https://doi.org/10.15813/kmr.2022.23.3.007

디지털 전환: D.N.A.(Data, Network, AI) 키워드를 활용한 토픽 모델링

Digital Transformation: Using D.N.A.(Data, Network, AI) Keywords Generalized DMR Analysis

안세환 (Sehwan An) 한양대학교[1)]
고강욱 (Kangwook Ko) 한양대학교[2)]
김영민 (Youngmin Kim) 한양대학교[3)]

〈 국문초록 〉

디지털 전환의 핵심 인프라로서 데이터·네트워크·인공지능(D.N.A.) 분야의 확산과 유망 산업의 등장은 경제 전반에 걸쳐 활발한 디지털 혁신의 기반이 되고 있다. 본 연구에서는 텍스트마이닝 방법론을 적용하여 WoS 데이터베이스의 SCIE급 색인에 해당하는 연구의 초록, 출판연도 및 연구분야를 입력변수로 활용하여 주요 토픽을 도출하였다. 우선, 단어 출현 빈도에 기반한 TF 및 TF-IDF 분석을 통해 주요 키워드를 확인하고, 이어서 g-DMR(Generalized Dirichlet-Multinomial Regression)을 이용하여 토픽 모델링을 수행하였는데, 다양한 형태의 변수를 메타정보로 활용 가능한 해당 토픽 모형의 이점으로 단순하게 토픽을 도출하는 것 이상의 의미를 적절하게 탐색할 수 있었다. 분석 결과에 따르면, 비즈니스 인텔리전스, 제조 생산 시스템, 서비스 가치 창출, 원격 진료, 디지털 교육 등의 토픽들이 디지털 전환에서 주요 연구주제인 것으로 식별되었다. 토픽 모델링의 결과를 요약하자면, 1) COVID-19 이후 비즈니스 인텔리전스를 주제로 하는 연구가 전 영역에서 활발하게 수행되고 있으며, 2) 제조 분야에서 지능형 제조 솔루션 및 메타버스 등의 이슈가 등장함에 따라 제조 생산 시스템에 관한 주제가 다시 한번 주목받고 있음을 확인하였다. 마지막으로, 3) 주제어 자체는 기술과 서비스의 측면에서 분리하여 볼 수 있지만, 다수의 연구에서 해당 기술들을 접목하여 적용된 다양한 서비스를 포괄적으로 다루고 있으므로 이를 별개로 해석하는 것이 바람직하지 못하다는 점을 알 수 있었다.

주제어: 디지털 전환, 데이터, 네트워크, 인공지능, 토픽 모델링

1) 제1저자, hwan86@hanyang.ac.kr
2) 제2저자, system@hanyang.ac.kr
3) 교신저자, yngmnkim@hanyang.ac.kr

논문접수일 2022년 05월 15일 | 1차 수정 2022년 08월 08일 | 게재확정일 2022년 08월 15일

1. 서론

4차 산업혁명과 디지털 대전환(Digital Transformation, DT 또는 DX) 과정에서 인공지능(AI), 인터넷기반자원공유(Cloud), 가상/증강현실(VR/AR), 사물인터넷(IoT) 등 혁신 성장을 주도할 미래 유망 산업이 부각되고 있다. 특히, 디지털 전환의 핵심 인프라로서 데이터·네트워크·인공지능(Data, Network, AI, D.N.A.) 분야의 확산과 유망 산업의 등장은 경제 전반에 걸쳐 활발한 디지털 혁신의 기반이 되고 있으며, 글로벌 기업 중 하나인 아마존의 경우 데이터, 네트워크 및 인공지능 효과를 통해 업계를 디지털 방식으로 변화시키는 데 앞장서며, 2021년까지 미국 전자상거래 시장 점유율의 약 40%를 차지하고 있다(Siebel, 2018). 디지털 전환의 속도는 전 세계적으로 빠르게 나타나고 있으며, 이는 디지털 경제에서 가장 큰 비중을 차지하는 전자상거래의 규모가 2019년 1조 9천억 달러에서 2021년 2조 7천억 달러로 약 40% 증가할 것으로 예측된다. 또한 같은 기간 게임·음악·영상 등 디지털 콘텐츠 규모는 1천 8백억 달러에서 2천 1백억 달러로 약 20% 증가하고, 배달 플랫폼 규모는 1천 1백억 달러에서 1천 5백억 달러로 42% 증가할 것으로 예측된다(정준화, 박소영, 2020 재인용).

디지털 전환은 'D.N.A.' 생태계를 바탕으로 AI, IoT, 모바일, 빅데이터, 클라우드 컴퓨팅 등 디지털 기술의 발전에 따라 산업구조가 변화하고, 조직, 프로세스, 전략, 비즈니스 모델 등 기업 경영 전략의 모든 것들에서 변혁을 의미한다(손권상, 권오병, 2021; 이웅배 등, 2021). 포스트 코로나 시대 정보통신기술(ICT)을 중심으로 한 디지털 전환의 중요성이 높아지고, 디지털 뉴딜 등 주요 정책이 추진되면서 4차 산업혁명이 가속화되고 있다. 이처럼 4차 산업혁명 및 디지털 전환이라는 키워드가 경제·사회의 모든 영역에서 화두로 떠오르면서 관련 산업과 기술에 대한 키워드에 기반한 다양한 연구들이 이루어지고 있다(김태후, 한능호, 2021). 다양한 종류의 대규모 데이터에 대한 수집, 분석, 시각화 등을 특징으로 하는 빅데이터 기술의 발전은 이미지 및 텍스트와 같은 비정형 데이터로부터 의미를 추출하고 결과를 분석함으로써 가치 있는 정보 제공을 가능하게 하였다.

최신 연구동향이나 기술 분야의 이슈 등을 파악하기 위해 정량적 분석 기법 중 하나인 텍스트 마이닝(Text Mining)이 활발하게 사용되고 있으며, 이러한 텍스트 분석 기법에는 클러스터 분석(Cluster Analysis), 소셜 네트워크 분석(Social Network Analytics), 토픽 모델링(Topic Modeling) 등이 있다.

본 연구는 토픽 모델링 기법 중 일반화된 디리클레 다항 회귀(Generalized Dirichlet Multinomial Regression, G-DMR)를 적용하여 분석한다. 가장 기초적인 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)모델의 경우 문헌별 주제 분포만을 입력으로 활용하는 반면, DMR은 메타데이터(제목, 저자, 저널명, 출판연도 및 피인용수 등)를 포함하여 주제 분포를 추정할 수 있다는 장점이 있다.

본 연구에서는 디지털 전환의 'D.N.A.' 키워드 중심의 토픽 분석을 위해 디지털 전환과 토픽 모델링의 개념에 대해 고찰하고, Web of Science(WoS)의 데이터베이스에서 디지털 전환과 관련된 데이터, 네트워크 및 인공지능 중심의 키워드를 포함하는 연구 자료에서 출판연도 및 연구분야에 해당하는 메타데이터를 입력 데이터로 선정하였다. 구체적으로, 디지털 전환 분야에 대해 빈도 기반의 텍스트 분석과 함께 문헌별 부가적 변수(정보)에 해당하는 메타데이터를 활용하여 토픽을 도출할 수 있는 G-DMR 모형을 이용한 심층 분석을 수행하였다.

2. 이론적 배경

2.1. D.N.A. 중심의 디지털 전환 개념

디지털 시대의 핵심 인프라인 데이터, 네트워크, 인공지능을 기반으로 경제·사회 전반에서 디지털 전환을 모색하기 위한 전략의 일환으로 '디지털 뉴딜'의 개념이 소개되었다. 특히, 포스트 코로나 시대에 접어들면서 비대면 사회로의 전환을 위해 디지털 전환이 필수적으로 요구되었고, 이에 발맞추어 디지털 뉴딜이 적극적으로 추진되었다. 한국형 디지털 뉴딜의 핵심 과제는 데이터의 수집, 개방, 활용에서부터 인공지능을 접목한 서비스 창출에 이르기까지 데이터의 전 주기 생태계를 강화하기 위한 과제, 산업 현장에 5G 통신 및 인공지능 기술을 접목하는 융합 과제, 5G 통신 업무망과 클라우드 기반의 공공 스마트 업무환경을 구현하는 지능형 과제 등으로 구성되었다(박문우, 2020).

진정한 의미의 디지털 전환을 위해서는 디지털 시대에서 필요한 데이터가 무엇이고, 어떻게 그러한 데이터를 확보하고 품질을 높일 수 있는지에 대해 지속적인 논의가 요구된다. 또한 정보를 저렴하고 빠르게만 전달하는 것이 중요했던 기존의 네트워크와 달리 디지털 시대의 네트워크는 자율주행차·스마트시티와 같이 중요한 가치 또는 가치와 직결되는 데이터를 전달하는 것이므로 단순한 통신망이 아니라 디지털 시대의 '삶의 기반'으로 인식하는 것이 필요하다. 자율주행차, 스마트공장, 생활안전, 에너지 절감 등 다양한 분야에서 인공지능 기술 도입을 확대하여 기존 산업을 고도화하고, 국민이 체감할 수 있는 지능화 혁신이 이루어져야 한다. 이처럼 데이터가 디지털 경제의 원유라면, 그 엔진이 될 AI와 고속도로가 될 네트워크가 결합하여 전 산업의 지능화가 진전되고, 사회 전반의 디지털 포용 산업 및 서비스의 등장과 확산이 본격화될 것으로 전망된다(문형돈, 2021).

이처럼 인공지능이 지원하는 데이터 중심의 의사결정과 스마트 네트워크는 생산 현장의 작업 환경에 지속적인 변화를 가져왔다(Ralph & Stockinger, 2020). 인공지능, 빅데이터 등의 최신 디지털 기술을 빠르게 조직 내로 도입하여 변화할 수 있었던 기업들은 이를 활용한 타겟 마케팅 등으로 경영을 최적화하고 이윤을 극대화할 수 있었지만, 그렇지 못한 기업들은 점점 도태되어 격차가 확대되고 있는 실상이다(노규성, 2020). 궁극적으로 D.N.A. 환경에서 기업의 경쟁력을 확보하기 위해서는 데이터의 수집 및 가공 단계에서부터 AI 모델링 및 적용, AI 인프라 구축, 그리고 AI 서비스 확대 등의 절차가 핵심이고, 이 절차를 얼마나 신속하게 처리할 수 있느냐가 관건이다.

