Journal of Korean Library and Information Science Society
Vol. 56, No. 2

IoT 센서 데이터 기반 Topical N-gram 기법을 활용한 도서관 이용자의 이동 경로 패턴 분석*

Analyzing Movement Path Patterns of Library Users Using a Topical N-gram Method Based on IoT Sensor Data

정 도 헌(Do-Heon Jeong)**
김 규 환(Gyuhwan Kim)***

〈 목 차 〉

요 약 : 본 연구는 IoT 기반 카메라 센서를 통해 수집된 공공도서관 이용자의 실시간 이동 데이터를 분석하여, 공간 내 잠재 이동 경로 패턴을 정량적으로 도출하고 시각화하는 자동 분석 체계를 구축하고자 하였다. 이를 위해 센서를 통해 수집된 연속적 이동 데이터를 N-gram 방식으로 구조화한 후, LDA(Latent Dirichlet Allocation) 토픽 모델링 기법을 적용하였다. 이때 TF-IDF와 Word2Vec 기반의 두 가지 용어 가중치 방식을 bigram과 trigram 모델에 각각 결합하여 총 4종의 분석 모델을 구성하였으며, 각 모델의 토픽 분포를 비교하고 이동 흐름의 구조적 특성을 시각화하였다. 또한 기존 LDA 모델의 한계인 이동 방향성과 순서 정보의 미반영 문제를 보완하기 위해, Topical N-gram 기법을 적용한 분석 방법을 함께 제안하였으며, 각 모델의 분석 결과는 코사인 유사도와 JSD(Jensen-Shannon Divergence)를 활용한 앙상블 방식으로 통합하였다. 실험 결과, 단순 통계 방식으로는 확인하기 어려운 의미 있는 반복 이동 흐름이 토픽 단위로 도출되었으며, '출입구'와 '안내데스크'의 안내 및 참고 서비스 기능을 중심으로 주요 이동 경로들이 명확히 식별되었다. 본 연구는 실시간 이동 데이터를 기반으로 이용자 행태를 정량적으로 해석하고, 공공 서비스 운영과 기획에 활용 가능한 통합 분석 체계를 제시하였다는 점에서 의의가 있다.

주제어 : 이동 경로 패턴, IoT, LDA 토픽 모델링, N-gram, 앙상블 기법

ABSTRACT : This study aims to establish an automated analysis framework that quantitatively derives and visualizes latent movement path patterns within a public library, using real-time movement data collected through IoT-based camera sensors. To this end, continuous movement data captured by the sensors were structured using an N-gram approach and analyzed using Latent Dirichlet Allocation (LDA) topic modeling. Two term weighting methods—TF-IDF and Word2Vec—were each combined with bigram and trigram models, resulting in four analytical models. Topic distributions from each model were compared, and the structural characteristics of movement flows were visualized. To address the limitation of conventional LDA in capturing directionality and sequential information, the study also proposed an analysis method based on the Topical N-gram technique. The analysis results from each model were integrated using an ensemble approach based on cosine similarity and Jensen-Shannon Divergence (JSD). The experimental results revealed meaningful and repetitive movement patterns that are difficult to detect using simple statistical methods. In particular, key user routes centered around the 'entrance' and the 'information desk'—both serving as guidance and reference service hubs—were clearly identified. This study is significant in that it presents an integrated analysis framework capable of quantitatively interpreting user behavior based on real-time movement data, offering practical applications for the operation and planning of public services.

KEYWORDS : Movement Path Pattern, Internet of Things(IoT), LDA Topic Modeling, N-gram, Ensemble Method

* 본 연구는 2024년도 덕성여자대학교 교내연구비 지원에 의해 이루어졌음(3000010067).
** 덕성여자대학교 글로벌융합대학 문헌정보학전공 부교수
(doheonjeong@duksung.ac.kr / ISNI 0000 0004 6099 1600) (제1저자)
*** 인천대학교 문헌정보학과 교수(gyuhwan@inu.ac.kr / ISNI 0000 0004 6428 1251) (교신저자)

• 논문접수: 2025년 5월 16일 • 최초심사: 2025년 6월 1일 • 게재확정: 2025년 6월 9일
• 한국도서관·정보학회지, 56(2), 177-196, 2025. http://dx.doi.org/10.16981/kliss.56.2.202506.177

Ⅰ. 서 론

최근 IoT 센서 기술과 인공지능 기반의 데이터 분석 기법의 발전에 따라, 도서관을 비롯한 공공 영역에서는 이용자의 이동 경로 및 공간 이용 행태를 정량적으로 분석하려는 연구가 활발히 이루어지고 있다. 이러한 접근은 기존의 설문조사나 관찰 중심 방식으로는 포착하기 어려운 실시간 이동 패턴과 미세한 행태 변화를 효과적으로 파악할 수 있다는 점에서 주목받고 있다. 특히, 실시간 데이터 기반 분석은 서비스 운영의 효율성 제고는 물론, 안전 관리와 공간 배치 최적화에 실질적인 기초 자료로 활용될 수 있다.

선행연구에 따르면, AI 기반 카메라, Wi-Fi, Beacon 센서 등 다양한 장치를 활용하여 공공도서관 및 공공시설 내에서 이용자의 위치 정보, 체류 시간, 이동 경로 등을 수집·분석한 사례가 다수 수행된 바 있다(김규환, 정도헌, 2023; 박성재, 2019; Liu & Hsu, 2018; Qu, 2024). 이들 연구는 공간별 혼잡도 시각화, 맞춤형 서비스 제안, 응급 대응 전략 등 실무적 활용 가능성을 입증하고 있으며, 정보 서비스 분야의 데이터 기반 의사결정 환경의 확산에도 기여하고 있다.

그러나 센서를 통해 수집된 이동 경로 데이터는 복잡한 고차원 구조를 가지므로, 이를 단순한 빈도 분석이나 클러스터링 기법만으로 해석하는 데에는 한계가 있다. 이러한 한계를 극복하기 위한 방안으로, 최근에는 LDA(Latent Dirichlet Allocation) 토픽 모델링 기법을 적용한 실증 연구들이 주목받고 있다. 이 기법은 이용자 이동 데이터를 '문서', 경로를 '단어'로 간주하여 데이터를 구조화하고, 내재된 잠재 이동 패턴을 주제(topic) 단위로 추출할 수 있어 복잡한 행태 분석에 적합한 모델로 평가받고 있다(조아 외, 2015; Chu et al., 2014; Liu & Hsu, 2018; Mohamed et al., 2014).

본 연구는 도서관 내 이용자의 실시간 이동 데이터를 수집·분석하고, LDA 기반 토픽 모델링을 활용한 행태 분석의 프레임워크를 제안하고자 한다. 또한, 기존 모델에서 지적된 '단어 간 방향성과 순서 정보가 반영되지 않는' 한계를 보완하기 위해 Topical N-gram 기법을 적용함으로써, 연속적인 이동 경로의 흐름을 더욱 정밀하게 반영하고자 한다. 이러한 시도는 도서관을 포함한 다양한 공공 서비스 환경에서 사용자 중심의 공간 설계 및 운영 개선에 기여할 수 있는 실증적 기반을 제공하며, 데이터 기반 의사결정 과정에 새로운 기술적 방법론을 제시한다는 점에서 그 의의가 있다.

2장에서는 관련 이론과 선행 연구 사례를 검토하고, 3장에서는 센서를 통해 수집된 데이터를 가공하여 잠재된 이동 패턴을 도출하는 데이터 전처리 방안을 설명한다. 4장에서는 텍스트마이닝 기법과 효과적인 분석 도구를 통해 주요 이동 경로 패턴을 추출하고 이를 해석하며 자동화된 통합 모델을 생성하는 앙상블 기법을 제안한다. 마지막으로 본 연구의 결과와 의의, 향후 연구 방향 등을 논의하고자 한다.

