다. 구체적으로, 전체 기간을 3개월 단위로 나눈 총 4개의 기간을 대상으로 하여 기간별 이슈의 흐름을 분석하기 위한 실험을 수행하였다. 개별 기간에 대해 25개의 이슈를 도출하였으며, 각 이슈는 5개의 키워드로 기술하였다.

그림 6은 전체 기간에 대한 토픽 모델링 결과로 도출된 이슈 중 (드라마, 시청률, 연기, 시청자, 사랑)의 세부 기간별 이슈 흐름을 도식화한 결과이다. 본 이슈 흐름도에는 대응도가 15% 이상인 이슈들만 표시하였으며, 흐름도에서 진한 색으로 구별된 이슈는 대응도가 70% 이상인 핵심 이슈를 뜻한다. 전체 기간의 이슈인 (드라마, 시청률, 연기, 시청자, 사랑)에서 드라마의 시청률과 시청자의 사랑 등이 이슈가 되었음은 알 수 있지만, 이 이슈가 어떤 드라마 및 배우에 대한 내

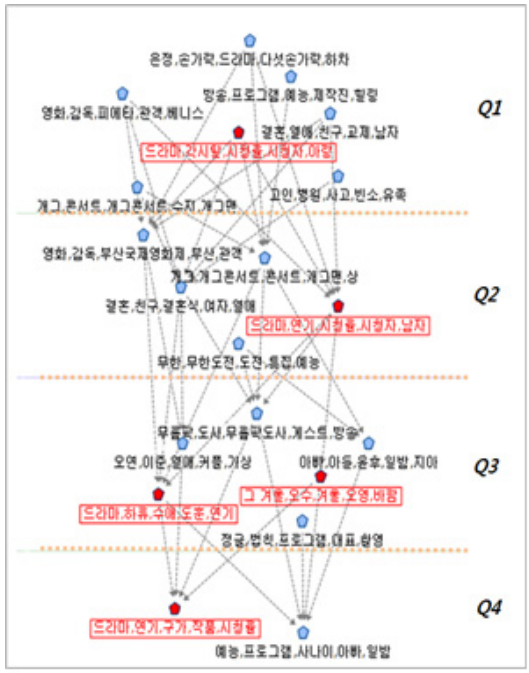


그림 6. 장기 이슈 중 (드라마, 시청률, 연기, 시청자, 사랑)에 대한 상세 이슈 흐름도
Fig. 6. Results of detailed issue flow diagram (Drama, Ratings, Performance, Viewer, Love)

용인지 구체적 내용은 파악할 수 없다. 하지만 제안 방법론을 통한 분석에서는 해당 이슈가 Q1 ~ Q4의 기간에 걸쳐 (드라마, 각시탈, 시청률, 시청자, 아랑), (드라마, 연기, 시청률, 시청자, 남자), (그 겨울, 오수, 겨울, 오영, 바람), (드라마, 하류, 수애, 도훈, 연기) 그리고 (드라마, 연기, 구가, 작품, 시청률)의 세부 이슈로 형성된 것임을 확인할 수 있다.

3.3 사용자 리뷰의 평가기준 별 이슈 식별 방법론: 호텔 리뷰 사이트를 중심으로[119]

3.3.1 연구 개요

많은 여가활동 정보 사이트는 각 상품에 대한 평균 평점 및 상세 리뷰를 제공함으로써 해당 상품에 관심이 있는 잠재고객의 의사결정을 지원하고 있다. 이들 사이트에서 각 평가 기준의 세부 항목에 대한 특징과 평가 기준별 주요 이슈를 파악하기 위해서는 고객이 직접 수많은 리뷰를 읽어야 한다는 불편이 따른다. 예를 들어 특정 호텔의 접근성이 4점이고 객실이 2점이라는 정보는 접근성 중 특히 지하철 역과의 거리, 객실 중 특히 욕실의 상태에 관심을 갖는 사용자에게 충분한 정보가 되지 못한다. 따라서 이 연구에서는 사용자 리뷰를 내용에 따른 평가 기준별로 자동 분류하고, 이에 기반을 두어 기준별 주요 이슈를 발굴하고 요약하는 방안을 제시하였다.

3.3.2 연구 동기

이 연구의 동기는 기존 여가활동 정보 사이트에서 제공되는 정보를 보다 구체화하여 제공하자는 것이며 그림 7을 통해 설명된다. 예를 들어 그림 7에서 사용자는 접근성 관점에서는 지하철역이 가까워야 한다는 요구사항을 갖고 있다. 그래프에서 해당 호텔의 접근성은 9.5점으로 매우 높게 나타났지만, 이 호텔이 지하철역과 가까워야 한다는 사용자의 요구사항을 충족시키지는 못한다. 따라서 이 호텔의 접근성이 높은 점수를 획득한 이유가 지하철역과 가깝기 때문인지, 아니면 호텔을 지나는 버스가 많거나 도심에 위치하기 때문인지를 파악하기 위해서는 상당수의 리뷰를 직접 읽어봐야 한다는 불편함이 있다. 따라서 각 평가 기준별 이슈를 요약하여 제시함으로써, 사용자에게 보다 구체적인 정보를 제공할 필요가 있다.

