Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 1 pp. 414-420, 2022

https://doi.org/10.5762/KAIS.2022.23.1.414
ISSN 1975-4701 / eISSN 2288-4688

텍스트 마이닝에 기반한 Expert Systems with Applications 저널의 토픽 모델링 및 분석

조현우
대구대학교 융합산업공학과

Topic Modeling and Analysis of Expert Systems with Applications Using Text Mining

Hyun-Woo Cho
Department of Industrial Convergence Engineering, Daegu University

요 약 본 연구의 목적은 텍스트 마이닝의 토픽 모델링 기법을 이용하여 Expert Systems with Applications(이하 ESWA) 저널에 게재된 논문들의 주요 연구 토픽과 토픽 변화 추이를 분석하는 것이다. 2000년부터 2021년까지 게재된 총 11,918편의 ESWA 저널 논문을 대상으로 논문의 제목, 키워드, 초록 데이터를 수집하였으며 파이썬과 잠재 디리클레 할당모형(LDA: latent Dirichlet allocation, 이하 LDA) 알고리즘에 기반한 토픽 모델링 분석을 수행하여 최대 일관성 값(0.409)을 가지는 10개의 최적 토픽 그룹을 도출하였다. 10개의 토픽은 최적화 방법론 (16.4%), 지식 모델 (15.7%), 분류 활용 (10.8%), 특성 선택 (10.4%), 예측 모델 (10.0%) 등의 비중 순서로 나타났다. 이러한 토픽들의 시기별 변화 추이에서는 최적화 방법론, 지식 모델, 분류 활용 3가지 토픽들이 특정 시기를 기점으로 활발하게 연구되어 진 것을 알 수 있었다. 더불어 토픽별 논문의 피인용 분석을 통해 특성 선택의 피인용 값이 2009년~2012년 기간 타 토픽 대비 최대값을 보여주었다.

Abstract This research aims to analyze the major research topics of the Expert Systems with Applications (ESWA) journal and the trend changes in the topics by using the topic modeling method of text mining. Eleven thousand nine hundred eighteen ESWA articles published from 2000 to 2021 were automatically collected with titles, keywords, and abstracts. This text data was analyzed by Python and the latent Dirichlet allocation (LDA) algorithm to obtain ten topic groups with the highest coherence score (0.409). The ten topics obtained were optimization methodology (16.4%), knowledge model (15.7%), classification applications (10.8%), feature selection (10.4%), prediction model (10.0%), and so forth. From the viewpoint of the transition of these topics with time, it was found that the three topics of optimization methodology, knowledge model, and classification applications were actively studied from a specific time. In addition, by numerical analysis of the impact factor of articles, the topic of feature selection showed the maximum value compared to other topics from 2009 to 2012.

Keywords : Convergence, ESWA, LDA, Text Mining, Topic Modeling

*Corresponding Author : Hyun-Woo Cho(Daegu Univ.)
email: hwcho@daegu.ac.kr
Received September 15, 2021 Revised October 18, 2021
Accepted January 7, 2022 Published January 31, 2022

1. 서론

4차 산업혁명의 핵심인 인공지능, 사물 인터넷, 클라우드 컴퓨팅, 빅 데이터 분석 등의 디지털 기술은 기존의 제조와 서비스를 뛰어넘는 새로운 연구 분야와 신산업을 만들어 냄으로써 우리의 삶에 혁명적인 변화를 주고 있다[1]. 이미 상용화된 드론이나 AI 스피커를 비롯하여 자율 주행차, 핀 테크, 스마트 팩토리, 헬스케어, 사물 인터넷 등의 분야에서 혁신적인 서비스와 제품이 출시되어 시장 점유율을 높이고 있다. 이러한 배경에는 개인별 맞춤과 추천이 가능하도록 지능화된 알고리즘과 방대한 실시간 데이터를 분석하는 기법들의 연구와 개발이 있다고 할 수 있다[2]. 이러한 신산업에서의 연구 개발은 과거 타 분야 대비 기존 학문 간 경계가 무의미해지고 다양한 분야의 기술들이 융합되어 적용되는 융복합 연구를 필요로 하는데 이러한 추세는 국내에서 이미 문과와 이과의 구분이 없는 통합 고교과정이나 대학의 융복합 학과 신설 등에서도 확인되고 있다. 또한 개인용 통신기기인 스마트 폰과 SNS의 급격한 보급으로 과거의 정형화된 데이터가 아닌 문서, 사진, 동영상 등의 비정형 데이터가 급격히 증가하고 있는 것도 특징이라 할 수 있다[2].

