Journal of Korean Library and Information Science Society
Vol. 56, No. 4

임베딩 모델 기반의 MARC 레코드 중복검증*

Embedding Model-Based Approach to Duplicate Verification in MARC Records

이 순 영(Soon-Young Lee)**
송 민 건(Min-Geon Song)***
이 수 상(Soo-Sang Lee)****

〈 목 차 〉

요 약 : 본 연구는 AI 기술을 적용해 MARC 레코드의 중복검증 알고리즘 성능 향상을 도모하였다. 기존의 규칙 기반 알고리즘의 한계를 극복하기 위해 텍스트의 의미적 유사성에 기반하는 AI 임베딩 모델을 활용하여 MARC 레코드를 벡터화하고, 유사도 검색을 통해 의미적 유사도를 분석하여 중복레코드를 탐지하였다. 구체적인 연구 방법으로는 첫 번째, 임베딩 모델에 기반한 벡터 유사도 검색으로 MARC 레코드의 중복을 탐지하는 알고리즘을 구현해 선행연구와 동일한 데이터로 성능 평가를 수행하였고, 두 번째, 앞선 실험의 평가 결과를 반영해 임베딩 방식의 장점을 극대화할 수 있는, 즉 문자열 표기 차이로 인한 중복레코드를 식별하는 알고리즘을 구현, 이를 위해 새롭게 구축한 실험데이터와 평가 지표로 알고리즘의 성능을 평가하였다. 실험데이터는 실제 도서관 현장에서 나타날 수 있는 표기 방식의 차이를 반영하여 8가지 변형 규칙을 적용해 구성하였다. 첫 번째 실험 결과, 동일 집단을 중복으로 판정하는 비율이 선행연구보다 개선되었으나, 권호 정보가 다른 다권본 자료를 유사하다고 판정하는 등 숫자나 특수기호의 정확한 매칭을 요구하는 영역에서는 임베딩 방식의 한계를 보였다. 임베딩 방식의 장점을 검증하기 위한 두 번째 실험 결과, 전체 실험데이터에 대해 복본 레코드와 변형 규칙을 100% 식별하는 것으로 나타났다.

주제어 : 인공지능, 임베딩 모델, 벡터 유사도 검색, MARC 레코드, 중복검증

ABSTRACT : This study aimed to improve the performance of duplicate verification algorithms for MARC records by applying AI technology. To overcome the limitations of existing rule-based algorithms, we utilized AI embedding models based on semantic similarity of text to vectorize MARC records and verify duplicate records through similarity search and semantic similarity analysis. The specific research methodology consisted of two phases. First, we implemented a duplicate verification algorithm for MARC records based on vector similarity search using embedding models and evaluated its performance using the same dataset as the prior study. Second, reflecting on the evaluation results of the initial experiment, we implemented an algorithm that maximizes the advantages of the embedding approach—specifically, identifying duplicate records caused by variations in string notation. For this purpose, we evaluated the algorithm's performance using newly constructed experimental data and evaluation metrics. The experimental dataset was designed to reflect notational variations that may occur in actual library settings, applying eight transformation rules. The results of the first experiment showed that the rate of correctly identifying identical groups as duplicates improved compared to the prior study. However, the embedding approach revealed limitations in areas requiring precise matching of numbers and special characters, such as incorrectly judging multi-volume materials with different volume information as similar. The results of the second experiment, designed to validate the advantages of the embedding approach, demonstrated 100% identification of both duplicate records and transformation rules across the entire experimental dataset.

KEYWORDS : AI, Embedding Models, Vector Similarity Search, MARC Records, Duplicate Verification

* 이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.
** 부산대학교 문헌정보학과 강사(libry@pusan.ac.kr / ISNI 0000 0004 7598 0218) (제1저자)
*** 부산대학교 문헌정보학과 박사수료(mgs207@pusan.ac.kr / ISNI 0000 0005 1420 3658) (공동저자)
**** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr / ISNI 0000 0000 6434 9851) (교신저자)

• 논문접수: 2025년 11월 19일 • 최초심사: 2025년 12월 7일 • 게재확정: 2025년 12월 18일
• 한국도서관·정보학회지, 56(4), 1-20, 2025. http://dx.doi.org/10.16981/kliss.56.4.202512.1

Ⅰ. 서 론

현재 도서관이 직면한 주요 과제 중 하나는 인공지능(Artificial Intelligence, 이하 AI) 기술의 급속한 발전과 이로 인한 AI 서비스의 대중화일 것이다. 2022년 OpenAI의 생성형 AI 서비스인 ChatGPT가 출시된 이후, AI의 역할은 '분석'에서 '생성'으로, 고도의 '전문 분야'에서 자연어로 무엇이든 만들어내는 '일상의 서비스'로 자리 잡았다. 실제로 ChatGPT는 2022년 11월 공개된 직후 약 2개월 만인 2023년 1월에 월간 활성 이용자(monthly active users)가 약 1억 명에 도달(Reuters, 2023)했고, 2025년 10월 기준 월간 활성 이용자는 전 세계 성인의 약 10% 수준에 도달(Business Inside, 2025)해 역사상 가장 빠른 속도로 성장하는 애플리케이션라는 평가를 받고 있다. 2023년 4월, 월스트리트저널의 기사 제목처럼 "ChatGPT로 촉발된 AI 경쟁에 아마존, 마이크로소프트, 구글(The Wall Street Journal, 2023)" 등의 글로벌 IT 기업들이 합류하고, 사업 분야도 제조업, 금융업, 교육업 등으로 다각화되면서 생성형 AI 서비스는 각 기업의 핵심 사업으로 추진되고 있다.

도서관 또한 업무와 이용자 서비스 전반에 AI 기술을 도입하려는 다양한 시도가 이루어지고 있다. Clarivate, EBSCO, Elsevier, DBpia와 같은 국내외 주요 학술 DB 벤더들은 자사 제품에 AI 기술을 통합해 차세대 도서관 서비스 플랫폼(Library Service Platform, LSP)과 연구자 지원 도구를 개발해 출시하였으며, 우리나라를 비롯한 핀란드, 독일, 중국, 미국, 두바이 등 각국 도서관에서도 AI 기술 기반의 주제색인, 자동 분류, 열람 관리, AI 검색 및 챗봇 서비스를 제공하고 있다(이수상 외, 2025, 72-89).

특히 클라우드 기반의 도서관 서비스 플랫폼인 ExLibris의 Alma는 목록 작업에 드는 시간과 노력을 절약할 수 있도록 목록 작성자를 지원하는 AI Metadata Assistant를 출시했다. AI Metadata Assistant를 활용해 제목, 저자, 내용 등의 서지 정보에 기반하여 목록 레코드를 생성하거나 기존 레코드를 보강할 수 있으며, AI가 제안한 메타데이터를 목록 작성자가 수락, 수정 또는 폐기할 수 있다(CENL News, 2022). 또한 핀란드국립도서관은 자동 서지 기술 도구인 아니프(Annif)를 기반으로 자동화된 주제색인 서비스인 Finto AI를 출시하였는데, Finto AI는 범용 핀란드어 온톨로지 YSO를 기반으로 현재 핀란드어, 스웨덴어, 영어로 주제색인 서비스를 제공하고 있다(National Library of Finland, 2023). 이처럼 AI가 메타데이터 자동 생성과 자동색인을 포함한 도서관의 다양한 업무 영역에 도입되고 있으나, MARC 레코드 중복검증에 있어서는 여전히 규칙 기반 방식을 따르고 있다.

MARC 레코드 중복검증에의 AI 기술 도입의 필요성에 관해 서술하기에 앞서, MARC 레코드 중복검증은 다양한 기관 간 종합목록을 구축하는 과정에서 발생한 중복레코드 식별 문제에서 시작되었다고 볼 수 있다. 특히 국내 공공도서관의 경우 서지데이터의 반입과 함께 복본의 수만큼 서지데이터를 구축하는 '1책 1레코드' 구조이기 때문에 한 도서관 내에서도 동일한 자료에 대해

다양한 형태의 중복레코드가 존재하는 상황이다. 서지데이터 품질 통제와 처리의 효율성 증대, 그리고 이에 기반한 검색서비스의 신뢰성 확보를 위해서는 서지데이터 중복검증을 통한 대표서지 혹은 마스터 레코드 축적 후 개별 소장 정보를 추가하는 서지공유형으로 종합목록을 운영할 필요가 있다(노지현, 이은주, 2023). 또한 RDA, BIBFRAME 등 차세대 서지 구조로 전환에 대응하기 위해서는 개별 도서관부터 개별자료 기반의 목록 레코드를 구현형 기반의 목록 레코드로 통합하는 작업이 시급하다(송민건, 이수상, 2024).