3.3.3 분석 데이터 및 실험 결과

이 연구는 글로벌 호텔 정보 사이트인 'H'사 사이트에서 2005년 8월에서 2016년 5월까지의 기간에 작성된 총 423건의 리뷰를 수집하여 실험을 수행하였다.

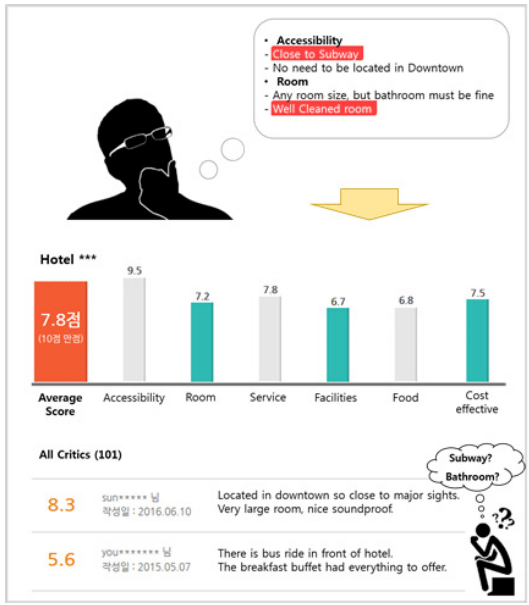


그림 7. 연구 동기 - 사용자 리뷰의 평가 기준별 이슈 식별
Fig. 7. Research motivation – Identifying issues of user reviews from the perspective of evaluation criteria

수집된 리뷰는 6개의 평가 기준에 대한 총 4,860개의 유닛으로 분리되었다. 이 연구는 하나의 문서를 분석 단위로 사용하는 기존의 토픽 모델링 연구와는 달리, 문서를 문장 단위로 분리하고 문장 단위의 분석을 수행하였다.

그림 8은 제안 방법론을 통한 활용 시나리오를 보이고 있다. 예를 들어 그림 8(a)의 요약 테이블을 통해 각 호텔의 평가 기준별 주요 이슈를 파악할 수 있으며, 특정 이슈를 선택함으로써 그림 8(b) 또는 그림 8(c)와 같이 해당 이슈에 대응되는 리뷰 원문을 조회할 수 있다.

3.4 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법론[120]

3.4.1 연구 개요

대중들은 일반적으로 특정 주제에 대한 정보 수집을 위해 SNS, 인터넷 뉴스, 블로그 등 여러 매체를 동시에 활용하고 있다. 하지만 다양한 매체를 통해 유통되는 문서들은 서로 유사한 내용일지라도 매체에 따라 상이한 카테고리로 분류되는 경우가 비일비재하다. 따라서 이 연구는 서로 다른 분류 체계를 갖는 이질적인 매체에 대해, 기존의 물리적 분류 체계를 그대로 유지하면서도 전체를 통합하여 관리할 수 있는 논리적 체계 구축을 위한 방안을 제안하였다. 또한 카테고

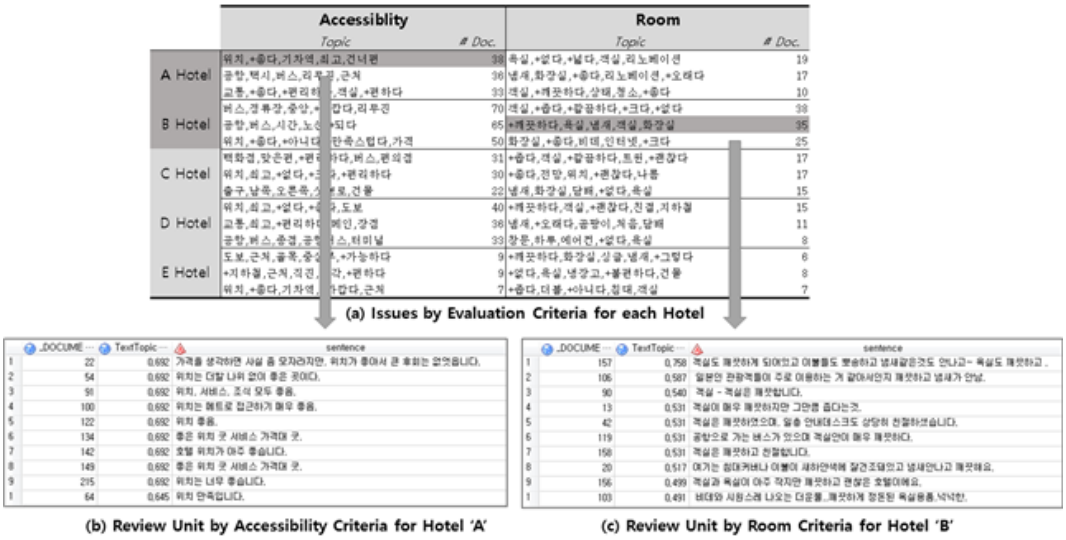


그림 8. 이슈 및 대응 리뷰
Fig. 8. Issues and corresponding reviews

리 부여 과정에서 분류의 정확도 향상을 위해 준지도 학습을 사용하였으며, 이 과정에서 기준 문서와 동일한 출처의 문서뿐 아니라 다른 출처의 문서를 학습 데이터 보강에 사용한 점은 매우 새로운 시도로 인정받을 수 있다.