텍스트 마이닝은 자연어 처리(natural language processing) 기법을 사용하여 대용량의 텍스트 데이터를 기계 학습(machine learning)에 적합한 표준화되고 구조화된 데이터로 정형화하여 의미 있는 특성과 정보를 추출하는 기술을 의미한다[3]. 이러한 텍스트 마이닝을 통해 문서 데이터의 분류(classification), 군집화 (clustering), 시각화 (visualization), 특성 추출(feature extraction)을 수행하게 된다. 텍스트 마이닝은 개인화 마케팅, 고객 불만사항 분석, 번역, 소비자 감성 분석 등 다양한 분야에서 빅 데이터 분석의 주요한 툴로서 활용되고 있다[2-6]. 한편, 텍스트 마이닝 기법의 하나인 토픽 모델링(topic modeling)은 텍스트 데이터로 이루어진 문서들에서 사용된 주제어들의 사용 패턴을 분석하여 문서들을 대표할 수 있는 숨겨진 주제 또는 토픽을 자동으로 추출하는 기법이다. 비지도 학습 방법인 토픽 모델링을 구현하기 위하여 LDA, LSA(latent semantic analysis), CTM(correlated topic model) 등의 알고리즘이 제안되었으며 편의성과 범용성 측면에서 장점이 있는 LDA가 주로 사용되고 있다[7]. LDA는 문서에서 발견된 단어 빈도 분포를 분석하여 문서들의 잠재된 토픽들을 찾아내는 확률적 모형으로서 제조, 정보통신, 에너지, 특허, 음악, 환경, 마케팅 분야 등 다양한 연구 분야에서 활용되고 있다[4-10]. 그러나 융복합 분야 논문의 연구 주제 분석에 있어 토픽 모델링을 활용한 연구는 거의 진행된 바가 없다.

본 연구에서는 융복합 분야의 국제 학술지 중 하나인 ESWA에 게재된 2000년부터 2021년까지 논문들을 대상으로 LDA에 기반한 토픽 모델링 분석을 수행하여 2000년 이후 주요 연구 주제를 파악하고자 한다. ESWA 저널은 5.45의 5년 평균 영향력 지수(impact factor)를 가진 융복합 분야의 선도 저널로서 데이터 마이닝과 인공지능 등의 지능형 시스템과 기술을 대상으로 공학, 경영, 서비스, 의학, 금융, 물류 등 다양한 융합 관점의 논문들이 게재되고 있다. 해당 기간 ESWA에 게재된 논문들의 제목, 키워드, 초록 데이터들을 자동 수집하고 전처리 과정과 토픽 모델링 분석을 통해 연구 주제들의 상호 관련성과 비중을 분석하고 시기별 연구 주제 변화 추이를 살펴보고자 한다. 핵심어와 연구 주제들 그리고 시간에 따른 변화 추이의 분석 결과는 제조 분야를 포함한 다양한 산업 분야에서 진행될 향후 융복합 연구의 방향과 주제 설정에 기초 분석 자료로 활용될 수 있을 것이다. 본 논문의 구성은 우선 LDA 방법론에 대한 간략한 소개를 시작으로 대상 논문 데이터에 대한 자료 수집 절차 및 전처리 과정을 제시하고 ESWA 게재 논문을 대상으로 수행된 토픽 모델링 분석 결과와 결론으로 이어진다.

2. 방법론

LDA는 일련의 문서들에 어떤 토픽이 존재하는지를 알려주는 토픽 모델링 알고리즘이다. 일반적으로 문서는 여러 개의 토픽을 가질 수 있으며 토픽들은 디리클레 분포(Dirichlet distribution)을 따른다고 가정한다[12]. LDA는 문서의 토픽 가중치와 토픽의 단어 가중치를 고려하여 d라는 문서 내 i번째 단어의 토픽이 j에 할당될 확률을 구하게 되는데

$$p(z_{d,i}=j|z_{-i},w) = \left\{\frac{n_{d,k}+\alpha_k}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)}\right\}\left\{\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}v_{k,j}+\beta_j)}\right\} \quad (1)$$

여기서 α와 β는 잠재 파라미터이며 $n_{d,k}$와 $v_{k,j}$는 단어의 빈도를 나타낸다. 전체 토픽 수는 K, 전체 문서 수는 D, d라는 문서의 단어 수는 N이다. LDA 알고리즘을 그림으로 표현하면 Fig. 1로 나타낼 수 있다[12]. α는 d번째 문서의 토픽의 비율인 θ_d를 결정하는 디라클레 분포

의 파라미터이며, γ는 k번째 토픽 단어들의 분포를 나타내는 β_k값을 결정하는 파라미터이다.

