공공도서관에 적용 가능한 SBERT 기반의 지식 확장형 도서 추천 모델 연구

A Study on the SBERT-based Knowledge Expansion Book Recommendation Model Applicable to Public Library

양 지 수 (JiSu Yang)*
백 재 은 (JaeEun Baek)**

초 록

정보를 빠르고 쉽게 획득하는 맞춤형 도서 추천 서비스가 적극 도입되는 가운데 기존 키워드 매칭 방식의 한계인 의미적 격차를 극복하고, 이용자의 지적 외연 확장을 위해 본 연구에서는 SBERT 기반의 맥락적 도서 추천 모델을 제안하고자 하였다. 도서관 정보 나루의 전국 통합 대출 빅데이터에서 6,376권의 도서 메타데이터를 수집 및 정제하여 이를 바탕으로 도서 요약문을 고차원 벡터로 임베딩하고, 문맥 기반 유사도를 산출하여 벡터 데이터베이스(ChromaDB)를 통해 효과적인 시맨틱 검색이 가능한 모델을 구축하였다. 실험 결과, 본 연구에서 제안한 모델은 이용자의 기초 관심 주제 대비 KDC 기준 약 45.6%의 지식 확장 지수를 기록하였고, 특히, TF-IDF 모델과의 비교 실험을 통해 단순 단어 매칭에서 발생하는 맥락 없는 가짜 확장 사례를 제어하면서 의미적 정합성을 유지하며 타 학문 분야로의 지식 도약을 유도하는 성능을 입증하였다. 이는 추천의 정밀도와 의외성 사이의 균형을 확보하며 알고리즘에 의한 필터버블 현상을 억제하고 공공도서관의 본질적 가치인 지적 성장을 지원하는 기술적 토대를 마련할 수 있을 것이라 기대한다.

ABSTRACT

In order to overcome the 'semantic gap' which is the limitation of the traditional keyword-matching-methods and to expand users' intellectual horizons, this study proposed a contextual book recommendation model based on SBERT while actively introducing customized book recommendation services that acquire information quickly and easily. Utilizing 6,376 refined book records from the 'Library Information Naru' national integrated loan big data, a model implemented an effective semantic search architecture by embedding book summaries into high-dimensional vectors and utilizing a vector database(ChromaDB). Experimental results demonstrated that the proposed model achieved a Knowledge Expansion Ratio of 45.6% according to the KDC. In particular, comparative experiments with the TF-IDF model and semantic verification verified that the proposed system effectively suppresses 'Pseudo-Expansion'-contextless recommendations arising from simple word overlaps-while facilitating intellectual leaps into diverse academic fields with semantic consistency. By securing a practical balance between recommendation precision and serendipity, this study practically mitigates the 'filter bubble' effect. Ultimately, this study will establish a technical foundation for supporting the core mission of public libraries fostering intellectual growth and broadening users' exploratory boundaries.

키워드: 도서관, 도서 추천 모델, 도서 추천 시스템, 의미적 격차, 지식 확장 지수, 필터 버블, 키워드 매칭, 한국 십진분류법
Library, Book Recommendation Model, Book Recommendation System, Semantic Gap, Knowledge Expansion Ratio, Filter Bubble, Keyword-Matching, Korean Decimal Classification

* 덕성여자대학교 문헌정보학전공 학사과정(yangjs0217@duksung.ac.kr) (제1저자)
** 덕성여자대학교 문헌정보학과 교수(jaeeunb@duksung.ac.kr / ISNI 0000 0004 7875 8452) (교신저자)
논문접수일자 : 2026년 5월 19일 논문심사일자 : 2026년 5월 26일 게재확정일자 : 2026년 6월 4일
한국비블리아학회지, 37(2): 261-282, 2026. http://dx.doi.org/10.14699/kbiblia.2026.37.2.261

1. 서 론

1.1 연구의 배경 및 필요성

현대 사회는 정보기술의 비약적 발전과 방대한 데이터 축적으로 개별 이용자의 요구를 정교하게 분석하여 반영하는 초개인화 환경으로 진입하고 있다. 이러한 흐름 가운데 도서관에서도 과거의 보편적 도서 제공 서비스에서 탈피하여 이용자의 대출 이력과 선호도를 분석한 '맞춤형 도서 추천 서비스'를 강화하여 이용자 만족도를 제고하고 있다.

최근의 도서 추천 시스템은 협업 필터링을 통해 유사 이용자 그룹의 데이터를 활용하거나, 형태소 분석 기반의 키워드 매칭을 통해 이용자가 선호하는 주제와 유사한 도서를 제안하는 방식으로 변화하며 고도화되고 있다. 그러나 이와 같은 개인화 추천 알고리즘은 이용자가 선호하는 정보 범주 내에 고립시키는 필터버블 현상을 야기한다는 비판에 직면해 있다. 실제로 여러 소셜 미디어 플랫폼에 대한 대규모 실증 연구에서 사용자들이 동질적인 클러스터로 집합하여 유사한 관점만을 접하게 되는 에코 챔버 효과가 지배적임을 보여주고 있다(Cinelli et al., 2021). 과거의 이력에 과도하게 의존하는 방식은 추천 시스템의 피드백 루프를 통해 이용자가 이미 인지하고 있는 영역의 정보만을 반복적으로 노출함으로 선택의 다양성을 감소시키고(Jiang et al., 2019) 기존의 편향 지식 체계를 강화하는 결과로 이어지게 된다. 그리고 전통적인 키워드 매칭 방식은 단어의 단순 일치 여부에만 치우쳐 데이터의 물리적 특징과 인간의 개념적 해석 사이의 단절인 의미적 격차 문제에 직면하게 된다. 최근 딥러닝 기반 자연어 처리 기술의 발전으로 Transformer와 같은 모델에 의해 텍스트의 의미를 이해하고 전달하는 능력이 크게 향상되었으나(Xie et al., 2021), 이러한 기술은 이용자에게 익숙한 분야만을 제안하여 지식의 편식을 심화시키는 한계를 가져와 모든 종류의 정보에 대한 공평한 접근을 보장하고 개인의 지적 성장을 도모해야 한다는 공공도서관의 본질적인 역할(IFLA, 2012)을 저해할 위험을 크게 가지고 있다. 따라서 단순히 과거 이력을 반복하는 추천에서 벗어나, 도서의 내재적 맥락을 정교하게 이해하고 이용자에게 신선하고 유용한 지적 발견의 기회를 제공할 수 있는 의외성 기반의 지능형 추천 시스템 연구가 절실히 요구되는 시점이라고 볼 수 있다. 이에 본 연구에서는 이용자의 선호도를 반영하면서 지식의 외연을 확장할 수 있는 SBERT(Sentence-BERT, 이하 SBERT) 기반의 문장 임베딩 기술(Reimers & Gurevych, 2019)을 활용하여 공공도서관에 적용할 수 있는 도서 추천 모델을 제안하고자 한다.

기존의 키워드 매칭 방식이 가진 기술적 한계를 극복하고 도서 요약문 등의 비정형 데이터에서 의미론적 유사도를 정교하게 추출함으로 이용자의 잠재적 독서 의도를 파악하기 위해, 첫째, 문맥 이해 능력이 뛰어난 SBERT 모델을 활용하여 이용자의 관심 키워드와 도서 간의 의미적 연결을 강화한 맥락 기반 추천 모델을 설계하였다. 둘째, 도서관의 주제 분류에서 일반적으로 사용하는 한국십진분류법(Korea Decimal Classification, 이하 KDC)을 바탕으로 한 주제 분류에 고착되지 않고 의미적으로 연관된 타 분야의 도서를 발견할 수 있도록 유도

하여 추천 모델의 의외성을 실현하였다. 그리고 셋째, 제안한 모델을 통해 도출된 추천 결과가 실제 이용자의 주제 분포를 어떻게 변화시키는지 분석하여 공공도서관 서비스로서의 지식 확장 가능성을 실증적으로 검증하였다. 이와 같은 과정을 통해 본 연구에서는 최종적으로 이용자에게 개인화된 만족을 제공함과 동시에, 지식의 편식을 해소하고 균형 잡힌 독서 경험을 지원하는 새로운 도서관 추천 서비스 모델을 제시하고자 하였다.

1.2 연구의 범위 및 방법

본 연구는 다음과 같이 대상과 대상의 범위, 방법론을 바탕으로 실행하였다.

우선, 국내 공공도서관에서 수집된 도서 데이터를 대상으로 총 6,376권의 도서 메타데이터를 수집하고 정제하여 실험 모델의 기초 자료로 한정하였다. 단, 데이터의 수집 과정에서 발생할 수 있는 인기도 편향을 해결하기 위해 중복된 인기 도서와 과도하게 편중된 데이터를 정제하였고 이를 통해 실험의 객관성을 확보하고자 하였다. 특히 추천 알고리즘이 상대적으로 인지도가 높은 소수의 아이템에 과도하게 집중할 수 있는 인기도 편향을 완화하지 않으면 롱테일에 속한 다수의 도서가 추천 목록에서 계속 배제되는 부작용이 발생할 수 있어(Klimashevskaia et al., 2024), 이와 같은 시각을 연구에 적용하였다.

