교육과정평가연구
The Journal of Curriculum and Evaluation
2024, Vol. 27, No. 3, pp. 197~221
DOI: https://doi.org/10.29221/jce.2024.27.3.197

텍스트 마이닝을 활용한 국내외 프로세스 데이터 연구 동향 분석[1)]

최진수 (충남대학교 박사수료)*
정혜원 (충남대학교 교수)**

요약

본 연구에서는 교육학 분야에서 프로세스 데이터 관련 국내외 연구 동향을 살펴보기 위해 국내 46편, 국외 20편의 학술논문을 연구 대상으로 선정하여 텍스트마이닝 분석하였다. 분석 결과는 다음과 같다. 첫째, 국내 연구에서는 학습 분석, 교수학습 유형 분석, AI학습(교과), 학업성취, 국외 연구는 평가 과정(학생 태도), 탐구 기반 학습 행동 분석, 문제해결 과정, LMS를 활용한 학습 유형 분석으로 각각 4개 토픽이 나타났다. 둘째, 국내에서는 학업성취 관련한 연구의 비중이 가장 높았고, 상향(hot) 토픽인 것으로 나타났으며, 국외에서는 4개 주제의 비중이 비슷하고, 평가 과정(학생 태도), 탐구 기반 학습 행동 분석의 주제가 새롭게 등장한 주제였다. 셋째, 국내에서는 learning, data, time, 국외에서는 student, process, performance가 토픽을 연결하는 주요 키워드였다. 특히 국내에서는 learning, 국외에서는 student가 토픽 간 긴밀하게 연결하고 중심성 정도가 높았다. 이를 통해 프로세스 데이터와 관련된 국내외 연구 동향을 파악하고 향후 연구의 방향성 및 정책 수립에 기초자료로 활용할 수 있을 것이다.

주제어 : 프로세스 데이터, 로그 데이터, 연구 동향, 텍스트 마이닝

1) 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 연구되었음(NRF-2022M3J6A1084843).
* 제1저자, jins1204@gmail.com
** 교신저자, chw7@cnu.ac.kr

I. 서 론

학생이 문제 상황에 접했을 때 해결 과정을 이해하는 것은 학생의 인지적 특성과 학업성취도를 파악하는 것과같이 중요하다. 이와 관련하여 국내에서도 개정 교육과정을 통해 과정 중심 평가를 강조하고 있다. 2015 개정 교육과정에서는 기존의 결과 중심 평가와 대비되는 과정 중심 평가를 도입하여 패러다임 전환을 꾀하고 있는 것으로 나타났고, 2022 개정 교육과정에서도 학생의 문제해결 및 사고의 과정, 학습의 과정을 점검하는 평가를 강조하고 있다(교육부, 2017; 교육부, 2020; 함은혜, 2022). 그러나 이러한 패러다임 전환을 8년간 지속하고 있고, 미래 교육의 방향을 반영한다는 점에서 한국 교육현장에 과정 중심 평가 방식이 안정적으로 정착할 필요가 있다(최훈원, 최윤정, 2024).

최근 온라인 학습이 보편화되면서 학업성취도 평가와 더불어 학생의 문제해결 과정을 이해하기 위한 컴퓨터 기반 검사로의 평가 체계가 전환되고 있다(Fang & Ying, 2020). 컴퓨터 기반 검사는 지필평가로 수집하기 어려운 과정적 정보를 수집할 수 있고, 이와 같은 프로세스 데이터에 대한 접근성이 좋기에 주목받고 있다. 관련 분야 전문가들은 컴퓨터 기반 검사를 통해 학생 평가와 관련된 새로운 분석과 해석의 가능성을 열어줄 수 있어 차세대 평가의 핵심 역할을 할 것으로 기대한다(전성균, 상경아, 2023; 함은혜 2022). 피험자의 문제해결 과정을 이해하기 위해서는 문제해결 과정 중에 발생하는 판단과 결정의 일련의 과정을 파악하고 평가할 수 있는 적절한 평가 방법이 필요하다. 즉 컴퓨터 기반 검사를 통해 수집한 프로세스 데이터로 개인의 문제해결 역량을 측정하고 분석할 수 있는 새로운 평가 체제로의 가능성이 열린 것이다(김연후, 2023).

교육 분야에서 데이터 활용과 자료 분석은 학생의 특성을 파악하고, 학습을 설계하는데 중요한 방법이다. 최근에는 교육과 학습 상황을 더 깊이 이해하기 위해 프로세스 데이터를 통한 다양한 분석 방법이 적용되고 있다(송민정, 2013; 안미리, 2016; 조용상 외, 2013). 국내에서는 주로 LMS를 통한 학습자의 학습 과정을 객관화하고 학생들의 학습 상황과 성과 예측, 교수학습을 개선하는데 프로세스 데이터를 활용하고 있었다(김영수 외, 2016). 프로세스 데이터를 활용한 학습자의 동기와 인지적 과정을 이해하는 교육 분야 연구는 국제적으로 크게 증가하고 있다(함은혜, 2022). 이와 관련하여 주로 응답시간을 활용한 인지 능력 수준에 대한 추정의 정확도, 검사의 타당화에 관심이 있고, 최근에는 학습자의 정의적 특성과 관련지어 분석, 활용하는 연구가 증가하고 있다(이소라, 2019; Kyllonen & Zu, 2016).

이와 같이 국내외 프로세스 데이터에 대한 교육 분야의 관심이 높아지고 있음에도 불구하고 연구 동향 및 트렌드를 분석하여 연구 양상을 살펴본 연구는 없었다. 교육 분야에서 프로세스 데이터를 활용한 연구의 다양성과 누적이 요구됨에 따라 최근까지 프로세스 데이터 관련하여 수행된 국내외 연구 동향을 비교 분석하고 후속 연구를 위한 아이디어와 방향성을 제시할 필요가 있다.

본 연구에서는 최근 연구 동향 분석에 주로 활용되고 있는 토픽모델링과 의미연결망 분석을 활용하여 프로세스 데이터 연구 동향을 살펴보고자 한다. 이러한 연구 기법은 연구자의 주관이 많이 개입될 수 있는 기존의 연구 방법의 단점을 보완하고 더 객관적이고 밀도있게 연구 동향을 분석할 수 있다는

장점이 있어 최근 사회과학 분야에서도 연구 동향 분석에 주로 활용되고 있다(김대영, 2021; 최훈원, 최윤정, 2024). 본 연구에서는 특히 과거부터 현재까지 국내외 연구 동향을 비교 분석하여 국내와 국외에서 관심이 있는 연구 주제를 각각 파악하고 연구 주제별 트렌드 변화 양상을 살펴봄으로써 향후 관련 분야의 연구 방향성과 시사점을 제시하고자 하였다. 이를 통해 컴퓨터 기반 검사로 수집한 프로세스 데이터가 교육 분야에서 어떠한 연구 주제로 활용되고 있는지 살펴보고, 2022 개정 교육과정에서 강조하고 있는 학생의 문제해결 과정에 대한 인지 과정 및 과정 중심 평가와 관련된 정책적 방향을 제시하는데 기초자료로 활용될 수 있을 것으로 기대된다. 이에 본 연구의 문제는 다음과 같다.

연구 문제 1. 프로세스 데이터 연구의 핵심 키워드와 토픽은 무엇인가?

연구 문제 2. 프로세스 데이터 연구의 토픽별 변화 양상은 어떠한가?

연구 문제 3. 프로세스 데이터 연구의 토픽별 핵심 키워드의 중심성과 관계는 어떠한가?

II. 이론적 배경

1. 컴퓨터 기반 검사(Computer Based Assessment)

컴퓨터 기반 검사는 학생의 평가 수행을 포함한 전 과정이 컴퓨터 시스템을 통해 이루어지는 평가를 의미한다. 최근에는 온라인 학습이 보편화됨에 따라 학업성취도 평가와 문제해결 과정을 파악하기 위해 컴퓨터 기반 검사로 전환되고 있다. 대규모 평가 프로그램에서도 컴퓨터 기반 평가의 중요성을 인식하고 적극적으로 평가 체제를 변환하는 추세에 있는데, 국내에서도 학업성취도 평가를 수행하고 교육정책을 수립하기 위해 참여하는 국제 학업성취도 평가인 PISA와 TIMSS가 각각 컴퓨터 기반 검사로 전환되어 운영되고 있다. PISA는 2015년부터 컴퓨터 기반 검사가 전면 도입되었고, TIMSS도 2023년부터 도입되어 운영하고 있다. 컴퓨터 기반 검사는 응답 과정 중에 프로세스 데이터를 수집할 수 있다는 장점이 있다(Oranje et al., 2017). 이와 같이 컴퓨터 기반 검사는 문제해결의 중요성이 강조됨에 따라 도입되었으며, 문제해결은 문제를 이해하고 해결하는데 요구되는 개인의 인지적 능력을 파악하는데 중요하다(송미영 외, 2014).

2. 프로세스 데이터(process data)

프로세스 데이터는 넓은 의미로는 디지털 기기가 기록할 수 있는 모든 데이터를 의미하고, 이때 프로세스 데이터는 로그 데이터(log data)로서 디지털 시스템과 피험자 상호작용에 대한 상세 기록 및 사용자 또는 서버가 생성하는 이벤트의 타임스탬프이고(Jiang et al., 2021), 컴퓨터 기반 검사로 평가하는 과정에서 수집되는 모든 정보를 의미한다(김연후, 2023; Provasnik, 2021). 좁은 의미에서 프로세스 데이터는 문제해결 과정에서 기록된 응답자의 연속 행동을 의미한다. 액션 시퀀스(action

sequence)를 의미한다. 프로세스 데이터는 하나의 평가에서 문제해결 과정이 다른 평가에서는 결과 데이터가 될 수 있어 생성적 정의에 모호함이 있다(Bergner & Davier, 2019; Levy, 2020). 액션 시퀀스로서의 프로세스 데이터는 타임 스탬프와 함께 기록된 작업 수행 데이터이고 평가 문항에 대한 작업 과정을 반영하는 경험적 데이터로 인지적, 비인지적 잠재 구조를 반영하고 있는 것으로 판단하다(Provasnik, 2021; Zhan & Qiao, 2022). 본 연구에서는 로그 데이터 및 액션 시퀀스를 포함한 넓은 의미로서의 프로세스 데이터로 개념을 정의하였다.