2.2. 토픽 모델링

LDA와 같은 토픽 모델은 문서 집합 및 기타 개별 데이터의 통계 분석에 유용한 도구이며, 문서 집합 내에 존재하는 잠재적 주제에 대해 단어들의 분포를 기반으로 식별해내는 분석 기법이다(Blei et al., 2003). 토픽 모형은 사용자가 지정한 토픽의 개수만큼 문서별 토픽 확률을 도출하고, 특정 문서에서 가장 높은 확률을 갖는 토픽을 해당 문서의 주제로 간주할 수 있다.

토픽 모델은 모형의 통계적 가정에 따라 초기의 LDA 모형으로부터 확장된 형태의 DTM(Dynamic Topic Model), ATM(Author Topic Model), CTM(Correlation Topic Model), STM(Structural Topic Model) 등의 기법으로 진화하였으며, 다양한 텍스트 자료의 분석에 활용되고 있다(Blei et al., 2003; Blei & Lafferty, 2007; Roberts et al., 2014). 김선주, 김병수(2021); 윤혜정 등(2021)의 연구에서는 LDA를 이용하여 서비스업에서

의 고객 충성도 요인과 사용자 인식을 분석하였고, 안재영 등(2022)은 메타버스 개념 등장 이전과 이후의 연구동향 분석을 위해 LDA를 활용하였다. 박영욱, 정규엽(2021)은 상관토픽모델을 활용하여 유튜브 크리에이터 소비자의 온라인 구전 특성을 연구하였고, 최현홍, 심동녘(2020)은 ICT 유관기관에서 발간한 보고서 자료에 대하여 구조적 토픽모형 분석을 적용하여 ICT융합 이슈를 분석하였다. 이현상 등(2021)은 DETM과 STM을 두 가지 모델을 활용하여 섬유소재 분야에 특허 기술 동향을 분석하였다.

논문과 같은 연구정보 데이터에는 일반적으로 저자, 출판사 및 출판연도와 같은 메타정보가 수반되며, 이러한 메타데이터를 기반으로 변화되는 주제 분포를 파악하기 위한 DMR(Dirichlet-Multinomial Regression) 토픽 모델링 기법이 제안되었다(Mimno & McCallum, 2012). 박영욱, 정규엽(2021); 원종호 등(2021)의 연구에서는 학문 분야의 동향과 온라인 게시물을 활용한 인식을 분석하기 위해 시간 정보(연도, 월)인 메타데이터를 포함하였다. 이처럼 DMR 모델은 저자, 국가, 연구분야, 출판지역 등 메타데이터별 주제 분포를 추정할 수 있다는 이점을 가지고 널리 활용되고 있지만, 모델이 가질 수 있는 메타데이터 변수가 범주형(Categorical Variable)으로 제한된다는 점에서 한계를 가지고 있다. 피인용수나 출판연도와 같은 데이터는 비율과 구간을 나타낼 수 있는 양적 자료로써 DMR 모델에서는 수치형 변수(Numeric Variable)를 기준으로 주제 분포를 산출하는 것에 제약이 있다. 이러한 DMR의 한계를 극복하고 개선하기 위해 Lee and Song(2020)은 g-DMR(Generalized DMR) 모델을 제안하였다. g-DMR 모델은 임의의 다항식 차수를 갖는 다양한 연속 변수를 처리할 수 있으므로 보다 동적인 주제의 경향을 나타낼 수 있다.

본 연구에서는 디지털 전환과 관련된 연구정보 데이터로 초록을 포함하여 출판연도와 연구분야의 메타정보를 입력변수로 활용하기 위해 연속형 변수 투입에 제약이 있는 DMR의 개선된 모델인 g-DMR을 이용하여 토픽 분석을 수행하였다.

2.2.1. DMR topic model

LDA 모형은 문헌이 여러 주제들의 조합으로 이루어져 있고, 각각의 주제는 여러 개의 단어가 포함된다고 가정하며, 각 문헌별 주제 분포만을 살펴볼 수 있다. 그러나 많은 연구자들이 문헌이 가지는 부가적인 정보를 활용하여 메타데이터별 주제 분포를 추정하고자 함에 따라서 Mimno and McCallum(2012)에 의해 DMR 모형이 제안되었다.

DMR 토픽 모형은 메타데이터에 민감한 주제를 생성하기 위해 각 주제 t에 대한 벡터 $\boldsymbol{\lambda}_t$가 추가되고, 문헌의 메타데이터에 따라 α값이 달라질 수 있다는 점에서 하이퍼 파라미터 α가 모든 문헌에 대해서 같고 대칭적이었던 LDA와 다르게 나타난다. DMR은 다음과 같은 문서 생성 프로세스에 따라 확장될 수 있다.

(1) 각 토픽 t에 대해

(A) $\boldsymbol{\lambda}_t \sim N(0, \sigma^2 I)$

(B) $\phi_t \sim Dir(\beta)$

(2) 각 문서 t에 대해

(A) 각 토픽 t에 대해 $\alpha_{d,t} = \exp(\boldsymbol{\chi}_d \boldsymbol{\lambda}_t)$

(B) $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}_d)$

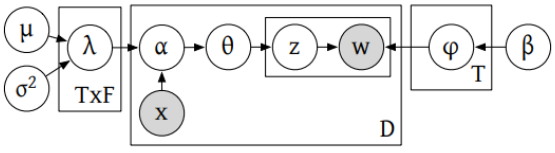
(C) 각 단어 i에 대해

i. $z_i \sim Mult(\boldsymbol{\theta}_d)$

ii. $w_i \sim Mult(\boldsymbol{\phi}_{z_i})$

여기서 σ^2은 λ_t에 대한 사전 분산을 나타내고, ϕ_t는 토픽 t에 대한 주제별 단어 분포 벡터를 의미하며, 이는 매개변수 β를 따르는 대칭적인 디리클레 분포에서 그 값을 뽑는 것을 의미한다. 또한 문서 d의 메타데

이터를 의미하는 χ_d는 메타데이터가 총 F개가 있다고 가정하면, 이는 F차원의 벡터로 나타낼 수 있다. 그리고 α_d는 λ_t 및 χ_d로부터 도출된 주제 분포에 대한 $\alpha_{d,t}$의 집합을 나타내는 파라미터이다. 핵심적으로 주제 개수를 T, 메타데이터 개수를 F라고 한다면, 이에 따라 T x F 차원의 행렬로 표현되는 λ가 앞선 식의 변수가 되고 이 값에 따라 변화하는 확률을 최대로 하는 λ를 추정하는 것이다. 이러한 DMR 모형의 구조는 <그림 1>에서 확인할 수 있다.



〈그림 1〉 DMR topic model
(source: Mimno and McCallum, 2012)

2.2.2. Generalized DMR (g-DMR)

DMR 토픽 모델에서 $\alpha_{d,t}$가 $\exp(\chi_d \cdot \lambda_t)$로 정의되므로 메타데이터 활용에 대한 단조성의 한계를 극복하기 위해 Lee and Song(2020)은 보다 일반화된 주제분포함수(Topic Distribution Function, TDF)로 $\alpha_{d,t}$를 대체할 수 있는 g-DMR 모형을 제안하였다.

$$\boldsymbol{\alpha}_{d,t} = \exp(f_t(\boldsymbol{\chi}_d)) + \varepsilon \quad (1)$$

f_t함수는 문서 d의 메타데이터 벡터인 χ_d를 전달받아 그 문서의 주제 t에 해당하는 가중치를 반환하게 된다. 또한 계산 정밀도의 제약으로 인해 종종 지수함수의 값이 0이 되는 것을 보정하기 위해 평활 매개변수로 ε을 추가하였다.

최대한 다양한 형태를 나타낼 수 있는 범용적인 함수가 입력데이터에 대한 적합한 결과 값을 산출할 수 있으므로, 이를 위해 다항식 근사(Polynomial Approximation)를 적용하였다. 다항식 근사에는 단순 다항식 근사(Simple Polynomial Approximation), 푸리에 근사(Fourier Approximation), 르장드르 다항식 근사(Legendre Polynomial Approximation) 등 연속함수를 근사하기 위해 활용되는 다양한 방식이 있지만, 해당 모델에서는 최적화를 위해 르장드르 다항식을 가정하고 있다. 즉, 정확한 형태를 예측할 수 없는 주제분포함수 $f_t(\chi_d)$를 르장드르 다항식이라 가정하여 작은 오차로 원본 형태를 최대한 근사할 수 있다는 것이다. 분석 목적에 따라 연속형 메타데이터에 대해서는 주제분포함수를 다음의 식 (2), (3)과 같이 설정할 수 있다.

$$f_t(\chi) = \sum_n \lambda_{t,n} L_n(\chi_d) \quad (2)$$

$$f_t(\boldsymbol{\chi}) = \sum_{m,n} \lambda_{t,m,n} L_m(\chi_1) L_n(\chi_2) \quad (3)$$

르장드르 다항식의 근사를 사용할 때 무엇보다 차수를 설정하는 것이 무엇보다 중요하며, 차수를 높일수록 실제의 원본 형태와 유사한 함수가 되겠지만 계산량이 너무 증가하지 않도록 적당하게 큰 수로 설정해야 한다. 또한 식3의 2차원 르장드르 다항식을 사용하면 메타데이터가 2개 이상인 경우에도 대응하기에 용이하다는 이점이 있다. 이 경우에도 마찬가지로 변수 χ_1과 χ_2를 근사하는 차수만 결정해준다면, 2개 이상의 변수로도 쉽게 확장할 수 있다.

LDA를 이용하여 각 문서별 토픽 분포를 구하고, 특정 메타데이터를 가진 문서들을 취합하여 메타데이터별 토픽 분포를 계산할 수 있다. 그러나 이는 확률 모델 외부에서 추가적인 분석을 수행하는 것이고, 연구모델에서는 메타데이터에 따라 문서 토픽 분포의 사전 분포가 영향을 받을 수 있다는 가정을 전제하고 있으므로, 각각의 모델에서 도출된 토픽에 큰 차이가 있다. 즉, 연구분야라는 범주별 토픽 트렌드를 세밀히 관측하기 위해서 g-DMR 모형이 더욱 적합하다고 할 수 있다.

2.2.3. Perplexity 및 Coherence score

토픽 모델링은 문서 집합 내 단어들이 동시에 출현하는 것을 바탕으로 같은 의미 부류에 속하는 단어들을 주제별로 그룹화해준다. 그러나 토픽 모델링과 같은 비지도학습 기반의 분석 기법은 자동으로 처리된(labeling) 결과의 신뢰성을 높이기 위해 모델링 결과에 대한 성능평가가 필요하다. 군집화 기법을 평가하는 방법은 크게 내재적인(Intrinsic) 방식에 해당하는 복잡도 점수(Perplexity score)와 이를 개선한 일관성 점수(Coherence score)가 있다.

