https://doi.org/10.5392/JKCA.2018.18.08.178

국내 학술논문 주제 분류 알고리즘 비교 및 분석

Comparison and Analysis of Subject Classification for Domestic Research Data

최원준, 설재욱, 정희석, 윤화묵
한국과학기술정보연구원 콘텐츠 큐레이션센터

Wonjun Choi(cwj@kisti.re.kr), Jaewook Seol(wodnr754@kisti.re.kr),
Heeseok Jeong(hsjeong@kisti.re.kr), Hwamook Yoon(hmyoon@kisti.re.kr)

요약

학술정보 성과물을 서비스하기 위하여 논문 단위의 주제 분류는 필수가 된다. 하지만 현재까지 저널 단위의 주제 분류가 되어 있으며 기사 단위의 주제 분류가 서비스되는 곳은 많지 않다. 국내 성과물 중에서 학술 논문의 경우 주제 분류가 있으면 좀 더 큰 영역의 서비스를 담당할 수 있고 범위를 정해서 서비스 할 수 있기 때문에 무엇보다 중요한 정보가 된다. 하지만, 분야 별 주제를 분류하는 문제는 다양한 분야의 전문가의 손이 필요하고 정확도를 높이기 위해서 다양한 방법의 검증이 필요하다. 본 논문에서는 정답이 알려져 있지 않은 상태에서의 정답을 찾는 비지도 학습 알고리즘을 활용해서 주제 분류를 시도해 보고 연관도와 복잡도를 활용해서 주제 분류 알고리즘의 결과를 비교해 보고자 한다. 비지도 학습 알고리즘은 주제 분류 방법으로 잘 알려진 Hierarchical Dirichlet Precess(HDP). Latent Dirichlet Allocation(LDA), Latent Semantic Indexing(LSI) 알고리즘을 활용하여 성능을 분석해 보았다.

■ 중심어 : | 과학기술정보 | 논문 데이터 | 학술논문 | 주제 분류 | 정보 서비스 |

Abstract

Subject classification of thesis units is essential to serve scholarly information deliverables. However, to date, there is a journal-based topic classification, and there are not many article-level subject classification services. In the case of academic papers among domestic works, subject classification can be a more important information because it can cover a larger area of service and can provide service by setting a range. However, the problem of classifying themes by field requires the hands of experts in various fields, and various methods of verification are needed to increase accuracy. In this paper, we try to classify topics using the unsupervised learning algorithm to find the correct answer in the unknown state and compare the results of the subject classification algorithms using the coherence and perplexity. The unsupervised learning algorithms are a well-known Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) algorithm.

■ keyword : | Science Technology Information | Research Data | Academic Paper | Subject Classification | Information Service |

* 본 연구는 한국과학기술정보연구원 과학기술콘텐츠 큐레이션 체제 구축 연구과제로 수행되었습니다.
접수일자 : 2018년 07월 06일 심사완료일 : 2018년 07월 24일
수정일자 : 2018년 07월 24일 교신저자 : 윤화묵, e-mail : hmyoon@kisti.re.kr

I. 서 론

국내뿐만 아니라 해외에서도 학술정보 유통을 위한 과학기술정보 데이터 서비스 사업에 대한 지원이 확대되고 있다. 이와 같은 현상은 국가 과학기술정보 데이터가 그만큼 중요하다는 의미를 내포하고 있으며, 해당 데이터의 활용 서비스도 국가 R&D 사업에 큰 일조를 할 것이다. 과학기술정보 중에서도 논문 성과물은 가장 많은 범위를 차지하고 있으며, 논문 성과물 데이터를 확인하고 가치 정보를 제공하는 서비스가 진행되고 있다. 국내에서는 과학기술문헌의 수집과 효과적인 제공을 위해서 NDSL(National Digital Science Library)[1]에서 연구자들에게 국내외 학술저널 및 프로시딩 정보 및 원문을 제공해 주는 서비스를 하고 있다. 하지만 현재의 부족한 부분은 저널 단위의 논문 분야가 서비스되고는 있지만 논문단위로는 서비스가 되고 있지 않아서 보다 심도 있는 데이터 분석에 어려움이 있다는 것이다. 또한, 학술정보에 대한 관심이 증가하면서 논문 검색에 분류 알고리즘을 활용한 사례도 있다[2]. 논문들 간의 주제를 분류하는 방법 중에는 토픽 모델링 방법도 많이 활용하는데 Hierarchical Dirichlet Precess(HDP)[2] 방법과 Latent Dirichlet Allocation(LDA)[3] 방법 그리고 Latent Semantic Indexing(LSI)[3] 방법이 있다. LSI 알고리즘은 PLSI또는 PLSA라고도 한다. LDA는 각 문서의 단어 분포와 문서 집합 전체의 단어 분포를 통해 주제를 도출하고 각 주제에 속할 수 있는 단어의 확률을 계산한다. LSA는 코퍼스를 가지고 문서들의 유사도를 계산하지만 LDA는 토픽 집단을 생성하는 것이 다른 점이다. 본 논문에서는 논문 단위의 주제 분류를 위해 주제 분류로 많이 활용되는 주제 분류 알고리즘을 비교해 보고 성능을 평가해 보고자 한다.