현재 국가자료종합목록(이하 KOLIS-NET)과 한국교육학술정보원(이하 KERIS) 종합목록의 중복검증 알고리즘은 서지 요소의 완전일치 혹은 일치 정도에 점수를 부여하는 규칙 기반 방식이다. 먼저 KOLIS-NET은 MARC 레코드의 기관 제어번호와 ISBN, 표제/발행자/발행년 순으로 단계적 비교하되 각 단계에서 서지 요소 및 권차를 검증하는 방식이고, KERIS 종합목록은 고유 식별자를 통해 중복 가능성이 있는 후보 레코드 추출한 다음, 신규 레코드와 후보 레코드에서 9가지 비교 요소를 추출하여 직접 비교 후, 각 요소별 점수를 산정하여 중복 여부를 판정하는 방식이다(송민건, 이수상, 2024). 이러한 규칙 기반 방식은 도서관별・목록 작성자별 표기 방식의 차이, 레코드 입력 오류, 다국어 처리 문제 등으로 인해 동일한 자원임에도 중복으로 처리되지 않는다는 명확한 한계가 있어, 연구진의 선행연구(송민건, 이수상, 2025)에서 제안한 것처럼 AI 기술을 적용해 중복검증 알고리즘의 성능을 개선하는 연구가 요구된다.

이에 본 연구에서는 기존의 규칙 기반 중복검증 알고리즘의 한계를 극복하기 위해 텍스트의 의미적 유사성에 기반한 새로운 중복검증 방식을 제안하고자 한다. 이를 위해 AI 임베딩 모델을 활용하여 MARC 레코드를 벡터화하고, 벡터 유사도 검색을 통해 MARC 레코드의 의미적 유사도를 분석함으로써 중복레코드를 탐지하였다. 먼저 AI 임베딩 기반 유사도 방식으로 MARC 레코드 중복검증 알고리즘을 구현해 기존의 규칙 기반 방식과 동일한 데이터를 기반으로 성능 평가를 수행하였고, 두 번째, 앞선 실험의 평가 결과를 반영해 임베딩 방식의 장점을 부각할 수 있는, 즉 문자열 표기 차이로 인한 중복레코드를 식별하기 위해 실험데이터를 구축하고 알고리즘을 구현, 평가 지표를 마련해 성능을 평가하였다.

Ⅱ. 이론적 배경

1. 임베딩 모델

임베딩(embedding) 모델은 문장, 단락, 트윗과 같은 원시 데이터를 의미(semantic meaning)를 담고 있는 고정 길이 벡터로 변환한다. 이러한 벡터를 통해 기계는 정확한 단어가 아닌 의미를

기반으로 텍스트를 비교하고 검색할 수 있는데, 예를 들면 '머신러닝'이라는 단어만 매칭하는 것이 아니라 '기계학습', 'ML'과 같이 서로 다른 표현을 사용하더라도 유사한 개념을 가진 텍스트들이 벡터 공간에서 서로 가깝게 배치된다(LangChain Docs, n.d.). 벡터는 다차원 공간의 정보를 나타내는 숫자 값으로 기계학습 모델이 희소하게 분포된 항목 간의 유사점을 찾는 데 도움을 준다. 다시 말해, 실제 정보를 ML 모델이 해석할 수 있는 연속된 값인 벡터로 변환하도록 훈련된 알고리즘을 임베딩 모델이라고 하는 것이다(AWS, 발행년불명). 현재 분류, 검색, 클러스터링, 추천, 이상 탐지 등 다양한 영역에서 임베딩 모델이 사용되고 있으며, OpenAI의 text-embedding-3, Microsoft의 E5, all-MiniLM-L6, Alibaba Group의 gte-large 등이 대표적인 임베딩 모델로 거론되고 있다(현유경, 2025).

2. 벡터 유사도 검색

상기에 언급한 것처럼, 임베딩 모델을 통해 텍스트를 벡터로 변환한 이후에는 벡터로 변환된 항목 간 유사점을 찾아내는 것이 다음 단계이다. 이를 벡터 검색(vector search) 혹은 벡터 유사도 검색(vector similarity search)이라고 하는데, IBM에서는 기존 검색과 비교하며 벡터 검색의 이점을 다음과 같이 설명하고 있다(IBM, n.d.).

> 최고의 피자 레스토랑에 대한 정보를 찾기 위해, 전통적인 키워드 검색 엔진은 "최고", "피자", "레스토랑"과 정확히 일치하는 단어를 포함하는 페이지를 검색해 "최고의 피자 레스토랑" 또는 검색자의 위치 "근처 피자 레스토랑"과 같은 결과만 반환한다. 전통적인 키워드 검색은 검색의 맥락이나 의도보다는 키워드의 '매칭'에 중점을 두기 때문이다.
>
> 반면, 시맨틱 벡터 검색(semantic vector search)은 질의의 의미와 맥락을 이해하기 때문에, 콘텐츠에서 "최고의 피자 레스토랑"이라는 정확한 단어를 사용하지 않더라도 평점이나 추천도가 가장 높은 피자 가게를 검색해낼 수 있다. 따라서 다양한 지역의 고품질 피자 가게를 소개하는 기사나 가이드를 포함하는 '맥락적으로 보다 적합한 결과'를 제시할 수 있다.

벡터 유사도 검색은 대규모 데이터 세트에서 유사한 항목을 효율적으로 검색할 수 있다. 특히 차원 축소를 통해 기존의 검색 방식이 처리하기 어려웠던 고차원 데이터를 처리할 수 있고, 이로 인해 텍스트 데이터뿐만 아니라 다차원적 특성에 기반하는 이미지, 동영상 같은 비정형, 반정형 데이터를 처리할 수 있다(ENCORD, 2023).

벡터 유사도 검색의 과정은 다음과 같다. 첫 번째, 적절한 임베딩 모델을 통해 처리 대상 항목을 벡터로 표현한다. 두 번째, 유클리드 거리, 코사인 유사도 등 유사도 공식을 적용해 벡터 간 유사도

를 정량화한다. 세 번째, 효율적인 유사도 검색을 위해 최근접 이웃(Nearest Neighbor, 이하 NN) 알고리즘을 적용해 처리 대상 항목과 가장 가까운 벡터를 검색한다. 대표적인 NN 알고리즘은 kNN(k-Nearest Neighbors) 알고리즘, HNSW(Hierarchical Navigable Small World) 알고리즘, Faiss(Facebook's similarity search) 알고리즘 등이 있다.

특히 본 연구에서 활용한 Faiss는 Meta가 개발한 페이스북 유사도 검색 라이브러리로, 역파일 구조, 벡터 양자화, IVFADC(Inverted File with Approximate Distance Calculation) 등 여러 인덱싱 기법을 조합해 검색의 속도와 검색 결과의 정확성에 균형을 맞춘 알고리즘으로 평가받고 있다(ENCORD, 2023; Meta, 2017). 다시 말해 정확한 매칭에 기반을 둔 역파일 기법과 벡터를 그룹화해서 빠르게 비교하는 벡터 양자화 기법, 그리고 벡터를 그룹으로 나눈 후에 그룹 안에서 거리를 계산하는 IVFADC 기법을 활용하여 빠르게 검색하면서도 정확도 손실은 최소화했다는 것이다.

3. 선행연구

본 연구는 기존의 중복검증 알고리즘의 한계점을 극복하기 위해 MARC 레코드를 임베딩 모델로 벡터화하고 벡터 유사도 검색을 통해 MARC 레코드의 동일성을 검증하고자 한다. MARC 레코드 중복검증에 관한 연구는 연구진의 선행연구(송민건, 이수상, 2024; 2025)에서 분석하여 제시하였으므로, 여기에서는 유사도 기반으로 저작 식별을 시도한 문헌정보학 분야의 연구와 다양한 임베딩 모델을 제안 및 적용하고 그 성능을 확인하기 위한 목적으로 서지데이터 임베딩을 수행한 연구물들을 분석하였다.