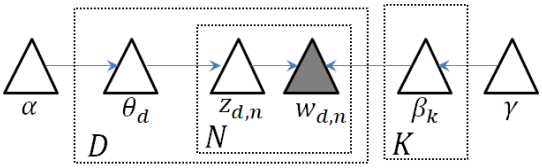

Fig. 1. A diagram of LDA model

Fig. 1에서 보는 바와 같이 θ_d라는 토픽 비율의 변화에 따라 d번째 문서 내 단어들의 토픽 $z_{d,n}$이 결정되고 최종적으로 토픽 내 단어 분포 β_k와 $z_{d,n}$로부터 $w_{d,n}$을 구하게 된다. 이러한 LDA 분석에서는 군집 분석에서 군집의 수를 지정하듯이 토픽의 수를 사전에 지정해주어야 한다. 토픽 수 선정을 위한 지표로써 일반적으로 혼란도(perplexity)와 일관성(coherence) 지표를 사용하고 있다[12].

2000년부터 2021년까지 ESWA에서 게재된 11,918편 논문들의 발간년월, 논문 제목, 키워드, 초록을 수집하였다. 파이썬으로 작성한 스크래핑 프로그램을 이용하여 구글의 학술 논문 검색 사이트에서 필요 정보를 자동 수집하였다. 수집된 논문들의 연도별 대상 논문 수를 Fig. 2에 나타내었다. 그림에서 보이듯이 2008년~2009년을 기점으로 논문 수가 크게 증가하였는데 이는 저널 주제의 범위 확대, 연간 저널 발행 횟수의 증가, 특별호 편성 등의 영향으로 판단된다.

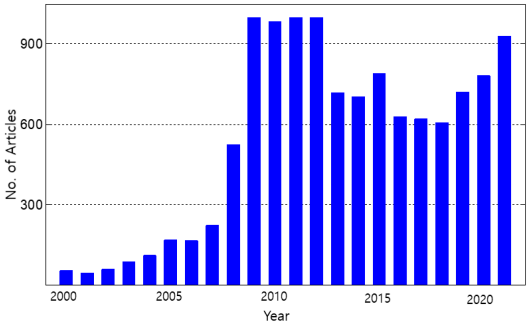

Fig. 2. A plot for number of articles analyzed

수집된 비정형 텍스트 데이터를 바탕으로 LDA 토픽 모델링을 수행하기 전에 토픽 모델링의 정확성을 향상시키기 위한 전처리 작업은 필수적이다. 전처리의 첫 단계는 토큰화 (tokenization)로서 형태소에 따라 단어 분리가 다르므로 분석에 필요한 단위로 단어를 분리하게 된다. 다음으로는 분석에 불필요한 단어를 삭제하는 불용어 제거이다. 대명사, 관사, 전치사 등을 우선 제거하였으며 논문의 특성상 자주 나타나지만 토픽 모델링과는 관련성이 떨어지는 'study'와 'result' 등의 명사와 'discuss', 'calculate', 'determine' 등의 동사를 불용어 리스트에 포함해 제거하였다. 마지막으로, 같은 의미이나 형태만 다른 단어들을 하나의 표제어로 정규화시키는 표제어 추출(lemmatization)을 수행하였다.

앞에서 언급하였듯이 LDA 분석에 앞서 사전에 토픽의 수를 결정하여야 하는데 본 연구에서는 일관성 지수를 사용하여 토픽의 수를 결정하였다. 여기서 사용하지 않은 혼란도 지수는 특정 확률 모델이 실제 문서의 결과를 적절히 학습하였다는 사실만을 보여줄 뿐 실제 해석에서는 한계가 있다고 알려져 있다[7]. 이의 단점을 해결한 일관성 지수는 실제 사람이 해석하기 적합한 척도로서 이 지수가 높을수록 LDA 분석의 각 토픽이 일관되며 유사한 단어들로 구성되어 있음을 나타낸다. 본 연구에서는 최적의 토픽 수 선정을 위해 토픽 수에 따른 일관성 지수를 도출하였다. 토픽의 수에 따른 일관성 지수를 나타낸 Fig. 3에서 보이듯이 본 연구에서는 최대의 일관성 지수 0.409를 달성한 10개로 토픽의 최적 개수를 선정하였다.

