토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구*

A Study of Research on Methods of Automated Biomedical Document Classification using Topic Modeling and Deep Learning

육지희 (JeeHee Yuk)**
송 민 (Min Song)***

초 록

본 연구는 LDA 토픽 모델과 딥 러닝을 적용한 단어 임베딩 기반의 Doc2Vec 기법을 활용하여 자질을 선정하고 자질집합의 크기와 종류 및 분류 알고리즘에 따른 분류 성능의 차이를 평가하였다. 또한 자질집합의 적절한 크기를 확인하고 문헌의 위치에 따라 종류를 다르게 구성하여 분류에 이용할 때 높은 성능을 나타내는 자질집합이 무엇인지 확인하였다. 마지막으로 딥 러닝을 활용한 실험에서는 학습 횟수와 문맥 추론 정보의 유무에 따른 분류 성능을 비교하였다. 실험문헌집단은 PMC에서 제공하는 생의학 학술문헌을 수집하고 질병 범주 체계에 따라 구분하여 Disease-35083을 구축하였다. 연구를 통하여 가장 높은 성능을 나타낸 자질집합의 종류와 크기를 확인하고 학습 시간에 효율성을 나타냄으로써 자질로의 확장 가능성을 가지는 자질집합을 제시하였다. 또한 딥 러닝과 기존 방법 간의 차이점을 비교하고 분류 환경에 따라 적합한 방법을 제안하였다.

ABSTRACT

This research evaluated differences of classification performance for feature selection methods using LDA topic model and Doc2Vec which is based on word embedding using deep learning, feature corpus sizes and classification algorithms. In addition to find the feature corpus with high performance of classification, an experiment was conducted using feature corpus was composed differently according to the location of the document and by adjusting the size of the feature corpus. Conclusionally, in the experiments using deep learning evaluate training frequency and specifically considered information for context inference. This study constructed biomedical document dataset, Disease-35083 which consisted biomedical scholarly documents provided by PMC and categorized by the disease category. Throughout the study this research verifies which type and size of feature corpus produces the highest performance and, also suggests some feature corpus which carry an extensibility to specific feature by displaying efficiency during the training time. Additionally, this research compares the differences between deep learning and existing method and suggests an appropriate method by classification environment.

키워드: 문헌 분류, 자질 선정, 텍스트 범주화, 토픽 모델, 딥 러닝, LDA, Doc2Vec, 텍스트 마이닝
document classification, feature selection, text categorization, topic model, deep learning, LDA, Doc2Vec, text mining

---

* 이 논문은 석사학위논문(2017년 12월)의 축약본임.
** 연세대학교 일반대학원 문헌정보학과(jeeheeyuk@gmail.com) (제1저자)
*** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2018년 5월 20일 ■ 최초심사일자 : 2018년 6월 11일 ■ 게재확정일자 : 2018년 6월 19일
■ 정보관리학회지, 35(2), 63-88, 2018. 〔http://dx.doi.org/10.3743/KOSIM.2018.35.2.063〕

1. 서 론

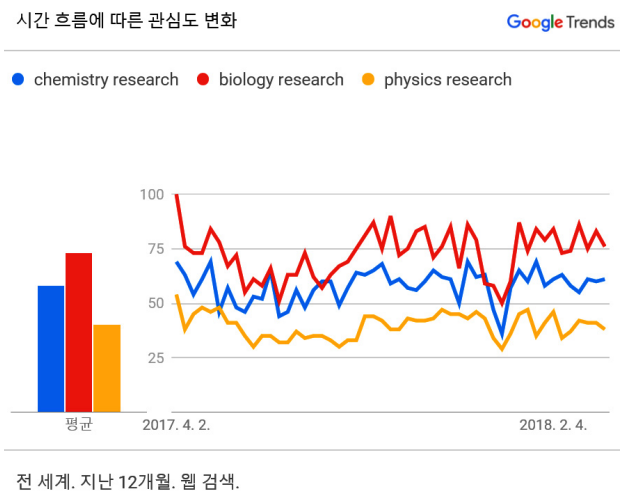
문헌자동분류의 연구 목적은 대량의 문헌을 보다 효율적으로 분류하는 방법을 제안하는 데 있으며 사전에 정의된 범주의 라벨을 각 문헌에 부여함으로써 문헌을 자동으로 분류하는 문헌범주화 과정을 포함한다. 최근 인터넷 환경에서 방대한 양의 텍스트 정보가 축적됨에 따라 이를 사전에 조직하고 분류하기 위한 자질선정 및 분류 기법에 관한 연구의 필요성이 지속적으로 제기되고 있다(Lilleberg, Zhu, & Zhang, 2015; Jiang, Lewris, Voltmer, & Wang, 2016; Bhushan, Danti, & Fernandes, 2017).

문헌으로부터 중요한 정보를 추출하는 텍스트 마이닝 기술의 발달은 단순 통계량을 이용한 이전의 방법보다 문헌의 내용에 근거한 자질을 추출할 수 있는 가능성을 제시하였다(Cohen & Singer, 1999). 자질선정은 분류 과정에서 문헌의 내용을 대표할 수 있는 색인어를 추출하는 단계이다. 자질선정 방법은 자질집합의 크기와 분류 알고리즘 등과 더불어 분류 성능에 많은 영향을 미치는 분류 성능 요소로써 중요하게 연구되어져왔다.

학술논문은 정형화된 텍스트이다. 학술논문의 구조적 특성을 고려하는 것은 텍스트 마이닝에 있어서 매우 중요하다(정영미, 2012). 핵심 정보가 문헌의 위치에 따라 구분될 수 있다는 가정에 근거하여 표제, 초록 등 학술논문의 위치를 기준으로 단어를 추출하고 자질로 선정하여 이용하려는 시도는 지속되어왔다.

표제와 초록은 학술논문을 이용한 문헌자동분류 실험에 가장 빈번하게 이용되어왔다. 전문(full-text)은 문헌의 내용으로부터 문맥적 요소를 반영한 의미론적 특성(semantics features)을 선정할 수 있으나 방대한 양의 데이터를 처리하는 과정에서 효율성이 저하될 수 있는 문제로 인하여 실험에 비교적 활발하게 이용되지 않았다. 다양한 텍스트 마이닝 기술의 발달과 처리 속도의 향상으로 보다 많은 양의 데이터 처리에 소요되는 시간이 절감되었다. TREC genomic에서는 텍스트 범주화 실험에 이용할 실험문헌을 생의학 학술문헌의 전문으로 구축하여 제시하였다. TREC 2004, TREC 2005에서는 세 개의 학술지(Journal of Biological Chemistry, Journal of Cell Biology, Proceedings of the National Academy of Science)로부터 2002년부터 2003년 사이에 출판된 총 11,880건의 학술논문을 수집하고 4개의 범주(alelle, expression, GO annotation, tumor)로 구분하였다. 그러나 TREC에서 제안한 이후로 생의학 학술문헌의 전문을 문헌분류실험에 이용할 수 있도록 구축한 연구는 미흡하다. MEDLINE/PubMed 데이터베이스에서 2017년 12월 검색 결과, 생의학 분야의 학술문헌은 약 2천 6백만(26,759,399)건의 인용이 일어났다. 또한 2017년 4월 2일부터 2018년 2월 4일 사이의 대표적인 자연 과학 분야인 화학, 물리학과 비교한 〈그림 1〉의 구글 트랜드 조사 결과, 연구가 지속적으로 활발하게 이루어지고 있다. 따라서 논문의 출판 속도와 활용도가 높은 생의학 학술문헌의 전문을 포함한 실험문헌 집단을 구축하여 생의학문헌의 자동분류 연구에 이용할 필요성이 있다.

최근의 문헌자동분류에 활용되는 대표적인 텍스트 마이닝 기법은 토픽 모델(topic model)과 딥 러닝(deep learning)을 적용한 단어 임베딩(word embedding) 기반의 Word2Vec 및 자



〈그림 1〉 자연 과학 분야 구글 학술 트랜드 조사 결과

질선정과 분류가 동시에 가능한 Doc2Vec 방법이 있다(Wang, Xu, Xu, Tian, Liu, & Hau, 2016; Atlig, Reyyan, & Yigit, 2017; Hughes, Li, Kotoulas, & Suzumura, 2017). 그러나 각각의 방법을 학술논문의 단어 출현 위치에 따라 구분하여 적용한 연구는 미흡하며 분류 성능에 상대적으로 큰 영향을 미치는 성능 요소를 심층적으로 분석하여 규명하지 못했다는 점에서 연구의 제한을 가진다. 또한 딥 러닝을 활용한 최근의 문헌분류 연구에서는 문헌의 특성에 따라 딥 러닝 이외의 방법이 분류의 효율성에 적합할 수 있음을 간과하였다. 최근 국내의 연구에서도 LDA(Latent Dirichlet Allocation) 토픽 모델과 딥 러닝을 적용한 분류 연구가 이루어지기 시작했다. 그러나 학술논문의 표제, 키워드, 초록과 본문을 대상으로 각 기법 간 성능에 영향을 미치는 요소를 단어의 출현 위치, 자질집합의 크기, 학습 방법에 따른 분류 알고리즘의 성능으로 기존 연구에 이용되어온 방법과 종합적으로 비교하고 평가한 연구는 매우 제한적이다.

본 연구의 목적은 국내의 문헌자동분류에 제한적으로 적용되어온 LDA 토픽 모델과 딥 러닝 방법인 Doc2Vec을 학술문헌의 구조에 따른 단어의 위치, 자질집합의 크기, 학습 방법에 따른 분류 알고리즘에 따라 구분하여 각 기법의 특징과 차이점을 심층적으로 분석하는데 있다. 이를 위하여 LDA 토픽 모델과 Doc2Vec을 이용하여 단순 통계량에 근거한 방법보다 의미론적 특성을 추출하고 자질집합의 크기를 조정하여 NB, k-NN, SVM, Ridge 분류 알고리즘을 이용한 성능을 살펴보고자 한다. 또한 딥 러닝 기반의 Doc2Vec을 이용하여 학습 횟수 및 차원 수에 따라 비교하고 특정 질병 단어를 기준으로 심층 학습 방법의 특징인 문맥 정보 제공의 유무에 따른 성능의 차이를 분석한다. 연구를 통하여 LDA와 Doc2Vec 기법의 특징과 차이점을 비교하고 학술문헌의 구조에 따른 단어의 출현 위치에 따라 표제, 키워드, 초록과 전문으로 구분하고 자질집합의 크기 및 학습 방법

에 따른 분류 알고리즘의 특성을 제시함으로써 향후 문헌자동분류 연구에 시사점을 제공할 수 있을 것으로 기대한다.