Ⅱ. 이론적 배경

1. 이용자 행태 및 이동 패턴 분석 연구

최근 디지털 전환과 IoT 기술 등 ICT(Information and Communication Technology) 제반 환경의 발전으로 인해 도서관을 비롯한 공공 분야에서 이용자의 행태 및 이동 패턴 분석 연구의 기술적 진보가 이루어지고 있다. 이용자 데이터의 정밀 분석은 단순히 공간 내 분포를 파악하는 것에 그치지 않고, 서비스 개선, 공간 효율성 증대, 안전 관리 및 비상 상황 대응 등 다양한 측면에서 실질적인 의사결정의 근거 자료로 활용될 수 있다. 국내에서는 공공도서관에 AI 카메라를 설치하여 이용자의 성별, 연령대, 체류 공간, 이동 동선을 분석하고 그 결과를 기반으로 공간 재배치, 운영시간 조정, 군집 기반 맞춤형 서비스 기획에 대한 시사점을 도출한 연구(김규환, 정도헌, 2023)와 스마트폰의 Wi-Fi 신호를 활용해 도서관 이용자의 공간 활용 패턴을 분석하고 개선 전략을 제안한 연구(박성재, 2019) 등이 수행된 바 있다. 해외에서는 공공 인프라 분야에서 센서 네트워크와 빅데이터 분석을 접목한 이용자 행태 분석 연구가 활발히 이루어지고 있다. 영국의 공공도서관에서 Wi-Fi 인터넷 접근 제한 정책이 이용자 행동에 미치는 영향을 분석하고 정보 접근성과 이용자 만족도 간의 관계를 탐색하였다(Spacey et al., 2017). 미국의 대학도서관에서 Wi-Fi 연결 로그를 분석하여 공간별 체류 패턴과 혼잡도 시각화를 통해, 도서관 관리자에게 실시간 방문자 대시보드를 제공하는 방식을 구현하였다(Qu, 2024). 이와 같이, 도서관 및 공공기관에서의 이용자 행태와 이동 패턴 분석은 공간 운영 효율성과 안전 관리뿐 아니라, 이용자 경험 개선에 있어서도 핵심적인 역할을 수행하고 있으며, 이를 위한 데이터 기반 분석 기술의 실전적 연구는 지속적으로 확대될 전망이다.

2. LDA 토픽 모델링 기법과 Topical N-gram 기법을 이용한 패턴 분석 연구

LDA(Latent Dirichlet Allocation)는 문서 집합 내에 잠재하는 토픽들을 추론하기 위한 확률 기반 주제 모델로, 각 문서는 여러 토픽의 혼합으로 표현되며, 각 토픽은 특정 단어의 분포로 나타난다(Blei et al., 2003). 이러한 특성 덕분에 LDA는 자연어 텍스트 분석뿐 아니라, 시공간적 구조를 가지는 이동 경로 데이터 분석에도 효과적으로 적용될 수 있으며 기존의 방법으로는 포착하기 어려운 미세한 이동 패턴과 행태 변화를 밝혀내는 연구에 활용되고 있다. LDA 기법은 도서관, 공공기관, 교통, 문화시설 등 다양한 분야에서 이용자 행태 분석에 적용되어 왔다. 도시 교통카드 데이터를 활용하여 LDA 모델을 적용한 연구에서는 각 승객의 승하차 기록을 문서로, 이동 패턴을 단어로 간주하여 분석한 결과, 주거지역에서 상업지역으로의 통행, 시외지역에서 중심지로

의 유입 등의 주요 유형을 도출하고 도시계획과 교통정책 수립에 활용될 수 있음을 보여주었다(조아 외, 2015). 유사한 연구 사례로, 택시의 GPS 트랜잭션 데이터를 기반으로 토픽을 추출하고 시각화하였으며(Chu et al., 2014), 프랑스 Rennes 지역에서 교통카드와 인구통계 데이터를 결합하여 이동 유형을 분석하기도 하였다(Mohamed et al., 2014). 도서관 분야에서 LDA 기법이 적용된 사례로, 대만의 대학도서관에서 비콘(beacon) 센서를 통해 수집된 실내 위치 데이터를 LDA로 분석하여 열람 공간, 정보 검색 공간, 휴게 공간 등 공간 이용 행태를 토픽 단위로 시각화하였다(Liu & Hsu, 2018).

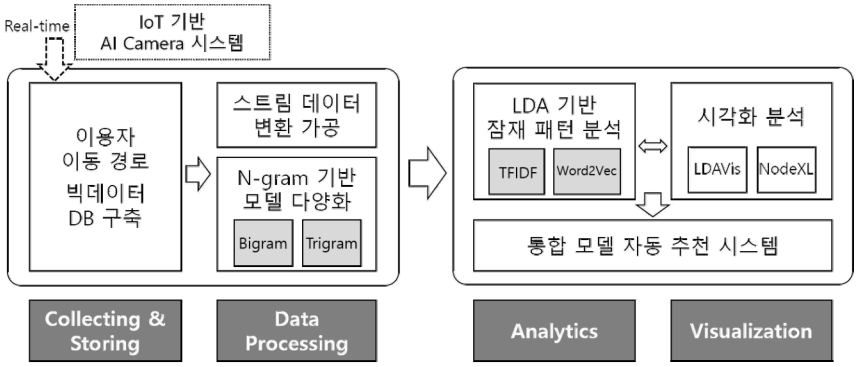
이러한 연구들은 LDA가 단순 통계 분석보다 더 세밀하게 이용자의 복합적 행태를 파악하는 데 유용함을 보여준다. 특히, 공간 운영 효율성 제고, 긴급 상황 대응, 정보 서비스 최적화 등의 실무적 활용 가능성이 강조되고 있다. 그러나, LDA 토픽 모델링 기법은 대량의 데이터로부터 인간이 파악하기 어려운 잠재된 행태의 패턴 추출에는 효과적이지만, 시계열 변화나 연속적인 행위 흐름을 반영하는 데는 구조적인 한계를 가진다. 이를 보완하고자 DTM(Dynamic Topic Model)이 제안되기도 하였다(Blei & Lafferty, 2006). DTM은 시간 단위로 분할된 문서 집합에서 각 토픽의 단어 분포가 시간에 따라 점진적으로 변화하도록 모델링함으로써, 토픽의 시간적 진화를 설명할 수 있다. 그러나 DTM 역시 단어의 동시 발생에 기반한 unigram 기반의 모델로, 단어 간 순서나 맥락적 흐름을 고려하지 못한다는 점에서 한계를 지닌다. 특히 이용자의 실제 이동 경로나 행동 순서를 분석할 필요가 있는 경우에는 충분한 설명력을 제공하지 못한다. 이를 보완하기 위한 방법으로 제안된 것이 Topical N-gram 모델이다. 이 모델은 단어 간 순서를 반영하며, 연속된 행동이나 이벤트의 흐름을 고려한 분석이 가능하며 시간적으로 연결된 이용자의 이동 행위나 장소 전이 패턴을 정밀하게 추적할 수 있다는 점에서 주목을 받고 있다(Lin et al., 2010; Wallach, 2006; Wang et al., 2007). 이와 같이, 도서관을 비롯한 다양한 공공 분야에서 진행된 다양한 연구 사례들은 Topical N-gram을 비롯한 LDA 기반의 응용 기법들이 이용자의 동선 파악 및 공간 활용 분석에 있어 강력한 도구로 활용될 수 있음을 보여준다.

Ⅲ. 이용자의 이동 경로 데이터 처리 방안

1. 시스템 구성 및 연구 절차

〈그림 1〉은 이용자 이동 패턴 분석 프레임워크 구축을 위한 시스템 구성 및 실험 절차를 요약적으로 설명한다. 전체 과정은 데이터의 수집 및 저장, 실시간 데이터에 대한 가공 처리, 대용량 데이터의 분석, 실험 결과의 해석과 시각화의 4단계로 구성된다. 우선 AI 카메라를 기반으로 한

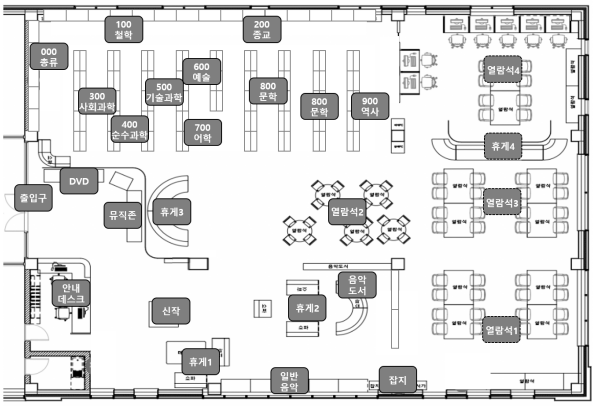
IoT 센서 시스템을 통해 실시간으로 획득한 대량의 데이터를 1차 가공하여 이용자의 이동 경로 단위의 데이터를 구축한다. 이때 실시간 스트림 데이터는 기준 시간 이상을 머문 데이터로 필터링한 후 일련의 이동 경로를 표현하는 데이터 형식으로 구축한다. 본 연구에서는 지속시간 5초 이상의 데이터를 필터링하여 사용한다. 데이터 처리 과정에서 N-gram 모델을 사용하였으며, 실험에 사용한 두 가지 모델은 bigram과 trigram이다. 분석가가 대량의 데이터를 효과적으로 분석하기 위한 빅데이터의 차원 축소 방법으로 LDA 토픽 모델링 기법을 사용한다. 이 과정에서 입력 단위인 단어(Word)의 가중치 부여 방식을 단순 가중치 모델인 TF-IDF와 단어 간의 의미 관계를 반영하는 모델인 Word2Vec의 두 가지 모델을 사용한다. 이와 같이 본 연구에서는 N-gram 모델과 용어 가중치 모델을 각각 2종씩 사용하여 총 4종의 분석 모델을 구축하여 실험하고, 분석 결과의 시각화를 통해 각각의 특징을 살펴본 후, 최종 통합 모델을 자동으로 구축하는 앙상블 시스템을 제안하고자 한다. 시각화를 통한 분석 방법으로는 직관적으로 토픽을 분석할 수 있는 LDAVis를 활용하고, NodeXL을 이용해 그래프 기반의 네트워크 분석을 수행한다.



