https://doi.org/10.5392/JKCA.2021.21.12.095

토픽 모델링을 활용한 한국콘텐츠학회 논문지 연구 동향 탐색

An Exploratory Research Trends Analysis in Journal of the Korea Contents Association using Topic Modeling

석혜은[*], 김수영[**], 이연수[*], 조현영[*], 이수경[***], 김경화[****]
독립 연구자[*], 제일기획[**], Kurvv.AI[***], 청주대학교 국어교육과[****]

Hye-Eun Seok(graceseok@gmail.com)[*], Soo-Young Kim(tabit75@naver.com)[**], Yeon-Su Lee(yeonsu.leeyoon@gmail.com)[*], Hyun-Young Cho(teresackim@gmail.com)[*], Soo-Kyoung Lee(tototong@naver.com)[***], Kyoung-Hwa Kim(joyedu95@naver.com)[****]

요약

본 연구의 목적은 한국콘텐츠학회 논문지에 게재된 9,858건의 논문을 대상으로 토픽 모델링을 활용하여 지난 20년간 연구동향을 탐색함으로써 콘텐츠 연구개발에서의 주요 토픽을 도출하고 학술적 발전방향을 제공하는데 있다. 추출된 토픽의 신뢰성과 타당성을 확보하기 위해 양적 평가기법 뿐만 아니라 정성적 기법을 단계적으로 적용하여 연구자들이 합의한 수준의 말뭉치가 생성될 때까지 이를 반복적으로 수행하였으며 이에 따른 구체적인 분석 절차를 제시하였다. 분석 결과 8개의 핵심 토픽이 추출되었다. 이는 한국콘텐츠학회가 특정 학문 분야를 한정하지 않고 다양한 분야의 융·복합 연구 논문을 발간하고 있음을 보여준다. 또한 2012년 이전 상반기에는 공학기술 분야 토픽 비중이 상대적으로 높게 나타난 반면, 2012년 이후 하반기에는 사회과학 분야 토픽 출현 비중이 상대적으로 높게 나타났다. 구체적으로 '사회복지' 토픽은 상반기 대비 하반기에 약 4배수 증가세가 나타났다. 토픽별 추세분석을 통해 추세선의 변곡점이 나타난 특정 시점에 주목하여 해당 토픽의 연구동향에 영향을 미친 외적 변인을 탐색하였고 토픽과 외적 변인 간 관련성을 파악하였다. 본 연구결과가 국내 콘텐츠 관련 연구 개발 및 산업 분야에서 진행되고 있는 활발한 논의를 진행하는데 시사점을 제공할 수 있기를 기대한다.

■ 중심어 : | 콘텐츠 연구 | 한국콘텐츠학회 논문지 | 연구동향 | 토픽 모델링 | LDA |

Abstract

The purpose of this study is to derive major topics in content R&D and provide directions for academic development by exploring research trends over the past 20 years using topic modeling targeting 9,858 papers published in the Journal of the Korean Contents Association. To secure the reliability and validity of the extracted topics, not only the quantitative evaluation technique but also the qualitative technique were applied step-by-step and repeated until a corpus of the level agreed upon by the researchers was generated, and detailed analysis procedures were presented accordingly. As a result of the analysis, 8 core topics were extracted. This shows that the Korean Contents Association is publishing convergence and complex research papers in various fields without limiting to a specific academic field. Also, before 2012, the proportion of topics in the field of engineering and technology appeared relatively high, while after 2012, the proportion of topics in the field of social sciences appeared relatively high. Specifically, the topic of 'social welfare' showed a fourfold increase in the second half compared to the first half. Through topic-specific trend analysis, we focused on the turning point in time at which the inflection point of the trend line appeared, explored the external variables that affected the research trend of the topic, and identified the relationship between the topic and the external variable. It is hoped that the results of this study can provide implications for active discussions in domestic content-related R&D and industrial fields.

■ keyword : | Contents Research | the Journal of the Korea Contents Association | Research Trend | Topic Modeling | LDA |

접수일자 : 2021년 11월 12일
수정일자 : 2021년 12월 06일
심사완료일 : 2021년 12월 06일
교신저자 : 석혜은, e-mail : graceseok@gmail.com

I. 서 론

오늘날은 각 분야의 지식이나 정보가 다양한 콘텐츠(contents)로 생산되며 소비되고 있다. 이러한 지식이나 정보는 산업과 연계되어 경제적인 영향력이 커지고 있어 중요성이 더욱 높아지고 있다. 콘텐츠는 정보・통신의 측면에서 지식과 정보를 의미하는 용어로, 표준국어대사전에서는 '콘텐츠'란 "정보・통신 분야에서 인터넷이나 컴퓨터 통신 등을 통해 제공되는 정보나 내용들"이라고 정의하고 있다. 그리고 콘텐츠 산업 진흥법 제1장 제2조 1항(법률 제 10369호, 2010. 6. 10. 전부개정)에 따르면 '콘텐츠'란 부호·문자·도형·색채·음성·음향·이미지 및 영상 등의 자료 또는 정보를 의미한다. 이는 콘텐츠가 특정 분야에서 생산된 지식이나 결과에만 국한되지 않고 지속적으로 해당 범위가 확장되고 있으며 생산 및 소통 방식 등도 다양해지고 있음을 의미한다.

이러한 콘텐츠 및 콘텐츠 산업의 중요성이 높아지고 있으며, 특히 포스트 코로나 시대에 앞으로의 콘텐츠의 영역에 대한 미래지향적인 논의와 준비가 필요한 시점이다. 그런데 국내에서 '콘텐츠'로 어떤 내용들이 연구되었으며 어떻게 변화하였는지에 대해 체계적으로 분석한 연구는 찾아보기 어렵다. 이러한 문제의식에서 출발한 본 연구는 한국콘텐츠학회의 논문지에 게재된 논문을 대상으로 토픽 모델링을 이용하여 심층적으로 분석해봄으로써 국내 콘텐츠와 관련된 하위영역의 주요 연구 토픽을 추출하고 이러한 토픽들이 지난 20년간 어떻게 변화되었는지 탐색한다. 한국콘텐츠학회지가 과학기술분야 콘텐츠뿐만 아니라 융·복합 학술분야 연구를 출간하는 학술지로 성장해온 것을 고려해볼 때, 한국콘텐츠학회지에 게재된 연구들의 연구동향을 살펴봄으로써 콘텐츠와 관련된 학문적 논의와 중요한 쟁점 및 문제의식의 변화를 거시적으로 살펴보는 것은 의미가 있다.

이를 위하여 본 연구에서는 토픽 모델링을 활용하여 콘텐츠 연구의 연구 동향을 살펴보고자 한다. 토픽 모델링은 텍스트에 숨겨진 의미 구조를 발견하기 위해 대규모 텍스트 자료로 구성된 빅데이터에서 토픽(즉, 일관된 단어 군집)을 추출하는 텍스트 마이닝 및 개념 추출 방법이다[1]. 이 방법은 대표적인 비지도 머신러닝 기법(Unsupervised Machine Learning)으로 양적 연구방법에 해당한다. 그런데 양적 연구방법은 비교적 적은 비용과 시간으로 대규모 자료를 효율적으로 분석할 수 있다는 장점이 있는 반면 '콘텐츠'라는 구성개념(construct)에 내재된 깊은 수준의 단면들을 탐색하기 어렵다는 한계가 있다. 이러한 방법만으로는 중요하지만 아직 발견되지 않은 새로운 현상이 여전히 간과되기 쉽다. 이에 양적 방법과 질적 방법을 상호 보완한다면 각 방법의 단점을 극복하여 대규모 자료로부터 심층적인 통찰과 타당한 결과를 얻을 수 있을 것이다.

