25-2 Data Science Lab Modeling Project - 추천시스템 팀 보고서

13기 이채원, 곽도윤, 송채은 | 14기 구기현, 조재우

1. 서론 (Introduction)

최근 몇 년간 건강과 웰빙에 대한 사회적 관심이 높아지면서, 러닝은 가장 대중적인 야외 운동으로 자리 잡았다. 단순한 유행을 넘어 러닝은 체력 증진, 다이어트, 스트레스 해소 등 다양한 목적을 가진 생활 속 활동으로 정착하였다. 이와 더불어 사용자들은 단순히 일정 거리를 완주하는 것을 넘어, 자신의 체력 수준과 러닝 목적, 그리고 선호하는 환경적 조건(예를 들어 경사도, 안전성, 경관, 편의시설 등)에 최적화된 맞춤형 러닝 코스를 찾는 데에 관심을 가지게 되었다.

그러나 기존의 경로 탐색 서비스는 대부분 최단 거리 또는 최소 시간을 기준으로 설계되어 있어, 러닝과 같이 운동 목적을 가진 활동의 특성을 충분히 반영하지 못하는 한계가 존재하였다. 특히 러닝 코스는 단순한 이동 경로가 아니라, 경사도, 야간 안전도, 자연경관, 횡단보도 유무, 루프 형태의 완결성 등 복합적인 요인을 종합적으로 고려해야 한다는 점에서 차별화된 접근이 필요하였다.

2. 파이프라인 (Pipeline)

