The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 267–278

DOI: https://doi.org/10.5351/KJAS.2021.34.2.267
Print ISSN1225-066X / Online ISSN 2383-5818

Research trends in statistics for domestic and international journal using paper abstract data

Jong-Hoon Yang[a], Il-Youp Kwak[1,a]

[a]Department of Applied Statistics, Chung-Ang University

Abstract

As time goes by, the amount of data is increasing regardless of government, business, domestic or overseas. Accordingly, research on big data is increasing in academia. Statistics is one of the major disciplines of big data research, and it will be interesting to understand the research trend of statistics through big data in the growing number of papers in statistics. In this study, we analyzed what studies are being conducted through abstract data of statistical papers in Korea and abroad. Research trends in domestic and international were analyzed through the frequency of keyword data of the papers, and the relationship between the keywords was visualized through the Word Embedding method. In addition to the keywords selected by the authors, words that are importantly used in statistical papers selected through Textrank were also visualized. Lastly, 10 topics were investigated by applying the LDA technique to the abstract data. Through the analysis of each topic, we investigated which research topics are frequently studied and which words are used importantly.

Keywords: text mining, word embedding, topic modeling

1. 서론

최근 국내외 빅데이터 및 분석 시장이 해마다 성장하고 있음에 따라 통계학 분야에서의 연구가 활발히 이루어지고 있다. 또한 코로나19로 인한 4차 산업 혁명의 가속화와 AI 산업에 대한 관심이 높아져 이에 따른 통계학 분야의 연구 동향 역시 변화가 있을 것으로 예상된다. 시선을 2010년대로 돌려보면 머신러닝과 딥러닝에 대한 관심이 커진 시기로 공공과 기업에서 빅데이터 관련 부서를 만들고 활용하면서 2010년대를 대표하는 기술이 되었다. 정보기술의 발달은 데이터양의 증가, 데이터 형태의 다양성, 데이터가 쌓이는 속도에 맞추어 데이터 속에 의 미있는 가치를 찾기 위해 통계분석의 중요성을 강조하고 있다. 따라서 통계학은 점점 더 폭넓은 분야의 연구가 이루어지고 있는 학문이므로 연구 동향에 대한 분석이 필요하다.

기존에 국내에서는 여러 분야에 대해 논문 데이터를 이용한 연구 동향 분석이 이루어져 왔다. Kim (2020)은 미국산업응용수학 학회지 논문의 제목과 초록 데이터를 수집한 후, LDA기법과 시계열 회귀 모형을 적용해 연구의 흐름을 파악하며 향후 교육과정과 연구 방향에 대해 시사점을 제시하였다. Jeon 등 (2017)은 인공지능 분야의 연구 동향을 파악하기 위해 인공지능 관련 논문의 초록 데이터를 수집한 후, 전처리 과정을 거친 텍스트를 활용하여 형태소 분석을 통한 단어들을 추출하였다. 이 단어들을 통해 연도별, 국가별 동시 출현

This research was supported by the Chung-Ang University Research Scholarship Grants in 2020.
[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjakgu, Seoul 06974, Korea. E-mail: ikwak2@cau.ac.kr

Published 30 April 2021 / journal homepage: http://kjas.or.kr

단어의 빈도를 이용하여 인공지능 분야의 연구동향을 제공하였다. Choi와 Lee (2017)는 태권도학 연구논문의 키워드를 활용하여 시계열 관점에서 태권도학 연구의 지식구조에 대해 정보를 제공하고 키워드 분석을 통해 연구 동향에 대한 정보를 제공하였다. Lee 등 (2015)은 제조시스템 분야의 논문을 수집한 후, 논문 키워드의 빈도를 기반으로 주제어를 선택하여 시각화를 통해 연구 동향을 파악하면서 향후 연구 방향을 제시하였다.

몇몇 학문 분야에 대해 연구 동향 파악을 위해 텍스트마이닝 기법을 이용하여 분석한 연구들이 있었으나 주로 키워드 빈도를 기반으로한 분석이 주가 되었으며, 통계학 분야 논문 데이터를 사용하여 연구 동향에 대한 분석은 진행되지 않았다. 또한 국내외의 연구 동향을 비교하는 연구 역시 이루어지지 않아 국내의 연구가 해외의 연구와 비교하여 어떤 유사점과 차이점이 있는지 파악할 필요가 있다.

이에 따라 본 연구에서는 2002년부터 2020년 6월 10일까지의 키워드 및 초록 데이터를 활용하여 통계학 연구 분야의 전반적인 동향을 파악하고자 한다. 키워드 데이터를 이용하여 국내와 해외에서 저자들이 선정한 주요 키워드를 제시하며 공통으로 관심이 있는 분야와 차이가 있는 분야를 제시하고자 한다. 초록 데이터를 이용하여 시각적으로 주요 키워드 간의 관계를 살펴보며 분야별 토픽을 묶어 파악하려 한다.