이에 본 연구에서는 추출된 토픽의 신뢰성과 타당성을 확보하기 위해 양적 평가기법뿐만 아니라 정성적 기법을 단계적으로 적용하여 연구자들이 합의한 수준의 말뭉치가 생성될 때까지 데이터 전처리 과정을 반복적으로 수행하였으며 이에 따른 구체적인 분석 절차를 제시하였다. 최종 단계에서는 생성된 말뭉치를 이용하여 주요 토픽을 추출한 후 토픽별로 지난 20년간 연구 동향을 분석하였다. 이처럼 토픽과 시대적 변인 간 관련성을 확인하는 작업을 통해 추출된 토픽의 신뢰성과 타당성을 검증함으로써 텍스트 마이닝을 통해 찾아낸 숨겨진 패턴과 새로운 정보는 콘텐츠와 관련된 학문적 영역의 미래지향적 논의를 위한 기초자료로서 의미가 있다.

II. 이론적 배경

1. 콘텐츠 관련 연구현황

콘텐츠의 생산과 유통 그리고 소비의 속도가 빨라지고 있어 콘텐츠의 동향과 앞으로의 콘텐츠 개발과 소비의 방향성에 대한 연구가 중요해지고 있다. 콘텐츠 산업의 연구개발(Research and Development; 이하 R&D)은 2010년 이전까지는 콘텐츠의 신규 서비스 개발과 콘텐츠 서비스의 효율적 운영 등 비기술적 지식창출에 해당하는 서비스 관점과 문화산업기술(Culture Technology; CT)을 포함한 기술접목과 관련된 R&D를 모두 포괄하는 개념이었다. 그러나 콘텐츠 R&D 개념은 인문사회과학 및 예술 등 창의성의 근간이 되는

기초 R&D의 개념 도입과 범위 확장을 통해 보완의 필요성이 대두되었다[2].

이와 같이 콘텐츠에 대한 관심은 학계에서도 꾸준히 지속되어져 왔다. 일례로, 2021년 10월 기준으로 구글 학술검색에서 '콘텐츠'라는 검색어로 조회하면 약 2,610,000건의 학술자료가 검색된다. 기간을 설정하여 검색하면 2000년부터 약 203,00 건, 2010년부터 약 139,000건, 2020년부터 약 19,400건으로 매년 약 10,000여건의 '콘텐츠' 관련 학술자료가 발표되고 있음을 확인할 수 있다. 특히 2000년 이후 최근 학술자료의 경우 제목에 '콘텐츠'라는 중심어가 상당 비율로 나타났고, 학술자료 발간 분야는 과학기술, 인문·사회, 예술 분야 등 폭넓은 영역을 포괄하고 있는 것으로 나타났다. 이는 콘텐츠와 관련된 학술연구가 거의 모든 영역에서 진행되고 있다고 해도 과언이 아님을 뒷받침하는 자료라 할 수 있다. 이러한 현실적 상황을 고려해볼 때 콘텐츠 관련 연구 자료를 통합하고, 체계적인 발전 방향을 모색하는 것은 상당한 어려움이 따르는 일이다.

현재 우리 사회는 코로나19를 겪으면서 급속한 변화를 경험하고 있다. 코로나19 팬데믹에서 비대면 환경의 중요성이 높아지면서 인터넷 환경, 디지털 환경에서의 소통이 급속히 증가하였다[3]. 이러한 사회의 변화와 환경의 변화는 콘텐츠에도 중요한 영향을 미친다. 경제적 가치와 밀접한 콘텐츠는 학문적인 지식이나 정보에만 머물지 않고 중요한 산업과 밀접한 관련을 맺으며 그 영향이 커지고 있다. 이에 포스트 코로나를 준비하는 현시점에서 콘텐츠와 콘텐츠 산업의 내용이나 방향에 대한 분석과 앞으로의 방향에 대한 논의가 필요하다. 이러한 논의가 진행되는 중에 콘텐츠 연구개발에서의 주요 토픽과 연구동향 분석결과는 콘텐츠 관련 정책을 개발하고 진행하는데 중요한 시사점을 줄 수 있을 것이다.

2. 토픽 모델링 관련 연구현황

토픽 모델링은 텍스트 마이닝 및 자연어 처리(Natural Language Processing; NLP)를 위한 비지도 머신러닝 기법으로 텍스트로 구성된 빅데이터에서 숨겨진 구조를 밝히는데 사용되는 통계적 언어모형(Statistical Language Model)이다[4]. 토픽 모델링에서 토픽은 통계적으로 유의미하게 발생하는 단어 목록이고, 텍스트는 이메일, 책, 블로그 게시물, 저널, 신문 기사 및 모든 종류의 비정형 텍스트를 포함할 수 있다. 토픽 모델링을 활용한 국내 연구는 토픽 모델링 기법을 활용한 연구와 토픽 모델링 기법을 개선하는 연구로 크게 분류해 볼 수 있다.

먼저, 최근 토픽 모델링을 활용한 학술적 연구는 문헌정보학 분야뿐만 아니라 의료, 정보, 교육, 관광 분야 등에서 활발하게 진행되고 있다. 토픽 모델링을 활용한 연구를 진행한 연구자들은 해당 분야의 기존 연구들이 미시적이고 연구자의 주관성이 상당 부분 개입되었다는 한계점이 있었던 것에 비해 대규모 텍스트 자료를 이용한 토픽 모델링 활용 연구는 상대적으로 거시적이고 객관적이라는 강점이 있음을 주장하고 있다[5]. 최근에는 논문 초록을 주된 데이터로 분석하고 있으며 R과 같은 범용 툴을 사용하여 토픽 모델링 외에 시계열 분석, 네트워크 분석 등과 같은 추가적인 분석기법을 활용하는 사례가 있다[5-7]. 또한 토픽 모델링은 텍스트 형태의 온라인 리뷰를 분석할 때도 효과적인 것으로 알려져 있다[8][9].

다음으로, 토픽 모델링 기법을 개선하는 연구가 있다[표 1]. 토픽 모델링 알고리즘을 개선하는 기술적인 연구를 진행한 연구자들은 입력 텍스트 전처리과정에서 SVD(Singular Value Decomposition) 방법을 도입하거나, LDA 알고리즘과 Word2Vec을 혼합한 방법, Word2Vec과 클러스터링 기법을 혼합한 방법 등을 제안하였다[10-12]. 이 연구를 진행하였던 연구자들은 제안한 방법들이 토픽 모델링의 성능을 일정부분 개선하는 효과는 있었지만 소량의 텍스트 자료에 적용하는 경우에는 그 효과가 크지 않았다고 보고하였다[8].