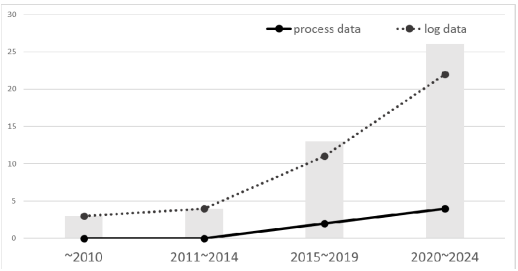
III. 연구 방법

1. 분석 대상

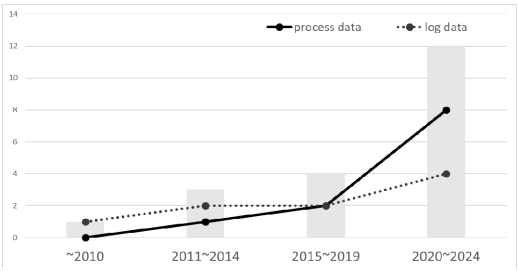
본 연구에서는 프로세스 데이터와 관련한 국내외 연구 동향을 살펴보기 위해 학술 연구 정보 서비스(RISS)에 탑재된 국내외 학술지 정보를 검색하였고, 본 연구에서는 넓은 의미로서의 프로세스 데이터를 정의하였으나, 다른 연구에서 좁은 의미로서 로그 데이터와 구분하여 사용하거나, 과정적 수행 기록에 따른 데이터에 대한 표현이 다를 수 있기 때문에, 연구 대상을 선정함에 있어서 '프로세스 데이터(process data)'와 '로그 데이터(log data)', '액션 시퀀스(action sequence)'를 키워드로 검색하여 관련 논문을 수집하였다. 이때, 액션 시퀀스는 해당 키워드로 발간된 국외 논문이 없기 때문에 국내외 연구 동향 비교를 위해 이는 분석에서 제외하였다. 국내에서는 한국학술진흥재단(KCI) 등재지로 선정되고 교육학 분야에서 발간된 논문 74편, 국외는 SSCI로 등재된 학술지 중 교육학 분야에서 발간된 논문 34편이 검색되었다. 이 중에서 주제와 관련이 없는 논문과 영문 초록이 없는 경우를 제외하여 최종적으로 국내 46편, 국외 20편이 연구 대상으로 선정하였고, 국내외 연구 동향을 비교하기 위해서 각 논문의 영문 초록을 분석에 활용하였다. 연구대상 논문의 정보는 다음의 〈표 1〉과 같다. 국내외 논문의 process data와 log data의 키워드로 검색된 논문을 각 연도별로 구분하여 도식화한 결과는 [그림 1], [그림 2]와 같다. 국내 연구의 경우, process data보다 log data를 키워드로 더 많은 연구가 이루어지고 있었고, 최근에도 크게 증가하는 추세인 것으로 파악된다. 국외 연구의 경우에는 process data와 log data로 발간된 논문의 수는 비슷한 수준이나 최근에는 process data를 키워드로 한 논문이 증가하고 있는 것으로 보인다.

〈표 1〉 분석 대상

구분	국내 논문				국외 논문				소계 (개)
	Process data		Log data		Process data		Log data		
	논문(개)	비율(%)	논문(개)	비율(%)	논문(개)	비율(%)	논문(개)	비율(%)	
~2009	0	0.0%	3	7.5%	0	0.0%	1	9.1%	4
2010~2014	0	0.0%	4	10.0%	1	9.1%	2	18.2%	7
2015~2019	2	33.3%	11	27.5%	2	18.2%	2	18.2%	17
2020~2024	4	66.7%	22	55.0%	8	72.7%	4	36.4%	38
합계	6	100.0%	40	100.0%	11	100.0%	9	81.8%	66



[그림 1] 국내 발행 논문 수



[그림 2] 국외 발행 논문 수

2. 연구 방법

본 연구는 프로세스 데이터 연구 동향을 분석하기 위해서 텍스트마이닝을 실시하였다. 키워드를 추출하여 빈도 분석하고, 문서 내 단어의 중요도를 의미하는 가중치인 TF-IDF(Term Frequency-Inverse Document Frequency)를 활용하여 토픽모델링 및 데이터를 시각화하였다. 분석에는 NetMiner 4.0 program을 사용하였고, 기술통계와 연도별 시계열 분석에는 SPSS 27.0 for window를 사용하였다.

가. 데이터 전처리

본 연구에서는 학술 연구 정보 서비스(RISS)에 탑재된 학술지 중 '프로세스 데이터(process data)'와 '로그 데이터(log data)'를 키워드로 검색한 논문의 영문 초록을 수집하여 전처리를 실시하였다. 텍스트마이닝에서 전처리 과정은 잘못된 단어 집합이나 의미 없는 분석 결과가 나타나지 않도록 사전 점검하는 단계로 연구 과정에서 가장 중요하다고 할 수 있다. 문서에서 명사를 추출하되, 의미 분석이 어려운 1글자 단어를 제외하고 유사한 의미의 단어는 동의어로, 주요 용어는 복합명사로 처리하여 연관어 분석 시 단순 명사로 분리되지 않도록 지정어, 유사어, 제외어 사전을 작성하여 전처리하였다(최진수, 정혜원, 2022). 이를 토대로 TF-IDF와 N-gram을 사용하여 키워드를 추출하였으며, 본 연구에서

는 복합명사 처리 시에 두 단어, 세 단어가 함께 등장하는 관계를 모두 포함하여 지정어 사전을 작성하였으므로 n-gram 방식을 활용하였다고 볼 수 있다. 또한, 많은 문서에서 반복적으로 사용되는 경우 유의한 단어 집합을 만들기 어렵기 때문에 TF-IDF 값이 0.1 이상인 단어만 추출하여 분석에 활용하였다. 이때 TF 값은 문서 내부 단어의 출현 빈도를 모든 단어의 출현 횟수로 나누어 정규화한 형태이며, DF는 로그 정규화를 통해 산출되므로 TF와 IDF 인자를 곱한 TF-IDF 가중치는 정규화된 것으로 간주할 수 있을 것이다(이성직, 김한준, 2009). 유예림(2017)의 연구에서도 문서 길이 차이로 발생할 수 있는 영향을 최소화하기 위해 정규화된 TF-IDF를 활용하고 있다. 본 연구에서는 정규화된 TF-IDF 값을 합산하여 제시하였다. 개별 단어가 각 문서에서 가지는 각각의 TF-IDF 가중치 값을 모두 더한 TF-IDF 가중치는 단어의 중요도를 평가할 수 있는 비교적 신뢰가 가능한 지표로 평가된다. TF-IDF 가중치가 높을수록 단어의 중요도가 높고 주요 키워드로 추출할 수 있다(이예은, 장정현, 2022; 장은아, 정혜원, 2024). TF-IDF 가중치는 다음의 〈수식 1〉과 같이 산출할 수 있다. TF-IDF는 TF와 IDF의 곱으로 표현되며, TF는 단어 i가 j번째 문서에서 등장하는 빈도를 j에 등장한 모든 단어의 수로 나눈 값이며, IDF는 전체 문서의 수를 단어 i를 포함한 문서의 수로 나눈 값에 log를 취한 값이다. 본 연구에서 사용한 sumTF-IDF는 개별 문서에서 도출된 단어 i의 개별 TF-IDF 값을 합하여 단어 i에 대한 전체 문서에서의 TF-IDF 값을 의미한다.

$$(TF-IDF)_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad \cdots \langle\text{수식 } 1\rangle$$

$$TF_{(i,j)} = \frac{\text{문서 } j\text{에서 단어 } i\text{가 등장한 횟수}}{\text{문서 } j\text{에서 등장한 모든 단어의 수}},\ IDF_i = \log\left(\frac{\text{총 문서의 개수}}{\text{단어 } i\text{를 포함하는 문서의 수}}\right)$$

나. 토픽모델링

텍스트마이닝 기법 중 토픽모델링은 방대한 텍스트로부터 숨어있는 중요한 토픽 또는 주제를 찾아내는 기법이며, 관측치가 각 토픽에 할당될 확률을 모두 제시하여 분석한다는 점에서 군집분석과 차이가 있다. 또한, 토픽모델링은 잠재 변수로서 토픽을 임의 추출하기 때문에 이론적 배경과 내용학적 이론이 뒷받침되어야 한다. 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 알고리즘을 활용한 토픽모델링 방법은 토픽 수 결정이 임의적으로 이루어지지만, 사회과학 분야 텍스트마이닝 연구에서 현재까지 가장 많이 활용되고 있다(유진은, 2021). 문서 내 단어를 최소한의 토픽에 할당하면서 가능한 적은 수의 단어들이 특정 토픽에 속할 확률을 높이는 통합 확률을 기반으로 정밀도를 높였다(김기옥, 2020). 도출된 토픽에 대해 시기별로 각 연구 주제가 차지하는 비중을 살펴보고 시계열 회귀 분석을 실시한다. 토픽별 비중의 시간에 따른 변화 추이를 살펴보는 것은 교육정책의 흐름을 반영한 결과이고, 학문적 관심이 시간에 따라 어떻게 변화하는지를 파악할 수 있다(백영민, 2020). 그 결과, 표준화 계수가 양수로 나오면 상향(hot) 연구 주제, 음으로 나오면 하향(cold) 연구 주제로 구분할 수 있고, 이를 통해 주제별 트렌드 변화를 파악하였다(박종순, 김창식, 2019).