먼저 서지레코드를 저작 단위로 군집화하기 위한 기초연구로써, 동일한 저작을 식별하기 위해 저자명 유사도와 표제 유사도를 계산해 각각의 유사도가 일정 수준 이상인 경우 동일한 저작으로 그룹화하는 방법을 채택한 윤재혁 외(2020)의 연구가 있다. FRBR의 저작 단위로 서지레코드를 군집화하는 과정에서 저자 데이터를 명확하게 식별할 수 있도록 책임표시사항의 역할어의 목록을 생성하고 표준화하는 과정을 적용하였다. 그 결과, 저작을 창작하는데 기여한 사람이 다수인 레코드를 대상으로 선행연구에 비해 높은 정확도로 저작을 군집화하였으나, 표제의 불일치로 인하여 저작 단위의 군집 형성이 제대로 이루어지지 못한 문제점이 확인되었다. 연구는 데이터 오류가 있는 상황에서 최대한의 정확도를 끌어내기 위한 방법을 모색하려는 시도로, 가장 우선적으로 서지레코드의 품질을 최대한 일관되게 고수준으로 유지하는 것이 필요하다고 제안하였다.

나상오 외(2025)는 한국문학분야의 KORMARC 서지레코드를 활용하여 개선된 저작 식별 방안을 제안하였다. 동일 저작이라도 표기 방식 등의 차이로 식별되지 못하는 문제를 해결하기 위한 데이터 전처리 과정을 수행하고 네트워크 분석 기법을 활용하여 동일 표제 또는 저자명을 가진 레코드를 연결하여 저작 식별을 시도하였다. 하지만 해당 연구에서 역시 텍스트 완전일치 기준을 적용함에

따라 필드 입력 방식의 차이로 동일 저작이 분리되는 등의 한계가 존재하였다.

상기 두 편의 선행연구에서는 공통적으로 서명과 저자명만을 활용하여 저작 식별을 시도하였고, 그 결과 필드 입력 방식의 차이나 입력 오류로 동일 저작이 군집되지 못하는 사례가 빈번하게 발견되었다. 기존의 중복검증 알고리즘의 문제점과 마찬가지로 다양한 입력 방식이 혼재하는 작금의 상황에서는 서명과 저자명만을 활용하거나 완전일치 방식을 적용하기에는 한계가 있다는 것이다.

한편, 서지데이터의 요소를 완전일치로 비교하는 방식이 아닌 다양한 임베딩 모델을 적용하여 유사도를 비교한 연구도 수행되었는데, Mikolov et al.(2013)은 대규모 텍스트 데이터에서 단어의 유사도를 파악하고 임베딩을 적용하는 Word2Vec 모형을 제안하였으며 이를 활용하거나 응용한 선행연구가 다음과 같이 진행되었다.

Barkan과 Koenigstein(2016)은 자연어 처리 분야의 Skip-gram with Negative Sampling(SGNS, word2vec) 기법을 협업 필터링에 적용한 Item2Vec 모델을 제안했다. 많은 추천 알고리즘이 이용자와 아이템 사이의 관계를 임베딩하는 것에 반해 아이템 사이의 직접적인 이용 관계를 임베딩하여 이용자 정보 없이도 아이템 간 관계를 추론할 수 있는 것을 확인, SVD 기반 유사도 모델과 비교했을 때 장르 일관성 측면에서 우수한 성능을 나타냈다.

이정훈과 정윤경(2019)은 Word2Vec 임베딩 모델을 응용하여, 이용자가 읽은 책의 목록을 기반으로 각 책마다 벡터를 부여하여 유사한 책을 추천하는 시스템을 설계하였다. 연구에서는 서지레코드의 텍스트를 벡터 임베딩하여 비교하지 않고 이용자의 독서 이력만을 통해 유사도를 분석하였다.

이처럼 Word2Vec 기반의 도서 추천 모델에 관한 연구는 최근까지도 이어지고 있는데, 콘도 코이치 외(2025)는 효과적인 도서 추천 시스템 개발을 목표로 Word2Vec 기반의 SCDV(Sparse Composite Document Vectors)를 통해 도서 정보를 벡터화하고 아이템 기반 협업 필터링을 적용한 추천 모델을 제안하였다. Book-Crossing Community 데이터셋과 Google Book API를 활용하여 202,468권의 도서를 대상으로 실험한 결과, 정확률과 재현율 간 상충관계를 확인하였으며, 이를 통해 Word2Vec과 SCDV 기반 협업 필터링 기법이 도서 추천 분야에서 효과적임을 실증적으로 입증하였다.

임베딩 모델을 주제 분류에 적용한 연구도 수행되는데, 먼저 강우진 외(2023)는 공공도서관 대출데이터에 Item2Vec 기법을 적용하여 분류기호를 벡터로 변환하고 총 522개의 분류기호로 네트워크를 생성, 공공도서관 이용자가 동시에 대출한 도서들의 주제 분야 간의 연관성을 파악하고자 하였다. 연구 결과, 15개 커뮤니티가 네트워크에서 식별되었고, 두 개 이상의 커뮤니티에서는 요목 수준에서 주제적 연관성이 나타나, Item2Vec 기법이 함께 대출될 가능성이 높은 자료의 주제를 파악하는 데 도움이 됨을 확인하였다.

또한 이용구(2023)는 단행본 서명에 다양한 단어 임베딩 기법을 적용하여 단어 벡터를 추출하고, 이를 분류 자질로 활용하여 단어 임베딩 기법에 따른 자동분류의 성능을 분석하였다. Word2vec,

fastText, Skip-gram와 같은 임베딩 모델을 활용하였으며, 실험 결과 세 모델 모두 TF-IDF 자질보다 자동분류 성능에서 우수한 성능을 보였다. 대체로 우수한 성능을 보인 모델은 fastText 모델이었으며, Word2Vec 모델은 낮은 차원 또는 작은 크기의 샘플에서 우수한 성능을 보였다. 해당 연구에서는 도서 자료에 단어 임베딩을 적용한 사례가 없어 가장 기본적이고 단순한 서명 데이터에 대해 단어 임베딩을 적용하였으며, 다양한 메타데이터 요소를 활용한다면 더 나은 성능을 기대할 수 있을 것이라 하였다.

Word2Vec과 Item2Vec 외에도 서지데이터 임베딩을 위한 다양한 모델들이 연구되었다. Ganesh et al.(2016)은 공동 저자 네트워크에서 저자 간 임베딩 유사도 모델인 Author2Vec 모델을 제안하였고, Yoneda et al.(2017)은 서지데이터의 여러 요소 간의 관계를 나타내는 새로운 임베딩 모델인 Bib2Vec 모델을 제안하였다. Bib2Vec 모델은 제목과 초록 등의 텍스트 요소와 저자, 참고문헌과 같은 비텍스트 요소로 구분하여 임베딩을 적용하였다. Anvari와 Amirkhani(2018)는 이용자의 독서 이력만을 활용해 도서를 임베딩하는 Book2Vec 모델을 제안하였다. Goodreads 소셜 네트워크에서 도서 데이터를 수집해 분석하였으며, 시각화를 통한 분석 결과 동일 저자의 작품과 동일 시리즈의 도서가 벡터 공간에 근접하게 배치됨을 확인하였다. 앞서 언급한 이정훈과 정윤경(2019)의 연구처럼 이용자 독서 이력만으로 도서 간 의미 있는 관계를 포착할 수 있음을 확인하였다.

Zhang은 3편의 논문(Zhang et al., 2017; Zhang, Wang, Zhao, et al., 2018; Zhang, Wang, Xu, et al., 2018)을 통해 다층 클러스터링과 지역적 구조에 대한 기본 개념을 만든 후, 이를 확장해 도서 간 지역성과 계층성을 벡터에 반영하는 모델을 개발, 최종적으로 도서 분류 체계와 같은 계층 구조 데이터를 벡터화하는 Tree2Vector 모델을 제안하였다. Tree2Vector는 실제 저자 추천과 이미지 검색에서 우수한 성능이 나타남을 입증하였다.

Zhao et al.(2020)은 기존 Tree2Vector 모델을 확장하여 실시간 추천 시스템에 적합하도록 설계된 DTree2Vec 모델을 제안하였다. 이 모델은 온라인 연재 소설과 같이 완결되지 않은, 즉 연재 중인 도서의 동적 계층 트리 구조를 구축하고 코사인 기반 지역 재구성 모델(Cosine-type Local Reconstruction Model)을 통해 통합된 의미적 특징(semantic features)으로 표현한다. 연재 중인 도서의 업데이트 상태를 추적하고 후속 콘텐츠를 실시간으로 추가해 추천 품질을 보장하는데, 실험 결과 미완결 연제 도서에 대해 더 높은 추천 정확도를 달성하여 도서의 실시간 추천 문제를 효과적으로 해결할 수 있음을 보여주었다.