2. 이론적 배경

2.1 문헌 분류 연구 개관

문헌 분류 기법을 구성하는 기본 요소는 문헌 집단, 분류자질, 분류기이다(정영미, 2012). 문헌집단으로 이용되는 텍스트의 유형은 뉴스 기사, 학술문헌, 상품에 관한 리뷰, 회의록 등 다양하다. 자질선정의 목적은 범주 간 식별성이 높은 단어를 선정함으로써 분류 성능을 향상시키는 데 있다. 자질은 단어 또는 구문(phrase)으로 선정할 수 있으며 일반적으로 문헌의 내용에서 핵심이 되는 단어를 추출하여 구성한다(Harter, 1975). 자질을 선정할 때 고려되는 두 가지 중요한 요소는 문헌 식별성과 자질집합의 크기이다(Lewis, 1992). 문헌의 식별성이란 추출한 특성이 문헌의 내용과 문헌이 속하는 범주를 잘 대표할 수 있는지를 의미한다. 자질집합의 크기는 단어의 개수를 의미한다. 각 요소에서 논의되는 사항은 다음과 같다.

자질을 선정하기 위해서는 문헌의 의미론적 특성을 추출해야 한다. 이 과정에서 다양한 텍스트 마이닝 기법이 이용된다. 차원의 축소(dimension reduction)는 문헌분류 연구에서 논의되는 중요한 문제이다(John, Kohavi, & Pfleger, 1994). 자질집합의 크기를 줄이는 것을 차원의 축소라 하며 이는 문헌의 대표성을 훼손하지 않는 범위 안에서 이루어져야 한다. 분류 실험에 최적화(optimization)된 자질집합의 크기를 발견하는 두 가지 목적은 첫째, 분류 알고리즘의 복잡성 감소이며 둘째, 실험의 효율성 증진이다(Fuhr & Buckley, 1991).

일반적으로 문헌은 높은 차원으로 벡터 공간에 표현된다. 높은 차원으로 문헌을 표현할 경우 분류 알고리즘의 복잡성을 높이므로 분류 성능을 저하시킬 수 있다(Koller & Sahami, 1996). 또한 모든 단어가 문헌의 내용을 대표하는 자질로서 적합한 것은 아니다(Luhn, 1957). Mladenic과 Grobelnik(1999)은 모든 단어를 분류 자질로 이용하는 것보다 문헌의 내용을 대표할 수 있는 단어를 사용할 때 분류 정확도가 높음을 입증하였다. 각 문헌의 대표성을 가지는 자질을 선정하고 최적화된 자질집합의 크기를 확인함으로써 차원의 저주(curse of dimensionality)로 인한 성능 저하를 피할 수 있다.

김판준(2016)은 기계학습을 이용한 자동분류 연구에서 분류 성능 요소에 따른 차이를 분석하였다. 또한 자질선정이 학습 집합의 크기, 가중치 부여 기법 및 범주 할당 방법과 함께 가장 활발하게 연구되는 분류 성능 요소 중 하나임을 명시했다. 실험문헌집단은 국내 학술문헌의 제목과 초록을 수집하여 구축하였으며 별도의 자질 선정을 수행하지 않고 추출한 모든 자질을 실험에 사용하였다. 최상희, 이재윤(2012)은 정형화된 텍스트의 위치에 따라 문헌 클러스터링에 사용할 때 식별성이 높은 단어와 불용어를 선정하기 위하여 학술논문의 초록을 수집하고 분석하였다.

자질선정 방법은 분류 알고리즘의 특성에 따라서도 차이를 가질 수 있다. 이재윤(2005)은 문헌 간 벡터 유사도를 측정하고 SVM 분류기

의 성능 향상에 적합한 벡터 자질집합을 구성하는 방법을 제안하였다.

2.2 선행 연구

문헌의 특성을 추출하고 자질을 선정하는 방법은 학습방법에 따라 문헌빈도를 이용한 연구, 토픽모델을 이용한 연구, 딥 러닝 기반 단어 임베딩을 이용한 연구로 구분할 수 있다.

2.2.1 문헌 빈도를 활용한 문헌 분류

BOW(bag-of-words)는 문헌을 벡터 상에 표현하는 전통적인 방법이다. 이는 문헌 내에 단어의 출현 빈도만을 산술한 방식으로 단어의 순서 정보를 고려하지 않는 제한을 가진다(Hofmann, 1999).

기존에 자주 이용되어온 대표적인 방법은 단어빈도(term frequency, TF)와 역문헌빈도(inverse document frequency, IDF)를 이용한 문헌빈도 방법이다(Salton & McGill, 1983). 이 밖에도 정보획득량(information gain), 카이제곱 통계량(χ^2 statistic)과 역범주빈도(inverse category frequency)가 자주 이용되는 기법이다(Forman, 2003). 본 연구에서는 최근까지 분류 실험을 진행한 연구에서 성능 비교를 위하여 자주 이용되는 TF-IDF를 사용하여 다른 자질 선정 방법과의 차이를 비교하였다(Kusner, Sun, Kolkin, & Weinberger, 2015; Wadbude, Gupta, Mekala, Jindal, & Karnick, 2016).

2.2.2 토픽모델을 활용한 문헌 분류

토픽 모델은 텍스트나 이미지로부터 추출한 주제 정보로 대용량의 데이터를 올바른 범주에 분류하는 데 이용되고 있다. Blei(2012)의 확률 그래프 기반 LDA 모델이 현재 가장 빈번하게 이용되고 있다.

LDA 모델은 특성이 전체 범주 내에서 그리고 각 문헌 내에서 얼마만큼의 비율로 출현하는지 알 수 있다는 점에서 단어의 출현 빈도만을 고려하는 BOW의 대안으로 제시되어왔다(Torkkola, 2004; Wei & Croft, 2006; Wang & Manning, 2012). 이는 Deerwester, Dumais, Furnas, Landauer, Harshman(1990)이 제안한 잠재의미색인(Latent Semantic Indexing, LSI)보다 의미론적 특성을 선정할 수 있는 가능성을 제시했다.

Blei, Ng, Jordan(2003)은 LDA 모델을 이용하여 문헌을 표현하고 상위 토픽으로 순위화한 특성을 이용하여 자질집합의 크기를 최대 99.6%까지 축소하였다. 문헌에 등장한 모든 단어를 자질로 이용한 경우와 LDA 모델을 통하여 선정한 50개의 주제를 자질로 이용한 경우를 SVM으로 실험한 결과, LDA로 추출한 토픽을 자질로 활용할 때 pLSI 등 기존의 자질선정 방법보다 성능이 향상되었음을 보고했다.

Wang과 Quian(2008)은 LDA 토픽 모델을 이용하여 특성을 추출한 후 잠재의미색인과 NB, k-NN, SVM을 이용하여 실험하였다. 실험 결과, LDA로 추출한 특성을 SVM으로 분류한 경우에 기존의 방법보다 1-6% 가량 높은 성능을 나타냈다. 이외에도 Le와 Bernardi(2012)가 LDA를 이용하여 자질을 선정하고 SVM를 이용하여 질문을 분류하는 등 다양한 텍스트 유형을 대상으로 토픽을 분류자질로 활용한 연구가 이루어졌다.

그러나 토픽 모델은 문헌의 전체적인 주제

구조를 분석하는 데는 적합하지만 단어의 출현 확률에 근거한 추론 모델로서 단어의 문맥 정보를 활용하지 못한다는 제한을 가진다(Liu, Liu, Chua, & Sun, 2015). 예를 들어, 생의학 문헌에서 특정한 질병을 유발하는 유전자를 자질로 선정할 경우에 해당 유전자가 질병을 유발하는 원인인지, 억제하는지에 관한 정보는 고려하지 않는다.

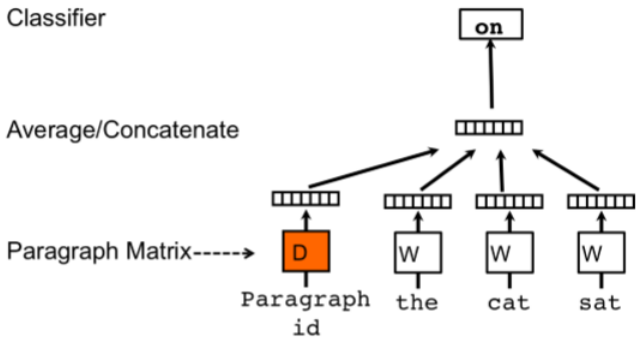
2.2.3 딥 러닝 기반의 단어 임베딩을 활용한 문헌 분류

단어 임베딩은 단어의 관계를 추론하여 실수 공간상의 벡터로 전사하는 기법이다. 이는 문장 속 단어들의 관계를 추론하며 유사한 의미의 단어를 유사한 벡터 값으로 표현한다. 단어 임베딩의 원리는 분포적 의미(distributional semantics)에 기반을 둔다. 특정한 단어의 의미와 주변단어의 분포가 관계성이 있음을 가정하며 각 단어 벡터 값의 연산을 통해 단어 간 관계를 만든다(Turian, Ratinov, & Bengio, 2010; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). 심층 신경망을 이용한 단어 임베딩 기반 학습은 Bengio, Ducharme, Vincent, Jauvin(2003)이 NNLM(neural network language model)을 제안한 이후 분류에 활발하게 적용되고 있다(Zhang & Zhou, 2006; Collobert & Weston, 2008; Tang, Qin, & Liu, 2015).