II. 선행 연구

국내 기록 관리학 연구 분야[3]에서는 HDP 토픽 모델링과 LDA 토픽 모델링을 응용한 결과를 비교 분석하였다. LDA 토픽 모델링의 경우 전체 토픽이 과도하게 특정 부분에 집중되어 있지 않고 일부 소 영역을 이루며 고르게 분포되어 각 토픽의 특징을 파악할 수 있는 고유한 키워드가 많으면 연구 주제영역을 뚜렷하게 구분할 수 있다. HDP 토픽모델링은 세부 주제별 연구 동향 분석을 하는 데에도 효율적으로 알려져 있다. 토픽 분석을 위해서는 용어 가중치의 연산을 활용한 벡터 공간 모델을 적용하는데 주로 벡터 공간 모델의 TF(Term Frequency)와 IDF(Inverse Document Frequency)를 계산하여 이들을 각 용어의 가중치로 계산한다. 비정형 의료 데이터 분석에 LDA 모델을 적용한 경우[4] LDA 알고리즘의 계산량이 많이 걸리며, 토픽에 나타난 단어 경향성이 크지 않는 경우 사용한 데이터의 특성 때문에 토픽 간 단어의 경향성에 차이가 없는 경우 또는 토픽을 충분히 다양화하지 않는 문제로 발생한다고 한다. 후보자 추천 시스템에 LDA 알고리즘을 적용한 경우도 있다[5]. LDA 알고리즘은 주제의 개수를 미리 알고 있어야 한다. 주어진 문서에서 발견된 단어의 수에 대한 분포를 분석함으로써 해당 문서가 어떤 주제를 함께 다루고 있을지를 예측할 수 있다. 문서의 잠재 키워드를 추출하는데 LDA 모델을 활용한 사례[6]도 있다. 주제란 다수의 문서에서 공통적으로 기술하고 있는 추상적인 내용을 뜻한다. 이런 관점에서 문서는 한 가지 주제에 대해서만 기술하기 보다는 여러 가지의 주제를 부분적으로 기술하고 있다. LDA 모델은 확률모델이기 때문에 여러 번 반복 실험을 실시하여 수행되었으며 논문의 제목, 초록, 키워드 등의 100개미만의 단어는 제외하고 소문자로 모두 변경 후 문장 부호를 삭제하고 불용어(전치사, 대명사, 관사 등)를 제거하고 어간 추출 방법을 활용하였다. 법률 판례 분류 시스템에 LDA를 활용한 사례[7]도 있다. 또한, 토픽 별 카테고리를 분류하기 위해 트위터의 토픽을 LDA 모델에 적용하였다[8]. 논문의 주제어와 초록의 관계를 LDA로 학습한 사례도 있다[9]. 학습을 위해 불용어를 제거하고 논문의 저자 주제어와 초록을 쌍으로 만들어 학습 데이터를 만든 후 학습 군과 테스트 군의 데이터 분류를 시도하였다.

이외에도 토픽 모델링을 이용한 유사 시청 사용자 그룹핑 및 TV 프로그램 추천 알고리즘에 적용한 사례

[10]와 토픽 모델링을 활용한 국내 문헌정보학 연구동향 분석[11] 그리고 소셜 이미지 annotation[12], IoT 어플리케이션에서의 QoS 서비스에서 link relationship 을 규명하기 위해 활용[13], 확장성 고려[14], semi-supervised 알고리즘을 활용한 텍스트 분류에 활용[15], 온톨로지 기반의 토픽 모델링 구현[16] 등이 있다. 토픽 모델에서 방대한 양의 문서가 주어졌을 경우, 토픽 학습 시간이 오래 걸린다는 문제가 있다. 토픽 확장 모델에서는 각각의 토픽이 전체 어휘 사전을 공유하는 것이 아니라, 토픽별로 독립적인 어휘 사전을 가진다. 새로운 단어가 업데이트되면 토픽에 새로운 단어를 추가하고 토픽 별로 현재 토픽의 단어 값을 계산한다. 마지막으로 상위 k개만 남긴다[17].