복잡도 점수는 특정 확률모델이 실제로 관측되는 값을 얼마나 잘 예측하는지를 평가할 때 사용하며, 해당 값이 작을수록 토픽 모델이 실제 문헌 결과를 잘 반영하고, 학습이 잘 되었다고 평가할 수 있다. 또한 토픽 모델에서 적절한 주제 개수를 선정하기 어려울 때, 여러 주제 개수로 학습을 반복함으로써 최소의 복잡도 점수를 갖는 주재 개수를 토픽으로 선정할 수 있다. Chang et al.(2009)의 연구에 따르면 낮은 복잡도 점수는 학습이 잘 된 것을 의미하지만 이것이 절대적으로 해석에 적절한 결과를 나타내는 것은 아니라고 주장하였다. 따라서 사람이 해석하기에 적합한지를 파악하기 위한 척도로써 주제 일관성(Topic Coherence)이 Newman et al.(2010)에 의해 제안되었다. 본 연구에서는 토픽 분석의 결과로 뽑힌 주제들에서 상위 단어 간에 얼마나 의미상 유사한지를 파악하기 위해 단어 간 코사인 유사도(Cosine Similarity)와 정규화 점별 상호정보량(Normalized Pointwise Mutual Information, NPMI) 등을 이용하여 일관성 점수를 산출하고 최적 토픽 수를 결정하였다.

3. 연구방법

3.1. 데이터 수집

본 연구의 분석을 위해 Web of Science(WoS)의 데이터베이스에서 키워드로 “Digital Transformation”을 필수적으로 포함하고, “Data”, “Atificial Intelligence(AI)”, 및 “Network”를 선택적으로 포함하는 검색 조건을 설정하였다. 해당 키워드에 대한 SCI-EXPANDED(Science Citation Index Expanded), SSCI(Social Sciences Citation Index), A&HCI(Arts & Humanities Citation Index) 색인의 오픈 엑세스를 포함하는 논문으로 수집 대상을 한정하였다. 자료의 수집 기간은 2010년부터 1월부터 2022년 3월까지 총 12년 3월의 기간에 해당하며, 토픽 분석을 위해 초록뿐만 아니라 제목, 키워드, 연구 분야, 피인용수, 출판연도와 같은 메타정보를 함께 추출하였다. 총 1,090건의 문헌 리스트에서 초록이 누락된

〈표 1〉 데이터 수집 요약

분류	데이터 필터링 항목
데이터베이스	https://www.webofscience.com (WoS database)
키워드	"digital transformation" (주제) and "data" OR "artificial intelligence" OR "AI" OR "network" (주제)
논문 색인	A&HCI(Arts & Humanities Citation Index), SCI-EXPANDED(Science Citation Index Expanded), SSCI(Social Sciences Citation Index)
출판연도	2010/1/1 - 2022/3/31
주요 메타정보	Abstract, Times Cited, Publication Year, Research Areas, Author Keywords, Article Title
문서 수	#1,072

문헌들을 제거하고, 1,072건의 연구정보 데이터를 최종적으로 수집하였다. 수집된 데이터의 요약 정보는 <표 1>과 같다.

3.2. 데이터 전처리

텍스트와 같은 비정형 데이터를 분석하기 위해서는 다양한 전처리 작업이 선행되어야 하며, 본 연구에서는 g-DMR을 적용하기 위해 논문들의 초록을 입력데이터로, 연구분야와 출판연도를 메타정보로 활용하였다. 영문 초록을 대상으로 텍스트 전처리를 수행하여 토픽 모델의 입력 말뭉치(Corpus)로 사용하였다. 텍스트 전처리 과정은 코퍼스를 사용하는 텍스트 분석 연구에서 결과의 신뢰도에 많은 영향을 미치는 중요한 부분으로, 일반적으로 토큰화(Tokenization)와 어간(Stemming) 및 표제어(Lemmatization) 추출, 불용어(Stopword) 제거 등의 과정을 거친다(정유경, 2020). 기본적인 전처리에 앞서 영문 이외의 문자를 삭제하고, 숫자 및 특수 문자를 제거하였다. 이어서 영문 초록의 전처리를 위해 케라스(Keras)의 text_to_word_sequence로 토큰화하였고, NLTK의 WordNetLemmatizer와 pos_tag를 이용하여 명사에 해당하는 표제어만을 추출하였다. 전처리된 입력데이터의 최종 문서 수와 총 어휘 수는 각각 1,072, 55,453개로 나타났다. 또한 메타정보인 연구분야의 초기 범주가 80가지로 매우 많아서 분석에 적합하지 않았기 때문에 이를 <표 2>와 같이 7가지의 범주로 재구성하였다.

구체적으로, 토픽 모델링을 위해 두 가지 절차에 따라 범용어 처리를 수행하였다. 첫 번째, 범용어 처리를 위해서 (“purpose”, “paper”, “article”, “background”, “chapter”, “case”, “result”, “review”, “model”, “use”, “study”, “research”, “framework”, “method”, “methodology”, “approach”, “literature”, “field”, “researcher”)와 같이 논문 작성에 빈번하게 사용되는 어휘를 중심으로 1차적인 불용어 제거를 수행하였고, 이어서 모델링 과정의 rm_top (최상위 빈도 단어 제거)의 파라미터 입력값을 2로 설정함으로써 제목에 포함되는 어휘인 “transformation”을 포함하여, 최상위 빈도 단어 2개(“transformation”, “technology”)를 제거하였다. 코퍼스로 활용된 문서 수가 1,072건으로 비교적 많지 않고, 결과 해석을 통해 토픽 레이블링에 충분한 도메인 지식을 반영하고자 범용어 처리를 보수적으로 진행하였다.

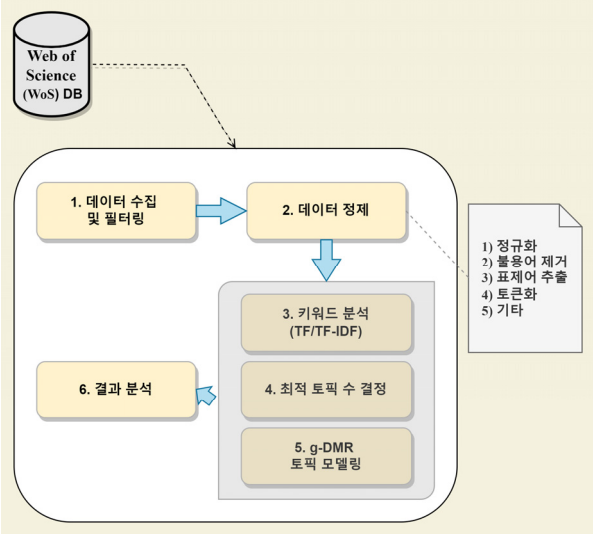
3.3. 분석 프레임워크

본 연구에서는 수집된 텍스트 자료를 기반으로 텍스트마이닝 분야에서 가장 빈번하게 적용되는 키워드 빈도수 기반의 빈출 순위 분석(TF)과 단어 빈도-역 문서 빈도 분석(TF-IDF)을 일차적으로 수행한다. 다음으로 최적 토픽 수 결정을 위한 일관성 점수(Coherence

〈표 2〉 재구성된 연구분야

No.	연구분야	문서 수
1	Business & Economics	265
2	Engineering/Applied Science	249
3	ICT	201
4	Health Care	124
5	Natural Sciences	99
6	Humanities/Others	70
7	Social Sciences	64

score)를 산출하고, 메타정보를 이용하여 키워드 간 관계를 세부적으로 탐색할 수 있고, 맥락적 의미 파악이 가능한 g-DMR 모형을 통해 토픽 분석을 수행한다. 마지막으로 각 토픽별 레이블링을 통한 연구동향을 분석한다. 구체적인 데이터 분석 프레임워크는 다음의 <그림 2>와 같다.



〈그림 2〉 데이터 분석 프레임워크

4. 분석 결과

4.1. 빈도 기반 키워드 분석

본 연구는 토픽 분석을 수행하기 전 문서 집합 내 단어의 출현 빈도 순위에 대한 기초적인 분석 결과를 검토함으로써 분석 대상 문헌의 전반적인 키워드 결과와 그 경향성에 대해 파악해보고자 하였다. 키워드 분석을 위해 단순하게 문서 전체에 나타난 단어들의 출현빈도를 의미하는 TF(Term Frequency)와 하나의 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치로써 각각의 문서에서 자주 출현하는 단어에 높은 가중치를 주고, 모든 문서에서 자주 등장하는 단어에 대해서는 역가중치(페널티)를 부여하는 방식인 TF-IDF(Term Frequency-Inverse Document Frequency)을 사용하였다. 즉, TF-IDF는 문서의 주제를 가늠할 수 있는 핵심 단어가 높은 점수를 나타내는 경향이 크다고 할 수 있다.

우선, <표 3>에서 분석 대상은 수집한 연구자료의 초록이고, 전체 기간(2010/01~2022/03)에 대한 TF 및 TF-IDF 순위를 함께 제시하였다. 분석 결과에 따르면, 디지털 전환의 핵심 기술이라고 할 수 있는 데이터, 네트워크. 인공지능과 관련된 연구의 특성에 따라서 다소 기술과 관련된 범용적인 키워드인 'transformation', 'technology', 'industry', 'system', 'analysis', 및 'network' 등의 단어 빈도가 높은 것으로 관찰되었다. 또한, 'health'와 'care'라는 키워드를 통해 헬스케어와 관련된 디지털 전환 이슈가 많았음을 추측해볼 수 있다. TF 및 TF-IDF 기준 순위가 다소 다르게 나타났는데, 앞선 헬스케어 관련 키워드의 경우, TF-IDF에서 TF보다 높은 순위로 나타났기에 특정 문서의 주제를 추정할 수 있는 핵심 키워드라고 할 수 있다.