3.4.2 연구 동기

이 연구의 동기는 유사한 내용을 다루고 있는 문서를 매체와 관계없이 동일한 카테고리에서 제공하자는 것이며, 그림 9를 통해 설명 가능하다.

그림 9는 "해외여행 어플리케이션"에 대한 글이 매체에 따라 "IT", "여행", "생활" 등으로 상이하게 관리되고 있는 현상을 보이고 있다. 이에 대해 기존의 물리적 분류 체계는 그대로 유지한 채 그림에서 화살표로 나타낸 것과 같은 논리적 매핑을 수행함으로써,

그림 9. 연구 동기 - 이종 매체 카테고리의 다중 매핑
Fig. 9. Research motivation - Mapping categories of heterogeneous sources

다양한 매체의 모든 문서들을 통일된 카테고리 기준 하에서 용이하게 접근할 수 있다.

3.4.3 분석 데이터 및 실험 결과

이 연구는 국내 포털 사이트 'N' 사와 'O' 사에서 배포된 인터넷 뉴스 기사 6,000개를 대상으로 실험을 수행하였다. 토픽의 수는 50개로 지정하였으며, 문서가 포함하는 용어 중 핵심 용어만을 문서의 구조화에 활용하기 위해 명사만을 분석에 사용하였다. 제안 방법론에 따라 개별 기사에 대해 매체와 카테고리 정보로 구성된 2차원 레이블을 부여했으며, 매체 간, 지도 학습과 준지도 학습 간, 동질 학습 데이터와 이질 학습 데이터 간의 정확도를 비교하였다. 그 결과 매우 흥미롭게도 일부 카테고리에서 이질 학습 데이터를 사용한 준지도 학습의 분류 정확도가 지도 학습 및 동질 학습 데이터를 사용한 준지도 학습의 분류 정확도보다 높게 나타나는 현상을 발견하였다.

3.5 텍스트 분석을 활용한 정보의 수요 공급 기반 뉴스 가치 평가 방안[56]

3.5.1 연구 개요

대부분의 새로운 정보는 뉴스 기사를 통해 대중에게 노출되고, 대중은 SNS 등 다양한 매체를 통해 이 기사에 대한 의견 또는 추가 정보를 표현함으로써 해당 정보를 확산시킨다. 이러한 측면에서 언론사에 의해 뉴스가 제공되는 행위를 정보의 공급으로 인식할

수 있고, 해당 정보에 대한 대중들의 SNS 활동은 그 정보의 소비 수요를 표출하는 것으로 이해할 수 있다. 이 연구에서는 정보 공급과 수요의 대표 매체로 인터넷 뉴스 기사와 트위터를 선정하고, 특정 이슈에 대해 뉴스가 갖는 정보로서의 가치를 이와 관련된 트윗의 양으로 평가하는 방안을 제시하였다. 언론사는 제안 방법론을 통해 가치가 높은 이슈를 식별하고 해당 정보의 생산에 집중함으로써, 수요에 부응하는 양질의 정보를 풍부하게 제공할 수 있을 것으로 기대한다.

3.5.2 연구 동기

이 연구의 동기는 뉴스 기사의 가치를 이와 관련된 트윗의 양으로 평가하고, 이를 통해 정보의 수급 불균형 현상에 대한 이해의 폭을 넓히는 것이다. 그림 10의 좌측과 같이 많은 정보가 인터넷 뉴스 등 언론을 통해 공급되며, 대중은 SNS 활동을 통해 이러한 정보에 대한 수요를 표출한다. 하지만 어떤 이슈는 뉴스를 통해 꾸준히 다루어지지만 대중에게 외면당하기도 하고, 반대로 어떤 이슈는 소수의 기사를 통해서만 다루어지지만 대중이 이에 대한 관심을 SNS를 통해 폭발적으로 나타내기도 한다. 따라서 어떤 이슈가 더욱 집중적으로 다루어질 가치가 있고 어떤 이슈가 불필요하게 많이 다루어졌는지를 각 이슈와 관련된 뉴스 기사의 수와 트윗 수의 비율에 기반을 두어 평가하는 것은 매우 의미 있는 시도라 할 수 있다.

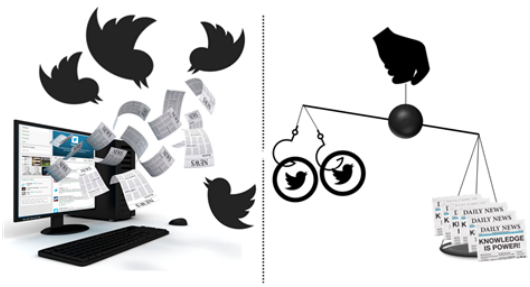


그림 10. 연구 동기 - 정보의 수요 공급 기반 뉴스 가치 평가
Fig. 10. Research motivation – Evaluating news value based on supply and demand of information

3.5.3 분석 데이터 및 실험 결과

이 연구의 실험은 2014년 6월 22일부터 2014년 7월 5일까지 2주의 기간 동안 국내 한 뉴스 포털 사이트에 게재된 뉴스 기사 387,014건과 해당 기간에 발생한 트위터 데이터 31,674,795건을 대상으로 수행되었다. 그림 11은 분석 결과를 나타내며, 그림에서 세로 축은 뉴스 당 트윗 수의 상대적 비율을 나타낸다.