표 1. 토픽 모델링 기법 개선 연구

논문	내용
Kyung Im Kim 외 (2009)	입력 텍스트 전처리 과정에서 SVD(Singular Value Decomposition) 방법 도입
Moody (2016)	LDA 알고리즘과 Word2Vec을 혼합하여 LDA2Word 알고리즘 제안
Won-joon Choi and Euhee Kim (2019)	Word2Vec기법과 클러스터링 기법, 토픽모델링을 혼합한 형태의 텍스트 분석 방법 제안
윤상훈, 김근형 (2021)	Word2Vec을 이용한 토픽 모델링 확장모형 제안

3. LDA 알고리즘

LDA(Latent Dirichlet Allocation) 알고리즘은 토픽 모델링에서 효과적으로 수행되는 것으로 알려져 있기 때문에 소셜 네트워크, 소프트웨어 공학, 정치학, 의학, 언어 과학 등 텍스트 마이닝 연구에 널리 빈번하게 활용되어왔다[13]. 이에 최근 국내에서도 LDA 알고리즘을 활용한 토픽 분석 연구가 활발하게 보고되고 있다[5]. NLP에서 LDA 알고리즘은 관찰된 텍스트 데이터 셋인 말뭉치(corpus)를 잠재 집단으로 설명할 수 있는 생성확률모형(Generative Probabilistic Model; 혹은 생성통계모형이라고도 함)이다. 여기서 잠재 집단은 토픽을 의미하며 데이터의 일부가 유사한 이유를 설명한다. LDA 알고리즘의 기본 발상은 모든 문서들이 특정 개수의 토픽을 기반으로 생성되고 문서에 포함된 각 단어들은 해당 토픽 단어로부터 무작위로 선택된다고 가정한다[14][15]. 또한 모든 문서의 토픽 분포가 공통으로 디리슈레 사전분포(Dirichlet Prior Distribution)를 공유한다고 가정한다. 즉 LDA 알고리즘의 각 잠재 토픽은 단어들에 대한 확률분포로 표현되며, 토픽들의 단어 분포도 디리슈레 사전분포를 공유한다.

[그림 1]은 LDA 문서 생성 절차를 도식화한 것이다. 짙은 색상으로 표시된 w 노드는 관찰값인 단어들을 의미하며 나머지 노드들은 LDA 진행 과정에서 결정된다. θ_d는 알파(α)에 의해, z_n은 θ_d에 의해, ϕ_t는 베타(β)에 의해 결정된다. 노드를 둘러싼 네모박스는 단어의 수(N), 문서의 수(D), 토픽 수(T) 만큼 각각 반복 수행을 통해 알고리즘이 실행됨을 의미한다. 해당 절차에서 α, β, 토픽 수는 연구자가 직접 설정해야 하는 하이퍼 파라미터(Hyper-Parameter)이며, 이 값들을 제외한 나머지 잠재변수는 추정해야 할 모수들이다. 구체적으로, 말뭉치 W와 개별 문서 d가 주어졌을 때, LDA 알고리즘의 절차는 다음과 같다. 먼저, 모수가 β인 디리슈레 분포로부터 토픽 $t(t\in\{1,....,T\})$에 대한 다항분포 ϕ_t를 선택한다. 다음으로, 모수가 α인 디리슈레 분포로부터 문서 $d(d\in\{1,....,D\})$에 대한 다항분포 θ_d를 선택한다. 마지막으로, 문서 d에 포함된 단어 w_n $(n\in\{1,....,N_d\})$에 대해, θ_d로부터 토픽 z_n을 선택하고, ϕ_{z_n}으로부터 단어 w_n을 선택한다. 여기서 개별 문서 d는 단어 N_d 개($d\in1,...,D$)를 포함하고, 말뭉치 W는 문서 D개의 집합이다. 한편, 잠재변수의 모수를 추론하기 위해 관찰값 W의 사후확률은 식(1)과 같이 계산하며 이를 최대화한다.

$$P(W|\alpha,\beta) \tag{1}$$
$$=\prod_{d=1}^{D}\int P(\theta_d|\alpha)(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(Z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta))d\theta_d$$

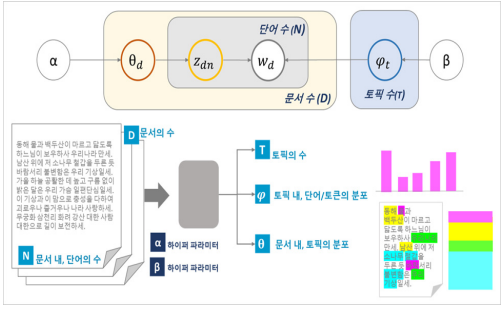

그림 1. LDA 문서 생성 절차 및 예시

Ⅲ. 연구 방법

연구범위는 한국콘텐츠학회에서 2001년부터 2021년 3월까지 출판된 국내 논문 총 9,860편의 학술논문의 초록이다. 본 연구는 한국콘텐츠학회 편집위원회로부터 학술연구를 위한 목적으로 한국콘텐츠학회 논문 자료를 활용할 것을 승인받고 진행하였다. 분석 프로그램은 파이썬(Python)을 활용하였다. [그림 2]는 본 연구에서의 주요 연구 절차를 도식화한 것이다. 먼저 웹 크롤링을 활용하여 9,860건의 초록으로 구성된 텍스트 데이터 셋을 생성한 후, 데이터 전처리를 단계적으로 수행함으로써 분석용 말뭉치를 생성하였다. 그리고 나서 최종 말뭉치를 활용하여 LDA를 이용한 토픽 분석을 실시한 후, 추출된 토픽에 대한 지난 20년간 추세 분석을 실시하였다.

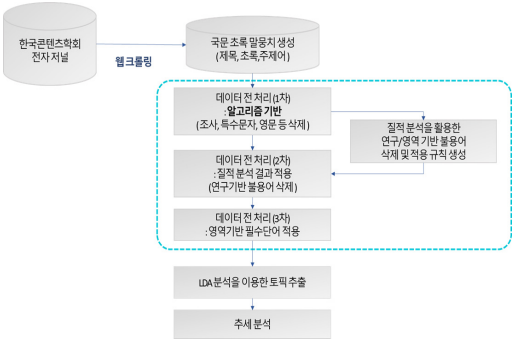

그림 2. 연구 절차

1. 크롤링을 활용한 국문 초록 말뭉치 생성

KoreaScience 웹페이지에서 크롤링 기법을 활용하여 한국콘텐츠학회 논문지의 국문 초록, 서명, 주제어를 수집하여 말뭉치를 생성하였다. 초록이 잘 읽혔는지 검증한 방법은 다음과 같다. 먼저, 읽힌 자료 중에 잘린 내용을 확인하였다. 이후에는 일부 표본을 선택하여 원래 파일과 대조하여 초록 중 일부가 특수 기호의 영향으로 잘리거나, 초록이 누락된 경우 원문을 확인하여 자료를 정비하였으며, 국문 초록이 결측된 2건을 제외한 총 9,858 건의 초록이 데이터 셋으로 입력됨을 확인하였다.

2. 질적 분석을 활용한 데이터 전처리

웹크롤링으로 수집한 텍스트 데이터는 분석을 위해 데이터 전처리 과정이 요구된다. 한국콘텐츠학회 논문지에 게재된 논문들의 토픽은 다양한 융·복합 분야와 관련이 있다. 이에 논문들의 토픽을 보다 뚜렷하게 분별하면서도 추출된 토픽들에 대한 해석의 타당성을 확보하기 위한 전략으로 토큰화(Tokenization), 정제(Cleaning), 정규화(Normalization)를 포함한 양적 분석과 융·복합 학문 기반 불용어 및 필수단어 목록 규칙을 생성하는 질적 분석을 병행하였으며, 연구자들이 합의한 의미 있는 말뭉치가 생성될 때까지 반복적으로 실시하였다.