다음으로, <표 4>에서는 시간적 특성이 반영될 수 있도록 연도별(5년 주기) TF-IDF 순위를 제시하였다. 분석결과를 살펴보면, 각각의 연도별 및 주기별 키워드 순위에 있어서 상이한 결과가 나타났다. 첫 번째 기간에는 디지털 전환이라는 추상적인 개념이 막 도입되어, 'business', 'value', 'creation', 'knowledge', 'managerial'의 용어와 함께 'network', 'web', 'scm' 등의 키워드가 상위권에 등장하였다. 이러한 결과를 토대로 온라인 서비스와 물류, 공정 등의 분야에서 디지털 혁신에 관한 주제로 연구가 수행되었음을 알 수 있다. 두 번째 기간의 2015년도에는 'health care', 및 'security'의 키워드를 통해 헬스케어와 보안과 관련한 디지털 전환 이슈가 주요 키워드로, 그리고 2016년도에 'performance', 'technology'의 키워드와 함께 'ceo', 'cio', 'organization'

〈표 3〉 TF 및 TF-IDF 분석 결과 (상위 30위 키워드)

NO.	전체 TF	TF_counts	전체 TF-IDF	TF-IDF_scores
1	transformation	1571	technology	31.82207
2	technology	1407	transformation	28.495
3	system	862	industry	25.90691
4	industry	860	system	23.91266
5	process	699	business	23.06615
6	business	695	service	20.19919
7	service	608	process	20.07678
8	analysis	557	health	18.99701
9	information	555	innovation	18.07729
10	health	537	information	17.01494
11	management	522	management	16.50471
12	innovation	478	analysis	16.19501
13	value	476	company	16.00121
14	development	469	value	15.6836
15	company	386	development	14.87451
16	network	379	organization	14.35345
17	organization	368	capability	14.30815
18	performance	343	network	13.78866
19	capability	321	performance	12.92387
20	strategy	311	firm	12.64338
21	application	307	strategy	11.94041
22	role	298	application	11.67534
23	level	292	care	11.42972
24	design	290	change	11.30455
25	change	288	chain	11.19959
26	firm	281	level	11.12976
27	practice	280	practice	10.78793
28	care	270	role	10.71163
29	implementation	267	sector	10.61446
30	time	266	knowledge	10.61165

및 'institution' 등의 키워드가 상위권에 올랐다. 2016년에는 기술혁신 분야 세계 최고 권위상인 '2016 CIO 100 Awards'의 개최 결과로 관련 키워드가 상위권에 다수 등장하였다. 마지막 기간에는 'health', 'education', 'platform', 'supply chain' 'analysis' 등의 단어가 상위권에 출현하며, COVID-19 이후 원격 진료, 온라인 교육 등의 수요가 증가함에 따라 비대면 플랫폼 기반의 디지털 전환 이슈가 주요 키워드로 등장하고 있음을 확인할 수 있다. 또한 'manufacturing'의 키워드가 등장한 것은 기존 제조 분야에서 다양한 형태로 디지털 전환이 추진되어왔음에도 불구하고, 2020년에 들어서, 다시 한번 제조 디지털화에 대한 이슈가 등장하고 있음을 짐작할 수 있다.

전체 기간과 주기별 결과를 비교하며, 전반적으로

〈표 4〉 연도별(5년 주기) TF-IDF 순위

NO.	TF-IDF(period 1)		TF-IDF(period 2)		TF-IDF(period 3)	
	2010	2011	2015	2016	2020	2021
1	business	resource	expertise	organizati-on	industry	technology
2	network	web	care	information	technology	transformation
3	business network	transformation	health	ceo	transformation	industry
4	value	extent web	technology	frequency	business	system
5	production	scm	health care	industry	system	business
6	print production	extent	privacy	business	service	process
7	print	banking	security	force	health	innovation
8	simulation	web scm	issue	adoption	information	service
9	design	banking project	system	cio	process	health
10	value creation	project	change	service	management	analysis
11	creation	resource fit	holocaust	ceo cio	development	information
12	network value	capability	study	performance	company	value
13	opportunity	fit	infrastructure	institution	value	management
14	process	implication	work	transformation	strategy	company
15	demand	governance	question	role	analysis	firm
16	technology	governance mechanism	service	internet	network	supply
17	product	influence	security privacy	water	capability	chain
18	service	activity	way	point	level	development
19	product service	investment	demand	society	firm	organization
20	activity	intensity	information	technology	performance	supply chain
21	value business	intensity managerial	communication	opportunity	innovation	capability
22	industry	knowledge	care system	implication	education	performance
23	firm	managerial	treatment	safety assessment	challenge	level
24	transformation	managerial knowledge	cost	assessment	platform	network
25	level	mechanism	transformation	safety	application	digitalization
26	process print	mechanism extent	offer	chemical	organization	manufacturing
27	process describe	scm association	study holocaust	practice	practice	design
28	level report	information	author	network	change	sector
29	work	effort	methodology	partnership	manufacturing	knowledge
30	opportunity exploitation	information intensity	foundation	opinion	knowledge	model

디지털 전환과 관련된 연구 키워드의 대략적인 트렌드를 파악할 수 있다. 예를 들어, 헬스케어에 관한 키워드는 전체 기간에서 TF-IDF 순위가 8위로 상위권에 포진하였고, 2015년을 기점으로 2020년과 2021년까지 3위→7위→9위로 다소 순위가 하락했지만, 꾸준하게 관심받는 연구분야로 나타났다. 또한 프로세스 및

공급사슬관리와 같이 제조공정에 관한 키워드 전체 기간에서 각각 7위와 25위로 상대적으로 높은 중요도를 보였고, 프로세스의 경우 14위(period 1)에서 6위(period 3)로 TF-IDF 순위가 다소 상승하였고, 공급사슬관리는 3위(period 1)에서 16위(period 3)로 TF-IDF 순위가 다소 하락한 것을 확인할 수 있다.

4.2. 토픽 분석

4.1절에서 다룬 빈도 기반 키워드 분석은 그 자체만으로 유용한 인사이트를 제공하지만, 문서 자료에 내재한 세부적인 맥락을 파악하기 어렵다는 한계가 있다. 따라서 이러한 한계점을 보완하고, 메타데이터를 이용한 주제 분포를 추정하기 위해서 g-DMR 모형을 적용하여 토픽 모델링을 수행하였다. 토픽 분석을 위해 프로그래밍 언어에 Python 3.8과 통합 개발 환경으로 Jupyter Notebook을 사용하였다. 또한 토픽 모델링을 위한 라이브러리로 tomotopy 12.0, 그리고 tomotopy로 추출한 토픽의 시각화를 위해 pyLDAvis, 영문 처리를 위해 nltk 등을 활용하였다.

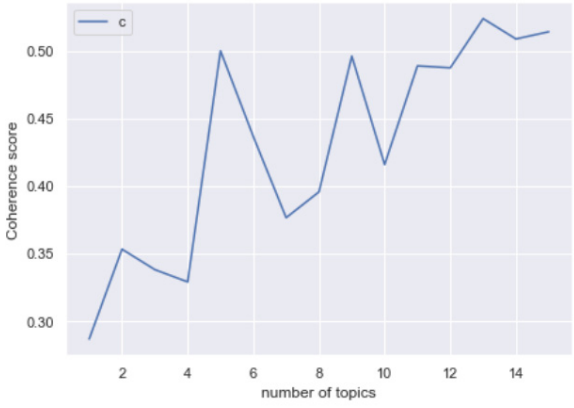
4.2.1. 최적 토픽 수

일반적으로 토픽의 개수가 증가하면 더욱 세분화된 주제의 분석이 가능하지만, 비중이 매우 작아서 해석이 용이하지 않은 토픽이 도출될 수 있다. 반면에 토픽의 개수가 감소하면 결과의 해석이 용이해지고 각 토픽이 일정의 비중을 갖지만, 유의미한 토픽이 도출되지 않을 수도 있다.

본 연구에서는 최적의 토픽 수를 선정하기 위해 토픽의 수를 1부터 15지 증가시키면서 일관성 점수 중 c_v 값을 측정하였고, 구체적인 결과는 다음의 <표 5>, <그림 3>과 같다.

〈표 5〉 토픽 일관성 점수

토픽 수	Coherence score
1	0.286
2	0.353
3	0.338
4	0.329
5	0.5
6	0.436
7	0.376
8	0.395
9	0.496
10	0.416
11	0.489
12	0.487
13	**0.524**
14	0.509
15	0.514



〈그림 3〉 최적 토픽 수 (topic_n=1~15)

토픽 일관성 점수의 산출 결과는 토픽이 13개일 때 0.524으로 가장 높게 나타났고, 이에 따라 g-DMR 토픽 모형 분석을 위해 토픽 수를 13개로 설정하였다. 토픽 수가 적을 때도 일관성 점수가 일부 높게 나타나는 경우가 있음에도 모델의 최적화 과정을 거쳐 산출된 <표 8>의 평균 일관성 점수를 토픽 개수가 적을 때와 비교하여 토픽이 13개일 때, 가장 높은 평균 토픽

일관성 점수를 갖는 것으로 나타났다. 개별 토픽 단위로 볼 때, 일관성 점수가 높지만 토픽 개수가 적어서 도출되지 않는 유의미한 토픽들이 발생하여 최종적으로 토픽을 13개로 설정하였다.

4.2.2. g-DMR 모델링 결과

토픽 분석의 성능을 높이기 위해서 적절한 파라미터를 선택하는 것이 중요하며, g-DMR 모형에 투입한 파라미터 값에 대한 정보는 <표 6>과 같다. 일반적으로 β는 0.01과 0.1 사이에서 결정되고, Mimno and McCallum(2012)의 연구에 따르면 DMR의 두 모델에 대해 β를 0.01로 설정하고 있다. σ 및 σ_0에 해당하는 파라미터는 메타데이터에 의한 주제 분포의 변화에 영향을 미치기 때문에 이 값이 작으면 작을수록 메타데이터에 의한 전체 분포가 더 균등해지게 되고, 이는 토픽 간에 구별을 어렵게 만드는 요인이 될 수 있다. Lee and Song(2020)의 연구에서는 3 이상의 σ 또는 σ_0 값은 메타데이터에 의한 분포 차이를 최대화하고, 모델의 깁스 샘플링(Gibbs sampling) 과정을 불안정하게 만들어 실험에 실패할 수 있기에 σ와 σ_0 값을 3 이하로 설정하였다. 또한 σ는 고차 항의 계수 분포를 결정하므로 전체 모델의 수렴에 도움이 되도록 σ_0보다 작은 값으로 설정되어야 한다. 본 연구에서는 여러 차례의 실험을 통해 적절한 σ 및 σ_0 값을 0.5와 2로 설정하였다. 그 외에도 토픽 수, 단어의 최소 문헌 빈도, 제거할 최상위 빈도 단어의 개수, 토픽 분포 함수(TDF)로 쓰일 르장드르 다항식의 차수, 단어 가중치 등의 입력 파라미터가 있다. 본 연구에서 주제의 개수는 앞선 실험을 통해 13개의 토픽 수를 선정하였으며, 단어가 출현한 최소 문서 빈도를 2개로 설정하였고, 최상위 빈도 단어 2개('transformation', 'technology')를 제거하였다. 이어서 연속형 메타데이터를 처리하기 위한 르장드르 다항식의 차수와 단어의 가중치 기법을 각각 5, PMI로 설정하였다. 실험은 초기 Burn-in 단계에서 200회 및 모델 훈련을 위해 총 2,000회를 반복 수행하였다.

