연구논문/작품 중간보고서

2018 학년도 제 학기

제목	연구 동향 분석을 위한 텍스트 마이닝 기법 적용 분석 A Analysis of the Application of Text Mining for the Studies of Research Trends	○ 논문( ) 작품( ) ※해당란 체크
GitHub URL	https://github.com/CHOOJUNGMOOK/CHOO	
평가등급	지도교수 수정보완 사항	팀원 명단
	○ ○ ○	추 정 묵 (인) (학번: 2013311399)

2018년 9월 일

지도교수 : 김 응 모 서명

요약

최근 빅 데이터를 연구 및 분석에 활용하는 경우가 많아지고 있다. 따라서 다양한 분야에서 빅 데이터를 이용하기 위한 시도가 많아지고 있는 실정이다. 그 중에서 다양한 연구 분야의 연구 동향을 빅 데이터 분석을 통해 알아내기 위한 다양한 방법이 제시되고 있고, 그러한 기술 중 하나인 텍스트 마이닝이 주로 사용되고 있는 추세이다. 따라서 본 논문에서는 이러한 텍스트 마이닝의 기법을 간략히 소개한 후 연구 동향 분석에서 사용된 텍스트 마이닝 기술에 대해 분석해보고, 향후 활용 방안에 대해서 논의한다.

1. 서론

최근 정보통신 기술이 급격하게 발전하면서 인터넷 공간에서 문서를 비롯한 수많은 정보들이 대량으로 만들어지고 동시에 유통되고 있다. 그에 따라 인터넷 공간에서 만들어진 다양한 데이터들을 활용하려는 데이터 마이닝(data mining) 기술이 발달하게 되었다. 다양한 데이터 마이닝 기술 중에서 사회연결망서비스(sns : social netwok service)나 검색엔진, 온라인 쇼핑몰 등에서 방대하게 생산되는 텍스트 데이터를 활용하려는 텍스트 마이닝(text mining) 기술이 빠르고 다양하게 발전하고 있다. 계량화 되지 않은 비정형 데이터인 텍스트 데이터 자연어 처리 기술에 기반하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술인 텍스트 마이닝은 비정형의 자연어 문서를 대상으로 하는 데이터 마이닝의 한 분야로서, 자연어 처리를 통한 파싱, 형태소 분석, 품사 태깅, 관계 추출, 의미 추출과, 언어모델링을 통한 언어 감지, 규칙기반 개체명 인식, 기계 학습 알고리즘을 통한 습득 정보 사용능력 개선, 마이닝 기술을 통한 통계적 기법 활용 정보 분석 등 문서에 담긴 고급 정보들을 탐색 하는 분야이다. 이러한 텍스트 마이닝 기술은 상품 리뷰 텍스트를 분석하여 순위를 예측하거나, 비정형 데이터 수집을 통해 비즈니스 모델을 만들거나, 텍스트로부터 특정 감정을 추출하는 연구들에 활용되거나 [6], 연구 논문은 많은 부분이 텍스트 데이터로 되어 있다는 특징에 주목되어 다양한 분야의 연구 동향을 파악 및 분석하기 위해 활용되게 되었다. 한국교육학술정보원(RISS)에서 '텍스트 마이닝 동향' 검색 키워드를 제목과 본문에 포함하는 수도 200개가 넘는다. 이와 같이 텍스트 마이닝의 활용도 및 중요성이 더욱 높아지고 있음을 알 수 있다. 따라서 본 연구에서는, 우선 텍스트 마이닝의 다양한 기법들의 특징 및 구현 절차를 간단하게 소개하였다. 그 후 국내의 연구 동향 분석 분야 전반에 활용되는 텍스트 마이닝 기법들을 특정 분야에 국한되지 않고 분석하고 파악하여, 향후 텍스트 마이닝을 통해 연구 동향 분석을 할 때에 활용될 수 있도록 하였다. 본 연구의 구성은 2장

에서 주요 데이터 마이닝 기법들에 대해 소개하고, 3장에서 텍스트 마이닝 기법이 활용된 연구 동향 분석들을 살펴보고 정리한다. 마지막으로 4장에서는 이들을 바탕으로 향후의 연구 동향 분석 분야에서의 데이터 마이닝 기법의 발전 방향 및 전망에 대해 논의한다.