〈그림 1〉 이용자 이동 패턴 분석 프레임워크

2. 데이터 수집 및 저장

본 연구에서 제안한 공공기관 이용자의 잠재된 이동 경로 패턴을 자동으로 분석하기 위한 실험용 데이터셋을 수집하였다. 인공지능 기반 분석 전문 기업인 트리플렛(Triplet, https://triplllet.com/)의 AI 카메라 센서 기술을 이용하여 인천 소재 B 공공도서관에 직접 설치・운영하여 획득한 이용자 데이터를 활용하였다. 〈그림 2〉는 B 도서관의 평면도를 바탕으로 가공한 것이며 이용자 이동 경로를 획득하기 위한 센서들의 위치를 표시하고 있다. 센서를 통해 5개의 카테고리에 대해 총 16개의 세부 위치를 기록한 레이블 정보를 수집하였다. 또한, 〈그림 2〉의 평면도에서 '열람석1', '열람석2', '휴게1' 등 이용자 공간의 일부 위치 정보는 자료 제공 및 수집에 제한을 두었음을 밝힌다.



Category	Place
출입구	출입구
안내데스크	안내데스크
서가	100_철학
	200_종교
	300_사회과학
	400_순수과학
	500_기술과학
	600_예술
	700_어학
	800_문학
	900_역사
	DVD
	뮤직존
	신작
	음악도서
	일반음반
	잡지
열람석	열람석1
	열람석2
	열람석3
휴게	휴게1
	휴게2
	휴게3
	휴게4

〈그림 2〉 도서관 평면도 내 센서 위치(좌) 및 레이블 정보(우)

AI 카메라 센서로부터 수집된 실시간 데이터의 구조는 〈그림 3〉과 같다. 실시간 데이터는 고유 id가 자동 부여되며, 기관 id(store_id)를 가지고 있다. 출입 시 자동으로 개체가 식별되어 부여되는 이용자 번호(guest_code)는 시작 시점부터 종료 시점까지 실시간으로 일정 시간 간격에 따라 연속적으로 위치 정보를 저장한다. 이용자에 대한 상세 정보는 별도의 DB 테이블로 관리된다. 이용자 데이터 확보를 위한 시스템의 시범 운영 기간은 2023-01-03 08:56:44부터 2023-02-28 21:59:36까지 약 두 달간이며, 최종 구축된 실시간 캡처 데이터는 총 4,524,762건, 이동 경로 패턴 수(고유한 guest_code 수)는 총 5,403건이었다.

id	store_id	guest_code	created_datetime	category	place
54792118	16	20230103_0904_81175	2023-01-03 9:25	책장	뮤직존
54792119	16	20230103_0904_81175	2023-01-03 9:25	책장	뮤직존
54792120	16	20230103_0904_81175	2023-01-03 9:25	책장	뮤직존
54792121	16	20230103_0904_81175	2023-01-03 9:25	책장	300_사회과학
54792586	16	20230103_0904_81598	2023-01-03 9:38	안내데스크	안내데스크
54792587	16	20230103_0904_81598	2023-01-03 9:38	책장	일반음반
54792588	16	20230103_0904_81598	2023-01-03 9:38	책장	일반음반
54792589	16	20230103_0904_81598	2023-01-03 9:38	책장	800_문학
54792590	16	20230103_0904_81598	2023-01-03 9:38	책장	800_문학
54792591	16	20230103_0904_81598	2023-01-03 9:38	책장	200_종교
54792592	16	20230103_0904_81598	2023-01-03 9:38	책장	800_문학
54792593	16	20230103_0904_81598	2023-01-03 9:38	출입구	출입구

〈그림 3〉 센서 데이터의 이용자별 위치 정보 DB 테이블 구조

3. 이동 패턴 정보 생성을 위한 데이터 처리

이용자의 이동 패턴 분석 성능을 높이기 위해, 본 연구에서는 두 가지의 데이터 가공 방안을 제안한다. 첫 번째는 N-gram 기법을 활용한 방향성 있는 연속적 데이터 구성이다. 기존의 LDA 모델은 입력 데이터로 unigram 기반의 단어 분포를 활용하기 때문에 시간적 순서나 방향성을 반영하지 못한

다는 한계를 가진다. 특히, 이동 경로와 같은 연속적 데이터에서는 이러한 방향성 부재가 의미 있는 패턴 분석을 저해할 수 있다. 이를 보완하기 위해 본 연구에서는 연어(collocation) 기반의 N-gram 기법을 바탕으로 한 LDA 응용 모델인 Topical N-gram 모델을 사용하였다. 이동 경로 데이터를 추출하는 방법으로 bigram과 trigram을 사용하며 quadgram(또는 4-gram)은 후술한 이유로 인해 실험에서 제외하였다. 각 모델의 특징과 한계를 설명하면 다음과 같다. Bigram 모델은 인접한 두 지점 간의 이동 정보(A-B)를 그대로 추출하여 명확하게 이동 흐름을 얻을 수 있다. 단일 그래프(graph)만 생성하여 네트워크를 생성하기 때문에, 최종 네트워크가 다소 단조로울 가능성이 있다. 이에 비해, trigram 모델은 세 지점 간의 연속된 이동 정보(A-B-C)를 단위로 데이터를 생성하므로, 반복 패턴과 이동 정보를 더욱 풍부하게 생성할 수 있다. quadgram 모델은 A-B-C-D의 4단계의 긴 연속 관계를 생성할 수 있는 반면, 윈도우 크기의 과다 설정으로 인해 중요 패턴의 누락 가능성과 데이터 과생성으로 인한 희소성 문제가 증가할 수 있어 본 연구에서는 제외하였다. 이러한 데이터 과생성과 희소 현상은 다음 장의 추출 데이터 통계(〈표 1〉 참조)를 통해 부가 설명한다. 이와 같은 N-gram 기법 기반의 분석 방안은 이용자의 실제 이동 경로에 내재된 방향성과 연속성을 직접 반영함으로써, 이용자의 이동 경로와 같은 연속적 데이터에 대한 토픽 모델링 기법의 해석력을 향상시킨다.

이용자 이동 패턴 분석의 성능 향상을 위한 두 번째 데이터 처리 방안은 출발지(SP, Starting Point) 및 도착지(EP, End-Point)에 대한 더미(dummy) 플래그 삽입 방법이다. 센서로부터 수집된 이동 경로 데이터의 원본은 센서에서 획득한 위치 및 전이 정보만을 제공하므로, 특정 지점이 이용자 이동 경로의 시작점인지 중간 경로인지 구분하기 어렵다. 〈그림 4〉의 (예시 1)을 통해, 이동 경로가 (1) 'A-B-C' 데이터와 (2) 'C-A-B' 데이터에서 bigram 정보를 추출한다고 가정할 때, 모두 A에서 B로 이동했다는 'A-B' 정보는 제공하지만, 이용자가 서비스를 최초로 이용했다는 시작점 A((1)의 경우)에 대한 정보를 확인할 수 없다는 것을 알 수 있다. 이를 해결하고자 제공 서비스의 시작과 끝임을 알려주는 더미 플래그인 SP, EP를 각각 삽입하였다. 이러한 더미 플래그 삽입 기법은 외부로부터의 유입 경로를 추적하거나 출입 현상을 분석하는 데 유용하게 활용될 수 있다. 또한, 〈그림 4〉의 (예시 2)와 같이 이용자가 단일 지점만 방문한 경우에도 SP, EP의 삽입으로 인해 trigram 데이터가 생성되므로 데이터의 누락 없이 bigram과 trigram 정보를 추출할 수 있음을 확인할 수 있다.