다. 의미연결망 분석

본 연구에서는 주요 키워드 간의 관계를 파악하고자 의미연결망 분석(Semantic Network Analysis, SNA)을 활용하였다. 이는 단어의 연결 중심성을 도출하여, 한 단어에 연결된 단어가 많을수록 그 단어의 중심성이 크고 해당 문서에서 중요한 단어로 판단한다(강주연, 이이든, 김지수, 2020; 오창우, 2017; 지미선, 2018; 최진수, 정혜원; 2020). 네트워크 분석은 방대한 양의 텍스트 데이터에서 잠재된 의미구조와 키워드 간 관계적 속성을 파악할 수 있기 때문에 사회적 이슈에 대한 빅데이터나 텍스트데이터를 분석하고자 하는 목적으로 활용되고 있다(강주연 외, 2020; 김용희, 한창근, 2019; 이신영, 2018). 본 연구에서는 도출된 키워드를 기반으로 토픽별 주요 키워드에 대한 중심성의 순위와 시각화한 결과를 제시하였다.

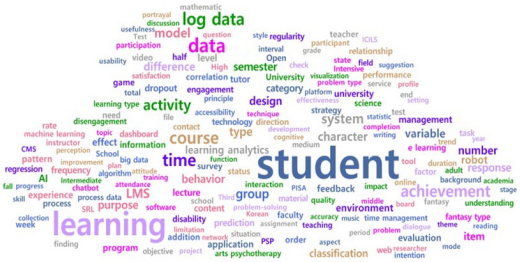
IV. 연구 결과

1. 키워드 추출 및 빈도 분석

본 연구의 연구 대상은 프로세스 데이터와 로그 데이터를 키워드로 발간된 국내외 논문 총 66편이다. 국내 논문 46편에서 명사를 추출하여 도출된 단어는 689개였으며, 전처리를 통해 최종 도출된 단어는 총 639개, 국외 논문 20편에서 도출된 단어는 424개였고, 전처리를 통해 최종 393개의 단어가 추출되었다. 각각 단어의 빈도 기준 순위와 TF-IDF 기준 순위를 나타낸 결과는 〈표 2〉와 〈표 3〉과 같다. 상위 200개 키워드를 추출하여 워드 클라우드로 시각화한 결과는 [그림 3], [그림 4]와 같다.

〈표 2〉 국내 연구의 단어 순위

단어 순위						TF-IDF를 통한 단어 순위					
	단어	빈도		단어	빈도		단어	TF-IDF		단어	TF-IDF
1	student	180	11	LMS	30	1	course	4.60	11	LMS	4.01
2	learning	126	12	group	30	2	achievement	4.52	12	group	4.01
3	data	63	13	character	29	3	activity	4.45	13	purpose	4.00
4	course	53	14	type	28	4	time	4.35	14	system	3.98
5	time	53	15	difference	26	5	model	4.24	15	design	3.93
6	log data	52	16	variable	26	6	variable	4.14	16	response	3.90
7	achievement	48	17	number	25	7	difference	4.12	17	environment	3.87
8	activity	43	18	behavior	24	8	character	4.11	18	number	3.87
9	system	37	19	design	24	9	semester	4.05	19	type	3.86
10	model	33	20	purpose	24	10	behavior	4.04	20	university	3.82



[그림 3] 빈도기반 워드클라우드

[그림 4] TF-IDF기반 워드클라우드

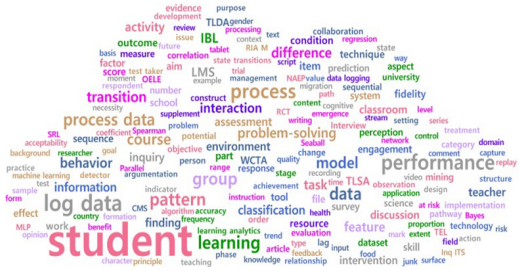
먼저, 국내 연구에서 전체 논문을 대상으로 상대적으로 많이 출현한 단어는 student(180), learning(126), data(63), course(53), time(53), log data(52), achievement(48), activity(43), system(37), model(33) 등과 같았다. 학생과 학습의 키워드 빈도수가 다른 단어들보다 많이 출현하고 있고, 다음으로 학기, 시간, 로그 데이터와 같은 데이터를 활용한 연구, 학생의 성취와 활동에 관심을 갖는 연구가 많이 이루어지고 있었다.

TF-IDF를 통한 단어의 중요도 순위를 살펴본 결과, course(4.60), achievement(4.52), activity(4.45), time(4.35), model(4.24), variable(4.14), difference(4.12), character(4.11), semester(4.05), behavior(4.04) 등이 나타났다. 빈도수와 비교하여 model, semester, response, environment, university의 키워드가 우선 순위로 등장한 점이 특징이다. 이러한 결과를 토대로 살펴보면 국내 연구에서는 프로세스 데이터와 관련하여 교육기관 중 고등교육에 해당하는 대학의 시스템 환경에 관심이 많고, 학기, 강의별 학생의 학습 유형과 성취도 수준을 파악하는데 많은 관심을 보이는 것으로 판단된다.

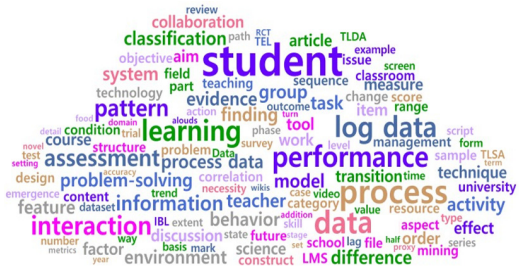
다음으로 국외 연구의 결과는 〈표 3〉와 같으며, 각각 단어의 빈도 기준 순위와 TF-IDF 기준 순위를 나타냈다. 상위 200개 키워드를 추출하여 워드 클라우드로 시각화한 결과는 [그림 5], [그림 6]와 같다. 상대적으로 많이 출현한 단어는 student(67), log data(27), performance(25), process(24), learning(20), pattern(20), group(20), data(18), process data(18), course(16) 등과 같았다. 국외 논문도 학생 키워드가 가장 많이 출현하였고, 다음으로 로그 데이터와 프로세스, 프로세스 데이터와 같은 키워드의 출현 빈도가 높았다. 또한, pattern이나 group의 키워드가 우선순위로 나타난 것으로 보아 프로세스 데이터를 활용한 학생의 학습 유형을 파악하는데 관심이 많은 것으로 보인다. 국내 연구에서는 빈도수를 기준으로 achievement, activity, character, type이 우선순위로 나타난 것과 비교하여 국외 논문에서는 performance, process, process data, problem-solving, IBL(Inquiry Based Learning)이 우선순위로 나타난 것으로 보아 학업성취 수준보다는 탐구 기반 학습이나 문제해결 과정에서 학생의 해결 과정에 대한 정보 및 수행 태도에 관심을 보이는 것으로 예측할 수 있다.

〈표 3〉 국외 연구의 단어 순위

단어 순위						TF-IDF를 통한 단어 순위					
	단어	빈도		단어	빈도		단어	TF-IDF		단어	TF-IDF
1	student	67	11	model	16	1	pattern	1.95	11	problem-solving	1.81
2	log data	27	12	transition	14	2	group	1.95	12	feature	1.78
3	performance	25	13	behavior	13	3	process data	1.92	13	task	1.78
4	process	24	14	difference	13	4	log data	1.92	14	interaction	1.78
5	learning	20	15	problem-solving	13	5	performance	1.89	15	transition	1.77
6	pattern	20	16	IBL	12	6	model	1.88	16	information	1.77
7	group	19	17	feature	12	7	course	1.83	17	activity	1.76
8	data	18	18	interaction	12	8	learning	1.83	18	assessment	1.75
9	process data	18	19	task	12	9	behavior	1.81	19	classification	1.74
10	course	16	20	LMS	11	10	difference	1.81	20	data	1.71



[그림 5] 빈도기반 워드클라우드



[그림 6] TF-IDF기반 워드클라우드

TF-IDF를 통한 단어의 중요도 순위를 살펴본 결과, pattern(1.95), group(1.95), process data(1.92), log data(1.92), performance(1.89), model(1.88), course(1.83), learning(1.83), behavior(1.81), difference(1.81) 등이 나타났다. 빈도수와 비교하여 course, information, activity, assessment, classification의 키워드가 우선순위로 등장한 점이 특징이다. 이러한 결과를 토대로 살펴보면 국외 연구에서는 평가 중 얻을 수 있는 학생의 활동 정보에 관심이 있고, 이를 통한 유형화 연구가 주로 이루어지고 있는 것으로 파악된다. 국내 연구에서는 TF-IDF를 기준으로 LMS, semester, university가 등장한 반면에 국외 연구에서는 problem-solving, task, assessment가 우선 순위로 나타난 것으로 보아, 국내 연구가 학습 설계에 초점이 맞춰져 있다면, 국외 연구에서는 학습보다는 평가, 과제 수행 및 문제해결 과정에 대한 관심이 더욱 높은 것으로 보인다.

2. 토픽모델링

토픽모델링은 확률적 생성 모델인 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 알고리즘을 사용하였다. LDA에서는 토픽의 수에 따라 분석 및 결과 해석이 달라지기 때문에 토픽 수를 결정하는 것이 매우 중요한 과정이다(유진은, 2021). 적절한 사전 모수를 선택하고 토픽 수를 선정하는데 다양한 방법이 있으나 본 연구에서는 선행연구를 토대로 클러스터 평가 방법인 silhouette 방법을 활용하여 최적의 토픽 수를 탐색하였다. silhouette 방법은 k-means 클러스터링에 필요한 통계적 모수를 비교하여 토픽의 독립성을 판단하는 방법이다. silhouette 계수가 1에 가까울수록 도출된 각 토픽들이 독립적으로 잘 구분되어 적합한 모델인 것을 의미한다(Panichella, et al., 2013).