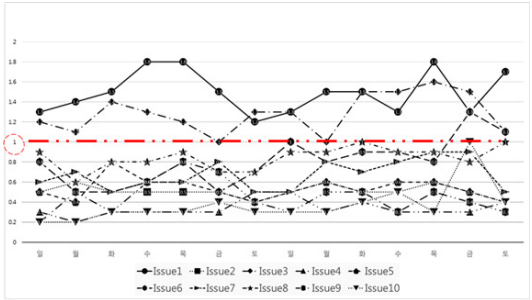


그림 11. 이슈별 NVI 추이
Fig. 11. Trend of NVI(News Value Index)

즉 전체 뉴스 수의 전체 트윗 수에 대한 비율을 기준 비율인 '1'로 설정하였으며, 기준 비율보다 높은 값을 나타내는 이슈는 뉴스 당 트윗 수가 평균보다 높음을 나타낸다. 실험 결과 매우 흥미롭게도 10개 중 8개 이슈에 대한 뉴스는 대중의 관심에 비해 과잉 공급되고 있으며, 단지 2개 이슈만이 뉴스 공급에 비해 대중의 높은 관심을 받고 있는 것으로 나타났다.

Ⅳ. 결 론

본 연구에서는 최근 관심 및 활용도가 급증하고 있는 텍스트 분석 관련 주요 기술 및 연구 동향을 살펴보고, 특히 토픽 모델링을 활용하여 다양한 분야의 문제를 해결한 최근 연구 사례를 소개하였다. 특히 벡터 공간모델, 용어 가중치, 차원 축소, 의미적 유사성 등 텍스트 분석의 핵심 기술 및 빈도 분석, 토픽 모델링, 문서 분류 등 다양한 활용 방법론을 요약하였을 뿐 아니라, 텍스트 데이터의 수집, 대상 문서 추출, 용어 사전의 활용 등 분석 결과의 품질에 영향을 끼칠 수 있는 다양한 요소를 실제 분석 관점에서 소개하였다는 점에서 본 연구의 의의를 찾을 수 있다. 본 연구에서 소개한 방법론의 활용을 통해 다양한 매체를 통해 빠르게 생성되는 방대한 양의 텍스트 데이터로부터 더욱 의미있는 지식과 통찰을 얻을 수 있을 것으로 기대한다.

References

[1] G. Salton, *The SMART retrieval system – experiments in automatic document processing*, Prentice-Hall, 1971.

[2] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing,"

Commun. ACM, vol. 18, no. 11, pp. 613-620, Nov. 1975.

[3] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Develop.*, vol. 1, no. 4, pp. 309-317, Oct. 1957.

[4] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 28, no. 1, pp. 11-21, Jan. 1972.

[5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. & Management*, vol. 24, no. 5, pp. 513-523, Dec. 1988.

[6] D. L. Lee, H. Chuang and K. Seamons, "Document ranking and the vector-space model," *IEEE Softw.*, vol. 14, no. 2, pp. 67-75, Mar. 1997.

[7] K. Pearson, "On lines and planes of closest fit to systems of point in space," *Philosophical Mag.*, vol. 2, pp. 559-572, 1901.

[8] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educational Psychol.*, vol. 24, no.6, pp. 417, Sept. 1933.

[9] I. Jolliffe, *Principal Component Analysis*, John Wiley & Sons, 2002.

[10] J. E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons, 2005.

[11] G. Saporta and N. Niang, *Principal component analysis: application to statistical process control*, Data Analysis, 2009.

[12] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Rev.*, vol. 35, no. 4, pp. 551-566, Dec. 1993.

[13] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, Oct. 1999.

[14] D. L. Lee, H. Chuang, and K. Seamons, "Document ranking and the vector-space model," *IEEE Softw.*, vol. 14, no. 2, pp. 67-75, Mar. 1997.

[15] M. B. Koll, "WEIRD: An approach to concept-based information retrieval," *ACM SIGIR Forum*, vol. 13, no. 4, pp. 32-50, Apr. 1979.

[16] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391-407, Sept. 1990.

[17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Research and Development in Inf. Retrieval*, pp. 50-57, Berkeley, USA, Aug. 1999.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Machine Learning Res.*, vol. 3, pp. 993-1022, Jan. 2003.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint,* arXiv:1301.3781, Jan. 2013.

[20] Y. Goldberg and O. Levy, "Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint*, arXiv:1402.3722, Feb. 2014.

[21] O. N. Park and H. J. Park, "A study on the international research trends in electronic records management: InterPARES 3 and ITrust achievements," *J. Records Management & Archives Soc. Korea*, vol. 16, no. 1, pp. 89-120, Feb. 2016.

[22] S. JU and M. S. Myoung, "The analysis of core contents in consumer area from 1st to 2009 revised middle school home economics textbooks," *J. Korean Home Econ. Edu. Assoc.*, vol. 27, no. 4, pp. 37-50, Dec. 2015.

[23] V. Ingle and S. Deshmukh, "Live news streams extraction for visualization of stock market trends," in *Proc. Int. Conf. Sign., Netw., Comput., Syst.*, pp. 297-301, New Delhi, India, Feb. 2016.