최근 Mikolov 외(2013)에 의해 제안된 Word2Vec은 단어 임베딩을 활용한 텍스트 및 이미지 분류, 자연어 처리, 질의응답 시스템 연구에 폭넓게 활용되고 있다. Word2Vec은 은닉층의 개수를 줄임으로 모델을 단순화시켜 학습 시간을 단축하고 단어 간의 관계를 추론할 수 있다. 학습에 이용되는 모델은 CBOW와 Skip-Gram 모델이 있으며, 이 중 대용량 문헌 처리 시 Skip-Gram 모델이 자주 이용된다.

Li, Wang, Zhang, Sun, Ma(2016)는 자질 추출에 LDA 토픽 모델과 Word2Vec을 결합한 방법을 제안하였다. 이 때 각 문헌의 토픽을 유클리디안 거리 분포로 표현하였으며 최적화된 자질집합의 크기를 확인하기 위하여 토픽의 개수를 조정하였다. Xing, Wang, Zhang, Liu(2014)는 LDA와 Word2Vec 모델을 이용하여 sohu 연구 센터에서 제공하는 IT, 중국 자동차 기사 등 9개의 범주에 해당하는 웹 문서의 분류 자질을 추출하였다. 기존의 연구와 달리 범주의 의미를 잘 표현할 수 있도록 문서가 속한 범주에 특화된 CSGMM(class-specific GMM)과 범주에 속한 각각의 문서 의미를 벡터 공간에서 잘 표현할 수 있도록 SSA(semantic space allocation) 모델을 개발하고 Word2Vec과 LDA를 이용한 실험과 분류 성능을 비교하였다.

Doc2Vec은 비정형화된 텍스트를 문장, 구문 또는 문헌 단위로 벡터 공간에서 표현하는 알고리즘이며 Paragraph Vector(Le & Mikolov, 2014)와 동일한 개념이다. Word2Vec이 단어 임베딩 모델로 제안되었듯이 Doc2Vec은 문헌 임베딩 모델(document embedding model)을 활용하여 고정된 길이의 벡터 공간 안에 문헌을 표현하는 기존 벡터 공간 모델의 제한점을 극복하였다. 이는 문헌을 대표할 수 있는 식별성이 높은 자질을 상대적으로 낮은 차원에 나타냄으로써 자질 크기로 인한 분류 알고리즘의 학습수행 저하의 문제를 해결한 분류 모델이다. Paragraph Vector 모델은 〈그림 2〉와 같은 원리로 자질 선정과 분류가 가능하다. 이는 자질



〈그림 2〉 Paragraph Vector Model (Le, Mikolov 2014, fig. 2)

선정 단계에만 이용되는 Word2Vec과 차이점을 가진다.

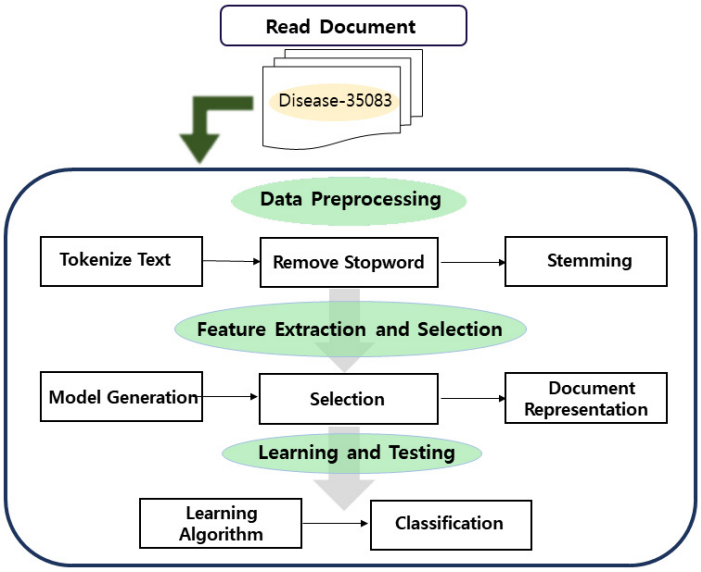
최근 Doc2Vec의 활용 가능성을 타진하는 연구가 이루어지고 있다(김도우, 구명완, 2017; Dai, Olah, & Le, 2015; Lau & Baldwin, 2016). Doc2Vec은 학습 속도와 자질 선정에 적용하기에는 실험적인 단계라는 이유로 Word2Vec에 비교할 때 상대적으로 빈번한 연구가 이루어지지 않았다. 그러나 자질 선정과 분류가 동시에 가능함으로써 연구에 소요되는 시간을 절감하고 심층 학습의 장점을 동시에 사용할 수 있다는 점은 Doc2Vec의 강점이며 높은 활용성이 기대되어 연구의 가치를 가진다.

Le와 Mikolov(2014)는 Paragraph Vector를 분류에 이용하고 BOW 방법으로 자질을 추출한 결과와 비교하였다. 실험 결과, NB와 SVM 복합 분류기를 bi-gram으로 학습했을 때 오류 확률이 8.78%(Wang & Manning, 2012)보다 낮은 7.42%로 나타났다. Jiang 외(2016)는 TripAdvisior와 Yelp의 리뷰 데이터로 Word2Vec과 Doc2Vec을 자질 추출 및 선정에 활용한 결과 유사한 성능을 나타내어 Doc2Vec이 자질 추출 및 선정 단계에 이용될 수 있는 가능성을 시사했다.

3. 연구 방법

3.1 실험 개요

본 연구는 실험을 통하여 LDA와 Doc2Vec 기법을 분류 성능 요소에 따른 차이점과 함께 분석하였다. 실험에서 살펴본 성능 요소는 문헌의 위치에 따라 구분한 자질집합의 종류 및 자질집합의 크기와 분류 알고리즘이다. 이를 통하여 문헌 분류자질 선정에 효율성을 가지는 자질선정방법과 분류자질로서 가치를 가지는 자질집합의 종류 및 적합한 크기를 확인하고 분류 알고리즘에 따라 이들 변인이 미치는 영향을 성능을 통해 평가하였다. 본 연구는 〈그림 3〉과 같은 순서로 진행되었다. 먼저 생의학 학술문헌을 세 가지 질병 범주로 구분하여 PMC로부터 수집하고 Disease-35083 실험문헌집단을 구축하였다. 또한 학습문헌의 위치에 따라 표제, 키워드, 초록과 전문으로 자질집합을 구성하고 크기에 따른 분류 성능에 미치는 영향의 차이를 확인하였다. LDA를 이용한 실험에서는 자질집합의 종류 및 크기에 따른 차이를 살펴보았다. 마지막으로 Doc2Vec을 이용한 실



〈그림 3〉 연구 절차

험에서 학습 반복횟수와 차원 수가 성능에 미치는 영향을 확인하였다.

3.2 연구문제

본 연구는 LDA, Doc2Vec 기법을 활용하여 생의학 학술논문의 자동분류에 영향을 미치는 성능 요소들의 특성을 규명하였다. 문헌의 위치에 따라 표제, 키워드, 초록과 전문을 나누어 각각 자질을 추출하였으며 실험에 이용한 학습 방법에 따라 NB, SVM 등의 분류기와 딥 러닝 방법인 Doc2Vec의 특성을 분류 성능의 차이를 통해 도출하였다. Doc2Vec의 특징을 구체적으로 살펴보기 위하여 학습 횟수와 차원 수, 문맥 정보 반영 요소를 중심으로 다각도로 분류 성능을 분석하였다. 실험을 통하여 아래의 세 가지 연구문제를 규명하고자 한다.

- 연구문제 1. 자질집합의 크기가 달라질 경우 학습 방법에 따른 분류 성능은 어떤 차이를 가지는가?
- 연구문제 2. 문헌 내 위치가 달라질 경우 학습 방법에 따른 분류 성능은 어떤 차이를 가지는가?
- 연구문제 3. 문맥 정보와 학습 횟수가 달라질 경우 심층 학습 방법을 이용한 분류 성능은 어떤 차이를 가지는가?

연구문제에 사용한 용어를 정의하면 다음과 같다. 연구문제 1의 자질집합의 크기란 분류자질의 개수를 의미한다. 연구문제 2의 자질집합의 종류란 학술 문헌으로부터 자질을 추출한 위치에 따라 네 가지로 구분한 자질집합을 의미한다. 구체적으로 표제, 저자가 부여한 키워드, 초록, 본문으로 구분된다. 연구문제 3의 문맥 정보란 질병 단어의 주변에 분포한 단어 간

관계의 추론 과정이 반영되었는지를 의미한다.

3.3 실험 절차

3.3.1 실험 문헌집단

본 연구에서 실험에 사용한 실험 문헌집단은 PubMed Central(PMC)에서 제공하는 암, 당뇨병, 에이즈 질병에 관한 학술지에 수록된 논문 30,869편을 대상으로, 학습문헌집합에 27,870편(90%), 검증문헌집합에 3,099편(10%)으로 구분하였다. PMC는 미국 국립 보건원 산하 미국 국립 의학 도서관(NLM/NIH)에서 운영하는 생의학 및 생명 공학 학술논문 전문을 제공하는 데이터베이스로 전문을 제공하는 학술지 2,104개와 4천 7백만 편의 학술논문을 보유하고 있다. Disease-35083의 범주는 암, 당뇨병, 에이즈 세 가지 질병이다. 범주를 질병으로 구분한 이유는 특정한 질병에 관한 생의학 연구가 빈번하게 이루어지고 있으며 검색 시 질병 명칭을 키워드로 사용한다는 점을 고려한 것이다. 암, 당뇨, 에이즈는 미국 내 질병 사망원인 또는 질병으로 지출되는 비용에서 높은 순위를 차지하고 있다. 또한 특정 질병에 관련하여 전문적으로 다루는 학술지가 다수 출판되어서 학습에 필요한 충분한 양의 데이터를 확보할 수 있다.

