Smart Media Journal / Vol.15, No.6 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.6.82

추론 가속화를 위한 부가 정보 기반 조건부 플로우 매칭 콜드스타트 아이템 추천 모델

(Accelerating Inference in Cold-Start Item Recommendation Models via Side Information-Based Conditional Flow Matching)

김건한*, 김대로**, 천세진***, 한정규***
(Geonhan Kim, Daro Kim, Sejin Chun, Jungkyu Han)

요 약

추천 시스템에서 신규 아이템은 사용자 상호작용 데이터 부재로 인해 협업 필터링 적용이 어려운 콜드스타트 문제를 야기한다. 최근 이를 해결하기 위해 아이템 부가 정보를 활용하는 생성형 모델, 특히 확산 모델이 주목받고 있으나, 반복적인 노이즈 제거 과정으로 인한 느린 추론 속도가 실시간 서비스 적용의 구조적 한계로 지적된다. 이에 본 논문은 확산 모델의 연산 병목을 해소하고 추론 효율성을 개선한 부가 정보 기반 조건부 플로우 매칭 추천 모델을 제안한다. 제안 모델은 아이템 콘텐츠 정보를 조건으로 목표 사용자 상호작용 벡터를 직접 예측하는 궤적을 학습함으로써, 이력 데이터가 없는 신규 아이템에 대해서도 유의미한 추천을 수행한다. 추론 스텝 조절을 통해 확산 모델 대비 약 10배 이상의 추론 속도 향상을 달성하였다. 이는 추천 품질과 연산 효율성 간의 실용적 트레이드오프를 입증한 것으로, 제안 모델이 대규모 실시간 콜드스타트 추천 환경에 보다 적합하고 확장 가능한 방법론임을 확인하였다.

■ 중심어 : 추천 시스템 ; 콜드스타트 ; 추론 ; 플로우 매칭 ; 부가 정보

Abstract

In recommender systems, new items cause the Cold-Start problem due to the lack of user interaction data, making collaborative filtering difficult to apply. While generative models leveraging side information, particularly Diffusion Models, have gained attention as a solution, their iterative denoising process remains a structural bottleneck that limits real-time service deployment. To address this, we propose a side information-based Conditional Flow Matching recommendation model that resolves the computational bottleneck of Diffusion Models and improves inference efficiency. The proposed model learns trajectories that directly predict target user interaction vectors conditioned on item content, enabling meaningful recommendations even for new items without interaction history. By adjusting inference steps, the model achieves at least 10× faster inference compared to the Diffusion Model. This demonstrates a practical trade-off between recommendation quality and computational efficiency, confirming that the proposed model is a more suitable and scalable methodology for large-scale, real-time Cold-Start recommendation environments.

■ keywords : Recommender System ; Cold-Start; Inference ; Flow Matching ; Side Information

Ⅰ. 서 론

OTT, 전자상거래, 뉴스 플랫폼 등 현대 디지털 서비스에서는 매일 방대한 양의 신규 콘텐츠가 등록되며, 이를 적시에 사용자에게 노출하는 것이 서비스 경쟁력의 핵심 요소로 부상하였다.

특히 신규 콘텐츠 공개 시점에 다수 사용자를 대상으로 추천 후보를 산출해야 하므로, 콜드 아이템에

* 준회원, 동아대학교 컴퓨터공학과
** 준회원, 동아대학교 컴퓨터공학과
*** 정회원, 컴퓨터AI공학부
*** 종신회원, 컴퓨터AI공학부
이 논문은 2023년도부터 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2023-0-00076, SW중심대학(동아대학교))

접수일자 : 2026년 05월 06일
수정일자 : 2026년 05월 26일
게재확정일 : 2026년 06월 01일
교신저자 :한정규, 천세진 e-mail : jkhan@dau.ac.kr, sjchun@dau.ac.kr

