

## 토픽모델링을 활용한 실감미디어 분야 국가 R&D 과제 연구 동향 분석

서 효 기<sup>1</sup> · 김 영 준<sup>2\*</sup><sup>1</sup>고려대학교 기술경영전문대학원 박사수로<sup>2</sup>고려대학교 기술경영전문대학원 교수

# Research Trend Analysis of National R&D Projects in Immersive Media Using Topic Modeling

Hyo-Gi Seo<sup>1</sup> · Youngjun Kim<sup>2\*</sup><sup>1</sup>Ph.D. Candidate, Graduate School of Management of Technology, Korea University, Seoul 02841, Korea<sup>2</sup>Professor, Graduate School of Management of Technology, Korea University, Seoul 02841, Korea

### [요 약]

최근 실감미디어 산업의 부상에 따라 국내에서도 국가 전략과 육성 체계가 마련되었다. 정책의 복잡성이 증가함에 따라 데이터 분석 기반 정책 수립의 중요성은 더욱 커지고 있으나 실감미디어 분야의 국가 R&D 동향을 분석하는 연구는 부족한 실정이다. 이에 본 연구는 2002년부터 2023년까지 수행된 18,504개의 NTIS 국가 R&D 과제를 웹 스크래핑 기술을 이용하여 수집하여 연구 현황을 살펴보았다. 수집된 과제에서 독립된 단어 19,503개를 추출하고 토픽모델링 분석방법을 이용하여 주된 연구 방향을 식별하는 분석을 진행하였다. 실감미디어 관련 R&D 과제에서 도출된 주요 토픽 9개는 다음과 같다. 사용자 경험기반 실감 콘텐츠, 공간데이터 기반 환경 모델링, 통신 및 에너지 전송시스템, 디지털 혁신과 교육 서비스, 소자, 재료 및 공정 등 기반기술, 원격 및 자동화 시스템, 플랫폼 설계 및 유지관리, 영상 데이터 분석 및 처리, 영상기반 의료 서비스로 각 토픽에 적합한 주제를 부여하였다.

### [Abstract]

With the rise of the immersive media industry, South Korea has established national strategies and support systems. As policy complexity grows, data-driven policy development has become increasingly important. However, research diagnosing national R&D trends in immersive media remains limited. This study applied topic modeling analysis to identify key research directions. Through web-scraping, data from 18,504 projects conducted between 2002 and 2023 were collected from the NTIS National R&D database, yielding a total of 19,503 unique words. The analysis identified nine major R&D topics related to immersive media, including user experience-based immersive content, spatial data-driven environment modeling, communication and energy transmission systems, digital innovation and education services, foundational technologies (such as devices, materials, and processes), remote and automated systems, platform design and maintenance, image data analysis and processing, and image-based medical services.

**색인어** : 실감미디어, 토픽모델링, 텍스트마ining, 국가과학기술지식정보서비스, 연구 동향분석**Keyword** : Immersive Media, Topic Modeling, Text Mining, NTIS, Research Trend Analysis<http://dx.doi.org/10.9728/dcs.2024.25.11.3409>

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 31 August 2024; Revised 02 October 2024

Accepted 18 October 2024

**\*Corresponding Author; Youngjun Kim**

Tel: +82-2-3290-4872

E-mail: [youngjkim@korea.ac.kr](mailto:youngjkim@korea.ac.kr)

## 1. 서론

### 1-1 연구의 배경 및 필요성

실감미디어(Immersive Media)는 인간의 감각기관을 통해 몰입감과 현장감을 제공하는 미디어로, 가상현실(VR), 증강현실(AR), 혼합현실(MR) 등 다양한 형태로 구현되는 차세대 디지털 미디어며 최근 실감 미디어 산업이 빠르게 부상하고 있다. 가상환경을 체계적으로 이해하기 위한 개념적인 틀의 구성요소를 제안한 연구[1], 미디어의 진화와 개념 확장과 관련된[2] 및 메타버스의 정의와 의미에 대한 분석[3] 등을 통해 관련한 선행연구들이 활발히 진행되고 있다. 이러한 기술들은 엔터테인먼트, 교육, 의료, 군사 등 다방면의 산업에서 혁신을 이끌고 있다[4],[5]. 다수의 국가와 글로벌 기업들은 실감미디어 기술의 발전을 기업 및 국가의 핵심역량으로 설정하고[6], 이를 지원하기 위한 다양한 정책과 육성 체계를 마련하고 있다. 국내에서도 실감미디어산업의 중요성이 강조되고 있는 실정에 따라 관련 R&D 투자 확대 및 연구개발의 중장기적인 전략 수립을 위한 방안을 마련하고 있다[7].

그러나 실감미디어 분야의 연구개발 현황을 체계적으로 분석한 연구는 수행되지 않았다. 특히 정책의 복잡성과 기술의 다변화로 인해 데이터 기반의 정책수립이 필수적임에도 불구하고 국가 R&D 동향을 종합적으로 분석하고 진단한 연구의 부족은 국가적 전략 수립과 관련된 의사결정 과정에서 타당한 방향성을 제안하는 데 어려움이 존재한다. 본 연구는 방대한 데이터를 분석하는데 활용성이 높은 텍스트마이닝 기법을 사용하여 연구의 동향을 분석하였고, 실증적인 연구인 국가과제를 분석의 대상으로 하여 실감미디어 연구 분야에 처음으로 적용하였다는 점에서 큰 의미가 있다.

### 1-2 연구의 목적 및 방법

국가과학기술지식정보서비스(NTIS, National Science & Technology Information Service, 이하 NTIS)는 국가 R&D 기획에서 성과 활용까지 연구 및 개발하는 사람뿐 아니라 과학기술 지식정보에 관심 있는 모든 국민에게 열려있는 개방형 국가과학기술 지식정보 허브이다. NTIS 포털에는 사업, 과제, 인력 성과 등 국가 R&D 사업에 대한 정보와 논문, 특허, 연구과제, 정책 동향 등 과학기술 지식정보를 통합 서비스하는 국가과학기술 지식정보 포털이기에 국가 R&D 과제 분석에 중요한 역할을 할 수 있다[8].

특히나 논문 분석이 기술 발전의 정도와 연구 성과 및 흐름을 파악하는 데 유용한 방법이나[9],[10], 상세한 연구 핵심 내용은 명확히 포함하지 않는 경우가 많고 실질적인 기술의 상용화나 시장 반응에 대한 정보는 부족하다[11]. 반면 국가 R&D 과제데이터는 특정 기술이나 산업에 대한 국가의 전략적 방향성과 자원 배분 현황을 명확하게 드러내며, 이는 단순

히 개별 기술의 발전 흐름을 이해하는 것을 넘어, 국가 차원의 연구개발 전략과 정책적 목표를 이해하는 데 필수적인 정보를 제공한다[12]. 특히, 국가 R&D 과제는 연구개발의 초기 단계부터 최종 성과까지 전 과정을 포함하고 있어, 특정 기술의 연구 동향뿐만 아니라 해당 연구 분야의 단계별 연구 동향을 파악할 수 있으며[13], 전반적인 발전 경로와 정책적 방향성과 유망 기술탐색에 중요한 역할을 한다[14].