〈표 6〉 입력된 파라미터 값

Parameters	Values
Term Weight	PMI
min_df (Minimum document frequency of words)	2
rm_top (The number of top words to be removed)	2
k (The number of topics)	13
degrees (Topic Distribution Function)	5
alpha (Exponential of mean of normal distribution for 'lambdas')	0.1
beta (Dirichlet distribution for topic – word)	0.01
sigma, sigma0 (Non-constant, constant terms of 'lambdas')	0.5, 2.5
Iterations, Burn-in steps	2000, 200

g-DMR의 주제 추정에 반영한 문서의 메타정보로는 크게 두 가지로 구분할 수 있는데, 첫 번째 메타데이터는 연구분야에 대한 정보이다. 각 연구분야별 범주와 방향성에 적합한 주제가 반영되는 것은 당연함으로 도메인 정보의 투입으로 결과 해석에 큰 의미를 제공하지 못하지만, 모형 추정에 용이하도록 지침을 준다는 측면에서 의미가 있다. 두 번째 메타데이터는 문서(논문)의 출판연도이다. 출판연도 정보는 g-DMR 모형 추정에 지침을 줄 뿐만 아니라, 토픽의 해석 측면에서도 상당한 의미를 가질 수 있다. 특히 시간의 흐름에 따른 연구 트랜드 변화를 분석하여 본 연구에서 중점적으로 살펴보고자 하는 디지털 전환에서의 데이터, 네트워크 그리고 인공지능 중심의 연구 경향성 탐색에 유용하게 활용할 수 있다. 일정 수의 연구자료가 존재하고, 디지털 전환이라는 용어가 본격적으로 사용되기 시작한 시점(2010년)을 기준으로 출판연도 정보를 연속 변수로 모형에 활용하였다.

본 절에서는 g-DMR 토픽 분석을 통해 13가지의 토픽과 각각에 대해 토픽을 구성하는 상위 키워드(단어)를 20위까지 도출하였고 다음의 <표 7>과 같다. 해당 키워드는 토픽의 구성 확률이 높은, 즉, 토픽 생성에 강한 영향을 미치는 핵심 단어를 의미한다. 다음으로 앞서 도출된 토픽별 키워드를 통해 각각의 주제를 요약하여 레이블링(Labeling)하였고, 상위 키워드 5개와 그 비중을 <표 8>에 요약하였다. 또한 최적화된 모델의 토픽별 일관성 점수를 산출하여 각각의 토픽이 얼마나 의미론적으로 높은 일관성을 갖는지 파악하였다. 토픽 레이블은 각각의 문서별로 토픽 분포를 확인

〈표 7〉 토픽별 상위 20개 키워드 (K=13)

Key words	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
1	chain	energy	construction	industry	food	laboratory	city	service	country	health	employee	water	education
2	supply	system	information	process	agriculture	network	image	business	policy	care	work	welfare	factor
3	capability	device	bim	management	code	control	purchase	customer	readiness	healthcare	leadership	visit	student
4	innovation	iot	pharmacy	organization	waste	pathology	music	value	economy	hospital	cyber security	community	intention
5	performance	production	function	knowledge	product	source	consumer	firm	capital	patient	job	video	university
6	enterprise	security	asset	information	retailer	land	mobility	marketing	growth	medicine	resilience	expertise	adoption
7	firm	maintenance	mining	digitalization	store	port	program	platform	company	disease	worker	child	school
8	economy	time	twin	company	definition	property	classification	ecosystem	enterprise	system	occupation	sport	acceptance
9	relationship	manufacturing	airport	decision	radiology	edge	airline	creation	level	record	battery	custom	competence
10	effect	privacy	uncertainty	development	platform	obesity	death	market	effect	surgery	vulnerability	journalism	literacy
11	governance	application	trust	analysis	face	slide	velocity	product	performance	implementation	skill	treatment	influence
12	resilience	network	design	challenge	conflict	reconfiguration	audience	design	region	population	cio	stress	crisis
13	agility	sensor	processing	project	recognition	maritime	teacher	sale	indicator	treatment	substitution	procurement	perception
14	orientation	cost	project	tool	crop	news	television	consumer	intensity	covid	hbm	scene	competency
15	maturity	technique	procurement	business	container	radio	service	provider	labor	outcome	executive	competence	survey
16	mechanism	machine	registration	change	farm	state	medium	bank	productivity	quality	board	modality	skill
17	integration	resource	measurement	sector	ad	infrastructure	bridge	audit	index	safety	wage	play	hotel
18	role	chemical	system	intelligence	resale	ontology	convenience	implication	sustainability	democracy	organization	die	attitude
19	strategy	process	turbulence	opportunity	system	depot	channel	client	income	drug	change	justice	evaluation
20	coordination	layer	topic	strategy	grocery	party	classroom	innovation	efficiency	breach	probability	movement	face
:	:	:	:	:	:	:	:	:	:	:	:	:	:
30	activity	engineering	cable	issue	climate	simulation	inspection	demand	metal	delivery	selection	disaster	publisher

〈표 8〉 토픽 분석 요약

주제 #	일관성 점수	상위 키워드 및 비중					주제 요약
1	0.686	chain (0.073)	supply (0.073)	capability (0.063)	innovation (0.043)	performance (0.037)	공급망 디지털화 성과
2	0.651	energy (0.027)	system (0.022)	device (0.02)	iot (0.016)	production (0.015)	제조 생산 시스템
3	0.681	construction (0.071)	information (0.043)	bim (0.03)	pharmacy (0.028)	function (0.028)	정보 기술
4	0.591	industry (0.012)	process (0.007)	management (0.007)	organization (0.007)	knowledge (0.007)	비즈니스 인텔리전스
5	0.704	food (0.11)	agriculture (0.042)	code (0.035)	waste (0.026)	product (0.026)	감지 (모니터링) 시스템
6	0.690	laboratory (0.034)	network (0.032)	control (0.03)	pathology (0.029)	source (0.024)	네트워크 기술 (클라우드)
7	0.695	city (0.093)	image (0.048)	purchase (0.024)	music (0.022)	consumer (0.019)	데이터 분석 기술
8	0.651	service (0.055)	business (0.046)	customer (0.04)	value (0.037)	firm (0.028)	서비스 가치 창출
9	0.709	country (0.026)	policy (0.024)	readiness (0.02)	economy (0.02)	capital (0.02)	지속 가능 성장 (효율성)
10	0.744	health (0.11)	care (0.06)	healthcare (0.041)	hospital (0.026)	patient (0.026)	원격 진료
11	0.736	employee (0.068)	work (0.048)	leadership (0.038)	cybersecurity (0.036)	job (0.036)	자동화 기술 (노동력 대체)
12	0.756	water (0.053)	welfare (0.041)	visit (0.038)	community (0.035)	video (0.034)	복지 시스템
13	0.734	education (0.045)	factor (0.038)	student (0.036)	intention (0.027)	university (0.026)	디지털 교육
Ave. = 0.694		–					

하고, 높은 토픽 비중을 차지하는 토픽별 추출된 키워드의 의미적 연관성을 반영하여 연구자가 직접 선정하였다.

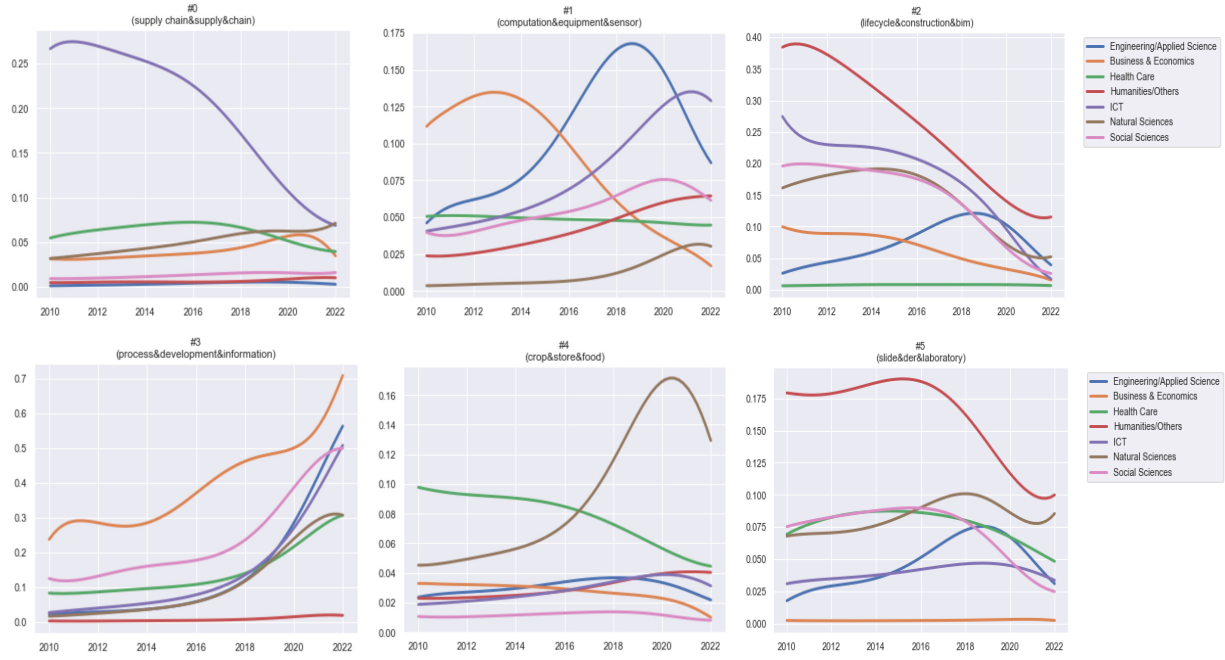
디지털 전환을 필수적으로 포함하며 데이터, 네트워크 그리고 인공지능 기술을 중심으로 연구된 논문의 초록을 분석하여 도출된 주요 주제는 '공급망 디지털화 성과', '제조 생산 시스템', '정보 기술', '비즈니스 인텔리전스', '감지(모니터링) 시스템', '네트워크 기술', '데이터 분석 기술', '서비스 가치 창출', '지속 가능 성장', '원격 진료', '자동화 기술', '복지 시스템', '디지털 교육'인 것으로 나타났다. 제조 생산 시스템, 정보 기술, 네트워크 기술 등과 같이 구체적 기술 분야를 나타내는 토픽들과 공급망 디지털화 성과, 지속 가능 성장, 원격 진료 및 디지털 교육 등과 같이 디지털화로 인한 비즈니스 생태계 구축과 이에 수반되는 정책, 그리고 활용 성과 관점에서의 다양한 토픽이 관찰되었다.