그리고 이들을 대상으로 크게 3단계의 수행 방법으로 구분하여 다음과 같이 연구를 진행하였다.

(1) 총 6,376권의 수집된 도서 줄거리 정보를 정제하여 학습 및 검색에 적합한 형태로 가공하는 전처리 단계를 실행한다.

(2) 한국어의 문맥적 의미 파악에 특화된 SBERT 모델을 활용하여 각 도서의 텍스트 데이터를 고차원 벡터로 변환하는 문장 임베딩 과정을 수행한다. 여기서 생성된 벡터 데이터는 효율적인 검색과 유사도 연산을 위해 벡터 데이터베이스 ChromaDB(Chroma, 2023)에 구축하며, 이용자가 입력한 관심 키워드를 동일한 벡터 공간으로 투영한 후 도서 벡터 간의 코사인 유사도를 측정하여 최적의 추천 리스트를 산출한다.

(3) 추천된 도서들의 KDC 분류를 전수 조사하여 이용자가 입력한 기초 관심 주제의 분류군 대비 추천 리스트가 얼마나 다양한 주제로 확장되었는지 분석한다. 특히 입력 키워드의 KDC 범주를 벗어나 새롭게 제안된 도서들의 비율을 정량적으로 산출함으로 제안된 모델이 이용자의 지적 외연을 확장하는 의외성을 어느 정도 확보하고 있는지 실증적으로 검증하는 절차를 함께 실행한다.

2. 도서 추천 알고리즘

2.1 도서 추천 알고리즘 비교 분석

2.1.1 협업 필터링 및 연관규칙 기반 추천

도서 추천 시스템은 정보의 과잉 속에서 이용자의 취향과 목적에 부합하는 저작물을 선별하여 제시하는 핵심적인 서비스이다. 초기의 도서 추천은 주로 대출 빈도가 높은 베스트셀러 또는 신간 도서를 중심으로 이루어졌으나, 최근 빅데

이터와 기계학습 기술의 발전에 따라 이용자 개별 맞춤형 서비스로 진화하고 있다.

가장 보편적으로 활용되는 알고리즘은 협업 필터링(Collaborative Filtering, 이하 CF)으로, 유사한 성향을 지닌 이용자들의 과거 이력을 바탕으로 선호도를 예측하는 방식이고 딥러닝의 발전과 함께 행렬 분해 및 신경망 기반 변형 모델로 계속 진화해 왔다(Zhang et al., 2019). 하나의 예시로 알라딘의 '추천 마법사'와 같은 초기 상용 서비스는 이용자의 구매 이력, 클릭 로그, 블로그 활동 등을 분석하여 취향이 유사한 고객군을 분류하고 도서를 추천하는 사용자 기반 및 아이템 기반 CF 기술을 활용하며 발전해 왔다.

반면에, 평점 데이터가 부족한 도서관 환경에서는 대출 데이터의 빈도와 동시 발생 확률을 분석하는 연관규칙 알고리즘이 주로 사용되고 있다. 그래서 본 장에서는 도서 간의 연관성을 도출하고 매개 중심성 기반의 추천을 수행하여 성능을 개선한 사례를 살펴보며 기술하였다.

2.1.2 콘텐츠 기반 필터링 및 키워드 추출

콘텐츠 기반 필터링(Content-Based Filtering, 이하 CBF)은 아이템이 보유한 고유 속성과 이용자의 과거 선호 프로파일 간의 유사성을 산출하여 추천하는 방식으로, 협업 필터링과 함께 추천 시스템 분야에서 가장 핵심적인 패러다임 중 하나로 평가된다(Zhang et al., 2019).

기존에는 주로 키워드 추출 방식에 의존하였으나, 최근 도서관 현장의 경우, 연관규칙 분석 등 빅데이터 기법을 도입하여 이용자의 잠재적 요구를 파악하려는 시도가 지속되고 있다(임정훈 외, 2022). 그러나 이러한 통계적 방식은 대출 이력 등 정형 데이터에 편중되어 도서 내부의 풍부한 맥락을 반영한 개인화 서비스에 여전히 한계가 존재한다. 이는 이용자 간의 상관관계를 이용하는 협업 필터링과 달리, 아이템 자체가 보유한 특징 공간을 정의하고 이를 벡터화하여 매칭하는 기술적 기제에 기반하기 때문이다.

초기 CBF는 도서관 시스템의 표준 서지 메타데이터(저자, 장르, 주제어 등)와 같은 정형 데이터에 의존하였으나, 최근에는 텍스트 마이닝 기법을 활용한 비정형 데이터 분석으로 외연을 확장하고 있다. 대표적으로 밀리의 서재에서 제공하는 'AI 독파밍'은 도서 추천, 본문 검색, 독서 기록 자동화 등을 결합하여 독자의 잠재적 의도를 파악하는 서비스로(서민음, 2025), 이 시스템은 자연어 질의에 따라 적합한 도서를 제안하거나 도서 요약 및 페르소나 챗봇과의 대화 기능을 제공함으로 비정형 텍스트 데이터의 가치와 소비를 확장하는 기술적 혁신 사례를 보여주고 있다. 특히 독자의 리뷰에 도출된 'AI 스마트 키워드'를 메타데이터 화하여 추천 모델의 입력값으로 활용하는 방식은 정성적인 독서 데이터를 정량적인 추천 지표로 변환하여 시스템의 정확도와 개인화 수준을 높이는 핵심 기제로 작용하고 있다.

기존의 콘텐츠 기반 필터링은 TF-IDF(Term Frequency-Inverse Document Frequency, 이하 TF-IDF)와 같은 단어 빈도 기반의 키워드 가중치 산출 방식에 의존해 왔으나, 이러한 방식은 단어의 표면적 일치 여부에만 집중하기에 언어의 중의성을 파악하지 못하는 의미적 격차 문제에 직면하게 된다. 이에 이러한 한계를 극복하기 위해 샴 네트워크 구조를 활용하여 문장 전체의 맥락을 고차원 벡터로 변환하는 SBERT와

같은 임베딩 기술이 점차 주목받고 있다.

2.1.3 공공도서관 관점에서 기존 시스템의 한계 및 문제점

앞서 기술한 협업 필터링과 콘텐츠 기반 필터링 기술은 도서 추천의 효율성을 제고하였으나, 공공도서관 서비스의 본질적인 목적으로 볼 수 있는 정보 접근 및 활용의 다양성과 지식의 확장이라는 측면에서 기술적 한계를 가져올 수 있다. 이를 간단히 정리하면 다음과 같다.

첫째, 데이터 희소성 및 콜드 스타트 문제이다. 협업 필터링 방식은 충분한 이용자 피드백(평점, 대출 이력 등)이 축적되지 않은 신간 도서나 신규 이용자에게 추천을 적절하게 하지 못한다. 그리고 콜드 스타트 문제는 신규 사용자 유입과 신규 아이템 노출 측면 모두에서 추천 시스템의 실용적 가치를 저해하는 핵심 장애 요인으로 계속 보고되고 있어 딥러닝 기반의 다양한 접근법이 이를 완화하기 위해 제안되고 있다(Panda & Ray, 2022). 특히 평점 시스템이 활성화되지 않은 공공도서관 환경에서 이러한 제약은 더욱 두드러지며, 이는 대출 빈도가 높은 특정 인기 도서에만 추천이 집중되는 인기도 편향 현상을 심화시키게 된다(Yuan & Hernandez, 2023).

둘째, 의미적 격차와 문맥 파악의 한계이다. 기존의 콘텐츠 기반 필터링은 주로 형태소 분석을 통한 키워드 매칭 방식에 의존하고 있다. 이는 단어의 단순 일치 여부에 치우쳐 있어, 동의어나 유의어와 같은 언어적 다양성을 수용하지 못하는 한계가 있다. 예를 들면, 도서 줄거리 내의 개념이 유사하더라도 사용된 어휘가 다를 경우 이를 동일한 맥락으로 인식하지 못해 결과적으로 문맥적 유사성이 높은 도서를 추천 목록에서 누락시키는 결과를 초래하게 된다.

셋째, 선호 왜곡 및 필터 버블현상이다. 이용자의 과거 이력이나 인구통계학적 정보에 과도하게 의존하는 방식은 이용자의 실제 의도와 무관한 오(誤) 추천 결과를 유발할 수 있다. 예를 들어 성인 이용자가 아동 도서를 대출한 이력이 시스템에 반영될 경우, 이후 도서 추천 결과에서 아동 도서 위주로 편향되는 선호 왜곡 문제가 발생할 수 있게 되고, 유사한 성향의 아이템만을 반복적으로 노출하여 이용자의 지식 범위를 좁히는 필터 버블현상은 다양한 학문 분야를 탐색해야 하는 공공도서관 이용자의 지적 성장을 저해하는 요소로 작용이 될 수 있다.

이처럼, 기존의 키워드 기반 매칭이나 단순 이력 기반 추천 방식으로 도서의 내재적 의미를 온전히 파악하는 것은 어렵다. 도서관 이용자에게 다채롭고 의미 있는 독서 경험을 제공하기 위해서는 단어의 표면적 일치를 넘어 문장의 맥락을 정교하게 이해할 수 있는 임베딩 기술의 도입이 필수적이다. 이에 다음에서는 임베딩 기술, 특히 SBERT에 대해 살펴보고 기술하였다.