2. 텍스트 마이닝 기법

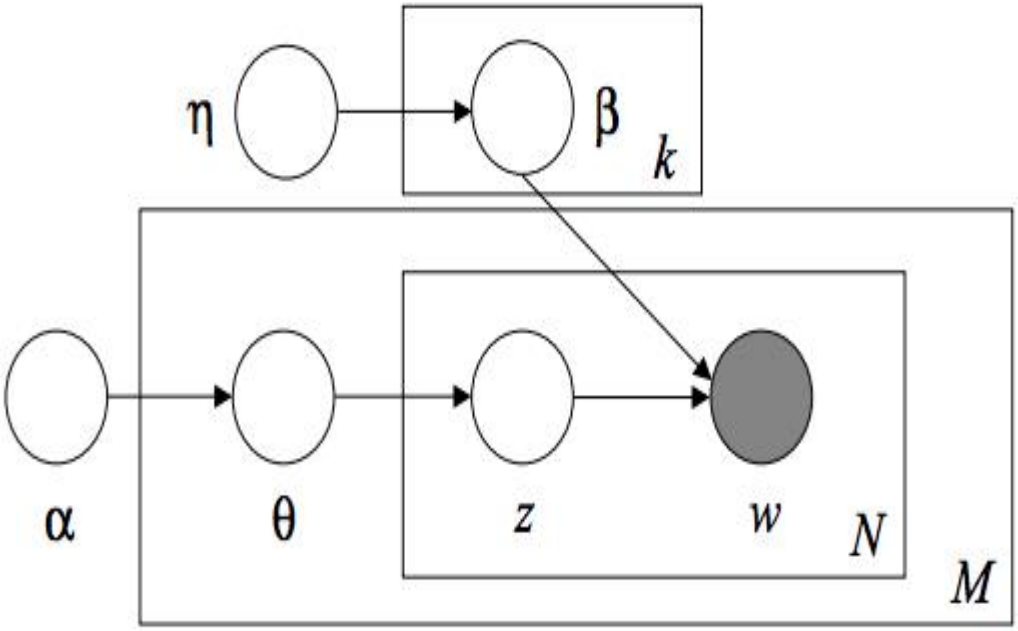
2.1 토픽 모델링 기법

토픽모델링은 방대한 양의 텍스트 문서 집합으로부터 중요 주제를 추출해주는 확률적 모델 알고리즘으로써, 주제에 대응되는 데이터를 분석하여 제공한다(Blei, Ng and Jordan 2003). 일반적인 군집화(clustering) 기법에선 하나의 텍스트 문서가 하나의 데이터로만 취급받는데 반해, 토픽모델링은 하나의 텍스트 문서가 여러 텍스트 문서에 동시에 대응 가능 하므로 현실 세계의 데이터 모델링에 적합하다는 특징을 가지고 있다. 특히 특정 분야의 연구 동향 분석에 주요 기법으로써 활용되고 있다 [3]. 대표적인 알고리즘으로는 LSA(Latent Semantic Allocation), pLSA(Probablistic Latent Semantic Analysis)와 같은 모델링 기법이 있고 Blei, Ng and Jordan(2003)이 고안한 LDA(Latent Dirichlet Allocation)모델링 기법이 기존 알고리즘의 정성적 한계를 보완한 방법으로써 현재 가장 널리 사용되고 있는 추세이다. 또한 이를 변형한, Teh et al.(2007)가 고안한 HDP(Hierarchical Dirichlet Process)모델링 방법 또한 주요 기법으로 사용되고 있다. 국내에서는 주로 LDA 토픽 모델링 기법을 활용한 연구들이 주로 진행되고 있다. 다음으로는 이렇게 가장 많이 활용되는 LDA 모델링 기법과 LSA, HDP 모델링 방법에 대해 살펴보도록 한다.

2.1.1. LDA

LDA(Latent Dirichlet Allocation)는 가장 대표적인 토픽모델링의 방법론으로 주어진 데이터 내의 이산적인 주제들에 대해 가정하는 확률 생성 모델이다. 문서 내부의 주제들의 확률분포와 각 주제를 구성하는 단어의 확률 분포를 통해 문서를 구성하는 주제를 확률적으로 택하고 그 주제에 존재하는 단어를 확률적으로 선택하는 샘플링과정을 반복함으로써 문서집합 내의 잠재된 토픽들을 찾아내는 기법이다 [1]. 2003년 Blei가 토픽 모델링 초기에 사용되던 잠재 의미 분석(LSA)과, 이를 변형한 확률기반 잠재 의미 분석(pLSA)알고리즘을 발전시켜 고안한 LDA 알고리즘을 발표하였고, 그 이후 토픽모델링의 주요 기법으로 LDA가 사용되고 있다 [3]. LDA 알고리즘을 통해 문서의 텍스트와 같은 관찰할 수 있는 변수를 이용하여 문서 데이터의 구조나 문맥과 같은 관찰할 수 없는 변수를 추론할 수 있고, 이를 활용하여

전체적인 문서의 집합들의 주제와 특정 단어들이 주제 안에 있을 확률, 문서들 안에서의 주제의 비율과 같은 데이터를 알아낼 수 있다. LDA 알고리즘의 모델은 <그림 1>과 같다.



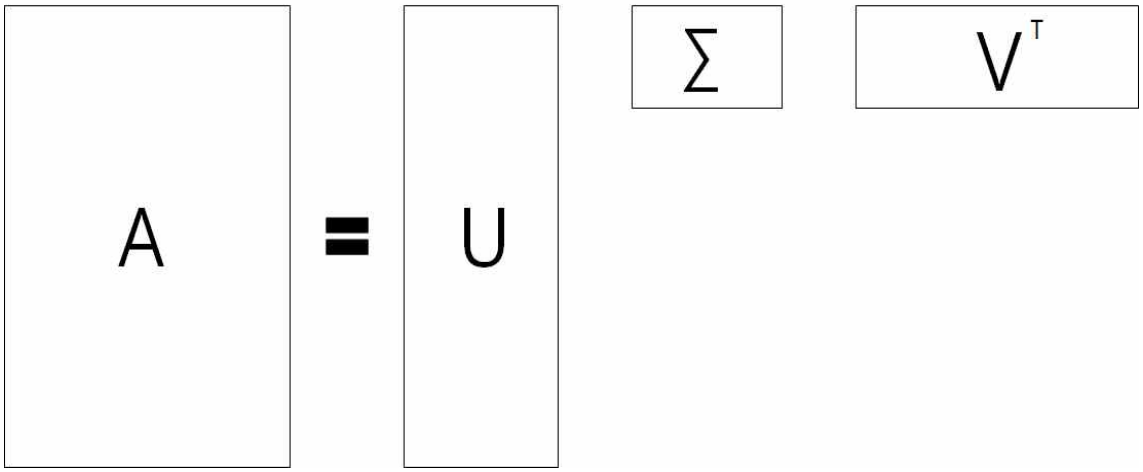
<그림 1> LDA 그래프 모델

<그림 1>에서 우선 k는 토픽의 개수이다. θ는 문서별 토픽의 비율(topic proportions)이고 α는 θ값을 결정하는 파라미터이다. β는 토픽별 단어 w의 생성비율(per-corpus topic distributions)이며 η는 β값을 결정하는 파라미터이다. z는 문서의 단어의 토픽이며, w는 문서의 단어로 문서에서 관측되는 변수(observed variable)를 의미한다 [3]. Dirichlet 분포를 따르는 θ는 각 문서집합에 대한 주제 비율로써 이 θ값에 따라 단어들의 토픽이며 문서 데이터 내에 존재하는 z가 결정된다. 그리고 단어 w는 각 토픽의 단어 생성 비율인 β값과 각 단어의 토픽을 의미하는 z값에 따라 결정 된다. 이와 같이 사전에 정한 변수들의 값에 따라 결과가 변하는 것이 LDA 기법의 특징이다. 그러므로 변수 값을 사전에 적절하게 정하는 것이 매우 중요하다. 또한 소량의 문서 데이터를 분석할 때에 낮은 효율을 보이기 때문에 그런 경우에는 다른 토픽 모델링 기법을 활용해야 한다는 한계점이 있다.