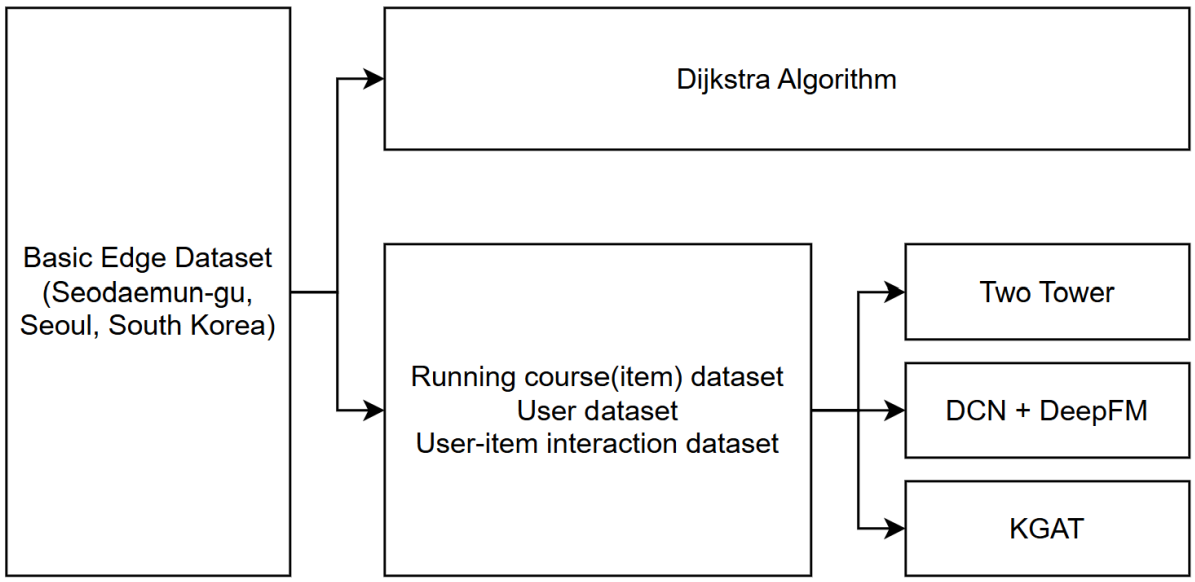


Figure1. 프로젝트 파이프라인

이에 본 프로젝트는 러닝 코스 제공의 두 가지 접근법을 병렬적으로 탐구하였다. 첫 번

째는 사용자의 즉각적 요구사항을 반영한 **규칙 기반 러닝코스 생성**(Dijkstra Algorithm)이며, 두 번째는 과거 선호 패턴을 학습한 **데이터 기반 러닝코스 추천**(Two Tower, DCN+DeepFM, KGAT)이다. 러닝코스 생성 트랙에서는 경사도, 안전성, 편의시설을 고려한 맞춤형 비용 함수를 구현하였고, 추천 시스템 트랙에서는 500개 노드 기반 38,000개 러닝코스와 125명의 온라인 설문 데이터로 상호작용 데이터셋을 구축하여 세 개의 딥러닝 모델을 Warm-start와 Cold-start 시나리오에 따라 학습 및 평가하였다. 본 연구는 데이터 구축, 경로 생성, 추천 모델링 전 과정을 직접 구현함으로써 러닝 도메인에 특화된 실질적 개인화 추천의 가능성을 검증하였다.

2-1. 데이터셋 (Dataset)

본 프로젝트의 데이터셋 구축은 기초 도로 데이터 구축부터 최종 상호작용 데이터셋 생성까지 단계적으로 진행되었다.

(1) 기초 도로 데이터셋 구축

서울시 서대문구의 보행자 네트워크를 기반으로 러닝에 필요한 종합적인 도로 정보를 구축하였다. OSMnx를 통해 추출한 도로 네트워크 데이터에 서울시 공공데이터(CCTV 위치, 가로등 현황, 등고선 데이터)와 OSM 내부 태깅 데이터(편의점, 화장실, 조경, 지하철 위치정보)를 결합하였다. 도로별 평균 경사도, CCTV 밀도, 조도, 편의시설 유무 등의 물리적 데이터를 표준화하여 기초 도로 데이터베이스(Mega DB_processed.csv)를 구축하였다.

(2) 러닝 코스 데이터셋(아이템 데이터셋) 생성

서대문구 전역을 최대한 균등하게 커버하는 500개의 주요 노드를 선정하고, 각 노드를 출발점으로 하는 다양한 러닝 코스를 생성하였다. 일반적으로 선호되는 러닝 환경(보행로 선호, 골목길 회피, 직진 선호, 루프 형태)을 고려하여 약 38,000개의 러닝 코스 후보군을 자동 생성하였으며, 이를 generated_routes_final.csv로 정리하였다.

(3) 코스 특징 계산

생성된 각 코스에 대해 Dijkstra 알고리즘에서 사용한 비용 함수 로직을 활용하여 특징을 부여하였다. 각 경로별 난이도(경사도 기반), 안전도(CCTV 및 가로등 밀도), 자연경관 비율, 편의시설 접근성 등의 종합 점수를 계산하고 정규화하여 df_route_capped_normalized.csv를 생성하였다.

(4) 사용자 데이터셋 구축

잠재적 사용자들의 러닝 관련 선호도를 수집하기 위해 온라인 설문조사를 실시하였다. 리서치 매칭 플랫폼인 Surveasy를 통해 2025년 9월 22일부터 24일까지 2일간 진행되었

으며, 총 125명이 참여하였다. 설문 문항은 러닝 목적(체력 증진, 다이어트, 마라톤 준비, 스트레스 해소 등), 러닝 시간대, 거리 선호도, 경사도 선호, 안전 인식 등 다양한 개인적 요인을 포함하였다. 수집된 데이터는 사용자 피처(user features)로 변환되어 추천 모델의 입력 변수로 활용되었다.

(5) 사용자-코스 상호작용 데이터셋 구축

사용자의 선호 피처(난이도, 안전성, 경치, 편의성)와 각 코스의 대응 피처 간의 유사도 거리를 계산하여 상호작용 점수를 산출하였다. 유클리드 거리를 기반으로 유사할수록 높은 점수를, 상이할수록 낮은 점수를 부여하였다. 특히 달리고 싶은 거리 차이가 큰 경우(±1km 이상)에는 후순위로 간주하도록 패널티를 적용하였으며, 횡단보도 개수나 야간 조도 등 불편 요소가 많은 코스는 자동으로 가중 감점되도록 설계하였다. 이를 통해 user_preferred_route.csv를 생성하여 사용자-코스 간의 선호도를 정량적으로 표현하였다.