일반적으로 분류 실험에서는 모든 범주의 문헌을 사용하기보다는 실험 목적에 따라 학습할 범주를 선택한다. 동일한 문헌집단 내에서도 범주의 개수 및 실험 문헌의 수는 실험 목적에 따라 다르다. Lilleberg, Zhu와 Zhang(2015)은 20 newsgroup의 범주를 2개, 3개, 4개로 구분하여 실험한 결과 2개와 4개 간 정확도의 차이가 나타났으나, 3개와 4개 범주 간 성능의 차이는 크지 않았음을 확인하였다. Jiang 등(2016)은 총 5개의 범주를 가지고 TripAdvisor와 Yelp 리뷰 데이터를 실험에 사용할 때 0-3점의 review는 0, 4-5점에 해당하는 review는 1로 구분하여 각 2개 범주로 구성하였다.

전술한 연구의 목적은 자질선정 방법에 관한 성능을 비교하는 데 있다. 따라서 2개의 범주를 분류 실험에 사용하여 정확률이 높은 것과 관계가 없이 실험 결과 성능을 평가하여 우수한 자질선정 방법에 관하여 확인할 수 있었다.

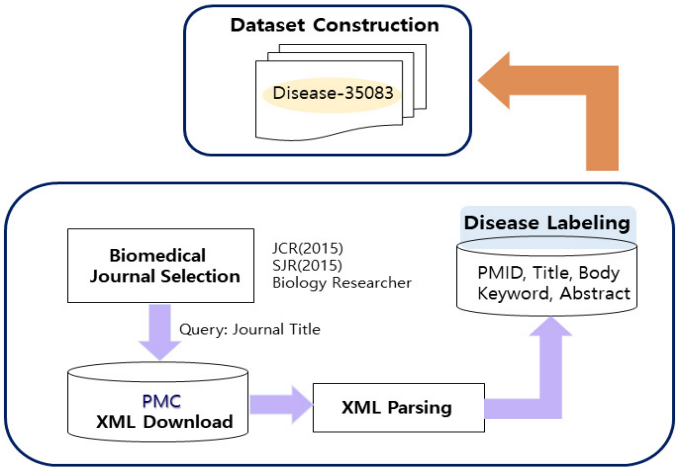
본 실험의 목적은 새로운 분류기를 개발하여 성능의 우수성을 입증하는 데 있지 않고, 기존에 이용되는 분류 알고리즘의 원리를 이용하여 구축한 분류기를 통하여 자질선정 방법에 따른 성능의 차이와 특성을 분석하는 것이다. 따라서 Disease-35083 문헌집단의 질병 범주 개수를 3개로 구분한 것은 실험의 목적에 부합하다고 판단하였다.

KEGG 질병 범주 체계에 근거하여 암, 당뇨병, 에이즈의 하위 질병분류체계를 고려하여 학술지 선정 시 목록에 포함할 수 있도록 하였다. 암에 관한 구체적인 질병 명칭은 C00-D48, 당뇨병은 E00-E90, 에이즈는 B20-B24에 해당하며 PMC에서 전문을 제공하고 2015 JCR(Journal Citation Reports)과 2015 SJR(Scientific Journal Rankings)에서 학술지 영향력(Impact Factor)을 기준으로 구성한 학술지 목록 중에서 생물학 전문 연구원의 검증을 통해 암, 당뇨병, 에이즈 질병에 관한 상위 10개의 학술지를 최종적으로 선정하였다. 선정한 학술지에 게재된 논문들은 모두 해당 질병에 관한 학술적인 가치를 인정받은 논문으로 판단하였다. 전체

범주별 학술지당 평균 논문 수는 1,028편이다. 실험문헌수집 및 평가용 문헌집단 구축 과정은 〈그림 4〉와 같다.

문헌 수집은 선정한 학술지의 전문을 PMC에서 XML로 수집하였다. 선정한 학술지에 수록된 모든 문헌을 가능한 많이 수집하기 위하여 문헌의 수집 연도에 제한을 두지 않고 제공하는 전체 연도를 대상으로 수집하였다. 이후 암, 당뇨병, 에이즈 질병 범주에 따라 수집한 문헌을 학습문헌집단과 검증문헌집단으로 분할하였다. 수집한 XML 파일에서 실제로 문헌의 내용이 존재하지 않을 경우 문헌집단에서 제외하였다. 〈표 1〉, 〈표 2〉, 〈표 3〉은 질병별 학술문헌 수집 결과이다. 학습에 사용한 문헌과 검증에 사용한 문헌의 규모를 범주별로 구분하여 〈표 4〉에 제시하였다.

〈표 1〉은 암 전문 학술지명과 PubMed Central에서 원문이 제공되는 연도범위 및 문헌 수에 관한 내용이다. KEGG 질병 분류 체계에 근거하여 암의 하위 질병분류체계를 고려하였다. 상위 질병인 Cancer 이외에도 Oncology 등 대표적으로 암에 관련하여 다루는 학술지를 목록으로 선정하였다. 수집 결과, 총 16,414건의 문헌이 수집되었으며 10개의 학술지가 모두 암에 관한 학술문헌으로 구성되기 때문에 학술 지간 문헌 수의 편차는 고려하지 않았다. 〈표 2〉는 당뇨병에 관한 학술지명과 문헌 제공연도 및 수집 건수를 기술한 것이다. Obesity(비만)는 질병이 아닌 질병을 유발하는 상태를 지칭한다. 비만은 당뇨병을 일으키는 원인이 되는 신체적 상태이며 당뇨병 연구에서 빈번하게 다루어지고 있다. 따라서 당뇨병과 관련된 연구가 다수 게재되고 인용력 지수가 높은 학술지는 수집 목록으로 선정하였다. Endocrinology(내분비학)과 Metabolism(물질대사, 대사량)은 생물체 내에서 일어나는 호르몬 등 물질의 분해나 합성과 같은 물질적 변화에 관한 내용을 다룬다. 당뇨병에 관하여 수집된 문헌 수는 총 9,604건이다. 〈표 3〉은 에이즈에 관한 학술지 목록과 수집한 문헌 제공연도 및 문헌 건수를



〈그림 4〉 실험문헌 수집 및 평가용 문헌집단 구축 과정

〈표 1〉 암 범주 데이터 기초 통계

번호	학술지명	제공연도	논문 수
1	Nature Reviews Cancer	2010-2016	207건
2	Cancer Cell	2002-2016	770건
3	Journal of Clinical Oncology	1984-2016	2,295건
4	Journal of the National Cancer Institute	1992-2016	1,175건
5	Cancer Research	1985-2017	4,051건
6	Clinical Cancer Research	1995-2016	2,835건
7	Leukemia	2001-2016	924건
8	Oncogene	1996-2016	2,262건
9	American Journal of Cancer Research	2010-2017	814건
10	Molecular Cancer Therapeutics	2003-2017	1,081건
합계	학술지당 평균 논문 수(1,641건)	수집범위(1984-2017)	16,414건

〈표 2〉 당뇨병 범주 데이터 기초 통계

번호	학술지명	제공연도	논문 수
1	Endocrine Reviews	2002-2016	133건
2	Trends in Endocrinology and Metabolism	2006-2016	24건
3	Diabetes Care	1998-2016	4,053건
4	Diabetes	1990-2016	3,114건
5	Obesity Reviews	2001-2016	113건
6	Diabetologia	2009-2016	897건
7	World Journal of Diabetes	2010-2017	447건
8	Diabetes Therapy	2010-2017	251건
9	BMJ Open Diabetes Research & Care	2013-2017	212건
10	Diabetes Technology & Therapeutics	2003-2017	360건
합계	학술지당 평균 논문 수(940건)	수집범위(1990-2017)	9,604건

〈표 3〉 에이즈 범주 데이터 기초 통계

번호	학술지명	제공연도	논문 수
1	AIDS and behavior	1997-2017	1,208건
2	AIDS (London, England)	1993-2017	1,317건
3	The Open AIDS Journal	2007-2017	70건
4	AIDS care	2002-2016	713건
5	Journal of the International Association of Providers of AIDS Care	2012-2016	62건
6	Current HIV/AIDS reports	2006-2016	152건
7	AIDS Research and Therapy	2004-2017	435건
8	AIDS Patient Care and STDs	2004-2017	490건
9	HIV/AIDS (Auckland, N.Z.)	2009-2017	172건
10	AIDS Research and Treatment	2009-2017	232건
합계	학술지당 평균 논문 수(495건)	수집범위(1993-2017)	4,951건

기술하였다. 에이즈는 질병의 특성 상 발병 이후의 치료 과정에 관한 연구가 다수 진행되었다. 질병의 치료법과 발병 환자에 대한 치료를 다루는 학술지가 상위 영향력 지수를 가지는 현상도 다른 질병과의 차이점이다. 또한 에이즈와 동일하게 다루어지고 있으나 다른 질병 명칭으로 HIV가 있다. 에이즈 학술지 목록을 선정할 때 HIV를 사용한 학술지도 포함하였다. 수집된 총 문헌 수는 4,851건이다.

〈표 4〉는 수집한 문헌을 학습에 9, 성능 평가에 1로 배분한 결과이다. 분류 실험을 위한 문헌의 벡터 표현은 분류 결과 값, 자질과 자질 값의 형태로 나타난다. 분류 결과 값이란 문헌이 각 범주에 할당된 값을 의미한다. Disease-35083 실험문헌집단의 범주는 3가지이다. 따라서 분류 결과 값을 Cancer는 0, Diabetes는 1, AIDS는 2의 라벨을 부여하여 각 범주의 분류 결과 값을 확인하여 옳은 범주에 분류가 이루어졌는지 구분할 수 있도록 하였다.

3.3.2 텍스트 전처리

먼저 수집한 문헌을 대상으로 SAX 파서를 이용하여 실험에 사용할 XML 요소를 추출하였다. SAX 파서는 대용량 문헌 처리에 이용되는 방법으로 문서 전체를 메모리에 올려 처리하는 DOM 파싱과 달리 스트림 방식으로 문헌을 읽으면서 추출 대상인 XML 요소만을 이벤트 방식으로 처리하기 때문에 비교적 적은 공간의 메모리로 처리가 가능하다는 이점을 가진다. 추출한 XML 요소는 PMID, 학술지명, 키워드, 표제, 초록과 본문으로 총 6가지이다. XML 추출 과정에서 본문 및 키워드가 제공되지 않은 문헌은 실험문헌집단에서 제외하였다. 〈표 5〉는 수집한 문헌으로부터 추출한 XML 요소를 기술하였다.