데이터 전처리 과정을 요약하면 다음과 같다. 1단계에서는 마침표를 포함한 특수 문자들을 기준으로 문자열을 분리한 후, NLP 연구에서 일반적으로 사용되는 불용어인 특수 기호, 영어, 숫자, 문장 부호 등을 삭제하였다. 형태소 분석은 Konlpy 패키지의 okt 알고리즘을 이용하여 분석하였으며, 품사는 명사만 추출하여 약 28,000여 개의 고유 단어를 추출하였다. 2단계에서는 1단계에서 생성된 말뭉치 중, 출현 빈도가 20회 이상인 4,400여 개의 단어를 대상으로 연구·영역 기반 불용어 삭제 및 필수단어 목록을 위한 규칙을 생성한 뒤 2단계 규칙을 적용한 말뭉치를 다시 생성하였다.

이에 대하여 다시 설명하면, 1단계에서 생성된 말뭉치는 특정 단어의 출현 빈도가 매우 높을지라도 문서가 포함하고 있는 실제 토픽 간 차이를 실제적으로 반영하지 못하는 한계가 있을 수 있다. 예를 들어, 연구, 분석, 결과, 이용 등의 단어는 상위 10개 범위 내에 출현할 만큼 빈도가 높지만 대부분의 연구물에서 반복적으로 사용되기 때문에 융·복합 연구 토픽들을 섬세하게 분류하는데 변별력이 낮은 단어라고 할 수 있다. 이러한 이유로 심리학 박사 1인, 국어교육학 박사 1인, 광고 분석 전문가 1인이 연구·영역 기반 불용어 삭제 목록 및 필수단어 적용규칙을 생성하였다. 이 후, 데이터 사이언스 및 빅데이터 분석 전문가 3인과 합의를 통해 규칙을 확정하고 연구·영역 기반 삭제 목록 규칙을 적용한 말뭉치를 생성하였다.

마지막으로, 3단계에서는 2단계에서 생성한 연구·영역 기반 필수단어 목록을 적용하고 유사어를 통일하였다. 필수단어의 경우 기술 및 디지털 기반 연구와 밀접한 관련이 있는 5G, VR, 4차 산업 등의 외래어와 숫자가 포함된 핵심 단어를 말뭉치에 포함하였다. 그리고 유사어의 경우 '코비드'와 '코로나', '컨텐츠'와 '콘텐츠'와 같이 문맥상 동일한 의미를 갖는 유사어는 한 개 단어로 통일함으로써 추가적인 정제 작업을 반복적으로 시도하였다. 이상의 데이터 전처리 작업을 통해 약 3,000여 개의 고유 단어를 추출하였다.

3. LDA를 이용한 토픽 추출 및 타당성 검토

토픽 분석은 gensim 라이브러리의 LdaModel 함수를 이용하였다. 적절한 토픽 개수를 설정하기 위하여 혼잡도(Perplexity)와 간섭성(Coherence) 값을 계산하였다. 혼잡도의 기본 발상은 모형이 이전에 본 적이 없는 새로운 단어의 출현에 대해 얼마나 자연스럽지 않은지 포착하는 것이다. 혼잡도는 보류된 검증 데이터

셋의 정규로그우도(Normalized Log-Likelihood of a Held-Out Test Set)로 산출되며, 이 값이 작을수록 언어 모델의 성능은 좋다는 것을 의미한다. 간섭성은 토픽 내 단어들이 얼마나 의미론적으로 일관성이 있는지 판단하는 지표이며, 이 값이 클수록 의미론적 일관성이 높음을 의미한다. 이에 임의의 토픽 개수 20, 30, 50, 100을 할당하여 패턴을 탐색한 결과, 토픽 수가 적을수록 유의미한 값이 나타났음을 확인하였다.

이후 여러 번의 사전 분석 결과와 토픽 내 연관 단어들의 유의미성을 종합적으로 고려하여 최종 토픽 수는 8개를 선정하였다(혼잡도 = -6.816, 간섭성 = 0.514). 또한, 분류된 토픽의 타당성을 확보하기 위해 pyLDAvis 라이브러리의 시각화 분석기법을 이용하여 토픽 간 거리와 토픽별 연관 단어 상위 30개의 상대 점유율을 산출하였다[16][17].

4. 추출된 우세 토픽을 활용한 추세 분석

한국콘텐츠학회에서 과거 혹은 현재 활발하게 수행된 연구와 그렇지 않은 연구 토픽을 살펴보기 위해 2011년을 기점으로 2001년부터 2011년까지는 상반기로, 2012년부터 2021년 3월까지는 하반기로 분류한 후 워드 클라우드 및 토픽별 빈도 분석을 실시하였다. 또한 2001년부터 2021년까지 연도별 토픽 주제의 변화를 탐색하기 위해 단순선형회귀분석 및 추세 그래프 분석을 실시하였다. 선형회귀분석에서 추정된 회귀 계수는 단순히 증가 혹은 감소하는 직선 형태의 패턴만을 반영하므로 해당 토픽에 대한 연도별 논문 수의 실제 변화 패턴을 확인하는데 한계가 있다. 이에 추세 그래프 분석을 통해 지난 20년간 토픽별 변화 패턴을 심층적으로 탐색하였다.

Ⅳ. 연구 결과

1. 탐색 자료 분석

한국콘텐츠학회의 논문지에 게재된 연도별 논문 게재 건수는 [그림 3]과 같다. 2013년까지는 지속적으로 논문 게재 편수가 증가하였고, 2009년 이후에는 매년 600편 이상 논문이 게재되었다(단, 2021년 자료는 3월까지 수집된 결과임). 논문 게재 편수가 2009년에 급격히 증가한 현상이 나타났는데 이는 한국콘텐츠학회 논문지가 2008년에 KCI 등재지가 된 것과 관련이 있음을 시사한다. [그림 4]는 본 연구에서 생성한 최종 말뭉치에 포함된 빈출 단어를 워드 클라우드로 시각화한 것이다. 빈도가 높은 단어일수록 단어 크기는 더 크게 표현되는데, 본 연구에서는 '교육', '정보', '사회', '개발'의 빈출 빈도가 상대적으로 높은 것으로 나타났다. 이는 콘텐츠의 영역에서 '교육', '정보', '사회', '개발'과 관련된 쟁점과 논의가 활발히 진행되고 있고 콘텐츠 산업 분야에서도 해당 분야가 관련이 높을 가능성이 있음을 시사한다.