[24] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context-preserving, dynamic word cloud visualization," *IEEE Comput. Graphics Appl.*, vol. 30, no. 6, pp. 42-53, Nov. 2010.

[25] D. Scanfeld, V. Scanfeld, and E. L. Larson,

"Dissemination of health information through social networks: Twitter and antibiotics," *Am. J. Infection Control*, vol. 38, no. 3, pp. 182-188, Apr. 2010.

[26] W. Seo, H. Park, and J. Yoon, "An exploratory study on the korean national R&D trends using co-word analysis," *J. Inf. Technol. Appl. & Management*, vol. 19, no. 4, pp. 1-18, Dec. 2012.

[27] G. E. Heo and M. Song, "Examining the intellectual structure of a medical informatics journal with author co-citation analysis and co-word analysis," *J. Korean Soc. Inf. Management*, vol. 30, no. 2, pp. 107-225, Jun. 2013.

[28] S. Seo and E. Chung, "Domain analysis on the field of open access by co-word analysis," *J. Korean Biblia Soc. Library Inf. Sci.*, vol. 24, no. 1, pp. 207-228, Mar. 2013.

[29] B. Kang and J. H. Park, "Profiling and co-word analysis of teaching korean as a foreign language domain," *J. Korean Soc. Inf. Management*, vol. 30, no. 4, pp. 195-213, Dec. 2013.

[30] H. Choi and H. Varian, "Predicting the present with google trends," *Econ. Record*, vol. 88, no. 1, pp. 2-9, Jun. 2012.

[31] C. Graeme, "Googling the present," *The Labour Gazette*, vol. 4, no. 12, pp. 59-95, Dec. 2010.

[32] N. Askitas and K. F. Zimmermann, "Google econometrics and unemployment forecasting," *Appl. Econ. Quarterly*, vol. 55, no. 2, pp. 107-120, Jun. 2009.

[33] N. Khanh-Ly, B. J. Shin, and S. J. Yoo, "Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information," in *Proc. BigComp*, pp. 223-230, Hong Kong, China, Jan. 2016.

[34] R. Kaushik, S. A. Chandra, D. Mallya, J. N. V. K. Chaitanya and S. S. Kamath, "Sociopedia: An interactive system for event detection and trend analysis for twitter data," in *Proc. Int. Conf. Advanced Comput., Netw. Informatics*, pp. 63-70, Bhubaneswar, India, Sept. 2015.

[35] J. B. Yi, C. K. Lee, and K. J. CHA, "An analysis of IT trends using tweet data," *J. Intell. Inf. Syst.*, vol. 21, no. 1, pp. 143-159, Mar. 2015.

[36] H. Steinhaus, "Sur la division des corps materiels en parties," *Bull. Acad. Polon. Sci.*, vol. 4, no. 12, pp. 801-804, Oct. 1956.

[37] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics Probability*, vol. 1, no. 14, pp. 281-297, Berkeley, USA, Jun. 1967.

[38] K. Bae, J. Hwang, Y. Ko, and J. Kim, "A search-result clustering method based on word clustering for effective browsing of the paper retrieval results," *J. KISS : Software and Appl.*, vol. 37, no. 3, pp. 214-221, Mar. 2010.

[39] S. Jung, S. H. Lim, J. H. Jeon, B. M. Kim, and H. A. Lee, "Web search result clustering using snippets," *J. KISS : Databases*, vol. 39, no. 5, pp. 321-331, Oct. 2012.

[40] J. H. Kim, J. S. Lee, M. Lee, W. Kim, and J. S. Hong, "Term mapping methodology between everyday words and legal terms for law information search system," *J. Intell. Inf. Syst.*, vol. 18, no. 3, pp. 137-152, Sept. 2012.

[41] S. Han, "A comparative study on clustering methods for grouping related tags," *J. Korean Soc. Library Inf. Sci.*, vol. 43, no. 3, pp. 399-416, Sept. 2009.

[42] M. Steyvers and T. Griffiths, *Probabilistic topic models : Handbook of latent semantic analysis*, Lawrence Erlbaum Associates, 2007.

[43] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. National Academy Sci.*, vol. 101, no. suppl 1, pp. 5228-5235, Apr. 2004.

[44] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *J. Am. Statistical Assoc.*, vol. 101, no. 476, pp. 1566-1581, Jan. 2012.

[45] Q. Yao, Z. Song, and C. Peng, "Research on text categorization based on LDA," *Comput.*

Eng. Appl., vol. 47, no. 13, pp. 150-153, May 2011.

[46] S. Y. Yu, "Exploratory study of developing a synchronization-based approach for multi-step discovery of knowledge structures," *J. Inf. Sci. Theory Practice*, vol. 2, no. 2, pp. 16-32, Jun. 2014.

[47] M. Lim and N. Kim, "Investigating dynamic mutation process of issues using unstructured text analysis," *J. Intell. Inf. Syst.*, vol. 22, no. 1, pp. 1-18, Mar. 2016.

[48] S. A. Jin, C. E. Heo, Y. K. Jeong, and M. Song, "Topic-network based topic shift detection on twitter," *J. Korean Soc. Inf. Management*, vol. 30, no. 1, pp. 285-302, Mar. 2013.

[49] J. H. Park and M. Song, "A study on the research trends in library & information science in korea using topic modeling," *J. Korean Soc. Inf. Management*, vol. 30, no. 1, pp. 7-32, Mar. 2013.