2. 분석방법

2.1. 자료수집

한국 및 해외에서 통계학 연구들이 어떻게 진행되어 왔는지 동향을 파악하기 위해 국내와 해외의 통계학 논문 초록 데이터를 이용하였다. 국내 통계학 논문 자료는 한국통계학회에 등록된 통계학 학술지인 응용통계연구, Communications for Statistical Applications and Methods (CSAM), 그리고 Journal of the Korean Statistical Society (JKSS)에서 2002년부터 2020년 6월 10일까지 등재된 논문을 검색하였다. 각각 응용통계연구 1,432편, CSAM 1,239편, JKSS 832편으로 총 3,503편이 선정되었다. 해외 통계학 논문을 대표하는 학술지로는 해외 최고수준의 연구들과 비교하기 위하여, Annals of Statistics와 Journal of the American Statistical Association (JASA)를 선택하여 국내와 마찬가지로 2002년부터 2020년 6월 10일까지 등재된 논문을 검색하였다. 각각 Annals of Statistics에서 2,061편과 JASA에서 2,801편으로 총 4,862편을 선정하였다. 최종적으로 국내 3,503편과 해외 4,862편으로 총 8,365편의 논문 자료를 이용해 초록 데이터 분석을 진행하였으며 논문별로 저자, 발행 연도, 영문초록, 키워드, 인용 횟수 등의 자료를 수집하였다. 해당 논문 자료들을 활용하여 빈도 분석, Word Embedding, LDA, t-SNE 등의 분석을 활용하였다.

2.2. 빈도분석

국내외 통계학 연구들에 대해 초록 데이터와 키워드를 통한 빈도 분석을 수행하였다. 국내와 외국 논문 키워드 비교의 통일성을 위해 국내 논문에서 한국어 키워드와 영어 키워드 중 영어 키워드를 이용하였다. Python Numpy, Pandas, Collections 모듈을 활용해 키워드별 사용 횟수, 연도별 논문 수, 빈도수 상위 키워드들의 연도별 사용 횟수를 파악할 수 있는 프로그램을 작성하였다. 분석과정에 있어 동일한 의미를 가지지만 대문자, 소문자, 띄어쓰기, 약어 등으로 다르게 파악되는 키워드들은 사전에 병합하여 분석하였다.

2.3. TextRank

TextRank는 Google의 PageRank의 알고리즘을 이용한 것으로 PageRank (Brin과 Page, 1998)는 하이퍼링크를 가지는 웹 문서의 중요도에 따라 가중치를 매겨 Rank를 부여한 후, Rank가 높은 사이트가 다른 사이트에서 많이 참고하는 페이지로 해석된다. 이러한 알고리즘을 바탕으로 TextRank (Mihalcea와 Tarau, 2004)는 문서

내의 문장 또는 단어를 이용하여 문장 또는 단어에 가중치를 부과하여 Ranking을 매긴다. 본 연구에서는 논문 초록 문서를 이용하여 각 단어의 중요도를 계산한 뒤, 중요도가 높은 단어를 추출하였다.

2.4. Word Embedding

Word Embedding이란 단어를 n 차원의 밀집 벡터(dense vector)의 형식으로 나타내는 것을 말하며 그 종류로는 LSA, Word2Vec, FastText, Glove 등이 있다 (Yin와 Shen, 2018; Landauer 등, 1998; Goldberg와 Levy, 2014; Joulin 등, 2016; Pennington 등, 2014). 본 연구에서는 단어 간 유사도를 반영할 수 있도록 단어의 의미를 벡터화하기 위해 Word2Vec을 사용하였다. Word2Vec에서는 Continuous Bag of Words (CBOW)와 Skip-Gram 두 가지 방식이 존재한다 (Mai 등, 2019; Mikolov 등, 2013). CBOW는 주변에 있는 단어들을 이용하여 중간에 있는 단어들을 예측하는 방법이며 Skip-Gram은 중간에 있는 단어로 주변 단어들을 예측하는 방법이다. 이후의 과정에서 중간에 있는 단어를 표적 단어, 주변 단어를 문맥 단어라고 지칭한다. 여기서 Skip-Gram은 하나의 은닉층(hidden layer)로 이루어진 간단한 뉴럴 네트워크(neural network) 구조를 이루고 있다. 타겟 단어를 입력값으로 받아 임베딩하려는 단어의 개수를 V, 노드(node)의 개수를 N으로 했을 때, $V \times N$ 가중치 행렬 W를 업데이트하면서 학습이 이루어진다. 식 (2.1)을 최대화하는 방향으로 학습이 진행되며 식 (2.1)의 좌변은 조건부 확률로, 표적 단어(t)가 주어졌을 때, 문맥 단어(m)가 나타날 확률을 나타낸다. 여기서 v_t는 입력층(input layer)과 은닉층 사이의 가중치 행렬 W에서 주어진 타겟 단어에 해당하는 행 벡터이고, u_w는 은닉층과 출력층(output layer) 사이의 가중치 행렬 W'의 문맥 단어에 해당하는 열벡터를 의미한다.

$$P(m|t) = \frac{\exp\left(u_m^T v_t\right)}{\sum_{w=1}^{W} \exp\left(u_w^T v_t\right)}. \tag{2.1}$$

우변의 분자 값을 키우는 것은 표적 단어의 벡터와 문맥 단어의 벡터의 내적을 키우는 것으로 벡터의 내적은 코사인값과 연관되어 있어서 두 단어의 코사인 유사도를 높이는 것으로 볼 수 있다. 이러한 학습 과정을 거친 파라미터 행렬 W가 최종 결과물이 된다. 본 연구에서는 성능 비교를 진행한 여러 논문의 자료를 기반으로 CBOW보다 성능이 좋은 Skip-Gram을 사용하였다 (Rong, 2014).

2.5. t-SNE

t-distributed stochastic neighbor embedding (t-SNE)는 데이터 차원축소(dimesionality reduction)와 시각화(visualization)의 방법으로 고차원의 데이터를 저차원 공간에 시각화하는 알고리즘이다 (Maaten과 Hinton, 2008). $x_1, x_2, \ldots, x_N$의 데이터가 있을 경우 x_i와 x_j의 유사도를 식 (2.2)라고 했을 때,

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}. \tag{2.2}$$

고차원에서 x_i와 x_j의 서로에 대한 유사도 값 p는 식 (2.3)과 같이 정의한다.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{2.3}$$

그리고 사영시키고자 하는 저차원에서 y_i와 y_j의 서로에 대한 유사도는 식 (2.4)로 정의한다.