선행연구에 따르면 분석에 필요한 모수에 대해서는 연구자와 연구 분야에 따라 사용하는 모수의 값이 다를 수 있고, 많은 연구에서 다양한 값들이 제시되고 있다(Naili, Chaibi & Ghézala, 2017). 최근에는 통계적 방법과 해석적인 방법을 함께 사용하는 추세이다. 본 연구에서도 통계적인 방법과 해석적인 방법을 함께 사용하여 효율적으로 토픽모델링을 수행하고 모델의 정확성을 확보하고자 하였다(이수상, 2016; 이정훈, 2019; Lu et al., 2011). 즉, 기존 연구 결과를 토대로 통계적 모수, 토픽 수, TF-IDF의 가능한 범위를 설정하여 범위 내의 수치들을 1,000회 반복하였고, silhouette 계수가 1에 가까운 수치를 유지하는 모델이면서 맥락적인 의미 해석이 가능한 모델을 최종 모델로 선택하고자 하였다. 그 결과는 다음의 〈표 4〉, 〈표 5〉와 같다. α=0.2, β=0.01을 모수로 적용하였고, 토픽의 수는 4개인 model 3이 최종 모델로 선정하였으며, 최종 모델의 silhouette 계수는 국내 연구는 0.781, 국외 연구는 0.742로 1과 가까우므로 적합한 모델이라고 할 수 있다.

〈표 4〉 국내 연구의 모델별 Silhouette 계수

Type	model #	TF-IDF	Topic	α	β	Silhouette
Word	3	0.1	4	0.2	0.01	0.781
Word	7	0.1	5	0.2	0.01	0.732
Word	11	0.1	6	0.2	0.01	0.661
Word	15	0.1	7	0.2	0.01	0.689

〈표 5〉 국외 연구의 모델별 Silhouette 계수

Type	model #	TF-IDF	Topic	α	β	Silhouette
Word	3	0.1	4	0.2	0.01	0.742
Word	7	0.1	5	0.2	0.01	0.708
Word	11	0.1	6	0.2	0.01	0.727
Word	15	0.1	7	0.2	0.01	0.660

또한, 토픽 구분이 적절히 이루어졌는지 판단하기 위해 토픽별 cosine 유사도 거리를 측정하여 확인할 수 있는데, 본 연구에서는 cosine 유사도 거리 적용 시, 단어의 출현 빈도와 중요도를 반영하고

자 자연어 처리에서 일반적으로 사용되는 Bag of Words(BoW) 모델에 TF-IDF 가중치를 사용하였다(김상도 외, 2009). 그 결과는 다음의 〈표 6〉과 같다. cosine 유사도 거리는 -1에서 1까지 나타나며 1에 가까울수록 토픽이 서로 유사하며, -1에 가까울수록 반대되는 토픽임을 의미한다. 본 연구 결과 국내 연구의 토픽별 cosine 유사도 거리는 최소 0.1076에서 최대 0.2272로, 국외 연구의 토픽별 cosine 유사도 거리는 최소 0.0502에서 최대 0.4087로 0에 가까운 것을 볼 수 있다. 따라서 토픽의 구분이 적합하게 이루어졌다고 볼 수 있다. 그 중에서 국외 연구의 토픽 중 topic1과 topic4는 다른 토픽과 비교하여 유사도가 다소 높은 것으로 나타났는데, 두 토픽의 주제 특성이 유사함을 알 수 있다.

〈표 6〉 토픽별 cosine 유사도 거리

	국내				국외			
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
Topic 1								
Topic 2	.1076				.0502			
Topic 3	.1222	.1560			.1393	.1130		
Topic 4	.1675	.1868	.2272		.4087	.1211	.1522	

다음의 〈표 7〉은 토픽별로 각 토픽에서 높은 빈도로 출현하는 단어와 토픽의 명칭을 나타냈다. LDA를 활용한 토픽모델링의 경우 도출된 토픽에 대해 포함된 단어를 포괄할 수 있는 주제명을 부여해야 한다. 토픽에 포함될 확률이 높은 논문을 토대로 해당 문서를 파악할 수 있다(유진은, 2021). 본 연구에서는 도출된 토픽에 포함된 단어와 각 토픽에 포함될 확률이 높은 논문으로 해당 토픽의 주요 주제를 파악하였으며, 이를 토대로 연구자 간 합의하여 토픽 주제를 명명하였다. 먼저, 국내 연구의 경우 topic 1의 키워드는 data, response, item, learning analytics, teacher, model, process, web, system, server, process data으로 나타났다. 이를 토대로 학습분석학과 관련된 프로세스 데이터 관련 연구들이 분류되었으며, topic 1을 '학습 분석'과 관련한 연구 주제인 것으로 명명하였다. 이와 같이 topic 2는 '교수학습 유형 분석', topic 3은 'AI학습(교과)', topic 4는 '학업성취'로 명명하였다.

〈표 7〉 토픽별 키워드

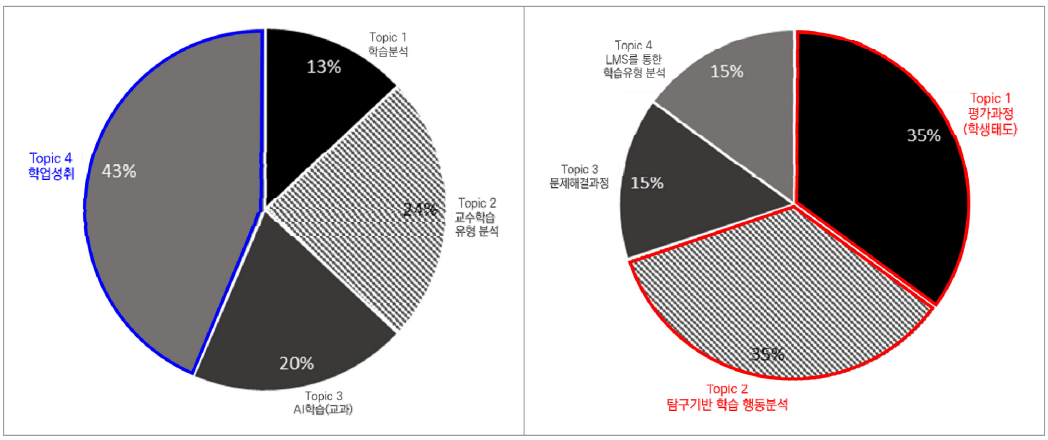
Topic		Topic key Word
Topic 1	학습 분석	data, response, item, learning analytics, teacher, model, process, web, system, server, process data
Topic 2	교수학습 유형 분석	character, type, time, variable, classification, system, faculty, number, addition, course, science
Topic 3	AI학습(교과)	robot, design, learning, pattern, AI, survey, program, purpose, disability, number, contact
Topic 4	학업성취	learning, achievement, log data, activity, course, data, time, group, behavior, semester, model

다음의 〈표 8〉은 국외 연구의 토픽별 키워드를 나타낸 결과이다. topic 1의 키워드는 student, process data, activity, assessment, classroom, teacher, fidelity, intervention, score, effect, engagement로 평가 과정 중 학생의 참여도, 기여도, 태도에 관심이 있는 연구들이 분류되었음을 알 수 있었으며, 이를 토대로 topic 1의 주제를 '평가 과정(학생 태도)'로 명명하였다. 이와 같이 topic 2는 '탐구기반 학습 행동분석', topic 3은 '문제해결 과정', topic 4는 'LMS를 통한 학습유형 분석'으로 명명하였다.

〈표 8〉 토픽별 키워드

Topic		Topic key Word
Topic 1	평가 과정 (학생 태도)	student, process data, activity, assessment, classroom, teacher, fidelity, intervention, score, effect, engagement
Topic 2	탐구기반 학습 행동분석	group, pattern, transition, difference, IBL, process, TLSA, discussion, condition, technique, resource,
Topic 3	문제해결 과정	process, problem-solving, behavior, task, data, item, performance, outcome, part, school, information
Topic 4	LMS를 통한 학습유형 분석	student, log data, performance, course, model, learning, feature, LMS, classification, interaction, data

국내외 연구의 각 토픽별 문서 비율을 시각화한 결과는 [그림 7]과 같다. 가장 문서의 비율이 높은 주제는 국내의 경우 43%를 차지한 topic 4-학업성취였다. 다음으로 topic 2-교수학습 유형분석이 24%, topic 3-AI학습(교과)이 20%, topic 1-학습분석이 13%로 나타났다. 국외의 경우에는 topic 1-평가과정(학생태도)와 topic 2-탐구기반 학습 행동분석의 주제가 각각 35%로 논문의 수가 가장 많았고, topic 3-문제해결 과정, topic 4-LMS를 통한 학습유형 분석이 각각 15%로 나타났다.