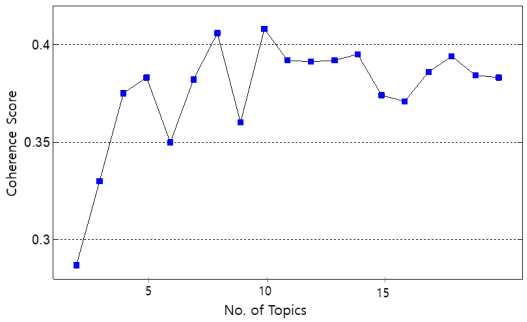

Fig. 3. Coherence score plot

3. 토픽 모델링 결과

2001년부터 2020년까지 ESWA에 게재된 11,918편 논문들의 발간년월, 논문 제목, 키워드, 초록에 대한 전처리 과정이 완료된 이후 핵심어에 대한 빈도 분석을 수행하였다. 전체 대상 텍스트의 핵심어에 대하여 빈도가 높은 순으로 상위 20위까지의 핵심어 단어를 Table 1에

나타내었다. 빈도순으로 살펴보면 'approach'가 8,006회로 가장 많이 사용되었으며 이후 'performance'는 6,412회, 'feature'는 6,242회, 'network'는 5,792회의 빈도순으로 사용되었다.

본 연구에서 수집된 비정형 텍스트 데이터로부터 중심이 되는 토픽과 핵심 단어를 도출하기 위하여 LDA 토픽 모델링을 수행한 결과를 Table 2에 나타내었다. 앞에서 일관도 지수를 통해 결정된 10개의 토픽 전체에 대하여 각각의 토픽을 표현하는 고빈도 핵심 단어 20개씩을 추출하였으며 이러한 토픽별 핵심 단어들의 연관성을 바탕으로 토픽의 이름을 정하였다.

토픽1은 'optimization', 'solution', 'search', 'performance' 등으로 이루어져 있으며 최적화 방법론과 관련된 주제이다. 제조 공정은 물론 물류나 컴퓨터 네트워크, 서비스 등 다양한 프로세스에서의 최적화 문제와 이를 해결하기 위한 새로운 방법론 그리고 최적화 성능 향상 등에 관한 연구들로 구성된다.

Table 1. Top 20 keywords

No.	Keyword	Freq.	No.	Keyword	Freq.
1	approach	8,006	11	analysis	4,128
2	performance	6,412	12	decision	3,737
3	feature	6,242	13	application	3,385
4	network	5,792	14	knowledge	3,335
5	information	4,850	15	accuracy	3,206
6	time	4,765	16	image	3,180
7	classification	4,538	17	solution	3,169
8	process	4,417	18	number	3,053
9	fuzzy	4,388	19	user	2,992
10	technique	4,312	20	optimization	2,889

토픽2는 'user', 'knowledge', 'information', 'recommendation' 등으로 이루어져 있으며 지식 모델과 관련된 주제이다. 사용자나 전문가의 지식, 경험, 정보들을 모델링하여 해석이 어려운 오작동인 공정이나 기계와 특정 작업 불량의 진단 등에 관한 연구들로 구성된다.

토픽3은 'image', 'classification', 'datasets', 'performance' 등으로 이루어져 있으며 분류 방법론의 활용과 관련된 주제이다. 분류 문제로서 주로 다루어졌던 공정 데이터의 분류를 포함하면서 최근 폭발적으로 증가한 지문, 얼굴, 그림 등 이미지 데이터에 대한 다양한 분류 문제 적용에 관한 연구들로 구성된다.

토픽4는 'feature', 'classification', 'classifier', 'selection' 등으로 이루어져 있으며 분류에서의 특성 선택과 관련된 주제이다. 분류 데이터 내 변수의 증가 및 변수들의 상관관계를 극복하기 위하여 분류에 유리한 선택된 특정 변수들로 분류기(classifier)를 구성함으로써 분류 정확도를 높이면서 처리 속도를 향상시킬 수 있는 연구들로 구성된다.