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}, \tag{2.4}$$

여기서 y_i, $i = 1, \ldots, N$은 차원 축소된 공간에서의 x_i값들이다. 위 두가지 수식을 통해 고차원 상에서 x_i들의 분포와 저차원 상에서 y_i들의 분포에 대한 유사도를 KL divergence 식 (2.5)를 이용하여 측정한다.

$$\mathrm{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2.5}$$

이 KL divergence 식 (2.5)는 y_i, $i = 1, \ldots, N$에 대한 함수이며, 이 식을 손실함수로 정의하고, 손실함수를 최소화 시키는 방향으로 학습을 진행한다. 이 학습의 결과로 추정된 $\hat{y}_i$가 t-SNE를 통해 차원 축소된 공간의 데이터이다.

Multi-dimensional scaling (MDS), ISOMAP, locally linear embedding (LLE) 모두 고차원의 정보를 저차원으로 축소하는 알고리즘이며 pincipal components analysis (PCA) 또한 차원축소의 기능을 가지고 있지만 t-SNE가 다른 알고리즘들에 비하여 저차원 공간상에 군집을 더 잘 시각화 시켜주는 경향이 있어 최근 많이 이용되고 있다 (Cox와 Cox, 2000; Roweis와 Saul, 2000; Jolliffe, 1986).

2.6. LDA

토픽모델링(topic modeling)은 문서의 집합에서 어떤 토픽이 존재하는지 알아내는 알고리즘으로 Papadimitriou 등 (1998)에 의해 처음 제안되었다. 토픽모델링을 위해서 잠재의미분석(latent semantic analysis; LSA), 확률적 잠재의미분석(probabilistic latent semantic analysis; PLSA), 잠재 디리클레 할당(latent dirichlet allocation; LDA) 등의 방법이 제안되었고, 그 중 Blei 등 (2003)이 제안한 LDA가 가장 많이 활용되고 있다. 알고리즘으로 문서 안에 토픽들이 존재한다는 가정을 하며, 전체 문서에 Gibbs Sampling 기법을 이용하여 토픽에 단어가 나타날 확률을 최대화해주는 토픽을 찾는 것이 최종 목표이다.

LDA는 α와 β를 모수로 가지는 디리클레(Dirichlet) 분포를 따르는 문서의 주제분포와 주제의 단어분포를 가정한다. 여기서 α는 문서 내 주제의 밀집도를 나타내고 β는 주제 내 단어의 밀집도를 나타내며 이 값들이 1에 가까워질수록 밀집도가 높아지는 것을 의미한다. 실제 문서에 포함된 단어를 관측해 나가면서 문서의 주제분포에서 하나의 주제를 정해 단어에 부여한다. 그리고 다시 이 주제에서 단어를 선택하고 다시 앞에서부터 반복하며 디리클레 분포를 업데이트하는 방식으로 문서 생성 과정을 모델링하는 것이다. 이 과정을 통해 문서 내의 단어들이 어떤 토픽에 배정되어야 하는지 추측할 수 있다.

LDA 기법에서는 단지 단어들의 빈도에만 중점을 둔 텍스트 데이터의 수치화 표현 방법인 Bag of Words (BoW), 여러 문서에 등장하는 각 단어들의 빈도를 행렬로 표현한 Document-Term Matrix (DTM)과 DTM 내의 각 단어들의 중요한 정도를 가중치로 주는 Term Frequency-Inverse Document Frequency (TF-IDF) 행렬을 입력값으로 받는다 (Brownlee, 2020). 이는 LDA는 단어의 순서와는 관계없이 오직 빈도만으로 계산되는 알고리즘임을 알 수 있다.

3. 분석결과

3.1. 데이터 특성

본 연구에 사용된 데이터는 2002년부터 2020년 6월 10일까지의 통계학 논문 국내 3,503편과 해외 4,862편으로 총 8,365편의 자료를 이용하였다. 통계학은 꾸준히 연구되어 왔던 분야 중 하나로 국내 및 해외에서 매년 100편 이상의 논문이 발표됐다. Figure 1은 국내외 연도별 논문 수에 대한 막대그래프이다. 국내 통계학 연구의 경우 2009년에 가장 많은 수의 논문이 발표되었으며 2008년을 기점으로 200편을 돌파하며 활발한 연구가 이루어졌다. 해외 통계학 연구의 경우 마찬가지로 2009년에 301편으로 가장 많은 수의 논문이 발표되었으며