예시 1	Problem	A → B → C vs. C → A → B
	Solution	**SP** → A → B → C → **EP**
예시 2	Problem	A
	Solution	**SP** → A → **EP**

〈그림 4〉 SP(starting point)와 EP(end-point) 더미 플래그 삽입 효과(예시)

4. 이동 패턴 N-gram 정보의 구축 결과 분석

〈표 1〉은 데이터베이스에 구축된 실험용 데이터의 통계 분석 결과를 요약한 것이다. 이용자의 이동 경로 수는 이동 세션 단위의 패턴 수를 의미하며, 입장하여 퇴장할 때까지의 동선이므로 이용자가 여러 번 방문하면 방문 수만큼 세션이 생성된다. 고유한 N-Gram 수는 unigram에서 trigram으로 갈수록 패턴의 다양성이 많아지므로 급격하게 증가하는 모습을 보인다. 평균, 중앙값, 최대, 최소는 unigram에서 trigram으로 갈수록 윈도우 크기가 1씩 증가하므로 값이 그만큼 감소한다.

특히 N-gram의 빈도 분포 표준 편차를 통해 몇 가지 사실을 확인할 수 있다. N-gram의 빈도 분포 표준 편차는 각 N-gram의 등장 빈도가 평균으로부터 얼마나 분산되어 있는지를 나타내는 지표로 빈도의 평균과 제곱 편차를 기반으로 계산된다. 높은 빈도 분포 표준 편차를 보이며 특정 N-gram이 매우 높은 빈도로 등장하는 경우, 이들의 영향력이 커지며 데이터 내 특정 패턴이 두드러진다는 의미이다. 반면 낮은 빈도 분포 표준 편차는 다수의 N-gram이 비슷한 빈도로 등장하는 것으로, 데이터가 균등하게 분포되어 있음을 나타낸다. 이는 데이터가 특정 패턴에 치우치지 않고 고르게 분포되어 있음을 의미하며, 패턴의 다양성이 높아진 trigram으로 갈수록 이러한 현상이 두드러짐을 확인할 수 있다. 빈도 분포 표준 편차는 데이터의 다양성과 집중도를 이해하는 데 중요한 역할을 하며, 텍스트 분석 및 이용자 행동 패턴 분석 등 다양한 분야에서 데이터의 특성과 모델의 성능 향상에 기여할 수 있다.

〈표 1〉 이동 패턴 데이터의 N-gram 추출 결과 통계 분석

	Unigram	Bigram	Trigram
이용자 이동 경로 수	5,403	5,403	5,403
N-gram 수	27,525	22,122	16,719
고유 N-gram 수	19	237	1,239
평균 N-gram 수	5.09	4.09	3.09
중앙값	4.0	3.0	2.0
최대 N-gram 수	80	79	78
최소 N-gram 수	3	2	1
경로당 N-gram 수 표준 편차	3.78	3.78	3.78
N-gram 빈도 분포 표준 편차	1,654.88	227.70	48.34

Ⅳ. 이동 패턴 분석과 통합 모델 자동 생성

1. LDA의 용어 가중치 모델

본 연구에서는 센서 기반 실시간 빅데이터의 차원 축소를 위해 LDA 토픽 모델링 기법을 사용한다. 앞서 언급한 바와 같이 LDA는 입력 단위로 단어(word)를 사용하므로 데이터의 방향성과 연속성을 반영하기 어렵다. 이러한 한계를 보완하고자, 본 연구에서는 Topical N-gram 모델을 제안하였으며, bigram과 trigram을 실험에 적용하였다. 또한 LDA의 단어 가중치 방식으로 널리 사용되는 TF-IDF(박주연, 정도헌, 2022; Blei & Lafferty, 2009)와 함께 Word2Vec을 추가 제안하여 총 4가지의 분석 모델을 생성하고 결과를 비교하고자 하였다(〈표 2〉 참조). Word2Vec은 의미적으로 유사한 단어 간의 분포를 벡터 공간에 반영하여, LDA의 입력 단어 간 의미적 연관성을 반영한 가중치를 부여하는 데 사용된다. 이로써 LDA는 더욱 의미 일관성이 높은 토픽을 학습하여 품질을 향상시킬 수 있다(Nguyen et al., 2015). 본 연구에서는 Word2Vec 구현을 위해 파이썬의 Gensim 라이브러리를 사용하여 단어 임베딩을 수행하였다. 실험 시, Skip-gram에 비해 학습 속도가 빠르고 고빈도 선호 경향을 갖는 CBOW(sg=0)를 기본 모델로 사용하였다. 단어 벡터는 100차원(vector_size=100)으로 설정하였으며, 중심 단어를 기준으로 앞뒤 5개의 단어(window=5)를 문맥으로 고려하였다. 모든 단어를 학습에 포함(min_count=1)하였으며, 성능 향상을 위해 4개의 스레드를 병렬 처리(workers=4)하였다.

〈표 2〉 이동 패턴 분석을 위한 4가지 LDA 모델 생성 방안

		LDA의 용어 가중치 모델	
		TF-IDF	Word2Vec
패턴 데이터 추출 방법 (N-gram)	Bigram	**Model 1 (BI_TFIDF_LDA)**	**Model 2 (BI_W2V_LDA)**
	Trigram	**Model 3 (TRI_TFIDF_LDA)**	**Model 4 (TRI_W2V_LDA)**

2. 시각화 및 패턴 분석(Analytics)

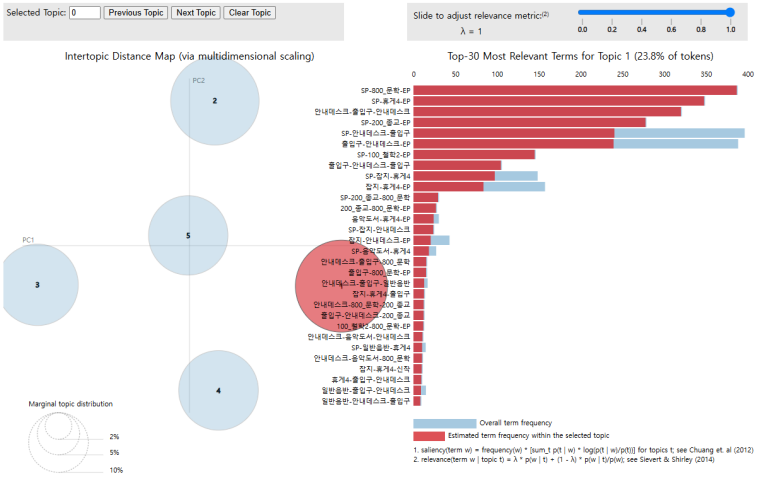
본 장에서는 앞서 제시한 4가지 모델의 분석과 결과 해석을 수행한다. 가장 단순한 베이스라인 모델인 TF-IDF 기반 bigram 모델(BI_TFIDF_LDA)로부터, 가장 복잡한 Word2Vec 기반 trigram 모델(TRI_W2V_LDA)까지 분석 프로세싱을 수행한 후, 주요 패턴의 특징을 중심으로 논의하고자 한다. LDA의 토픽 수(K)는 일반적으로 Perplexity, Coherence 등 정량적 지표를 고려하기도 하나(박주연, 정도헌, 2022), 본 연구는 해석의 용이성과 정책적 활용성을 고려하여

전문가가 직관적으로 해석할 수 있는 수준의 토픽 수가 효과적인 것으로 판단하였다. LDAvis의 분석 결과 데이터를 활용한 사전 실험을 통해 3~10개의 토픽 수를 비교한 결과, 중복성이 낮고 패턴이 다양하게 나타나는 5개의 토픽 수를 최종적으로 채택하였다. LDA를 통해 생성된 토픽의 예시는 〈표 3〉과 같다. 모든 연산 과정에서는 전체 단어를 이용하나, 지면의 제약으로 상위 7개 단어만을 표기하였다. 본 연구에서는 파이썬의 Gensim 라이브러리를 사용하여 토픽 모델링을 수행하였으며, 추론을 위해 Variational EM(Expectation-Maximization) 알고리즘을 사용하였다(Blei et al., 2003; Jeong & Song, 2014). E-step(iterations=400)은 변분 추론(Variational Inference)을 통해 주어진 단어 집합(문서)을 통해 문서별 토픽 분포를 추정하고, M-step은 이를 기반으로 토픽-단어 분포를 업데이트하며, 이 추론 과정을 지정된 횟수(passes=20)를 반복하였다. 또한 하이퍼파라미터로 alpha='auto' 및 eta='auto'를 설정하여 학습 데이터를 기반으로 '문서-토픽' 및 '토픽-단어' 분포에 대한 비대칭 디리클레 사전 분포(Asymmetric Dirichlet Prior Distribution)를 자동으로 학습하여 데이터의 특성에 맞는 최적의 분포를 추정할 수 있도록 하였다.