<표 8>를 살펴보면, 일관성 점수가 0.7 이상인 토픽들로는 5, 9, 10, 11, 12, 13에 해당하는데, 이는 해당 토픽들의 단어 그룹이 전체 문헌에서 70% 이상의 확

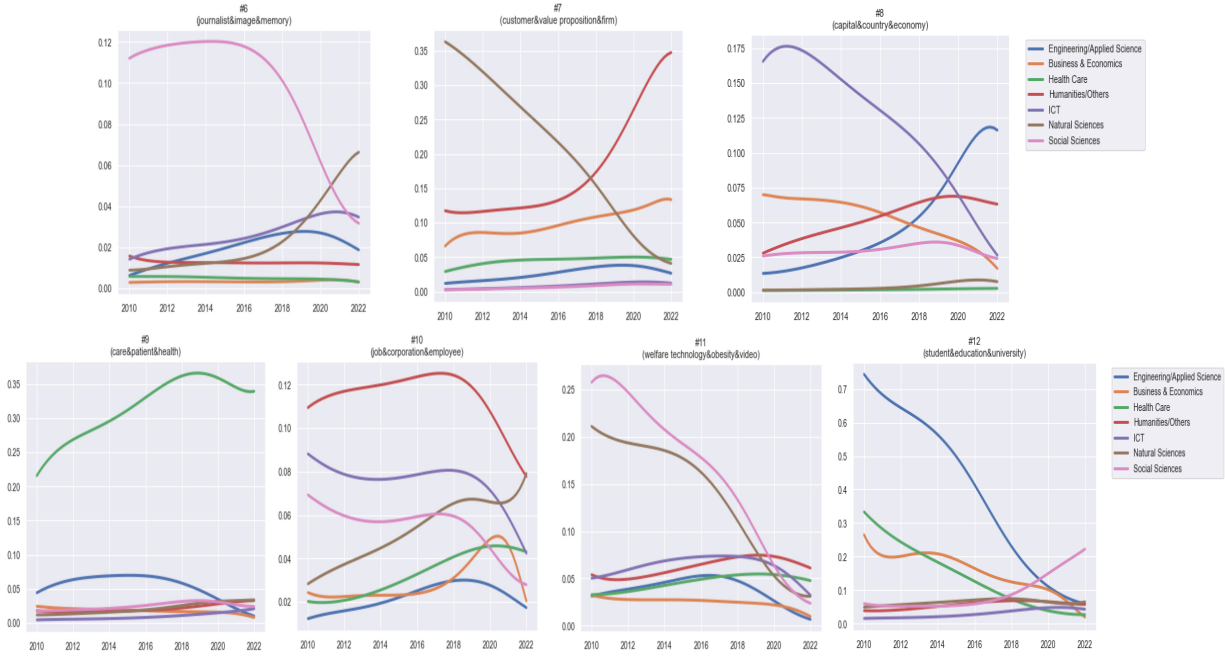
률로 동시에 출현하는 것을 의미한다. 보다 구체적으로 살펴보면, 토픽 5를 구성하는 키워드로 ‘food’의 비중이 가장 높게 나타났고, 그 외에 출현한 ‘agriculture’, ‘product’, ‘crop’, ‘farm’과 같은 키워드를 통해 농업 분야의 농작물을 추론할 수 있다. 여기에 ‘platform’, ‘face’, ‘recognition’, ‘system’ 등의 용어들과 함께 사용되어 스마트 팜 영역에서의 감지(모니터링) 시스템과 관련된 다수의 연구가 수행되었음을 추론할 수 있다. 다음으로 토픽 9를 구성하는 상위권 키워드는 전반적으로 비슷한 비중을 보였으며, 토픽을 구성하는 ‘company’, ‘readiness’, ‘economy’, ‘capital’ 등의 키워드로부터 기업에서 효율성과 생산성을 강조하는 지속 가능 성장에 관한 맥락을 파악할 수 있다. 이어서 토픽 10의 경우에는 토픽 내에 ‘health’, ‘care’, ‘healthcare’, ‘hospital’, ‘patient’, 및 ‘medicine’과 같은 키워드가 직접적으로 출현했을 뿐만 아니라, ‘system’, ‘treatment’, ‘covid’의 키워드를 통해 헬스케어 분야에서 원격 진료에 대한 주제가 활발하게 연구되고 있음을 확인할 수 있다. 토픽 11과 13을 구성하는 키워드로는 각각 ‘employee’, ‘work’, ‘leadership’, ‘cybersecurity’, ‘job’과 ‘education’, ‘factor’, ‘student’, ‘intention’, ‘university’와 같은 단어가 상위권에 포진되어 있는데, 이는 자동화 기술 등으로 인해 단순 노동력에 대한 수요가 줄어들고, 코로나19 이후 비약적으로 발전하고 있는 디지털(온라인) 교육에 관한 주제를 포괄적으로 다루고 있다. 마지막으로 토픽 12와 같이 ‘water’, ‘welfare’, ‘visit’, ‘community’, ‘video’ 등의 여러 범주의 구체적인 키워드를 포함하고 있어서 통합된 주제 도출이 어려운 경우에는 해당 토픽이 높은 비중을 차지하는 문서(코퍼스)들을 출력함으로써 관련 내용을 검토하여 복지 시스템이라는 주제를 도출할 수 있었다.

문서에서 가장 높은 비중을 차지하고 있는 토픽에 따라 전체 문서를 분류하여 토픽별 문서 수를 살펴보면, 비즈니스 인텔리전스(Business Intelligence)에 해당하는 문서가 627개, 그리고 제조 생산 시스템과 공급망 디지털화 성과에 해당하는 문서는 각각 77, 68개로 상위 3가지 토픽이 전체 문서의 약 72%를 차지하며 매우 높은 비중을 차지하는 것으로 나타났다. 이는 메타정보를 이용하는 DMR 모형의 특성에 따라 출판연도 및 연구분야의 편중 현상이 반영된 결과이다. 출판연도를 기준으로 보면 2019년에 발생한 코로나19 이후부터 최근까지 디지털 전환과 관련된 연구가 급격하게 증가하였고(2019년 기점으로 전체 문헌의 약 90% 이상), 이에 해당하는 연구분야 또한 Business & Economics 및 Engineering 분야(전체 문헌의 약 48% 이상)에서 더욱 두드러지게 나타났다.

다음의 <그림 4>와 <그림 5>은 메타정보인 연구분야별 시간(출판연도)의 흐름에 따라 토픽 발현 비중 변화를 시각화하였다. 이를 통해 연구분야별로 해당 토픽에서의 상대적 위치뿐만 아니라 각 토픽의 전반적인 경향성(추세) 역시 관찰할 수 있으며, 이에 대한 주요 요약표는 다음 <표 9>와 같다. 전술한 바와 같이 디지털 전환이라는 키워드를 필수적으로 포함하는 연구가 코로나19 이후에 가장 활발하게 진행되고 있기에 다수의 연구가 비교적 최근에 편중되어있고, 연구분야의 범주도 주로 Business & Economics 및 Engineering 분야에서 수행된 연구가 절반에 육박하기 때문에 특정 토픽의 비중이 매우 높게 나타났다. 비즈니스 인텔리전스와 관련된 토픽 4의 비중이 약 53%를 차지하며 매우 높은 비중을 나타냈고, 해당 토픽을 구성하는 연구분야 대부분이 증가 추세를 보이며 최근까지 활발하게 연구되고 있음을 알 수 있었다. 이어서 제조 생산 시스템과 서비스 가치 창출과 관련된 토픽 2와 토픽 8이 각각 9%와 7.1% 비중을 차지하였는데, 제조 생산 시스템과 관련한 토픽을 구성하는 연구분야 중에 전반적으로 기술과 관련된 Engineering과 ICT 분야에



〈그림 4〉 g-DMR 분석결과(topic_#1-6)



〈그림 5〉 g-DMR 분석결과(topic_#7-13)

서는 증가 추세를, 그리고 Business & Economics 분야에서는 감소 추세를 나타냈다. 서비스 가치 창출 관련 토픽을 구성하는 연구분야에서 특징적인 것은 Social Sciences 분야의 비중이 큰 폭으로 감소하였고, Natural Sciences 분야의 비중은 점진적으로 상승 추세를 보이는 점이다. 그 밖에도 낮은 비중(1.4%)으로 나타난 감

〈표 9〉 연구분야별 토픽 비중 추이

주제 # (비중)	주제어	연구분야별 토픽 비중						
		Business & Economics	Engineering /Applied Science	Health Care	Humanities/ Others	ICT	Natural Sciences	Social Sciences
1 (0.054)	공급망 디지털화 성과	증가	–	감소	–	감소	증가	–
2 (0.09)	제조 생산 시스템	감소	증가	감소	증가	증가	증가	증가
3 (0.021)	정보 기술	감소	증가	–	감소	감소	감소	감소
4 (0.53)	비즈니스 인텔리전스	증가	증가	중가	증가	증가	증가	증가
5 (0.014)	감지 (모니터링) 시스템	감소	감소	감소	증가	증가	증가	–
6 (0.015)	네트워크 기술 (클라우드)	–	증가	감소	감소	증가	증가	감소
7 (0.015)	데이터 분석 기술	–	증가	감소	–	증가	증가	감소
8 (0.071)	서비스 가치 창출	증가	증가	증가	증가	–	감소	–
9 (0.05)	지속 가능 성장 (효율성)	감소	증가	–	증가	감소	증가	–
10 (0.056)	원격 진료	감소	감소	증가	–	증가	증가	–
11 (0.024)	자동화 기술 (노동력 대체)	감소	증가	증가	감소	감소	증가	감소
12 (0.013)	복지 시스템	감소	감소	증가	증가	감소	감소	감소
13 (0.047)	디지털 교육	감소	감소	감소	증가	증가	–	증가

지(모니터링) 시스템과 관련된 토픽 5의 경우에는 스마트 팜과 같은 특정 산업에서의 연구가 주를 이루었기에 Natural Sciences 분야가 높은 비중을 차지했고, 비교적 최근부터 활발하게 연구되고 있음을 확인할 수 있다. 또한 비대면 진료(토픽 10, 5.6%) 관련 토픽은 다수의 연구가 Health Care 분야에서 수행되었고, 그 추세 또한 꾸준하게 증가하고 있음을 알 수 있다.