대한 추천 정확도뿐 아니라 대규모 실시간 추론 효율성이 동시에 요구된다. 이러한 콘텐츠 노출의 효과를 극대화하는 개인화 추천을 위해 협업 필터링이 가장 보편적인 기법으로 자리 잡았으며[1, 2], 최근에는 지식그래프 확장[14], 대화형 추천[15], 지식증류 기반 경량화[16] 등 다양한 딥러닝 기반 추천 기법이 연구되고 있다. 하지만 이러한 연구는 충분한 상호작용 데이터 없이는 효과적으로 작동하기 어렵다는 한계를 지닌다. 특히 신규 아이템은 상호작용 이력이 전무하여 콜드스타트(Cold-Start) 문제가 발생하며, 이는 콘텐츠 생성 주기가 빠른 현대 플랫폼 환경에서 서비스 품질을 저해하는 심각한 요인으로 지적된다[3]. 기존 연구들은 아이템의 부가 정보(Side Information)를 활용하는 확산 모델(Diffusion Model) 기반 접근을 시도하였으나[4, 5], 수백 단계의 반복적 노이즈 제거 과정으로 인해 실시간 서비스 적용에 구조적 한계가 있다. 한편 플로우 매칭 기반 추천 연구는 추론 효율성을 개선하였으나[7], 상호작용 이력이 전무한 콜드 아이템에는 적용이 어렵다. 즉, 추론 효율성과 콜드스타트 아이템 추천을 동시에 해결하는 방법론은 아직 충분히 연구되지 않은 상황이다. 이에 본 논문은 플로우 매칭의 빠른 추론 특성을 유지하면서, 아이템 부가 정보를 조건으로 활용하는 아이템 기반 구조로 재설계한 조건부 플로우 매칭(CFM: Conditional Flow Matching) 콜드스타트 추천 모델을 제안한다. 제안 모델은 베르누이 플로우(Bernoulli Flow) 설계와 추론 스텝 조절 전략을 통해 목표 분포로 향하는 최적의 추론 궤적을 학습함으로써, 두 가지 문제를 동시에 해결한다. 본 논문에서는 제안 모델과 기존 확산 모델(DDPM[10]) 간의 추론 속도 및 추천 성능을 비교 분석한다.

II. 관련 연구

1. 확산 모델을 이용한 콜드 아이템 추천

Han과 Chun[5]은 부가 정보를 조건으로 가우시안 노이즈로부터 콜드 아이템의 협업 필터링(Collaborative Filtering, CF) 임베딩을 직접 생성하는 확산 모델 기반 2단계 프레임워크를 제안하여 우수한 콜드스타트 추천 성능을 달성하였다. 그러나 수백 단계의 반복적 노이즈 제거 과정(T=200~400 스텝)으로 인해, 확산 모델의 확률 기반 생성 방식의 장점을 이용하기 위한 대량의 콜드 아이템 CF 임베딩의 실시간 생성 시, 추론에 수 초 이상 소요되나, 추론 효율화 방안은 제시되지 않았다. 본 연구는 플로우 매칭을 도입하여 소수의 스텝만으로 목표 분포에 수렴함으로써 이러한 추론 병목을 직접적으로 해결하고자 한다.

2. 플로우 매칭을 이용한 추천 시스템 연구

Liu 등[7]은 최적 전송 기반 벡터장 회귀와 행동 유도 사전 분포를 적용한 플로우 매칭 기반 협업 필터링 프레임워크 FlowCF를 제안하여 확산 모델 대비 추론 속도를 크게 단축하면서도 높은 추천 정확도를 달성하였다. 그러나 사용자의 전체 상호작용 행렬을 입력으로 하는 사용자 기반(User-based) 구조로 설계되어 있어, 상호작용 이력이 전무한 콜드 아이템 추천에 플로우 매칭을 활용하기 어렵다. 반면 본 연구는 아이템 부가 정보를 조건으로 활용하는 아이템 기반(Item-based) 접근으로 재설계하여, 추론 효율성과 콜드 아이템 추천을 동시에 해결한다.