[50] J. S. Oh, "Identifying research opportunities in the convergence of transportation and ICT using text mining techniques," *J. Transport Res.*, vol. 22, no. 4, pp. 93-110, Dec. 2015.

[51] S. T. Na, J. H. Kim, M. H. Jung, and J. E. Ahn, "Trend analysis using topic modeling for simulation studies," *J. Korea Soc. Simulation*, vol. 25, no. 3, pp. 107-116, Sept. 2016.

[52] J. Bae, N. Han and M. Song, "Twitter issue tracking system by topic modeling techniques," *J. Intell. Inf. Syst.*, vol. 20, no. 2, pp. 109-122, Jun. 2014.

[53] J. Bae, J. Son, and M. Song, "Analysis of twitter for 2012 south korea presidential election by text mining techniques," *J. Intell. Inf. Syst.*, vol. 19, no. 3, pp. 141-156, Sept. 2013.

[54] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 424-433, Philadelphia, USA, Aug. 2006.

[55] S. A. Jin and M. Song, "Topic modeling based interdisciplinarity measurement in the informatics related journals," *J. Korean Soc. Inf. Management*, vol. 33, no. 1, pp. 7-32, Mar. 2016.

[56] D. Lee, H. Choi, and N. Kim, "A method for evaluating news value based on supply and demand of information using text analysis," *J. Intell. Inf. Syst.*, vol. 22, no. 4, pp. 45-67, Dec. 2016.

[57] H. J. Hwang, H. R. Shim, and J. Choi, "Exploration of user experience research method with big data analysis : Focusing on the online review analysis of echo," *J. Korea Contents Assoc.*, vol. 16, no. 8, pp. 517-528, Aug. 2016.

[58] G. Kim and H. Yun, "Topic modeling approach to understand changes in customer perceptions on hotel services in seoul," *J. Korea Serv. Management Soc.*, vol. 17, no. 3, pp. 217-231, Sept. 2016.

[59] J. D. Park, "A study on mapping users' topic interest for question routing for community-based q&a service," *J. Korean Soc. Inf. Management*, vol. 32, no. 3, pp. 397-412, Sept. 2015.

[60] D. Jeong, J. Kim, G. Kim, J. U. Heo, B. W. On, and M. Kang, "A proposal of a keyword extraction system for detecting social issues," *J. Intell. Inf. Syst.*, vol. 19, no. 3, pp. 1-23, Sept. 2013.

[61] B. Noh, Z. Xu, J. Lee, D. Park, and Y. Chung, "Keyword network based repercussion effect analysis of foot-and-mouth disease using online news," *J. Korean Inst. Inf. Technol.*, vol. 14, no. 9, pp. 143-152, Sept. 2013.

[62] J. An, K. Ahn, and M. Song, "Text mining driven content analysis of ebola on news media and scientific publications," *J. Korean Soc. Library Inf. Sci.*, vol. 50, no. 2, pp. 289-307, May 2016.

[63] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. Advances in Neural Inf. Process. Syst.*, pp. 288-296, Vancouver, Canada, Dec.

2009.

[64] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proc. Eur. Conf. Inf. Retrieval*, pp. 338-349, Dublin, Ireland, Apr. 2011.

[65] C. Apté, F. Damerau, and S. M. Weiss, "Automated learning of decision rules for text categorization," *ACM Trans. Inf. Syst. (TOIS)*, vol. 12, no. 3, pp. 233-251, Jul. 1994.

[66] G. Weikum, "Foundations of statistical natural language processing," *ACM SIGMOD*, vol. 31, no. 3, pp. 37-38, Sept. 2002.

[67] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1-47, Mar. 2002.

[68] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proc. 7th Int. Conf. Inf. Knowledge Management*, pp. 148-155, Maryland, USA, Nov. 1998.

[69] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[70] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer Verlag, 1982.

[71] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Machine Learning*, pp. 137-142, London, UK, Apr. 1998.

[72] Y. H. Kang and Y. B. Park, "Design of automatic document classifier for IT documents based on SVM," *J. IKEEE*, vol. 8, no. 2, pp. 186-194, Dec. 2012.

[73] C. Lee, S. Lim, and H. Kim, "Korean semantic role labeling using structured SVM," *J. KIISE*, vol. 42, no. 2, pp. 220-226, Feb. 2015.

[74] D. D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," in *Proc. 3rd Annu. Symp. Document Anal. Inf. Retrieval*, pp. 81-93, Las Vegas, USA, Apr. 1994.

[75] J. H. Roh, H. Kim, and J. Y. Chang, "Improving hypertext classification systems through WordNet-based feature abstraction," *The J. Soc. e-Business Stud.*, vol. 18, no. 2, pp. 95-110, May 2013.

[76] K. P. Kim and Y. S. Kwon, "Performance comparison of naive bayesian learning and centroid-based classification for e-mail classification," *IE Interfaces*, vol. 18, no. 1, pp. 10-21, Mar. 2005.

[77] M. Kam and M. Song, "A study on differences of contents and tones of arguments among newspapers using text mining analysis," *J. Intell. Inf. Syst.*, vol. 18, no. 3, pp. 53-77, Sept. 2012.