user_id	rank_1	rank_2	rank_3	rank_4	rank_5	rank_6	rank_7	rank_8	rank_9	rank_10
0	5038	33738	5611	8085	3397	10533	10535	37149	7683	29250
1	32382	65	11951	13797	1417	1413	9717	4333	11946	32858
2	17647	7697	7729	22462	1088	1077	32288	32303	21064	34679
3	5032	13305	5034	5028	5029	5033	32807	28696	23100	33724
4	35798	35717	35713	16768	5079	6890	2490	16759	24016	24015
5	6597	16433	16436	16430	204	6598	6595	36175	36170	16319

Figure2. user_preferred_route.csv 일부 (사용자별 1위부터 38,000위까지의 route id를 표기하였다)

(6) Warm-start와 Cold-start 데이터 구성

추천 시스템 학습 및 평가를 위해 구축된 데이터셋을 Warm-start와 Cold-start 시나리오로 분할하였다.

Warm-start는 기존 사용자의 활동 기록이 축적된 상태를 가정하였다. 개별 사용자의 상위 선호 코스 1,000개 중 800개를 학습용, 200개를 테스트용으로 분할하였다. 학습 데이터는 긍정 샘플과 부정 샘플을 1:4 비율로, 테스트 데이터는 현실적 평가를 위해 1:50 비율로 구성하였다. 이러한 구성을 통해 모델이 충분히 다양한 코스와의 관계를 학습하고, 사용자가 아직 경험하지 않은 코스에 대한 선호를 얼마나 정확히 예측하는지를 평가할 수 있도록 하였다.

Cold-start는 신규 사용자를 대상으로 기본 피처만으로 추천하는 시나리오를 가정하였다. 전체 사용자 중 100명을 학습용, 25명을 테스트용으로 분리하여, 테스트 그룹은 학습 단계에 전혀 포함되지 않도록 하였다. 이를 통해 모델의 일반화 성능을 검증하였다.

2-2. 모델 (Models)

본 프로젝트는 러닝 코스 추천의 특수성을 반영하기 위해 경로 생성용 최적화 알고리즘(Dijkstra)과 세 가지 딥러닝 추천 모델(Two-Tower, DCN+DeepFM, KGAT)을 구현하였다.

(1) Dijkstra Algorithm

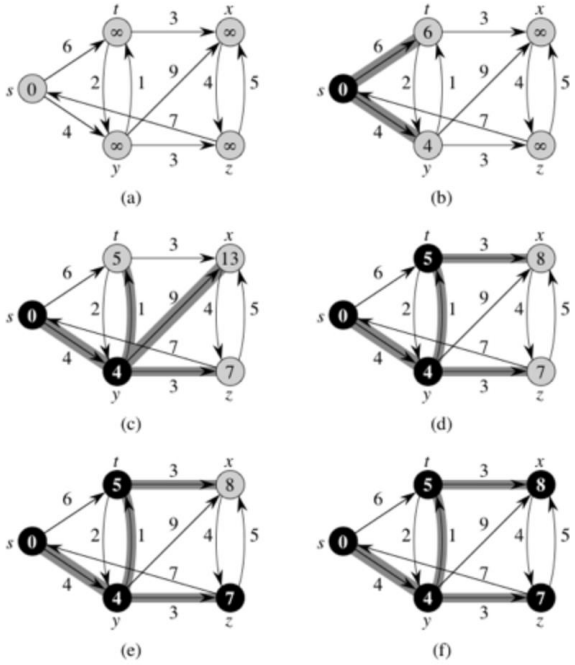


Figure3. 다익스트라 알고리즘 예시

Dijkstra 알고리즘은 그래프 이론에서 단일 출발점으로부터 각 노드까지의 최단 경로를 탐색하는 대표적인 최적화 기법이다. 본 연구에서는 이를 러닝 코스 생성에 적용하되, 단순히 물리적 거리를 최소화하는 것이 아니라 사용자의 러닝 선호도를 반영한 맞춤형 비용 함수(customized cost function)를 정의하여 최적 경로를 탐색하였다.