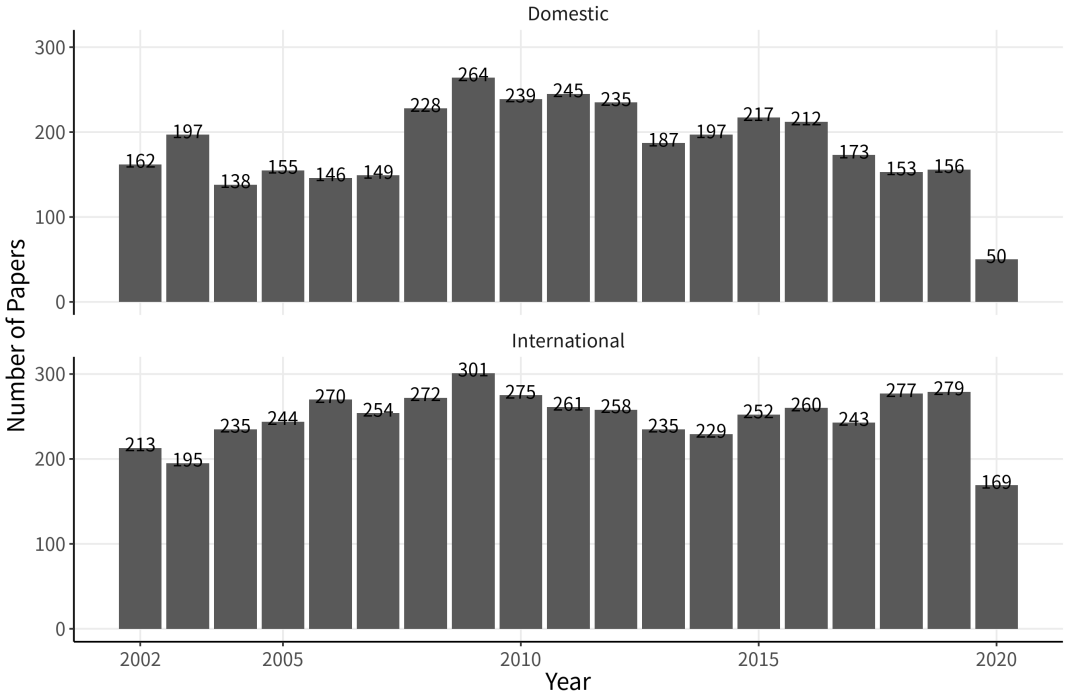

Figure 1: *Number of papers by year.*

이후에도 큰 변동 폭 없이 꾸준한 연구가 진행 중이다. 국내와 해외 통계학 연구 모두 2000년대 초중반보다 2000년대 후반에 들어서 더 많은 연구가 이루어지고 있으며 국내 연구에서 확연한 차이를 보인다. 하지만 국내에서는 2009년을 기점으로 양적인 연구의 성과가 조금씩 떨어지나 해외에서는 꾸준히 200편 이상의 논문이 발표되며 높은 성과를 보인다.

3.2. 키워드 분석 결과

저자들이 논문에 제시한 키워드들의 빈도를 분석한 결과를 Table 1에서 나타낸다. 국내 논문의 경우 총 3,503편의 논문에서 12,808개의 키워드를 뽑았으며 중복된 키워드들을 제외하면 8,609개의 키워드가 있다. 해외 논문의 경우 총 4,862편의 논문에서 24,158개의 키워드를 뽑았으며 중복된 키워드들을 제외하고 11,608개의 키워드가 있다. 이 중 상위 60개의 키워드를 보며 약어 또는 기호에 의해서 다르게 구별된 키워드들을 정리하여 다시 국내외 각각 상위 30개의 키워드를 정리하여 2002년부터 2020년까지의 활발한 통계학 연구 분야들에 대해 살펴보았다.

국내에서는 Maximum Likelihood Estimator가 71회로 가장 높은 빈도수를 가진 키워드였으며 Variable Selection이 52회, Outliers와 Bootstrap이 각각 43회와 40회로 2, 3, 4번째에 있다. 해외에서는 Marcov Chain Monte Carlo가 171회로 가장 높은 빈도수를 가진 키워드로 나타났으며 High-Dimensional Data가 153회 Sparsity, Nonparametric Regression가 각각 131회, 128회로 뒤를 따랐다. 국내외에서 모두 높은 빈도를 보이는 키워드는 Marcov Chain Monte Carlo, Maximum Likelihood, Nonparametric, Bayesian Inference, Variable Selection, Bootstrap, Consistency, Principal Component Analysis 등이 있었다. 해당 연구주제들은 국내외에서 모두 활발하게 연구가 이루어지고 있는 것으로 보인다. 국내에는 적지만 해외에서 더 많이 등장하는 키워드

Table 1: Keyword frequency (Top 30)

Rank	Keyword (Domestic)	Rank	Keyword (International)
1	Maximum Likelihood Estimator(71)	1	Marcov Chain Monte Carlo(171)
2	Variable Selection(52)	2	High-Dimensional Data(153)
3	Outliers(43)	3	Sparsity(131)
4	Bootstrap(40)	4	Nonparametric Regression(128)
5	Bayesian Inference(40)	5	Bayesian Inference(122)
6	EM Algorithm(38)	6	Variable Selection(115)
7	Logistic Regression(34)	7	Model Selection(114)
8	Clustering(34)	8	Functional Data Analysis(106)
9	Marcov Chain Monte Carlo(34)	9	Bootstrap(100)
10	Classification(31)	10	Consistency(97)
11	Principal Component Analysis(25)	11	Principal Component Analysis(94)
12	Gibbs Sampling(25)	12	False Discovery Rate(90)
13	Regression(25)	13	Causal Inference(89)
14	Bias(24)	14	Lasso(88)
15	Mean Squared Error(23)	15	Maximum Likelihood(86)
16	Goddness-of-fit Test(23)	16	Asymptotic Normality(86)
17	Survival Analysis(23)	17	Missing Data(72)
18	Longitudinal Data(21)	18	Multiple Testing(70)
19	Support Vector Machine(21)	19	Empirical Process(70)
20	Missing Data(21)	20	Dimension Reduction(69)
21	Linear Regression(20)	21	Robustness(65)
22	Consistency(20)	22	EM Algorithm(61)
23	Imputation(20)	23	Mixture Model(60)
24	Malliavin Calculus(19)	24	Classification(58)
25	Random Forest(19)	25	Hypothesis Testing(58)
26	Lasso(19)	26	Gaussian Process(56)
27	Random Effects(19)	27	Regularization(53)
28	Nonparametric(18)	28	Density Estimation(51)
29	Data Mining(18)	29	Measurement Error(50)
30	Quantile Regression(18)	30	Smoothing(50)