2.1.2. LSA

잠재 의미 분석(LSA: Latent Semantic Analysis) 기법은 초기에 고안된 토픽 모델링 기법으로써, 잠재 의미 색인 (Latent Semantic Indexing)으로도 불린다. LSA 기

법의 기반은 벡터 공간 모델이며, 문서 데이터의 집합 내에서의 단어들이 사용됨에 있어서 잠재적인 구조(Latent Structure)를 반드시 가진다고 가정하여 실행하는 기법이다 [7]. 단어들은 다의성(Polysemy)과 동의성(Synonymy)이라는 특징을 가지는데, 이러한 특성은 텍스트 마이닝에서 큰 장애물로 작용한다. 따라서 LSA 기법에선 문서 데이터 내의 단어들이 잠재적인 구조를 가지는 것이 필요하다. 다의성은 눈이라는 단어와 같이 하나의 단어가 다른 두 개 이상의 가지 뜻을 가지는 특성이며, 차와 자동차와 같이 다른 별개의 단어들이 동일한 의미를 가지는 특성이 동의성이다. 결국 LSA 기법은 문서 데이터 내부의 단어들의 동의성과 다의성이라는 장애물을 극복하고자 하려는 기법이라고 할 수 있다 [7]. 또한 LSA 기법은 문서 데이터의 문맥(context)를 파악하기 위해 잠재적인 구조를 파악하는데, 이를 위해서 단어들의 공기(Co-occurrence)성을 활용한다. 공기성은 하나의 문서에서 높은 빈도로 함께 단어들이 나타나는 관계를 말한다.



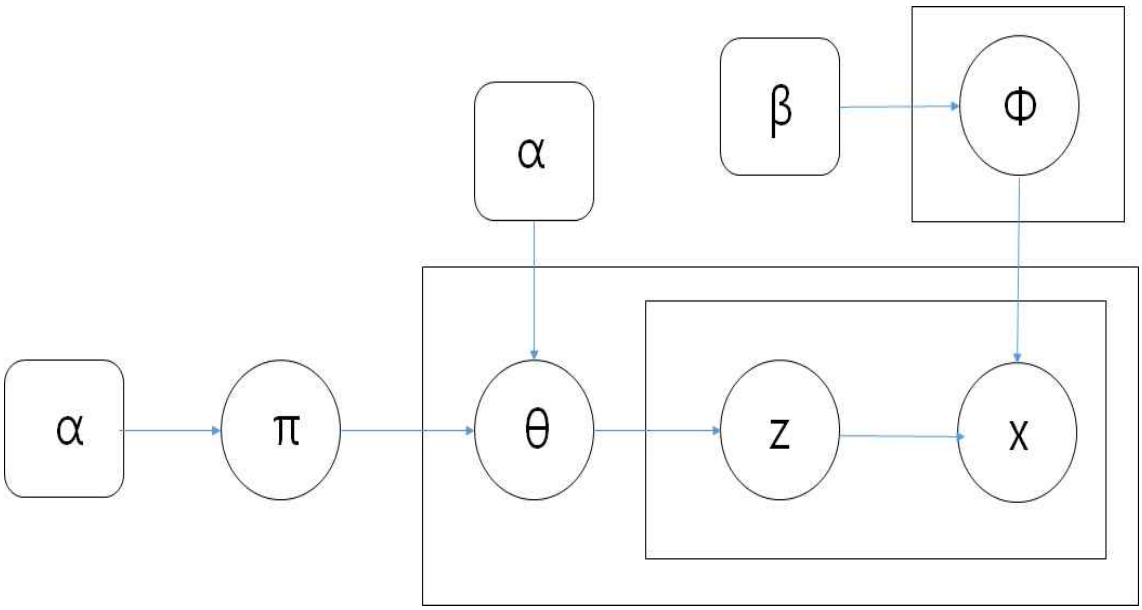
<그림 2> SVD기법

LSA의 알고리즘은 다변량 통계 분석(Multivariate Statistics) 기법의 하나인 특이값 분해(SVD: Singular Value Decomposition)를 활용한다. <그림 2>와 같이, SVD는 높은 차원의 A 매트릭스를 세 개의 매트릭스의 곱으로 분해하여 축소하는 기법이다. 따라서 차원을 나타내는 변수인 k 값을 미리 설정해야 하고, 이 값을 적절히 설정하는 것이 매우 중요하다는 특징을 가지고 있다.

2.1.3. HDP

HDP(Hierarchical Dirichlet Process)기법 랜덤 프로세스(Random Process)에 기반한 DP(Dirichlet Process)를 계층적으로 적용하는, Teh et al.(2007)이 고안한 토픽 모델링 기법이다 [8]. 2.1.1.절에서 소개한 LDA기법은 문서의 토픽의 개수 k는 미리 지정되며, 그 값을 기반으로만 데이터가 분포를 형성할 수 있다. 하지만 HDP기법

에선 모분포에 따라 토픽의 개수 k를 생성할 수 있어서 미리 설정하지 않아도 되는 기법이다. 따라서 사용자가 정한 k, α, β값에 따라 결과 값의 변동이 큰 LDA 기법과 다르게, HDP기법은 DP를 통해 적절한 k 값을 알아서 찾아낼 수 있다는 장점이 있다.