〈표 5〉 PMC 문헌 XML 추출 요소 정의

XML 요소명	정의
〈journal-title〉	학술지명
〈article-id pub-id- type =“pmid”〉	논문 PMID
〈kwd-group〉	논문 키워드
〈article-title〉	논문 표제
〈abstract〉	논문 초록
〈body〉	논문 전문

추출한 텍스트는 Java 프로그래밍 언어 기반의 형태소 분석기인 Stanford CoreNLP(Manning, Surdeanu, Bauer, Finkel, & McClosky, 2014)를 이용하여 전처리하였다. Stanford CoreNLP는 단어의 품사 식별이 가능하고, 불용어 제거 및 단어의 원형 복원을 가능하게 한다. 문장 분할 단계를 거친 후 토큰화한 텍스트를 대상으로 특수문자와 숫자를 제거하였다. 이후 불용어 목

〈표 4〉 Disease-35083 실험 문헌집단 통계

번호	범주	학습문헌 수	검증문헌 수	총 문헌 수
1	Cancer	14,772	1,642	16,414
2	Diabetes	8,643	961	9,604
3	AIDS	4,455	496	4,951
총 문헌 수		27,870	3,099	30,969

록을 제거하고 포터 스테밍 알고리즘을 이용하여 스테밍 작업을 진행하였다. 스테밍 작업이란 단어의 어간을 추출하여 동일한 의미를 가지나 단수, 복수 등 형태가 다른 단어를 하나의 의미로 인식하도록 처리하는 작업이다.

본 실험에서 진행한 텍스트 전처리는 분류 실험을 위한 자질집합을 구성하는 전 단계에 수행하였다. 전처리 시 문헌 빈도(DF)가 3미만인 단어를 제거하여 차원을 축소하였다.

다음으로 LDA 토픽모델과 Doc2Vec을 이용하여 자질을 추출 및 선정하였다. 이 때, 선정방법에 따라 동일한 단어도 다른 자질 값을 가질 수 있다.

실험에 이용한 자질집합의 종류를 학술문헌의 위치에 따라 표제, 키워드, 초록과 본문으로 구분하고 이를 학습문헌과 검증문헌으로 분할하였다. 〈표 6〉은 자질집합의 종류별 문헌집단 현황과 추출한 단어의 수를 나타냈다.

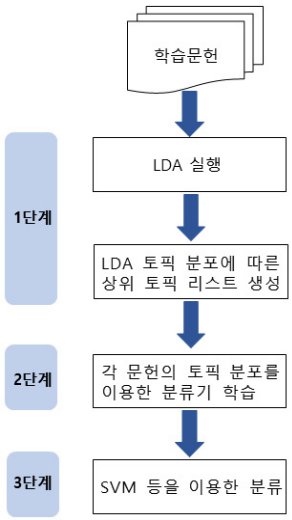
LDA와 Doc2Vec을 이용한 실험은 토픽 개수와 차원 수에 따라 자질집합 크기에 따른 분류 성능을 비교하였으며 각각의 실험 단계에 구체적인 내용을 기술하였다.

3.3.3 자질 추출 및 선정

자질 선정을 위한 방법으로 LDA 토픽 모델과 Doc2Vec을 활용하였다.

본 연구에서는 토픽 자질을 생성하는데 JAVA 기반의 Mallet(http://mallet.cs.umass.edu/)을 활용하였다. 또한 Doc2Vec 벡터 기반 학습 자질을 생성할 때 Python 언어 기반의 gensim을 활용하였다.

LDA는 문헌의 전체적인 주제 구조를 단어로 나타내며 단어 분포에 기반을 둔 문헌 표현이 가능하다. LDA 토픽 모델을 이용한 분류 실험 절차는 〈그림 5〉와 같다.



〈그림 5〉 LDA를 이용한 실험 절차

LDA는 문헌에서 각 주제별로 중요한 단어를 순위화한 결과를 확인할 수 있다. 주제별로

〈표 6〉 학술문헌의 위치에 따른 집합 구성 현황

종류	학습문헌	검증문헌	단어 수
표제	27,557	1,294	20,883
키워드	11,634	1,294	11,942
초록	24,507	2,724	61,304
본문	24,507	2,724	61,304

상위에 나타난 단어를 확률 분포에 근거하여 자질로 선정하여 차원을 축소하였다. LDA로 단어 분포를 산출한 후에 토픽별로 상위 N개의 단어를 사전으로 구성한다. 예를 들어 토픽 1로 구분된 상위 토픽 분포의 순위 단어는 토픽 1의 1번째 순위 단어, 토픽 2에 해당하는 3번째 상위 단어는 토픽 2의 3번째 순위 단어로 토픽 분포에 근거하여 확인할 수 있다.

3.3.4 분류기 학습 및 검증

실험에 사용한 분류 알고리즘은 나이브 베이즈(navie bayes, 이하 NB), k-최근접 이웃(k-nearest neighbor, 이하 k-NN), 지지 벡터 기계(support vector machine, 이하 SVM), Ridge와 Doc2Vec으로 총 5가지이다. NB, k-NN, SVM과 Ridge 분류기는 JAVA 언어로 공개된 기계학습 실험 패키지인 Weka 3.6 버전을 사용하였고, Doc2Vec은 Python 언어를 이용하여 구현하였다. 분류기 학습을 위한 입력 자질 구성에 이용한 자질선정 방법은 LDA 토픽 모델과 Doc2Vec이다. LDA를 활용하여 추출한 자질은 NB, k-NN, SVM, Ridge 분류기를 이용하여 성능 평가에 활용하였다.

3.3.5 성능 평가

분류 성능 평가 척도로 마이크로 평균 정확률(micro average precision)을 사용하였다. 마이크로 평균 정확률은 전체 범주의 할당 건수 중에서 올바르게 분류된 문헌에 대한 비율을 산출한 것이다. 본 연구에서 사용한 실험문헌은 분류라벨이 각 문헌마다 하나씩만 할당되어 있으므로 마이크로 평균 정확률과 마이크로 평균 재현율, 그리고 이를 결합한 마이크로 평균 F1 척도가 모두 동일한 값을 가진다. 이 경우 마이크로 평균 정확률을 평가 척도로 이용한다(이재윤, 2005).

일반적으로 분류기의 성능은 하나의 범주에 대하여 측정한 결과로만 평가하기는 어렵다. 분류 성능은 각 범주별 분류 결과에 대한 측정값을 종합하여 평가해야 한다. 따라서 문헌이 올바른 범주로 분류되었는지 측정한 값은 각 범주의 분류 성능을 평균하여 평가하는 것이 적합하다. 성능 평가에 이용한 마이크로 평균 정확률의 수식은 다음과 같다.

$$\text{마이크로 평균 정확률} = \frac{\text{검증문헌의 적합 범주 할당 횟수}}{\text{검증문헌의 총 할당 횟수}}$$

이밖에도 매크로 평균(macro average)을 이용한 정확률 및 재현율 척도가 있으나 일반적으로 문헌자동분류 실험의 성능 평가에는 마이크로 평균이 이용된다(Yang & Liu, 1999). 분류 방법에 따른 학습속도는 많은 양의 문헌을 처리하는 데 있어서 중요한 요소이다. 따라서 성능 평가에서 정확률과 함께 학습속도를 간략하게 살펴보고자 한다.

4. 연구결과 분석 및 평가

4.1 자질집합의 크기에 따른 실험 결과

학습자질집합의 크기에 따른 성능을 비교하기 위하여 두 가지 접근방법을 채택했다. 첫 번째 접근방법은 LDA 토픽 모델로 탐지한 k개의 토픽에 포함된 n개의 단어를 핵심 단어로

추출하는 방법이고 두 번째는 Doc2Vec을 이용하여 각 문헌을 N차원으로 전사하여 표현하는 방법이다. 일반적으로 학습자질집합의 크기가 증가할 경우에 분류 성능이 향상되는 경향이 있으나 일정한 크기를 넘어서게 되면 오히려 분류 성능이 저하될 수 있다. 따라서 실험을 통하여 최적화된 자질집합의 크기를 파악하고자 하였다.

본 실험에서는 분류기를 학습하고 분류 성능을 살펴보았다. 이 때 토픽으로 등장한 k개의 단어는 학습에 이용한 자질의 수를 의미한다. 자질집합의 크기에 따른 분류 성능의 차이를 확인하기 위하여 토픽 개수를 10, 20, 30개로 조정하여 정확률을 살펴보았다. 〈표 7〉은 LDA 토픽 모델 수행 결과, 자질집합의 크기와 위치에 따른 성능의 차이를 나타낸 결과이다.

LDA 토픽 모델은 사전에 추출할 토픽의 개수와 각 토픽에 포함되는 단어의 수를 설정할 수 있다. 일반적으로 한 개의 토픽에는 20개의 단어를 포함하도록 설정하며 본 실험에서도 이와 같은 방법으로 토픽의 개수를 설정하고 특성을 추출하였다. 예를 들어 10개의 토픽은 200개의 단어가 포함되어 집합을 구성한다.

토픽 모델을 이용하여 구성한 자질집합의 크기를 조정하며 실험 환경에 변화를 주고 성능을 살펴본 결과, 토픽 개수가 30개일 때 성능이 가장 높은 것으로 나타났다. 토픽 개수가 30개일 때 표제, 키워드, 초록과 본문에서 가장 높은 분류 정확률을 가졌다.

자질집합의 종류에 따른 분류 실험 결과, 본문을 이용한 경우 표제, 키워드, 초록보다 높은 성능을 보였다. 본문은 학습문헌 수가 표제 및 키워드에 비교하여 많고 범주를 표현하는 양질의 단어집합으로부터 분류자질을 선정할 수 있다는 점이 영향을 미친 것으로 볼 수 있다. 논문의 초록은 단어 수와 길이가 상대적으로 일정하고 본문과 비교하여 적은 수의 단어를 포함하므로 학습속도에 효율성을 가진다.