III. 주제 분류 알고리즘 특징 및 학습 진행

문헌 집합을 표현하는 직관적인 방법은 문헌들에 어떤 용어들이 있는지를 숫자로 표현하는 것이다. 불용어를 제외하여 단어의 빈도를 문헌 별로 나타낼 수 있는데 단점은 0 개인 단어가 많이 존재한다는 것이다. 따라서 행렬의 크기는 큰데 여러 분석 기법들이 제대로 작동되지 않을 수 있다[18]. 주성분분석(Principal component Analysis) 은 높은 차원의 데이터들을 분석하여 그 중에서 핵심이 되는 차원을 뽑아내서 차원을 낮추는 기법이며 특이 값 분해(Singular Value Decomposition)도 행렬 데이터의 차원을 축소하는데 자주 사용된다. 잠재 의미 분석(Latent Semantic Analysis) 는 압축된 행렬에서 함축된 의미를 도출해 낼 때 사용된다. pLSA(Probabilistic Latent Semantic Analysis)는 문헌-용어 행렬을 문헌 내에 특정 용어가 등장한 횟수를 기반으로 하는 것이 아니라, 문헌 내에 특정 용어가 등장할 확률을 기반으로 해서 구축하는 것이다. 따라서 행렬 내에 음수도 사용하는 SVD를 사용하지 못하며 대신 음수 미포함 행렬 분해 기법이나 기대값 최대 기법 등을 활용한다. 입력은 주로 문헌-용어만을 받으며 문헌 내에 주제가 어떻게 분포하고 있을지는 고려하지 않는다. LDA는 문서 별 주제 분류와 주제별 단어라는 2가지 내용을 고려한다. pLSA는 새로운 문헌을 생성해 내는 것을 잘 다루지 못했는데 이러한 약점을 보완하기 위해서 LDA가 등장했다. LSA는 구현과 검증이 직관적인 장점이 있다. 그리고 결과도 비교적 안정적이다. LDA 구현은 어렵지 않지만 검증이 쉽지 않다. Topic의 개수를 사전에 튜닝해야 하며, 각각의 단어요소에 너무 민감하다. 뉴스기사와 같은 형태는 LSA가 안정적이며 새로운 이슈분석과 같은 형태는 LDA가 적합하다.

토픽 모델링의 결과를 평가하는 방법은 내재적인 것(Intrinsic)과 외재적인 것(Extrinsic)으로 나뉘는데 내재적인 기법 중 고전적인 방법이 Perplexity(혼란도)이며, Perplexity의 한계를 극복하기 위해 제시된 Topic Coherence가 있다. 대게 깁스샘플링 과정에서 반복횟수가 증가할수록 Perplexity는 감소하는 경향을 나타낸다. 그러다가 특정 시점을 지나면 더 이상 Perplexity는 감소하지 않고 증가, 감소를 반복하며 요동치는 지점이 등장하는데 이때를 해당 깁스 샘플링의 수렴 지점으로 보고 샘플링을 멈추는 경우가 많다. 그리고 이때의 Perplexity가 해당 모델의 최종 Perplexity가 된다. 이 값이 작으면 작을수록 해당 토픽 모델은 실제 문헌 결과를 잘 반영한다는 뜻이므로 학습이 잘 되었다고 평가를 할 수 있다. 이 값은 LDA 등에서 적절한 주제 개수를 정하기 어려울 때 유용하게 쓰인다. 다양한 주제 개수로 학습을 진행해보면 주제 개수가 몇 일때 Perplexity가 최소가 되는지 알 수 있기 때문이다. 문제는 Perplexity 값이 작은 것이 학습이 잘 되었다는 의미이다. 그러나 Chang의 논문에 따르면 낮은 Perplexity 값이 늘 해석에 적절한 결과를 보이지는 않는다고 한다. 따라서 사람이 해석하기에 적합한지를 확인하기 위해 다른 척도가 필요했는데, 이렇게 해서 제시된 척도가 Topic Coherence(주제 일관성)입니다. 토픽 모델링 결과로 나온 주제들에 대해 각각의 주제에서 상위 N개의 단어를 뽑는다. 모델링이 잘 되었을수록 한 주제 안에는 의미론적으로 유사한 단어가 많이 모여 있게 마련이다. 따라서 상위 단어 간의 유사도를 계산하면 실제로 해당 주제가 의미론적으로 일치하는 단어들끼리 모여 있는지 알 수 있다. Ramage[9]가 고안한 Labeled

LDA는 LDA의 대표적인 지도학습 버전이다. LDA가 문헌 집합 내에서 K개의 주제를 자동으로 분류해 준다면, L-LDA는 복수 개의 주제가 라벨된 문헌들의 집합을 분석하여 주제별 단어 분포를 학습한다. Supervised LDA는 문헌에 붙은 응답 변수(response variable)가 해당 문헌을 어떤 주제로 분류해야할지 힌트를 제공한다.

TF-IDF(Term Frequency - Inverse Document Frequency)는 가중치를 구하는 알고리즘인데 주로 문서 간 비슷한 정도, 특정 단어가 문서에서 얼마나 중요한지의 척도, 문서 내 단어들에 척도들을 계산해서 핵심어를 추출, 검색엔진에서 검색결과의 순위를 결정 등에 사용된다. TF는 문서 내 특정 단어의 빈도를 말한다. IDF는 특정 단어가 전체 문서 집합 내에서 얼마나 공통적으로 많이 등장하는지를 나타내는 값으로 표현된다. 따라서 TF-IDF는 특정 문서 내에서 단어 빈도가 높을수록, 전체 문서들엔 그 단어를 포함한 문서가 적을수록 TF-IDF 값이 높아지게 된다. 그래서 이 값을 이용해서 모든 문서에 나타나는 흔한 단어들을 걸러내며, 특정 단어가 가지는 중요도를 알 수 있게 된다. TF-IDF의 목적은 다른 문서에 자주 언급되지 않고 해당 문서에는 자주 언급되는 토큰에 대해 점수를 높게 부여하는 것이다.