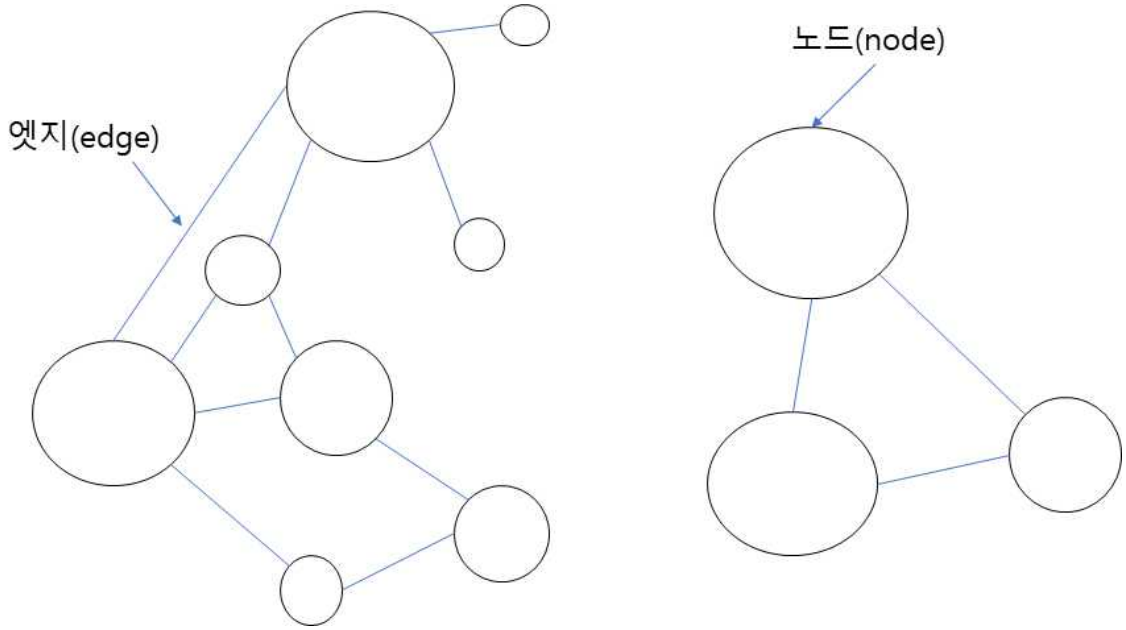


<그림 3> HDP 그래프 모델

HDP 알고리즘의 모델은 <그림 3>과 같다. <그림 3>에서 특정 문서에 대한 토픽의 분포는 노드 θ이며, 특정 토픽은 Φ이고 전체 토픽의 분포는 π로 나타난다. α는 θ값을 결정하는 파라미터이며 β는 Φ값을 결정하는 파라미터이고 각각 토픽과 단어의 사전 확률 분포를 나타낸다. z는 문서의 단어에 해당되는 토픽이다. 따라서 이 파라미터 값들을 통해 z와 θ를 추정하는 것이 목적이다. HDP 알고리즘은 토픽의 개수를 나타내는 k 값을 알아서 설정하기에, LDA 알고리즘과 달리 그래프 모델에 나타나지 않는 것을 볼 수 있다.

2.2 텍스트 네트워크 분석 기법

텍스트 네트워크 분석(Text Network Analysis)이란 텍스트 마이닝을 통해 얻어진 텍스트 데이터에 네트워크 분석을 적용한 것으로, 네트워크의 형태로 단어들 간의 관계를 나타내어 분석하는 기법이라 정의할 수 있다 [12]. 특히 텍스트로 이루어진 문서들을 구조적, 계량적으로 분석할 수 있게 해줌으로써 자료 분석의 신뢰성을 높일 수 있다. 텍스트 네트워크 분석 기술을 통해 주요 토픽과 단어의 관계를 네트워크의 형태로 시각적으로 확인할 수 있게 되고, 그 관계의 근접성을 바탕으로 주요 토픽을 군집화 하는 등의 분석도 가능하다. <그림 4>는 네트워크 구조를 간단히 나타낸 것이다. 네트워크는 네트워크 내의 개념들인 노드(node)와 개념들 간의 관



<그림 4> 네트워크의 구조

계를 나타내는 엣지(edge)로 구성된다 [4]. 노드 간 연결의 강도에 따라 엣지의 굵기가 변하고, 노드의 중심성에 따라 노드의 크기가 변한다. 텍스트 네트워크 분석에서 노드는 데이터의 토픽, 단어이며 그 단어들이 같은 문서에서 동시에 포함되어 있는 경우 엣지가 표시되며, 이렇게 동시에 연관된 단어가 출현하는 것을 특정한

관계를 가지는 동시성(co-occurrence)의 개념을 사용 한다 [10]. 텍스트 네트워크 분석은 두 가지의 큰 특징을 가지고 있다. 첫째, 특정 단어가 전체 네트워크에 대하여 어떤 영향력을 끼치는지 파악할 수 있다 . 텍스트 네트워크 분석을 통해 각 단어들의 영향력과, 서로에게 미치는 관계성을 시각적으로 표현하므로 단어를 포함한 텍스트 데이터에 대해 명확한 이해를 할 수 있게 하는 것이다 [9]. 둘째, 연관성이 높은 단어들이 어떤 군집을 형성하는지 파악 가능하게 된다. 네트워크 구조에서 군집들이 어떤 식으로 이루어져 있는가를 분석하는 것은 텍스트 데이터의 이해에 큰 영향을 준다. 또한 다른 텍스트 마이닝 기법에선 생략되는, 낮은 동시발생 군집들도 텍스트 네트워크 분석을 통해서는 파악할 수 있다 [9].