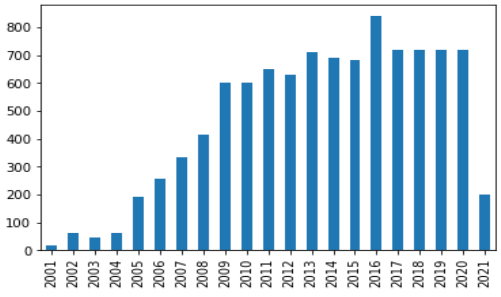

그림 3. 연도별 논문 게재 건수

그림 4. 빈출 단어 워드 클라우드

2. 토픽 분석 및 추출된 토픽의 타당성 검토

[표 2]는 토픽 분석을 통해 추출된 8개 우세 토픽에 대해 각 토픽별로 해당 토픽을 대표하는 관련성이 높은 연관 단어들 중 상위 10개를 순차적으로 정리한 것이다. 토픽의 비중을 살펴보면, 8개의 토픽 중 '산업·정책' 토픽(1,737 건)의 비중이 17.6 %로 가장 높게 나타났고, 그 뒤를 이어 '영상'(1,315 건)과 '정보기술'(1,309

건)의 비중이 13.3 %로 나타남으로써 과학·기술 분야 토픽이 상위권에 위치하였다. 반면 인문·사회 및 의학·보건 분야 토픽이 10 % 정도의 높은 비중으로 추출된 점도 주목할 만하다. 구체적으로, '사회복지'(1,244 건)는 12.6 %, '교육·인력개발'(1,192 건)은 12.1 %, '문화콘텐츠'(1,202 건)는 12.2 %, '마케팅·서비스'(1,012 건)는 10.3 %, '의료·보건'(847 건)는 8.6 %로 분류되었다.

표 2. 토픽 분석 추출 결과 – 8개 토픽과 주제어

토픽	우세 토픽	논문 수 (건)	비율 (%)	연관 단어(%)
1	산업·정책	1,737	17.6	기업(.017), 지역(.017), 산업(.015), 정책(.012), 사회(.010), 관리(.009), 기술(.008), 사업(.008), 제도(.008), 개발(.008)
2	사회복지	1,244	12.6	사회(.031), 노인(.021), 건강(.017), 스트레스(.014), 가족(.012), 청소년(.012), 삶(.012), 활동(.011), 경험(.011), 여성(.010)
3	교육·인력개발	1,192	12.1	학습(.035), 교육(.032), 조직(.030), 직무(.029), 만족(.016), 몰입(.015), 성과(.012), 교사(.010), 리더십(.010), 능력(.009)
4	영상	1,315	13.3	영상(.027), 콘텐츠(.023), 디자인(.020), 게임(.017), 평가(.015), 개발(.013), 애니메이션(.010), 제작(.010), 공간(.009), 변화(.009)
5	문화콘텐츠	1,202	12.2	문화(.025), 영화(.021),미디어(.014), 콘텐츠(.012), 한국(.012), 표현(.012), 관광(.011), 의미(.010), 이미지(.010),작품(.007)
6	마케팅·서비스	1,012	10.3	의도(.021), 고객(.019), 정보(.019), 품질(.016), 브랜드(.016), 광고(.014), 소비자(.014), 인터넷(.013), 모바일(.013), 만족(.012)
7	정보기술	1,309	13.3	정보(.028), 데이터(.024),기법(.021), 네트워크(.019), 알고리즘(.013), 모델(.013), 기술(.012), 성능(.012), 설계(.012), 효율(.012)
8	의료·보건	847	8.6	프로그램(.025), 환자(.021), 교육(.020), 의료(.020), 대학(.016), 간호(.016), 치료(.013), 병원(.012), 대학생(.012), 운동(.011)
전 체		9,858	100.0	

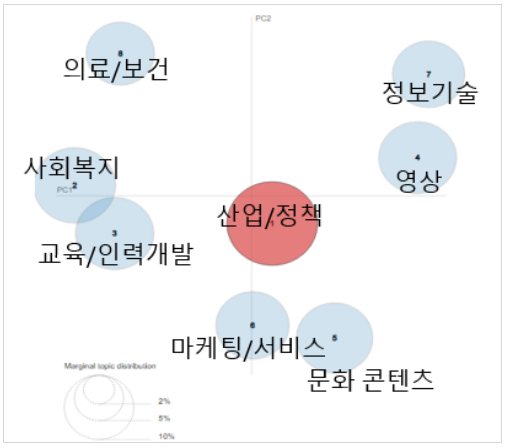

그림 5. 토픽 간 거리 맵

[그림 5]는 pyLDAvis 라이브러리를 이용하여 토픽 간 거리를 시각화한 것이다[16][17]. 토픽 모델링은 유용한 방법이지만 단어의 조합을 찾는 것만으로 이해하기 어렵다. pyLDAvis는 분류된 8개 우세 토픽 간 거리를 계산하여 중심이 결정되는 2차원 평면의 원으로 그린 다음, MDS(Multi-Dimensional Scaling)기법을 활용하여 토픽 간 거리를 2차원으로 투영함으로써 시각화한 그림을 제공해준다. 해당 그림에서 각각의 원은 추출된 8개의 토픽이고 원의 크기는 토픽의 비중을, 원과 원 사이의 거리는 유사성의 정도를 의미한다. 즉, 원의 크기가 클수록 최종 생성된 말뭉치에서 추출된 토픽의 비중이 높고, 각 원 사이의 거리가 멀수록 유사성이 낮다고 해석한다[16]. 예를 들어 '사회복지'와 '교육·인력개발' 원은 겹쳐진 부분이 있기 때문에 다소 유사성이 있음을 보여주지만, 이 두 가지 토픽과 '정보기술'과 '영상'은 상당한 거리가 있는 것으로 나타났기 때문에 유사성이 매우 낮은 토픽이라고 해석할 수 있다. [그림 5]의 네 영역을 나누어 해석하는 것도 의미가 있다. 먼저, 1사분면에 위치한 '정보기술'과 '영상'은 매체의 속성이 유사하며, 3사분면과 4사분면에 위치한 '산업·정책', '마케팅·서비스'와 '문화 콘텐츠'는 산업적인 성향이 유사한 주제라고 볼 수 있다. 다음으로, 2사분면과 3사분면에 위치한 '의료·보건', '사회복지' 및 '교육·인력개발'은 인간 대상의 연구 영역으로 인간이 자신이 속한 사회에서 유의미한 삶을 살아갈 수 있도록 지원하고자 하는 유사한 목적을 갖고 있는 영역이다.

3. 상반기와 하반기 빈도분석

[그림 6]과 [그림 7]은 2011년을 기점으로 2001년부터 2011년까지는 상반기로, 2012년부터 2021년 3월까지는 하반기로 분류한 후, 상·하반기 시점별 워드 클라우드를 분석한 결과이다. 상반기에는 '시스템', '정보', '서비스', '기술', '개발' 등 과학기술 용어가 굵고 큰 글씨로 표기되어 있는 반면, 하반기에는 '사회', '교육', '인식', '문화', '만족' 등 인문·사회과학 분야 용어가 굵고 큰 글씨로 표기되었다.

[표 3]은 상반기와 하반기를 구분하여 토픽별 논문 수와 해당 기간 내에서의 토픽 비중을 산출하여 정리한 것이다. 상반기에 출판된 논문 수는 3,230편, 하반

기에 출판된 논문 수는 6,628편으로 상반기에 비해 하반기에 게재된 논문 수는 약 2배 증가하였다. 상반기의 경우 정보기술(26.2 %)과 영상(19.4 %) 두 개 토픽이 전체의 약 45 %를 차지하며 정보기술 과학 영역이 우세한 반면, 하반기에는 산업·정책(18.7 %), 사회복지(16.4 %), 교육·인력개발(13.3%), 문화콘텐츠(13.2 %), 마케팅·서비스(10.7 %) 등 인문・사회분야 영역이 전반적으로 우세한 것으로 나타났다.