이상과 같이 서지데이터에 임베딩 모델을 적용한 다양한 연구가 수행되었으나 대부분 임베딩 모델을 적용해 도서, 주제, 저자를 추천하는 시스템을 개선하거나 주제 분류, 이미지 검색을 위한 목적으로 연구가 이루어졌으며, 서지데이터의 중복검증에 임베딩 모델을 적용한 연구는 찾아볼 수 없었다. 또한 선행연구 분석 결과, 기존의 규칙 기반의 동일성 검증은 서명, 저자 등 제한된 서지요소를 문자열 매칭 방식으로 비교하였기 때문에 표기 방식의 다양성이나 데이터 입력 오류

에 취약하고 성능이 저하되었다. 하지만 임베딩 모델에 기반한 연구들은 제목, 저자, 장르 등 핵심 서지 요소만을 활용해 도서의 의미적 표현을 학습하고, 실험 결과 정확도와 같은 각종 지표에서 우수한 성능이 나타나는 것을 알 수 있다.

Ⅲ. 연구 설계

서론에서 언급한 바와 같이 본 연구는 연구진의 선행연구(송민건, 이수상, 2025)의 후속 연구로써 두 가지 실험을 설계하였다. 먼저 실험 1에서는 규칙 기반의 중복검증 알고리즘과 임베딩 모델 기반 알고리즘의 성능을 비교 평가하고자 하였다. 임베딩 모델 기반 알고리즘의 성능을 검증하기 위해 연구진의 선행연구에서 사용했던 실험데이터를 동일하게 활용하였다. 실험 1의 연구 결과에 근거하여, 실험 2는 임베딩 기반 유사도 검색의 핵심적인 장점, 즉 “맥락적으로 보다 적합한 결과를 제시(IBM, n.d.)”하는지를 중점적으로 검증하는 것으로 설계하였다. 이를 위해, 실제 도서관 현장에서 발생하는 MARC 레코드 표기 방식의 차이에 주목했다. 구체적으로 표제, 저자, 출판사 등의 문자열 표기 차이로 인해 발생하는 중복 레코드 사례를 식별하고, 해당 사례에 기반한 8가지 변형 규칙을 적용하여 실험데이터를 구축한 후 알고리즘을 설계 및 구현하였다.

1. 실험데이터

앞서 언급한 것처럼, 규칙 기반 알고리즘과 중복검증 성능을 비교 평가하기 위해 실험 1은 연구진의 선행연구에서 구축한 실험데이터를 그대로 활용하였다. 이 실험데이터는 부산 지역 M 도서관의 99,957건의 MARC 데이터 중 일부 유사한 데이터가 존재하는 레코드 쌍을 중복 후보 레코드 쌍으로 추출한 것으로, 090 필드(자관 청구기호)가 일치하는 레코드 쌍 중 049 필드(소장사항)의 $v(권・연차기호)와 $f(별치기호)가 일치하는 동일 집단 2,521쌍, 불일치 집단 3,047쌍이다(송민건, 이수상, 2025, 296-297).

실험 2의 데이터는 총 3가지 집단으로 구성하였는데, 첫 번째 집단은 M 도서관이 소장하고 있는 자료 중 복본 수 5건 이상인 레코드 84건으로 복본 집단으로 구축하였고, 두 번째 집단은 도서관별・목록 작성자별 표기 방식의 차이를 반영하여 만든 8가지 변형 규칙을 적용한 데이터로, M 도서관의 MARC 레코드 중 이 변형 규칙을 적용할 수 있는 집단 1,000건을 원형 집단으로, 각 변형 규칙을 적용해 변형한 1,000건의 레코드를 변형 집단으로 구축하였다. 따라서 2,084건의 레코드가 하나의 JSON 파일로 구축되었으며, 8가지 규칙을 각각 적용한 8개 JSON 파일을 만들어 총 16,672건의 데이터가 실험 2의 실험데이터이다. 그리고 각 레코드마다 999 필드 $a에 집단명을

추가해 중복검증의 성능을 평가하는 기준으로 삼았다.

실험 2의 복본 집단을 구성하고 있는 각 레코드의 도서기호와 본표제, 복본 수는 아래 〈표 1〉과 같다.

〈표 1〉 실험 2의 복본 집단 데이터

도서기호	본표제	복본 수
큰글 813.7-63	[큰글자도서] 불편한 편의점	5
아동 813.8-7115	경성 기억 극장	5
유아 740-195	I'm the biggest thing in the ocean	5
가등록 911.058-28	돌아온 외규장각 의궤와 외교관 이야기	8
가등록 499.83-1	비숲	6
가등록 813.7-1	아몬드	6
가등록 814.7-1	오전을 사는 이에게 오후도 미래다	10
가등록 813.7-4	불편한 편의점	5
가등록 814.7-2	단어의 집	6
가등록 813.7-8	페퍼민트	6
가등록 911.89-2	망치질하는 어머니들 깡깡이마을 역사 여행	6
가등록 818-4	(세탁비는 이야기로 받습니다.) 산복빨래방	5
가등록 813.7-9	열다섯에 곰이라니	5
가등록 813.8-5	나를 찾아 줘!	5
계		84

실험 2에서 적용한 8가지 변형 규칙은 아래 〈표 2〉와 같다. 첫 번째 규칙부터 네 번째 규칙은 표제와 관련된 중복 레코드 표기의 다양성을 반영하였는데, 먼저 첫 번째는 본표제(245 $a) 내 괄호로 둘러싸인 관제를 포함하는 MARC 레코드를 원형 집단으로, 괄호 안의 내용을 완전 삭제하여 변형 레코드를 만들고 변형 집단으로 구축한 것이다. 두 번째는 표제관련정보(245 $b)가 존재하는 원형 레코드의 245필드 $b를 245필드 $a의 뒤로 이동하여 변형 레코드로 구축하였다. 세 번째는 대등표제(245 $x)가 존재하는 MARC 레코드의 245필드 $x를 245필드 $a의 뒤로 이동하여 변형 집단을 구축하였다. 네 번째는 권표제(245 $p)가 존재하는 MARC 레코드의 245필드 $p를 245필드 $a의 뒤로 이동하여 변형 집단을 구축하였다. 그리고 다섯 번째 규칙은 권차에 관한 내용으로 필드245 $n이 존재하는 MARC 레코드의 245필드 $n의 데이터를 삭제하여 변형 집단을 구축하였다.

다음은 저자와 관련된 표기의 다양성을 반영하여, 책임표시사항(245 $d $e)에서 반점(,), 세미콜론(;), 공백( )이 포함된 레코드의 반점, 세미콜론, 공백 이후의 텍스트를 모두 삭제하는 규칙과, 특정 역할어가 포함된 레코드의 표기 형식을 변경하는 규칙을 적용하였다.

마지막으로 출판사명은 단순 오탈자 외에도 약칭, 영문명과 한문 이형으로 인한 불일치 등의 사례도 존재하지만, 임프린트명 처리로 인한 불일치의 사례가 다양하고 빈번하게 나타나 이를

반영하여 변형 규칙을 만들었다. 이는 임베딩 모델이 표기상의 유사성을 효과적으로 포착할 수 있음을 검증하기 위한 것으로, 출판사명(임프린트명)의 형태로 입력된 MARC 레코드의 발행처(260 $b) 괄호 안에 두 글자 이상의 텍스트가 포함된 원형 레코드를 괄호 안의 데이터만 남기고 삭제해 변형 레코드로 구축하였다.

이상의 8가지 변형 규칙을 적용한 실험데이터 구축 사례는 아래 〈표 2〉에 정리하였다.