들은 High Dimensional Data, Sparsity, Functional Data Analysis, False Discovery Rate, Causal Inference 등이 있었고, 국내에서 비교적 더 많이 등장하는 키워드들에는 Logistic Regression, Gibbs Sampling, Bias, Mean Squared Error, Goodness of fit Test 등이 있었다. 또한, 국내 논문에서는 Survival Analysis, Support Vector Machine, Random Forest 등의 세부 분야에 대한 키워드들이 상위에 있는 반면 해외 논문에서는 통계학 분야에서 자주 사용되는 거시적인 의미를 가진 키워드들이 주로 나타났다.

3.3. Word Embedding

초록은 한 논문의 전체 내용을 대표할 수 있는 부분이기 때문에 초록 데이터를 따로 수집하여 Word Embedding에 적용했다. 국내 논문 3,503편 중에서 영문초록 데이터가 있는 논문 3,276편과 해외 논문 4,862편은 모두 영문초록 데이터가 존재하여 총 8,138편의 논문에 대하여 Word Embedding을 통한 단어 간의 관계를 확인하였다.

Figure 2는 Word Embedding을 통해 초록 데이터에 있는 명사 단어들을 100차원 벡터로 표현한 후, *t*-SNE

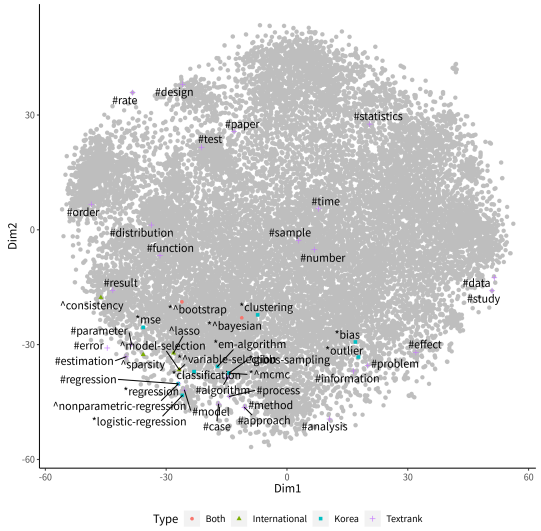

Figure 2: *Semantic correlations among words visualized by t-SNE.*

를 통해 2차원 데이터로 차원 축소하여 나타낸 그래프이다. 전체 명사 단어들은 회색 점으로 표현되었고, 선정된 주요 단어들을 그 위에 표기하였다. 한국 논문과 외국 논문 각각 키워드 분석을 통해 찾은 Table 1의 키워드 중 Top 15개씩 총 30개의 키워드와 Textrank 기법을 통해 가장 중요도가 높은 상위 30개의 키워드를 뽑아 전체 60개의 키워드가 선정된 주요 단어들이다. 표기된 단어 앞에, 해외 논문들에서 선정된 키워드는 ∧를, 국내 논문들로부터 선정된 키워드는 *표기를, Textrank로 선정된 키워드들은 #표기를 하였다. *t*-SNE를 통해 각 중요 키워드 간의 연관성을 시각적으로 관찰해 볼 수 있다.

결과를 보면 국내, 국외 논문의 저자들이 선정한 키워드들은 좌 하단에 몰려있는 것을 볼 수 있었고, 반면에 Textrank를 사용해 선정된 단어들은 비교적 고르게 퍼져 있는 것을 알 수 있었다. Word Embedding 기법에 의해서 단어들간의 의미론적 관계가 보전되도록 벡터화되기 때문에, 비슷한 단어들이 근처에 위치하게 된다. 통계학 분야 페이퍼의 저자들은 그들이 쓴 페이퍼의 통계 방법론과 관련된 단어들을 키워드로 사용하는데, 이 단어들이 의미론적으로 비슷한 맥락에서 많이 사용되다보니 좌하단에 함께 위치하게 된 것으로 보인다. 반면 Textrank 를 통해 선정된 단어들은 통계학 학문 분야 논문들에서 중요하게 사용되는 단어들이며, 전체 단어 공간에 고르게 퍼져있었다. 왼쪽 아래 부분의 방법론 키워드 외에 Textrank에 의해 선정된 단어로는 rate, design, test, time, sample, number, data, study 등이 있었다.

3.4. LDA

초록 데이터를 활용하여 LDA를 통해 문서들을 10개의 토픽으로 분류해 보았다. LDA를 통한 토픽 모델링 기법은 문서에 담긴 단어들을 통해 주제를 찾아내며 그 주제에 포함되는 키워드들을 나타내 주기 때문에 이를 통해 해당 토픽에 대한 의미를 찾아 해석할 수 있다. 토픽별로 TF-IDF 행렬을 통해 구해진 상위 10개의 단어를 살펴본 후, 토픽들이 각각 어떤 연구주제에 따라 분류되었는지 확인하였다.