2.2.1 중심성 분석과 모듈성 분석

위와 같은 텍스트 네트워크를 분석하기 위한 방법론으로는 중심성(centrality) 분석과 모듈성(modularity) 분석이 있다. 중심성은 네트워크에서 각 노드의 중요도를 나타내는 지표로서 매개 중심성(betweenness centrality), 근접 중심성(closeness centrality) 등 다양한 지표가 있지만, 그 중 연결 중심성(degree centrality)에 대해

서만 간략히 설명한다. 연결 중심성은 노드 사이의 연결만을 고려하여 중요도를 산출하는 지표이다. 식 (1)은 연결 중심성을 구하는 식이며(Freeman et al, 1978) n은 모든 노드의 수이고 노드 i와 k가 연결된 경우 a의 값은 1이 된다. 이러한 중심성을 바탕 으로 문서 내의 단어가 어떤 관계를 갖는지 알아낼 수 있다.

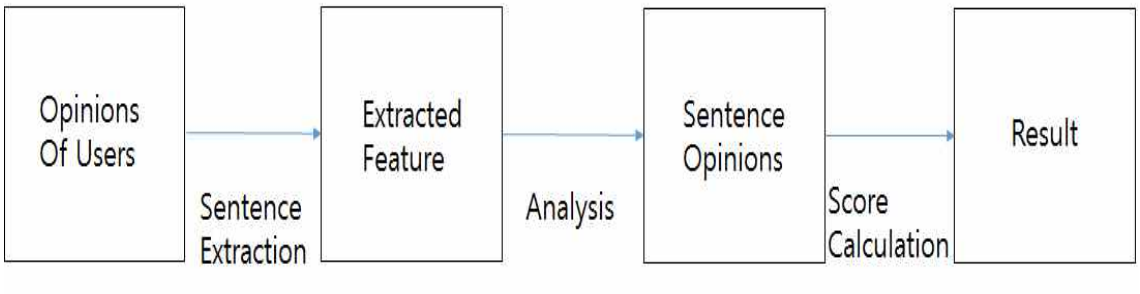
$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k) \qquad (1)$$

모듈성은 같은 말로 노드들의 군집을 뜻한다. 네트워크는 다양한 노드들의 군집으로 이루어져 있기 때문에 이러한 군집들의 경향을 파악하여 전체 네트워크의 흐름을 파악할 수 있다. 모듈성은 –1과 1사이의 값으로 1에 가까울수록 네트워크 내에 군집이 존재할 가능성이 높다는 것을 의미한다. 식 (2)는 모듈성을 구하는 식이다 (Blondel et al, 20008). A_{ij}는 노드 i와 j를 연결하는 엣지의 가중치를 뜻하며 k는 그 노드에 연결된 엣지의 가중치의 합이고 m은 $\frac{1}{2}\sum A_{ij}$ 와 같다 [9].

$$Q = \frac{1}{2m}\sum [A_{ij} - \frac{k_i k_j}{2m}]\delta(i, j) \qquad (2)$$

2.3 감성분석과 오피니언 마이닝 기법

감성분석(sentiment analysis)은 특정 문서 데이터의 긍정, 부정과 같은 극성에 대한 감정을 분석하고 분류하는 텍스트 마이닝 분석의 한 기법이다 [6]. 문서 내의 단어들의 감성극성(sentimentpolarity)을 최소 단위로써 사용한다. 문서에 대한 감정을 추출해내기 위해서 수집된 단어 데이터를 분류하는 단어들의 집합인 감성사전을 사용한다. 특정 단어가 감성사전에 정의되어 있다면 그 단어가 어떤 감정을 가진 것인지 판별하는 것이다. 하지만 단어는 동음이의어 등 판별에 어려움을 주는 특징들을 가지고 있기 때문에, 그를 위한 처리를 해주어야 한다. 그러한 처리 중에는 단어의 형태소를 분석하여 다른 단어와 연관성을 활용해 특정 극성을 어느 정도 가지고 있는지 분석하는 PMI 기술 등이 있다 [6]. 오피니언 마이닝(Opinion mining)은 감성 분석을 위한 기법 중 하나로써, 감성 사전을 통해 단어들 사이의 극성을 수치화 하여 극성을 분석하는 기법이다 [6]. 오피니언 마이닝은 세 가지 단계로 구성된다. 첫 번째 단계는 데이터에서 의견을 추출하는 것이고, 두 번째는 긍정과 부정의 극성을 분석하는 것이고, 마지막은 그로부터 최종적인 언어학적 결과를 정의하고 도출해내는 것이다 [1]. <그림 5>는 이러한 오피니언 마이닝의 흐름



<그림 5> 오피니언 마이닝의 흐름

을 간략히 나타낸 것이다. 이중에서도 가장 중요한 단계는 극성을 분석하는 부분의 감성사전의 질이며, 따라서 감성사전을 잘 구축하는 것이 오피니언 마이닝의 정확도를 올리는 방법이다. 또한 오피니언 마이닝은 상품 및 서비스에 대한 리뷰를 분류하는 데에 활발히 사용되고 있는데, 특정 상품에 대한 반응을 파악하여 기업은 그 상품에 장점을 극대화 하거나 단점을 보완할 수 있기 때문이다. 기존의 오피니언 마이닝은 BOW(Bag-of-words) 방법을 이용하는 n-gram 모델을 활용하였다. 식(3)은 n-gram 모델에서의 결합 확률(Joint Probability)를 구하는 식이다.