토픽5는 'network', 'prediction', 'performance', 'market' 등으로 이루어져 있으며 예측 모델과 관련된 주제이다. 신경망이나 기계 학습 알고리즘에 기반하여 조업데이터를 통한 제품의 품질 예측, 시장의 수요량이나 점유율 예측, 주가 지수 예측 등에 대한 예측 모델 연구들로 구성된다.

Table 2. Results of topic modeling

No.	Topic	Ratio	Top 10 words in each topic
1	optimization method	0.164	optimization, time, solution, performance, search, function, network, parameter, approach, control
2	knowledge model	0.157	user, knowledge, information, approach, recommendation, social, application, item, framework, task
3	classification application	0.108	image, classification, feature, datasets, performance, deep, approach, segmentation, technique, accuracy
4	feature selection	0.104	feature, classification, classifier, signal, performance, selection, accuracy, machine, approach, recognition
5	prediction model	0.100	network, prediction, approach, performance, market, time, stock, cluster, number, analysis
6	multi-object decision-making	0.097	solution, approach, cost, design, time, decision, multi-objective, agent, resource, research
7	business management	0.089	process, performance, service, decision, approach, portfolio, selection, criterion, evaluation, management
8	text mining	0.073	information, approach, decision, text, process, document, word, analysis, query, performance
9	fuzzy logic	0.069	fuzzy, approach, decision, value, function, measure, group, performance, information, number
10	anomaly detection	0.039	detection, risk, tree, human, credit, traffic, approach, attack, emotion, learning

토픽6은 'solution', 'cost', 'multi-objective', 'decision' 등으로 이루어져 있으며 다중 목적 의사 결정과 관련된 주제이다. 복수의 목적 함수에 대한 최적 의사

결정 방법론과 이의 활용에 관한 연구로서 품질 비용 최적화, 공급망 관리에서의 공급업체 선정, 재고관리 계획 수립 등으로 구성된다.

토픽7은 'process', 'service', 'decision', 'portfolio' 등으로 이루어져 있으며 지능형 경영 관리와 관련된 주제이다. 경영 전략 수립 및 경영 지원 시스템 관점에서 정보기술, 빅 데이터 분석, 지능형 방법론 등을 활용하여 목표 시장 설정과 수요 예측, 사업 및 투자 포트폴리오 분석, 특허 관리와 기술 경영, 리스크 관리 등에 관한 연구들로 구성된다.

토픽8은 'information', 'decision', 'text', 'document', 'word' 등으로 이루어져 있으며 텍스트 마이닝과 관련된 주제이다. 기존의 정형화된 데이터 마이닝에 비해 급격하게 늘어나고 있는 비정형 텍스트 데이터를 분석하는 방법론으로서 웹 문서, 논문, SNS 데이터, 사용자 리뷰 등의 웹 콘텐츠를 적극 활용하여 소비자 감성 분석, 부동산 가격 분석, 주가 모델링, 특정 산업 및 기술 트렌드 분석, 연구 및 특허 토픽 모델링 등에 관한 연구들로 구성된다.

토픽 9는 'fuzzy', 'decision', 'value', 'function' 등으로 이루어져 있으며 퍼지 로직과 관련된 주제이다. 공정 제어 용도의 지능제어 및 fuzzy 제어기 설계 및 성능 검증, 신경망 등 다른 지능형 방법론과 결합하여 선박, 모터 속도, 전력량 등의 제어 시스템에 관한 연구들로 구성된다.

토픽10은 'detection', 'risk', 'credit', 'traffic' 등으로 이루어져 있으며 이상 감지와 관련된 주제이다. 데이터에 기반한 제조 공정 모니터링을 확장하여 다양한 환경에서의 특이점을 감지하는 연구로서 우주 및 항공 시스템 이상 감지, 해킹 등 네트워크 침입 및 유해 트래픽 감지, FDS (fraud detection system)과 같은 신용카드 이상 거래 감지, 지능형 cctv를 통한 범죄 상황 감지 등의 연구들로 구성된다.