〈표 2〉 실험 2의 8가지 변형 규칙

rule_1	본표제(245 $a) 내에 괄호'()'로 둘러싼 관제가 포함된 레코드의 괄호 안 내용을 모두 삭제	
	원형 레코드	245 {"a": "(서울대 교수와 함께하는)10대를 위한 교양 수업.", "n": "1.", "p": "유성호 교수님이 들려주는 법의학 이야기/", "d": "유성호.", "e": "신병근 그림"}
	변형 레코드	245 {"a": "10대를 위한 교양 수업.", "n": "1.", "p": "유성호 교수님이 들려주는 법의학 이야기/", "d": "유성호.", "e": "신병근 그림"}
rule_2	표제관련정보(245 $b)가 존재하는 레코드의 245 $b를 245 $a의 뒤로 이동	
	원형 레코드	245 {"a": "희망을 찾는가:", "b": "전혀 다른 방식으로 세상을 바꾸는 대안 노벨상 수상자들 이야기/", "d": "게세코 폰 뤼프케;", "e": "김시형 옮김"}
	변형 레코드	245 {"a": "희망을 찾는가: 전혀 다른 방식으로 세상을 바꾸는 대안 노벨상 수상자들 이야기/", "d": "게세코 폰 뤼프케;", "e": "김시형 옮김"}
rule_3	대등표제(245 $x)가 존재하는 레코드의 245 $x를 245 $a의 뒤로 이동	
	원형 레코드	245 {"a": "도넛을 구멍만 남기고 먹는 방법=", "x": "Can you eat a doughnut while keeping the hole?/", "d": "오사카대학 쇼세키카 프로젝트 지음;", "e": "김소연 옮김"}
	변형 레코드	245 {"a": "도넛을 구멍만 남기고 먹는 방법= Can you eat a doughnut while keeping the hole?/", "d": "오사카대학 쇼세키카 프로젝트 지음;", "e": "김소연 옮김"}
rule_4	권표제(245 $p)가 존재하는 레코드의 245 $p를 245 $a의 뒤로 이동	
	원형 레코드	245 {"a": "생각이 크는 인문학.", "n": "15.", "p": "빅데이터/", "d": "정용찬 글;", "e": "이진아 그림"}
	변형 레코드	245 {"a": "생각이 크는 인문학. 빅데이터/", "n": "15.","d": "정용찬 글;", "e": "이진아 그림"}
rule_5	권차(245 $n)가 존재하는 레코드의 245 $n 데이터를 삭제	
	원형 레코드	245 {"a": "(상상초월)포켓몬 과학 연구소.", "n": "2/", "d": "야나기타 리카오 글;", "e": "정인영 옮김"}
	변형 레코드	245 {"a": "(상상초월)포켓몬 과학 연구소.", "d": "야나기타 리카오 글;", "e": "정인영 옮김"}
rule_6	책임표시사항(245 $d $e)에서 반점(,), 세미콜론(;), 공백( )이 포함된 레코드의 반점, 세미콜론, 공백 이후의 텍스트를 모두 삭제	
	원형 레코드	245 {"a": "시가 쑥덕쑥덕:", "b": "부산 교동초등학교 어린이들의 동시 모음집 2/", "d": "부산교동초등학교 어린이들 글씀;", "e": "조주현 [공]엮음"}
	변형 레코드	245 {"a": "시가 쑥덕쑥덕:", "b": "부산 교동초등학교 어린이들의 동시 모음집 2/", "d": "부산교동초등학교", "e": "조주현"}
rule_7	책임표시사항에서 특정 역할어가 포함된 레코드의 표기 형식을 변경	
	원형 레코드	245 {"a": "그림 그리는 할머니 김두엽입니다/", "d": "김두엽 지음"}
	변형 레코드	245 {"a": "그림 그리는 할머니 김두엽입니다/", "d": "지은이: 김두엽"}

rule_8	발행처(260 $b)에서 괄호'()'안에 두 글자 이상의 텍스트가 포함된 레코드의 괄호 안의 데이터만 남기고 삭제	
	원형 레코드	260 {"a": "서울:", "b": "생각나눔(기획실크),", "c": "2022"}
	변형 레코드	260 {"a": "서울:", "b": "기획실크", "c": "2022"}

2. 중복검증 알고리즘

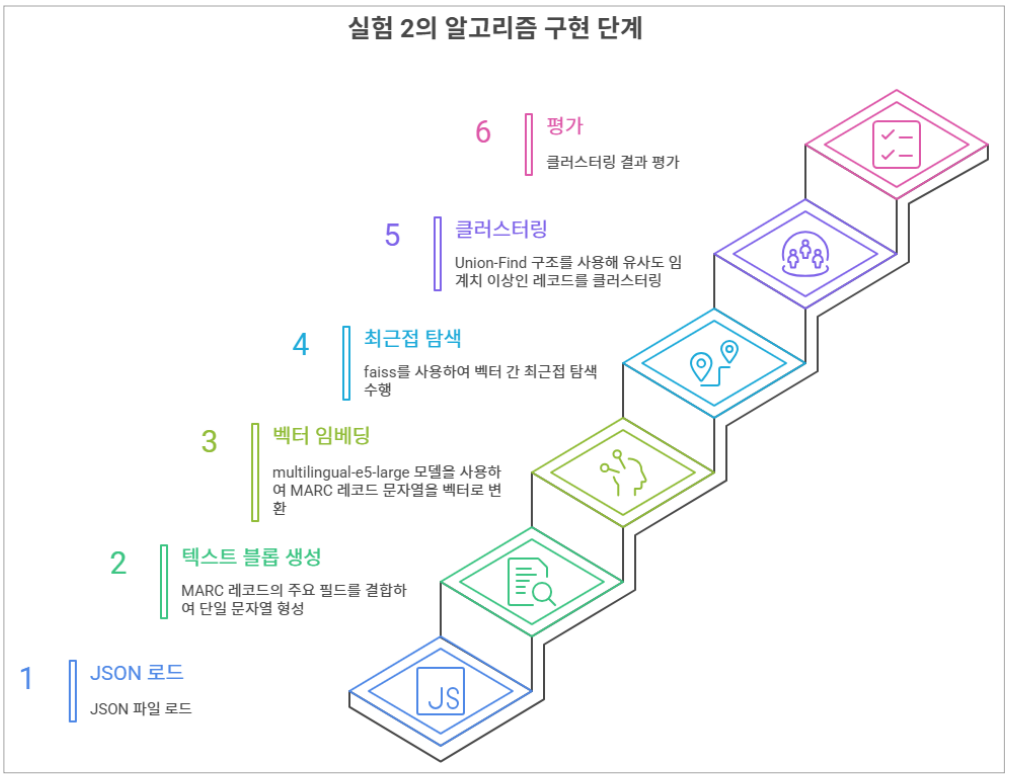
본 연구는 위에서 언급한 두 가지 실험데이터에 따라 알고리즘 테스트를 수행하였다.

먼저 규칙 기반 중복검증 알고리즘과의 비교를 위해 설계한 실험 1은 임베딩 모델을 처음 적용하는 초기 단계로 실험을 진행하는 과정에서 분류 성능 평가를 진행하고 그 결과에 따라 알고리즘을 조금씩 수정하는 방식으로 진행하였다. 이를 위해 첫째, JSON 파일을 불러와 MARC 레코드의 주요 정보를 추출해 하나의 텍스트 블롭(blob)으로 저장하고, 둘째, 임베딩 모델을 이용해 벡터 임베딩을 수행하고, 셋째, 두 레코드 간 코사인 유사도를 계산해 유사도 임계치에 따라 매우 유사, 유사, 가능성 있음, 다름을 비율로 확인하는 단계로 구현하였다.

이 과정에서 MARC 레코드를 텍스트 블롭으로 저장한 것은 필드별 임베딩 대비 연산의 효율성 측면의 장점도 있지만, 다양한 필드와 지시기호 등으로 분산된 서지데이터를 하나의 문자열로 통합해 유사도 비교의 정확도를 높이기 위한 목적이다. 그리고 임베딩 모델의 경우 GPT-3.5에 기반한 text-embedding-ada-002로 시작해 성능 향상을 이유로 GPT-4에 기반한 text-embedding-3-large로 변경하여 벡터 임베딩을 수행하였다.

다음으로 실험 2는 8가지 변형 규칙을 적용한 데이터로 실험한 것으로 첫째, JSON 파일을 불러와 MARC 레코드의 주요 필드에 대해 하위 필드를 결합한 하나의 텍스트 블롭으로 묶고, 둘째, multilingual-e5-large 모델을 활용하여 MARC 레코드 문자열을 벡터로 임베딩한 후, 셋째, Faiss를 적용해 벡터 간 최근접 탐색을 수행하였다. 마지막으로 서로 연결된 요소를 하나의 그룹으로 묶기 위해 Union-Find 구조를 적용해 유사도 임계치 이상인 레코드들을 동일 클러스터로 묶고 실제 복본을 잘 찾아냈는지, 원형과 변형이 같은 클러스터로 묶였는지를 평가하였다. 아래 〈그림 1〉은 실험 2의 알고리즘을 Napkin AI를 활용해 도식화한 것이다.