[그림 7] 토픽별 문서의 비율

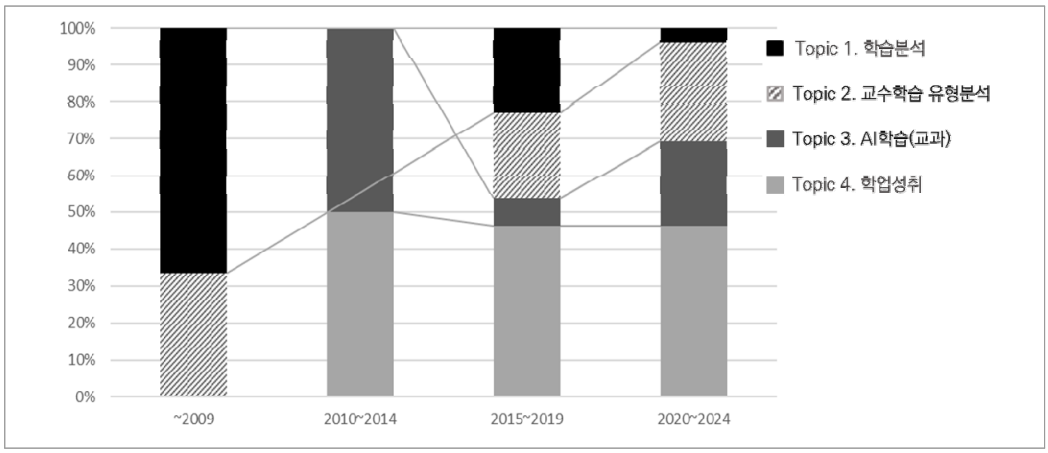
3. 토픽 트렌드 분석

토픽모델링 결과에 따라 연도별 논문의 출현 빈도를 살펴보았다. 연도별 트렌드를 살펴보기 위해 국내외 공통적으로 참여하고 있는 대표적인 국제 학업성취도 평가인 PISA에 컴퓨터 기반 검사가 전면 도입된 2015년을 기점으로 5개년씩 구분하여 연구 트렌드를 분석하고자 하였다. 그 결과는 다음의 〈표 9〉와 같다. 국내 연구에서는 2009년도 이전에는 topic 1-학습 분석과 topic 2-교수학습 유형 분석과 관련한 연구가 등장하였고, 2010년부터 2014년까지는 topic 3-AI학습(교과)과 topic 4-학업성취를 주제로 한 연구가 새롭게 등장하였다. 2015년도 이후에는 4개 토픽의 주제가 다양하게 등장하였으나 topic 4-학업성취와 관련된 연구가 가장 많이 나타났고, 2020년부터 현재까지 연구 논문의 수를 살펴보았을 때도 topic 4-학업성취의 연구 주제로 발간된 논문의 수가 가장 많았다. 다음으로는 topic 2-교수학습 유형 분석과 topic 3-AI학습(교과)이 많은 것으로 나타났다.

〈표 9〉 연도별/토픽별 논문출현 빈도

Topic		~2009	2010~2014	2015~2019	2020~2024	전체	
						N	비율(%)
Topic 1	학습 분석	2	0	3	1	6	13.0
Topic 2	교수학습 유형 분석	1	0	3	7	11	23.9
Topic 3	AI학습(교과)	0	2	1	6	9	19.6
Topic 4	학업성취	0	2	6	12	20	43.5
합계		3	4	13	26	46	100.0

이를 토대로 논문출현 변화 추이를 그래프로 표현한 결과는 다음의 [그림 8]과 같다. 2009년도 이전에 나타난 topic 1-학습 분석과 topic 2-교수학습 유형 분석의 주제는 2015년도 이후로 점차 감소되고 있으며, topic 3-AI학습(교과)은 증가하는 추세였다. 2010년도 이후로 등장한 topic 4-학업성취의 연구 주제는 현재까지 그 논문의 수와 비중이 계속 증가하고 있는 추세로 파악된다.



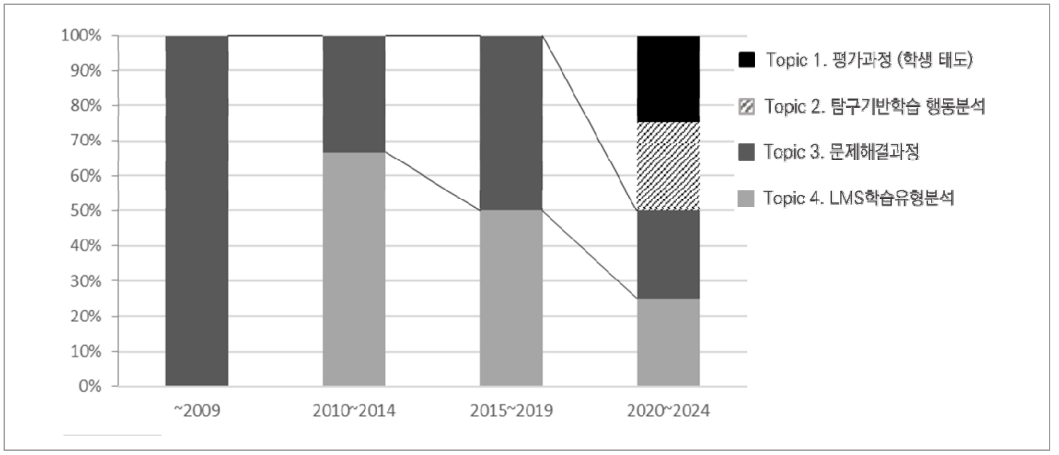
[그림 8] 연도별/토픽별 논문출현 변화 추이

국외 논문의 연도별 트렌드는 다음의 〈표 10〉과 같다. 국외 연구에서는 2009년도 이전에 topic 3-문제해결 과정과 관련된 연구가 1건 발간되었고, 2010년부터 2019년까지는 topic 3-문제해결 과정과 topic 4-LMS를 통한 학습유형 분석과 관련된 연구가 일부 발간되었음을 확인하였다. 2020년 이후에는 4개 토픽의 주제가 다양하게 등장하고 있다.

〈표 10〉 연도별/토픽별 논문출현 빈도

Topic		~2009	2010~2014	2015~2019	2020~2024	전체	
						N	비율(%)
Topic 1	평가 과정 (학생 태도)	0	0	0	3	3	15.0
Topic 2	탐구기반 학습 행동분석	0	0	0	3	3	15.0
Topic 3	문제해결 과정	1	1	2	3	7	35.0
Topic 4	LMS를 통한 학습유형 분석	0	2	2	3	7	35.0
합계		1	3	4	12	20	100.0

[그림 9]는 국외 연구의 논문출현 변화 추이를 그래프로 표현한 결과이다. 시간이 흐름에 따라 연구 주제가 다양하게 등장하고 있었으며, 매년 발간되는 논문의 숫자도 점차 증가하고 있는 것으로 파악된다. 특히 2020년 이후부터는 topic 1-평가과정(학생 태도), topic 2-탐구기반 학습 중 행동분석의 연구 주제가 새롭게 등장한 것이 특징이다.



[그림 9] 연도별/토픽별 논문출현 변화 추이

다음으로 시간의 흐름에 따라 프로세스 데이터 관련 연구 주제의 변화를 살펴보기 위해 시계열 회귀분석을 실시하였다. 토픽모델링 결과에서 연도별 토픽 비중을 기반으로 SPSS의 선형회귀분석을 활용할 수 있다. 토픽별 시계열 분석 결과는 다음의 〈표 11〉, 〈표 12〉와 같다. 시계열 분석에서 회귀계수가 통계적으로 유의하다고 할 때, 표준화된 회귀계수가 양수이면 연구 주제가 주목받고 있는 상향 토픽

(hot topic)이라고 볼 수 있다. 반면에 회귀계수가 음수이면 하향 토픽(cold topic), 회귀계수의 값이 유의하지 않으면 중립 토픽(Neutral topic)으로 구분한다(박종순, 김창식, 2019). 다만, 시계열 자료를 회귀 분석하려는 경우에는 이전 시점의 오차항에 의한 자기상관이 존재하기 때문에 잔차의 독립성을 검증하여 회귀 분석의 결과가 타당하다는 것을 확인하기 위해 Durbin-Watson 검사 결과를 함께 제시하였다. Durbin-Watson 검사 결과의 d 통계량은 0에서 4까지의 값을 갖는데, 2에 가까울수록 잔차의 독립성 가정을 만족한다고 판단할 수 있다. 국내 연구의 시계열 분석 결과는 1.754에서 2.762까지, 국외 연구의 경우에는 1.897과 2.519로 나타나 결과가 타당하다고 판단하였다.

〈표 11〉 국내 연구 토픽 시계열 분석

Topic		β	t	d	p	Hot/Cold
Topic 1	학습 분석	.300	.546	2.386	.623	
Topic 2	교수학습 유형 분석	.567	1.376	1.754	.241	
Topic 3	AI학습(교과)	.442	.852	2.762	.457	
Topic 4	학업성취	.683	2.288	1.811	0.062*	Hot

*p<.1

〈표 12〉 국외 연구 토픽 시계열 분석

Topic		β	t	d	p	Hot/Cold
Topic 1	평가 과정 (학생 태도)	-	-	-	-	
Topic 2	탐구기반 학습 행동분석	-	-	-	-	
Topic 3	문제해결 과정	.449	.871	1.897	.448	
Topic 4	LMS를 통한 학습유형 분석	.301	.705	2.519	.512	

〈표 11〉과 같이 국내 연구의 토픽을 시계열 분석한 결과를 살펴보면, 모든 토픽의 회귀계수가 양수로 나타났다. 그러나 topic 4-학업성취를 제외한 다른 토픽들은 통계적으로 유의하지 않아 중립 토픽으로 판단할 수 있다. topic 4-학업성취의 경우에는 유의수준 90%를 기준으로 회귀계수가 통계적으로 유의한 것으로 나타나 Hot topic인 것으로 나타났다. 국내 연구에서는 프로세스 데이터를 활용한 학업성취 연구가 2010년 이후 처음 등장한 이후로 꾸준히 증가하고 있으며 누적된 논문 발간 수도 가장 많은 것으로 나타나 국내에서는 학생의 성취 수준에 가장 많은 관심을 보이고 있는 것으로 판단할 수 있다.