III. 본 론

1. 문제 정의

본 연구에서 사용자 집합 $U=\{u_1,u_2,u_3,....,u_{|U|}\}$와 아이템 집합 $I=\{i_1,i_2,.....,i_{|I|}\}$은 다음과 같이 정의된다. 사용자와 아이템 간의 암시적 피드백(Implicit Feedback)은 이진 상호작용 행렬로 $y_{u,i}=\begin{cases}1, u가\ i와\ 상호작용한\ 경우\\0, 그렇지\ 않은\ 경우\end{cases}$ (단, $Y\in\{0,1\}^{|U|\times|I|}$) 나타내며, 협업 필터링 관점에서 개별 아이템 x_i를 식 (1)과 같은 상호작용 행렬의 i번째 열벡터로 표현할 수 있다.

$$x_i=[y_{1,i},y_{2,i},...,y_{|U|,i}]^T\in\{0,1\}^{|U|} \quad (1)$$

일반적인 추천 시스템은 상호작용이 풍부한 기존 아이템에 대해 x_i를 직접 활용하여 아이템의 상호작용 특성을 학습한다. 그러나 콜드 아이템 i_{cold}는 관측된 사용자 상호작용이 전무하여 $x_{cold}=[0]^{|U|}$가 되므로, 기존의 CF 임베딩 방식을 적용할 수 없다. 이에 본 연구는 아이템 등록 시점에 확보할 수 있는 부가 정보 임베딩 $c_i \in \mathbb{R}^d$로, 조건부 생성 모델 $p_\theta(x|c_i)$를 최적화한다. 여기서 c_i는 제목, 장르, 줄거리 등 아이템의 부가 정보를 임베딩한 d차원의 벡터이다. 용이한 설명을 위해 기존 아이템의 부가 정보는 c_i^{warm}, 콜드 아이템의 부가 정보는 c_i^{cold}로 표기한다. 조건부 생성 모델 $p_\theta(x|c_i)$의 최적화는 식 (2)와 같이, 부가 정보 c_{cold}를 생성 가이드로 삼아 초기 상태인 사전 노이즈 x_0로부터 생성함수 $f(\theta)$를 찾는 문제로 귀결된다. 나아가 본 연구는 추론 시 사용되는 스텝 수 S를 최소화하면서도 추천 성능을 유지하는 효율적 생성 궤적을 학습하는 것을 추가 목표로 한다.

$$\hat{x}_{cold}=f_\theta(x_0,c_{cold}) \tag{2}$$

2. 조건부 플로우 매칭

플로우 매칭은 상미분 방정식(ODE: Ordinary Differential Equation)을 기반으로 단순한 사전 분포를 복잡한 데이터 분포로 변환하는 생성 모델 프레임워크로[6], 초기 노이즈에서 실제 데이터로 향하는 확률 경로와 그 방향을 지시하는 벡터장을 학습한다. 확산 모델이 수백 스텝의 곡선형 경로를 따르는 것과 달리, 플로우 매칭은 직선 경로를 학습하므로 적은 스텝으로 동일한 목표 분포에 도달할 수 있다.

가. 확률 경로와 손실 함수

플로우 매칭은 생성하여야 할 목표 데이터 x_1이 주어졌을 때, 사전 노이즈 x_0에서 x_1으로 향하는 조건부 확률 경로를 정의한다. 특히 최적 전송 이론[6]을 결합해 시간 $t\in[0,1]$에서의 중간 상태 x_t를 가장 직관적인 선형 보간으로 식 (3)과 같이 계산한다.

$$x_t=(1-t)x_0+tx_1 \tag{3}$$

이러한 직선 경로를 형성하는 이상적인 목표 벡터장 $u_t(x_t|x_1)$는 데이터의 변화량인 x_1-x_0로 도출된다. 신경망 모델 $v_\theta(x_t,t)$는 주어진 시간 t와 중간 상태 x_t에서 이상적인 목표 벡터장 u_t의 방향을 근사하도록 파라미터 θ를 학습하며, 손실 함수는 식 (4)와 같이 두 벡터장 간의 MSE를 최소화하는

$$L_{CFM}(\theta)=E_{t,x_0,x_1}\left[\,\|\,v_\theta(x_t,t)-u_t(x_t|x_1)\,\|^2\,\right] \tag{4}$$

형태로 정의된다.