알고리즘의 구현과 실험은 Google Colab PRO 환경에서 Python 3으로 수행하였고, 실험 1에서는 OpenAI의 text-embedding-ada-002과 text-embedding-3-large(유료 API) 모델을 활용해 모델별 임베딩 성능의 차이를 살펴보았고, 실험 2에서는 한국어를 포함한 다국어 처리에 더욱 적합하고 의미적 검색에 특화된 Microsoft의 multilingual-e5-large(Hugging Face 오픈소스) 모델을 활용하였다.



〈그림 1〉 실험 2의 알고리즘 구현 단계

Ⅳ. 연구 결과

1. 실험 1: 동일 집단과 불일치 집단 간의 비교 실험

실험 1의 데이터는 090 필드가 일치하는 동일 집단 2,521쌍과 불일치 집단 3,047쌍이며, 앞서 언급한 것처럼 실험을 진행하는 과정에서 분류 성능 평가를 진행하고 그 결과에 따라 알고리즘을 조금씩 수정하는 방식으로 실험하였다. 앞선 실험은 표제 관련 정보, 저자 관련 정보, 출판사항, ISBN 등의 주요 필드를 추출해 텍스트 블롭을 생성하고 GPT-3.5 기반의 text-embedding-ada-002 모델로 임베딩하였다. 그 결과 코사인 유사도 0.95 이상인 매우 유사 레코드가 2,482쌍, 98.45%로 나타나 선행연구의 결과인 98.10%(송민건, 이수상, 2025, 297)보다 0.35% 높게 나타났으나, 불일치 집단을 유사하다고 판단한 경우가 매우 높게 나타났다.

이와 같은 실험의 결과에 따라, 불일치 집단만을 대상으로 재실험을 수행하였다. MARC 레코드의

주요 필드가 아닌 전체 필드를 추출하고 GPT-4 기반의 text-embedding-3-large로 임베딩 모델을 변경하였다. text-embedding-ada-002는 1,536차원의 벡터를 생성하는 데 반해, text-embedding-3-large는 최대 3,072차원의 벡터를 생성할 수 있어 보다 많은 정보를 임베딩 벡터에 담을 수 있으며, 차원 크기를 유연하게 조절할 수 있다는 장점이 있다(OpenAI Platform, 2024).

〈표 3〉 실험 1의 유사도 값별 매칭 비율 비교

분류	주요 필드 임베딩/GPT-3.5	전체 필드 임베딩/GPT-4
0.95 이상 Exact (매우 유사)	372건 (12.21%)	114건 (3.74%)
0.85 이상 Probable (유사)	973건 (31.93%)	274건 (8.99%)
0.75 이상 Possible (가능성 있음)	1,636건 (53.69%)	334건 (10.96%)
0.75 미만 Distinct (다름)	66건 (2.17%)	2,325건 (76.30%)

불일치 집단만을 대상으로 재실험한 결과, 〈표 3〉과 같은 결과가 나타났다. 먼저 표제 관련 정보, 저자 관련 정보, 출판사항, ISBN 등의 주요 필드를 text-embedding-ada-002로 임베딩한 결과, 코사인 유사도 0.95 이상의 매우 유사 레코드가 372건 12.21%, 0.85 이상의 유사 레코드가 31.93%로 나타났다. 다음으로 MARC 레코드의 전체 필드를 대상으로 text-embedding-3-large 모델로 임베딩한 결과, 코사인 유사도 0.75 미만으로 다름 판정을 받은 레코드 수가 66건 2.17%에서 2,325건 76.30%로 늘어나 임베딩 성능이 크게 향상된 것으로 나타났다. 하지만 여전히 불일치 집단을 유사하다고 판정한 사례들이 다수 나타나, '매우 유사'하다고 나타난 MARC 레코드의 사례를 살펴본 결과, 아래 〈그림 2〉와 같이 문자열은 일치하지만 권호 정보가 다른 다권본인 경우인 것을 알 수 있었다.

여기에서 임베딩 기반 유사도 검색의 한계가 드러났는데, 바로 의미적 유사성에 기반해 유사도를 판단하기 때문에 제목과 저자가 같지만 권호 정보가 다른 경우 중복으로 판정한다는 점이다. 다시 말해 임베딩 기반 유사도만을 적용한 경우에는 숫자나 특수 기호 처리에는 한계가 있으며, 필요에 따라서는 ISBN을 포함한 숫자형 데이터는 직접 비교하고 문자열 데이터에 대해서는 임베딩 유사도 검색을 병행하는 방식이 필요하다는 것이다. 하지만 서론에서 언급한 것처럼 향후 차세대 서지 구조로의 전환을 위해서는 개별자료 기반의 목록 레코드를 구현형 기반의 목록 레코드로 통합할 필요가 있기 때문에 임베딩 유사도 검색의 장점을 극대화하는 방안도 고려할 필요가 있다고 판단하였다.

245_a	245_b	245_n	245_p	245_d	260_a	260_b	260_c	300_a	300_c	500_a	653_a	700_a	950_b	700
암살자:	이원호 장편소설.	1,	아프간/	이원호 지음	서울:	한결미디어	2021	336p.;	23cm	표제관련정보:아프	장편소설	이원호	₩15000	
암살자:	이원호 장편소설.	2,	서울/	이원호 지음	서울:	한결미디어	2021	337p;	23cm	표제관련정보:아프	장편소설	이원호	₩15000	
(설민석의) 만만 한	재미 만점 ★ 효과 만점 ★	2 ,	남북국 시대부터	설민석 ,	서울:	아이세움,	2020	195 p.;	28 cm	감수: 단꿈 연구소	역사학습만화		₩12000	[{'a': '신지은
(설민석의) 만만 한	재미 만점 ★ 효과 만점 ★	4 ,	조선 후기/	설민석 ,	서울:	아이세움,	2021	195 p.;	28 cm	감수: 단꿈 연구소	역사학습만화		₩12000	[{'a': '신지은
(설민석의) 만만 한	재미 만점 ★ 효과 만점 ★	3 ,	조선 전기/	설민석 ,	서울:	아이세움,	2021	195 p.;	28 cm	감수: 단꿈 연구소	아동한국사		₩12000	[{'a': '신지은
(설민석의) 만만 한	재미 만점 ★ 효과 만점 ★	4 ,	조선 후기/	설민석 ,	서울:	아이세움,	2021	195 p.;	28 cm	감수: 단꿈 연구소	역사학습만화		₩12000	[{'a': '신지은
(이상한 과자 가게)전천당.		8/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	189p.:	20cm		아동		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게)전천당.		9/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	165p.:	20cm		아동		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게)전천당.		8/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	189p.:	20cm		아동		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게) 전천당.		7/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	181 p.:	19 cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '히로시
(이상한 과자 가게)전천당.		9/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	165p:	20cm		아동		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게)전천당.		1/		히로시마 레이코 글;	서울:	길벗스쿨,	2019	144p:	19cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게)전천당.		9/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	165p.:	20cm		아동		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게) 전천당.		7/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	181 p.:	19 cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '히로시
(이상한 과자 가게)전천당.		2/		히로시마 레이코 글;	서울:	길벗스쿨,	2019	152p:	19cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '김정은
(이상한 과자 가게)전천당.		1/		히로시마 레이코 글;	서울:	길벗스쿨,	2019	144p:	19cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게)전천당.		2/		히로시마 레이코 글;	서울:	길벗스쿨,	2019	152p:	19cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '김정은
(이상한 과자 가게) 전천당.		7/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	181 p.:	19 cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '히로시
(이상한 과자 가게)전천당.		1/		히로시마 레이코 글;	서울:	길벗스쿨,	2019	144p:	19cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '쟈쟈'
(이상한 과자 가게) 전천당.		7/		히로시마 레이코 글;	서울:	길벗스쿨,	2020	181 p.:	19 cm	원저자명: 廣嶋玲	판타지동화		₩12000	[{'a': '히로시
(교과서보다 먼저 읽는)첫 세계사.		1/		한정영 글;	서울:	북멘토,	2020	181p.:	23cm	감수: 김민수, 김동	청소년		₩15000	[{'a': '한정영
(교과서보다 먼저 읽는)첫 세계사.		2/		한정영 글;	서울:	북멘토,	2020	213p.:	23cm	감수: 김민수, 김동	청소년		₩15000	[{'a': '한정영
슬로하이즈의 신:	즈지무라 미즈키 장편소설	2/		즈지무라 미즈키 지음;	서울:	용실북스,	2020	431p.;	20cm	원저자명: 辻村深	장편소설		₩15000	[{'a': '즈지무
슬로하이즈의 신:	즈지무라 미즈키 장편소설	1/		즈지무라 미즈키 지음;	서울:	용실북스,	2020	319p.;	20cm	원저자명: 辻村深	장편소설		₩15000	[{'a': '즈지무
미스터션샤인=	드라마 원작소설.	2/		김은숙 극본;	서울:	알에이치코	2020	401p.;	19cm		원작소설		₩14800	[{'a': '김은숙
미스터션샤인=	드라마 원작소설.	1/		김은숙 극본;	서울:	알에이치코	2018	409p.;	19cm		원작소설		₩14800	[{'a': '김은숙