표제와 키워드를 이용한 결과, 각각 81.42%와 87.48%의 성능을 기록하였으며 전반적인

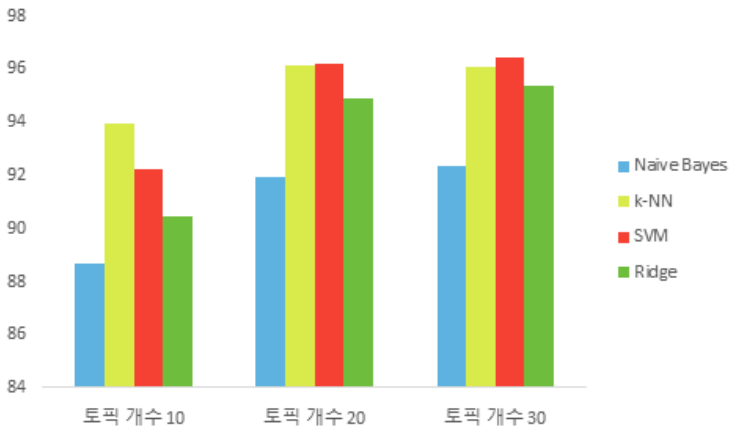
〈표 7〉 LDA 토픽 개수에 따른 분류 성능 비교

		표제	키워드	초록	본문
NB	t=10	72.61	81.09	85.57	88.70
	t=20	72.98	81.17	87.13	91.95
	t=30	73.17	81.01	87.65	92.37
k-NN	t=10	78.93	85.87	89.93	93.97
	t=20	79.90	86.49	92.58	96.16
	t=30	81.40	**87.48**	92.14	96.10
SVM	t=10	79.73	85.82	88.95	92.21
	t=20	80.59	86.21	92.73	96.18
	t=30	**81.82**	87.19	**93.46**	**96.42**
Ridge	t=10	79.43	85.44	87.72	90.48
	t=20	80.39	86.51	91.18	94.90
	t=30	81.39	86.51	92.07	95.36

분류 성능이 초록과 본문에 비교하여 낮게 나타났다. 표제와 키워드는 초록과 본문에 비하여 한정된 단어 수와 단편적인 정보로 이루어지기 때문에 충분한 학습 양을 제공하지 못하는 점을 고려할 수 있다. 그러나 표제와 키워드는 본문을 대표할 수 있는 중요한 단어가 집약적으로 표현되어 나타난 부분이다. 또한 적은 개수의 단어로 실험하였을 때도 80% 이상의 성능을 보여 분류자질로써 충분한 가치를 가짐을 알 수 있다. 따라서 표제와 키워드를 자질 선정에 이용할 때, 출현한 단어를 바로 선정하는 것보다는 본문에서 자질 선정 시 표제와 키워드에 출현한 단어에 일정 값의 가중치를 부여하는 방법을 활용할 수 있다. 특히 학습 양이 본문에 비해 많지 않으나 87% 가량의 성능을 나타낸 키워드를 표제와 합하여 단어빈도 등의 순위를 이용하고 본문에서 해당 단어에 가중치를 주는 기준으로 활용한다면 자질의 대표성을 확보할 수 있을 뿐만 아니라 선정에 소요되는 노력도 절감할 수 있을 것이다.

〈그림 6〉은 본문을 이용한 자질집합의 크기와 분류 알고리즘에 따른 성능을 도식화한 것이다. SVM으로 실험한 결과 본문과 초록의 성능 차이는 2.96%이다. 분류 알고리즘에 따른 성능을 살펴보면 SVM이 표제, 초록, 본문에서 가장 높은 성능을 나타냈으며 k-NN은 키워드 자질집합을 이용한 경우 가장 높은 성능을 보였다. 가장 낮은 성능을 보인 것은 NB이다. NB는 단어를 독립적으로 판별하여 처리 속도의 효율성을 가지며 재현율이 높게 나타나는 특성이 있으나 정확률에서는 다른 분류기에 비하여 성능이 낮은 경향을 보인다. 실험 결과, 표제에서 가장 높은 성능을 나타낸 SVM과 8.65% 차이를 보였고, 본문에서는 4.05%의 성능 차이를 나타냈다. Ridge와 k-NN이 SVM 다음으로 높은 성능을 보였으며 일반적으로 k-NN이 근사한 차이로 Ridge보다 높은 정확률을 나타냈다.

SVM과 Ridge 분류기를 이용하였을 때 표제, 키워드, 초록과 본문에서 모두 토픽 30개의 자질집합이 높은 성능을 보였다. 키워드를 NB로 분류한 결과, 토픽 20개를 학습에 이용한 경우에 토픽 30개를 이용하였을 때보다 0.16% 높



〈그림 6〉 토픽 자질집합의 크기에 따른 분류 성능 비교

은 성능을 나타냈다. Ridge를 이용한 결과, 토픽 20개와 토픽 30개가 동일한 성능을 보였다.

k-NN 분류기로 실험한 결과 초록과 본문에서 토픽 20개의 자질집합이 토픽 30개보다 각각 0.42%와 0.16% 높은 성능을 보여 가장 적합한 크기를 확인할 수 있다.

Doc2Vec을 이용한 실험을 통해서 자질집합의 크기 및 종류와 학습 횟수가 분류 성능에 어떤 영향을 미치는지 확인하고자 하였다. 〈표 8〉은 Doc2Vec의 학습 횟수와 자질집합의 크기 및 종류에 따른 분류 성능을 비교한 것이다.

구체적으로 학습 횟수가 자질집합의 크기 및 종류에 따라 분류 성능에 미치는 영향이 차이를 가지는지를 확인하고자 한다. 또한 딥 러닝을 활용할 경우, 기존의 방법을 이용한 경우와 비교하여 어떤 성능 차이를 가지는지 살펴보고자 한다.

학습 횟수는 20과 30번 두 가지로 설정하였다. D는 차원을 의미하는 Dimension의 약자이며 분류자질의 개수이다.

키워드를 제외한 표제, 초록과 본문은 학습 횟수가 30번일 때 더 높은 성능을 나타냈다. 또한 다른 자질집합을 사용할 때 보다 본문을 이용한 경우에 월등하게 높은 성능을 보였다. 키워드는 학습 횟수가 20번일 때 0.51%에서 1.06% 사이로 학습을 30번 진행할 때보다 높게 나타났다. 학술문헌에서 키워드는 단어 그 자체이며 문맥적인 추론 요소가 거의 없다. 하나의 논문은 약 5개의 저자 키워드가 부여되는데 이들 간의 관계성을 유추하여 자질 선정에 반영하는 과정은 문맥 추론 과정과 차이를 가진다.

4.2 문헌 출현 위치에 의한 자질집합 종류에 따른 실험 결과

학술문헌의 위치에 따라 표제, 키워드, 초록, 본문을 자질집합으로 구성하여 실험한 결과 LDA와 Doc2Vec 모두 본문에서 가장 높은 성능을 나타냈다.

〈표 9〉는 LDA와 Doc2Vec 실험에서 토픽 개수와 벡터 차원 수를 조정하며 실험한 결과, 자질집합의 종류에 따라 가장 높은 성능을 나타낸 결과를 정리하였다. Doc2Vec은 본문을 이용한 실험에서 100개의 차원으로 LDA로 구성한 600개의 자질을 이용하여 NB, k-NN, Ridge로 분류한 경우보다 더 높은 성능을 나타냈다. LDA 수행 결과 산출된 단어를 SVM으로 분류한 실험에서는 Doc2Vec을 이용한 경우보다 0.14% 높은 정확률을 보였다. 그러나 Doc2Vec이 100개 차원을 이용한 결과와 비교할 때 LDA는 더 많은 개수의 단어를 필요로 한다는 점에서 제한을 가진다.

범주 할당 시 분류 알고리즘 연산의 복잡성

〈표 8〉 Doc2Vec을 이용한 학습 횟수 및 자질집합의 크기와 종류에 따른 분류 성능 비교

학습 횟수	표제		키워드		초록		본문	
	D=50	D=100	D=100	D=200	D=100	D=200	D=100	D=200
20	58.42	59.65	**63.76**	**65.10**	89.68	87.95	94.96	92.83
30	**58.44**	**60.03**	63.25	64.04	**90.31**	**89.16**	**96.28**	**95.57**

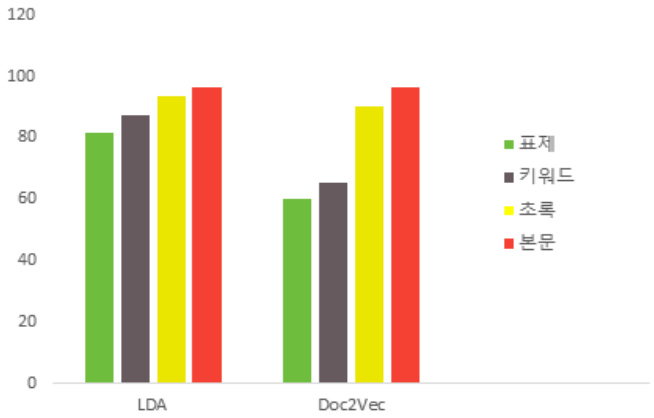
〈표 9〉 자질집합 종류에 따른 분류 성능 비교

		표제	키워드	초록	본문
LDA (feature: 600)	NB	73.17	81.17	87.65	92.37
	k-NN	81.40	87.48	92.58	96.16
	SVM	81.82	87.19	93.46	96.42
	Ridge	81.39	86.51	92.07	95.36
Doc2Vec (feature: 100)		60.03	65.10	90.31	96.28

을 줄이고 차원의 저주를 피하기 위하여 학습 방법에 따른 성능을 비교하였다. Doc2Vec을 표제와 키워드로 실험한 결과, 약 60%에서 65%의 성능을 나타냈으며 초록은 약 90%, 본문은 약 96%를 보였다. 자질집합의 종류에 따른 편차는 약 36%이다. 이는 LDA의 편차가 약 15%로 나타난 것과 비교할 때 상대적으로 큰 차이임을 알 수 있다.