결과적으로 높은 비중을 차지하는 토픽, 즉 상대적 중요도가 높은 토픽의 경우, 토픽을 구성하는 키워드가 가장 빈번하게 발생한 것을 의미하며, 낮은 비중을 차지하는 토픽의 경우 일부 연구(문서)에 따라 해당 연구분야의 비중이 급격하게 변화하거나 토픽을 구성하는 단어 키워드에 더욱 민감하게 작용할 수 있다. g-DMR 모형을 통해 도출된 토픽은 부가적인 입력변수로 출판연도와 연구분야와 같은 메타정보를 포함하고 있기 때문에 이러한 정보를 반영한 주제를 도출하고, 기본적인 빈도 분석이나 LDA와 같이 논문 초록만을 입력으로 투입하는 모델을 통해서 얻을 수 없었던 부수적 의미를 파악할 수 있다는 점에서 매우 유용하게 활용될 수 있다.

5. 결론

5.1. 논의 및 시사점

코로나 19로 인해 가장 주목받고 있는 변화는 재택근무, 원격 진료, 온라인 교육 등 비대면화의 확산이다. 병원 내 감염 확산을 방지하기 위해 한시적으로 전화 상담과 비대면 처방이 허용되면서 그동안 경험하지 못했던 원격 진료도 경험하고 있다. 생필품을 온라인으로 구매하거나 식품에 대한 배달 수요가 증가하고, 재택·원격근무, 화상회의 등의 스마트워크가 확산되는 등 전 분야의 디지털 전환이 가속화하고 있다(김정언, 2020).

디지털화 과정과 그 결과는 변화와 지속 가능한 사회의 형성을 가속화하고, 디지털 세계에서의 우리의 결정, 행동 및 존재는 데이터를 생성하며, 이는 비즈니스 방법과 관행 개선에 상당한 기회를 제공한다(Pappas et al., 2018). 따라서 진정한 디지털 시대로의 전진(도약)을 위해서는 지금까지의 데이터·네트워크·인공지능, 즉, D.N.A. 기술과 비즈니스 생태계에 대한 보다 상세한 이해가 필요하다. 본 연구는 토픽 모델링 결과로 도출된 다양한 주제를 탐색함으로써 디지털 전환하에서 D.N.A. 기술과 비즈니스 환경에 대한 전반적인 연구동향을 파악하고자 하였다. 이러한 맥락을 통해 g-DMR 모형의 분석 결과는 다음과 같은 시사점을 제공한다.

학문적 관점에서 살펴보면, 코로나19를 기점으로 2019년도부터 디지털 전환에 관한 연구가 전반적으로 증가하고 있음을 확인할 수 있다. 특히 전체 토픽 비중의 절반 이상을 차지하는 비즈니스 인텔리전스(토픽 4) 관련 주제가 Business & Economics, Social Sciences, Engineering, 및 ICT 등 다수의 연구분야에서 급격한 증가 추세를 나타내고 있다. 상위 키워드인 ‘industry’, ‘management’, ‘analysis’, ‘organization’, ‘knowledge’, ‘company’ 등을 통해 비즈니스 인텔리전스가 기업들이 보유한 수많은 데이터를 분석하여 의사결정을 지원하는 일련의 기술로서 다양한 범주와 주제로 널리 연구되고 있음을 짐작해볼 수 있다. Bordeleau et al.(2020)은 더 많은 기업이 BI&A(Business Intelligence and Analytics) 활동을 통해 디지털 시대로의 전환에 참여함으로써 데이터에서 가치를 창출해야 한다고 제안하였으며, 이러한 주장은 본 연구에서 도출된 비즈니스 인텔리전스(토픽 4), 데이터 분석 기술(토픽 7), 그리고 서비스 가치 창출(토픽 8) 등의 주제와 연계하여 생각해볼 수 있다. 특히 서비스 가치 창출에 해당하는 토픽의 연구 추이를 보면, ‘service’, ‘business’, ‘customer’, ‘value’, ‘firm’ 등이 상위 키워드를 구성하며 2018년도부터 ‘Humanities/Others’ 분야에서 빠른 속도로 비중이 증가하는 것을 알 수 있다.

제조 생산 시스템(토픽 2)에 해당하는 키워드는 여러 연구분야에서 2018년 시점부터 상승 추이에 접어드는 반면 정보 기술(토픽 3) 토픽은 반대의 양상을 나타내고 있는데, 이는 정보 기술에 해당하는 다수의 기술이 세분화되어 네트워크(토픽 6), 데이터, 그리고 자동화(토픽 11) 등의 세부 주제로 나뉘어 나타난 결과로 짐작할 수 있다. 또한 네트워크 기술과 관련된 주제의 경우 ‘Humanities/Others’ 범주에서 가장 높은 연구 비중을 차지하는 것을 확인할 수 있는데, 이는 네트워크 기술의 특성에 따라 학제간 연구가 활발하게 수행되었음을 의미한다. 최근 제조 분야에서의 디지털 전환에 대한 키워드가 다시 한번 강조되고 있는데, 이러한 이유로는 빅데이터 기반의 지능형 제조 솔루션과 메타버스로 구현되는 제조 시뮬레이션 등의 기술이 화두로 떠오르고 있음에 기인한다. 이는 빈도 기반의 분석 결과에서도 유사하게 나타난다.

연구분야별 주요 토픽과 경향성을 <표 10>과 같이

정리하였다. 디지털 전환과 관련된 모든 연구분야에서 “비즈니스 인텔리전스”에 해당하는 토픽이 활발하게 연구되고 있을 뿐만 아니라, 실무적으로 기업의 빅데이터 환경이 발전함에 따라 기업이 보유한 수많은 데이터를 정리하고 분석하여 의사결정을 돕는 BI 부문의 디지털화가 촉진되고 있다. 공학 및 정보통신 분야의 경우 공통적으로 “제조 생산 시스템”에 관한 토픽이 그 중요도가 상대적으로 높아지고 있으며, 실제 제조현장에서 데이터, 인공지능, 5G, 및 사물인터넷 등의 기술을 접목한 지능형 생산 시스템, 기업 간 협업 솔루션 등이 재조명받으면서 제조 분야의 디지털 전환 이슈를 이끌고 있다. 또한 헬스케어와 자연과학 분야는 “원격 진료”, “모니터링 시스템”, “데이터 분석 기술” 등이 주요 키워드로써 최근 더욱 활발하게 연구되고 있음을 확인할 수 있었고, “자동화 기술”에 해당하는 토픽은 두 가지 연구분야에 동시에 등장하고 있다. 디지털 기반의 스마트 병원은 ICT를 의료에 적용하여 입원환경 개선과 환자 및 보호자 교육을 통해 환자의 안전을 강화하며 의료서비스의 혁신을 가속화하고 있다. 그 밖에도 사회과학 분야에서 디지털 교육에 관한 이슈가 주요 토픽으로 나타났는데, 이는 코로나 19 이후 시행된 비대면 수업 등으로 디지털 리터러시 교육에 대한 수요가 증가하고, 변화하는 교육 및 학습 환경에 적응하기 위해서 디지털 교육이 주목받고 있음을 알 수 있다.

결론적으로 데이터와 네트워크, 그리고 인공지능에 이르는 세 가지 기술은 각기 다른 영역의 독립적인 기술로 간주할 수 있으나, 서비스 관점에서는 이를 떼어놓고 생각할 수 없다. 5G와 같은 네트워크 통신을 이용하는 IoT, 센서 등의 스마트 기기에서 수집된 데이터를 축적하고 가공·결합하여 자율주행차, 스마트공장, 디지털 정부, 원격 진료 등 여러 분야에 제공 및 활용하도록 지원하여 인공지능 기반의 다양한 서비스를 창출할 수 있기 때문이다. 이러한 맥락에서 제조 생산 시스템, 정보, 네트워크, 데이터 분석 기술과 관련된 토픽은 데이터와 네트워크, 인공지능의 기술적인 측면과 디지털 기술을 아우르는 원격 진료(토픽 10), 디지털 교육(토픽 13), 감지 시스템(토픽 5)의 토픽을 비즈니스 환경 측면으로 구분할 수 있지만, 사실 양 측면에서 도출된 토픽의 기술과 서비스는 상호 유기적으로 연계되어 있다고 볼 수 있다. 구체적으로, 원격 진료에 해당하는 토픽의 경우 헬스케어 산업에서 비대면 서비스의 형태로 바라볼 수 있지만, 결국 원격 진료 서비스가 얼굴, 피부, 모발, 장기 등의 신체 정보를 포함하는 다양한 이미지 및 영상 데이터를 수집하여 이를 딥러닝과 같은 인공지능 기술로 분석함으로써 디지털 기술 기반하에 제공되는 것을 의미한다. 마찬가지로 디지털 교육 관련 토픽은 D.N.A.기반의 인프라 강화를 통해 온라인 수업과 교육 콘텐츠와 같은 서비스를 확충함으로써 비즈니스 환경이 조성될 수 있음을 시사한다.

이상의 논의를 요약하자면, 데이터, 네트워크 및 인공지능을 중심으로 진행하였던 디지털 전환 연구의 토픽 모델링 결과로는 첫째, COVID-19 이후 비즈니스 인텔리전스를 주제로 하는 연구가 전 영역에서 활발하게 수행되고 있으며, 둘째, 공학 및 정보통신 분야에서 지능형 제조 솔루션 및 메타버스 등의 이슈가 등장함에 따라 제조 생산 시스템에 관한 주제와 헬스케어 분야의 원격 진료, 자동화 기술, 사회과학 분야의 디지털 교육 등이 주목받고 있음을 확인하였다. 마지막으로, 비록 토픽 모델링의 결과로 도출된 주제어 자체는 기술과 서비스의 측면에서 분리하여 볼 수도 있었지만, 결과적으로 다수의 연구에서 해당 기술들을 접목하여 적용된 다양한 서비스를 포괄적으로 다루고 있으므로 이를 별개로 해석하는 것은 바람직하지 못하다는 것을 알 수 있었다.

본 연구에서는 여러 텍스트마이닝 방법론을 적용하여 WoS 데이터베이스의 SCIE급 색인에 해당하는 연구의 초록, 출판연도 및 연구분야를 입력변수로 활용하여 주요 토픽을 도출하였다. 우선, 단어 출현 빈도에 기반한 TF 및 TF-IDF 분석을 통해 주요 키워드를 확인하고, 이어서 g-DMR 모형을 이용하여 토픽 분석을 수행하였는데, 다양한 형태의 변수를 메타정보로 활용 가능한 해당 토픽 모형의 이점으로 단순하게 토픽을 도출하는 것 이상의 의미를 적절하게 탐색할 수 있었다. 분석 결과에 따르면, 비즈니스 인텔리전스, 제조 생산 시스템, 서비스 가치 창출, 원격 진료, 디지털 교육 등의 토픽들이 디지털 전환에서 주요 연구주제인 것으로 식별되었으며, 분석 결과를 바탕으로 본 연구의 시사점을 도출하였다.