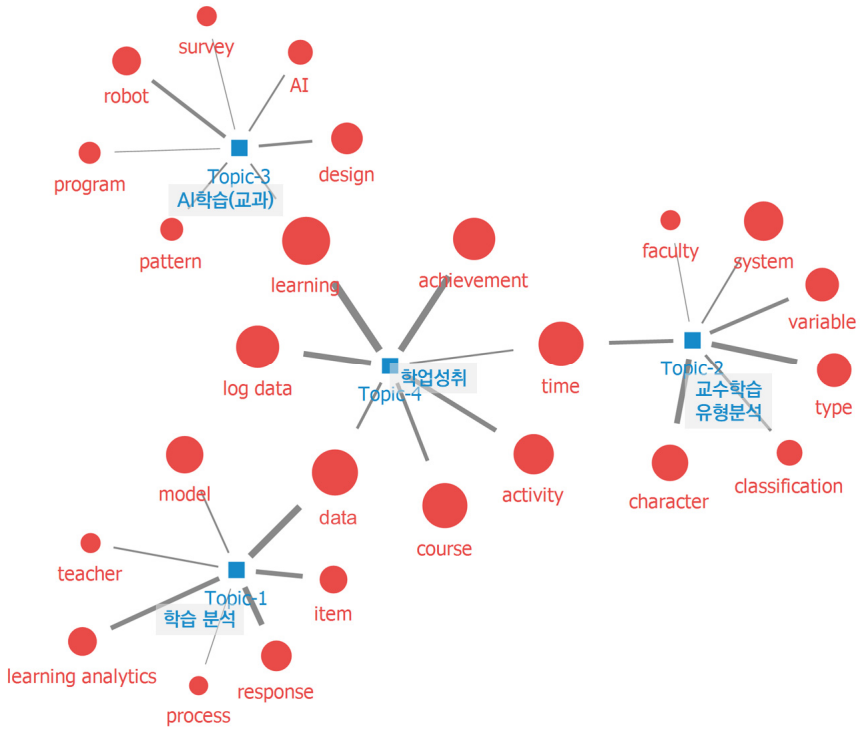
〈표 12〉는 국외 연구의 토픽을 시계열로 분석한 결과이다. topic 3-문제해결 과정과 topic 4-LMS를 통한 학습유형 분석과 관련된 연구는 회귀계수가 양수로 나타나 논문의 수가 점차 증가하지만, 통계적으로 유의하지 않기 때문에 중립 토픽으로 분류된다. 반면에 topic 1-평가과정(학생 태도)과 topic 2-탐구기반 학습 행동분석은 2020년 이후에 새롭게 등장하였지만, 이전에 발간된 논문의 수가 없기 때문에 회귀 분석 결과가 산출되지 않아 통계적 유의성을 판단할 수 없었다.

최근 국외 연구에서는 연구 주제별 비중이 비슷한 수준으로 다양한 연구가 수행되고 있으며, 특히

최근에는 평가 및 탐구기반 학습 과정 중 학생의 행동과 태도를 분석하기 위해 프로세스 데이터를 활용하는데 주로 관심이 있는 것으로 보인다.

4. 의미연결망 분석

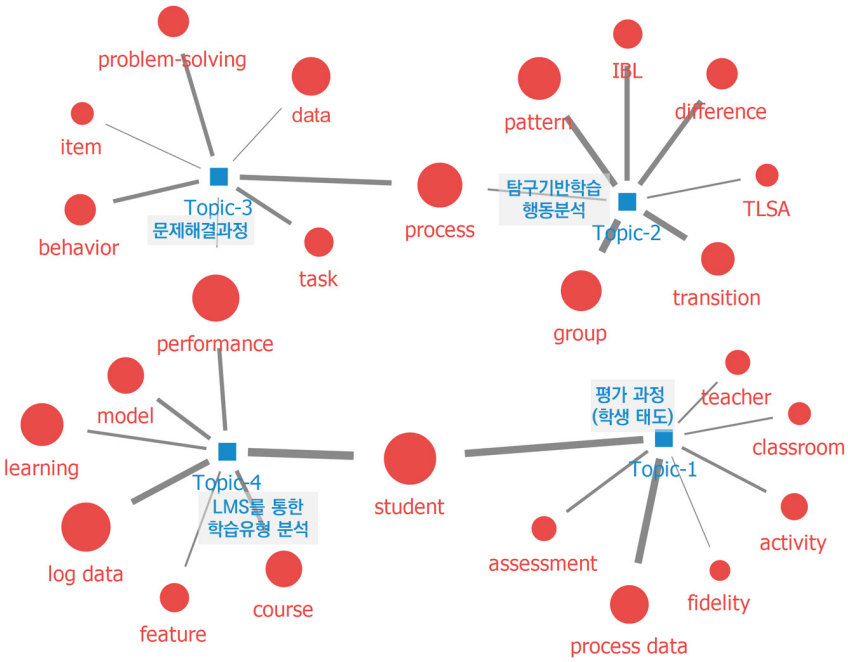
다음의 [그림 10], [그림 11]은 TF-IDF를 기반으로 핵심 키워드의 네트워크를 시각화한 결과이다. 이는 토픽 간 cosine 유사도 거리를 측정하여 네트워크 구성 및 군집화한 결과이다. 토픽 간 cosine 유사도 거리는 토픽모델링 결과를 평가하는데 보조적인 판단을 할 수 있지만, 이를 통해 토픽별 주요 키워드와 토픽 간의 관계를 유추할 수 있으며 관계성을 설명하기 위한 키워드도 확인할 수 있다는 장점이 있다. 토픽 간 cosine 유사도 거리 결과는 〈표 6〉에 제시하였다.



[그림 10] 토픽별 TF-IDF기반 키워드 네트워크

먼저, [그림 10]은 국내 연구의 키워드 네트워크이다. topic 4의 학업성취는 다른 주제들을 연결하는 핵심 토픽인 것으로 나타났다. topic 1의 학습분석과는 data, topic 2의 교수학습 유형 분석과는 time, topic 3의 AI학습(교과)과는 learning의 키워드와 긴밀하게 연결되어 있는 것을 볼 수 있다. 그 중에서도 〈표 6〉에 따르면 학업성취의 주제는 비교적 다른 토픽과 비교하여 topic 3의 AI학습(교과)과의 유사도가 다소 높은 것을 확인할 수 있었는데, 학습이라는 키워드를 공통으로 공유하면서 학업성

취 수준을 예측하고 이에 따른 학습을 설계하는 연구, 혹은 효과적으로 학업 성취 수준을 향상시키기 위한 AI프로그램 활용에 대한 연구들이 수행된 결과가 반영되었다고 볼 수 있다.



[그림 11] 토픽별 TF-IDF기반 키워드 네트워크

다음으로 국외 연구의 토픽별 키워드 네트워크를 시각화한 결과는 [그림 11]과 같다. 국내 연구의 키워드 네트워크이다. topic 2의 탐구기반 학습 중 행동분석 연구는 process를 키워드로 topic 3의 문제해결 과정 연구와 긴밀하게 연결되어 있었다. topic 3은 performance를 키워드를 topic 4의 LMS를 통한 학습유형 분석 연구와 공유하고 있었다. topic 4와 topic 1의 평가 과정(학생 태도)은 student를 공통된 키워드로 연결되어 있었는데 〈표 6〉에 따르면 cosine 유사도 거리가 가장 가까운 것으로 나타났다.

토픽별 주요 키워드를 분류하고 키워드의 중심성 정도를 〈표 13〉과 같이 작성하였다. 국내 연구에서 topic 1의 경우 data 키워드의 중심성이 .1842로 가장 높았으며, 다음으로 item, model, process, PISA, response, server, PSP, collection, tesk taker의 키워드 중심성이 높은 것으로 나타났다. topic 2는 type의 중심성이 .2609로 가장 높았으며, time, character, course, system, variable, addition, faculty, task, category의 순서로 나타났다. topic 3은 learning의 중심성이 .1905로 가장 높고, robot, design, pattern, survey, chatbot, game, system, AI, fantasy-type의 순서로 나타났으며, topic 4는 learning 키워드가 .3878로 가장 중심성이 높았으며, log data, achievement, activity, course, data, model, behavior, group, semester의 순서로 나타났다.

〈표 13〉 토픽별 주요 키워드 중심성

Topic	Topic 1		Topic 2		Topic 3		Topic 4	
	학습 분석		교수학습 유형 분석		AI학습(교과)		학업성취	
	키워드	중심성	키워드	중심성	키워드	중심성	키워드	중심성
1순위	data	.1842	type	.2609	learning	.1905	learning	.3878
2순위	item	.1579	time	.2391	robot	.1786	log data	.2449
3순위	model	.1184	character	.2283	design	.1429	achievement	.2347
4순위	process	.1184	course	.1848	pattern	.1190	activity	.1735
5순위	PISA	.1053	system	.1630	survey	.1190	course	.1735
6순위	response	.0789	variable	.1522	chatbot	.1071	data	.1735
7순위	server	.0789	addition	.1413	game	.1071	model	.1633
8순위	PSP	.0658	faculty	.1304	system	.1071	behavior	.1531
9순위	collection	.0658	task	.1304	AI	.0952	group	.1531
10순위	test taker	.0658	category	.1196	fantasy-type	.0952	semester	.1429

국외 연구의 토픽별 주요 키워드 중심성 정도는 〈표 14〉과 같다. topic 1의 경우 student 키워드의 중심성이 .4030으로 가장 높았으며, 다음으로 process data, fidelity, classroom, assessment, tablet, teacher, treatment, RCT, interaction의 키워드 중심성이 높은 것으로 나타났다. topic 2는 group의 중심성이 .2500으로 가장 높았으며, difference, student, TLSA, learning, performance, IBL, TLDA, transition, TEL이 차례로 나타났다. topic 3은 process의 중심성이 .2222였고, 다음으로 behavior, problem-solving, part, task, intervention, data, indicator, information, item이 나타났다. topic 4도 student의 키워드가 .3708로 중심성이 가장 높았으며, model, performance, log data, course, learning, LMS, data, feature, survey의 순서로 중심성이 높았다.