또한, 정보화 시대의 도래와 함께 디지털 데이터의 양이 폭발적으로 증가하고 있음에 따라 이러한 대량의 비정형 데이터들을 분석하고 유의미한 정보를 추출하는데 핵심적인 역할을 하는 텍스트마이닝 기법이 활발히 연구되고 있다[15]. 특히, 연구 동향분석에 있어 텍스트마이닝은 데이터를 기반으로 특정 분야의 연구주제, 핵심 키워드, 연구 발전의 흐름 등을 파악하기 위해 유용하게 사용된다. 이를 통해 연구자들은 최신 연구 동향을 신속하게 파악하고 연구 방향을 설정하는데 중요한 자료로 활용할 수 있다. 따라서 실감미디어 분야의 국가 R&D 과제를 텍스트 마이닝 기법을 사용하여 분석함으로써, 현재 산업의 상태뿐만 아니라 미래 방향성까지 종합적으로 진단할 수 있다. 이러한 분석은 정책 수립과 전략적 결정에서 필수적인 기초 자료로 활용될 수 있으며, 향후 연구개발 방향을 설정하는 데 있어서 중요한 역할을 할 것으로 기대할 수 있다.

본 연구는 이러한 목적과 배경에서 출발하였으며 NTIS 데이터 포털에서 수집 가능한 전체 기간인 2002년부터 2023년까지 22년간 수행된 20,209개의 실감미디어 관련 국가 R&D 과제를 웹 스크래핑 기술을 이용하여 수집하여 실감미디어 분야의 연구 현황을 분석하고자 한다. 본 연구는 토픽모델링의 대표적인 알고리즘인 LDA 기법을 사용하여 토픽을 추출하였고 파악한 주요 단어를 통해 주제를 유추할 수 있는 보조 지표로서 주요 단어들과 유사한 단어들을 벡터의 형태로 수치화할 수 있는 Word2Vec 모델을 이용하였다. Coherence Score, Held-out Likelihood 와 t-SNE 시각화 분석 결과를 바탕으로 토픽의 수를 결정하였으며 실감미디어 분야의 주요 연구주제를 보다 명확하게 분석하고 국가 R&D 동향을 종합적으로 진단하고자 한다. 이를 통해 실감미디어산업의 미래 방향성을 제시하고, 향후 연구 및 정책 수립을 위한 기초 자료를 제공할 수 있을 것으로 기대한다.

## II. 이론적 배경

### 2-1 실감미디어 연구 동향

실감미디어산업은 새로운 개념과 기술의 도입에 따라 다양한 연구가 진행되고 있다. 산업의 초기에는 주로 가상 및 증강현실 기술의 기술적 구현에 초점이 맞춰져 있었으나, 최근 보다 광범위한 영역에서의 연구가 활발히 진행되고 있으며 다양한 방식으로 선행연구가 진행되었다. 실감미디어의 몰입

감을 높이기 위한 선행연구로는, 새로운 영상 매체에 적합한 콘텐츠 제작 기법들에 대한 변화가 시도되고 있다[16],[17]. 시각뿐만 아니라 청각과 후각과 같은 다양한 감각 요소가 사용자에게 보다 나은 현실감을 전달할 수 있다. 그에 따라 구체적인 청각, 후각 정보가 없는 영상에도 시각 외의 다른 감각의 정보를 제공하는 방법에 관한 연구가 제안되었다[18]. 사용자와 미디어 간의 상호작용은 실감 경험의 질을 결정하는 중요한 요소 중 하나이며 이러한 상호작용을 통해 양방향 소통이 원활 해지고, 몰입이 확장되어 양방향 소통을 통한 새로운 경험을 할 수 있다. 이러한 기존의 방식과 차별화된 능동적인 재해석을 경험하는 순간 인지가 변화하는 경험을 할 수 있다[19],[20]. 이뿐만 아니라 다양한 방식의 입력 장치와 관련한 연구가 진행되고 있고 특히 사용자가 가장 편하게 사용할 수 있는 신체를 이용한 제스처 인식 방식의 인터페이스 적용과 관련한 많은 연구가 진행되고 있다[21].

### 2-2 토픽모델링 연구 동향

비정형 데이터는 구조가 정해져 있지 않고 형식이나 패턴이 없는 자유로운 형식의 데이터이며 이미지, 오디오, 비디오, 텍스트 등의 데이터가 이에 속한다. 데이터베이스의 테이블처럼 구조화된 형식을 갖추고 있는 정형 데이터와는 다르게 비정형 데이터 그 자체로는 분석이 어려울 수 있지만, 적절한 분석 기법을 사용하면 유용한 정보와 인사이트를 제공하는 강력한 데이터 자원이 될 수 있다. 텍스트마이닝(Text Mining)은 비정형 텍스트 데이터를 자연어처리기술(Natural Language Processing, NLP)을 통해 분석하여 통계적, 의미론적, 또는 패턴 기반의 정보들을 도출하여 데이터 내에 숨어있는 의미를 찾아가는 기술이다. 특히 텍스트 마이닝은 정보 검색, 감성 분석, 주제 추출, 연구 동향 등 다양한 분야에서 활발히 연구가 진행되고 있다[22].

텍스트 마이닝의 여러 기법 중 토픽모델링은 텍스트 데이터에서 잠재적인 주제, 토픽을 발견하는 기법으로 이 기법을 통해 문서 집합에서 숨겨진 문서 간의 관계를 이해하여 연구의 주제나 트렌드를 파악하기에 효과적이다. 토픽모델링의 대표적인 방법으로는 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)와 잠재 의미 분석(Latent Semantic Analysis, LSA) 등이 있다. LSA 기법은 문서-단어 행렬을 특이값 분해(Singular Value Decomposition SVD) 하여 단어의 의미 공간을 축소하는 방식으로 저차원 공간에서 문서와 단어 간의 연관성을 분석한다[23]. 반면 LDA 기법은 주어진 문서 집합은 여러 개의 토픽으로 이루어져 있다고 가정하며, 각 토픽은 단어들의 확률적 분포로 표현된다. 이를 통해 LDA는 각 문서가 특정 주제에 속할 확률 분포를 추정하며, 문서 내 단어들의 분포가 여러 주제에 걸쳐 어떻게 나타나는지 분석할 수 있다[24]. 이 외에도 사전 훈련된 언어모델을 바탕으로 문장이나 문서의 문맥적인 의미를 더 잘 파악할 수 있도록 설계된 BERT(Bidirectional Encoder Representations from Transformers) 모델[25] 및 시간에 따른 토픽의 변화를 추적하는 Dynamic Topic Models 등도

활발히 연구되고 있다[26]. 기존에 연구되지 않았던 실감미디어 분야의 연구 동향 및 연구 방향성을 파악하기 위해 텍스트마이닝 기법과 토픽모델링의 방식을 적용하였다.