[78] H. Lim and D. W. Kim, "Using mutual information for selecting features in multi-label classification," *J. KISS : Softw. Appl.*, vol. 39, no. 10, pp. 806-811, Oct. 2012.

[79] J. Yoon, J. Lee, and D. W. Kim, "Feature selection in multi-label classification using NSGA-II algorithm," *J. KISS : Softw. Appl.*, vol. 40, no. 3, pp. 133-140, Mar. 2013.

[80] J. S. Hong, N. Kim, and S. Lee, "A methodology for automatic multi - categorization of single - categorized documents," *J. Intell. Inf. Syst.*, vol. 20, no. 3, pp. 77-92, Sept. 2014.

[81] J. W. Seo, T. S. Shon, J. T. Seo, and J. S. Moon, "A study on the filtering of spam e-mail using n-Gram indexing and support vector machine," *J. Korea Inst. Inf. Security & Cryptology*, vol. 14, no. 2, pp. 23-33, Apr. 2004.

[82] I. Joe and H. T. Shim, "A SVM-based spam filtering system for short message service (SMS)," *The J. Korean Inst. Commun. Inf. Sci.*, vol. 34, no. 9, pp. 908-913, Sept. 2009.

[83] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Proc. AAAI Workshop on Learning for Text Categorization*, pp. 55-62, Wisconsin, USA, Jul. 1998.

[84] X. Jia, K. Zheng, W. Li, T. Liu, and L. Shang, "Three-way decisions solution to filter spam email: An empirical study," in *Proc. Int.*

Conf. Rough Sets and Current Trends in Comput., pp. 287-296, Chengdu, China, Aug. 2012.

[85] H. J. Kim, J. J. Jung, and G. S. Jo, "Spam - mail filtering system using weighted bayesian classifier," *J. KISS : Softw. Appl.*, vol. 31, no. 8, pp. 1092-1100, Aug. 2004.

[86] H. S. Lee, J. I. Cho, M. H. Jung, and J. S. Moon, "An approach to detect spam e-mail with abnormal character composition," *J. Korea Inst. Inf. Security & Cryptology*, vol. 18, no. 6A, pp. 129-137, Dec. 2008.

[87] S. Lee, "A splog detection system using support vector systems," *J. Korea Inst. Inf. Commun. Eng.*, vol. 15, no. 1, pp. 163-168, Jan. 2011.

[88] J. Jung and M. Yoo, "Tag search system using the keyword extraction and similarity evaluation," *The J. Korean Inst. Commun. Inf. Sci.*, vol. 40, no. 12, pp. 2458-2487, Dec. 2015.

[89] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Lang. Technol.*, vol. 5, no. 1, pp. 1-167, May. 2012.

[90] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL-02 Conf. Empirical Methods in Natural Lang. Process.-Volume 10*, pp. 79-86, Philadelphia, USA, Jul. 2002.

[91] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Computational Linguistics*, pp. 417-424, Philadelphia, USA, Jul. 2002.

[92] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in *Proc. 37th Annu. Meeting of the Assoc. Computational Linguistics on Computational Linguistics*, pp. 246-253, College Park, USA, Jun. 1999.

[93] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 168-177, Seattle, USA, Aug. 2004.

[94] H. Chen and D. Zimbra, "AI and opinion mining," *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 74-80, May 2010.

[95] N. Jindal and B. Liu, "Mining comparative sentences and relations," in *Proc. AAAI*, pp. 1331-1336, Boston, USA, Jul. 2006.

[96] J. Kamps, M. J. Marx, R. J. Mokken, and M. Rijke, "Using wordnet to measure semantic orientations of adjectives," in *Proc. LREC 2004*, pp. 1115-1118, Lisbon, Portugal, May 2004.

[97] S. M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proc. 20th Int. Conf. Computational Linguistics. Association for Computational Linguistics*, pp. 1367-1367, Geneva, Switzerland, Aug. 2004.

[98] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 8th Conf. Eur. Chapter of the Assoc. Computational Linguistics*, pp. 147-181, Madrid, Spain, Jul. 1997.

[99] X. Ding, B. Liu, and L. Zhang, "Entity discovery and assignment for opinion mining applications," in *Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1125-1134, Paris, France, Jul. 2009.

[100] E. Yu, Y. Kim, N. Kim, and S. R. Jeong, "Predicting the direction of the stock index by using a domain-specific sentiment dictionary," *J. Intell. Inf. Syst.*, vol. 19, no. 1, pp. 95-110, Mar. 2013.

[101] E. C. Dragut, H. Wang, P. Sistla, C. Yu, and W. Meng, "Polarity consistency checking for domain independent sentiment dictionaries," in *Proc. 50th Annu. Meeting of the Assoc. for Computational Linguistics: Long Papers-Volume 1*, pp. 997-1005, Jeju Island, Korea, Jul. 2012.

[102] S. Park, W. Lee, and I. C. Moon, "Efficient extraction of domain specific sentiment lexicon with active learning," *Pattern Recognition Lett.*, vol. 56, no. 15, pp. 38-44,

Apr. 2015.

[103] A. Go, R. Bhayani, and L. Huang, *Twitter sentiment classification using distant supervision,* CS224N Project Report, Stanford, 2009.

[104] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proc. 23rd Int. Conf. Computational Linguistics: Posters*, pp. 241-249, Beijing, China, Aug. 2010.