이러한 원인은 학습방법과 추론 과정의 차이에서 기인하며 심층 신경망을 사용하는 딥 러닝과 단순 확률 분포와 통계 값을 이용한 방법과의 차이에 근거한다.

〈그림 7〉은 LDA로 표제, 키워드, 초록과 전문에서 추출한 단어로 SVM 분류기를 학습하여 분류한 결과와 Doc2Vec을 사용하여 분류한 성능을 비교한 것이다. 표제, 키워드는 본문을 대상으로 분류한 결과와 비교할 때 성능 차이가 크다. 문헌의 내용을 심층적으로 탐색하여 표현하는 딥 러닝은 본문처럼 추론 정보가 많은 긴 텍스트를 처리할 때 표제와 키워드의 단편적인 정보를 처리하는 경우보다 분류 효율성에서 가치를 가진다. 처리할 문헌의 양이 적고 텍스트의 내용이 단순한 경우에는 LDA를 이용하는 것이 바람직하다. Doc2Vec을 이용한 실험은 방대한 양의 문헌과 복잡한 텍스트 내용 처리 시 보다 적은 개수의 단어로 내용에 밀착하여 표현함으로써 분류의 효율성을 가짐을 확인할 수 있다.



〈그림 7〉 자질집합 종류에 따른 분류 성능 비교

4.3 문맥 추론 정보에 따른 실험 결과

딥 러닝은 추론 과정이 복잡하기 때문에 단순한 정보를 반복적으로 학습할 경우에 오히려 분류 성능이 저하될 수 있다. 딥 러닝을 이용한 학습은 추론 수준이 높기 때문에 문맥 정보가 미미한 단순 정보로부터 문맥 상 의미를 살릴 수 있도록 심층적으로 내용을 학습하는 경우 오류 확률이 높아진다. 따라서 Doc2Vec을 분류 실험에 적용하는 경우에는 문맥 정보가 풍부하고 많은 양의 데이터를 학습할 때 성능이 높게 나타남을 알 수 있다.

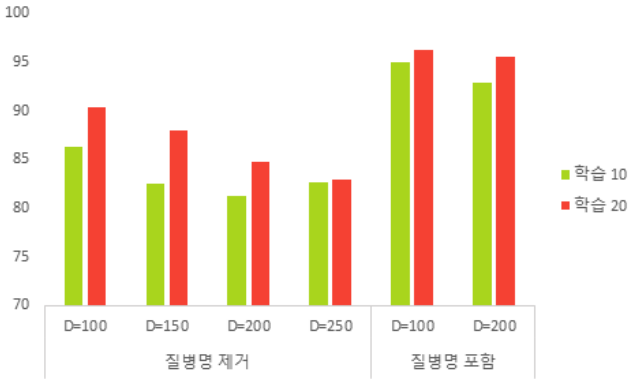
위의 결론을 확인하기 위한 실험으로 문헌에 등장한 강력한 범주 식별 정보인 질병 명칭을 분류자질에서 제거한 경우를 추가적으로 실험하여 살펴보았다.

본 실험에서는 질병 명칭을 범주로 사용하였으며 본문에 해당 단어가 출현한 경우 분류의 성능을 높이는데 영향을 미칠 것이라고 가정했다. 따라서 질병 명칭을 포함하지 않았을 경우의 분류 성능을 살펴보고 질병을 포함하였을 때와 비교하여 자질선정에 반영하고자 하였다.

〈표 10〉은 Cancer, Diabetes, Aids를 각 범주에 속한 문헌의 본문에서 제거한 후 분류한 결과를 학습 횟수 및 크기에 따라 실험한 결과이다. 실험 결과, 자질집합의 크기가 100이고 학습 횟수가 30번일 때 91.31%로 자질의 차원수가 150개인 경우보다 0.44% 높았다. 50개씩 자질의 개수를 증가하여 실험한 결과 점차 성능이 떨어지는 것을 볼 수 있었다. 〈그림 8〉은

〈표 10〉 질병 단어 제거 후 Doc2Vec 학습 횟수와 크기에 따른 분류 성능 비교

학습 횟수	질병명 제거			
	D=100	D=150	D=200	D=250
10	86.30	82.46	81.25	82.61
20	90.35	88.06	84.74	82.96
30	**91.31**	90.67	**89.03**	88.39
40	91.11	**90.87**	88.92	**88.68**



〈그림 8〉 질병명칭 제거 및 포함에 따른 성능 비교

본문에서 질병 명칭을 제거 및 포함한 분류 성능을 학습 횟수와 차원 수에 따라 나타냈다. 질병 명칭을 포함한 경우에 100개 차원이 96.28%를 나타낸 것에 비교하여 200개 차원이 95.57%를 나타내며 상대적으로 낮은 성능을 보였다.

질병 명칭은 문맥 추론 시 해당 질병과 관계된 단어 간의 유추에 많은 영향을 주었음을 실험을 통해 확인할 수 있었다. 단어가 제거된 것은 문맥의 추론에 해당 질병 단어와 관계된 유추가 이루어지지 못했음을 의미하며 생의학 문헌에서 질병명은 주변 문맥과 관계된 단어를 선별할 수 있다는 점에서 중요함을 알 수 있었다.

4.4 실험 결과 고찰

실험 결과를 종합하여 살펴보면 다음과 같다. LDA 토픽 모델은 600개의 단어를 활용하였을 때 표제, 키워드, 초록과 본문에서 단어의 개수를 10개(200개), 20개(400개)로 조정한 경우보다 높은 성능을 보였다. 본문을 이용하여 SVM으로 실험한 결과, 토픽 30개를 이용한 경우에 96.42%를 보였으며, 토픽 20개를 이용한 경우보다 정확률이 0.28% 높게 나타났음을 확인할 수 있다. Doc2Vec 실험은 차원 수가 100개일 때 96.28%의 성능을 나타냈다.

문헌의 위치에 의한 자질집합의 종류에 따른 실험에서 LDA를 활용한 경우 표제, 키워드와 초록 및 본문의 성능이 최고 15% 가량 차이를 가졌으며 Doc2Vec을 이용한 실험에서는 자질집합의 종류 간 차이가 최대 36%까지 나타났다. 그러나 본문을 자질 선정에 이용한 경우에 방법 간 성능 편차가 600개의 자질을 이용한 LDA 실험, 100개의 단어를 이용한 Doc2Vec 실험에서 3%미만으로 나타났다. 또한 초록에서 LDA가 93.46%, 본문에서 96.42%를 나타냄으로써 표제 및 키워드와 동일한 개수의 단어를 이용하여 초록과 본문에서 더 높은 성능을 보임을 확인할 수 있었다. LDA와 Doc2Vec이 문헌의 내용에 근거하여 추출할 단어를 선별하고, Doc2Vec은 단어 간의 관계를 추론하는 과정이 심층 학습으로 이루어지기 때문에 추론 과정이 복잡한 내용의 텍스트를 처리할 때, 단순한 내용의 텍스트를 처리할 때보다 성능에 큰 영향을 받지 않는 원리에 기인한 것이다.

LDA 방법은 각 문헌의 주제에 따른 토픽 분포를 단어의 선정 기준으로 활용하기 때문에 보다 내용에 근거한 단어가 선정된다. 따라서 상대적으로 문맥 정보를 담고 있지 않은 표제와 키워드는 초록 및 본문과 비교하여 차이를 가질 수 있다.

딥 러닝 방법인 Doc2Vec을 이용한 실험에서 이 차이를 더 명확하게 확인할 수 있었다. Doc2Vec 실험 결과 표제에서 60% 대의 성능을 나타낸 것과 달리 본문에서는 100개의 자질로 96% 가량의 분류 성능을 나타냈다. 실험 결과, Doc2Vec의 학습 횟수는 본문에서 20회로 진행할 때 가장 높았으며 차원 수는 100인 경우 가장 높은 성능을 나타냈다. 또한 차원 수를 50개씩 증가시키고, 학습 횟수를 10, 20회로 비교한 결과 자질집합의 종류에 따라 분류 성능의 차이가 나타남을 확인하였다. 표제, 키워드에서 추출한 단어는 학습 횟수와 차원 수가 증가할 때 분류 성능이 낮아지는 경우가 발생함을 알 수 있었다. 이는 Doc2Vec이 표제와 키워드에서 LDA 보다 낮은 성능을 나타내는 것과 동일한

원리에 기인한다.

이러한 특징은 질병명칭을 제거한 실험과 질병명칭을 포함한 실험 결과의 비교를 통해 더 명확하게 확인할 수 있었다. 질병명은 범주를 대표하는 단어이며 각 문헌에 출현하는 빈도가 높은 단어이다. 질병 단어의 제거는 해당 단어를 기준으로 주변에 위치한 단어 간의 관계 및 유사 단어에 대한 학습이 이루어지지 않은 경우를 의미한다. 따라서 질병 단어를 기준으로 주변의 단어를 통한 문맥 학습이 이루어지지 않음으로써 심층 학습의 강점인 추론 정보가 질병 단어를 기준으로 삭제되었을 때 성능이 낮아질 수 있음을 확인하였다. 심층 학습 방법을 이용한 연산 과정에서 단어 간의 관계를 추론할 때 문맥 정보는 중요한 영향을 미친다. 그런데 해당 질병 단어가 제거된 상태에서 단순히 단어 정보가 삭제된 것이 아니라 그 단어를 기준으로 추론할 수 있는 과정이 문헌을 표현할 때 반영되지 않음을 알 수 있다. 따라서 질병명은 문헌에서 중요한 단어이며 문맥 정보를 반영하여 성능에 영향을 주는 요소임을 확인할 수 있었다.

5. 결 론

본 연구는 LDA 토픽 모델과 최근 분류 연구에 도입된 딥 러닝 방법을 적용하고 성능의 차이를 자질집합의 크기와 문헌 출현 위치 및 분류 알고리즘에 따라 평가하였다. 실험 결과를 정리하면 다음과 같다.