## III. 연구 방법

### 3-1 연구 절차

국가 R&D 과제 데이터를 수집하기 위해 NTIS 포털에서 키워드 검색을 통해 과제를 검색하였고 웹 스크래핑 기술을 통해 텍스트 데이터를 수집하였다. 웹 스크래핑은 온라인상의 정보들을 탐색하고 수집하는 작업을 의미하며 Selenium, BeautifulSoup library를 이용하여 게시글의 html 페이지 소스 분석을 통해 필요한 데이터를 수집하였다. 수집된 한국어 데이터 분석을 위해 KoNLPy library를 이용하여 자연어 분석을 진행하였으며 KoNLPy 가 지원하는 여러가지 형태소 분석기 중 정확도와 분석 속도를 우선 고려하여 Mecab 사전을 선정하여 사용하였다. 수집한 데이터를 구조화, 정제, 정렬 등의 관계형 테이블 형태로 관리하기 위해 Pandas library를 이용하였다. 수집한 텍스트 데이터에서 연구 동향에 활용 할 수 있는 명사를 추출하였고 불용어 사전을 설정하여 분석결과에 노이즈로 영향을 주는 단어들 제외하는 클렌징 작업을 반복해서 수행하며 전반적인 데이터의 전처리 과정을 거쳤다. 본 연구에서는 탐색적 자료 분석의 과정으로 연도별 과제 수, 연구비의 추세, 미래유망신기술(이하 6T) 관련기술분류 등의 빈도분석을 선행하여 수행한 후 LDA 기반의 토픽모델링과 Word2Vec 임베딩 과정을 통한 키워드와 유사한 단어 분석 등의 기법을 통해서 주요 연구 트렌드 및 방향성에 대해 연구를 진행하였다. 데이터 수집에서 정제 및 분석까지 전체

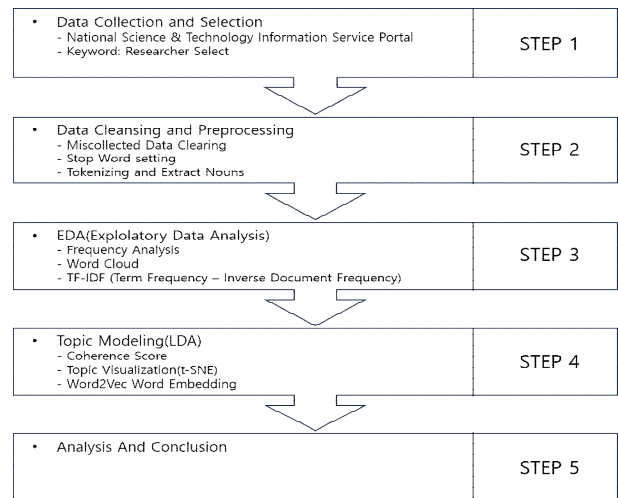


그림 1. 토픽모델링을 통한 연구 동향 분석 흐름도  
Fig. 1. Workflow of trend analysis with topic modeling

과정은 python 언어로 알고리즘을 작성하여 분석에 활용하였다. 전반적인 토픽모델링 분석의 흐름도는 그림 1과 같다.

### 3-2 연구 방법

#### 1) 연구과제 데이터 수집

실감미디어 관련 연구과제들의 동향을 분석하기 위하여 다양한 키워드로 검색한 데이터 수집이 필요하였다. 학자마다, 관련 분야마다 통일되지 않고 다양한 용어를 사용하고 있어 대표적인 한국정보통신기술협회에 기술된 실감미디어와 관련 있는 단어들을 참고하였다. NTIS 포털에서 “가상현실”, “증강현실”, “공간 미디어”, “공간 영상”, “디지털 트윈”, “실감 미디어”, “초실감”, “라이트필드”의 검색 키워드를 이용하였으며 데이터를 확보할 수 있는 전체 기간인 2002년부터 2023년까지 진행된 연구과제 데이터를 수집하였다. 수집한 과제 중 일부는 중복된 과제 및 실감미디어와 관련이 없는 과제가 포함되어 있었기에 연구자는 이러한 잘못 수집된 자료를 직접 검토하고 삭제하였다. 그 결과 수집한 20,206건의 과제 중 18,504건의 과제들을 분석 대상 자료로 선정하였다. 과제별로 수집된 데이터 중 기준년도, 연구비, 과제명, 한글키워드, 6T 관련기술분류, 연구목표요약 데이터를 분석에 사용하였다. 텍스트 데이터 말뭉치로서 과제명, 한글키워드, 연구목표요약을 하나의 문서로 합쳐 토픽모델링으로 분석할 원시 데이터로 설정하였다. 수집한 데이터 중 결측치가 많은 연구내용요약, 기대효과요약 항목은 분석에서 제외하였다.

#### 2) 데이터 전처리 및 자연어 분석

불용어 사전에는 포괄적이거나 일반적인 단어로서 의미를 유추해 내기 어려운 단어들, 가령 ‘시스템’, ‘문제’, ‘목표’ 등의 단어들에 추가되었다. 이는 분석에 무의미한 단어들인 분석 과정에 포함될 경우 모델의 효율성과 정확성에 부정적인 영향을 주기 때문이며 도메인 지식이 있는 연구자의 판단을 바탕으로 반복적인 작업을 통해 불용어 사전을 만들고 분석할 단어에서 삭제하는 반복적인 과정을 진행하였다. 비정형 데이터인 텍스트 데이터를 컴퓨터로 인식 및 분석이 가능한 개별적인 단위로 분할 하는 과정을 Tokenizing 이라고 하며 한국어 형태소 분석기를 이용하여 토큰화 작업을 진행할 수 있다. 한국어 형태소 분석기는 Hannanum, Kkoma, Okt, Mecab, Komoran 등이 있으나 수집한 데이터에서 임의로 샘플 데이터를 선정하여 형태소 분석을 실험해 본 결과 Mecab-ko이 속도 측면과 정확도 측면에서 가장 적절하다 판단되어 채택하여 분석에 사용하였다. 본 연구의 목적인 주제 분석과 키워드 추출을 위해 명사를 추출하여 분석에 사용하였다. 이와 같은 전처리 과정과 명사 추출 과정을 통해 18,504개의 과제에서 2,147,216개의 단어를 추출하였으며 그 중 독립된 단어 19,503개를 분석에 사용하였다.