[105] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach," in *Proc. 20th ACM Int. Conf. Information and Knowledge Management*, pp. 1031-1040, Glasgow, UK, Oct. 2011.

[106] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Computational Sci.*, vol. 2, no. 1, pp. 1-8, Mar. 2011.

[107] J. S. Lim and J. M. Kim, "An empirical comparison of machine learning models for classifying emotions in korean twitter," *J. Korea Multimedia Soc.*, vol. 17, no. 2, pp. 232-239, Feb. 2014.

[108] Y. Jung, Y. Choi, and S. H. Myaeng, "Determining mood for a blog by combining multiple sources of evidence," in *Proc. IEEE/WIC/ACM Web Intell., Int. Conf.*, pp. 271-274, California, USA, Nov. 2007.

[109] F. Keshtkar and D. Inkpen, "Using sentiment orientation features for mood classification in blogs," in *Proc. IEEE NLP-KE 2009*, pp. 1-6, Dalian, China, Sept. 2009.

[110] K. H. Y. Lin, C. Yang, and H. H. Chen, "Emotion classification of online news articles from the reader's perspective," in *Proc. 2008 IEEE/WIC/ACM Int. Conf. Web Intell. and Intell. Agent Technol.-Volume 01*, pp. 220-226, Washington, USA, Dec. 2008.

[111] M. Nam, E. Lee, and J. Shin, "A method for user sentiment classification using instagram hashtags," *J. Korea Multimedia Soc.*, vol. 18, no. 11, pp. 1391-1399, Nov. 2015.

[112] T. N. Phan and M. Yoo, "Facebook fan page evaluation system based on user opinion mining," *The J. Korean Inst. Commun. and Inf. Sci.*, vol. 40, no. 12, pp. 2488-2490, Dec. 2015.

[113] Y. Kim and M. Song, "A study on analyzing sentiments on movie reviews by multi-level sentiment classifier," *J. Intell. Inf. Syst.*, vol. 22, no. 3, pp. 71-89, Sept. 2016.

[114] C. Lee, D. Choi, S. Kim, and J. Kang, "Classification and analysis of emotion in korean microblog texts," *J. KISS : Databases*, vol. 40, no. 3, pp. 159-167, Jun. 2013.

[115] J. W Hwang and Y. Ko, "A korean sentence and document sentiment classification system using sentiment features," *J. KIISE : Comput. Practices and Lett.*, vol. 14, no. 3, pp. 336-340, May 2008.

[116] J. An, J. Bae, N. Han, and M. Song, "A study of 'Emotion Trigger' by text mining techniques," *J. Intell. Inf. Syst.*, vol. 21, no. 2, pp. 69-92, Jun. 2015.

[117] J. Moon, I. Jang, Y. C. Choe, J. G. Kim, and G. Bock, "Case study of big data-based agri-food recommendation system according to types of customers," *The J. Korean Inst. Commun. Inf. Sci.*, vol. 40, no. 5, pp. 903-913, May 2015.

[118] D. Kim, W. X. S. Wong, M. Lim, C. Liu, N. Kim, J. Park, W. Kil, and H. Yoon, "A methodology for analyzing public opinion about science and technology issues using text analysis," *J. Inf. Technol. Serv.*, vol. 14, no. 3, pp. 33-48, Sept. 2015.

[119] S. Byun, D. Lee, and N. Kim, "A methodology for identifying issues of user reviews from the perspective of evaluation criteria: Focus on a hotel information site," *J. Intell. Inf. Syst.*, vol. 22, no. 3, pp. 23-43, Sept. 2016.

[120] D. Kim and N. Kim, "Mapping categories of heterogeneous sources using text analytics," *J. Intell. Inf. Syst.*, vol. 22, no. 4, pp. 193-215, Dec. 2016.

김 남 규 (Namgyu Kim)

1998년 2월 : 서울대학교 컴퓨터공학과 학사
2000년 2월 : 한국과학기술원 경영공학 석사
2007년 2월 : 한국과학기술원 경영공학 박사
2007년 3월~현재 : 국민대학교 경영정보학부 교수
<관심분야> 텍스트마이닝, 데이터마이닝, 데이터베이스

이 동 훈 (Donghoon Lee)

2010년 2월 : 한국방송통신대학교 컴퓨터과학과 학사
2012년 2월 : 국민대학교 비즈니스IT전문대학원 석사
2011년 8월~2016년 2월 : 솔트룩스 Product Development Group
2016년 3월~현재 : 국민대학교 비즈니스IT전문대학원 박사과정
<관심분야> 텍스트마이닝, 데이터마이닝, 온톨로지

최 호 창 (Hochang Choi)

2017년 2월 : 국민대학교 경영학부 학사
2017년 3월~현재 : 국민대학교 비즈니스IT전문대학원 석사과정
<관심분야> 데이터마이닝, 텍스트마이닝

William Xiu Shun Wong

2011년 9월 : Universiti Sains Malaysia 컴퓨터공학과 학사
2015년 2월 : 국민대학교 비즈니스IT전문대학원 석사
2015년 3월~현재 : 국민대학교 비즈니스IT전문대학원 박사과정
<관심분야> 데이터마이닝, 텍스트마이닝, 감성분석