〈표 3〉 TRI_TFIDF_LDA 모델의 토픽 생성 결과 예 (상위 7개 단어만 표기)

Topic1	Topic2	Topic3	Topic4	Topic5
SP-800_문학-EP	SP-안내데스크-EP	SP-잡지-EP	SP-음악도서-EP	휴게4-잡지-휴게4
SP-휴게4-EP	SP-출입구-안내데스크	SP-안내데스크-일반음반	안내데스크-출입구-EP	잡지-휴게4-잡지
안내데스크-출입구-안내데스크	출입구-안내데스크-EP	안내데스크-일반음반-EP	SP-일반음반-EP	SP-안내데스크-800_문학
SP-200_종교-EP	음악도서-800_문학-음악도서	SP-안내데스크-음악도서	SP-안내데스크-출입구	안내데스크-800_문학-EP
SP-안내데스크-출입구	800_문학-음악도서-800_문학	SP-안내데스크-200_종교	신작-일반음반-신작	SP-출입구-EP
출입구-안내데스크-EP	SP-안내데스크-휴게4	휴게4-안내데스크-EP	일반음반-신작-일반음반	SP-휴게4-잡지
SP-100_철학-EP	음악도서-잡지-음악도서	안내데스크-200_종교-EP	SP-신작-일반음반	휴게4-잡지-EP

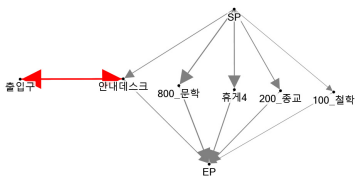
LDAvis(LDA Visualization)는 LDA의 결과를 시각적으로 표현하는 다양한 기능으로 인해 많은 연구에서 대표적인 시각화 도구로 활용되고 있다(Sievert & Shirley, 2014). 주요 기능은 토픽 간 상호 관계 시각화, 토픽의 주요 단어 분석, λ(람다) 파라미터 조정을 통한 데이터 필터링 등이다. λ의 값을 낮추면 해당 토픽의 대표 단어들이 강조되고, 값을 높이면 전체적으로 자주 등장하는 단어들이 강조되는 특징이 있다. 토픽의 크기는 전체 코퍼스(corpus)에서 그 토픽이 등장할 확률이 크다는 의미이다. 토픽 간의 거리는 토픽 간의 유사도를 의미하며 가깝게 위치한 토픽일수록 중복되는 단어나 내용이 많고, 멀리 위치한 토픽들은 서로 다른 주제를 다루며 이질적인 내용임을 나타낸다. 본 연구에서 생성한 〈그림 5〉의 경우, 각 토픽이 전체 영역에 고르게 분포되어 있어 분석이 균형 있게 이루어졌음을 확인할 수 있다. 본 연구에서는 Python 프로그래밍 언어와 pyLDAvis 라이브러리를 이용하여 LDA 토픽 생성과 동시에 시각화가 진행되도록 직접 프로그래밍하였다.

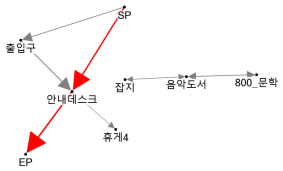
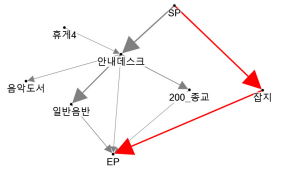
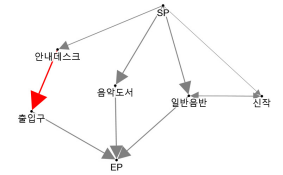
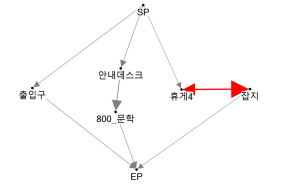


〈그림 5〉 TRI_TFIDF_LDA 모델의 LDAVis 시각화 예시

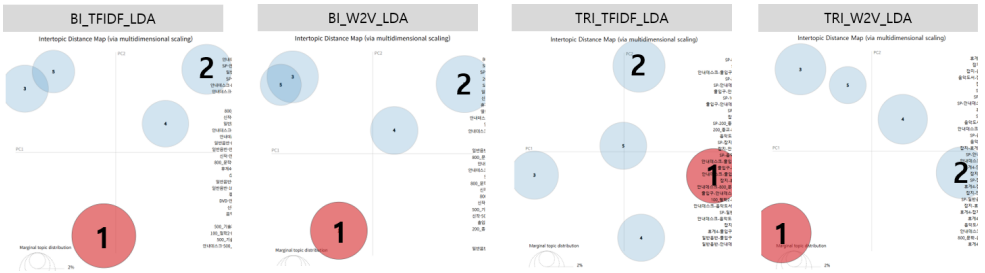
시각화를 통한 분석 결과의 구체적인 내용을 살펴보고자 한다. 본 연구에서는 N-gram을 기반으로 실제 이동 경로를 추출하고, 이를 단위 그래프(graph)로 구성하여 네트워크 시각화를 수행하였다. 주요 시각화 도구로는 MS Excel의 플러그인 소프트웨어인 NodeXL(https://nodexl.com/)을 사용하였다. 네트워크 데이터의 시각화는 다음과 같은 원칙에 따라 구성하였다. 시작점(SP)은 최상단에, 종단점(EP)은 최하단에 배치하였으며, 단방향 경로는 위에서 아래로 향하는 하향 화살표로, 상호 이동의 경우는 수평 방향의 양방향 화살표로 표현하여 시각적 직관성을 높이고자 하였다. 앞서 〈표 1〉에서 설명한 바와 같이, Trigram 모델은 Bigram 모델에 비해 더 풍부한 관계를 표현할 수 있는 특성을 지닌다. 이에 따라 시각화 분석에서는 trigram과 TF-IDF 가중치를 적용한 TRI_TFIDF_LDA 모델을 중심으로 전체 토픽에 대한 종합적 해석을 수행하였다(〈표 4〉 참조). 네트워크 내 주요 경로는 가장 높은 가중치를 갖는 굵은 간선으로 붉은색을 사용하여 강조하였고, 낮은 가중치를 가진 연결선은 얇은 지선으로 처리하여 보조적 경로로 해석하였다.

〈표 4〉 TRI_TFIDF_LDA 모델의 네트워크 시각화를 통한 상세 분석 (종합)

Topic Number	TRI_TFIDF_LDA 모델의 이동 경로 네트워크	토픽 경로 해석
T1		- 주요 경로: '출입구'와 '안내데스크' 간의 잦은 이동이 두드러짐. '출입구'의 다양한 공지 및 안내를 통한 '안내데스크'로의 이동이 빈번함. - 기타 경로: 열람실(문학, 종교, 철학)과 휴게실(4번)로 바로 이동한 후 퇴장하는 경로이며, 명확한 의도를 가진 시설 이용 행태를 보임.

Topic Number	TRI_TFIDF_LDA 모델의 이동 경로 네트워크	토픽 경로 해석
T2		- 주요 경로: 외부에서 바로 '안내데스크'로 이동하거나 '출입구'를 거쳐 '안내데스크'를 이용한 후 방문을 종료함. - 기타 경로: '잡지'-'음악도서'-'800_문학' 경로는 근거리 내에서의 서비스 간 상호 이동을 표현함(공간 구성의 효과 검토 측면).
T3		- 주요 경로: (1)'잡지' 코너로 직행하는 패턴이 가장 두드러짐. (2)입장 후 '안내데스크'를 거쳐 '일반음악' 코너로 행하는 경로가 대표적임. - 기타 경로: '안내데스크'를 기점으로 '음악도서', '일반음악', '200_종교' 등 다양한 서비스로 분기하는 패턴이 보임.
T4		- 주요 경로: '안내데스크'를 통해 다시 '출입구'로 이동하는 경향을 주로 보여줌. - 기타 경로: '음악도서' 및 '일반음악'으로 직접 이동과 '일반음악'-'신작' 코너 간의 상호 이동 경향도 확인 가능함.
T5		- 주요 경로: 입장 후 '휴게4'로 이동하여 '잡지' 코너와의 상호 이동 후 퇴장하는 경로가 가장 두드러짐(휴게공간의 효용성 검토 측면). - 기타 경로: '안내데스크'를 통해 '800_문학'으로 이동하는 경로 발견. 시설 외부에서 '출입구'의 공지 및 일정 등을 확인하는 행태를 보임.