#### 3) 토픽모델링

토픽모델링의 분석 결과는 연구자가 선정한 토픽의 수에 따라 토픽별 주요 단어들인 결과물로 나오게 된다. 이때 선정한 토픽의 수가 적으면 데이터의 주요 패턴을 찾기 어려운 반면 과하게 많으면 과적합의 문제를 발생시킬 수 있기에 연구자는 토픽의 수를 선정하는 데 있어 신중함을 기해야 한다. 본 연구에서는 토픽의 수를 선정하기 위해 다양한 보조지표들을 활용하여 적절한 토픽의 수를 찾기 위한 과정을 거쳤다. 우선, Coherence Score와 Held-Out Likelihood를 이용하여 각 토픽의 의미적 일관성을 평가하였다. 첫번째로 Coherence Score는 특정 토픽의 단어들끼리 서로 얼마나 자주 함께 나타나는지를 측정하여 같은 토픽에 속하는 단어들끼리 실제로 관련이 높은 단어들이지를 판단하여 토픽의 의미적 일관성을 측정하는 지표이다. 두번째로 Held-Out Likelihood 분석은 주어진 데이터 세트로부터 LDA 모델이 얼마나 적절하게 분석되었는지, 즉 모델의 일반화 여부를 측정하는 분석방법이다. 이를 통해 데이터 적합도를 평가하여 정량적인 지표를 기반으로 토픽의 수를 선정하는데 도움이 된다. 셋째로는 t-SNE(t-Distributed Stochastic Neighbor Embedding) 시각화 방법이다. 이 기법은 각 토픽간의 거리와 관계를 저차원의 영역으로 축소하는 방식으로서 토픽 간의 군집 구조를 시각화 할 수 있는 분석 기법이다. 또한 분석한 주요 단어로부터 토픽의 주제를 선정하는 라벨링 과정을 위해 주요 단어들과 유사한 단어를 분석하는 Word2Vec 임베딩 방식도 병행되었다. Word2Vec 기법은 각 단어들에 크기와 방향성을 갖는 벡터 형태로 변환하여 단어와 단어 간의 의미적 유사성을 수치화하는 기법으로 단어의 문맥적 의미를 고려하여 벡터 공간상에서 유사한 단어들끼리 가까운 위치에 매핑한다. 이와 같은 방법들을 이용하여 각 토픽의 의미를 명확히 하고 라벨링의 정확성을 높여 모델링 결과의 해석에 정교함을 높이고자 하였다. 최종적으로 토픽의 수와 주제의 결정은 모델의 결과를 분석하고 도메인 지식을 바탕으로 한 연구자의 주관적인 판단에 따라 이루어졌다.

## IV. 연구 결과

### 4-1 탐색적 자료 분석

수집된 자료는 2002년도에 174건의 과제가 연구되었으며 연구비는 약 967억의 연구비 규모이다. 하지만 2023년도에는 2,291건의 과제가 수행되었으며 그 연구비 규모 또한 1조 2,432억원 수준으로 과제 수와 연구비 모두 급격하게 증가함을 확인할 수 있었다. R&D 과제 특성상 다년간 이어지는 과제인 경우가 다수 수행되고 있으며, 수집한 NTIS 데이터 자료 중 과제비는 연구를 시작하는 첫해 년도 자료로 기록 관리를 고려하였을 때 연구비 규모는 더욱 커질 수 있다. 조사

기간동안 수행된 연구과제 수와 연구비 합계는 그림 2와 그림 3으로 그 경향성을 파악할 수 있다.

NTIS 데이터에는 6T 관련기술분류와 국가전략기술분류로 기술을 분류하여 관리하고 있는데 6T 관련기술분류체계는 기술 분야를 포괄적으로 나누어 기술의 발전을 종합적으로 분석하는데 중점을둔다. 따라서 실감미디어 관련 연구 동향을 분석하는 본 연구에 있어 분석에 사용하는 것이 적절하다고 판단되었다. 6T 분류모델에 따라 과제를 분류하여 경향성을 확인하였을 때 정보기술(IT) 분야가 월등히 빠른 증가 추세를 보였으며 그 뒤를 이어 생명공학(BT) 분야와 문화기술(CT) 분야가 실감미디어 분야에서 활발히 연구되고 있음을 확인할 수 있었다. 과제비 분석과 마찬가지로 다년간 수행되는 과제의 경우 기준년도는 연구 시작 시점으로 관리되는 점을 감안할 때 분류별 비중은 다소 차이가 날 수 있다. 그림 상의 ETC는 6T에 해당하지 않는 과제의 경우를 말한다. 6T 기술분류에 따른 연도별 과제 수는 그림 4에서 그래프로 확인할 수 있다.

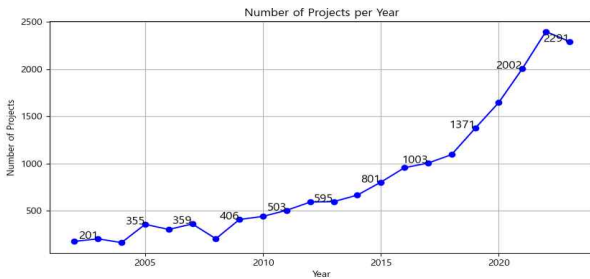


그림 2. 연도 별 연구 과제 수  
Fig. 2. Number of research per year



그림 3 연도 별 연구 과제비  
Fig. 3. Research fund per year

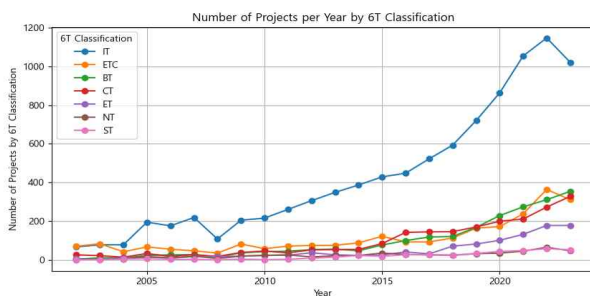


그림 4. 6T 기술분류에 따른 연도별 과제 수  
Fig. 4. Number of research per year by 6T classification

가장 최근 연구의 방향성을 파악해 보기 위해 2023년도의 자료를 대상으로 그 비중은 그림 5에서 파이 차트로 확인 할 수 있다. 분야별 연구과제수의 정량적인 수치는 전체 과제 중 약 1,020여개로 전체 과제수의 약 45%의 비중이 IT 관련 과제이었고 BT, CT 관련 과제는 각각 16%, 15% 정도의 비중을 차지함을 확인할 수 있다.