2.2 문장 임베딩 기술

전술한 키워드 기반 매칭의 한계를 극복하기 위해 최근 자연어 처리 분야에서는 단어의 표면적 일치를 넘어 문장 전체의 맥락을 고차원 벡터 공간에 투영하는 문장 임베딩 기술이 주목받고 있다(Reimers & Gurevych, 2019).

문장 임베딩 기술은 도서의 줄거리나 요약문과 같은 비정형 텍스트를 특징 공간의 좌표로

변환한다. 이 과정에서 AI, 인공지능, 머신러닝과 같이 어휘는 다르지만 의미적으로 연관된 개념들을 벡터 공간상에서 인접하게 배치함으로 의미론적 유사성을 식별할 수 있고, 구글이 제안한 BERT(Bidirectional Encoder Representations from Transformers, 이하 BERT)의 경우, 트랜스포머 구조를 기반으로 문맥의 양방향 의미를 학습하여 자연어 이해 분야에서 혁신적인 성능을 입증하였다(Devlin et al., 2018). 그러나 표준 BERT 모델은 두 문장 간의 유사도를 계산할 때 모든 문장 쌍을 모델에 입력해야 하는 교차 인코딩 방식을 사용하므로 대규모 데이터 세트에서 실시간 검색을 수행하기에는 연산 비용이 과다하다는 단점이 있다. 이를 해결하기 위해 제안된 SBERT는 BERT 구조에 샴 네트워크를 결합하여 문장을 고정된 크기의 벡터로 출력하는 구조를 취하고 있다(Reimers & Gurevych, 2019).

SBERT는 개별 문장을 독립적으로 인코딩하여 벡터화하므로, 생성된 도서 임베딩 값을 벡터 데이터베이스에 저장해 두면 코사인 유사도 연산을 통해 실시간에 가까운 고속 검색이 가능하다. 그리고 의미가 유사한 문장들을 벡터 공간상에서 근접하게 배치함으로 코사인 유사도 기반의 효율적인 의미론적 검색을 가능하게 한다(Reimers & Gurevych, 2019). 다시 말해, SBERT는 샴 네트워크 및 트리플렛 네트워크 구조를 통해 사전 학습된 BERT를 미세조정함으로 코사인 유사도를 통해 비교 가능한 유의미한 문장 임베딩을 생성한다. 이는 기계가 계산하는 수치와 인간의 의미 인식 사이의 간극인 의미적 격차를 근본적으로 해소하는 기술적 기제가 된다.

2.3 도서 추천 모델의 다양성과 의외성

2.1장과 2.2장에서 데이터 기반의 기존 도서 추천 모델과 내재적 의미 처리 문장 임베딩 기술을 살펴보고 분석하였다. 이를 통해 도서 추천 모델의 평가 지표는 전통적인 예측 정확도를 넘어 이용자의 정보 탐색 외연을 확장하기 위한 '다양성'과 '의외성'의 중요성이 강조된다고 판단되었다.

우선, 정확도가 이용자의 과거 선호도와 일치하는 아이템을 제시하는 척도라면, '의외성'은 정확도 중심의 추천 평가를 보완하는 핵심 지표로, 이용자가 예상하지 못한 유용하고 신선한 아이템을 발견하는 정도를 의미한다(Kotkov et al., 2020). 예를 들어 3,000명 이상의 이용자를 대상으로 수행한 대규모 실증 연구(Chen et al., 2019)에 따르면, 신규성, 예측 불가성, 관련성이 의외성에 유의미한 인과적 영향을 미치며, 의외성은 최종적으로 이용자의 만족도와 구매 의도를 향상시키는 것으로 나타났다. 지나치게 높은 정확도만 추구할 경우, 이용자는 이미 인지하고 있는 정보 범위 내에 갇히게 되는 필터 버블현상을 겪게 된다(Areeb et al., 2023).

그리고 '다양성'은 추천 리스트 내 아이템들이 얼마나 상이한 주제를 포괄하는지 나타내는 지표이다. Alhijawi et al.(2022)은 추천 시스템의 목표를 정확도, 다양성, 신규성, 의외성, 커버리지의 다섯 가지 차원으로 체계화하며 단일 정확도 목표 추구의 한계에 대해 지적하였다. 도서 추천에서 다양성의 확보는 이용자가 특정 장르에 편중되지 않고 다각적인 학문 분야를 접하게 함으로써 지적 성장을 도모하는 공공도서관의 공공적 가치와 직결된다. 더욱이, SBERT 기

반의 임베딩 기술은 텍스트의 맥락적 유사성을 유지하면서도 서로 다른 분류 기호를 가진 도서들을 연결할 수 있는 기술적 유연성을 제공하기에, 결과적으로 이용자의 독서 분포를 확장하는 핵심적인 역할로 수행할 것으로 볼 수 있다. 이에 본 연구에서는 공공도서관 이용자의 기존 관심사와 맥락적으로 연결되어 있으면서 주제 분류 측면에서 새로운 분야를 제안하는 '의외성 구현'과 '다양성 확보'를 통해 '지식 확장'이 달성될 수 있을 것으로 판단되어 이를 주된 관점으로 설정하여 3장에서 연구를 진행하였다.

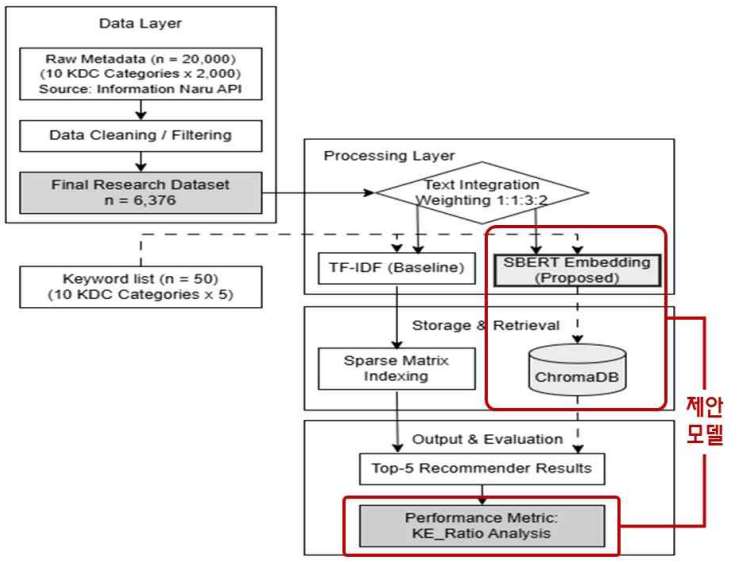
3. 지식 확장형 추천 모델 설계 및 구현

본 연구에서 구동하고자 하는 지식 확장형 추천 모델은 〈그림 1〉과 같이 구성하여 데이터 수집 → 전처리 → 임베딩 생성 → 벡터 DB 구축 → 유사도 검색의 단계를 거쳐 운영하였다.

3.1 데이터 세트 수집 및 통합 파이프라인

데이터 수집은 ① 도서관 정보 나루 기반의 실 이용 데이터를 확보하였고, ② 알라딘 API를 통한 도서 줄거리(Description) 메타데이터를 보강하였다. 이들을 바탕으로 ③ 데이터 정제를 진행하여 최종 데이터 세트를 확정하였다.

데이터 수집의 정확성과 공신력을 확보하기 위해 도서관 정보 나루의 오픈 API를 활용하였다. API 메뉴얼에 의하면 인기 대출 도서는 2014년 1월 1일 데이터부터 제공하고 있어, 본 연구에서는 2014년 1월 1일 ~ 2026년 1월 13일(조회일 하루 전) 사이 데이터를 대상으로 수집 기준이나 가중치는 따로 지정하지 않고 소급



〈그림 1〉 지식 확장형 추천 모델 아키텍처

된 데이터를 바탕으로 인기 대출 순위대로 산출하여 수집하였다. 그리고 KDC의 십진 분류 체계 간 데이터 불균형을 최소화하고자 대분류 기준 0~9까지 총 10개 카테고리를 순회하면서 수집을 실행하였고, 분류당 상위 20페이지를 요청하여 1페이지당 호출 건수(pageSize) '200'으로 지정, 대분류별 대출 상위 도서를 최대 4,000건으로 제한하는 임계치를 설정하여 총 40,000건 규모의 표집을 진행하였다. 이어서 도서관별 중복 레코드를 병합하고 서지 정보가 불충분한 데이터를 제외한 결과, 일차적으로 8,219건의 중간 데이터 세트를 구축하였다. 이때 총류(000) 등 정보 밀도가 높은 분야는 설정된 임계치를 충족하였으나, 문학(800) 분야의 경우 중복 제거 및 성인용 도서 선별 과정에서 가용 데이터 수가 연구에 필요한 최소 수준에 미치지 못함을 확인하였다. 이에 데이터 세트의 다양성과 추천의 형평성을 확보하기 위해, 해당 분야 800의 분야를 대상으로 반복적인 추가 API 호출 및 타겟팅 수집을 수행하여 특정 학문 분야에 치우치지 않는 균형 잡힌 추천 데이터베이스를 마련하였다. 그리고 수집된 기초 데이터의 줄거리 필드 누락 문제를 해결하기 위해 알라딘 오픈 API를 연동하였다.