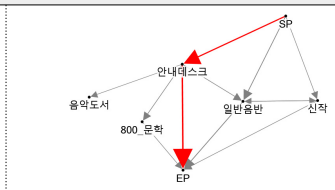
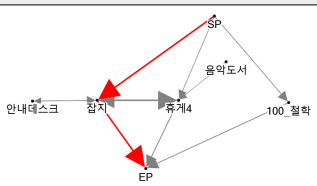
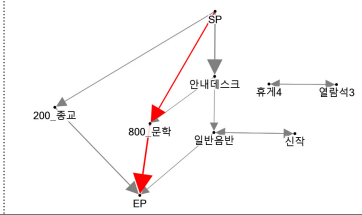
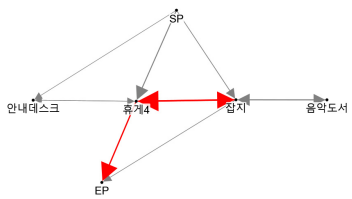
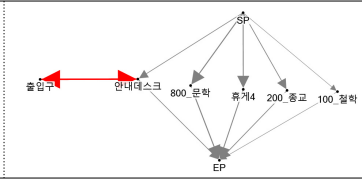
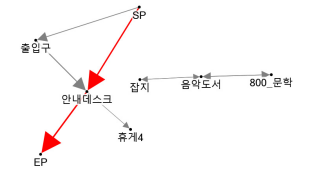
〈그림 6〉은 LDAvis 시각화를 활용하여 4가지 모델의 토픽 분포를 비교한 결과를 보여준다. 4분할된 그림 중 좌측에 위치한 bigram 모델들은 서로 유사한 분포를 보이며, 특히 좌측 상단에 위치한 3번과 5번 토픽은 높은 유사성을 나타낸다. 반면, 우측에 제시된 trigram 기반 모델은 토픽이 고르게 형성된 양상을 보이며, 이 중에서도 TRI_TFIDF_LDA 모델은 가장 균형 잡힌 토픽 분포를 나타낸다. 4개의 모델에서 토픽의 크기 정보를 통해 1번과 2번 토픽이 비중이 가장 높은 대표 토픽임을 확인할 수 있으며, 토픽 간 거리를 통해 두 토픽이 상이한 내용을 다루고 있음을 알 수 있다.

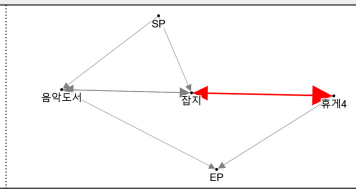
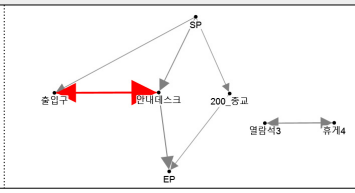


〈그림 6〉 LDAvis 시각화를 통한 4가지 모델의 토픽 분포 비교도

〈표 5〉는 4가지 모델의 대표 토픽인 1번과 2번 토픽을 시각화한 후, 주요 내용을 비교하여 정리한 결과를 제시하고 있다. 붉은색 간선을 중심으로 주요 경로의 유사점과 차이점을 비교하여 설명하였고 가중치가 낮은 보조 경로는 기타 경로로 분류하여 요약적으로 설명하였다. 파라미터 조정에 따라 생성되는 모델이 다양해질수록 해석의 풍부함이 증가하는 장점이 있다. 그러나 동시에 분석가가 선택해야 할 해석의 범위가 기하급수적으로 증가함에 따라, 해석 부담이 커지는 한계도 존재한다. 이에 따라 다양한 모델로부터 통합 모델을 자동으로 생성할 수 있는 추천 시스템의 필요성이 제기된다. 다음 장에서는 이러한 문제를 해결하기 위한 방안으로, 앙상블 기법을 활용한 통합 모델 생성 방법을 제안하고자 한다.

〈표 5〉 4가지 모델의 대표 토픽(T1 & T2) 시각화 및 경로 해석

모델 구분	토픽 경로 해석	토픽 경로 해석
BI_TFIDF_LDA	T1 	T2 
	- 주요 경로: 입장하여 '안내데스크' 이용 후 바로 퇴장하는 경로가 가장 두드러짐. - 기타 경로: 안내데스크를 중심으로 다양한 서비스 이용 행태를 보임. '일반음악'-'신작' 서비스 간의 상호 이동 경향이 나타남.	- 주요 경로: 입장 후 바로 '잡지' 코너로 이동 후 종료하는 경로가 대표적임. - 기타 경로: '잡지'와 '휴게4' 간의 잦은 이동 패턴이 발견되며, '음악도서'에서 '휴게4'로의 이동도 나타남(휴게공간의 활용 패턴 측면 검토).
BI_W2V_LDA	T1 	T2 
	- 주요 경로: 입장 후 바로 '800_문학' 코너로 이동하고 이용 후 바로 퇴장하는 패턴이 보임. - 기타 경로: '일반음악'-'신작', '휴게4'-'열람석3'의 상호 이동하는 경향들이 확인됨.	- 주요 경로: '휴게4'와 '잡지' 코너 간의 상호 이동과 '휴게4'에서 퇴장하는 경로가 두드러짐. - 기타 경로: '휴게4'-'잡지'-'음악도서' 간의 상호 이동이 패턴이 뚜렷하게 나타남.
TRI_TFIDF_LDA	T1 	T2 
	- 주요 경로: '출입구'의 각종 공지 및 안내를 통한 '안내데스크'로의 이동이 빈번함. - 기타 경로: 열람실(문학, 종교, 철학)과 휴게실(4번)로 바로 이동한 후 퇴장하는 경로(명확한 의도를 가진 시설 이용 행태)	- 주요 경로: 외부에서 '안내데스크'로 직접 이동하거나 '출입구'를 거쳐 '안내데스크'로 이동한 후 방문을 종료함. - 기타 경로: '잡지'-'음악도서'-'800_문학' 경로는 근거리 내에서의 서비스 간 상호 이동을 표현함(공간 구성의 효과 검토 측면).

모델 구분	토픽 경로 해석	토픽 경로 해석
TRI_W2V_LDA	T1 	T2 
	- 주요 경로: '잡지'-'휴게4'의 상호 이동이 가장 뚜렷하게 나타나는 이용 행태임. '잡지' 코너는 여러 서비스에 대한 연결점 역할을 하고 있음(허브 서비스의 중요성 검토). - 기타 경로: '음악도서'-'잡지'-'휴게4' 간의 이동이 확인됨. 입장 후 '음악도서'와 '잡지'로는 직행하나, '휴게4'로 직행하지는 않음.	- 주요 경로: '출입구'와 '안내데스크' 간의 상호 이동이 두드러짐(타 모델에서도 확인된 주요 경로 중 하나). '안내데스크' 문의 후 바로 퇴장하는 경로가 존재함. - 기타 경로: '200_종교'로의 직접적인 서비스 이용 행태와 '열람석3'과 '휴게4'간의 상호 이동 모습 등이 확인됨.