비용 함수는 경사도, 야간 안전성, 횡단보도, 자연경관 네 가지 요소를 가중하여 구성하였다. 경사도는 사용자의 목표 난이도 수준에 따라 동적으로 조정되었으며, 야간 모드에서는 CCTV 및 가로등 밀도가 낮은 구간에 높은 페널티를 부여하였다. 횡단보도는 존재 여부에 따라 제거하거나 추가 비용을 부여하는 두 가지 방식 중 선택할 수 있도록 하였고, 자연경관 요소는 공원 또는 수변 구간의 비율을 고려하여 비용을 할인하였다. 이후 Min–Max 스케일링을 통해 각 요소를 정규화하여, 모델이 다양한 사용자 설정에 유연하게 반응할 수 있도록 구성하였다.

또한 단순 왕복이 아닌 루프 형태의 경로를 생성하기 위해 앵커 노드를 기준으로 경로의 전반부와 후반부를 분리하고, 이미 탐색된 경로 주변에 추가 비용을 부여하는 Route Poisoning 기법을 적용하였다. 이를 통해 동일 구간의 중복 탐색을 방지하고, 시작점과

종료점이 동일한 현실적인 러닝 루프 경로를 생성할 수 있었다.

(2) Two-Tower 모델

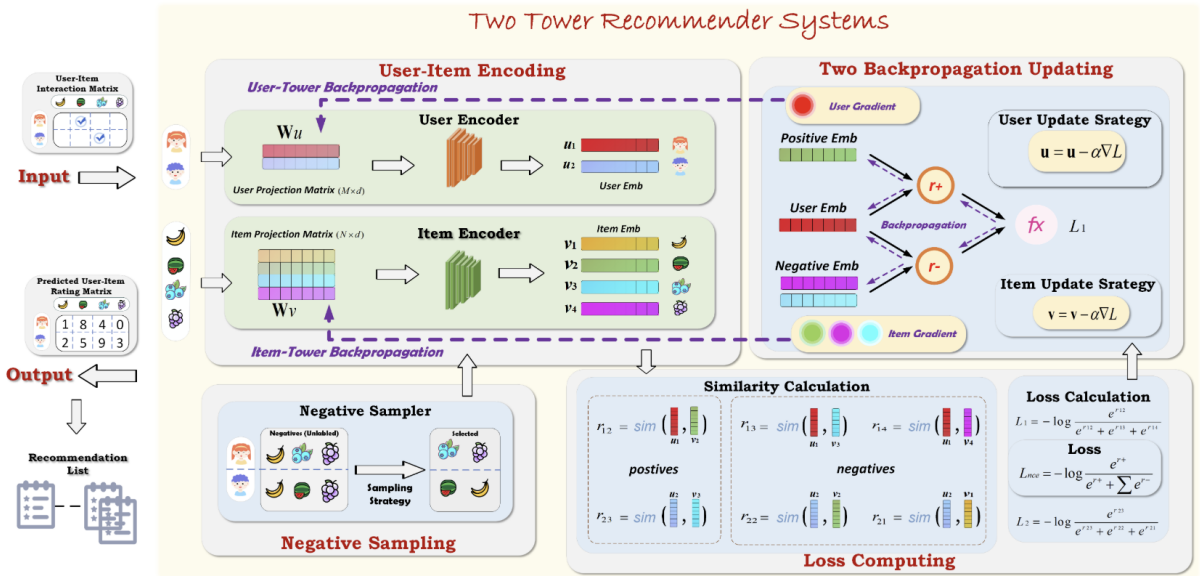


Figure4. Two-Tower의 전체 아키텍처

Two-Tower 모델은 사용자 타워(User Tower)와 아이템 타워(Item Tower)로 구성된 이중 구조의 딥러닝 추천 모델이다. 각 타워는 입력된 피처를 독립적으로 인코딩하여 동일한 차원의 임베딩 벡터를 생성하고, 이 두 벡터 간의 내적(dot product)을 통해 사용자와 코스의 선호 유사도를 계산한다. 이 구조는 대규모 추천 시스템에서 효율성과 확장성을 동시에 확보할 수 있다는 장점이 있다.

본 프로젝트에서의 사용자 피처는 러닝 레벨, 선호 경치 유형, 운동 시간대, 건강 목적 등 개인적 특성을 포함하였다.

아이템 피처는 코스의 거리, 경사도, 안전등급, 루프형 여부 등으로 구성되었다. 각 타워 내부에서는 입력 피처별 선형 변환 후 ReLU 활성화 함수를 적용하고, 드롭아웃을 포함한 다층 퍼셉트론(MLP)을 통해 64차원의 임베딩을 생성하였다.