수집된 도서 데이터베이스를 바탕으로 상세 서지를 호출하여 SBERT 임베딩에 필요한 풍부한 텍스트 데이터를 확인하였고, 이들 수집된 데이터를 정제한 결과, 최종적으로 6,376건의 도서 메타데이터를 확보하였다. 〈표 1〉에 제시하였듯이, 주류별 분포는 철학(100)과 사회과학(300) 분야의 비중이 높게 나타났고, 총류(000) 및 역사(900) 분야는 상대적으로 적은 분포를 보였지만, 추천 모델 구축을 위한 유의미한 표본 수를 충족하여 연구를 진행하였다.

3.2 데이터 전처리 및 품질 관리

앞서 3.1장에서 수집된 로우데이터의 노이즈를 제거하고 추천 모델의 학습 효율과 결과의 신뢰성을 높이기 위해 단계적인 전처리 과정을 진행하였다.

〈표 1〉 KDC 주류별 최종 데이터 세트 분포 현황

분류 기호	주류	권수(N)	비중(%)
000	총류	409	6.4
100	철학	885	13.9
200	종교	714	11.2
300	사회과학	838	13.1
400	자연과학	675	10.6
500	기술과학	713	11.2
600	예술	759	11.9
700	언어	484	7.6
800	문학	599	9.4
900	역사	300	4.7
합계		6,376	100.0

첫 번째, 아동용 도서와 저품질 메타데이터를 ISBN 및 텍스트 길이 기준을 바탕으로 필터링하여 제외하였다. 본 연구의 핵심 목적은 이용자의 지적 외연을 확장하고 학제적 탐색을 유도하는 것으로, 아동용 도서는 문장이 간결하고 주제의 복잡도가 낮아, 성인 이용자를 대상으로 한 지적 도약이나 맥락적 확장 등에 관한 효과를 측정하기에 정보 밀도가 부족하다고 판단하여 분석 대상에서 제외하였다. 확인 방식은 ISBN 부가 기호의 다섯 번째 자리가 '5'(아동용)로 분류된 도서 및 전집류를 자동 제외하는 ISBN 부가 기호 기반 필터링을 수행하였다. 이후 부가 기호만으로 걸러지지 않는 데이터를 위해 서명 및 저자 필드에서 '그림책', '만화', '애니메이션', '워크북' 등 아동용 자료를 시사하는 키워드를 추출하여 추가적으로 레코드를 소거하는 키워드 기반 정밀 필터링을 진행하였다. 그리고 마지막으로 도서 줄거리 데이터의 정보량 편차를 해소하고 연산 효율성을 제고하기 위해, 핵심 주제가 집중되는 서두 150자를 기준으로 텍스트를 정형화하였다.

두 번째, 출판사 통계 분석을 통한 총류(000) 데이터 편중 및 불균형을 해소하였다. 데이터 세트 구축 초기, KDC 대분류별 분포를 분석한 결과 총류(000) 분야가 타 분야에 비해 비정상적으로 밀집되어 나타나는 불균형 현상을 확인하였다. 해당 표본을 전수 조사한 결과, 공공도서관 대출 데이터 특성상 학습 만화 및 아동용 시리즈물이 총류(000) 및 기술 과학(400) 분야에 다수 포함되어 전체 분포를 왜곡하고 있음을 발견하였다. 이에 000대 도서를 발행하는 주요 출판사 리스트를 추출하여 데이터 정합성을 검토하였다. 그리고 최종적으로 '한빛미디어', '길벗', '이지스퍼블리싱' 등 실제 IT 및 컴퓨터 과학 전문 출판사의 데이터는 보존하되, 아동 전문 출판사 및 전집 위주의 발행사 데이터를 제거하는 전략적 필터링을 통해 데이터 세트의 학술적 순도를 높였다.

세 번째, KDC 분류 기호의 무결성 확보를 위한 자료형(String vs Float) 최적화를 수행하였다. 지식 확장 알고리즘의 정확도를 보장하기 위해 KDC 분류 체계의 기술적 결함을 수정하고 데이터를 표준화하였다. 데이터 수집한 초기, 일부 도서의 KDC 분류 기호가 '4'와 같이 단일 숫자로 기록되어 있거나 자릿수가 불일치하여 분류 체계의 위계성이 파괴된 데이터를 다수 발견하였다. 이에 이러한 비정형 데이터를 모델이 인식 가능한 3자리 표준 형식으로 변환하기 위해 다음과 같은 전처리를 수행하였다.

CSV 로드 과정에서 000~099번 대 데이터의 앞자리 '0'이 숫자형(Float/Int)으로 인식되어 소실되는 현상을 방지하고자 데이터 로드 시 「dtype={'kdc': str}」로 지정하여 데이터 왜곡을 차단하였고, 단일 숫자나 두 자리로 기록된 기호에 대해 「zfill(3)」 로직을 적용하여 '004', '018' 등 표준 3자리 코드로 정규화하였다.

3.3 모델의 아키텍처 및 알고리즘 구현

도서의 의미적 맥락을 정확하게 파악하기 위해 SBERT 구조를 채택하고, 좀 더 구체적인 구현을 위해 한국어의 다중 작업 학습이 완료된 「jhgan/ko- sroberta-multitask모델」(간정현, 2021)을 사용하여 본 연구에 실행하였다. 해당 모델을 선정한 기술적 근거와 특징은 다음과 같다.

기존의 통계적 방식인 Word2Vec 등은 단어의 표면적 일치 여부에 의존하여 문맥 파악에 한계가 있다. 이에 반해, SBERT는 문장 전체의 의미를 768차원의 밀집 벡터로 변환함으로 '비대면'과 '공간'처럼 어휘는 다르나 의미적 맥락이 유사한 도서를 포착하는 데 최적화되어 있다. 그리고 최신 거대언어모델은 뛰어난 성능을 보이나, 최종 정제된 6,300여 건의 도서 데이터를 매 검색 시 직접 비교하는 방식은 막대한 연산 비용과 토큰 제한으로 인해 실제 서비스 적용에 한계가 있다. 반면, SBERT는 고정된 벡터 추출을 통해 사전 인덱싱이 가능하여 효율적인 추천 모델을 구축하는 데 효율적이다.

이어서 임베딩된 고차원 벡터 데이터를 효율적으로 저장하고 검색하기 위해 오픈소스 벡터 데이터베이스 ChromaDB를 구축하였다. 기존의 단순 선형 검색 방식은 데이터 규모 증가에 따라 응답 시간이 기하급수적으로 늘어나는 한계가 있는 반면에, ChromaDB는 벡터 인덱싱 기술을 통해 수만 건의 데이터 중 사용자의 쿼리와 가장 유사한 벡터를 초 단위 이하로 탐색할 수 있는 공학적 확장성을 제공한다. 그래서 모델 작동 시 사용자의 입력 쿼리가 들어오면 동일한 SBERT를 통해 벡터화한 후, 〈식 3-1〉과 같이 데이터베이스 내 도서 벡터들과의 코사인 유사도를 계산하여 상위 추천 리스트를 산출하도록 설계하였다.

〈식 3-1〉

$$Similarity = \cos(\theta) = \frac{A \cdot B}{||A|| \, ||B||}$$

3.4 지식 확장을 위한 입력 데이터 전처리

3.4.1 서명 편향성 완화를 위한 줄거리 밀도 제어

우선, 초기 실험 단계에서 도서의 서명(Title), 저자명(Author), 줄거리(Description) 필드에 텍스트를 단순 결합하여 SBERT 모델의 입력 데이터를 구성하였다. 그러나 분석 결과, 검색어와 서명의 단어가 직접적으로 일치하는 도서들이 검색 결과의 상위권을 독점하는 제목 편향성 현상이 관찰되었다. 이는 사용자가 입력한 어휘의 표면적인 매칭에만 집중하게 하여 본 연구의 목적인 '심층 문맥 기반의 지식 확장'을 저해하는 요소로 작용하였다. 이에 본 연구에서는 이러한 서명 편향성을 완화하고 도서 본연의 맥락을 반영하기 위해, 입력 텍스트 구성 시 필드별 정보 밀도를 제어하는 전처리 방식을 도입하였다. 구체적으로는 도서의 주제와 핵심 내용이 가장 풍부하게 기술된 줄거리 필드의 토큰 비중을 인위적으로 확장하였으며, 수립된 가중치 기반의 텍스트 공식은 〈식 3-2〉와 같다.