〈그림 2〉 '매우 유사'하다고 나타난 MARC 레코드 사례

2. 실험 2: 8가지 변형 규칙을 적용한 실험데이터

실험 2는 임베딩 기반 유사도 검색의 장점에 집중하고자 실제 도서관 현장에서 나타나는 MARC 레코드 표기 방식의 차이 중에서 표제나 저자, 출판사를 기술하는 과정에서 문자열 표기 차이로 인해 중복레코드로 나타나는 사례를 실험데이터로 구축(〈표 2〉 참고)하고 알고리즘을 설계 및 구현하였다.

그리고 실험 2의 중복검증 알고리즘의 성능을 평가하기 위해 다음과 같이 두 가지 기준을 마련하였다. 첫 번째는 복본 집단이라고 명시되어 있는 레코드 중 실제 같은 클러스터에 속한 비율을 계산하는 복본 집단 탐지율로, M 도서관의 복본을 얼마나 잘 찾아내는가를 보기 위한 기준이다. 두 번째는 원형 집단과 변형 집단이 같은 클러스터에 속한 비율을 계산하는 원형 변형 매칭률로 8가지 변형 규칙을 적용한 레코드가 동일한 레코드라고 판정하는 즉, 표기 방식의 차이를 중복으로 잘 판정하는가를 보기 위한 기준이다. 8가지 변형 규칙을 적용한 8개 JSON 파일의 중복검증 성능 평가 결과는 아래 〈표 4〉와 같다.

〈표 4〉 실험 2 중복검증 결과

변형 규칙	최적 임계치	복본 집단 탐지율			원형-변형 집단 매칭률		
		레코드 수 (건)	매칭 수 (건)	비율 (%)	레코드 수 (건)	매칭 수 (건)	비율 (%)
rule_1: 관제 삭제	0.90	84	84	100	1,000	1,000	100
rule_2: 표제관련정보 표기	0.98	84	84	100	1,000	1,000	100
rule_3: 대등표제 표기	0.98	84	84	100	1,000	1,000	100
rule_4: 권표제 표기	0.98	84	84	100	1,000	1,000	100
rule_5: 권차 삭제	0.94	84	84	100	1,000	1,000	100
rule_6: 책임표시 간소화	0.98	84	84	100	1,000	1,000	100
rule_7: 역할어 표기	0.98	84	84	100	1,000	1,000	100
rule_8: 임프린트 표기	0.96	84	84	100	1,000	1,000	100

위 표에 나타난 바와 같이, 연구진이 구축한 실험데이터에 대해서는 모든 복본 레코드를 식별하고, 8가지 변형 규칙 전체에 대해 원형 레코드와 변형 레코드를 정확하게 동일한 클러스터로 분류하는 것으로 나타났다. 그리고 실험데이터마다 최적의 임계치를 계산하였는데 최저 0.90에서부터 최대 0.98까지 임계치의 변화가 나타나, 실험 1과 같이 유사도 임계치를 연구자가 임의로 설정하는 것보다는 평가하고자 하는 지표값을 최적화하는 임계치에 근거한 유사도 임계치를 제시하는 것이 바람직하다고 볼 수 있다.

Ⅴ. 결 론

본 연구에서는 MARC 레코드 중복검증을 위한 기존의 규칙 기반 알고리즘의 한계를 극복하고자 AI 임베딩 모델을 활용하여 중복검증 알고리즘을 구현, 실험데이터를 구축하고 성능을 평가하였다. 알고리즘은 1) MARC 레코드를 필드별로 구분해 하나의 문자열 덩어리로 만드는 텍스트 전처리 과정, 2) AI 임베딩 모델로 텍스트를 벡터 값으로 임베딩하는 과정, 3) 코사인 유사도를 계산하고 Faiss로 벡터 간 최근접 탐색을 수행하는 과정, 4) Union-Find 구조로 클러스터링하는 과정, 5) 평가 기준으로 성능을 평가하고 최적의 임계치를 찾는 과정으로 구현되었다. 이는 규칙 기반 중복검증 알고리즘을 연구한 연구진의 선행연구와 동일한 데이터를 가지고, 연구 초기(실험 1)에 구현한 알고리즘을 평가한 결과를 반영한 것이다.

실험 1에서는 MARC 레코드의 주요 필드를 추출하고 GPT-3.5 기반 모델로 임베딩한 결과와 전체 필드를 추출하고 GPT-4 기반 모델로 임베딩한 결과를 비교하였다. 먼저 주요 필드 추출 후 GPT-3.5 기반 모델로 임베딩한 결과 동일 집단의 경우 매우 유사하다고 판정된 레코드가 98.45%로 나타나 연구진의 선행연구의 결과보다 0.35% 높게 나타났다. 하지만 불일치 집단의 경우 다름으로 나타난 레코드 건수보다 유사하다고 판정된 레코드 건수가 월등히 높게 나타나, 전체 필드를 추출하고 GPT-4 모델로 임베딩을 변경하였다. 그 결과 코사인 유사도 0.75 미만의 다름 판정을 받은 레코드 수가 66건 2.17%에서 2,325건 76.30%로 늘어나 임베딩 성능이 크게 향상된 것으로 나타났다.

다만 동일하지 않은 레코드를 유사하다고 판정한 사례가 여전히 존재해 이를 분석해 보니 표제, 저자, 출판사 등 문자열 데이터는 일치하지만 권호 정보가 다른 '다권본'인 경우인 것으로 나타났다. 이에 따라 임베딩 모델 기반 유사도만을 적용한 알고리즘은 숫자나 특수 기호 처리에는 한계가 있음을 인지하였고, 실험 2에서 임베딩 기반 유사도 검색의 장점을 확인하기 위한 알고리즘을 구현하고 중복검증 성능을 평가하였다. 즉, 현장의 목록 레코드 작성 관행에 따른 중복 발생의 원인을 실험데이터로 구축하고 알고리즘이 이를 얼마나 중복으로 잘 판단하는지 평가하였다. 평가

결과 전체 실험데이터에 대해 복본 레코드와 변형 규칙을 100% 식별하는 것으로 나타났다.

이러한 실험을 통해 MARC 레코드 중복검증을 위해 임베딩 모델과 벡터 유사도 검색 기법을 활용하는 것의 이점과 한계가 명확히 드러났다고 볼 수 있다. 앞서 인용한 것처럼, 이러한 AI 기법의 적용은 "질의의 의미와 맥락을 이해하기 때문에, (중략) 맥락적으로 보다 적합한 결과를 제시"할 수 있으며, 임베딩 모델이 텍스트의 의미적 유사도에 기반한다는 점은 반대로 숫자 데이터의 정확한 매칭을 요구하는 영역에서는 한계로 작용할 수 있다는 것을 의미한다.

이에 본 연구의 실험에서 구축하지 못한 도서관 현장의 다양한 MARC 레코드 이형을 실험데이터로 확장할 필요가 있으며, 향후 이러한 다양한 이형에도 높은 정확도로 중복을 검증할 수 있는 알고리즘에 대한 설계 및 평가가 수행되어야 할 것이다. 특히 서론에서 언급한 것처럼 서지공유형 종합목록을 지원하고 차세대 서지 구조로의 전환에 대응하기 위해서는 개별 도서관이 아닌 지역 혹은 국가 단위의 MARC 레코드 중복검증의 정확도 향상을 목표로 알고리즘을 개선하고 프로토타입을 제작해 현장 데이터 기반의 단계적 검증이 필요할 것이다.