문서에서 단어의 중요성을 직관적으로 확인하기 위한 방법으로 단어빈도(TF, Term Frequency)와 역문서 빈도(IDF, Inverse Document Frequency)를 결합한 지표인 TF-IDF를 이용하여 단어의 상대적 중요성을 파악하였다. 단순히 단어가 등장한 빈도만 확인하였을 경우 단어의 빈도가 중요성을 대변한다고 볼 수 없고, 여러 문서에서 등장하는 같은 단어라 할지라도 문서마다 등장 빈도가 다를 수 있기 때문에 서로 다른 문서 간에 단어의 중요성을 비교하기 위해 TF-IDF 방식을 이용하였다. 그리고 그 결과를 그림 6의 WordCloud 방식으로 시각화하였으며 WordCloud 시각화 기법에서 글자의 크기는 해당 단어의 중요도를 나타낸다.

2023 Project Counts by 6T Classification

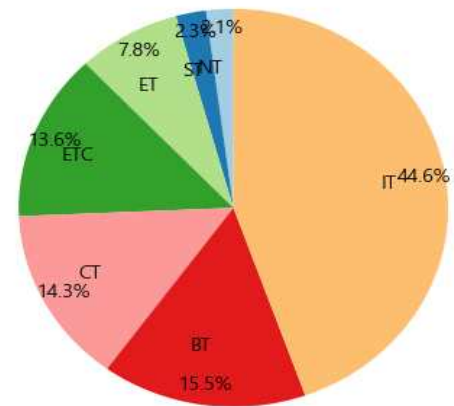


그림 5. 2023년도 과제의 6T 기술분류  
Fig. 5. Number of Research in 2023 by 6T Classification



\*Word Cloud for Korean-Based Project  
그림 6. 중요도 기반(TF-IDF) 상위 단어 시각화  
Fig. 6. Top Word Visualization based on TF-IDF

4-2 토픽모델링 분석

토픽모델링에서 적절한 주제의 수를 설정하기 위해 토픽의 의미적 일관성을 측정하는 Coherence Score 분석을 하였을 경우 그림 7에서 확인할 수 있듯 8개에서 12개의 토픽 수가 가장 적절할 것으로 예상하였다.

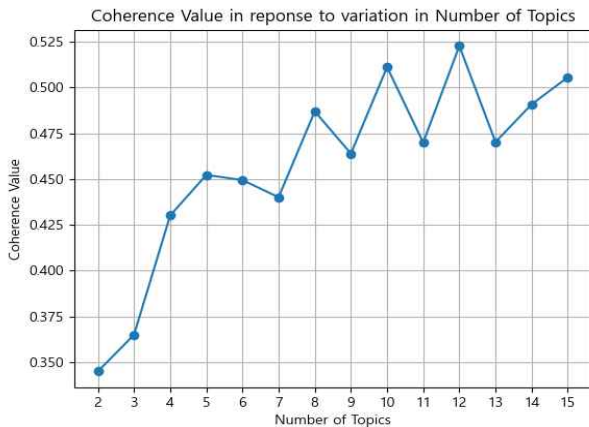


그림 7. 토픽 수에 따른 일관성 점수  
Fig. 7. Coherence score by topic number

두 번째 지표로서 토픽의 의미적 일관성을 확인하기 위해 Held-out Likelihood 분석을 통해 4개에서 9개의 토픽의 수가 적절한 것으로 판단 되었다. 이 결과는 그림 8을 통해 확인할 수 있다.

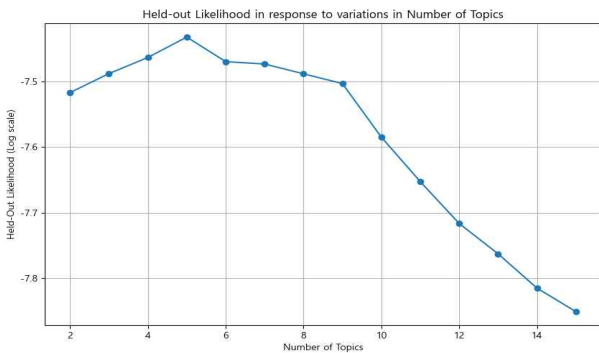


그림 8. 토픽 수에 따른 모델 일반화 평가  
Fig. 8. Held-out likelihood by topic number

토픽 간의 거리를 시각화하는 방법인 t-SNE 분석 기법을 통해 가장 적절한 토픽의 수는 7개에서 10개 정도임을 확인할 수 있었다. 해당 기법을 통해 토픽의 개수에 따른 분포 시각화 결과를 판단할 때는 군집이 분리되지 않고 겹치지 않는 토픽의 수를 선정하였다. 다만 이 분석 기법의 원리상 차원 축소로 인한 정보 손실이 발생하므로 토픽의 수를 결정하기 위한 보조지표로 활용하였다. 토픽의 수가 9개인 경우 t-SNE 시각화의 결과는 그림 9와 같이 확인할 수 있다.

Intertopic Distance Map (via multidimensional scaling)



그림 9. t-SNE를 이용한 토픽별 분포 시각화  
Fig. 9. Topic modeling visualization w/ t-SNE

다양한 지표들을 활용하여 토픽의 수의 범위를 산정하여 토픽의 수는 7~9개로 판단하였다. 그 후 토픽 수에 따른 각각의 토픽모델링 결과를 확인하며 각 주제를 명확하게 표현하고, 주제 간 겹치는 부분이 적어 해석의 용이성을 고려하여 연구자의 판단으로 최종적으로 토픽의 수는 9개로 선정하였다. 분류된 토픽별 주요 단어들을 기반으로 적절한 주제로서 라벨링을 위해 Word2Vec 기법 중 CBOW(Continuous Bag of Words) 방식을 통해 생성된 단어 벡터를 이용하여 토픽모델링의 결과 도출된 단어들과 벡터공간 상의 거리를 측정한다. 이 때 -1에서 1의 값을 갖는 코사인 유사도를 기반으로 거리를 측정하며 이 결과 값이 클수록 기준 단어와 유사한 단어라고 판단한다. 다만 토픽모델링의 주요 단어들 각각에 대해 유사한 단어를 선정하기 때문에 유사 단어를 채택 함에 있어 토픽의 주제를 잘 설명할 수 있는지 여부를 연구자가 판단하여야 한다.