〈식 3-2〉

$$\mathrm{Text}_{\mathrm{combined}} = \mathrm{Title} + \mathrm{Author} + (\mathrm{Description} \times 3)$$

본 방식은 입력단 수준에서 줄거리 데이터를 3회 중첩하여 전체 문자열 내 줄거리 토큰의 밀도를 의도적으로 높이는 전처리 조치이다. SBERT 모델이 텍스트를 인코딩할 때 단편적인 서명 키워드에만 과도하게 집중되는 현상을 억제하고, 줄거리 내에 내포된 심층적 유사도를 보다 우선적으로 반영하도록 유도하였고 이러한 전

처리 과정을 거친 결과, 서명에 검색어가 포함되지 않더라도 줄거리 내의 내용과 의미적 유사도가 높은 도서들이 추천 결과에 포함될 수 있는 장치를 마련하였다.

3.4.2 KDC 연계어 강화를 통한 의미적 노이즈 제어

이어서 가중치가 적용된 의미 유사도와 더불어 사용자를 새로운 학문 분야로 안내하기 위한 지식 확장 가중 점수 로직을 도입하였다. 이 로직은 단순한 유사도 순위 산출을 넘어 학제적 발견을 유도하기 위한 하이브리드 스코어링 방식으로, 최종 추천 점수($Total_{Score}$)는 〈식 3-3〉과 같은 산출식에 의해 결정된다. 여기서 각 변수의 정의와 역할은 다음과 같다.

우선, $Total_{Score}$(최종 추천 점수)는 사용자의 검색 의도에 부합하는 내용의 정확성과 주제의 다양성을 동시에 고려하여 재정렬한 추천 시스템의 최종 점수를 나타내고, $Semantic_{Similarity}$(의미 유사도)는 앞서 〈식 3-1〉에서 정의한 코사인 유사도 연산 값으로 SBERT 모델을 활용하여 사용자의 입력 쿼리와 도서 줄거리 간의 문맥적 유사성을 측정한 점수이다. 이 값은 0과 1 사이의 범위를 가지며, 사용자의 검색 의도와 도서의 의미가 일치할수록 1에 가깝고, 연관성이 없을수록 0에 가까운 값을 나타낸다. 그리고 $Is_{Expansion}$(지식 확장 여부 플래그)는 추천 후보 도서가 사용자의 기존 관심 분야를 벗어나, 새로운 학문적 발견을 유도할 수 있는 KDC 대분류에 속하는지 판별하는 변수를 보여준다. 후보 도서가 기존과 다른 새로운 KDC 대분류에 해당할 경우 1, 기존 분야와 동일할 경우 0의 값을 가지게 된다.

지식 확장 가중치(ω)는 의미적 유사성과 주제적 다양성 간의 균형을 조절하기 위한 조정 변수로 설정하였다. 본 연구에서 제안하는 50개의 검색 키워드는 단일 학문 분야에 완전히 종속되지 않고 여러 학제적 성격이 혼합된 융합형 키워드로 구성되어 있다. 이는 기계적인 KDC 대분류 일치 여부만을 기준으로 파라미터를 최적화할 경우, 실제 추천 목록의 내용적 연관성을 제대로 반영하지 못하는 한계가 있다. 이에 본 연구에서는 기계적인 KDC 분류 일치율에만 의존하는 파라미터 튜닝을 지양하고 최종 제안 모델을 통해 도출된 추천 도서 목록의 시맨틱 맥락 연관성을 연구자가 직접 전수 검증하는 정성적 최적화 프로세스를 거쳤다. 검증 결과, 이용자의 검색 의도를 정확하게 충족하는 의미적 연관성을 유지하면서도 분류 확장이 과도하게 발생하지 않는 균형점으로서 '$\omega = 0.15$'를 최종 값으로 설정하였다.

본 연구가 지향하는 지식 확장형 추천은 사용자의 본래 검색 의도를 유지하면서도 새로운 분야를 발견하도록 유도하는 것이다. 따라서 완전히 무작위한 분야가 추천되거나 기존 분야에만 고착되는 것을 방지하기 위해, 최종 추천 목록이 연관 도서와 타 분야 도서 간의 약 50% 내외의 균형을 이루도록 정성적 목표를 설정하였으며, $\omega = 0.15$는 이러한 실증적 관찰을 통해 도출된 경험적 최적값으로, 지식 확장을 유도하되 추천의 적합성을 저해하지 않기 위한 기준으로 활용되었다.

〈식 3-3〉

$$Total_{Score} = Semantic_{Similarity} + (\omega \times Is_{Expansion})$$

4. 실험 결과 및 분석

4.1 연구 실험 환경 및 성능 지표 정의

본 연구 제안모델의 성능을 객관적으로 검증하기 위해 3장에서 구축한 모델과 ChromaDB를 기반으로 성능 평가를 수행하였다. 실험 데이터는 최종 정제된 6,376건의 도서 메타데이터를 활용하였고, 제안모델의 우수성을 입증하기 위해 전통적인 키워드 매칭 방식인 TF-IDF 모델을 비교군으로 설정하여 동일한 실험을 수행하였다. 모델의 정확도와 지식 확장 가능성을 동시에 측정하기 위해 다음과 같은 지표를 활용하였고 각 지표의 산출식은 〈식 3-4〉, 〈식 3-5〉와 같다. 그리고 각 검색 키워드에 대해 출력하는 '상위 5개의 추천 도서'를 분석 대상으로 한정하여 모든 n은 5를 기준으로 산출되었다.

〈식 3-4〉

$$Semantic\ Recall = \frac{\text{검색 의도에 부합하는 도서 수}}{\text{전체 추천 도서 수}(N)}$$

〈식 3-5〉

$$KE_{Ratio} = \frac{\text{검색 키워드와 다른 } KDC \text{ 대분류 도서 수}}{\text{전체 추천 도서 수}(N)} \times 100$$

본 모델에서 사용자의 검색 의도를 정확히 반영하는지 확인하기 위해 검색어별로 추천된 상위 5권의 도서 중 검색 의도와 시맨틱 맥락이 일치하는 도서의 비율을 의미하는 의미적 재현율을 정의하였다. 동시에 지식의 외연 확장 정도를 측정하고자 지식 확장 지수를 고안하였다. 이는 추천된 5권의 도서 가운데 검색 키워드의 KDC 대분류와 일치하지 않는 타 분야 도서가 차지하는 비중을 계산한 수치를 나타낸다.

실험의 일관성을 위해 본 연구에서 키워드를 선정하였고, 키워드에 대한 기준은 국립중앙도서관의 주제명표목표를 참조하고 분류 번호를 우선하여 KDC 10개 주류 기준으로 분야별 5개씩 선정하였다. 본 연구의 목적 '지식 확장'을 실천하고자 단일 단어가 아닌 2~3개 어절로 이뤄진 단어이자 중심적인 의미를 나타내고 있는 키워드로 구성하고자 하였다. 단일 단어로 구성된 키워드는 주제 선정에 한계성을 보일 수 있기에, 이를 방지하고자 어절로 조합하였다. 예를 들면 단일 키워드 '인간 존재론' 관련 추천 도서를 찾을 경우, 결과적으로 주제가 철학 관련 도서에서 집중될 수 있어 조합의 형태 「단어 A + 단어 B」로 형성하였다. 단어 A가 총류일 경우, 단어 B는 총류(000)부터 역사(900)까지 설정하여 「인간 존재론(철학: 100) + AI(총류: 000)」 = 「AI 인간 관계론」과 같이 키워드를 작성하였고, 다각도적 관점에서의 키워드 선정을 위해 국립중앙도서관의 전거 브라우징, 포털 사이트 검색창 및 검색어 등을 활용하였다. 이와 같은 방법을 통해 본 연구에서 사용할 총 50개 최종 키워드를 구성하였다.

4.2 제안모델의 정량적 평가 결과

앞서 지표를 활용하여 제안모델에 대한 성능을 검증하였고 이를 측정한 결과, 의미적 정확성과 지식 확장성 측면에서 모두 유의미한 수치를 도출하였다. 특히, 전통적인 TF-IDF 방식과의 비교를 통해 제안모델이 가진 '유의미한 지식 확장'의 특성을 확인하였다. 정량적 평가 결과는 〈표 2〉와 같다.

〈표 2〉 실험 키워드(50개) 기반 성능 평가 요약

평가 지표	제안 모델 (SBERT)	비교모델 (TF-IDF)	비고
의미적 재현율	94.2%	67.2%	추천된 도서의 주제 적합성
지식 확장 지수	45.6%	63.2%	타 학문 분야 도서 추천 비율
검색어 당 평균 확장 권수	2.28권	3.16권	5권 중 평균 확장된 도서 수

우선, 평균 의미적 재현율은 94.2%로, 이는 SBERT 기반의 고차원 벡터 임베딩 모델이 검색어의 맥락을 정확히 파악하여 키워드가 직접적으로 포함되지 않은 도서라도 주제적으로 일치하는 결과를 도출하고 있음을 의미한다. 이에 반해 TF-IDF 방식은 형태소의 단순 일치 여부에 의존하는 특성으로 인해 의미적 재현율이 67.2%에 그치며 상대적으로 낮은 정밀도를 보였다. 그리고 지식 확장 지수는 제안모델이 45.6%, TF-IDF 모델이 63.2%로 나타났다. 수치상으로는 TF-IDF가 높은 확장성을 보이고 있으나, 이는 검색어 내 특정 단어, 예를 들면, 문화, 디지털 등과 같은 단어에 과도한 가중치가 부여되어 주제 맥락과 전혀 무관한 타 분야 도서가 추천되는 '가짜 확장' 현상이 합산된 결과로 분석된다. 결과적으로 제안 모델은 〈식 3-5〉에 의한 평균 2.28권의 확장을 기록하며, 의미적 정합성을 유지하는 범위 내에서의 정교한 확장을 달성했음을 확인하였다.