나. 추론

학습이 완료된 후, 새로운 데이터를 생성하는 추론(Inference) 과정은 상미분 방정식(ODE)을 푸는 수치적분 과정으로 이루어진다. 시간 $t=0$ 시점에 사전 분포 p_0로부터 무작위 노이즈 x_0를 추출한 뒤, 학습된 모델 v_θ가 지시하는 벡터장을 따라 $t=1$까지 궤적을 추적하면 식 (5)와 같이 최종 데이터 x_1이 생성된다.

$$x_1=x_0+\int_0^1 v_\theta(x_t,t)dt \tag{5}$$

실제 컴퓨팅 환경에서는 연속적 계산이 불가능하므로, 시간 t를 이산적인 스텝 Δt로 나누고, 식 (6)과 같이 오일러 방법(Euler method)으로 상태를 점진적으로 업데이트한다[6, 11].

$$x_{t+\Delta t}=x_t+v_\theta(x_t,t)\Delta t \tag{6}$$

3. 조건부 플로우 매칭 기반의 콜드 아이템 추천

본 연구는 신규 등록된 아이템의 상호작용 부재 문제를 해결하기 위해, 아이템의 부가 정보를 조건으로 활용하여 목표 상호작용을 직접 예측하는 조

건부 플로우 매칭 기반(CFM)의 콜드 아이템 추천 방법을 제안한다. 이때 암시적 피드백의 이진 특성을 보존하기 위해 선행 연구[7]을 따라 사전 분포를 베르누이로 채택하며, 이에 따라 식 (3)의 선형 보간을 이진 도메인에서 동치로 구현하는 마스킹 결합 방식을 함께 도입한다.

가. 이산적 상호작용을 위한 베르누이 사전 분포(Bernoulli Prior Distribution)

기존 확산 모델이 연속적인 가우시안 노이즈를 사용하는 것과 달리, 추천 시스템의 암시적 피드백 데이터는 사용자의 클릭 여부를 나타내는 이산적 이진값 $\{0,1\}$이다[7]. 상호작용 내역이 전무한 콜드 아이템의 공간을 만들기 위해 본 방법은 FlowCF[7]의 베르누이 사전 분포 설계를 콜드 아이템 추천 환경에 적용하여, 데이터의 이산적 특성을 보존하는 베르누이(Bernoulli) 분포를 초기 상태 x_0의 사전 분포로 채택한다. 배치(Batch) 내 콜드 아이템에 부여되는 초기 노이즈 x_0는 식 (7)과 같이 균등 확률을 반영한 확률 벡터 p를 기반으로 샘플링 된다.

$$x_0 \sim Bernoulli(p) \tag{7}$$

이를 통해 콜드 아이템은 연속적 노이즈가 아닌, 실제 사용자 상호작용과 동일한 이진(Binary) 도메인에서 시작한다.

나. 부가 정보 기반 조건부 모델 학습

학습 단계에서는 상호작용이 존재하는 기존 아이템들의 부가 정보 임베딩 c_i^{warm}와 실제 상호작용 벡터 x_1을 활용하여, 향후 콜드 아이템에 적용할 조건부 생성 모델 f_θ를 최적화한다. 연속적인 시간 $t \in [0,1]$는 N개의 이산적인 스텝으로 분할되며, 각 배치마다 균등 분포에서 정수 $k \sim \{1, N-1\}$을 추출하여 훈련 시간 t를 식 (8)과 같이 정의한다.

$$t = \frac{k}{N}, \quad k \in \{1,2,\dots,N-1\} \tag{8}$$

결정된 시간 t에서의 중간 상태 x_t는 t에 비례하는 확률로 활성화되는 이진 마스크 M_t를 생성하여 목표 데이터 x_1과 초기 노이즈 x_0를 식 (9)와 같이 확률적으로 결합(Masking)함으로써 구성된다.