LDA기법은 모수 통계라는 특성 상 학습에 앞서 원 데이터가 가지는 주제의 수 K를 설정해 주어야 한다. 이 K 값에 따라 LDA 토픽 모델링의 결과가 크게 달라지기 때문에 적절한 K값의 선정은 중요하다. 따라서 다양한 K값에 대해서 분석을 돌리고 perplexity값을 기준으로 적절한 k값을 선정한다. 이것이 LDA의 중요한 단점인데 데이터에 따라 적절한 주제를 찾아주도록 하는 방법이 디리클레 프로세스(Dirichlet Process) 또는 이를 응용한 계층적 디리클레 프로세스(Hierarchical Dirichlet Process)이다. 디리클레 프로세스는 디리클레 분포의 업그레이드 버전이다. 디리클레 분포는 주어진 하이퍼 파라미터에 따라 다항분포를 생성해 주는 분포에 대한 분포이다. 여러 개의 디리클레 프로세스들을 이어주기 위해 그 위에 디리클레 프로세스를 하나 더 씌운 과정이 HDP(Hierarchical Dirichlet Process)이다. 깁스 샘플링은 두 개 이상의 확률변수의 결합 확률분포로부터 일련의 표본을 생성하는 확률적 알고리즘이다. 결합 확률분포나 그에 관련된 확률 계산을 근사하기 위해 널리 사용되고 있다.

1. 주제 분류를 위한 학습 데이터 선정

주제 분류 알고리즘 성능을 비교하기 위해 학습 데이터 선정은 다음과 같이 진행하였다. 비지도 학습은 정답이 없는 학습데이터로 테스트가 이루어진다. 국내 학술 정보(OCEAN)[1] 메타데이터에서 기사 명(한글, 영문), 키워드(한글, 영문), 초록(한글, 영문), 저널 명(한글, 영문) 정보를 입력 값으로 선정하였다. 메타데이터의 추출 기준은 학술지 종 별 최소 700개의 논문 메타데이터를 선정하였다. 전체 종은 540 종정도 되며 메타데이터 수는 약 378,000 개의 논문을 활용하였다.

한글	영문	NLTK
이	jong	i
있	yun	me
하	park	my
것	ko	myself
들	il	we
그	cha	our
되	ae	ours
수	jung	ourselves
이	jae	you
보	sung	your
않	min	yours
없	sun	yourself
나	han	yourselves

그림 1. 불용어 리스트

저자명을 입력 값으로 넣으면 많이 나오는 성 이름이 상위 키워드로 분류가 되어 제대로 된 주제 분류가 나오지 않는 문제점이 있다. 입력 데이터의 개수는 저널별 주제 분류가 서비스 되고 있기 때문이 이 정보를 바탕으로 논문 주제 분류를 구분하기 위해 우선 저널 별 대분류 기준으로 논문의 수를 나누고 DDC 대분류를 기준으로 각 종 별 700개의 학습데이터, 700개의 테스트 데이터로 구분하여 수행하였다. 주제 분류 알고리즘은 Latent Semantic Indexing(LSI), Latent Dirichlet Allocation(LDA), Hierarchical Dirichlet Precess(HDP) 알고리즘으로 성능을 비교하였다. 불용어는 한글 불용어 약 200 개를 구분하고 영문 불용어 70여개 + NLTK의 불용어 리스트를 참조하였다. [그림 1]은 불용어 리스트의 일부를 나타낸 것이다.

2. 주제 분류 알고리즘 학습 단계

주제 분류 알고리즘 성능을 비교하기 위해 학습 데이터 선정은 다음과 같이 진행하였다. 학술 정보 메타데이터에서 기사 명(한글, 영문), 키워드(한글, 영문), 초록(한글, 영문), 저널 명(한글, 영문) 등의 기본 데이터를 가지고 학습데이터를 선정하기로 하였다. 학술정보 데이터 학습 단계는 다음과 같은 순서로 진행한다[그림 2]. 우선 학습 메타데이터를 가져와서 한글 문자에서 명사 형태소를 분석하기 위해 트위터 도구(https://github.com/twitter/twitter-korean-text)를 활용한다. 영어 문자의 경우에는 명사를 확인하기 위해 NLTK 툴에서 제공하는 방법을 활용한다. 다음 단계로 불용어(stopwords) 와 유의어를 제거한다. 한글과 영어에 대한 불용어를 각각 100개 정도로 설정했으며, 유의어를 제외한 명사추출 후 리스트 화 하였다. 문서 별 단어 사전을 구축하고 빈도수를 계산한 후 코퍼스를 생성해서 TF-IDF 모델에 적용하였다. 용어의 빈도 수 측정(Term Frequency) 뿐만 아니라 문서 내에 자주 사용되지는 않지만 주제 분류에 중요한 용어를 발견하기 위해 IDF(Inverse Document Frequency) 도 활용한다. 다음으로 생성된 TF-IDF 모델을 각 LSI, LDA, HDP 모델에 입력 값으로 넣어 학습 모델을 생성한다. 테스트 데이터 집합은 다음과 같이 진행한다. 논문 메타 데이터 수가 100개라면 학습 데이터를 80으로 하고 테스트와 검증 데이터를 10 개씩 학습하는 것이 적당하다. 데이터 학습을 위해서 DDC 주제 분류가 되어 있는 메타데이터를 선정하고 학습 데이터 수를 종 별 500개씩 선별하였다. 테스트와 검증을 위한 데이터는 각각 100 개씩 설정하였다.