〈표 14〉 토픽별 주요 키워드 중심성

Topic	Topic 1		Topic 2		Topic 3		Topic 4	
	평가 과정 (학생 태도)		탐구기반 학습 행동분석		문제해결과정		LMS를 통한 학습유형 분석	
	키워드	중심성	키워드	중심성	키워드	중심성	키워드	중심성
1순위	student	.4030	group	.2500	process	.2222	student	.3708
2순위	process data	.1791	difference	.1667	behavior	.1728	model	.2135
3순위	fidelity	.1493	student	.1167	problem-solving	.1605	performance	.2135
4순위	classroom	.1343	TLSA	.1000	part	.1111	log data	.1910
5순위	assessment	.1194	learning	.1000	task	.1111	course	.1798
6순위	tablet	.1045	performance	.1000	intervention	.0988	learning	.1685
7순위	teacher	.1045	IBL	.0833	data	.0864	LMS	.1461
8순위	treatment	.1045	TLDA	.0833	indicator	.0741	data	.1461
9순위	RCT	.0896	transition	.0833	information	.0741	feature	.1124
10순위	interaction	.0896	TEL	.0667	item	.0741	survey	.1124

V. 결론 및 논의

본 연구에서는 프로세스 데이터를 키워드로 과거부터 현재까지 발간된 연구의 동향을 살펴보았다. 이를 위해 국내외 총 66편의 영문 초록을 수집하여 분석 대상으로 선정하고, 연구 주제 추출 및 트렌드 분석을 위해 텍스트 마이닝 연구 기법을 활용하고, 토픽모델링과 의미연결망 분석을 적용하였다. 본 연구의 결과는 다음과 같다.

첫째, 프로세스 데이터 연구의 주요 키워드는 국내 연구의 경우 빈도수와 TF-IDF를 기준으로 하였을 때, 공통적으로 student, learning, time, log data, achievement, 국외 연구에서는 student, log data, assessment, process data, task 등이 주요 키워드로 나타났다. 국내 연구에서는 학생과 학습, 로그 데이터와 더불어 학생의 성취 키워드가 있었다면, 국외 연구에서는 학생, 로그 데이터 뿐만 아니라 과정적 정보를 담는 프로세스 키워드가 나타난 것이 특징이었고, 국내 연구와는 달리 평가과 과제 키워드가 등장하여, 국내에서는 평가 결과에 해당하는 성취 수준에 관심이 있다면, 데이터 수집 및 활용 분야가 평가 및 과제 수행 과정에 초점이 있는 것으로 보인다.

둘째, 프로세스 데이터 연구에서 추출된 키워드를 중심으로 토픽모델링을 실시한 결과, 총 4개의 토픽이 각각 도출되었다. 국내는 '학습분석', '교수학습 유형 분석', 'AI학습(교과)', '학업성취', 국외는 '평가과정(학생태도)', '탐구기반 학습 행동분석', '문제해결 과정', '학습유형 분석'으로 나타났다. 국내에서는 학생의 성취도를 예측하기 위한 데이터 분석과 학습 설계에 프로세스 데이터를 활용한 연구가 많은 것으로 나타났으며, 선행연구 결과와 마찬가지로 LMS를 통한 학습과정을 객관화하고, 교육적 데이터 마이닝을 통한 학습 처방, 학습 성과 통제에 주된 관심을 보이는 것을 확인할 수 있었다(안미리외, 2016). 국외에서는 프로세스 데이터를 활용한 과정 중심의 평가 관련한 연구가 주로 이루어진 것으로 판단된다. 또한 주로 응답시간을 활용한 인지 능력 수준에 대한 추정의 정확도, 검사의 타당화, 정의적 특성과 인지적 과정을 이해하는 교육연구가 증가하는 추세임을 추가로 확인할 수 있었다(이소라, 2019; 함은혜, 2022; kyllonen& Zu, 2016).

즉, 이와 같은 결과를 통해 국내외 연구에서는 프로세스 데이터를 바라보는 관점에 차이가 있었음을 확인하였다. 국내에서는 learning analytics(학습 분석학), faculty, teacher(교수, 교사), achievement(성취) 등 성취도, 교사의 수업설계와 교수학습 유형 분석에 관심을 보이며, 프로세스 데이터를 통한 학습분석학적 접근이 주로 이루어졌다. 반면 국외에서는 fidelity(충실도), engagement(참여), assessment(평가), task(과제) 등 평가 및 과제수행 중 학생의 참여 태도와 문제풀이 과정의 연속적 행동 패턴에 관심을 보이는 것으로 보아, 교육 평가적 접근 연구가 주로 이루어졌다고 평가할 수 있다.

또한, 프로세스 데이터와 로그 데이터를 중심으로 토픽들을 살펴보았을 때, 국내에서는 프로세스 데이터의 키워드가 topic 1의 학습분석 주제에 등장하고, web, system, server와 같은 단어들이 함께 등장했고, 로그 데이터는 topic 4의 학업성취의 주제에 등장하였다. 즉, 국내에서는 프로세스 데이터 정보 수집을 위한 시스템과 서버, 교사의 교수학습 등 방법적인 측면에 관심이 있었으며, 로그 데이

터와 관련해서는 학생 활동과 이에 따른 학업성취도 예측에 관심이 있는 것으로 보인다. 국외에서는 프로세스 데이터의 키워드가 topic 1의 평가 과정 중 학생의 태도에 나타났으며, fidelity, intervention, engagement와 같은 키워드가 함께 등장한 것으로 보아 학생의 문제해결 과정에서 얻을 수 있는 참여도, 태도 정보에 관심을 갖고 있는 것으로 보인다. 로그 데이터는 topic 4의 LMS를 통한 학습유형 분석 연구에서 나타났으며, 주요 키워드로 등장한 interaction, performance, feature 등을 통해 학습자 간 상호작용과 학습유형 분석에 관심이 있는 것으로 판단할 수 있다.

셋째, 프로세스 데이터 연구의 트렌드는 국내외 다른 양상을 보였다. 국내에서는 최근에는 학업성취도(topic4)와 관련된 연구가 가장 활발히 이루어지고 있으며, 비중이 점차 증가하는 추세로 hot topic으로 판단할 수 있었다. 학습유형 분석(topic2)과 AI학습(교과)(topic3) 연구의 수가 다소 증가하였으나 통계적으로 유의한 변화는 아니었다. 국외에서는 평가 과정 중 학생의 태도 분석(topic1)과 탐구기반 학습 중 행동유형 분석(topic2) 연구가 2020년 이후 새로 등장하여 주목받는 연구 주제로 판단할 수 있지만, 이전에 발간된 논문이 없어 통계적으로 유의성을 판단하기 어렵다.

넷째, 프로세스 데이터 연구의 토픽별 관계를 살펴보면, 주요 키워드를 중심으로 토픽이 연결되어 있었는데 국내에서는 learning, data, time, 국외에서는 student, process, performance의 키워드가 토픽을 서로 연결하는 핵심 키워드였다. 특히 국내에서는 학습을 의미하는 learning, 국외에서는 student 키워드가 토픽 간 긴밀하게 연결하고 중심성 정도가 가장 높은 키워드인 것으로 나타났다. 이러한 결과로 미루어 보아 국내에서는 교수학습과 교육 콘텐츠 설계 등 학습의 방법적인 측면에 관심을 가지고 프로세스 데이터를 활용하고자 한다면, 국외에서는 학생의 행동 분석과 문제해결 과정을 이해하기 위한 정보로서 학생의 행동 중심 프로세스 데이터에 관심이 있는 것으로 판단된다.

본 연구에서는 최근 주목받고 있는 프로세스 데이터를 키워드로 하여 국내외 연구 동향을 비교 분석하였다. 이를 토대로 한 연구의 제언은 다음과 같다.

첫째, 프로세스 데이터를 활용한 다양한 분야의 연구가 필요하다. 국내에서는 프로세스 데이터를 활용한 연구가 점차 증가하고 있으나, 로그 데이터를 활용한 학습분석학 연구에 집중되어 있었다. 최근 액션 시퀀스로서의 프로세스 데이터를 활용한 연구가 등장하고 있으나 여전히 학술적 연구는 부족하고 특정 연구 분야에 편중되어 있는 모습을 볼 수 있었다. 2022 개정 교육과정에서 과정 중심 평가로의 체제 변환을 강조하고 있는 것에 비해 관련 연구가 미비한 실정이다. 학생의 문제해결 상황에서 얻을 수 있는 연속적 행동 정보가 담긴 프로세스 데이터를 구축하여 이를 활용한 학술적 관심과 다양한 연구가 활발하게 진행되길 기대한다.

둘째, 프로세스 데이터를 수집하고 제공할 수 있는 연구의 기반이 마련되어야 한다. 국내에서는 대부분 적은 규모의 LMS를 통한 로그 데이터 수집, 적은 양의 데이터를 활용한 학습 효과 연구가 대부분이다(안미리 외, 2016). 그 중에서도 자체 LMS를 갖추고, 온라인 학습 활동이 활발하게 이루어진 대학 기관과 관련된 연구가 주로 이루어지고 있어, 그 외에 연구는 다소 부족한 것으로 보인다. 즉, 활용할 수 있는 프로세스 데이터의 한계로 연구 분야 및 대상에 제한적인 상황인 것이다. 따라서 학생의 문제해결 과정을 이해하기 위해, 국가 수준에서 주기적으로 수집하여 연구를 목적으로 제공할 수 있는 대규모 프로세스 데이터가 필요하다고 보았다. 특히 국외에서는 인지능력 추정의 정확도, 검사의 타당화, 정의적 특성과 인지적 과정을 이해하는 연구 등 다양한 분야에서 활용하고 있는 만큼 장기적으로

국내 학생들의 특성을 설명할 수 있는 국가 수준 컴퓨터 기반 검사와 프로세스 데이터의 수집 및 관리 체계를 마련할 필요가 있다.

본 연구는 온라인 학습과 컴퓨터 기반 검사 및 컴퓨터 적응검사가 확대되고 있는 현 시점에 프로세스 데이터를 활용한 국제적 연구 트렌드를 분석하고 관련 연구의 방향성을 제시하였다는데 의의가 있다. 특히 기존의 연구동향 연구와 비교해서 프로세스 데이터와 관련된 연구가 상대적으로 미비하며, 본 연구를 통해 프로세스 데이터의 용어 및 개념에 대한 이해를 높이고 관련 연구의 활성화를 위한 구체적인 주제를 제안할 수 있었다는 점에서 선행연구와의 차별성이 있다.