$$\begin{aligned} M_t &\sim Bernoulli(t) \\ x_t &= M_t \odot x_1 + (1 - M_t) \odot x_0 \end{aligned} \tag{9}$$

이를 통해 x_t의 모든 원소는 이진값을 유지하면서도, 기댓값은 식 (10)과 같이 연속적 플로우 매칭의 이상적인 직선 경로와 일치한다.

$$E[x_t] = t x_1 + (1-t) x_0 \tag{10}$$

조건부 생성 모델 f_θ는 현재 상태 x_t, 부가 정보 c_i^{warm}, 시간 임베딩(Timestep Embedding) t를 입력으로 받아 x_1을 직접 예측하며, 베이스라인 DDPM과 동일한 목적 함수 형태로 통일하여 생성 방식 차이에 따른 순수 효과를 비교하기 위해 벡터장 회귀 대신 직접 예측 방식을 채택하였다. 손실 함수는 식 (11)과 같이 모델의 예측값과 실제값 간의 MSE로 정의된다.

$$L(\theta) = E_{k, x_0, x_1, c}\left[\| x_1 - f_\theta(x_t, c_i^{warm}, t) \| \right. \tag{11}$$

식 (4)와 동일하게 MSE 기반 손실 함수를 적용하되, 여기서는 부가 정보 c_i^{warm}를 조건으로 추가한 형태이다.

다. 콜드 아이템 상호작용 추론

상호작용이 없는 콜드 아이템이 등록되면, 모델은 오직 해당 아이템의 부가 정보 c_i^{cold}만을 조건으로 입력받아 잠재적 상호작용 $\hat{x}_{cold}$를 추론한다. 베르누이 사전 분포에서 샘플링된 초기 상태 x_0에서 시작하여, 최적의 추론 스텝 수 $S\,(1 \le S \le N)$를 설정하고, $\Delta t = 1/S$ 단위로 t를 전진시키며 상태를 업데이트한다. S가 클수록 x_0에서 x_1까지의 궤적을 세밀하게 추적하며, S가 작을수록 넓은 간격으로 건너뛰어 추론 속도가 빨라진다. 각 스텝에

서 모델은 부가 정보 c_i^{cold}를 조건으로 목표 추정치 $\hat{x}_1$을 예측하고, 이를 활용해 현재 시점의 벡터장 v_t를 식 (12)와 같이 산출한다.

$$v_t = \frac{\hat{x}_1 - x_t}{1 - t + \epsilon} \tag{12}$$

산출된 벡터장 v_t에 시간 변화량 Δt를 곱하여 다음 상태인 $x_{t+\Delta t}$를 계산하는 오일러 적분(Euler Integration)을 반복 수행하는 식 (13)으로 나타낼 수 있다.

$$x_{t+\Delta t} = x_t + v_t \times \Delta t \tag{13}$$

식 (12) 분모의 ϵ은 시간 t가 1에 수렴할 때 분모 $(1-t+\epsilon)$이 0이 되어 발산하는 것을 방지하는 수치 안정성 상수이다. 이상의 제안 모델에 대한 성능 검증을 위해, 다음 장에서 실제 데이터셋을 활용한 실험을 수행한다.

Ⅳ. 실험 결과

1. 실험 데이터셋

표 1. 데이터셋 통계

데이터셋	분할	아이템	상호작용	콜드 아이템
ML1M	Train	2,788	667,161	0
	Valid	209	47,007	209
	Test	489	119,710	489
CiteULike	Train	13,584	164,210	0
	Valid	1,018	13,037	1,018
	Test	2,378	27,739	2,378

표 2. 데이터셋 부가 정보 통계

데이터셋	사용자	아이템	희소성	부가 정보 차원 (d)
ML1M	6,039	3,486	96.04%	400
CiteULike	5,551	16,980	99.78%	300