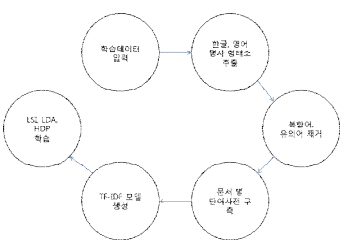

그림 2. 학술정보 데이터 학습 단계

3. 주제 분류 알고리즘 성능 비교

학습의 수행 시간은 약 1시간 정도 소요되었으며, 학습 도구는 GENSIM툴[19]을 사용하였다. 입력데이터는 전체 700개의 메타데이터에 대한 저널지 기준 대분류별 논문 수를 확인하였다. 종 별 700 개의 논문 메타 정보를 추출한 후에 각 논문이 어느 DDC 대분류에 속해 있는지를 확인하고 입력 값으로 넣었다. 총 7개의 대분류가 확인이 되었다[그림 3]. 저널지 분류로는 종교, 문학 분야보다는 컴퓨터 기술 분야의 논문이 많은 부분을 차지하고 있었다.

Arts	7700
Computer Science, Information & General Work	21700
Geography, History	2800
Philosophy and Psychology	1400
Science	56700
Social Sciences	18900
Technology	268800

그림 3. 종 별 저널지 대분류 기준 700 개의 논문 수

토픽을 7개로 설정하고 학습을 수행하여 학습된 모델을 비교한 결과는 다음과 같다[그림 4]. 괄호 안의 왼쪽에 있는 인덱스는 토픽 인덱스이며, 오른쪽은 상위 키워드 3개씩을 표시했으며, 가중치를 같이 표기하였다. LSI는 보다 다양한 분야의 키워드가 아쉬웠으며, HDP는 대분류보다는 중분류, 소분류 키워드가 존재한다. LDA는 각 토픽별 보다 다양한 키워드가 존재하는 것을 확인할 수 있다.

토픽을 10개로 설정하고 학습을 수행하여 학습된 모델을 비교한 결과는 다음과 같다[그림 5]. 지리학 같은 경우에 중복 키워드가 존재하므로 분류의 수가 큼을 확인할 수 있어 신뢰성이 낮아짐을 확인할 수 있다. 여러 주제에서 동일한 키워드가 반복되는 경우는 토픽의 수가 너무 크다는 것을 확인할 수 있다. 대체적으로 LSI는 주제별 중복 키워드가 확인되었으며, LDA와 HDP의 경우에는 주제 분류 별로 다양한 키워드가 존재하고 있음을 확인할 수 있다. [그림 6]는 LDA 알고리즘의 토픽 개수 별 일관도를 나타낸 것이다. 일관도를 측정하기 위해 각 토픽 별 상위 100개의 키워드를 활용하였다.

LSI
(0, '0.084*"inform" + 0.077*"research" + 0.076*"학회"')
(1, '-0.184*"방송공학" + -0.182*"broadcast" + -0.172*"회"')
(2, '0.221*"리아" + 0.221*"국비" + 0.221*"biblia"')
(3, '0.189*"과학" + 0.129*"technolog" + 0.121*"시료"')
(4, '-0.146*"emot" + -0.145*"sensibl" + -0.144*"디자인"')
(5, '0.175*"감성과학" + 0.173*"sensibl" + 0.172*"emot"')
(6, '-0.183*"지리학" + -0.138*"지역" + 0.114*"사용자"')
LDA
(0, '0.000*"플레이" + 0.000*"farm" + 0.000*"함의"')
(1, '0.000*"tip" + 0.000*"수소화물" + 0.000*"hydrid"')
(2, '0.000*"dischang" + 0.000*"trigger" + 0.000*"실리콘"')
(3, '0.001*"design" + 0.001*"societi" + 0.001*"research"')
(4, '0.002*"과학" + 0.002*"방송공학" + 0.002*"broadcast"')
(5, '0.000*"familiar" + 0.000*"ontolog" + 0.000*"온톨로지"')
(6, '0.000*"태깅" + 0.000*"상적" + 0.000*"폭소노미"')
HDP
(0, '0.001*emot + 0.001*감성과학 + 0.001*sensibl')
(1, '0.000*디자인 + 0.000*디자인학 + 0.000*중심')
(2, '0.000*부호화 + 0.000*time + 0.000*천북')
(3, '0.000*이북 + 0.000*didact + 0.000*busi')
(4, '0.000*peroneu + 0.000*videononehevc + 0.000*레저')
(5, '0.000*pilosa + 0.000*잠화 + 0.000*airflow')
(6, '0.000*뚜렷하 + 0.000*user + 0.000*고품격')