Figure 3은 Python의 pyLDAvis 모듈에서 Carson과 Kenneth가 제안한 LDAviz (2014)을 사용해, PCA를 통해 2차원 그래프로 토픽별 분류를 시각화한 자료와 해당 토픽의 상위 10개의 단어를 막대그래프로 나타낸 그림이다. 먼저 PCA 플롯은 토픽 벡터(topic vector)를 2차원으로 축소하여 각 토픽 간의 관계성을 파악할 수 있고, 비슷한 좌표에 위치한 토픽들은 비슷한 맥락을 가진다. 또한 표기된 토픽의 원의 크기로부터 해당 토픽

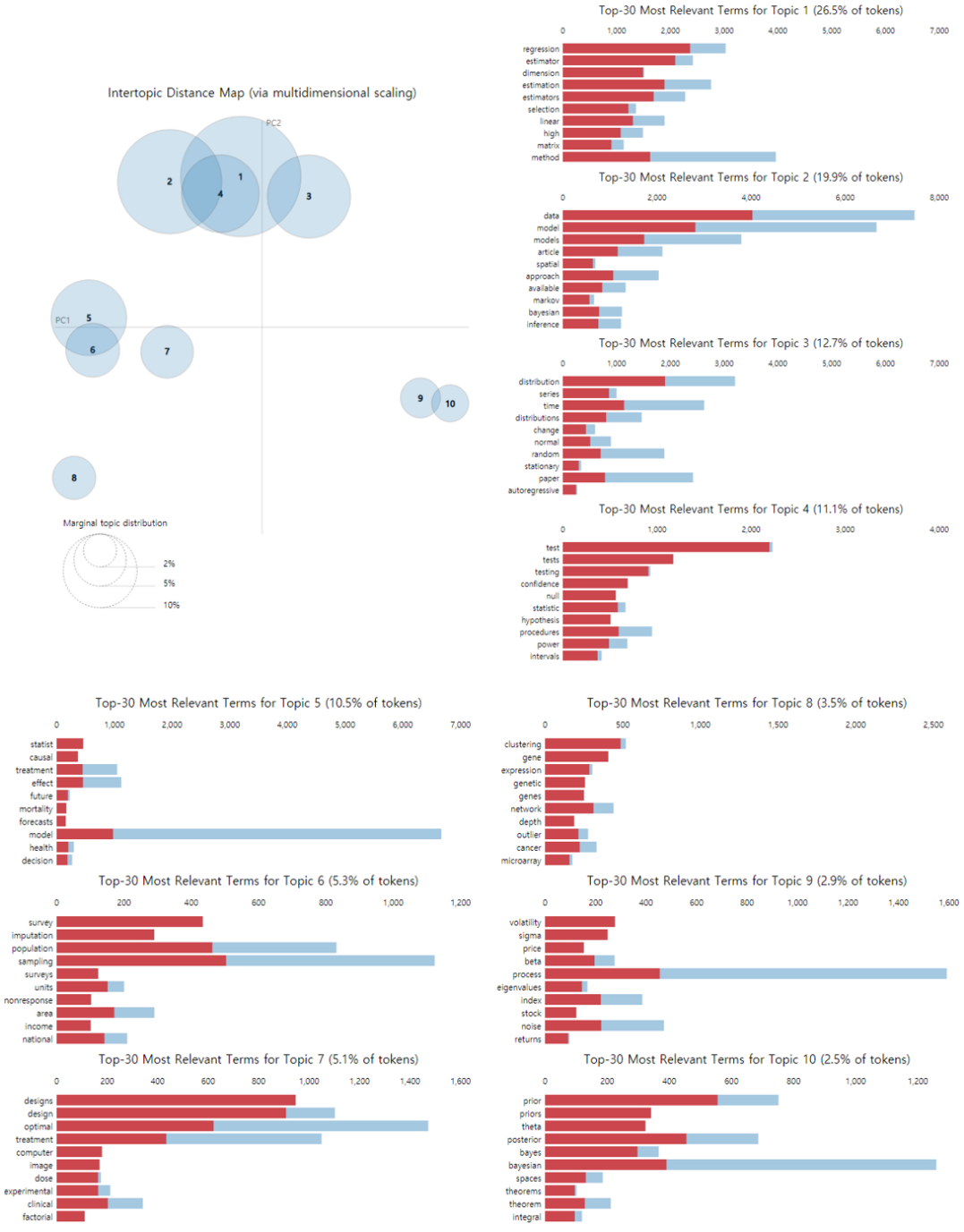

Figure 3: *Top 10 keywords of topics.*

을 갖는 문서 집합의 크기를 알 수 있다. 이를 위해 LDAvis는 특정 토픽에 대해서 해당 단어가 나타날 확률 벡터에 PCA를 적용하여 2차원의 형태로 나타낸다.

각 토픽별 상위 10개 단어를 살펴보면, Topic 1은 regression, estimator, estimation 등을 주요 단어로 가지며, 이를 통해 이 토픽은 회귀 및 추정에 관한 것으로 파악할 수 있다. Topic 2에서는 data, model, models, approach 등이 자주 나타나며 데이터 모델링에 관련된 토픽으로 지정하였다. Topic 3은 series, time, stationary, autoregressive 등이 영향을 많이 주는 단어로 나타나 시계열 관련 토픽임을 알 수 있다. Topic 4는 test, null, statistic, hypothesis, power 등의 단어들에 높은 가중치를 주어 가설검정을 주제로 선택하였다. Topic 5는 causal, treatment, effect 등의 단어를 통해 인과추론이라는 주제를 선택했다. Topic 6에서는 survey, imputation, population, sampling 등이 영향을 주는 단어로 나타나 표본조사 관련 토픽으로 설정하였다. Topic 7은 design, designs, optimal, treatment 등의 단어들이 높은 가중치를 가져 실험설계 관련 토픽임을 알 수 있다. Topic 8은 gene, genetic, cancer, micorarray를 통해 생물통계 관련 분야임을 파악했고 network, depth 등의 단어들을 통해 딥러닝 관련 주제들도 같은 토픽으로 분류된 것으로 보여 생물통계 및 딥러닝 이라는 주제로 지정하였다. Topic 9에서는 sigma, beta, process 등이 문서에 많이 나타나 공정관리 관련 토픽으로 설정했다. 마지막으로 Topic 10은 prior, posterior, bayes, bayesian 등이 토픽 분류에 영향을 주는 단어로 베이지안 관련 토픽임을 알 수 있다. 따라서 토픽에 해당하는 문서의 수가 가장 많은 토픽 순으로 번호를 지정하여 각 토픽의 의미는 “회귀 & 추정”, “데이터 모델링”, “시계열”, “검정”, “인과추론”, “통계 조사”, “실험 설계”, “생물통계 & 딥러닝”, “공정관리”, “베이지안”으로 정의하였다.