4.3 제안모델의 정성적 사례 분석

4.3.1 제안모델의 지식 확장 성공 사례

앞서 수치로 증명된 정량적 성과에 이어서 본 장에서는 제안모델이 실제 추천 목록을 생성하는 논리적 과정을 심층적으로 분석하여 기술하였다. 본 제안모델은 서명과 저자, 줄거리 필드에 대해 차등적인 가중치를 부여함으로써, 단어의 표면적 일치를 넘어 도서의 내재적 맥락을 정교하게 포착할 수 있었다.

가장 주목할 만한 사례로 첫 번째는 심리학적 키워드 '번 아웃 긍정 심리학(철학: 100)'의 검색 결과에서 관찰되었다. 제안모델은 단순히 개인의 심리 상태를 다룬 원론서에 머물지 않고, 무기력한 삶에 대한 실천적 개입과 사회적 행동을 다룬 사회과학(300) 분야의 도서 「내 인생 구하기」를 상위권으로 도출하였다. 이는 본 모델이 '번 아웃'이라는 현상을 개인의 내면적 문제를 넘어 사회적 관계와 환경 속에서 해결해야 할 과제로 맥락화 했음을 의미한다. 즉, '내면의 성찰(철학: 100)'에서 '사회적 실천(사회과학: 300)'으로 이용자의 지적 탐색 영역을 넓히는 유의미한 학제적 도약을 실현한 사례로 평가된다.

그리고 두 번째로 철학적 난제 키워드 '포스트 휴머니즘 기계 윤리(철학: 100)'는 질의에서 정교한 맥락 파악 능력을 입증하였다. 본 모델은 추상적인 윤리 담론을 넘어 실제 인공지능이 인간을 지배하거나 영향을 미치는 메커니즘을 다룬 총류(000) 분야의 도서 「알고리즘 포비아」를 추천하였다. 이는 형이상학적인 기계 윤리 개념이 현대 정보사회에서 알고리즘이라는 구체적인 기술로 어떻게 구현되고 작동하는지 의미론적으로 연결한 결과로, 철학적 사

유(100)를 정보 과학적 실체(000)로 확장하여 이용자에게 다차원적인 탐색 기회를 제공하는 의외성을 성공적으로 구현하였다.

마지막으로 예술(600) 분야의 '신화 서양 예술' 질의에 대해서도 정교한 맥락 파악 능력을 입증하였다. 모델에서는 단순한 서양 미술사 도서를 넘어, 예술 작품의 모티프가 되는 그리스·로마 신화의 서사를 깊이 있게 다룬 종교/신화(200) 분야의 도서 「미술관에서 읽는 그리스 신화」를 1순위로 제안하였다. 이는 시각 예술의 미적 감상을 그 근간이 되는 인문학적 서사와 연결한 결과이며, 예술(600)과 종교/신화 사이의 시맨틱 연결망을 통해 지식의 깊이를 더하는 효과적인 확장을 수행하였음을 보여준다.

4.3.2 비교모델의 한계 및 관련 사례

전통적인 통계 빈도 기반 모델 TF-IDF는 형태소의 단순 일치에 의존하기 때문에 KDC의 주제 분류상에서 타 분야로 이탈하더라도 의미적 정합성이 완전히 상실되는 '가짜 확장' 현상이 빈번하게 관찰되었다. 이는 단어의 중의성을 해소하지 못하고 통계적 수치에만 반응하는 기술적 한계에서 기인한다.

대표적인 사례로 키워드 '메타버스 기록 관리(총류: 000)'의 검색 결과에서 확인할 수 있었다. TF-IDF 모델은 검색어 내 '관리'라는 단어의 빈도에 과도하게 반응하여 정보 기술적 맥락과는 전혀 무관한 재테크 도서 「15억 작은 부자 현주씨의 돈 관리 습관」(사회과학: 300)을 추천하였다. 이는 수치상으로는 총류(000)에서 사회과학(300)으로의 확장이 일어난 것처럼 기록되나, 실제로는 디지털 데이터 보존 및 관리라는 이용자의 본질적 검색 의도를 상실한 정보 노이즈에 해당한다. 또한 키워드 '에듀테크 자기주도 학습(사회과학: 300)'의 질의에 대해서도 TF-IDF는 교육 기술적 맥락을 파악하지 못하고 단순히 '학습'이나 '교과서'라는 단어가 포함되었다는 통계적 근거만으로 역사/여행(900) 분야의 도서 「교과서가 쉬워지는 주말여행」을 제안하는 한계를 보였다. 이는 이용자가 의도한 현대 교육 공학의 탐색과는 무관한 기계적인 키워드 매칭의 결과이며, 단순한 수치상의 분류 이탈이 정보 접근의 다양성이나 이용자의 지적 외연 확장으로 직결되는 것이 아님을 보여준다. 즉, 결과적으로 비교모델은 익숙한 주제 내에서의 검색에 유효할 수 있으나, 의미론적 연결 고리가 필수적인 지식 확장 과정에서는 무분별한 이탈률을 생성하는 한계를 드러냈다. 이처럼 키워드 중심의 한계로 인해 발생하는 이러한 정합성 저하 문제는 도서의 내재적 맥락을 벡터 기반으로 분석하여 추천 품질을 유지하는 본 모델과 구별되는 점이라 할 수 있다.

4.4 시사점

기존의 도서 추천 시스템은 주로 협업 필터링이나 키워드 기반 콘텐츠 분석을 통해 이용자의 과거 선호를 반복적으로 반영하는 방식이다. 그러나 이러한 접근은 추천의 안정성과 예측 가능성을 확보하는 데에 효과적이나 이용자의 관심 영역을 구조적으로 확장하는 데에는 한계를 지니고 있다. 특히, 본 연구의 비교 실험을 통해 알 수 있듯이 전통적인 TF-IDF 방식은 형태소 일치에 의존함으로 검색어의 중의성을 해소하지 못하고 맥락 없는 도서를 추천하

는 확장의 한계를 보여주고 있다. 이에 비해 제안한 SBERT 기반 추천 모델은 도서 요약문의 문맥적 의미를 벡터 공간에서 해석함으로 어휘의 직접적 일치 여부와 무관하게 주제적 연관성을 포착할 수 있다는 점에서 차별성을 가지고, 단순히 카테고리 이탈률을 높이는 것을 넘어 추천 정확도와 지식 확장의 균형을 알고리즘적으로 구현했다. 즉, 비교모델에 비해 본 모델의 기술적 장점이 확인되었다고 할 수 있다. 이에 다음에서는 SBERT 기반 도서 추천 모델의 실험 결과를 바탕으로 크게 2개의 관점, 기술적 유효성과 공공도서관 서비스 측면에서 시사점을 적용하여 논의하고자 하였다.

첫 번째, 실험 결과를 통해 SBERT를 활용한 문장 임베딩 방식은 2.1.3장에서 지적한 의미적 격차를 효과적으로 해소할 수 있다는 것을 확인하였다. 기존의 키워드 기반 시스템이 '변화'나 '흐름'과 같은 일반 명사에 과도하게 반응하여 검색 품질이 저하되었던 것과 달리, 본 모델은 줄거리의 맥락을 벡터 공간에 투영함으로 어휘가 다르더라도 주제적 연관성이 높은 도서를 정확히 식별하였다. 이는 대규모 도서 데이터 세트(6,376권) 내에서 유사도 연산을 수행할 때, 단순한 키워드 매칭보다 훨씬 정교한 문맥적 유사성 산출이 가능함을 시사하며 의미론적 검색을 통한 기술적 한계를 극복했다고 볼 수 있다. 이어서 두 번째, 실험 과정을 통해 도서 추천 전후의 KDC 분포가 45.6%로 변화함으로 이용자가 기존에 선호하던 특정 장르나 주제에 고착되지 않고 새로운 학문적 분야로 관심을 넓혔다는 점을 정량적으로 증명하였다. 특히, 비교모델에서 나타난 무분별한 카테고리 이탈과 대조적으로 본 제안모델은 이용자의 검색 맥락을 보존하면서 타 학문 분야를 연결하는 통제된 의외성을 실현하였고, 추천 모델이 지나치게 정확도를 추구하는 경우 발생하는 필터 버블현상을 억제하는 동시에 의미론적 연관성을 전제로 한 지식 확장을 통해 추천 결과에 대한 정성적 적합성을 확보할 수 있었다.

마지막으로 세 번째, 본 연구의 결과를 통해 도서관의 맞춤형 서비스 고도화에 중요한 근거를 제공하였다. 평점 데이터가 부족한 도서관 환경에서 도서의 요약문만으로 단순 키워드 검색보다 월등한 품질의 고성능 추천이 가능하다는 점을 입증하였고, TF-IDF 방식에서 빈번하게 발생하는 맥락 없는 오(誤) 추천을 필터링함으로 이용자에게 좀 더 전문적이고 신뢰할 수 있는 학제적 탐색 경험을 제공할 수 있도록 하였다.