3. 앙상블 기법 기반 통합 분석 모델 생성

앞서 대량의 데이터에 대한 차원 축소 과정에서 N-gram, 용어 가중치 등 다양한 기법을 적용하여 결과를 비교하여 해석하였다. 모델이 다양해짐에 따라 해석의 폭이 풍부해지는 장점이 있는 반면, 사용할 모델이 많아질수록 분석가가 선택해야 할 해석 결과의 수가 기하급수적으로 증가하는 문제가 발생한다. 최적의 파라미터를 도출하여 단일 최종 모델을 제시하는 기존의 방법과 달리, 본 연구는 다양한 알고리즘에 기반한 다수의 분석 모델을 앙상블 기법으로 통합하고, 이를 기반으로 분석가에게 최적의 시뮬레이션 결과를 자동으로 추천하는 통합 모델 생성 방안을 제안한다. 이를 통해, 각 메트릭(metric) 모델의 장점을 결합함으로써, 단일 메트릭 모델을 사용할 때보다 안정적이고 강건한 결과를 얻을 수 있다. 본 연구에서 제안하는 앙상블 모델은 코사인 유사도(cosine similarity)와 JSD(Jensen-Shannon divergence)의 결합 모델이다. 토픽 유사도를 측정하기 위해 코사인, 자카드(Jaccard) 유사도를 비롯한 다양한 유사도 측정 모델이 활용되어 왔다. 특히, 코사인 유사도는 문헌 벡터의 크기를 반영하지 못한다는 단점이 존재하나, 이미 정규화된 단어의 확률 분포로 토픽을 구성하는 LDA 기법에서는 적합한 지표로 활용될 수 있다. 또한, 두 토픽의 확률 분포 차이를 측정하는 방법인 JSD는 대칭적인 특성으로 인해 KLD(Kullback-Leibler divergence)에 비해 안정적인 성능을 제공한다. 이러한 장점으로 인해 많은 연구에서 LDA 분석 및 평가에 활용되고 있다(Jeong & Joo, 2019; Kim & Oh, 2011; Liu & Hu, 2021). 코사인 유사도는 벡터 프로세싱을 통해 0과 1 사이의 값을 반환하며, 0은 불일치, 1은 완전 일치를 의미한다. 반면, JSD는 두 확률 분포(P, Q)의 차이를 계산하므로 0이면 완전히 일치하며, 값이 커질수록 불일치성이 증가함을 의미한다. 두 항을 결합한 앙상블 모델을 생성하기 위해, 코사인 유사도와 동일한 값의 범위를 갖도록 JSD 측정 결과를 0~1 범위로 정규화하고, 유사도 척도로 변환하는 과정이 필요하다. 〈공식 1〉은 자연로그 기반으로 계산된 JSD의 결과에 대해 로그 밑을 2로 변환하여 정규화하는 과정을 나타낸다. $M(x)$는 두 분포의 평균이며, JSD는 P와 Q가 그 평균 분포로부터의 얼마나 멀어져 있는지를 측정하는 값이다.

$$Normalized\_JSD(P \parallel Q) = \frac{1}{2}\sum_{x \in X}[P(x)\log_2(\frac{P(x)}{M(x)}) + Q(x)\log_2(\frac{Q(x)}{M(x)})]$$
$$M(x) = \frac{P(x)+Q(x)}{2}$$

〈공식 1〉 앙상블 모델 생성을 위해 정규화된 JSD

최종 생성된 앙상블 모델의 공식은 〈공식 2〉와 같다. 모델을 구성하는 두 척도는 0~1 범위의 동일한 스케일에서 작동하며, 가중치 부여를 통해 미세 조정이 가능하다. 여기서 A, B는 정규화된 확률 분포를 갖는 토픽 벡터이며, 코사인 유사도 $COS(A,B)$는 0과 1 사이의 값을 갖는다. 정규화된 지표인 $1-Normalized_JSD(A,B)$를 통해 코사인 유사도와 동일하게 0~1 범위로 정규화하였다. 각 항에 가중치 α와 β를 부여하며, 두 가중치의 합은 1이다. 본 연구에서는 기본 모델의 가중치로서 0.5를 부여하였다.

$$EnsembleModel(A,B) = \alpha \bullet COS(A,B) + \beta \bullet (1 - Normalized\_JSD(A,B))$$

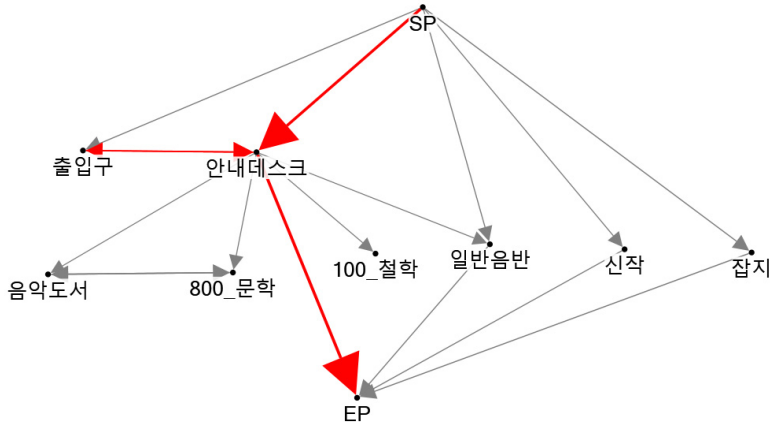
〈공식 2〉 통합 모델 생성을 위한 앙상블 모델

〈그림 7〉은 프로그래밍으로 자동 처리된 유사도 측정 결과를 스프레드시트 형식으로 정리한 것이다. 실험에 사용한 4개의 개별 모델 간의 조합(combination)을 통해 총 6(A~F)번의 두 모델 간 유사도 측정을 수행해야 하므로, 코사인과 JSD 각각에 대해 총 12(=6x2)회를 계산한다. 이때 각 모델은 5개의 토픽을 포함하므로, 총 625(=5x5x5x5)개의 경우의 수만큼 코사인 유사도와 JSD를 반복 계산하게 된다. 본 연구에서는 이와 같은 과정을 통해 총 7,500(=625x6x2)회의 계산을 수행하여 앙상블 모델(〈공식 2〉)에 따라 최종 계산 결과를 도출한다. 실험 결과, 시스템에 의해 자동 추천된 최종 통합 모델은 '*M1_T2(모델1의 토픽2) + M2_T5 + M3_T1 + M4_T3*'로 구성된 앙상블 결과이다. 입력 모델이 늘어날수록 계산 소요 시간은 지수적으로 증가한다.

M1	M2	M3	M4	A_jsd	A_cos	B_jsd	B_cos	C_jsd	C_cos	D_jsd	D_cos	E_jsd	E_cos	F_jsd	F_cos	avg_jsd	avg_cos	Ensemble_Score	Rank
2	**5**	**1**	**3**	**0.557**	**0.803**	**0.493**	**0.772**	**0.800**	**0.877**	**0.566**	**0.797**	**0.498**	**0.701**	**0.476**	**0.676**	**0.5650**	**0.7711**	**0.66804**	**1**
5	1	2	2	0.493	0.689	0.473	0.538	0.549	0.639	0.518	0.591	0.805	0.860	0.675	0.694	0.5855	0.6684	**0.62693**	**2**
2	5	1	5	0.557	0.803	0.493	0.772	0.630	0.741	0.566	0.797	0.454	0.507	0.364	0.507	0.5108	0.6879	**0.59938**	**3**
1	2	4	3	0.623	0.694	0.642	0.717	0.500	0.542	0.536	0.646	0.649	0.685	0.479	0.471	0.5714	0.6258	**0.59863**	**4**
2	2	1	3	0.584	0.572	0.493	0.772	0.800	0.877	0.283	0.299	0.649	0.685	0.476	0.676	0.5475	0.6469	**0.59722**	**5**
2	2	4	3	0.584	0.572	0.305	0.251	0.800	0.877	0.536	0.646	0.649	0.685	0.479	0.471	0.5588	0.5839	**0.57131**	**6**
4	4	4	4	0.767	0.743	0.424	0.516	0.736	0.770	0.448	0.527	0.527	0.551	0.337	0.359	0.5396	0.5778	**0.55868**	**7**
4	4	1	4	0.767	0.743	0.425	0.311	0.736	0.770	0.557	0.486	0.527	0.551	0.335	0.338	0.5577	0.5331	**0.54541**	**8**
2	4	1	3	0.223	0.302	0.493	0.772	0.800	0.877	0.557	0.486	0.381	0.482	0.476	0.676	0.4882	0.5991	**0.54366**	**9**
3	5	1	3	0.625	0.641	0.462	0.426	0.290	0.309	0.566	0.797	0.498	0.701	0.476	0.676	0.4863	0.5919	**0.53908**	**10**

〈그림 7〉 앙상블 모델 생성을 위한 메트릭 데이터 시트 (상위 10개만 표시)

〈그림 8〉은 최종 통합 모델에서 제시하는 이용자의 주요 이동 패턴을 시각화한 것이다. 외부에서 '안내데스크'로 직행한 후 바로 퇴장하는 패턴과 출입구의 공지를 확인한 뒤 '안내데스크'를 거쳐 다양한 서비스로 접근하는 패턴이 모두 관찰되며, 이는 '안내데스크'가 중요한 거점(허브) 역할을 수행하고 있음을 시사한다. 앞서 제시한 대부분의 세부 모델에서도 출입구의 정보 효용성이 매우 높은 것으로 나타났으며, 통합 모델에서도 이를 가장 중요한 경로 패턴으로 강조하고 있다. 통합 모델이 제시하는 세부 경로는 다음과 같다. 첫째, '음악도서'와 '800_문학' 코너 간의 상호 이동은 이용자의 유사 관심 분야 및 인접성을 고려한 공간 배치의 결과로 해석된다. 둘째, '신작', '잡지' 코너는 출입 직후 직접 방문되는 경우가 많으며, 이는 이용자가 명확한 목적을 갖고 방문하는 적극적 이용 패턴으로 볼 수 있다. 반면 '음악도서'와 '800_문학', '100_철학' 등은 '안내데스크'의 참고 서비스를 통해 해당 코너를 이용하는 모습을 보이고 있다. '일반음악' 코너의 경우, 직접 방문하는 행태와 안내를 통해 이용하는 행태를 모두 보이는 특징을 추가로 확인할 수 있다.