Figure 3의 PCA 결과를 통해, “회귀 & 추정”, “데이터 모델링”, “시계열”, “검정” 토픽들과 “인과추론”, “통계 조사”, “실험 설계” 토픽들 그리고 “공정관리”, “베이지안”이 근처에 위치한 것을 확인할 수 있다.

“회귀 & 추정”, “데이터 모델링”, “시계열”, “검정” 에서는 흔히 통계학 연구 분야에서 많이 쓰이는 단어들을 위주로 토픽이 이루어졌다.

Table 2는 각 토픽이 나타내는 연구주제와 이에 해당하는 국내외 논문의 개수를 요약한 것이다. 토픽별 문서의 수를 통해 통계학의 기본이 되는 “회귀 & 추정”이 가장 높은 빈도로 나타났으며 같은 맥락으로 “데이터 모델링”, “검정”, “인과추론” 등의 토픽이 빈도수 기반 상위 토픽에 있다. “시계열”, “생물통계 & 딥러닝”, “공정관리”, “베이지안”과 같이 세부적인 연구 주제들이 토픽으로 묶여있는 것을 보면 지난 18년 동안 통계학에서 해당 분야의 연구가 활발히 이루어진 것으로 볼 수 있었다. 국내 통계학 연구의 경우 시계열에 대한 연구가 23.6%의 비율로 가장 많이 이루어졌으며 그 뒤로 회귀 & 추정, 인과추론이 18.8%, 12.49%로 2002년부터 많은 연구가 진행되어 왔다. 데이터 모델링, 검정, 통계조사, 실험 설계 순으로 국내에서 활발한 연구가 있었고 상대적으로 생물통계 & 딥러닝, 공정관리, 베이지안과 같은 세부분야에 대해 관심이 높았다. 반면 해외 통계학 연구의 경우 41.97%와 24.53%의 비율로 반 이상의 연구가 회귀 & 추정에 대해 진행되었다. 다음으로 검정을 주제로 한 연구가 10.79%였으며 시계열과 인과추론을 주제로 한 연구가 8.95%, 6.39%를 차지하고 있다. 생물통계 & 딥러닝, 공정관리, 베이지안에 대해서는 국내에 비해 적은 비율의 연구가 이루어져 왔다. 이는 우리가 선정한 해외 논문들이 Annals of Statistics, JASA 두 저널에 한정되어 나타난 점도 있다.

4. 결론 및 시사점

본 연구에서는 국내외 통계학 분야의 논문에 대하여 Text Mining 기법들을 활용하여 국내와 외국에 대해 연구 동향을 비교 분석하였다. 기존에 통계학 분야에 대해서 연구 동향에 대한 연구가 부족하며 주로 키워드 위주의 분석이 이루어져 왔다. 본 연구에서는 키워드를 이용한 분석에 있어 Textrank 방법을 추가하여 키워드를 선정하였으며, 논문 초록 데이터를 이용해서 Word Embedding 기법과 Topic Modeling 기법의 하나인 LDA를 이용해서 연구 동향을 살펴보았다.

Table 2: The name of topics & number of topics

Topic	Topic name	Count(All)	Count(Domestic)	Count(International)
1	회귀 & 추정	2369 (31.77%)	616 (18.80%)	1753 (41.97%)
2	데이터 모델링	1415 (18.98%)	390 (11.90%)	1025 (24.53%)
3	시계열	1144 (15.35%)	770 (23.60%)	374 (8.95%)
4	검정	807 (10.83%)	356 (10.87%)	451 (10.79%)
5	인과추론	676 (9.07%)	409 (12.49%)	267 (6.39%)
6	통계 조사	365 (4.90%)	313 (9.55%)	52 (1.24%)
7	실험 설계	277 (3.72%)	150 (4.58%)	127 (3.04%)
8	생물통계 & 딥러닝	142 (1.91%)	110 (3.36%)	32 (0.77%)
9	공정관리	132 (1.77%)	79 (2.41%)	53 (1.27%)
10	베이지안	127 (1.70%)	83 (2.53%)	44 (1.05%)

키워드 데이터를 이용하여 살펴본 국내외 통계학 연구의 특징으로는 국내외 공통으로 Marcov Chain Monte Carlo, Maximum Likelihood, Nonparametric, Bayesian Inference, Variable Selection, Bootstrap, Consistency, Principal Component Analysis 등의 주제들이 주로 연구되는 것을 볼 수 있었고, 국외에서는 High Dimensional Data, Sparsity, Functional Data Analysis, False Discovery Rate, Causal Inference 등의 주제들이 더 많이 연구되고 있었다.

Word Embedding을 통해서 통계학 분야 논문의 키워드들은 방법론 위주로 선정된다는 것을 확인할 수 있었고, 방법론 단어들 외에도 Textrank를 통해 파악된 통계학 분야에서 잘 사용되는 단어들은 rate, design, test, time, sample, number, data, study 등이 있었다.