그림 4. 토픽 7개의 학습 모델 비교

LSI
(0, '0.084*"inform" + 0.078*"research" + 0.077*"biblia"')
(1, '0.186*"방송공학" + 0.185*"broadcast" + 0.173*"회"')
(2, '0.216*"biblia" + 0.216*"리아" + 0.216*"국비"')
(3, '0.183*"과학" + 0.123*"technolog" + 0.119*"시료"')
(4, '0.146*"design" + 0.145*"emot" + 0.145*"디자인"')
(5, '-0.201*"지리학" + 0.170*"감성과학" + 0.168*"emot"')
(6, '-0.252*"지리학" + -0.216*"글" + -0.153*"지역"')
(7, '0.137*"서비스" + -0.136*"교육" + 0.135*"servic"')
(8, '-0.143*"방송" + -0.138*"교육" + -0.126*"student"')
(9, '0.160*"글" + -0.113*"감성과학" + -0.111*"sensibl"')
LDA
(0, '0.003*"sensibl" + 0.003*"감성과학" + 0.003*"emot"')
(1, '0.001*"illumin" + 0.001*"감지" + 0.001*"변화율"')
(2, '0.001*"연산" + 0.001*"긴장" + 0.001*"질감"')
(3, '0.001*"player" + 0.001*"에이전트" + 0.001*"inconveni"')
(4, '0.002*"chromatographi" + 0.001*"슬픔" + 0.001*"koreanonenon"')
(5, '0.002*"memori" + 0.001*"컨텐츠" + 0.001*"스마트"')
(6, '0.002*"광고" + 0.001*"advertis" + 0.001*"transmiss"')
(7, '0.001*"변별" + 0.001*"회상" + 0.001*"손상"')
(8, '0.004*"영상" + 0.002*"simul" + 0.002*"색상"')
(9, '0.001*"직관" + 0.001*"염색" + 0.001*"dye"')
HDP
(0, '0.001*emot + 0.001*감성과학 + 0.001*sensibl')
(1, '0.001*감성과학 + 0.001*sensibl + 0.001*emot')
(2, '0.000*busi + 0.000*통해 + 0.000*지원')
(3, '0.000*처 + 0.000*배포 + 0.000*illeg')
(4, '0.000*interference + 0.000*busi + 0.000*clip')
(5, '0.000*gini + 0.000*perform + 0.000*토시')
(6, '0.000*store + 0.000*factor + 0.000*hnipc')
(7, '0.000*중심 + 0.000*배현경 + 0.000*량안')
(8, '0.000*특별 + 0.000*toxin + 0.000*대학병원')
(9, '0.000*손상 + 0.000*camphor + 0.000*중심성')

그림 5. 토픽 10개의 학습 모델 비교

비지도 방법으로 학술정보 데이터 학습을 수행 할 경우에는 정답 집합이 없기 때문에 주로 일관도와 복잡성을 평가해서 측정을 하게 된다. 본 논문에서는 일관도와 복잡도로 성능을 비교하였다. LDA 알고리즘에 대한 일관도는 주제 토픽이 7개 정도에서 값이 높음을 확인할 수 있다. 즉, 토픽의 개수가 7개가 될 때 최적의 토픽 분류가 된다는 것을 의미한다.

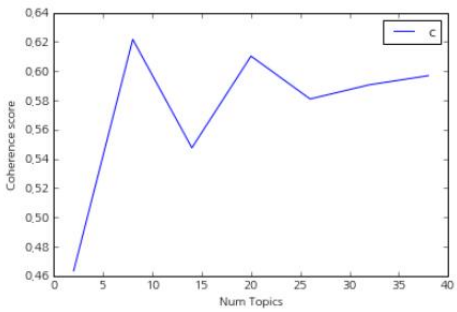

그림 6. 토픽 개수 별 일관도 비교(LDA)

[그림 7]는 LSI 알고리즘의 토픽 개수 별 일관도를 나타낸 것이다. 토픽의 개수가 적을 때는 일관도가 높지만 토픽의 개수가 증가하면 증가할수록 일관도가 떨어져 신뢰도가 낮아지는 추세를 확인할 수 있다.

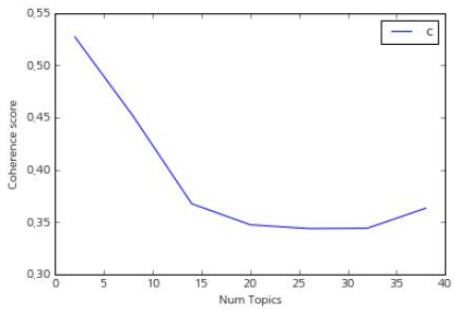

그림 7. 토픽 개수 별 일관도 비교(LSI)

[그림 8]는 HDP 알고리즘의 토픽 개수 별 일관도를 나타낸 것이다. 토픽의 개수가 적을 때는 일관도가 증가하는 행태를 보이지만 많을 때는 다시 일관도가 줄어들어 신뢰도가 낮아지게 된다.