토픽 분석을 통해 추출된 10개의 토픽에 대한 토픽 디스턴스 맵 (inter-topic distance map) 분석 결과를 Fig. 4에 나타내었다. Fig. 4의 좌측 그림인 토픽 디스턴스 맵은 토픽의 비중과 토픽들 사이의 거리를 보여주는 그림으로써 각각의 토픽들이 다른 토픽들과 가지는 연관성과 유사도를 시각적으로 보여준다. Fig. 4에서 원의 중앙에 위치한 숫자들은 도출된 10개 토픽을 표시하고 있다. 각각의 토픽들을 나타내고 있는 이 원들의 거리가 서로 가까우면 토픽 간 유사성이 높고 멀어질수록 유사성은 낮다.

Fig. 4의 우측 그림은 선택된 특정 토픽 (여기서는 토픽 1)에 대한 주요 단어들을 보여주고 있다. 이러한 토픽 디스턴스 맵을 통해 본 연구에서 추출된 10개 토픽은 일부 중첩 영역이 존재하지만 그리 크지 않은 것으로 보아 대부분의 추출된 토픽들이 상호 유사성이 낮아 명확하게 구분되어 도출된 것으로 판단된다. IDM 결과 그림에서 미미하지만 가장 큰 중첩 영역을 가진 토픽 2와 토픽 8은 각각 지식 모델과 텍스트 마이닝에 관한 주제를 나타내는데, 공통적으로 비정형화된 지식이나 문서들을 모델링하고 분석한다는 측면에서 보면 이 두 토픽의 공통점을 이해할 수 있다.

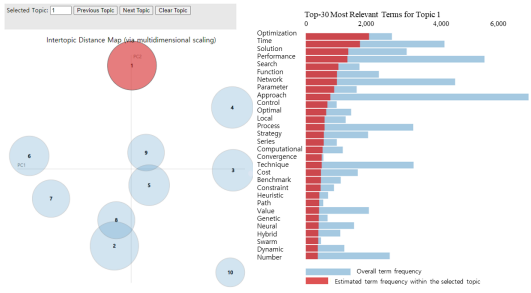

Fig. 4. IDM results

2000년부터 2021년까지 ESWA에 게재된 논문들로부터 얻은 10개의 토픽 각각이 시간에 따라 어떻게 변화하는지를 분석하기 위하여 Fig. 5에 토픽 트렌드의 변화 추이를 나타내었다. 이 그림은 개별 논문의 토픽을 연도별로 종합하여 10개 토픽에 대한 상대적 비율을 보여주고 있다. Fig. 5의 연도별 트렌드를 살펴보면 최적화 방법론 (토픽 1)과 분류 방법론의 활용 (토픽2) 연구의 경우 2000년에서 2010년까지의 기간에는 낮은 점유율을 보이다가 2008년 (토픽1)과 2015년 (토픽3)이후에는 상대

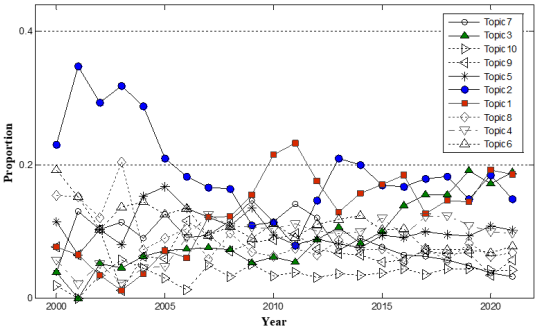

Fig. 5. Topic trend plot

적으로 크게 늘어난 상태를 최근까지도 유지하고 있음을 알 수 있다. 지식 모델 (토픽 2) 연구는 2009년까지 최대의 점유율을 보이면서 이후 다소 감소하였으나 최근까지도 꾸준히 발표되고 있다. 나머지 토픽들에 관한 연구는 2000년~2003년의 일부 토픽을 제외하면 큰 패턴의 변화 없이 지속되고 있다.

전체 조사 대상 기간 중 ESWA 저널 논문의 수가 급격히 증가하였으며 (Fig. 2) 연구 토픽의 변화가 상대적으로 집중되었던 시기인 2009년~2012년에 게재된 논문 추이를 토픽 1~5에 대해 살펴보았다. Fig. 6에 알 수 있듯이 최적화 방법론 (토픽1) 논문의 수가 155->212->232->176으로 2009년에서 2011년까지 크게 증가하다 2012년 다소 감소하는 추세에 있으며, 이와 반대로 지식 모델 (토픽2) 연구의 경우 110->112->79->146으로 감소하다 일시 증가한 것을 알 수 있다. 분류 방법론의 활용 (토픽3) 연구는 53->61->54로 유지되다가 2012년 88로 증가했으며 토픽 4와 5의 경우에는 일정한 수준을 유지하고 있다.