첫째, 자질집합의 크기에 관한 실험 결과, 자질 추출 방법에 따른 성능 편차가 있음을 확인하였다. 본문을 대상으로 실험하였을 때, LDA 실험에 600개 자질을, Doc2Vec 실험에 100개 자질을 이용하였다. LDA는 본문에서 토픽 20개를 이용하였을 때 96.42%, Doc2Vec은 100개 차원과 30번의 학습 횟수에서 96.28%를 나타냈다. 학습 문헌의 양이 많지 않고 비교적 간단한 예제를 판별하는 경우에는 딥 러닝 방법인 Doc2Vec을 적용하는 것보다 SVM 등의 기존 분류기를 활용하는 방법이 분류의 효율성을 고려할 때 적합한 방법이다.

둘째, 단어의 출현 위치에 따라 조정하며 실험한 결과, 본문에서 자질을 선정하여 이용하였을 때 LDA와 Doc2Vec을 이용한 모든 실험에서 가장 높은 성능을 나타냈다. 표제, 키워드, 초록을 대상으로 선정한 자질은 본문과 비교할 때 분류 성능이 낮게 나타난다. 그러나 표제, 키워드, 초록에서 선정한 자질은 본문의 핵심 정보를 포함하고 분류에 소요되는 시간이 짧아 그 자체로 일정 수준 이상의 정확률을 나타냈다는 점에서 각각의 집합만을 이용하여 실험하는 것보다 향후 분류자질 선정 시 가중치로 적용할 수 있는 가능성을 확인하였다.

셋째, 딥 러닝 방법인 Doc2Vec을 이용한 실험 결과, Doc2Vec 알고리즘이 어떤 특징을 가지는지 자세하게 살펴보고자 학습 횟수, 차원 개수와 자질집합의 종류에 따른 성능을 확인하였다. 또한 선행 연구에서 간과하였던 문맥 정보의 반영 여부를 확인하고자 범주 식별성이 높다고 판단한 질병 명칭을 제거하여 본문을 대상으로 수행하였다. 성능을 비교한 결과, 질병 단어를 제거하였을 경우에 학습 횟수와 차원의 개수와 관계없이 성능이 저하됨을 확인하였다. 본 연구의 의의는 다음과 같다. 첫째, 국

내의 문헌자동분류 연구에서 잘 다루어지지 않았던 LDA 토픽 모델과 최근의 딥 러닝 학습 방법을 적용한 Doc2Vec 기법을 적용하고 분류 알고리즘 간의 차이점을 세밀하게 분석하여 종합적으로 평가하였다. 이를 통해 기법 간 성능 요소에 따른 영향의 정도를 확인하고 분석하여 제시하였다는 점에서 향후 분류 연구에 활용할 수 있을 것으로 기대한다. 둘째, 딥 러닝을 자질 선정 및 분류에 이용하고 기존 방법과의 차이점을 심층적으로 평가함으로써 각 요인이 성능에 미치는 영향의 정도를 확인하였다.

본 연구는 학술문헌으로 구성된 하나의 실험 집단을 이용하고 범주의 개수가 제한적이라는 점에서 실험 결과를 일반화 할 수 있도록 학술 문헌 이외에 트위터, 기업의 상품 리뷰 데이터 등 텍스트의 유형에 따른 길이와 내용의 특성을 고려하여 자질선정 시에 세부적인 기준을 제시하고 접근하는 연구가 필요하다.

향후 연구에서는 딥 러닝과 토픽모델링을 이용한 자질선정 방법과 가중치 기법을 다양한 방법으로 결합하여 분류 성능 향상을 위한 새로운 기법을 제안하고자 한다. 토픽모델링을 이용한 기존의 학제성 측정 연구에 Doc2Vec을 적용하여 보다 정밀한 분석도 가능할 것으로 기대한다.

참 고 문 헌

김도우, 구명완 (2017). Doc2Vec과 Word2Vec을 활용한 Convolutional Neural Network 기반 한국어 신문 기사 분류. 정보과학회논문지, 44(7), 742-747.

김판준 (2016). 기계학습에 기초한 자동분류의 성능 요소에 관한 연구. 정보관리학회지, 33(2), 33-59. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

이재윤 (2005). 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구. 한국문헌정보학회지, 39(2), 123-146.

정영미 (2012). 정보검색연구(증보판). 서울: 연세대학교 출판문화원.

진설아, 송민 (2016). 토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구. 정보관리학회지, 33(1), 7-32. http://doi.org/10.3743/KOSIM.2016.33.1.007

최상희, 이재윤 (2012). 문서 클러스터링을 위한 학술지 논문의 구조적 초록 활용성 연구. 정보관리학회지, 29(1), 331-349. http://dx.doi.org/10.3743/KOSIM.2012.29.1.331

Atlig, C., Reyyan, K. O. C., & Yigit, T. A. K. A. (2017). Learning-based classification of natural science articles. International Journal of Scientific Research in Information Systems and Engineering (IJSRISE), 2(3), 20-26. http://www.ijsrise.com/index.php/IJSRISE/article/view/52

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3(Feb), 1137-1155.

Bhushan, S. B., Danti, A., & Fernandes, S. L. (2017). A novel integer representation based approach for classification of text documents. In Proceedings of the International Conference on Data Engineering and Communication Technology (pp. 557-564). Springer, Singapore.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84. http://dx.doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning (pp. 160-167). ACM. https://doi.org/10.1145/1390156.1390177

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3(Mar), 1289-1305.

Fuhr, N., & Buckley, C. (1991). A probabilistic learning approach for document indexing. ACM Transactions on Information Systems (TOIS), 9(3), 223-248.

Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. Journal of the American Society for Information Science, 26(5), 280-289. https://doi.org/10.1002/asi.4630260504

Hofmann, T. (2017, August). Probabilistic latent semantic indexing. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 211-218). ACM.

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. Stud Health Technol Inform, 235, 246-50.

Jiang, S., Lewris, J., Voltmer, M., & Wang, H. (2016, April). Integrating rich document representations for text classification. In Systems and Information Engineering Design Symposium (SIEDS), 2016 IEEE (pp. 303-308). IEEE. https://doi.org/10.1109/sieds.2016.7489319

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In Proceedings of the Eleventh International Conference on Machine Learning (pp. 121-129). https://doi.org/10.1016/b978-1-55860-335-6.50023-4

Koller, D., & Sahami, M. (1996). Toward optimal feature selection. Stanford InfoLab.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In Proceedings of the 32nd International Conference on Machine Learning (pp. 957-966).

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368.

Le, D. T., & Bernardi, R. (2012, July). Query classification using topic models and support vector machine. In Proceedings of ACL 2012 Student Research Workshop (pp. 19-24). Association for Computational Linguistics.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (pp. 1188-1196).

Lewis, D. D. (1992, February). Feature selection and feature extraction for text categorization. In Proceedings of the workshop on Speech and Natural Language for Computational Linguistics. https://doi.org/10.3115/1075527.1075574

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016, July). Topic modeling for short texts with auxiliary word embeddings. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 165-174). ACM. https://doi.org/10.1145/2911451.2911499

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015, July). Support vector machines and word2vec for text classification with semantic features. In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on (pp. 136-140). IEEE. https://doi.org/10.1109/icci-cc.2015.7259377

Liu, Y., Liu, Z., Chua, T. S., & Sun, M. (2015, January). Topical word embeddings. In AAAI (pp. 2418-2424).

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4), 309-317. https://doi.org/10.1147/rd.14.0309

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60). https://doi.org/10.3115/v1/p14-5010

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing

systems (pp. 3111-3119).

Mladenic, D., & Grobelnik, M. (1999). Predicting content from hyperlinks. In Proceedings of the ICML-99 Workshop on Machine Learning in Text Data Analysis, J. Stephan Institute.

PubMed Central (2017). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill. 24-51.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1422-1432). https://doi.org/10.18653/v1/d15-1167

Torkkola, K. (2004). Discriminative features for text document classification. Formal Pattern Analysis & Applications, 6(4), 301-308. https://doi.org/10.1007/s10044-003-0196-8

Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394). Association for Computational Linguistics.

Wadbude, R., Gupta, V., Mekala, D., Jindal, J., & Karnick, H. (2016). User bias removal in fine grained sentiment analysis. arXiv preprint arXiv:1612.06821.

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C. L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174, 806-814. https://doi.org/10.1016/j.neucom.2015.09.096

Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90-94). Association for Computational Linguistics.

Wang, Z., & Qian, X. (2008, December). Text categorization based on LDA and SVM. In Computer Science and Software Engineering, 2008 International Conference on (Vol. 1, pp. 674-677). IEEE. https://doi.org/10.1109/csse.2008.571

Wei, X., & Croft, W. B. (2006, August). LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 178-185). ACM. https://doi.org/10.1145/1148170.1148204

Xing, C., Wang, D., Zhang, X., & Liu, C. (2014, December). Document classification with distributions of word vectors. In Signal and Information Processing Association Annual Summit and

Conference (APSIPA), 2014 Asia-Pacific (pp. 1-5). IEEE.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information retrieval, 1(1-2), 69-90. https://doi.org/10.1109/apsipa.2014.7041633
http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Choi, Sanghee, & Lee, Jae-Yun (2012). Usability analysis of structured abstracts in journal articles for document clustering. Journal of Korean Society for Information Management, 29(1), 331-349. http://dx.doi.org/10.3743/KOSIM.2012.29.1.331

Chung, Yung-Mee. (2012). Research in information retrieval (Rev. ed.). Seoul: Yonsei University Press.

Jin, Seol A, & Song, Min (2016). Topic modeling based interdisciplinarity measurement in the informatics related journals. Journal of Korean Society for Information Management, 33(1), 7-32. http://doi.org/10.3743/KOSIM.2016.33.1.007

Kim, Dowoo, & Koo, Moung-Wan (2017). Categorization of Korean news articles based on convolutional neural network using Doc2Vec and Word2Vec. Journal of KIISE, 44(7), 742-747.

Kim, Pan-Jun (2016). An analytical study on performance factors of automatic classification based on machine learning. Journal of Korean Society for Information Management, 33(2), 33-59. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

Lee, Jae-Yun (2005). An empirical study on improving the performance of text categorization considering the relationships between feature selection criteria and weighting methods. Journal of the Korean Library and Information Science Society, 39(2), 123-146.