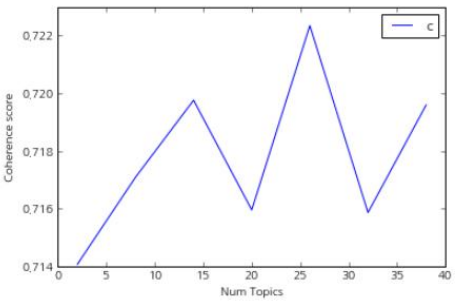

그림 8. 토픽 개수 별 일관도 비교(HDP)

주제 분류 알고리즘을 비교해 본 결과 입력 데이터의 대분류 개수와 비슷한 7개가 최적의 토픽으로 확인할 수 있다. 주제 분류 알고리즘의 성능은 LDA가 좋은 것으로 확인할 수 있다. [그림 9]는 LDA 알고리즘의 토픽 개수 별 복잡도를 나타낸 것이다. 복잡도는 다음과 같이 표기될 수 있다[그림 10]. w는 단어이며 d는 문서를 나타낸다. p(w)는 클수록 좋은 inference이므로 exp(-log(p(w)))는 작을수록 좋다.

$$Perplexity(w) = \exp[-\frac{\log p(w)}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{jd}}]$$

그림 10. 복잡도 수식 공식

토픽의 개수가 많아지면 복잡도는 낮아지지만 분류 수가 많아지므로 신뢰성이 낮아지기 쉽다.

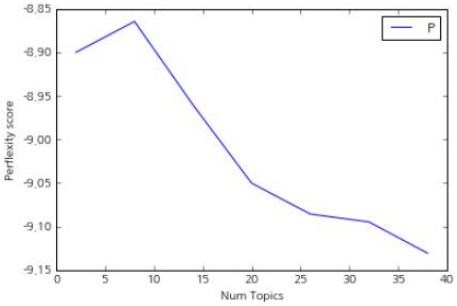

그림 9. 토픽 개수 별 복잡도(LDA)

[그림 11]은 각 주제 분류 알고리즘을 테스트 집합(종별 700개의 논문)에 적용시켰을 때에 대분류 별 논문 수를 표시한 것이다. HDP의 경우에는 많은 주제 분류가 나오긴 하지만 대부분 논문이 1개 또는 2개가 많기 때문에 최소 5개 이상인 분류만 표기하였다. HDP는 주로 세부 소분류에 적합한 알고리즘으로 생각된다.

대분류	LSI	LDA	HDP
Technology	2277	2995	3220
Science	762	264	77
Computer Science	195	182	16
Social Science	115	46	23
Arts	113	20	17
Geography, History	35	8	5
Philosophy and Psychology	3	5	3

그림 11. 주제 분류 알고리즘 테스트 집합 결과 비교

LSI와 HDP는 LDA보다 Social Science와 Arts 부분에서 차이가 있는 것으로 확인되지만 일관성이 적기 때문에 신뢰성이 다소 떨어진다. LDA, LSI, HDP 모델의 경우 특정 주제에 대해서 많은 논문이 몰린 것을 확인할 수 있다. 이 이유는 국가 학술정보(OCEAN) 데이터는 대부분 인문, 사회 분야 보다는 과학 분야의 논문이 많이 있다는 것을 반증하는 것이다. LDA 알고리즘은 일관도와 복잡도를 고려했을 때 다른 알고리즘보다 좋은 결과를 보였다.

III. 결론

본 논문에서는 국가 연구 성과물 중에 하나인 학술정보 메타데이터(OCEAN)를 활용한 주제 분류 알고리즘의 성능 테스트 및 비교를 해보았다. 실제 국가 논문 성과물 데이터를 바탕으로 주제 분류를 시도했다는 점에서 본 논문에서는 의의가 있으며, 데이터의 특성과 목적에 맞게 분류 알고리즘을 사용해야 한다는 것을 확인할 수 있었다. LSI는 보다 직관적인 데이터 집합에 활용하고 LDA는 다양한 키워드가 분류되어 새로운 용어를 적용하는데 유리하며, HDP는 보다 세분화된 분류 체계에 적용하는데 유리해 보인다. 본 연구의 한계점은 LDA 등의 알고리즘은 키워드에 민감한 결과를 보이기 때문에 키워드의 세밀한 정제가 필요하다는 것이다. 대분류를 기준으로 신뢰도가 향상이 되면 논문 단위의 주

제 분류가 되어 실제 다양한 분야의 기관과 연구자에게 필요한 주제 분류 서비스가 가능해 질 것이다. 알고리즘 비교 분석에서 데이터의 특성 별로 유용한 알고리즘은 존재하였으며, 목적에 맞는 올바른 알고리즘의 선택이 필수가 됨을 확인할 수 있었다. 향후에는 좀 더 효과적인 주제 분류를 위한 연구 개발을 진행할 계획이다.

참 고 문 헌

[1] 김무철, “과학기술용어 간 관계 도출을 위한 토픽 분석 연구,” 한국전자거래학회지, 제21권, 제1호, pp.119-129, 2016.

[2] 배덕호, 엄태환, 윤석호, 박정, 김상욱, “LDA를 이용한 논문 유사도 계산 방안의 성능 평가,” 한국통신학회 학술대회논문집, pp.356-357, 2013.