학습에 사용된 주요 파라미터값은 vector\_size = 200, window = 5, min\_count = 5, sg = 0 으로 설정하였다. Word2Vec 그 자체로는 LDA 토픽모델링에 직접적인 영향은 주지 않지만 단어 벡터의 표현력, 품질에 영향을 주어 토픽모델링에서 의미있는 결과를 도출해 내는데 간접적으로 활용된다. 그러므로 연구자는 컴퓨팅과위를 고려한 Word2Vec의 적절한 파라미터 변화 실험을 통해 토픽모델링의 결과를 비교해 보는 과정을 거치는 것이 좋다. 토픽모델링의 결과로 도출된 주요 단어들과 Word2Vec으로 분석한 유사단어들을 포괄적으로 고려하여 각 토픽의 주제를 선정한 결과를 표 1로 확인할 수 있다.

**표 1. 실감미디어 연구 관련 토픽 별 주제 표**  
**Table 1. Table of topics for immersive media research**

Topic No.	Proposed Topic label
1	사용자 경험기반 실감 콘텐츠 (User Experience Based Immersive Contents)
2	공간데이터 기반 환경 모델링 (Spatial Data Driven Environment Modeling)
3	통신 및 에너지 전송시스템 (Telecommunications and Energy Transmission Systems)
4	디지털 혁신과 교육 서비스 (Digital Transformation and Education Services)
5	소자, 재료 및 공정 등 기반기술 (Enabling Technologies, including Devices, Materials, and Processes)
6	원격 및 자동화 시스템 (Remote and Automated Systems)
7	플랫폼 설계 및 유지관리 (Platform Design and Maintenance)
8	영상 데이터 분석 및 처리 (Analyze and Process Video Data)
9	영상기반 의료 서비스 (Image-based Healthcare)

\*Topic Lables for Korean-Based Project

주제 1의 주요 단어로는 ‘기반’, ‘공간’, ‘정보’, ‘콘텐츠’, ‘서비스’, ‘사용’, ‘플랫폼’, ‘가상’, ‘실감’, ‘제공’이 있었고 관련된 상위 유사 단어(코사인 유사도 값)들로는 ‘사용자(0.7239)’, ‘인터랙티브(0.6328)’, ‘플랫폼(0.6194)’, ‘창작물(0.5896)’, ‘원자화(0.6026)’, ‘데이터베이스(0.6476)’ 등이 있다. 주제 1의 라벨을 선정하는 과정은 서비스, 기반, 사용자, 인터랙티브, 등의 주요단어와 유사 단어들 중 일부를 복합적으로 고려하여 사용자 경험기반 이라는 주제를 선정하였다. 실감, 창작물, 콘텐츠, 플랫폼 등의 단어를 통해 실감형 콘텐츠를 제공하는 개념의 주제를 선정하였다. 그 결과 주제 1의 라벨은 최종적으로 사용자 경험기반 실감 콘텐츠라고 정하였다. 다른 토픽의 주제들도 동일한 방법으로 선정하였다.

주제 2의 주요 단어로는 ‘공간’, ‘데이터’, ‘정보’, ‘환경’, ‘사용’, ‘모델’, ‘학습’, ‘구조’, ‘방법’, ‘양자’가 있었고 관련된 유사단어들로는 ‘알고리즘’, ‘그래프’, ‘데이터베이스’, ‘모델링’, ‘심층’, ‘추론’, ‘정보’, ‘공간’이 있었다. 데이터와 환경을 기반으로 한 모델링 방법에 초점을 맞추고 있는 것으로 보아 공간데이터 기반 환경 모델링으로 토픽의 라벨을 정하였다.

주제 3의 주요 단어로는 ‘기반’, ‘설계’, ‘제어’, ‘네트워크’, ‘통신’, ‘에너지’, ‘전송’, ‘무선’, ‘제조’, ‘선박’이 있었고 관련된 유사단어들로는 ‘기자재’, ‘송신’, ‘유선’, ‘통신망’, ‘패킷’, ‘공장’, ‘컨트롤’, ‘신재’ 등이 있었다. 이 토픽은 기술적인 부분으로 통신, 에너지, 네트워크 등 강조된 것으로 분석되어 주제 3의 라벨은 통신 및 에너지 전송시스템으로 정하였다.

주제 4의 주요 단어들은 ‘디지털’, ‘콘텐츠’, ‘구축’, ‘융합’, ‘교육’, ‘산업’, ‘분야’, ‘사업’, ‘양성’, ‘인력’이 있었고 관련된 유사단어들은 ‘인력’, ‘통합’, ‘혁명’, ‘구축함’, ‘산업’, ‘혁신’, ‘교육’ 등이 있었다. 이 토픽은 디지털 솔루션과 교육

의 융합산업을 다루는 것으로 보아 주제 4의 라벨은 디지털 혁신과 교육 서비스로 정하였다.

주제 5의 주요 단어로는 ‘구조’, ‘측정’, ‘효율’, ‘공정’, ‘나노’, ‘광학’, ‘소재’, ‘소자’, ‘물질’, ‘발광’ 등이 있었고 관련된 유사단어들로는 ‘계측’, ‘열처리’, ‘광소자’, ‘공진기’, ‘전계’, ‘화합물’ 등이 있었다. 이 토픽은 부품, 소자, 재료, 광학 등의 원천 기반기술에 중점을 두는 것으로 보아 주제 5의 라벨은 소자, 재료 및 공정 등 기반기술로 정하였다.

주제 6의 주요 단어로는 ‘설계’, ‘평가’, ‘제작’, ‘로봇’, ‘모듈’, ‘성능’, ‘검증’, ‘자율’, ‘드론’, ‘기관’ 등이 있었고 관련된 유사단어들로는 ‘화질’, ‘이펙터’, ‘모빌리티’, ‘무인기’, ‘프로토타입’, ‘시제품’ 등이 있었다. 이 토픽은 로봇 및 자율, 지동화 시스템의 설계, 제작, 평가와 관련된 측면에 중점을 두는 것으로 보아 주제 6의 라벨은 원격 및 자동화 시스템으로 정하였다.

주제 7의 주요 단어로는 ‘기반’, ‘데이터’, ‘정보’, ‘디지털’, ‘모델’, ‘구축’, ‘스마트’, ‘트윈’, ‘관리’, ‘운영’ 등이 있었고 관련된 유사단어들로는 ‘연계’, ‘알고리즘’, ‘통합’, ‘데이터베이스’, ‘유지관리’, ‘인공지능’, ‘정보’ 등이 있었다. 이 토픽은 디지털 플랫폼, 데이터 분석, 스마트 솔루션의 구축 및 운영에 관한 내용으로 보이며 주제 7의 라벨은 플랫폼 설계 및 유지관리로 정하였다.