학습과정에서는 InfoNCE 손실 함수를 적용하여 긍정 쌍(사용자가 실제 선호하는 코스)과 부정 쌍(비선호 코스) 간의 상대적 거리를 조정하였다.

모델은 Cold-start 학습으로 사전 훈련된 가중치를 불러와 Warm-start 환경에서 미세 조정(Fine-tuning)을 수행하였다. 이 접근을 통해 학습 속도를 단축하고 예측 정밀도를 향상시킬 수 있었다.

(3) DCN+DeepFM 모델

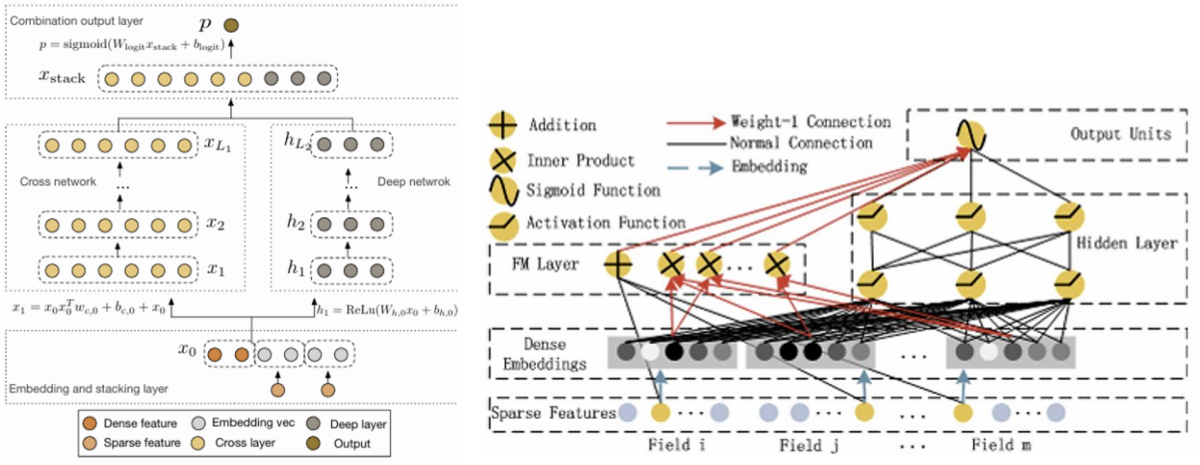


Figure4. DCN 아키텍처 일부와 DeepFM 아키텍처 일부

DCN+DeepFM 모델은 저차원 및 고차원 피처 상호작용을 동시에 학습하기 위해 설계된 하이브리드 구조이다.

DeepFM 모듈은 Factorization Machine(FM) 기반으로 2차 상호작용을 명시적으로 학습하며, DCN 모듈은 Cross Network를 통해 고차원 교차 항을 효율적으로 포착한다. 이 두 모듈의 출력을 Soft Ensemble 방식으로 결합하여, 비선형성과 일반화 성능을 동시에 강화하였다.

본 프로젝트에서 사용된 입력 피처는 사용자 측면에서는 러닝 빈도, 러닝 시간대, 경사 및 거리 선호도가 포함되었고, 아이템 측면에서는 코스의 총 거리, 평균 경사도, 안전등급, 루프형 여부 등이 포함되었다.

학습 구조상 DeepFM 모듈은 속성 간의 비선형 조합을, DCN 모듈은 피처 교차 항을 통해 저차·고차 상호작용을 명시적으로 계산하였다. 두 모듈의 출력은 각각 0.6 대 0.4의 비율로 결합되었으며, 순위 최적화를 위해 BPR Loss를 손실 함수로 사용하였다.

Warm-start 환경에서는 사용자 ID 임베딩과 속성 피처를 함께 활용하였고, Cold-start에서는 ID 임베딩을 제외하고 속성 피처만으로 학습을 진행하였다.