[3] 박준형, 오효정, “국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교,” 한국도서관·정보학회지, 제48권, 제4호, pp.235-258, 2017.

[4] 서경희, 이민수, 오상윤, “Spark 를 사용한 LDA 기반의 비정형 의료 데이터의 토픽 분석,” 한국통신학회 학술대회논문집, pp.61-63, 2016.

[5] 문병주, 송주의, 임현근, 정일품, “Hybrid-LDA 기반의 평가위원 추천시스템,” 한국통신학회 학술대회논문집, pp.1051-1052, 2016.

[6] 조태민, 이지형, “LDA 모델을 이용한 잠재 키워드 추출,” 한국지능시스템학회 논문지, 제25권, 제2호, pp.180-185, 2015.

[7] 심준식, 김형중, “LDA 토픽 모델링을 활용한 판례 검색 및 분류 방법,” 전자공학회논문지, 제54권, 제9호, pp.67-75, 2017.

[8] 정병문, 김태환, 이진, 김정선, “LDA 모델을 이용한 트위터 토픽 추출 및 토픽 카테고리 판단,” 한국정보과학회 학술발표논문집, pp.787-788, 2015.

[9] 봉성용, 황규백, “Labeled LDA를 이용한 저자 주제어 추천,” 한국정보과학회 학술발표논문집, Vol.37(1C), pp.385-389, 2010.

[10] 표신지, 김은희, 김문철, “토픽 모델링을 이용한 유사 시청 사용자 그룹핑 및 TV 프로그램 추천 알고리듬,” 한국방송미디어공학회 학술발표대회 논문집, pp.116-119, 2012.

[11] 박자현, 송민, “토픽모델링을 활용한 국내 문헌정보학 연구동향 분석,” 정보관리학회지, 제30권, 제1호, pp.7-32, 2013.

[12] L. Zheng, Z. Caiming, and C. Caixian, “MMDF-LDA: An improved Multi-Modal Latent Dirichlet Allocation model for social image annotation,” Expert Systems with Applications, Vol.104, pp.168-184, 2018.

[13] B. Cao, J. Liu, Y. Wen, H. Li, Q. Xiao, and J. Chen, “QoS-aware service recommendation based on relational topic model and factorization machines for IoT Mashup applications,” Journal of Parallel and Distributed Computing, 2018.

[14] Y. Papanikolaou and G. Tsoumakas, Subset Labeled LDA for Large-Scale Multi-Label Classification (2017, September 16), arXiv.org.

[15] M. Pavlinek and V. Podgorelec, “Text classification method based on self-training and LDA topic models,” Expert Systems with Applications, Vol.80, pp.83-93, 2017. http://doi.org/10.1016/j.eswa.2017.03.020

[16] M. Rani, A. K. Dhar, and O. P. Vyas, “Semi-automatic terminology ontology learning based on topic modeling,” Engineering Applications of Artificial Intelligence, Vol.63, pp.108-125, 2017. http://doi.org/10.1016/j.engappai.2017.05.006

[17] 곽창욱, 김선중, 박성배, 김권양, “무한 사전 온라인 LDA 토픽 모델에서 의미적 연관성을 사용한 토픽 확장,” 정보과학회 컴퓨팅의 실제 논문지, 제22권, 제9호, pp.461-466, 2016.

[18] 이호경, 양선, 고영중, “비격식 문서 분류 성능 개선을 위한 LDA 단어 분포 기반의 자질 확장,” 정보과학회논문지, 제43권, 제9호, pp.1008-1014, 2016.

[19] https://radimrehurek.com/gensim/

저 자 소 개

최 원 준(Won-Jun Choi) 정회원

- 1999년 3월 ~ 2006년 2월 : 원광대학교 수리통계학(학사)
- 2013년 3월 ~ 2017년 2월 : 한국과학기술연합대학교 과학기술정보학(공학석사, 공학박사)
- 2017년 2월 ~ 현재 : 한국과학기술정보연구원 연구원

<관심분야> : 과학기술정보, 네트워크 분석, 데이터 분석

설 재 욱(Jae-Wook Seol) 정회원

- 2007년 2월 ~ 2012년 8월 : 전북대학교 컴퓨터공학과(학사)
- 2012년 9월 ~ 2014년 8월 : 전북대학교 컴퓨터공학과(석사)
- 2014년 3월 ~ 현재 : 한국과학기술정보연구원 연구원

<관심분야> : 데이터마이닝, 기계학습, 자연어처리

정 희 석(Hee-Seok Jeong) 정회원

- 2002년 2월 : 충남대학교 컴퓨터공학과(공학사)
- 2004년 2월 : 충남대학교 컴퓨터공학과(공학석사)
- 2004년 3월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 데이터베이스 시스템, 빅 데이터, 분산처리

윤 화 묵(Hwa-Mook Yoon) 정회원

- 1997년 : 공주대학교 전자계산학과 졸업(석사)
- 2008년 : 배재대학교 컴퓨터공학과 졸업(박사)
- 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 텍스트마이닝, 지식관리, 빅 데이터