5.2. 연구의 한계점 및 향후 연구과제

디지털 전환이라는 비교적 최근에 주목받는 이슈를 중심으로 관련된 세부 기술을 데이터, 네트워크 및 인공지능으로 한정하여 연구정보를 추출하였기에 데이터의 절대적인 수가 충분하지 못하였다. 당연하게도 최근 시점에 연구된 자료의 비중이 매우 높게 나타났으며, 더불어 토픽 분석에 활용된 g-DMR 모형의 특성에 따라 특정 토픽이 편중되어 나타났다. 비교적 균등한 비중의 토픽을 도출하기 위해서는 사전 데이터 수집 단계에서부터 활용 가능한 메타정보의 데이터를 어느 정도 비슷한 양으로 추출하고, 용어 가중치 기법 등의 파라미터 조정을 통해 해결할 수 있을 것이다.

향후 연구에서는 더욱 다양하고, 폭넓은 분석을 위해서 특허, 보고서 등의 텍스트 자료를 활용하여 충분한 양의 데이터를 확보하고, 메타정보 또한 저자, 국가, 피인용수, 분류체계 등 세부적인 분석 목적에 적합한 자료 처리 등을 통해 다양한 형태로 연구의 저변을 넓힐 수 있을 것이다.

〈참고문헌〉

[국내 문헌]

1. 김선주, 김병수 (2021). 공유숙박업에서 고객 충성도에 영향을 미치는 요인: 구조 방정식 모형과 토픽 모델링 분석. **지식경영연구, 22**(3), 55-73.
2. 김정언 (2020). 디지털 뉴딜 주요 내용과 향후 과제. **정보처리학회지, 27**(2), 4-12.
3. 김태후, 한능호 (2021). 토픽모델링을 활용한 무역학 연구동향 분석: 4차산업혁명과 디지털 전환을 중심으로. **무역금융보험연구, 22**(3), 21-38.
4. 노규성 (2020). 포용적 혁신성장과 일자리 창출을 위한 디지털 뉴딜 전략에 관한 연구. **디지털융복합연구, 18**(1), 23-33.
5. 문형돈 (2021). **2021년 DNA 분야별 국내 디지털 혁신 전망** (No. 포커스1 FOCUS). 소프트웨어정책연구소.
6. 박문우 (2020). 한국판 뉴딜, 국가 디지털 전환을 위한 Data·Network·AI 기반 데이터 댐. **정보처리학회지, 27**(2), 13-20.
7. 박영욱, 정규엽 (2021). DMR(Dirichlet Multinomial Regression) 토픽모델링을 이용한 온라인 리뷰 빅데이터 기반 고객감성 분석에 관한 연구: 국내 5성급 호텔의 외국인 이용객 리뷰를 중심으로. **호텔경영학연구, 30**(2), 1-20.
8. 손권상, 권오병 (2021). 디지털 뉴딜 정책에 대한 언론 보도량과 주식 시장의 동태적 관계 분석: 4차산업혁명 관련 기업을 중심으로. **한국전자거래학회지, 26**(3), 33-52.
9. 안재영, 심소연, 윤혜정 (2022). 토픽 모델링 기법을 활용한 메타버스 증강현실 연구 동향 분석. **지식경영연구, 23**(2), 123-142.
10. 오혜라, 정윤재 (2021). 상관토픽모델을 활용한 유튜브 크리에이터 소비자의 온라인 구전 특성에 관한 연구. **한국광고홍보학보, 23**(3), 37-72.
11. 원종호, 이대호, 박인영 (2021). DMR 토픽 모델링을 활용한 배달 플랫폼 노동에 대한 인식 분석-디시인사이드 배민커넥트 갤러리 게시물과 신문기사 분석을 중심으로. **한국혁신학회지, 16**(2), 175-210.
12. 윤혜정, 안재영, 박상철 (2021). 토픽 모델링과 수정된 IPA를 활용한 O2O 주문·배달 앱에 대한 사용자 인식 연구. **지식경영연구, 22**(3), 253-271.
13. 이웅배, 이선웅, 정진섭 (2021). 디지털 트랜스포메이션에 따른 비즈니스 모델 혁신 메커니즘. **메커니즘 저널, 1**(1), 1-22.
14. 이현상, 조보근, 오세환, 하성호 (2021). 섬유소재 분야 특허 기술 동향 분석: DETM & STM 텍스트마이닝 방법론 활용. **정보시스템연구, 30**(3), 201-216.
15. 정유경 (2020). 디지털 인문학 분야의 국내외 연구 동향 분석. **정보관리학회지, 37**(2), 311-331.
16. 정준화, 박소영 (2021). 디지털 시대를 위한 D.N.A.(data, network, AI) 정책의 현황과 과제. **국회입법조사처 이슈와 논점, 1828**. Retrieved from https://www.nars.go.kr/report/view.do?cmsCode=CM0043&brdSeq=34607
17. 최현홍, 심동녘 (2020). 텍스트마이닝을 적용한 ICT융합 트렌드 분석. **한국혁신학회지, 15**(3), 257-281.

[국외 문헌]

18. Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. ***Annals of Applied Statistics, 1***(1), 17-35. https://doi.org/10.1214/07-AOAS114
19. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. ***Journal of Machine Learning Research, 3***(Jan), 993-1022.
20. Bordeleau, F., Mosconi, E., & De Santa-Eulalia, L. A. (2020). Business intelligence and analytics value creation in industry 4.0: A multiple case study in manufacturing medium enterprises. ***Production Planning and Control, 31***(2-3), 173.
21. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. ***Advances in Neural Information Processing Systems 22-Proceedings of the 2009 Conference***, 288.
22. Lee, M., & Song, M. (2020). Incorporating citation impact into analysis of research trends. ***Scientometrics, 124***(2), 1191. https://doi.org/10.1007/s11192-020-03508-3
23. Mimno, D., & McCallum, A. (2012). ***Topic models conditioned on arbitrary features with dirichlet-multinomial regression***. arXiv Preprint, arXiv:1206.3278.
24. Newman, D., Baldwin, T., Lau, J. H., & Grieser, K. (2010). Automatic evaluation of topic coherence. ***NAACL HLT 2010-Human Language Technologies: The 2010***

Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference, 100.

25. Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., & Lekakos, G. (2018). Big data and business analytics ecosystems: Paving the way towards digital transformation and sustainable societies. ***Information Systems & E-Business Management, 16***(3), 479-491.

26. Ralph, B., & Stockinger, M. (2020). Digitalization and digital transformation in metal forming: Key technologies, challenges and current developments of industry 4.0 applications. ***Paper Presented at the XXXIX Colloquium on Metal Forming***, Zauchensee, 13-23.

27. Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. ***American Journal of Political Science, 58***(4), 1064.

28. Siebel, T. M. (2018). ***Why digital transformation is now on the CEO's shoulders*** (Big data, Internet of Things, and Artificial intelligence No. 2018). McKinsey & Company.

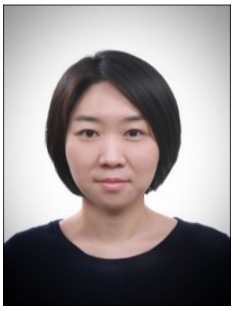
저 자 소 개

안 세 환 (Sehwan An)

단국대학교에서 경영학 학사, 한양대학교 대학원에서 산업공학 석사 학위를 취득하였고, 현재 기술경영학 박사과정을 수료하였다.
한국생산기술연구원 스마트제조기술그룹에서 병역특례 연구원으로 근무한 바 있고, 주요 연구분야는 제조 데이터 분석, 통계적 품질관리, 머신러닝, 텍스트 마이닝 등이다.

고 강 욱 (Kangwook Ko)

한양대학교 산업공학과에서 학사, 인디애나 대학교 비즈니스 스쿨에서 MBA를 취득하였고, 현재 한양대학교 대학원에서 기술경영학 박사과정을 수료하였다. 삼성전자 경영혁신센터에서 프로세스 혁신업무를 하고 있으며 주요 연구분야는 Supply Chain Management, 인공지능 기술, 텍스트 마이닝 등이다.



김 영 민 (Youngmin Kim)

한양대학교 산업공학과에서 학사, 석사 학위를 취득한 후 프랑스 Paris 6 대학 컴퓨터 공학과에서 석사, 박사 학위를 취득했다. Avignon 대학과 Lyon2 대학에서 박사후 연구원, 한국과학기술정보연구원에서 선임연구원으로 재직하였다.
2016년부터 한양대학교 기술경영전문대학원 교수로 재직하고 있다. 주요 연구분야는 기계학습, 확률 그래프모델, 정보 추출이다.

〈 Abstract 〉

Digital Transformation: Using D.N.A.(Data, Network, AI) Keywords Generalized DMR Analysis

Sehwan An*, Kangwook Ko**, Youngmin Kim***

As a key infrastructure for digital transformation, the spread of data, network, artificial intelligence (D.N.A.) fields and the emergence of promising industries are laying the groundwork for active digital innovation throughout the economy. In this study, by applying the text mining methodology, major topics were derived by using the abstract, publication year, and research field of the study corresponding to the SCIE, SSCI, and A&HCI indexes of the WoS database as input variables. First, main keywords were identified through TF and TF-IDF analysis based on word appearance frequency, and then topic modeling was performed using g-DMR. With the advantage of the topic model that can utilize various types of variables as meta information, it was possible to properly explore the meaning beyond simply deriving a topic. According to the analysis results, topics such as business intelligence, manufacturing production systems, service value creation, telemedicine, and digital education were identified as major research topics in digital transformation. To summarize the results of topic modeling, 1) research on business intelligence has been actively conducted in all areas after COVID-19, and 2) issues such as intelligent manufacturing solutions and metaverses have emerged in the manufacturing field. It has been confirmed that the topic of production systems is receiving attention once again. Finally, 3) Although the topic itself can be viewed separately in terms of technology and service, it was found that it is undesirable to interpret it separately because a number of studies comprehensively deal with various services applied by combining the relevant technologies.

Key words: Digital Transformation, Data, Network, Artificial Intelligence(AI), Topic Modeling

* hwan86@hanyang.ac.kr
** system@hanyang.ac.kr
*** yngmnkim@hanyang.ac.kr