마지막으로 초록 데이터를 통한 LDA 분석에서는 국내외 연구논문들의 분류를 10가지로 나누어 살펴보았고, “회귀 & 추정”, “데이터 모델링”, “검정”, “인과추론”, “시계열”, “생물통계 & 딥러닝”, “공정관리”, “베이지안” 등의 주제들로 활발한 연구가 이루어지고 있는 것을 확인할 수 있었다.

본 연구의 한계점 및 앞으로의 연구 방향으로는 좀 더 다양한 해외 학술지 데이터의 수집이 있겠다. 본 연구에서는 해외 최고수준의 연구들과 국내 논문의 비교를 위해 JASA와 Annals of Statistics 두 저널의 데이터만을 연구에 고려하였지만, 해외 통계 연구 전체를 대표한다고 보기에는 무리가 있다. 해외의 더욱 다양한 통계 분야의 논문들을 골고루 수집해 분석 할 수 있다면 또 다른 재미있는 국내외 연구 비교가 이루어질 수 있을 것이다.

References

Blei MD, Ng YA, Jordan IM (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Brin S and Page L (1998). The anatomy of a large-scale hypertextual web search engine. *In Proceedings of the Seventh International Conference on World Wide Web 7*, 107–117.

Brownlee J (2020). A gentle introduction to the bag-of-words model. *In Deep Learning for Natural Language Processing*

Choi CH and LEE JB (2017). The knowledge structure analysis on Taekwondo researches : Application of keyword network analysis, *The Korean Journal of Physical Education*, **56**, 627–644.

Cox FT and Cox MAA (2000). *Multidimensional scaling* 2nd ed, Chapman and Hall.

Goldberg Y and Levy O (2014). Word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv*

Jeon YB, Ryu SR, Song JH, and Kim HJ (2017), Analysis of research trends in artificial intelligence using text mining techniques, *Proceedings of the Korea Intelligent Information Systems Society*, 39–40.

Jolliffe IT (1986). Principal Component Analysis, *Springer Verlag*.

Joulin A, Grave E, Bojanowski P, Douze M, Jegou H, and Mikolov T (2016). Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651.

Kim SY (2020). Analysis on status and trends of SIAM journal papers using text mining, *Journal of the Korea Contents Association*, **20**, 212–222.

Landauer TK, Foltz PW, and Laham D (1998). An introduction to latent semantic analysis, *Discourse processes*, **25**, 259–284.

Lee IS, Park SH, Baek JG (2015). Identification of research trends in the manufacturing system field through text mining, *Proceedings of the Spring Conference of the Korean Institute of Industrial Engineers*, 4201–4205

Maaten L and Hinton G (2008). Visualizing data using t-SNE, *Journal of Machine Learning Research*, **9**, 2579–2605.

Mai F, Galke L, and Scherp A (2019). CBOW is not all you need: Combining CBOW with the compositional matrix space model, *CoRR*.

Mihalcea R, Tarau P (2004). TextRank: bringing order into texts. *In Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J (2013). Distributed rep-resentations of words and phrases and their compositionality, *Neural and Information Processing System (NIPS)*

Papadimitriou C, Raghavan P, Tamaki H, and Vempala S (1998). Latent semantic indexing: a probabilistic analysis. *In Proceedings of ACM PODS*, 159–168.

Pennington J, Socher R, and Manning CD (2014). Glove: global vectors forword representation, *EMNLP*, **14**, 1532–1543.

Rong X (2014). Word2vec parameter learning explained. arXiv.

Roweis TS and Saul KL (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, **290**, 2323–2326.

Sievert C, Shirley K (2014). LDAvis: A method for visualizing and interpreting topics, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

Yin Z and Shen Y (2018). On the dimensionality of word embedding, *Advances in Neural Information Processing Systems 31*, 895–906.

Received January 11, 2021; Revised January 25, 2021; Accepted January 27, 2021

초록데이터를 활용한 국내외 통계학 분야 연구동향

양종훈[a], 곽일엽[1,a]

[a]중앙대학교 응용통계학과

요 약

시간이 갈 수록, 정부, 기업, 국내, 해외를 막론하고 데이터의 양이 증가하고 있다. 이에따라 학계에서도 빅데이터에 대한 연구들이 늘어나고 있다. 통계학은 빅데이터 연구의 중심이 되는 학문들 중 하나이며, 늘어나는 통계학 분야 논문 빅데이터를 통해 통계학의 연구동향을 파악해 보는 것도 재미있을 것이다. 본 연구에서는 국내와 해외의 통계학 논문들의 초록데이터를 통해 어떤 연구들이 이루어지고 있는지 분석을 진행하였다. 저자들이 선정한 논문들의 키워드 데이터 빈도를 통해 국내외 연구 동향을 분석하였고, Word Embedding 방법을 통해 해당 키워드들의 관계성을 시각화 하였다. 여기서 저자들이 선정한 키워드들 외에 Textrank를 통해 선정된 통계학 분야 논문들에서 중요하게 사용되는 단어들도 추가적으로 시각화 하였다. 마지막으로 초록 데이터에 LDA 기법을 적용하여 10가지 토픽을 알아보았다. 각 토픽들에 대한 분석을 통해 어떤 연구 주제들이 자주 연구되며, 어떤 단어들이 중요하게 사용되는지 알아보았다.

주요용어: 텍스트 마이닝, Word Embedding, 토픽 모델링

이 논문은 2020년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과. E-mail: ikwka2@cau.ac.kr