(4) KGAT 모델

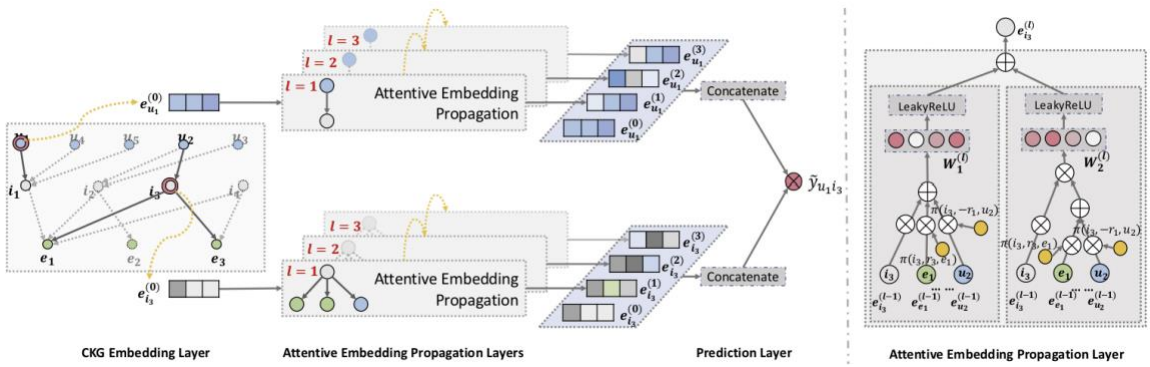


Figure5. KGAT의 전체 아키텍처

KGAT(Knowledge Graph Attention Network)는 사용자와 아이템 간의 직접적인 상호작용(user–item pair)만으로는 설명하기 어려운 잠재적 의미 관계를 보완하기 위해 지식그래프(Knowledge Graph)를 추천 모델에 통합한 구조이다.

기존의 협업 필터링 모델이 데이터 희소성(sparsity) 문제와 Cold-start 한계를 겪는 데 비해, KGAT는 아이템이 속한 지식그래프 내의 다양한 이웃 엔티티(예를 들어 장소, 경사, 안전도, 경관 등)를 활용하여 아이템의 표현을 더욱 풍부하게 학습할 수 있다.

본 프로젝트에서 구축된 지식그래프는 하나의 러닝 코스(Route)가 여러 속성(feature)과 특성(property)을 동시에 가지는 구조로 설계되었다. 모든 관계는 (head, relation, tail) 형태의 트리플(Triple)로 구성되었으며, 예를 들어 "Route_1 – has_feature:Park – Yes", "Route_1 – has_feature:Subway – Yes", "Route_1 – has_feature:Store – Yes", "Route_1 – has_property:Difficulty – Medium"과 같이 연결된다. 이러한 구조를 통해 모델은 각 코스가 어떤 환경적·물리적 요소를 포함하고 있는지를 명시적으로 학습할 수 있다.

즉, 단일 경로가 여러 이웃 노드(feature)와 연결되어 있으며, KGAT의 어텐션(attention) 메커니즘은 이들 중 사용자에게 더 중요한 속성에 동적으로 가중치를 부여하도록 설계되었다.

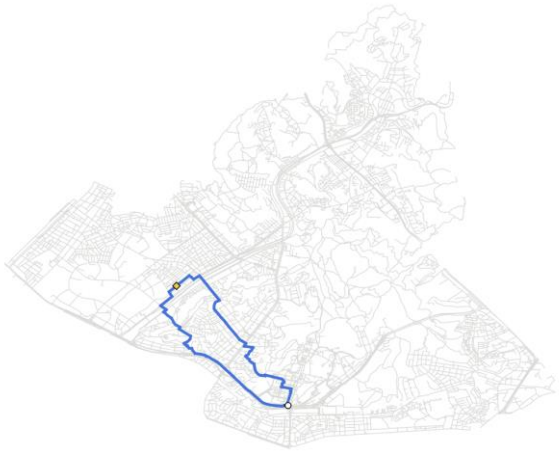
모든 노드는 연결된 이웃 엔티티의 중요도를 학습하고, 그 중요도에 따라 정보를 가중합하는 방식으로 임베딩을 업데이트하였다. 이를 통해 모델은 단순한 평균화보다 더 정교한 의미 기반 표현을 구성할 수 있었다.