〈그림 8〉 자동 추천을 통한 최종 통합 모델 (4개의 LDA 모델의 앙상블)

V. 결 론

본 연구는 공공도서관에서 수집된 실시간 센서 기반 데이터를 활용하여, 이용자의 이동 경로 및 공간 이용 행태를 정량적으로 분석하고, 이를 통해 사용자 중심의 공간 운영 및 서비스 개선에 기여할 수 있는 데이터 기반 분석 프레임워크를 제안하였다. 기존의 빈도 기반 통계나 단순 클러스터링 기법으로는 포착하기 어려운 미세한 이동 패턴과 행태 변화 탐지를 위해, LDA 토픽 모델링 기법을 적용하여 잠재적 이동 경로의 구조적 패턴을 효과적으로 추출하였다.

특히, 토픽 모델이 단어 간 순서나 방향성을 반영하지 못하는 한계를 보완하기 위해 Topical N-gram 기반의 새로운 모델을 설계하였고, 이를 통해 연속적인 이동 흐름과 시계열적 변화를 정밀하게 분석할 수 있음을 확인하였다. 이러한 접근은 단순한 이동 경로 분석을 넘어, 공공도서관 및 유사 공공시설의 공간 재설계, 맞춤형 서비스 개발 등 실무적 활용 가능성을 높이며, 데이터 기반 의사결정 환경의 고도화에 기여할 수 있다는 가능성을 제시하였다.

본 연구의 주요 의의는 다음과 같다. 첫째, 센서를 통해 대량으로 수집된 이동 패턴 데이터를 효과적으로 처리할 수 있는 전처리 전략과 함께, Topical N-gram 기법을 활용하여 반복적 이동 경로를 정밀하게 분석할 수 있는 새로운 분석 모델을 제시하였다. 둘째, 파라미터 기반으로 결정된 단일 모델이 아닌 다양한 분석 모델을 앙상블 기법으로 통합하고, 이를 기반으로 분석가에게 최적의 시뮬레이션 결과를 자동으로 추천하는 통합 분석 시스템 프레임워크를 구축하였다.

향후 연구에서는 본 연구의 방법론을 다양한 공공 서비스 영역에 적용함으로써, 공간적 특성과 이용자 행태의 차이를 정밀하게 규명하여 제안한 통합 프레임워크의 우수성을 검증해 보고자 한다. 또한, 이동 패턴 데이터에 인구통계학적 정보를 결합함으로써, 성별・연령・방문 목적 등 다양한 요소에 따른 행태 특성을 자동으로 분류하고 예측할 수 있는 고도화된 분석 기법으로 발전시킬 예정이다.

참 고 문 헌

김규환, 정도헌 (2023). AI 카메라를 활용한 공공도서관 이용자의 공간이용행태 분석 연구. 한국문헌정보학회지, 57(4), 333-351. https://doi.org/10.4275/KSLIS.2023.57.4.333

박성재 (2019). 스마트폰 무선신호를 이용한 공공도서관 이용자의 공간이용행태 분석. 정보관리학회지, 36(1), 295-313. https://doi.org/10.3743/KOSIM.2019.36.1.295

박주연, 정도헌 (2022). 텍스트마이닝 기법을 활용한 교육관점에서의 메타버스 관련 이슈 탐색 - 뉴스 빅데이터를 중심으로. 산업융합연구, 20(6), 27-35. https://doi.org/10.22678/JIC.2022.20.6.027

조아, 이경희, 조완섭 (2015). LDA 기법을 이용한 버스 승객의 잠재적 이동패턴 분석. 한국데이터정보과학회지, 26(5), 1061-1069. https://doi.org/10.7465/jkdi.2015.26.5.1061

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on machine learning (ICML), 113-120. https://doi.org/10.1145/1143844.1143859

Blei, D. M. & Lafferty, J. D. (2009). Topic Models. In A. Srivastava, & M. Sahami (Eds.),

Text Mining: Classification, Clustering and Applications, 71-93. Cambridge: Chapman and Hall/CRC. https://doi.org/10.1201/9781420059458

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022. https://dl.acm.org/doi/10.5555/944919.944937

Chu, D., Sheets, D., Zhao, Y., Wu, Y., Yang, J., Zheng, M., & Chen, G. (2014). Visualizing hidden themes of taxi movement with semantic transformation. 2014 IEEE Pacific Visualization Symposium, Yokohama, Japan, 137-144. https://doi.org/10.1109/PacificVis.2014.50

Jeong, D. H., & Joo, H. S. (2019). Topical prescriptive analytics system for automatic recommendation of convergence technology. Biotechnology and Bioprocess Engineering, 24(6), 893-906. https://doi.org/10.1007/s12257-019-0305-1

Jeong, D. H. & Song, M. (2014). Time gap analysis by the topic model-based temporal technique. Journal of Informetrics, 8(3), 776-790. http://dx.doi.org/10.1016/j.joi.2014.07.005

Kim, D. & Oh, A. (2011). Topic chains for understanding a news corpus. Computational Linguistics and Intelligent Text Processing - 12th International Conference, 20-26, https://doi.org/10.1007/978-3-642-19437-5_13

Lin, X., Li, D., & Wu, X. (2010). A Joint Topical N-Gram Language Model Based on LDA. In Proceedings of the 2nd International Workshop on Intelligent Systems and Applications, Wuhan, China, 1-4. https://doi.org/10.1109/IWISA.2010.5473439

Liu, C. & Hu, R. (2021). Hot topic discovery across social networks based on improved LDA Model. KSII Transactions on Internet and Information Systems, 15(11), 3935-3949, 2021. https://doi.org/10.3837/tiis.2021.11.004

Liu, D. Y. & Hsu, K. S. (2018). A study on user behavior analysis of integrate Beacon technology into library information services. Eurasia Journal of Mathematics, Science and Technology Education, 14(5), 1987-1997. https://doi.org/10.29333/ejmste/85865

Mohamed, K., Côme, E., Baro, J., & Oukhellou, L. (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data. Urban Computing Workshop ACM, NewYork.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. Transactions of the Association for Computational

Linguistics, 3, 299-313. https://doi.org/10.1162/tacl_a_00140

Qu, M. (2024) Exploring patron behavior in an academic library: a Wi-Fi-connection data analysis. Education and Information Technologies, 29, 11235-11256. https://doi.org/10.1007/s10639-023-12248-9

Sievert, C. & Shirley, K. (2014). LDAvis: a method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63-70. https://doi.org/10.3115/v1/W14-3110

Spacey, R., Muir, A., Cooke, L., Creaser, C., & Spezi, V. (2015). Filtering wireless (Wi-Fi) internet access in public places. Journal of Librarianship and Information Science, 49(1), 15-25. https://doi.org/10.1177/0961000615590693

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd International Conference on Machine Learning (ICML), 977-984. https://doi.org/10.1145/1143844.1143967

Wang, X., McCallum, A., & Wei, X. (2007). Topical N-Grams: phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), 697-702. IEEE. https://doi.org/10.1109/ICDM.2007.86

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

Cho, Ah, Lee, Kyung Hee, & Cho, Wan Sup (2015). Latent mobility pattern analysis of bus passengers with LDA. Journal of the Korean Data and Information Science Society, 26(5), 1061-1069. https://doi.org/10.7465/jkdi.2015.26.5.1061

Kim, Gyuhwan & Jeong, Do-Heon (2023) Analysis of space use patterns of public library users through AI cameras. Journal of the Korean Society for Library and Information Science, 57(4), 333-351. https://doi.org/10.4275/KSLIS.2023.57.4.333

Park, Ju-Yeon & Jeong, Do-Heon (2022). Exploring issues related to the metaverse from the educational perspective using text mining techniques: focusing on news big data. Journal of Industrial Convergence, 20(6), 27-35. https://doi.org/10.22678/JIC.2022.20.6.027

Park, Sung Jae (2019). Analyzing library space use patterns in a public library through

Smartphone WiFi. Journal of the Korean Society for Information Management, 57(4), 333-351. https://doi.org/10.4275/KSLIS.2023.57.4.333