한편, 이 시기의 논문들에 대해 피인용 횟수를 토픽별로 정리하여 논문 1편당 피인용 횟수를 도출해 보았다. 토픽 1의 경우 2009년 102.4를 시작으로 76.5->83.0->65.9로 변화하였는데 이러한 경향 즉, 최신 논문일수록 피인용 횟수가 감소하는 추세는 일반적이며 본 연구의 다른 토픽에서도 동일하게 관찰되었다. 주목할 사항은 분류의 특성 선택 (토픽 4)의 경우 타 토픽과 비교하였을 때 매년 가장 큰 피인용 수치를 보인다는 점이다.

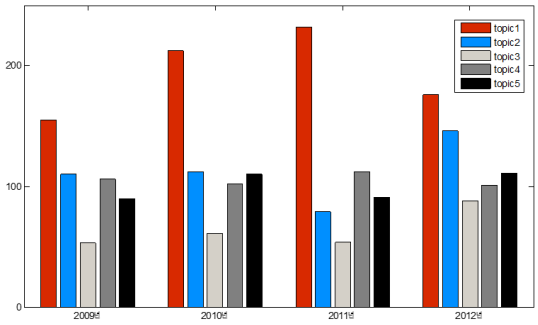

Fig. 6. A plot for top 5 topics from 2009 to 2012

구체적으로 살펴보면 2009년에 113.9를 시작으로 88.0->97.0->87.9의 값을 가져 매년 최대 피인용 값을 가지고 있다. 또한 논문 1편당 피인용 수치의 4년 평균값 (96.7)에서도 타 토픽 (토픽1: 81.9, 토픽2: 77.3, 토픽3: 77.0, 토픽5: 74.3) 대비 큰 격차를 보이고 있다. 이러한 특성은 빅 데이터 분석에서 볼 수 있듯이 분류 문제에서 발생하는 변수의 급격한 증가와 연관된 것으로 판단된다. 분류에 적합한 변수를 선택하고 나머지 변수는 분류에서 제외시키는 알고리즘을 통하여 분류의 성능과 속도를 향상시키는데 이러한 연구 주제들이 2010년 이후 활발하게 연구된 결과 이 시기 토픽 4의 피인용 값이 커지게 된 것으로 판단된다.

4. 결론

본 연구에서는 텍스트 마이닝의 LDA 알고리즘에 기반한 토픽 모델링을 활용하여 제조, 서비스, 금융, 물류 등 다양한 분야의 융복합 연구에 관한 논문들을 분석하였다. LDA에서 최대 일관성 수치를 보인 10개의 토픽을 도출한 결과 최적화, 지식 모델, 분류, 분류 특성 선택, 예측 모델 등의 순서를 보였다. 토픽들의 시기별 변화 추이를 분석한 결과 최적화 방법론, 지식 모델, 분류 활용 토픽은 특정 시기를 분기점으로 급격한 변화를 보인 반면 나머지 토픽들은 일관된 패턴을 보여주었다. 이러한 시기별 토픽 트렌드 추이는 빅 데이터, 스마트 제조, 사물 인터넷, 핀 테크 등 4차 혁명과 신산업의 성장과 관련이 있으며 신규 시스템의 효율적인 운용을 위해 필요한 연구 주제들과 밀접하게 관련되어 있다고 해석할 수 있다. IDM 분석에서는 본 연구에서 도출된 10개 토픽의 중첩된 영역이 크지 않아 조사 대상 기간 명확하게 구분된 토픽들이 유사한 비중으로 연구되고 있음을 알 수 있었다. 또한 2009년~2012년에 게재된 논문의 토픽을 정량적으로 분석한 결과 토픽 4인 분류의 특성 선택이 이 기간 동안 매년 가장 큰 피인용 수치를 보였으며 논문 1편당 피인용 평균값에서도 타 토픽 대비 큰 격차를 보여주었다.