5. 결론 및 제언

급변하는 정보기술의 발전으로 다양한 포털 시스템, 플랫폼, 데이터베이스 등에서 개인화 환경을 적극 도입하여 서비스하고 있다. 이와 같은 시스템은 도서관에서도 '도서 추천'으로 적용되어 이용자에게 이용자의 검색 데이터를 바탕으로 관련 키워드 혹은 주제 도서를 서비스하고 있다. 그래서 본 연구에서는 기존 도서관 추천 시스템이 가진 키워드 매칭의 한계를 극복하고, 이용자의 지식 외연을 확장할 수 있도록 도움을 줄 수 있는 도서 추천 모델을 제안하고자 하였다.

우선, 키워드 매칭이나 단순 필터링 방식에서 벗어나 도서의 줄거리와 요약문과 같은 비정형 텍스트를 분석하는 SBERT 기반의 문장

임베딩 기술을 핵심 방법론으로 채택하였고, 이 과정 가운데 추천 결과의 분석을 위한 분류 체계로 국내 공공도서관에서 주된 표준으로 활용되는 KDC를 기준으로 설정하여 추천 전후의 주제 분포 변화를 정량적으로 측정, 범위를 제한하였다. 이 연구를 통해 도출된 주요 결과와 시사점은 다음과 같다.

첫째, 도서의 제목이나 단순 키워드에 의존하던 기존 방식이 아닌 도서 요약문의 문맥적 의미를 파악하는 SBERT 모델을 적용함으로 '의미적 격차'를 효과적으로 해소하였고, 이용자가 입력한 검색어와 표면적인 단어가 일치하지 않더라도 내용 간에 연관성을 가진 도서를 정교하게 추출할 수 있다는 것을 확인하였다. 둘째, 추천 전후의 KDC 분포를 분석한 결과, 약 45.6%의 지식 확장 지수를 확인하였다. 비교 실험을 통해 전통적인 키워드 매칭 방식이 검색어 내 일반 명사에 반응하여 주제와 무관한 도서를 추천하는 가짜 확장의 한계를 보인 것과는 다르게 제안모델은 맥락적 유사성을 유지하며 도출된 통제된 의외성을 통해 추천의 정확도와 지식 확장이라는 두 지표 사이에서 기술적으로 정교한 균형을 확보하였다. 그리고 셋째, 결과적으로 제안모델을 통해 본 연구의 목적이자 도서관의 공공적 가치 '지적 탐색의 자유'와 공공도서관의 본질적 역할 '균형 잡힌 정보 접근', '지적 성장 지원'을 기술적으로 구현할 수 있음을 시사하였다.

다만, 공공도서관 추천 서비스의 지식 확장 가능성은 확인하였으나 실험 과정 및 환경적 제약으로 인해 다음과 같은 한계점이 발생하였다.

우선, 첫 번째 실험 데이터 세트가 대중적 인기 도서에 편중되어 구성되었다. 전국 통합 대출 데이터를 활용해 보편성을 확보하였으나 수집 방식의 특성상 인기도 편향이 발생하게 되었기에, 추후 학술적·학제적 가치가 높은 롱테일 도서나 전문 도서관의 장서 데이터를 포함하여 데이터 스펙트럼을 확장할 필요가 있다. 두 번째, 도서 요약문 데이터의 정보량 및 품질에 한계가 발생하였다. 활용한 API의 요약문이 200~300자 내외의 단문 위주로 구성되어 정교한 의미 추출에 제약이 있어 후속 연구에서는 서평, 목차, 리뷰 등 고밀도 텍스트 데이터를 결합하여 모델의 의미론적 변별력을 강화해야 할 것이다. 그리고 세 번째, 본 연구에서는 모델을 제안하였으나 실질적인 이용자 기반의 질적 평가를 실행하지 못했다. KDC 분포 변화를 통한 정량적 검증에 집중하여 실제 이용자의 만족도 조사나 장기적 형태 변화를 추적하지 못했기에 앞으로 현장 피드백을 반영한 실증 연구가 병행된다면 도서관 개인화 서비스의 실질적인 표준을 제시할 수 있을 것이고, 이 모든 과정은 기술적 관점에서 SBERT 모델을 도서관 도메인 특화 데이터로 파인 튜닝을 하거나 대형언어모델과의 결합을 통해 문맥 이해도를 한층 고도화하는 연구가 동시다발적으로 병행되어야 할 것이다. 마지막으로 네 번째, 본 연구는 연구자의 정성적 평가로 진행하였다. 모든 과정은 연구자의 객관적인 분석으로 실행했으나 이는 연구 결과의 신뢰성에 한계를 가져올 수 있는 부분이다. 이에 본 연구 결과의 정확성과 타당성을 제공하고 신뢰성을 높일 수 있도록 후속 연구를 진행하고자 한다. 본 연구 결과를 토대로 명확한 평가 기준 설계하고 실무 전문가의 검증을 실행하여 본 연구의 성과를 최종적으로 확인할 예정이다.

참 고 문 헌

간정현 (2021). Ko-sroberta-multitask. Hugging Face.
출처: https://huggingface.co/jhgan/ko-sroberta-multitask

국립중앙도서관 (2024). ISBN/ISSN한국문헌번호지침. 국립중앙도서관.
출처: https://www.nl.go.kr/

국립중앙도서관 (2024). 도서관 정보 나루: 도서관 빅데이터 분석 플랫폼. 국립중앙도서관.
출처: https://www.data4library.kr/

서민음 (2025.03.31). 책 추천・요약・정리까지 한번에…밀리의서재 "AI, 독서 경험에 혁신". 아시아경제. 출처: https://view.asiae.co.kr/article/2025033109225671336

임정훈, 조창제, 김종헌 (2022). 연관규칙을 활용한 학교도서관 도서추천시스템 개발에 관한 연구. 정보관리학회지, 39(3), 1-22. https://doi.org/10.3743/KOSIM.2022.39.3.001

한국도서관협회 (2013). 한국십진분류법 (제6판). 서울: 한국도서관협회.

Alhijawi, B., Awajan, A., & Fraihat, S. (2022). Survey on the objectives of recommender systems: measures, solutions, evaluation methodology, and new perspectives. ACM Computing Surveys, 55(5), 1-38. https://doi.org/10.1145/3527449

Areeb, Q. M., Nadeem, M., Sohail, S. S., Imam, R., Doctor, F., Himeur, Y., Hussain, A., & Amira, A. (2023). Filter bubbles in recommender systems: fact or fallacy—A systematic review. WIREs Data Mining and Knowledge Discovery, 13(5), e1512. https://doi.org/10.1002/widm.1512

Chen, L., Yang, Y., Wang, N., Yang, K., & Yuan, Q. (2019). How serendipity improves user satisfaction with recommendations? A large-scale user evaluation. In The World Wide Web Conference(WWW '19), 240-250. https://doi.org/10.1145/3308558.3313469

Chroma (2023). Chroma: the AI-native open-source embedding database. Chroma. Available: https://docs.trychroma.com/

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

IFLA (2012). IFLA code of ethics for librarians and other information workers. IFLA. Availbale: https://repository.ifla.org/items/2b9665ee-c48b-4adf-883d-8178c3775f8e

Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 383-390. https://doi.org/10.1145/3306618.3314288

Klimashevskaia, A., Jannach, D., & Elahi, M. (2024). A survey on popularity bias in recommender systems. User Modeling and User-Adapted Interaction, 34, 1777-1834. https://doi.org/10.1007/s11257-024-09406-0

Kotkov, D., Veijalainen, J., & Wang, S. (2020). How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm. Computing, 102(2), 393-411. https://doi.org/10.1007/s00607-018-0687-5

Panda, D. K. & Ray, S. (2022). Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review. Journal of Intelligent Information Systems, 59(2), 341-389. https://doi.org/10.1007/s10844-022-00698-5

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing(EMNLP), 3982-3992. https://doi.org/10.18653/v1/D19-1410

Xie, H., Qin, Z., Li, G. Y., & Juang, B. H. (2021). Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing, 69, 2663-2675. https://doi.org/10.1109/TSP.2021.3071210

Yuan, H. & Hernandez, A. A. (2023). User cold start problem in recommendation systems: a systematic review. IEEE Access, 11, 136958-136977. https://doi.org/10.1109/ACCESS.2023.3339320

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: a survey and new perspectives. ACM Computing Surveys, 52(1), 1-38. https://doi.org/10.1145/3285029

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Gan, Jung-Han (2021). Ko-sroberta-multitask. Hugging Face. Availbale: https://huggingface.co/jhgan/ko-sroberta-multitask

Korean Library Association (2013). Korean Decimal Classification and Relative Index (6th ed.). Seoul: Korean Library Association.