그림 6. 상반기(2001년 ~ 2011년) 워드 클라우드

그림 7. 하반기(2012년 ~ 2021년) 워드 클라우드

표 3. 상·하반기 추출 토픽별 해당 논문 수와 비중 비교

토픽	상반기 (2001~2011)		하반기 (2012~2021)	
	논문 수 (개)	비율 (%)	논문 수 (개)	비율 (%)
산업·정책	497	15.4	1,240	18.7
사회복지	154	4.7	1,090	16.4
교육·인력개발	310	9.6	882	13.3
영상	628	19.4	687	10.4
문화콘텐츠	327	10.1	875	13.2
마케팅·서비스	298	9.2	714	10.7
정보기술	847	26.2	462	6.9
의료·보건	169	5.2	678	10.2
전체	3,230	100.0	6,628	100.0

주목할 점은 상반기에 26.2 %로 가장 비중이 높았던 '정보기술' 토픽이 하반기에는 6.7 %로 가장 낮은 비중을 차지하였고, 상반기에 4.7 %로 가장 비중이 낮았던 '사회복지' 토픽이 하반기에는 16.4 %로 약 4배수 증가한 것이다. 또한 '영상' 토픽의 비중은 2배수 감소한 반면, '의료·보건' 토픽의 비중은 2배수 증가하였다.

[그림 8]은 상·하반기 토픽 논문 수를 토픽별로 비교한 것이고, [그림 9]는 상반기에 게재된 3,230편의 논문에서 8개 토픽별 게재 비율과 하반기에 게재된 6,628편의 논문에서 8개 토픽별 게재 비율을 비교한 그래프이다.

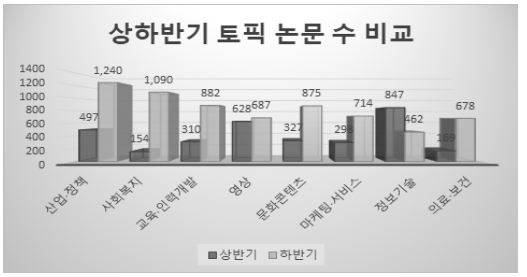

그림 8. 상·하반기 토픽별 논문 수 비교

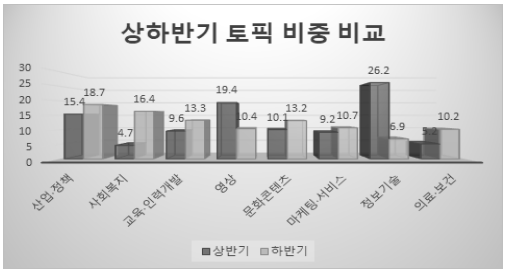

그림 9. 상·하반기 토픽 비중 비교

4. 단순선형회귀분석과 그래프를 이용한 추세 분석

[표 4]는 2001년부터 2021년까지 연도에 따른 해당 토픽별 논문 수에 대한 단순선형회귀분석을 실시한 후, 표준회귀계수(β)와 결정계수(R^2)를 정리한 것이다. 선형회귀분석에서 추정된 회귀 계수는 단순히 증가 혹은 감소하는 직선 형태의 패턴만을 반영하므로 해당 토픽에 대한 연도별 논문 수의 실제 변화 패턴을 확인하는 데 한계가 있다. 이에 [그림 10]의 추세 그래프에서 나타난 지난 20년 간 시간에 따른 변화 패턴을 '상승', '상승 후 하강' 으로 분류한 후, 이 결과를 [표 4]에 추가로 정리하였다.

표 4. 전체 기간 변수를 고려한 회귀분석 결과

토픽	β	R^2	패턴
산업·정책	9.19*	0.91	상승
사회복지	10.71*	0.93	상승
교육·인력개발	6.68*	0.92	상승
영상	1.55	0.22	상승 후 하강
문화콘텐츠	6.01*	0.90	상승
마케팅·서비스	5.04*	0.79	상승
정보기술	-3.38	0.66	상승 후 하강
의료·보건	5.20*	0.72	상승

주. *($p < 0.05$)

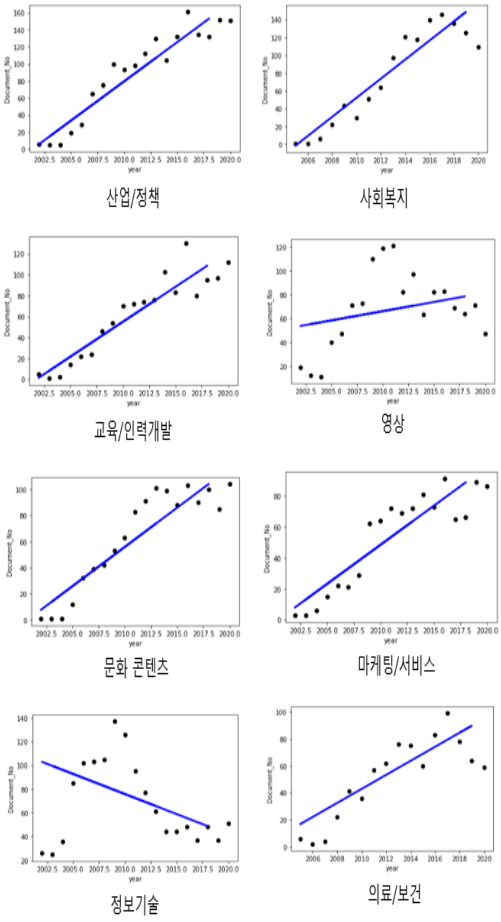

그림 10. 토픽 주제별 추세 그래프 및 회귀 직선

지난 20년간 토픽별 변화 패턴을 심층적으로 탐색한 결과, 8개의 토픽 중에서 최근까지 꾸준한 증가세를 유지한 토픽은 '사회복지(β = 10.71, R^2 = 0.93)', '산업·정책(β = 9.19, R^2 = 0.91), '교육·인력개발(β = 6.68, R^2 = 0.92)', '문화콘텐츠(β = 6.01, R^2 = 0.90)'로 나타났다. 또한 '사회복지'와 '의료·보건(β = 5.20, R^2 = 0.72)' 토픽이 최근까지 꾸준한 증가세를 유지하였으나 2017년 이후 꾸준한 감소세가 나타난 점이 주목할 만하다. 한편 '영상(β = 1.55, R^2 = 0.22)' 토픽은 2010년까지 꾸준히 증가하였으나 2011년 이후 점차 감소하는 경향이 나타났다. 그리고 '정보기술(β = -3.38, R^2 = 0.66)' 토픽의 경우 2008년까지 꾸준히 증가하였으나 2009년 이후 급격한 감소세를 보이다가 2013년부터 최근까지 게재된 논문 수는 상당히 미미한 것으로 나타났다.

V. 결론 및 논의

본 연구에서는 한국콘텐츠학회 논문지에 게재된 논문 자료를 이용하여 토픽 분석을 실시하였고, 추세 분석을 통해 지난 20년간 국내 콘텐츠 관련 중요한 관심사를 체계적으로 분석하였다. 본 연구의 의의는 다음과 같다.