주제 8의 주요 단어로는 ‘영상’, ‘기반’, ‘정보’, ‘차원’, ‘센서’, ‘인식’, ‘알고리즘’, ‘처리’, ‘카메라’, ‘비디오’ 등이 있었고 관련된 유사단어들로는 ‘코덱’, ‘필터링’, ‘라이다’, ‘매칭’, ‘센싱’, ‘고해상도’ 등이 있었다. 이 토픽은 센서를 기반으로 한 정보처리, 객체 인식 알고리즘, 영상기반 기술 관련된 내용을 포함하는 것으로 보아 주제 8의 라벨은 영상데이터 분석 및 처리로 정하였다.

주제 9의 주요 단어로는 ‘영상’, ‘기반’, ‘기능’, ‘변화’, ‘위성’, ‘의료’, ‘자료’, ‘진단’, ‘치료’, ‘환자’ 등이 있었고 관련된 유사단어들로는 ‘협진’, ‘약물’, ‘치과’, ‘임상’, ‘원격탐사’, ‘인공위성’, ‘치료법’, ‘질병’ 등이 있었다. 영상기술을 활용한 원격 탐사기술 관련된 주제일 수도 있지만 영상기술을 활용한 의료 진단, 치료, 모니터량 과 관련된 내용으로 판단되었다. 주제 9의 라벨은 영상기반 의료 서비스라고 정하였다.

## V. 결 론

본 연구는 디지털환경으로 급변하는 시대에 실감미디어에 대한 관심이 국가 R&D 연구에서는 어떻게 변화하고 있는지를 파악하기 위해 텍스트마이닝이라는 방법을 이용하여 정량적인 분석을 수행하였다. NTIS 포털에서 웹 스크래핑 방식을 이용하여 2002년부터 2023년까지 수행되었던 국내 실감미디어 분야의 국가 R&D 과제를 18,504건의 과제데이터를 수집하여 분석을 수행하여 실감미디어 분야의 연구 동향 및 주

요 연구과제의 방향을 파악하였다. LDA 기반의 토픽모델링 분석방법을 이용하여 연구 동향과 주제를 보다 섬세하고 명확하게 분류하기 위하여 Coherence Score, Held-Out Likelihood, t-SNE 시각화, Word2Vec Embedding 을 통한 유사단어 파악 기법 등 다양한 보조지표들을 함께 사용하였다. 최종적으로는 관련 산업의 도메인 지식이 있는 연구자의 판단으로, 토픽의 수와 라벨링 분석을 진행하였다. 실감미디어 기술의 발전과 함께 다양한 응용 분야에서의 연구가 활발히 진행되고 있음을 확인할 수 있었다. 분석 결과 실감미디어 분야의 연구 동향으로는 9개 주제를 도출하였으며 각각의 토픽별 레이블은 1. 사용자 경험기반 실감 콘텐츠, 2. 공간데이터 기반 환경모델링, 3. 통신 및 에너지 전송시스템 4. 디지털 혁신과 교육 서비스, 5. 소자, 재료 및 공정 등 기반기술, 6. 원격 및 자동화 시스템, 7. 플랫폼 설계 및 유지관리 8. 영상데이터 분석 및 처리, 9. 영상기반 의료 서비스로 연구동향의 흐름을 보인다. 즉 실감미디어는 가상현실(VR), 증강현실(AR), 혼합현실(MR) 등 첨단 기술의 융합이 중심이 되어, 교육, 의료, 산업 등 다양한 분야에서의 공간의 제약을 넘어 사용자들과 양방향 통신을 할 수 있는 사용자 편이성이 높은 플랫폼이 강조되고 있다고 파악된다. 그리고 이 플랫폼에서 사용할 수 있는 미디어, 콘텐츠의 몰입감을 높이기 위한 감각의 활용과 사용자와의 인터랙션에 그 중요성이 연구의 큰 부분을 차지하는 것으로 확인 할 수 있다. 마지막으로 새로운 미디어의 활용을 위해 소자 재료 등 장치에서부터 센서, 전력, 통신망 등의 하드웨어 인프라 구축에 관한 연구가 활발히 진행되고 있음을 확인할 수 있었다. 본 연구를 통해 실감미디어 산업의 미래 연구 방향을 제시하는 데 기여할 수 있기를 희망한다. 정부와 연구기관에 효율적인 R&D 자원 배부와 정책 수립에 도움이 되고 연구자에게 연구의 방향성과 수립하고 기술 혁신을 선도하는데 중요한 참고자료로 활용될 수 있을 것이다. 나아가, 실감미디어 기술의 빠른 발전에 발맞추어 지속적인 데이터 분석과 연구 동향 모니터링이 필요하다.

하지만, 본 연구에는 몇 가지 한계점이 존재한다. 우선 정적 연구를 넘어 시간에 따른 연구주제의 변화나 기술 동향의 흐름을 파악할 필요성이 있다. 약 20년 동안의 데이터 세트를 바탕으로 토픽모델링으로 연구의 동향을 분석한 이번 연구로는 관련 산업 기술의 변화 속도를 감안 했을 때 새로운 트렌드와 연구 분야의 변화를 놓칠 가능성이 있다. 그렇기에 시간적인 변수를 바탕으로 주제를 분석할 수 있는 다이내믹 토픽 모델링을 바탕으로 한 연구가 필요하다. 또한 토픽모델링의 결과를 보다 심층적으로 이해하고 분석하기 위해 토픽 간의 관계에 대한 분석이 미흡하다. 토픽 간의 중심성과 영향력을 평가하고 주제 간의 중요도와 영향력의 구조를 파악할 수 있는 추가적인 네트워크 분석을 통해 심층적인 연구로의 확장이 필요해 보인다. 마지막으로 학문적으로 심도 있고 실무적인 의미를 찾을 수 있도록 각 토픽별로 진행되고 있는 연구의 방향과 직면한 문제점에 대한 연구가 필요하다.