마지막으로 본 연구에서는 ESWA 저널의 논문만을 대상으로 토픽 분석을 수행하였기 때문에 국내 연구자들의 논문이 분석 대상에 일부 있었으나 타 국외 학술지와 국내 학술지가 포함되지 못한 한계를 가지고 있다. 향후 분석의 범위를 확장하여 더 다양한 학술지 논문들을 고려한 추가 연구가 필요할 것이다. 그리고 반도체 산업이나 회분식 공정(batch process) 등 특정 산업이나 공정에서 얻어진 텍스트 데이터의 토픽 모델링 분석에 관한 향후 연구를 제안한다. 이들 특화된 공정의 토픽 분석을 통해 다양한 지능형 융복합 방법론에 대한 분석이 제시된다면 고부가가치의 대규모 장치산업인 반도체/디스플레

이나 다품종 소량 생산 방식의 효율적인 공정 운용에 도움이 될 것으로 판단된다.

References

[1] Y. Duan, J. S. Edwards, Y. K. Dwivedi, "Artificial intelligence for decision making in the era of big data - evolution, challenges and research agenda", International Journal of Information Management, Vol. 48, pp. 63-71, 2019.
DOI:https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[2] M. Vanhala, C. Lu, J. Peltonen, S. Sundqvist, J. Nummenmaa, K. Järvelin, "The usage of large data sets in online consumer behaviour: a bibliometric and computational text-mining-driven analysis of previous research", Journal of Business Research, Vol. 106, pp. 46-59, 2020.
DOI:https://doi.org/10.1016/j.jbusres.2019.09.009

[3] A. Gupta, V. Dengre, H. A. Kheruwala, M. Shah, "Comprehensive review of text-mining applications in finance. Financial Innovation, Vol. 6, 39 (2020).
DOI:https://doi.org/10.1186/s40854-020-00205-1

[4] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", Multimedia Tools and Applications, Vol. 78, pp. 15169-15211, 2019.
DOI:https://doi.org/10.1007/s11042-018-6894-4

[5] A. Amado, P. Cortez, P. Rita, S. Moro, "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis", European Research on Management and Business Economics, Vol.24, No.1, pp.1-7, 2018.
DOI:https://doi.org/10.1016/j.iedeen.2017.06.002

[6] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, H. Mora, "A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making", Industrial Marketing Management, Vol.90, pp. 523-537, 2020.
DOI:https://doi.org/10.1016/j.indmarman.2019.08.003

[7] R. Alghamdi, K. Alfalqi, "A Survey of Topic Modeling in Text Mining", International Journal of Advanced Computer Science and Applications, Vol. 6, pp. 147-153, 2015.
DOI:https://dx.doi.org/10.14569/IJACSA.2015.060121

[8] S. Yoon, M. Kim, "topic modeling on fine dust issues using LDA analysis", Journal of Energy Engineering, Vol. 29, pp. 23-29, 2020.
DOI:https://doi.org/10.5855/ENERGY.2020.29.2.023

[9] J. Jeong, S. H. Kim, "Failure diagnosis using text mining and deep learning: development of prediction algorithm for responsible department", The Transactions of the Korean Institute of Electrical Engineers, vol. 69, pp. 1225-1236, 2020.
DOI:https://doi.org/10.5370/KIEE.2020.69.8.1225

[10] S. Moon, S. Chung, S. Chi, "Topic modeling of news article about international construction market using latent Dirichlet allocation", Journal of the Korean Society of Civil Engineers Vol.38, No.4, pp.595-599, 2018.
DOI:https://doi.org/10.12652/Ksce.2018.38.4.0595

[11] T. Yun and H. Ahn, "Fake news detection for Korean news using text mining and machine learning techniques", Journal of Information Technology Applications and Management, Vol.25, pp.19-32, 2018.
DOI:https://doi.org/10.21219/jitam.2018.25.1.019

[12] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

조 현 우(Hyun-Woo Cho) [정회원]

- 2003년 8월 : 포항공과대학교 기계산업공학부 (공학박사)
- 2003년 8월 ~ 2007년 8월 : 포항공과대학교, 조지아텍, 테네시주립대 연구원
- 2007년 9월 ~ 2011년 2월 : 삼성전자, 삼성디스플레이 책임연구원
- 2011년 3월 ~ 현재 : 대구대학교 융합산업공학과 교수

〈관심분야〉
공정모니터링, 빅데이터, 인공지능