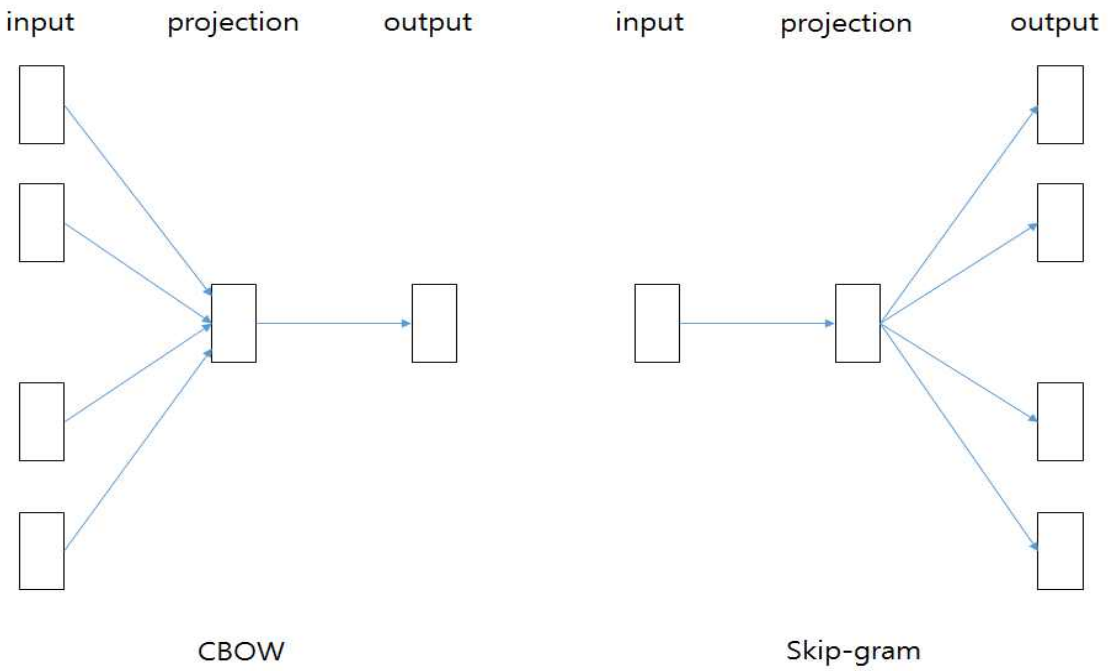
$$P(w_1,...,w_m) = \prod^{m} P(W_i|w_1,...,w_{i-1}) \qquad (3)$$

이러한 방법은 문장의 의미적 특성을 파악하지 못한다는 단점이 있었기에 최근엔 그러한 단점을 극복하여 벡터공간에 단어를 표현하는 벡터 공간 모델(Vecotr space model) 방법을 통한 워드임베딩(Word embedding)이 활발히 연구되고 있다. 벡터 공간 모델은 단어를 밀집된 형태로 단어를 나타내고, 의미적으로 비슷하면 가깝게 위치시켜 기존의 단점을 극복하며, 그 중 하나가 워드투벡(Word2vec)기법이다 [13].

2.4 워드투벡(Word2vec) 기법 [5]

word2vec은 2013년 구글의 Mikolov를 비롯한 구글의 연구원들이 작성한 "Efficient Estimation of Word Representations in Vector Space"에서 제안한 방법

을 알고리즘으로 구현한 것이다 [1]. word2vec은 특정 단어들 사이의 관계성을 잘 표현하는 특성을 가지고 있고, 그를 활용하여 단어들 사이의 관계를 식으로 나타낼 수 있다. word2vec의 원리는 텍스트 문서를 분석한 후 특정 단어에 대해 그 단어와 인접하여 나타나는 다른 단어를 관계성을 가진 단어로서 학습한다. 특정 단어들이 자주 인접하여 나타날수록 유사한 벡터 값을 가지도록 학습하는 것이다. LSA기법과 특정 단어가 벡터 값을 갖도록 분석하는 방식은 같지만, 특정 단어의 전후의 다른 단어들의 분포를 고려하여 학습한다는 점에서 차이가 있다 [1]. 기존의 LSA 기법에선 BOW, TF(Term frequency), TF-IDF(Term frequency-Inverse document frequency) 방법 등을 사용하였다 [13]. BOW 방법은 문서 텍스트 내에 특정 단어가 존재하면 1, 아니면 0으로 표현하고, TF 방법은 단어의 빈도수를 측정하여 표현하며, TF-IDF 방법은 단어의 중요도를 계산한다. 하지만 이러한 방법들은 단어 사이의 관계 파악이 힘들기 때문에, word2vec 기법에선 학습 알고리즘 모델로서 word2vec의 학습 알고리즘 모델은 CBOW(continuous bag of words)와 Skip gram을 사용한다 [1]. CBOW는 문서 텍스트에서 특정 단어의 주변단어를 이용하여 단어의 벡터를 생성하여 예측하는 알고리즘이고, Skip gram은 주어진 단어를 바탕으로 주변 단어의 벡터를 생성하여 예측하는 알고리즘이다. Word2vec기법은 분석결과로 각 단어에 대해서 최대 300차원의 벡터 값을 생성하기 때문에 기존의 기계학습 알고리즘을 적용할 수 있으며, 토픽모델링, 기계번역, 주가예측이나 추천시스템 등 다양한 분야에 있어서 활용이 가능하다 [1]. <그림 6>은 CBOW와 Skip gram 모델의 구조를 간략히 나타낸 것이다.