Lim, Jeong-Hoon, Cho, Chang-Je, & Kim, Jong-Heon (2022). A study on the development

of the school library book recommendation system using the association rule. Korean Society for Imformation Management. Journal of the Korean Society for Information Management, 39(3), 1-22. https://doi.org/10.3743/KOSIM.2022.39.3.001

National Library of Korea (2024). ISBN/ISSN Korean Literature Number Guidelines. National Library of Korea. Availbale: https://www.nl.go.kr/

National Library of Korea (2024). Library Bigdata. National Library of Korea. Availbale: https://www.data4library.kr/

Seo, Mideum (2025.03.31). Book recommendation, summary, and organization at once...Millie "AI, Innovates in Reading Experience". Asian Economy. Availbale: https://view.asiae.co.kr/article/2025033109225671336

[부록 1] 실험 키워드 50개에 대한 추천 도서 예시

쿼리	도서명	저자	KDC 대분류	유사도거리	유사도점수
도서관 디지털 리터러시	디지털 시대의 문헌정보학 연구 - 핵심 지식정보원안내	김태승(지은이)	총류	0.4196	0.5804
빅데이터 현대 저널리즘	문헌정보학의 연구방법론	임태삼(지은이)	총류	0.4044	0.5956
인공지능 연구 윤리	수상한 인공지능: AI는 세상을 어떻게 바꿀까?	스테퍼니 맥퍼슨 지음 ; 이가영 옮김	총류	0.3512	0.6488
메타버스 기록 관리	문헌정보학 연구의 현황과 과제	미타도서관 정보학회(지은이), 오동근(옮긴이)	총류	0.5658	0.4342
오픈소스 지식 공유	(비전공자를 위한 이해할 수 있는) IT 지식	최원영 지음	총류	0.4392	0.5608
AI 인간 존재론	AI 리터러시	김용성 지음	총류	0.4055	0.5945
번 아웃 긍정 심리학	나를 지켜내는 연습: 나를 단단하게 만드는 심리 처방전	브리애나 위스트 지음 ; 이상원 옮김	철학	0.4987	0.5013
실존주의 현대인 불안	어른이 되면 괜찮을 줄 알았다	김혜남, 박종석 지음	철학	0.4396	0.5604
인지 편향 의사결정	신경 끄기의 기술: 인생에서 가장 중요한 것만 남기는 힘	마크 맨슨 지음 ; 한재호 옮김	철학	0.5246	0.4754
포스트 휴머니즘 기계 윤리	인공지능은 선생님을 대신할까요?	이영호, 2DA 그림	총류	0.5248	0.4752
과학 종교 변화	박문호 박사의 빅히스토리 공부	지은이: 박문호	자연과학	0.4787	0.5213
명상 스트레스 관리	알아차림에 대한 알아차림	루퍼트 스파이라 지음 ; 김주환 옮김	철학	0.4275	0.5725
신화 서양 예술	세계의 신과 신화	조영경 글 ; 김혜연 그림	종교	0.3628	0.6372
기후 위기 생태 신학	기후 위기:지구 말고 지구인이 달라져야 해	소이언 지음 ; 김진화 그림	기술과학	0.4498	0.5502
동서양 사후 세계관	지적 대화를 위한 넓고 얕은 지식	채사장 지음	총류	0.4492	0.5508
저출산 복지 정책	정서적 흙수저와 정서적 금수저	최성애, 조벽 지음	기술과학	0.6378	0.3622
무역 갈등 공급망	'좋아요'는 어떻게 지구를 파괴하는가	기욤 피트롱 지음 ; 양영란 옮김	기술과학	0.6142	0.3858
SNS 정치 양극화	(SNS와 유튜브 1인 미디어를 위한) 스마트폰 활용법	김경수, 황세웅 지음	총류	0.586	0.414
에듀테크 자기주도 학습	혼자 공부하는 R 데이터 분석	강전희, 엄동란(지은이)	총류	0.4045	0.5955
프라이버시 헌법 가치	이토록 다정한 개인주의자	함규진 지음	철학	0.6443	0.3557
양자 역학 컴퓨터	컴퓨터 구조 - 프로그래밍 관점에서 바라보는 컴퓨터 구조	정기철(지은이)	총류	0.4443	0.5557
유전자 가위 윤리	AI 에이전트 시대의 새로운 상식	공훈의(지은이)	총류	0.6096	0.3904
기후 변화 해양 생태	이상한 기후, 그래서 우리는?	크리스티나 헬트만 지음 ; 유영미 옮김	기술과학	0.4264	0.5736

쿼리	도서명	저자	KDC 대분류	유사도거리	유사도점수
우주 망원경 은하 진화	브리태니커 지식 백과	엮음: 크리스토퍼 로이드 ; 옮김: 한국 백과사전연구소	총류	0.4117	0.5883
딥러닝 신약 개발	AI 버블이 온다 - 우리는 진짜 인공지능을 보고 있는가?	아르빈드 나라야난, 사야시 카푸르(지은이), 강미경(옮긴이)	총류	0.4577	0.5423
스마트 시티 디지털 트윈	(세상을 읽는 커다란 눈) 알고리즘	플로랑스 피노 글 ; 허린 옮김	총류	0.5794	0.4206
스마트 팜 식량 안보	공장식 농장, 지구가 아파요!	데이비드 웨스트, 이종원 옮김	기술과학	0.6258	0.3742
탄소 중립 수소 에너지	물 아저씨는 변신쟁이	글·그림: 아고스티노 트라이니 ; 번역: U&J	자연과학	0.6844	0.3156
원격 의료 공공 보건	알고 싶니 마음, 심리툰: 사람 마음이 약으로만 치료되나요?	지은이: 팔호광장	철학	0.634	0.366
3D 프린팅 제조 혁신	마인크래프트 건축 장인 시간 여행	구성: 제이크 터너, 번역: 아보카도	예술	0.5029	0.4971
AI 예술 저작권	된다! 생성형 AI 사진 & 이미지 만들기	김원석, 장한결 지음	총류	0.5076	0.4924
건축 미학 삶의 질	어디서 살 것인가 우리가 살고 싶은 곳의 기준을 바꾸다	유현준 지음	기술과학	0.4331	0.5669
K-팝 글로벌 전략	(보는 순간 팔로우하고 싶게 만드는) 인스타그램 브랜딩 레시피	김정은 지음	사회과학	0.5652	0.4348
영화 가상 현실 서사	프로젝트 헤일메리	앤디 위어 지음 ; 강동혁 옮김	문학	0.4909	0.5091
공연 예술 디지털 전환	(맛있는 디자인) 애프터 이펙트 CS6&CC	이수정 지음	총류	0.5864	0.4136
생성형 AI 기계 번역	AI 다음 물결 - 시뮬레이션을 넘어 현실로, 피지컬 AI 기반 자율주행·로봇의 미래	류원하오(지은이), 홍민경(옮긴이), 박종성(감수)	총류	0.3689	0.6311
인지 언어학 은유 분석	왜요, 그 말이 어때서요?: 나도 모르게 쓰는 차별의 언어	김청연 지음 ; 김예지 일러스트	언어	0.4901	0.5099
다문화 한국어 교육	아빠, 받아쓰기가 왜 어렵지?	노정임 글 ; 조승연 그림	언어	0.524	0.476
텍스트 마이닝 고전 분석	기울어진 문해력: 끊어진 대화의 시대, 텍스트와 세상을 새롭게 읽는 법	조병영 지음	총류	0.4269	0.5731
SNS 언어 예절	말하기 전에 생각했나요?: 당당하게 말하지만 상처 주지 않는 대화법	권민창 지음	철학	0.4924	0.5076
한국 근대 소설 지식인	거시기 머시기: 이어령의 말의 힘 글의 힘 책의 힘	이어령 지음	언어	0.4403	0.5597
SF 문학 미래 기술	컴퓨터 과학이 여는 세계	이광근 지음	총류	0.4332	0.5668
포스트모더니즘 서술 기법	(명화들이 말해주는) 그림 속 드레스 이야기	이정아 저	예술	0.5264	0.4736
아동 문학 도덕성 발달	내가 도와줄게: 다른 사람을 존중하고 배려하는 법	테드 오닐 ; 노은정 옮김	철학	0.4197	0.5803

쿼리	도서명	저자	KDC 대분류	유사도거리	유사도점수
디스토피아 사회 비판	고통 구경하는 사회: 우리는 왜 불행과 재난에서 눈을 떼지 못하는가	김인정 지음	사회과학	0.4902	0.5098
실크로드 문명 교류	문명으로 읽는 종교 이야기: 기독교, 유대교, 이슬람교, 불교, 힌두교 탄생의 역사	홍익희 지음	종교	0.5417	0.4583
일제 강점기 독립 운동	(알면 생생한) 한국 전쟁사	햇살과 나무꾼 글 ; 김유 그림	총류	0.4816	0.5184
인문 지리 도시 불평등	지도와 우리 고장: 동서남북 여기가 어디일까?	글: 김민경 ; 그림: 플러그	사회과학	0.5523	0.4477
르네상스 휴머니즘 의의	피렌체 서점 이야기: '세계 서적상의 왕' 베스파시아노, 그리고 르네상스를 만든 책과 작가들	로스 킹 지음 ; 최파일 옮김	총류	0.4851	0.5149
고고학 가야사 재조명	신라의 막판 뒤집기	글: 김해등 ; 그림: 신동민	역사	0.4521	0.5479