또한 Hard Negative Sampling을 적용하여 모델이 구분하기 어려운 유사 코스 간의 미세한 차이를 학습하도록 유도하였다. 이 접근은 특히 Cold-start 환경에서 모델의 일반화 능력을 향상시키는 데 효과적이었다.

3. 결과 (Results)

3-1. 경로 생성 알고리즘 성능

Dijkstra 알고리즘의 맞춤형 비용 함수 적용 결과를 다양한 사용자 시나리오에서 검증하였다. 다음은 두 명의 사용자 시나리오를 바탕으로 생성된 경로를 시각화한 것이다.



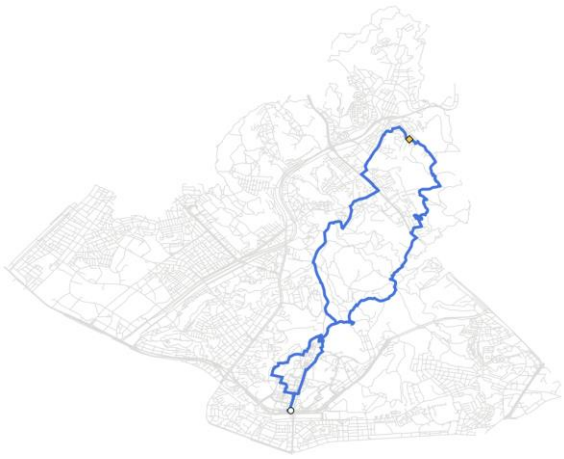


Figure6(좌). 연세대학교 정문 출발, 5km, 야간모드, 난이도 "최하"

Figure7(우). 연세대학교 정문 출발, 10km, 난이도 "중", 화장실 포함

3-2. 추천 시스템 성능평가

딥러닝 모델 성능 평가는 추천 시스템의 표준 지표인 Hit Ratio (HR@K)를 사용하여 수행하였다. HR@K는 실제 사용자가 선호한 코스가 추천된 상위 K개의 목록 내에 포함될 확률을 의미한다. Warm-start 환경에서는 개인화 정밀도(precision)를, Cold-start 환경에서는 일반화 성능(generalization)을 평가하는 데에 초점을 두었다.

	HR@1	HR@10
Two Tower	0.52	0.68
DCN + DeepFM	0.79	1.00
KGAT	0.32	0.85

Figure8. Warm-start 환경에서의 성능평가

	HR@1	HR@10
Two Tower	0.32	0.60
DCN + DeepFM	0.92	1.00
KGAT	0.56	0.84

Figure9. Cold-start 환경에서의 성능평가

Warm-start 환경에서 Two-Tower 모델은 HR@1이 0.52, HR@10이 0.68로 나타났다. KGAT 모델은 HR@1이 0.32, HR@10이 0.85로 비교적 높은 재현율을 보였다. 한편 DCN+DeepFM 모델은 HR@1이 0.79, HR@10이 1.00으로 계산되었으나, 이 결과는 지나치게 완벽하여 모델의 학습 과정에 문제가 있었을 가능성이 높았다.

Cold-start 환경에서도 유사한 경향이 관찰되었다. Two-Tower 모델은 HR@1이 0.32,

HR@10이 0.60으로 기본적인 일반화 성능을 보였으며, KGAT 모델은 HR@1이 0.56, HR@10이 0.84로 높은 성능을 유지하였다. 반면 DCN+DeepFM은 HR@1이 0.92, HR@10이 1.00으로 과도한 수치를 기록하여 실질적인 분석에서는 신뢰할 수 없는 결과로 판단되었다.

결과적으로 DCN+DeepFM은 모델링 과적합 사례로 평가되었고, 본 프로젝트의 핵심 분석은 Two-Tower와 KGAT 두 모델을 중심으로 진행되었다. Two-Tower 모델은 구조적 단순성과 실시간 처리 효율성 면에서 높은 활용 가능성을 보였으며, KGAT 모델은 속성 간 의미 기반 학습을 통해 Cold-start 상황에서의 일반화 성능을 입증하였다.

4. 결론 (Conclusion)

본 프로젝트는 러닝 코스 제공을 위한 두 가지 상호보완적 접근법을 병렬적으로 탐구하고 그 성능을 검증하였다.