<그림 6> word2vec 기법의 알고리즘 모델

3. 텍스트 마이닝 기법을 활용한 연구 동향 분석

3.1 토픽 모델링 기법을 활용한 동향 분석

박자현, 송민은 1970년도부터 2012년도까지 발표 논문 초록을 수집하여 LDA기반의 토픽 모델링 실험을 수행하였고, 그로부터 도출된 연구주제를 문헌정보학 주제 분류표와 비교·분석하여 국내 문헌정보학 연구동향을 분석하였다 [14]. 박준형, 오효정은 1997년부터 2016년까지 발표된 기록관리학 관련 논문 1,027건을 대상으로 DA 토픽모델링과 HDP 토픽모델링을 수행하여 국내 기록관리학 내에 거시적으로 대표되는 주제와 세부 주제별 미시적인 핵심 키워드를 도출하였다 [3]. 신규식, 최회련, 이홍철은 2006년 1월부터 2015년 6월까지의 신재생에너지학 언론기사 51,558건을 대상으로 LDA 토픽모델링을 통해 국내의 신재생에너지학 분야의 이슈 동향을 분석하였다 [15]. 윤지은, 서창진은 2001년부터 2018년 4월까지 Scopus에 게재된 스마트 헬스케어 관련 연구 2,690편을 대상으로 토픽모델링과 에고 네트워크 분석을 통해 스마트 헬스케어 연구동향을 파악하였다 [16]. 황서이, 황동열은 예술경영 연구 동향 분석을 위해 '예술경영연구'에 수록된 2001년부터 2017년까지 총 286편의 논문명과 주제어를 분석대상으로 하여 LDA 토픽 모델링을 활용하여 예술경영분야 연구의 주요 키워드를 추출하였다 [17]. 진설아, 송민(2016)은 2009년도부터 2013년까지 정보학 분야 학술지 6,545개의 논문을 대상으로 LDA 토픽모델링을 통해 학술지의 학제성을 측정하였다 [18]. 김창식, 최수정, 곽기영은 정보시스템분야 저널인 Asia Pacific Journal of Information Systems, Information Systems Review, The Journal of Information Systems에 출판된 논문의 초록 1,245편을 토픽모델링 기법을 활용, 분석 하여 주요 키워드를 도출하였다 [29]. 윤소연, 윤동근은 재난 민 안전관리에 관한 연구 동향 분석을 위해 '재난', '안전'을 키워드로 검색하여 27년간(1990년-2016년) 배포된 뉴스기사를 수집하여 토픽모델링을 통해 주요 주제어를 도출하였다 [20].

위에서 나타낸 연구들을 살펴보면 토픽 모델링 기법을 활용한 연구들의 대부분은 LDA 기법을 활용하여 각 분야의 동향을 분석한 것을 알 수 있다. 또한 토픽 모델링 기법과 함께 네트워크 분석과 시계열회귀분석 등 다양한 기법을 함께 활용한 것도 알 수 있다. 하지만 토픽 모델링이라는 기법을 단순히 적용하여 연구 동향을 분석한 연구가 대부분이어서, 다양한 텍스트 마이닝 기법을 여럿 적용하여 그 장단점과 차이를 분석한 연구는 거의 없다는 특징이 있었다.

3.2 텍스트 네트워크 기법을 활용한 동향 분석

조재인은 7년간 문헌정보학분야에 게재된 논문 1,752건을 대상으로 빈도 분석 및 네트워크 텍스트 분석을 통해 여러 주제 개념의 분포와 그 관계성을 도출하였고 그를 통해 문헌정보학의 최근 연구 경향과 변화 양상을 분석하였고 특정 개념이 종전보다 더 다양한 주제 개념과 연관이 있다는 결과를 얻었다 [23]. 김학실은 한국행정학회에서 이루어진 발표문과 논문을 중심으로 네트워크 텍스트 분석 기법을 활용하여 여성정책변동에 따른 여성정책의 연구경향성을 분석하였다. 그 결과로 주요 핵심어들을 도출하였고, 정책변동과 여성정책 연구경향이 궤를 같이 하는 경향이 있음을 알아내었다 [24]. 최지원은 노동시장에서의 청년층 직업 및 고용 관련 연구 동향을 분석하기 위해 2010~2015년에 발간된 250편의 연구물 대상으로 네트워크 텍스트 분석을 실시하였고 그 결과로 노동시장에서 특정 연령과 고용간의 관계성을 분석하였다 [25]. 최영출, 박수정은 한국행정학의 연구경향을 분석하기 위해 2005년부터 2009년까지의 한국행정학보에 게재된 논문들의 초록을 대상으로 네트워크 텍스트 분석방법을 적용하여, 행정학의 주요 연구 대상을 도출하고, 네트워크 텍스트 분석 방법이 여타 행정 분야 연구에 대해서도 적용가능하다는 가능성을 발견하였다 [26]. 이미진, 오순영, 최경숙은 국내·외의 종양유전간호 관련 연구를 분석하여 연구동향을 파악하기 위해 종양유전간호 관련 주제로 학술지에 게재된 논문들의 초록에서 제시한 핵심어들을 중심으로 한 텍스트 네트워크 분석을 실시하였고 그를 통해 관심이 높은 특정 주제어를 도출해내었으며 간호사의 중재프로그램 개발 방향 설정에 활용될 수 있다는 가능성을 찾아내었다 [27]. 서건우, 구강본은 최근의 무도 연구에 대한 동향을 파악하기 위해 2001년부터 20015년까지 무도 분야 논문 총 205편을 네트워크 텍스트 기법을 이용해 분석하여, 1425개의 단어 중 무도라는 단어가 가장 많이 사용되었음을 도출하였다 [28].
위에서 나타낸 연구들을 살펴보면 토픽 모델링 기법을 활용한 연구들과 마찬가지로 네트워크 텍스트 분석 기법만을 통해 분석한 연구가 대부분을 차지하며 다른 종류의 텍스트 마이닝 기법과 비교 분석하는 연구는 부족하였다.

4. 결론

본 연구에서는 국내에서 각 분야의 연구 동향 분석에 주로 사용되는 텍스트 마이닝 기법들에 대해 간략히 설명하였다. 또한 각 기법들이 어떤 분야에서 활용되고 있는지 그 연구들에 대해 정리해보았다. 연구 결과 연구 동향 분석을 위해 키워드와 토픽분석을 통한 토픽모델링, 오피니언 마이닝, 감성분석,등의 기법이 주로 활용되고 있음을 확인할 수 있었다. 또한 분석자료로는 논문의 초록, 특허정보 등이 있었으며, 분석 분야로는 기록관리학부터 무도연구까지 다양한 분야에서 텍스트 마이