제안 모델의 콜드스타트 추천 성능과 추론 속도를 검증하기 위해, 표 1, 2의 MovieLens1M(ML1M)과 CiteULike 데이터셋을 활용하였다. ML1M은 영화 추천 분야의 대표적인 벤치마크 데이터셋으로, 신규 영화를 개봉 초기에 관심 있는 사용자에게 신속히 노출하는 것이 흥행과 사용자 만족도에 직결되는 만큼, 콜드스타트 추천 연구에서 널리 활용되어 왔다[5, 7, 8]. CiteULike는 학술 논문 추천 분야의 벤치마크 데이터셋으로, 최신 논문을 출판 직후 관련 연구자에게 신속히 제시하는 것이 연구 생산성 향상에 기여하는 도메인 특성상 콜드스타트 문제가 빈번히 발생하며, 관련 연구에서 대표적으로 활용되어 왔다[5, 9]. 이에 본 연구에서는 두 데이터셋을 실험 데이터셋으로 선정하여 부가 정보 기반 콜드스타트 추천 성능을 평가한다. 검증 및 테스트 셋의 모든 아이템은 훈련 셋에 등장하지 않은 순수 콜드 아이템으로, 상호작용 이력 없이 부가 정보만을 활용한 추천 성능을 평가한다. 각 데이터셋의 부가 정보 임베딩은 ML1M의 경우 MovieLens Tag Genome[12]에서 제공하는 태그 관련성 점수 벡터를 400차원으로, CiteULike의 경우[9] 논문 제목 및 초록에 대한 TF-IDF 벡터를 300차원으로 구성하였다.

2. 베이스라인 모델 설정

제안 모델과의 추론 및 추천 성능 비교를 위해 DDPM을 베이스라인으로 설정하였다[10]. 원본 DDPM은 추천 시스템에 직접 적용하기 어려우므로, 원본 상호작용 벡터 x_1을 직접 예측하는 MSE 목적 함수와 역확산 매 스텝마다 부가 정보 임베딩 c_i^{cold}를 조건으로 주입하는 구조로 수정하였다. 이를 통해 두 모델은 동일한 네트워크 구조, 손실 함수, 조건 주입 방식을 공유하며, 생성 궤적의 추론 방식과 초기 분포(CFM: 베르누이, DDPM: 가우시안)에서 차이를 보인다. 따라서 두 모델 간 성능 차이는 네트워크 용량이나 학습 조건이 아닌, 순수하게 생성 방법의 차이에서 기인한다.

3. 평가 지표

추천 시스템의 Top-K 랭킹 성능을 평가하기 위해, *Recall@K* (*R@K*)와 *NDCG@K* (*N@K*)를 활

용하였으며, 본 실험에서는 $k \in [10, 20]$으로 설정하였다. $R@K$은 상위 K개 중 정답 아이템 내 포함 비율을, $N@K$은 정답 아이템의 순위가 높을수록 높은 점수를 부여하는 지표이다[13].