규칙 기반 경로 생성 측면에서, Dijkstra 알고리즘에 맞춤형 비용 함수를 적용한 결과 사용자의 명시적 요구사항(거리, 난이도, 안전성, 편의시설)을 반영하는 경로 생성이 가능함을 확인하였다. 야간 안전 모드나 특정 편의시설 경유와 같은 제약조건을 충족하는 루프 형태의 경로를 성공적으로 생성하였으며, 이는 실제 러닝 시나리오에 적용 가능한 실용적인 결과물이었다.

데이터 기반 경로 추천 측면에서는, 38,000개 코스와 125명의 사용자 데이터로 구축한 상호작용 데이터셋을 바탕으로 세 가지 딥러닝 모델을 평가하였다. Two-Tower 모델은 Warm-start에서 HR@10 0.68의 안정적인 성능과 구조적 단순성을 보였으며, KGAT 모델은 지식그래프 기반 학습을 통해 Cold-start 상황에서도 HR@10 0.84의 우수한 일반화 성능을 달성하였다. DCN+DeepFM의 비현실적 결과(HR@10 1.00)는 과적합 가능성을 시사하여 추가 검증이 필요한 것으로 판단되었다.

이 두 접근법은 각각 명확한 강점을 지닌다. Dijkstra 기반 경로 생성은 신규 사용자 즉시 대응, 특수 요구사항 보장, 설명 가능한 결과 제공이라는 장점을 가지며, 딥러닝 기반 추천은 암묵적 선호 패턴 학습, 과거 데이터 기반 개인화, 다양한 코스 풀에서의 추천이라는 강점을 보였다.

향후 연구에서는 이 두 방법론을 통합한 하이브리드 시스템 개발이 유망할 것으로 전망된다. 초기 사용자나 특수한 제약조건이 있는 경우 Dijkstra 알고리즘으로 즉각적인 경로를 제공하고, 사용 이력이 축적되면 딥러닝 모델 기반 개인화 추천으로 전환하는 적응형 시스템을 구현한다면, 러닝 코스 추천의 실용성과 정확도를 동시에 극대화할 수 있을 것이다. 특히 Two-Tower의 효율성과 KGAT의 일반화 능력을 결합한 앙상블 접근법은

Warm-start와 Cold-start 모두에서 균형 잡힌 성능을 제공할 것으로 기대된다.

4-2. 한계점 및 향후 과제 (Limitations and Future Work)

본 프로젝트는 러닝 코스 추천이라는 새로운 연구 영역을 개척하였으나, 다음과 같은 한계점이 존재한다.

첫째, 데이터 규모와 실제성의 한계이다. 125명의 설문 데이터는 통계적 일반화에 제한적이며, 무엇보다 실제 러닝 행동 데이터가 아닌 유사도 기반 시뮬레이션으로 상호작용을 생성하였다는 점에서 현실과의 괴리가 존재할 수 있다.

둘째, 모델 검증의 한계이다. DCN+DeepFM의 비현실적인 성능(HR@10 1.00)은 샘플링 또는 손실 함수 구현 과정의 오류를 시사하며, 이는 모델 구현 및 검증 과정에서 보다 엄격한 검토가 필요함을 보여준다.

셋째, 외부 요인 반영의 부족이다. KGAT의 지식그래프가 코스 내부 속성만을 포함하여, 날씨, 미세먼지, 지역 행사 등 실시간 외부 환경 요인을 고려하지 못했다.

향후 연구에서는 Strava, Nike Run Club 등 러닝 애플리케이션과 연동하여 실제 사용자 궤적 데이터를 확보하고, 시간대별 선호 변화와 계절적 요인을 반영한 Sequential Recommendation 모델을 도입해야 한다. 또한 HR 외에 Recall, NDCG, 다양성 지표를 통한 다각적 성능 평가가 필요하며, 궁극적으로는 Dijkstra의 즉각적 경로 생성과 딥러닝의 개인화 추천을 결합한 하이브리드 시스템 구현이 목표가 되어야 할 것이다.

본 연구는 데이터와 모델의 한계에도 불구하고, 규칙 기반과 데이터 기반 접근법의 상호 보완적 가능성을 실증적으로 확인하였다는 점에서 의의를 가진다.