첫째, 토픽 분석을 활용하여 지난 20년간 한국콘텐츠학회 논문지에 게재된 논문으로부터 8개 영역(산업·정책, 사회복지, 교육·인력개발, 영상, 문화콘텐츠, 마케팅·서비스, 정보기술, 의료·보건)의 우세 토픽을 추출하였다. 추출된 토픽별로 단순선형회귀분석 및 추세 그래프 분석을 통해 지난 20년간 토픽별 변화 패턴을 심층적으로 분석함으로써 한국콘텐츠학회 논문지 연구 자료에서 '콘텐츠'라는 개념이 내포하고 있는 토픽들의 특성을 도출하였다. 특히 추세선의 변곡점이 나타난 특정 시점에 주목하여 해당 토픽의 연구동향에 영향을 미친 외적 변인을 탐색하였고 토픽과 외적 변인 간 관련성을 파악하였다. 이러한 연구결과는 본 연구 결과의 신뢰성과 타당성을 지지하는 근거가 될 수 있다.

이를 구체적으로 설명하면 다음과 같다. 먼저, 논문 게재 편수가 2009년에 급격히 증가한 현상이 나타난 것은 한국콘텐츠학회 논문지가 2008년도 학술지 평가에서 KCI 등재학술지로 선정된 것과 관련이 있음을 시

사한다. 다음으로, 시점별로 2012년 이전에는 공학기술, 2012년 이후에는 사회과학 토픽 출현 비중이 높게 나타났다. 구체적으로, 상반기에는 '정보기술'과 '영상'이 약 50%를 차지하였으나, 하반기에는 인문·사회분야 토픽들이 전반적으로 우세하였다. 특히 '사회복지' 토픽의 경우 상반기 대비 하반기에 약 4배수 증가세가 나타난 것이 주목할 만하다. 이는 콘텐츠 R&D 개념이 인문·사회과학 및 예술 등 창의성의 근간이 되는 기초 R&D의 개념 도입과 범위 확장을 통해 보완의 필요성이 2010년에 제기됨으로써 2011년 이후 국내 연구에서 '콘텐츠' 개념이 인문·사회 연구영역을 포괄하는 개념으로 확장된 것과 관련이 있음을 보여준다[2].

마지막으로, '사회복지'와 '의료·보건' 토픽이 최근까지 꾸준한 증가세를 유지하였으나 2017년 이후 꾸준한 감소세가 나타난 점도 주목할 만하다. 이는 「생명윤리 및 안전에 관한 법률」 제 12조 제1항에 의거하여 2013년 2월부터 인간 대상 연구와 인체 유래물 연구를 수행하는 대학, 전문연구기관, 의료기관, 기업연구소 등에 기관생명윤리위원회(IRB, Institutional Review Board) 설치가 의무화됨에 따라 '의료·보건'과 '사회복지' 분야 연구자들의 연구가 한시적으로 위축되었을 가능성을 시사한다[18][19]. IRB는 연구대상자의 보호를 목적으로 연구계획서의 과학적 윤리적 타당성을 심사하기 위해 연구기관 내 설치하는 자율적 심의기구이다.

둘째, 본 연구에서는 의미 있는 연구결과를 얻기 위해 토픽 모델링 평가에 일반적으로 사용되는 계량적 평가 기법뿐만 아니라 산업 및 다학제 연구자 간 브레인스토밍을 통해 정성적 평가 기법을 도입하여 양질의 말뭉치를 생성하였고, 추출된 토픽들의 해석 가능성을 확보하였다. LDA와 같은 토픽 모형이 말뭉치의 예측 및 잠재 토픽을 제시해준다는 장점이 존재하지만, 동시에 비지도 머신러닝 프로세스로 인해 발견된 토픽의 정확성을 평가하기 어렵고 결과를 비교할 수 있는 표준 지표가 존재하지 않는다는 제한점도 있다[19]. 그럼에도 토픽 모형이 객관적으로 좋은지 혹은 나쁜지 식별하고 토픽 결과가 타당한지에 대한 객관적인 측정이 필요하다.

이러한 맥락에서 산업 및 다학제 기반 연구자들이 데이터 전처리 단계에서 연구 기반 필수단어 및 삭제 규칙을 생성하여 이를 적용함으로써 최종 말뭉치를 생성하였고, 혼잡도와 간섭성 지표를 이용하여 토픽 수를 결정하였다. 그리고 토픽별 핵심 단어 중에 중복되는 단어들의 비중이 낮고 토픽 내 상대점유율이 높은 단어들의 의미를 종합하여 토픽 주제를 선정하였다. 또한 시간에 따른 토픽 게재 빈도의 변화 패턴 분석을 통해 과거 시점의 중요한 쟁점과의 관련성을 도출하고 추출된 토픽들의 해석 가능성을 확인함으로써 실제적이면서도 유의미한 결과를 산출하였다.

셋째, 한국콘텐츠학회지 학술영역 범주화 및 전문가 구성 등 향후 발전적인 운영전략을 수립하는데 근거자료로 활용할 만한 가치가 있다. 한국콘텐츠학회는 한국콘텐츠 산업의 발전과 바람직한 지식정보화 사회의 발전에 기여하기 위한 목적으로 2000년 3월 16일에 창립되었으며, 현재 4차 산업혁명의 중심이 되는 융합 학문 분야를 리드하는 다양하고 폭넓은 주제들을 포함하고 있다. 이러한 학회의 지향점을 현실적으로 개선하기 위해서는 학술지와 학술대회 하위 분야의 구조화가 요구된다. 일례로, 2021년 10월 기준 한국콘텐츠학회 온라인 논문 투고 시스템의 분류는 콘텐츠공학(콘텐츠제작기술, 콘텐츠서비스기술, IT기반기술, 정보통신시스템기술), 콘텐츠디자인(콘텐츠기반이론, 인터랙티브콘텐츠, 동영상콘텐츠, 출판콘텐츠, 문화예술콘텐츠), 콘텐츠응용(사회과학콘텐츠, 생활문화콘텐츠, 과학기술정보콘텐츠, 교육콘텐츠), 기타 영역으로 나뉘어 있다[21]. 본 연구 결과에 따르면, 한국콘텐츠학회 논문지에 2010년 이후 사회복지, 마케팅·서비스 및 의료보건 분야 연구 비중이 공학 및 기술 분야 연구에 비해 증가하는 패턴이 확인되었다. 이에 사회과학 및 인문과학 영역을 세분화하고 전문가를 영입함으로써 양질의 우수한 논문이 지속적으로 생산될 수 있기를 바라는 바이다.

본 연구가 토픽 모델링을 이용하여 한국콘텐츠학회 논문지의 연구동향을 체계적으로 파악해보았다는 면에서 연구의 의의가 있지만, 토픽의 개수를 8개로 선정함으로써 본 연구에서는 규명하지 못한 중요한 관심사가 있을 수 있다. 예를 들어, 2010년대 하반기에 빅데이터와 인공지능이 사회적, 정치적으로 많은 관심을 받았음에도 추출된 8개의 토픽 혹은 연관 단어 목록에서는 관련 용어가 나타나지 않았다는 점은 주목할 만하다. 이는 주된 데이터로 논문 제목, 키워드, 논문 초록을 사용

하여 분석하였다는 분석의 한계점을 지적할 수 있다. 그럼에도 본 연구에서 한국콘텐츠학회 논문지에 2010년 이후 인문·사회과학 분야 연구비중이 공학·기술 분야에 비해 높게 분석된 점을 고려해 볼 때, 실제적으로 인공지능이나 빅데이터 관련 연구의 투고 비중이 사회적 관심에 비해 낮았다고 유추할 수 있다.