참 고 문 헌

AWS (발행년불명). 기계 학습에서 임베딩이란 무엇인가요? 출처: https://aws.amazon.com/ko/what-is/embeddings-in-machine-learning/

강우진, 정인영, 이종욱 (2023). 공공도서관 동시 대출 도서의 주제 연관성 분석 연구. 한국도서관·정보학회지, 54(3), 33-55.

권희정 (2018). [알고리즘] Union-Find 알고리즘. 출처: https://gmlwjd9405.github.io/2018/08/31/algorithm-union-find.html

나상오, 강우진, 이종욱 (2025). 한국문학 분야 KORMARC 서지레코드를 활용한 저작 식별 개선 방안 연구. 한국도서관·정보학회지, 56(2), 109-132.

노지현, 이은주 (2023). 공공도서관 서지데이터의 품질 제고 방안 - 부산시 공공도서관을 중심으로 -. 한국도서관·정보학회지, 54(3), 105-128. https://doi.org/10.16981/kliss.54.3.202309.105

송민건, 이수상 (2024). 공공도서관 목록데이터의 중복검증에 관한 연구 - 부산 지역 G도서관 사례를 중심으로 -. 한국도서관·정보학회지, 55(1), 1-26.

송민건, 이수상 (2025). 공공도서관 MARC 데이터 중복검증 알고리즘 개선 방안 연구 - 부산 지역 M도서관 사례를 중심으로 -. 한국도서관·정보학회지, 56(1), 289-305.

이수상, 이순영, 정철, 송민건 (2025). AI 기반의 도서관 서비스 혁신 방안 (이슈리포트). 두드림 & 부산대학교.

윤재혁, 도슬기, 오삼균 (2020). 저자역할용어사전 구축 및 저작군집화에 관한 연구. 정보관리학회지, 37(2), 197-223.

이용구 (2023). 단행본 서명의 단어 임베딩에 따른 자동분류의 성능 비교. 정보관리학회지, 40(4), 307-327.

이정훈, 정윤경 (2019). Word2vec 임베딩을 이용한 책 추천 시스템. 한국정보과학회 학술발표논문집, 922-924.

콘도 코이치, 황세웅, 황석형, 정영애 (2025). Word2Vec 유사도 기반의 협업 필터링을 이용한 도서 추천 시스템 제안. Journal of Platform Technology, 13(1), 31-42.

현유경 (2025). Embedding 모델. AI Interview Guide, 5-6. 위키독스. 출처: https://wikidocs.net/293578

Anvari, S. & Amirkhani, H. (2018). Book2Vec: representing books in vector space without using the contents. 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 176-182. doi: 10.1109/ICCKE.2018.8566329

Barkan, O. & Koenigstein, N. (2016). Item2VEC: Neural item embedding for collaborative filtering. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1603.04259

Business Insider (2025, October 10). ChatGPT is now being used by 10% of the world's adult population. Available: https://www.businessinsider.com/chatgpt-users-growth-openai-growth-sam-altman-ai-llm-2025-10/

CENL News (2022, May 23). German National Library launched its "Cataloguing machine". CENL. Available: https://www.cenl.org/german-national-library-launched-its-cataloguing-machine/

ENCORD (2023). What is Vector Similarity Search? Avaliable: https://encord.com/blog/vector-similarity-search/

Ganesh, J., Ganguly, S., Gupta, M., Varma, V., & Pudi, V. (2016). Author2Vec: learning author representations by combining content and link information. WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web, 49-50. https://doi.org/10.1145/2872518.2889382

IBM (n.d.). What is vector search? Avaliable: https://www.ibm.com/think/topics/vector-search

LangChain Docs (n.d.). Embediing Models. Avaliable: https://docs.langchain.com/oss/python/integrations/text_embedding/index#emb

edding-models

Meta (2017). Faiss: A library for efficient similarity search. Avaliable: https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1301.3781

National Library of Finland (2023, April 9). Automatic subject indexing tool Annif ready for wider production use. National Library of Finland News. Available: https://www.kansalliskirjasto.fi/en/news/automatic-subject-indexing-tool-annif-ready-wider-production-use

OpenAI Platform (2024, January 25). New embedding models and API updates. Avaliable: https://openai.com/index/new-embedding-models-and-api-updates/

Reuters (2023, February 3). ChatGPT sets record for fastest-growing user base - analyst note. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Yoneda, T., Mori, K., Miwa, M., & Sasaki, Y. (2017). Bib2vec: embedding-based search system for bibliographic information. Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 112-115. https://doi.org/10.18653/v1/e17-3028

Zhang, H., Wang, S., Wang, E. K., Li, Y., Zhang, Y., & Chu, D. (2017). Recommending e-books by multi-layer clustering and locality reconstruction. 2022 IEEE 20th International Conference on Industrial Informatics (INDIN), 1056-1061. https://doi.org/10.1109/indin.2017.8104919

Zhang, H., Wang, S., Zhao, M., Xu, X., & Ye, Y. (2018). Locality reconstruction models for book representation. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1873-1886. https://doi.org/10.1109/tkde.2018.2808953

Zhang, H., Wang, S., Xu, X., Chow, T. W. S., & Wu, Q. M. J. (2018). Tree2Vector: learning a vectorial representation for Tree-Structured data. IEEE Transactions on Neural Networks and Learning Systems, 29(11), 5304-5318. https://doi.org/10.1109/tnnls.2018.2797060

Zhao, H., Wu, H., Li, J., Zhang, H., & Wang, X. (2020). DTree2VEC: a high-accuracy and dynamic scheme for real-time book recommendation by serialized chapters and local fine-grained partitioning. IEEE Access, 8, 23197-23208. https://doi.org/10.1109/access.2020.2968220

• 국한문 참고문헌의 영문 표기
(English translation / Romanization of references originally written in Korean)

AWS (n.d.). What is embeddings in machine-learning? Available: https://aws.amazon.com/ko/what-is/embeddings-in-machine-learning/

Hyeon, Yu-gyeong (2025). Embedding Model. AI Interview Guide, 5-6. Wikidocs, Available: https://wikidocs.net/293578

Kang, Woojin, Jeong, Inyeong, & Lee, Jongwook (2023). A study on the topical associations of simultaneously borrowed books in public libraries. Journal of Korean Library and Information Science Society, 54(3), 33-55.

Koichi, Kondo, Hwang, Se-Woong, Hwang, Suk-Hyung, & Jung, Young-Ae (2025). Proposal of a book recommendation system using collaborative filtering based on Word2Vec similarity. Journal of Platform Technology, 13(1), 31-42.

Kwon, Heejeong (2018). [Algorithm] Union-Find algorithm. Available: https://gmlwjd9405.github.io/2018/08/31/algorithm-union-find.html

Lee, Jung-Hoon & Cheong, Yun-Gyung (2019). Book recommendation system using Word2vec. Proceeding of Korean Institute of Information Scientists and Engineers, 922-924.

Lee, Soosang, Lee, Soon-Young, Jung, Chul, & Song, Min-Geon (2025). Strategies for AI-based library service innovation (Issue Report). Do Dream & Pusan National University.

Lee, Yong-Gu (2023). Performance comparison of automatic classification using word embeddings of book titles. Journal of the Korean Society for Information Management, 40(4), 307-327.

Na, Sangoh, Kang, Woojin, & Lee, Jongwook (2025). A study on improving work identification using KORMARC bibliographic records in the field of Korean literature. Journal of Korean Library and Information Science Society, 56(2), 109-132.

Rho, Jee-Hyun & Lee, Eun Ju (2023). Improving the quality of bibliographic data in public libraries: focusing on public libraries in Busan metropolitan city. Journal of Korean Library and Information Science Society, 54(3), 105-128. https://doi.org/10.16981/kliss.54.3.202309.105

Song, Min-Geon & Lee, Soosang (2024). A study on duplication verification of public library catalog data: focusing on the case of G library in Busan. Journal of Korean Library and Information Science Society, 55(1), 1-26.

Song, Min-Geon & Lee, Soosang (2025). A study on improving duplicate verification algorithm for public library MARC data: focusing on the case of M library in Busan. Journal of Korean Library and Information Science Society, 56(1), 289-305.

Yun, Jaehyuk, Do, Seulki, & Oh, Samgyun (2020). Designing a FRBR work grouping algorithm of bibliographic records using a role term dictionary of authors. Journal of the Korean Society for Information Management, 37(2), 197-223.