닝 기법이 활용됨을 알 수 있었다. 논문과 연구결과는 주로 텍스트로 이루어져있다는 특징을 가졌기에, 텍스트 마이닝 기법은 연구 동향 분석에 대해 매우 넓은 활용 가능성이 있다. 하지만, 토픽 모델링과 네트워크 텍스트 분석 등 몇몇 기술이 주로 사용되는 점과, 그러한 기술을 하나만 사용하여 함께 사용되었을 경우의 더 넓고 정확한 분석 결과를 얻지 못하고 있는 점 등이 한계점으로 남아있는 상황이다. 이러한 문제의 이유 중 하나로는 아직 영어에 비해 한국어로 된 텍스트 분석의 정확도가 상대적으로 부족하다는 점을 들 수 있을 것이다.
본 연구가 기여하는 바는 다양한 텍스트 마이닝 기법들을 소개하였고, 또 그를 활용한 다양한 연구를 소개하였기 때문에 앞으로 연구 분야에 따라 활용 하려는 텍스트 마이닝 기법을 판단, 선정 하는데 도움을 줄 수 있다.
본 연구의 한계는 다양한 텍스트 마이닝 기법을 소개하기 위해 깊은 전문성을 가진 내용을 담지 못했다는 점과, 같은 분야에서의 연구 동향 분석에 대해선 언급하지 않았기에 비교해볼 수 없다는 점이다.
후속 연구로는 해외 텍스트 마이닝 현황 및 영역, 목적 등을 조사하여 이를 국내 연구현황과 비교한다면, 국내 텍스트 마이닝 연구의 발전방향을 모색하는데 기여할 수 있을 것이라 생각된다.
향후 연구에서는 특정 연구 동향 분석에서 여러 종류의 텍스트 마이닝 기법을 활용하여 비교 및 분석을 진행한다면 이를 기반으로 더욱 발전된 연구 동향 분석 기술이 만들어 질 수 있을 것이라 생각된다.

■ 참고문헌

[1] 김성근, 조혁준, 강주영, "학술연구에서의 텍스트 마이닝 활용 현황 및 주요분석 기법", 정보기술아키텍처연구, 2016, p.317-329.

[2] 장재영, "텍스트 마이닝을 위한 그래프 기반 텍스트 표현 모델의 연구 동향", 한국인터넷방송통신학회, 2013, p.37-47.

[3] 박준영, 오효정, "토픽모델링을 활용한 국내 기록관리학 연구동향 분석", 전북대학교대학원, 2018.

[4] 김연경, 신선애, 송해덕 "네트워크 텍스트 분석(network text analysis)을 통한 국내 수행공학의 연구동향 탐색", 한국인력개발학회, 2017, p.35-64.

[5] 김윤덕, "Word2Vec을 이용한 위키피디아 텍스트 데이터 분석 시스템 구현", 숭

실대학교대학원, 2017.

[6] 최은정, 김동근 "오피니언마이닝을 이용한 사용자 맞춤 장소 추천 시스템", 한국정보통신학회논문지, 2017, p.2043-2051.

[7] 변성훈, "기능적 요구사항과 비기능적 요구사항 간의 추적성 구축에 있어서 잠재 의미 색인 기법과 잠재 디리클레 할당 기법의 비교 연구", 아주대학교대학원, 2017.

[8] 정영섭, "계층적 디리클레 프로세스를 활용한 미지원 도메인 검출", 한국컴퓨터정보학회 논문지 제23권 제1호(통권 제166호), 2018, p.17-24.

[9] 나채원, "텍스트 네트워크 분석과 AHP를 적용한 품질기능전개에 관한 연구"

[10] 고득환, 박세훈, "텍스트 네트워크 분석(TNA)을 통한 한국 비교교육 연구경향 분석", 교육종합연구소, 2017, p.127-152

[11] 김혜영, 이도길, 강범모, "사건명사의 공기어 네트워크 구성과 분석", 언어와언어학, 2011, p.81-106.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv preprint arXiv:1301.3781, 2013.

[13] 어균선, 이건창, "Word2vec을 이용한 오피니언 마이닝 성과분석 연구", 한국콘텐츠학회종합학술대회, 2018, p.7-8.

[14] 박자현, 송민, "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석", 정보관리학회지, 2013, p.7-32.

[15] 신규식, 최회련, 이홍철, "신재생에너지 동향 파악을 위한 토픽 모형 분석", 한국산학기술학회, 2015, p.6411-6418.

[16] 윤지은, 서창진,. "토픽모델링과 에고 네트워크 분석을 활용한 스마트 헬스케어 연구동향 분석", 한국디지털콘텐츠학회, 2018, p.981-993.

[17] 황서이, 황동열, "토픽모델링과 의미연결망 분석을 통한 예술경영 연구동향 분석", 예술경영연구, 2018, p.5-29.

[18] 진설아, 송민, "토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구", 정보관리학회지, 2016, p.7-32.

[19] 김창식, 최수정, 곽기영, "토픽모델링과 시계열회귀분석을 활용한 정보시스템 분야 연구동향 분석", 한국디지털콘텐츠학회, 2017, p.1143-1150.

[20] 윤소연, 윤동근, "토픽모델링을 이용한 재난 및 안전관리 동향 분석", 한국지형공간정보학회지, 2017, p.75-85.

[21] Blei, D.M., "Probabilistic Topic Models", Communications of the ACM, Vol. 55 No. 4, 2012, pp. 77-84.

[22] Ohana, B. and B. Tierney, "Sentiment Classification of Reviews Using Sentiwordnet", paper presented at the 9th. IT & T Conference, 2009.

[23] 조재인, "네트워크 텍스트 분석을 통한 문헌정보학 최근 연구 경향 분석", 정보관리학회지, 2011, p.65-83.

[24] 김학실, "여성정책변동과 연구경향 분석", 한국행정학보, 2012, p.241-264.

[25] 최지원, "네트워크 텍스트 분석을 통한 노동시장에서의 청년층 직업 및 고용 연구 동향 분석", 농업교육과 인적자원개발, 2016, p.183-207.

[26] 최영출, 박수정, "한국행정학의 연구경향 분석 : 네트워크 텍스트 분석방법의 적용", 한국행정학보, 2011, p.123-139.

[27] 이미진, 오순영, 최경숙, "유전종양간호 관련 연구경향", 한국콘텐츠학회논문지, 2018, p.47-56.

[28] 서건우, 구강본, "텍스트 네트워크 분석을 통해 살펴본 최근 무도연구 동향", 한국웰니스학회지, 2016, p.407-417.