그럼에도 불구하고 국내외 프로세스 데이터를 키워드로 현재까지 발간된 연구의 수가 적어 분석 대상의 수가 많지 않다는 점에서 본 연구의 결과를 일반화하여 비교하는데 연구의 한계가 있을 수 있다. 후속 연구에서는 프로세스 데이터 관련 연구 분야의 키워드를 포함하거나 또는 학술지, 학위논문, 연구보고서 등으로 연구 대상을 확대하여 분석한다면 이러한 한계점을 개선할 수 있을 것으로 보인다. 특히 본 연구에서는 연구 대상 키워드를 프로세스 데이터와 로그 데이터로 연구의 범위를 한정하였는데, 프로세스 수행 기록에 따른 데이터를 의미하는 용어가 국내외 연구에서 분야에 따라 다양하게 사용될 수 있으므로 연구 대상을 선정함에 있어서 키워드 뿐만 아니라 연구 주제와 분야를 함께 고려할 필요가 있을 것이다. 또한, 향후 관련 연구가 다수 축적된다면 프로세스 데이터와 로그 데이터의 개념에 따라 각각의 연구 트렌드를 구분하여 분석하고 비교한다면 심층적 동향 분석이 가능할 것으로 기대한다.

참고문헌

강주연, 이이든, 김지수(2020). 텍스트 마이닝을 활용한 'Z세대' 관련 뉴스데이터 의미연결망 분석. **미래청소년학회지, 17**(2), 25-48.

교육부(2017). **2015 개정 교육과정 총론 해설: 고등학교.** 세종: 교육부.

교육부(2020). **2022 개정 교육과정 총론 해설: 고등학교.** 세종: 교육부.

김기옥(2020). 텍스트마이닝을 활용한 소비자학 연구 동향 분석. **소비자학연구, 31**(5), 19-47.

김대영(2021). 키워드 네트워크 분석을 통한 2010년대「교육 과정 연구」의 동향 분석. **교육방법연구, 33**(2), 271-292.

김상도, 윤희근, 박성배, 박세영, 이상조(2009). 문자열 커널을 이용한 인터넷 영화평의 감정 분석. **한국정보과학회 언어공학연구회: 학술대회 논문집**, 2009.10a, 56-60.

김연후(2023). 컴퓨터 기반 평가에서 프로세스 데이터를 활용한 피험자의 문제해결 행동 분석. 박사학위논문, 서울대학교.

김영수, 허희옥, 김현진, 계보경, 박연정, 김영희, 이현영, 두민영(2016). **교수 메시지 설계: 교육용 자료 제작 원리.** 서울: 교육과학사.

김용희, 한창근(2019). '수저 계급' 관련 웹 뉴스 기사에 대한 의미연결망 분석. **한국사회복지학, 71**(3), 55-81.

박종순, 김창식(2019). 빅데이터 연구동향 분석: 토픽모델링을 중심으로. **디지털산업정보학회 논문지, 15**(1), 1-7.

백영민(2020). **R를 이용한 텍스트 마이닝(개정판).** 서울: 한울아카데미.

송미영, 김성숙, 구자옥, 임혜미, 박혜영, 한정아, 손수경(2014). **OECD 국제 학업성취도 평가 연구: PISA 2012 컴퓨터 기반 평가 결과 분석.** 한국교육과정평가원. RRE 2014-4-2.

송민정(2013). 빅데이터(Big Data)를 활용한 비즈니스모델 혁신. **과학기술정책, 19**(2), 86-97.

안미리, 최윤영, 배윤희, 고윤미, 김민하(2016). 학습분석학 국내 문헌 고찰: 로그 데이터를 이용한 실증연구를 중심으로. **공학교육연구, 32**(2), 253-291.

오창우(2017). 한국에서 사회갈등 논의의 의미연결망 분석: 주요 포탈에서의 핵심어 간 네트워크를 중심으로. **정치커뮤니케이션연구, 45**(-), 37-67.

유예림 (2017). 빅데이터 분석 기법을 활용한 2015 개정 교육과정 정책에 대한 언론보도 분석. 박사학위 논문, 서울대학교.

유진은(2021). **AI시대, 빅데이터 분석과 기계학습.** 서울: 학지사.

윤희진(2020). 텍스트마이닝을 활용한 다문화 멘토링 관련 연구 동향 분석. **문화교류와 다문화 교**

육, **9**(1), 27-50.

이성직, 김한준(2009). TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법. **한국전자거래학회지, 14**(4), 59-73.

이소라(2019). 비대칭문항반응모형의 문항복잡성 타당화 가능성 연구: PISA 2012 프로세스 데이터의 응답 시간·횟수 변인을 중심으로. **한국교육학연구, 25**(3), 113-134.

이수상(2016). 독후감 텍스트의 토픽모델링 적용에 관한 탐색적 연구, **한국도서관정보학회지, 47**(4), 1-18.

이신영(2018). 평생교육학 연구의 주제어 연결망 분석과 지식구조 탐색. 박사학위논문, 동의대학교.

이예은, 장정현(2022). 비정형 데이터 수집과 TF-IDF를 통한 아동학대 분석 및 키워드 추출. **한국범죄심리연구, 18**(4), 171-182.

이정훈(2019). 2019년 강원도 화재 보도에 대한 언어망 분석: 미디어 의제 분석을 중심으로, **한국콘텐츠학회논문지, 19**(11), 153-167.

장은아, 정혜원(2024). 텍스트마이닝을 활용한 직업계 고등학교 연구동향 분석: 2010년 2023년까지. **교육학연구, 62**(4), 119-154.

전성균, 상경아(2023). ICILS 2018 프로세스 데이터를 활용한 문항 응답 행동 특성 분석. **컴퓨터교육학회 논문지, 26**(3), 1-13.

조용상, Abel, J., 유재택, 신성욱(2013). **표준화 이슈리포트: 학습분석 기술 활용 가능성 및 전망**. 서울: 한국교육학술정보원. RM 2013-15.

지미선(2018). 의미연결망 분석을 통한 에니어그램 키워드 중심 국내 학술지 연구 동향 분석. **에니어그램연구, 15**(2), 39-60.

최진수, 정혜원(2022). 토픽모델링과 의미연결망 분석을 활용한 영재교육 연구 동향 분석. **교육학연구, 60**(4), 1-28.

최훈원, 최윤정(2024). 키워트 네트워크 분석을 활용한 과정중심평가의 연구동향 분석. **교육과정평가연구, 27**(2), 251-277.

함은혜(2022). 프로세스데이터를 활용한 과제참여도 측정가능성 탐색. **교육평가연구, 35**(1), 23-48.

Bergner, Y., & von Davier, A. A. (2019). Process Data in NAEP: Past, Present, and Future. *Journal of Educational and Behavioral Statistics, 44*(6), 706-732.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine *Learning Research, 3*(-), 993-1022.

Fang, G., & Ying, Z. (2020). Latent Theme Dictionary Model for Finding Co-occurrent

Patterns in Process Data. *Psychometrika, 85*(3), 775-811.

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G., & Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. *Large-scale Assessments in Education, 9*(1), 1-31.

Kyllonen. P. C., & Zu. J. (2016). Use of Response Time for Measuring Cognitive Ability. *Journal of Intelligence, 4*(4), 1-29.

Levy, R. (2020). Implications of considering Response Process Data for Greater and Lesser Psychometrics, *Educational Assessment, 25*(3), 218-235.

Y. Lu, M. Qiaozhu & Z. Cheng Xiang. (2011). Investing task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Informaion Retrieval, 14*(2), 178-203.

M. Naili, A. H. Chaibi, & H. B. Ghezala. (2017). Arabic topic identification based on empirical studies of topic models. *ARIMA Journal, 27*(-), 45-59.

Oranje, A. , Gorin, J. , Jia, Y. , & Kerr, D. (2017). *Collecting, analyzing, and interpreting response time, eye-tracking, and log data*. In K.Ercikan & J. W.Pellegrino (Eds.), Validation of score meaning for the next generation of assessments.

A. Panichella, B. Dit, R. Oliveto, M. D. Penta, D. Poshynanyk, & A. D. Lucia. (2013). *How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms.* 2013 35th International Conference on Software Engineering (ICSE).

Provasnik, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest?. *Large-scale Assessments in Education, 9*(1), 1-17.

Zhan, P., & Qiao, X. (2022). Diagnostic Classification Analysis of Problem-Solving Competence using Process Data: An Item Expansion Method. *Psychometrika, 87*(4), 1529-1547.

• 논문접수 : 2024.07.05. / 수정본접수 : 2024.07.31. / 게재승인 : 2024.08.12.

ABSTRACT

Analysis of Research Trends in Process Data using Text Mining

Jinsu Choi
Ph.D. candidate, Chungnam National University

Hyewon Chung
Professor, Chungnam National University

The purpose in this study is to analysis the research trends for process data in the field of pedagogy. To do this, these were subjects of the study that a total of 60 Korean and foreign research papers. And topic modeling based on the Latent Dirichlet Allocation(LDA) algorithm and semantic network analysis(SNA) methods were used for this study. The results are as follows. First, the research on process data were divided into each four topics: Korean research-'learning analysis', 'teaching and learning type analysis', 'AI learning', 'academic achievement', foreign research-'evaluation process (student attitude)', 'inquiry-based learning behavior analysis', 'problem-solving process', and 'learning type analysis using LMS'. Second, as a result of analysis the trend of the topics in time, 'academic achievement' topic was a hot topic in Korea, it is hot in foreign, 'evaluation process (student attitude)' and 'inquiry-based learning behavior analysis'. Third, the topics of process data research were linked to the keywords: Korean research-learning, data, time, foreign research- student, process, performance. Specially, 'learning' was found to be an important keyword connecting the most topics in Korea, while it is 'student' in foreign. Based on these results, we can use them as basic data for future research direction and policy establishment.

Key Words: *Process Data, Log Data, Research Trends, Text Mining*