이상의 논의를 통해, 본 연구는 토픽 분석을 체계적으로 활용함으로써 실제 연구현황을 내포하고 있는 빅데이터에서 숨겨져 있는 패턴을 발견하였다는 점에서 의의가 있다. 이에 후속 연구에서는 '콘텐츠'라는 개념을 포괄하는 자료를 추가로 수집하여 우리나라 콘텐츠 관련 연구의 전반적인 연구동향을 파악해 볼 수 있을 것이다. 본 연구 결과가 국내 콘텐츠 관련 연구 개발 및 산업 분야에서 진행되고 있는 활발한 논의에 학술적인 시사점을 제공할 수 있기를 기대한다.

참 고 문 헌

[1] G. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet, and D. Delen, "Practical text mining and statistical analysis for non-structured text data applications," Elsevier Science & Technology, Waltham, https://doi.org/10.1016/C2010-0-66188-8, 2012.

[2] 한국콘텐츠진흥원, *디지털 콘텐츠 발전을 위한 인문·사회과학 통합형 R&D 모델 개발 기초 연구*, kocca 연구보고서, pp.10-47, 2010.

[3] 이수범, "콘텐츠 산업의 포스트 코로나19 이슈 탐색 연구 : 신문기사의 텍스트 마이닝 분석을 중심으로," 언론문화연구, 제30호, pp.35-70, 2021.

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

[5] 원용국, 김영우, "토픽 모델링을 활용한 한국 영어교육 학술지에 나타난 연구동향 분석," 한국콘텐츠학회논문지, 제21권, 제4호, pp.50-59, 2021.

[6] 박주섭, 홍순구, 김종원, "토픽모델링을 활용한 과학기술동향 및 예측에 관한 연구," 한국산업정보학회논문지, 제22권, 제4호, pp.19-28, 2017.

[7] 진설아, 송민, "토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구," 정보관리학회지, 제33권, 제1호, pp.7-32, 2016.

[8] 윤상훈, 김근형, "Word2Vec를 이용한 토픽모델링의 확장 및 분석사례," 정보시스템연구, 제30권, 제1호, pp.45-64, 2021.

[9] 장재윤, 최연재, 강지연, "국내 ICT 업종 종사자들의 직장에 대한 불만 요인 분석 및 전/현직자 간 차이 분석: 토픽 모델링 적용," 한국심리학회지: 일반, 제39권, 제3호, pp.445-480, 2020.

[10] K. Kim, N. C. T. Hai, and H. R. Park, "SVD-LDA: A Combined Model for Text Classification," JIPS(Journal of Information Processing Systems), Vol.5, No.1, pp.5-10, 2009.

[11] C. E. Moody, "Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec," arXiv Preprint arXiv: 1605.02019, 2016.

[12] W. Choi and E. Kim, "A Large-scale Text Analysis with Word Embeddings and Topic Modeling," Journal of Cognitive Science, Vol.20, No.1, pp.147-187, 2019.

[13] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and Topic modeling: models, applications, a survey," Multimedia Tools and Applications, Vol.78, Issue 11, pp.15169-15211, 2019.

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

[15] A. Gruber, Y. Weiss, and M. Rosen-Zvi, "Hidden topic Markov models," Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, Vol.2, pp.163-170, 2009.

[16] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp.63-70, 2014.

[17] J. Chuang, D. Ramage, C. Manning, and J. Heer, "Interpretation and trust: Designing model-driven visualizations for text analysis," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,

pp.443-452, 2012.
[18] 보건복지부, "기관생명윤리위원회 지원을 위한 포털 사이트 오픈," 보건복지부 보도자료, 2013. 01. 15.
[19] J. Kim and J. Kim, "Institutional review board and research ethics," THE JOURNAL OF THE KOREAN ACADEMY OF PEDIATRIC DENTISTRY, Vol.41, Issue 2, pp.187-192, 2014.
[20] L. Tay, S. Woo, L. Hickman, and R. Saef, "Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining," European Journal of Personality, Vol.34, No.5, pp.826-844, 2020.
[21] 한국콘텐츠학회, 온라인 논문투고 시스템, http://acoms.atit.co.kr:7090/kocon/index.jsp?publisher_cd=kocon&cid_year=null&cid_seq=null&lang=null&menu=null, 2021. 10. 26.

저 자 소 개

석 혜 은(Hye-Eun Seok) 종신회원

- 1998년 2월 : 이화여자대학교 통계학과(이학석사)
- 2020년 2월 : 이화여자대학교 심리학과(박사, 석박통합과정)
- 현재 : KAIST 빅데이터 엣지 클라우드 서비스 연구센터 자문위원, 대한창상학회 통계편집위원

〈관심분야〉 : 심리측정, 연구 방법론, 융·복합 연구, 머신러닝, 자연어처리(NLP) 등

김 수 영(Soo-Young Kim) 정회원

- 1997년 2월 : 이화여자대학교 통계학과(이학사)
- 1999년 2월 : 이화여자대학교 통계학과(이학석사)
- 2000년 ~ 현재 : 제일기획 미디어 플래너

〈관심분야〉 : 미디어플래닝, 광고마케팅 효과 측정 및 분석, 빅데이터, 머신러닝, 자연어처리(NLP)

이 연 수(Yeon-Su Lee) 정회원

- 1997년 2월 : 이화여자대학교 통계학과(이학사)
- 1999년 2월 : 이화여자대학교 통계학과(이학석사)
- 현재 : 프리랜서 데이터 사이언티스트

〈관심분야〉 : 빅테이터, 머신러닝, 자연어처리(NLP)신용리스크

조 현 영(Hyun-Young Cho) 정회원

- 1997년 2월 : 이화여자대학교 통계학과(이학사)
- 1999년 2월 : 이화여자대학교 통계학과(이학석사)
- 현재 : 프리랜서 데이터 사이언티스트

〈관심분야〉 : 머신러닝, 빅데이터, 자연어처리(NLP), CRM

이 수 경(Soo-Kyoung Lee) 정회원

- 1998년 2월 : 이화여자대학교 통계학과(이학사)
- 2000년 2월 : 이화여자대학교 통계학과(이학석사)
- 2000년 ~ 2001년 : SAS Korea
- 2001년 ~ 2005년 : Bearingpoint Korea
- 2019년 ~ 현재 : Kurvv.ai

〈관심분야〉 : 딥러닝, 머신러닝, 자연어처리(NLP)

김 경 화(Kyoung-Hwa Kim) 정회원

- 2002년 8월 : 이화여자대학교 교육학과(문학석사)
- 2009년 2월 : 고려대학교 국어교육학과(교육학석사)
- 2014년 2월 : 고려대학교 국어교육학과(교육학박사)
- 2020년 9월 ~ 현재 : 청주대학교 국어교육과 조교수

〈관심분야〉 : 국어교육, 리터러시 교육, 작문 교육, 독서 교육 등