## 참고문헌

- [1] W. S. Kim and K. Nah, "A Study on Conceptual Framework of Realistic Media Types," *Journal of the Korean Society of Design Culture*, Vol. 25, No. 2, pp. 93-106, June 2019. <https://doi.org/10.18208/ksdc.2019.25.2.93>
- [2] J.-A. Kim and D. Kim, "A Study on the Screen Evolution and Expansion of the Concept," *The Journal of Korean Institute of Next Generation Computing*, Vol. 8, No. 2, pp. 87-98, April 2012.
- [3] Deloitte. The Metaverse Overview: Vision, Technology, and Tactics [Internet]. Available: <https://www2.deloitte.com/cn/en/pages/technology-media-and-telecommunications/articles/metaverse-report.html>.
- [4] S. H. Lee, "A Study of the Realistic Media of Smart Glass in the Public Relations Hall of a Small Model House Less than Three Pyeong," *Journal of Korean Society of Media & Arts*, Vol. 20, No. 3, pp. 75-87, June 2022. <https://doi.org/10.14728/KCP.2022.20.03.075>
- [5] J. W. Kim, J. H. Maeng, J. Y. Joo, and K. H. Im, "Learning System for Scientific Experiments with Multi-Touch Screen and Tangible User Interface," *Journal of the Korea Contents Association*, Vol. 10, No. 8, pp. 461-471, August 2010. <https://doi.org/10.5392/JKCA.2010.10.8.461>
- [6] Korea Creative Content Agency. Immersive Media Industry Trends of Russia [Internet]. Available: <https://www.kocca.kr/kocca/bbs/view/B0158950/2004528.do?menuNo=204160>.
- [7] Gyeongju Smart Media Center, Realistic Media Product Promotion Group, Ministry of Science, ICT and Future Planning, Gwacheon, TRKO201700017290, July 2017.
- [8] National Science & Technology Information Service. NTIS User Guide [Internet]. Available: <https://www.ntis.go.kr/>.
- [9] M.-N. Shim, "Analysis of Research Trends on the Domestic Metaverse Using Topic Modeling: Focusing on Research from 2007 to 2022," *Journal of Digital Contents Society*, Vol. 23, No. 12, pp. 2457-2468, December 2022. <https://doi.org/10.9728/dcs.2022.23.12.2457>
- [10] J. Kim, C. Yang, and S. Park, "A Study on the Research Trends for Tunnel Traffic Safety Using Text Mining," *Journal of Digital Contents Society*, Vol. 23, No. 10, pp. 2075-2083, October 2022. <https://doi.org/10.9728/dcs.2022.23.10.2075>
- [11] S. M. Kim and Y. J. Kim, "Research Trend Analysis on Living Lab Using Text Mining," *Journal of Digital Convergence*, Vol. 18, No. 8, pp. 37-48, August 2020. <http://doi.org/10.14400/JDC.2020.18.8.037>
- [12] S. J. Park, K. H. Kim, and S. Jeong, "The Study on the Analysis of Efficiency of Governmental R&D Programs



- Regarding to the S&T Outcomes,” *Journal of Korea Technology Innovation Society*, Vol. 14, No. 2, pp. 205-222, June 2011.
- [13] J.-H. Kim and S.-S. Kim, “A Study on the Analysis of Agricultural R&D Keywords Using Textmining Method,” *Journal of the Korea Academia-Industrial Cooperation Society*, Vol. 22, No. 2, pp. 721-732, February 2021. <http://doi.org/10.5762/KAIS.2021.22.2.721>
- [14] J.-Y. Lee, H.-I. Na, B.-H. Lee, and T.-H. Kim, “Discovering the Knowledge Structure of Graphene Technology by Text Mining National R&D Projects and Newspapers,” *Journal of the Korea Contents Association*, Vol. 21, No. 2, pp. 85-99, February 2021. <http://doi.org/10.5392/JKCA.2021.21.02.085>
- [15] B.-H. Lee and T.-H. Kim, “Needmining and Topic Modeling of National R&D Projects to Build LEO Satellite Communication Data/Contents Industry Ecosystems,” *Journal of the Korea Contents Association*, Vol. 24, No. 2, pp. 1-14, February 2024. <https://doi.org/10.5392/JKCA.2024.24.02.001>
- [16] H.-B. Ko, “Issues in the Video Content Production and Theory due to the Appearance of Realistic Media,” *Asia-Pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol. 8, No. 5, pp. 237-250, May 2018. <http://doi.org/10.35873/ajmahs.2018.8.5.024>
- [17] J. Lee and Y. Yoon, “FMM: Fusion Media Middleware for Actual Feeling Service,” *Journal of Korea Multimedia Society*, Vol. 13, No. 2, pp. 308-315, February 2010.
- [18] G.-H. Lee, H.-C. O. Li, D. Bang, C. H. Ahn, M. S. Ki, and S. W. Kim, “The Effect of Matching between Odor and Color on Video Reality and Sense of Immersion,” *Journal of Broadcast Engineering*, Vol. 19, No. 6, pp. 877-895, November 2014. <https://doi.org/10.5909/JBE.2014.19.6.877>
- [19] H. K. Cho, “A Study on the Use of Immersive Media Contents Design Based on Extended Reality(XR) Technology in Digital Transformation Era,” *Journal of the Korean Society of Design Culture*, Vol. 26, No. 4, pp. 497-507, December 2020. <https://doi.org/10.18208/ksdc.2020.26.4.497>
- [20] D.-J. Kim, “A Study on the Spatial Characteristics of Realistic Media Exhibition Content,” *Journal of Digital Art Engineering & Multimedia*, Vol. 10, No. 2, pp. 221-229, June 2023. <https://doi.org/10.29056/jdaem.2023.06.08>
- [21] Y. Yoon and S. Chin, “Touch Sensory in Virtual Clay Art Using Leap Motion and Arduino,” *Asia-Pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol. 8, No. 2, pp. 467-474, February 2018. <http://doi.org/10.35873/ajmahs.2018.8.2.046>
- [22] J.-H. Ryu and Y.-Y. You, “The Fourth Industrial Revolution Core Technology Association Analysis Using Text Mining,” *Journal of Digital Convergence*, Vol. 16, No. 8, pp. 129-136, August 2018. <https://doi.org/10.14400/JDC.2018.16.8.129>
- [23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391-407, September 1990. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)
- [24] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, March 2003.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis: MN, pp. 4171-4186, June 2018. <https://doi.org/10.18653/v1/N19-1423>
- [26] D. M. Blei and J. D. Lafferty, “Dynamic Topic Models,” in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, Pittsburgh: PA, pp. 113-120, June 2006. <https://doi.org/10.1145/1143844.1143859>



### 서효기(Hyo-Gi Seo)

2009년 : 고려대학교 전기전자전파공학부 (공학학사)  
2011년 : 고려대학교 전기전자전파공학과 (공학석사)

2011년~현 재: LG전자 광학연구소  
2019년~현 재: 고려대학교 기술경영전문대학원 박사과정  
※관심분야 : 기술경영, 텍스트마이닝, 토픽모델링, 영상광학, 공간영상



### 김영준(Youngjun Kim)

2008년~2011년: Texas A&M International University 조교수  
2011년~2012년: 서울대학교 기술경영경제정책대학원 초빙교수  
2012년~현 재: 고려대학교 기술경영전문대학원 교수  
※관심분야 : 기술경영경제, 기술재무, 기술이전사업화, 전략적제휴