4. 실험 설정

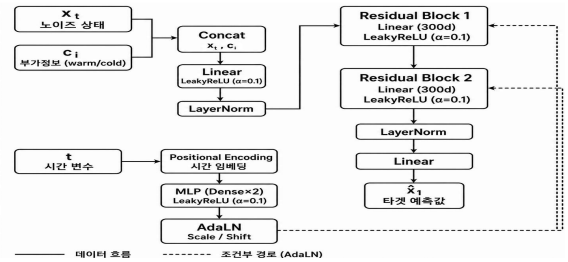

그림 1. 모델 네트워크 구조

그림 1은 모델의 네트워크 구조이다. 입력으로 노이즈 상태 x_t와 부가 정보 c_i(학습 시 c_i^{warm}, 추론 시 c_i^{cold})를 결합하고, 시간 변수 t는 위치 인코딩을 통해 AdaLN의 Scale 및 Shift 파라미터로 변환한다. 300차원 잔차 블록 2개를 거쳐 목표 상호작용 벡터 $\hat{x}_1$을 예측한다. 추론 시 예측값을 $[0,1]$로 제한하는 클리핑을 두 모델에 동일하게 적용하였으며, 확률적 샘플링에 따른 분산을 줄이기 위해 5회 반복 평균을 산출하였다. 하이퍼파라미터는 학습률 $\{0.001, 0.0005\}$, 임베딩 차원 $\{32, 64\}$, 드롭아웃 $\{0.0, 0.1\}$을 대상으로 그리드 서치를 수행하였으며, 검증 셋의 $Recall@20$을 기준으로 최적값을 선정하였다. 그 결과 ML1M에서는 CFM이 학습률 0.0005, 임베딩 차원 32로, DDPM이 학습률 0.001, 임베딩 차원 32로 선정되었으며, CiteULike에서는 CFM이 학습률 0.001, 임베딩 차원 64로, DDPM이 학습률 0.001, 임베딩 차원 32로 선정되었다. 드롭아웃은 모든 설정에서 0.1이 최적이었다. 추론 시간 측정 전 1회 워밍업(warm-up)을 선행하였다. 모든 실험은 AMD Ryzen 7 9700X, 128GB RAM, RTX A6000 (48GB VRAM) 환경에서 TensorFlow 2.15.0 및 CUDA 12.2 기반으로 수행되었다.

5. 평가 결과

본 실험은 세 단계로 구성된다. 가절에서는 파레토 프론티어 분석을 통해 각 모델의 대표 성능 기준인 훈련 시간 스텝 수(N)와 추론 스텝 수(S)를 도출한다. 나절에서는 도출된 값을 바탕으로 모델을 재학습한 뒤, CFM의 추론 스텝 수를 단계적으로 줄여가며 DDPM 대비 속도-성능 트레이드오프를 분석한다. 다절에서는 두 모델의 최종 추천 성능을 비교한다.

가. 최적 하이퍼파라미터 탐색

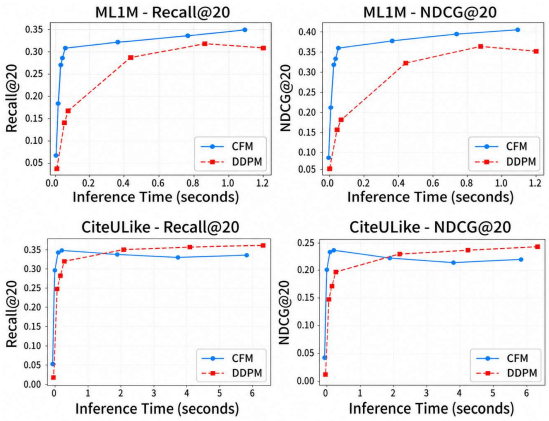

그림 2. 추론 시간에 따른 추천 성능 비교 그래프

그림 2는 각 데이터셋에서 모델의 추론 시간 대비 추천 성능($R@20$, $N@20$)의 상관관계를 나타내는 파레토 프론티어이다. 훈련 시간 스텝 수N을 $\{1,2,5,6,7,10,100,200,300\}$으로 변화시키며, 동일한 N 조건에서 추론 스텝 $S=N$으로 고정하여 평가하였다. 분석 결과, ML1M에서 CFM은 $N=10$ 이후 약 3~4%의 완만한 성능 향상만 보여 초기 구간 대비 성능 향상 폭이 둔화된다. CiteULike에서 CFM은$N=10$ 이후 스텝을 늘릴수록 최대 5.04% 성능이 하락한다. 반면 DDPM은 ML1M에서 $N=200$까지 성능이 증가하다가 $N=300$에서 2.70% 하락하며, CiteULike에서 $N=200$ 이후 $N=300$까지 0.92%의 미미한 성능 향상만 보인다. 훈련 스텝 N은 식(3)의 선형 보간 경로를 얼마나 세밀하게 학습하는지를 결정한다. CFM은 선형 경로를 학습하므로 적은 N으로도 목표 분포에 수렴하며, N이 일정 수준을 초과하면 추가적인 성능 향상이 둔화된다. 반면 DDPM은 곡선 경로를 따르므로 N이

부족하면 각 스텝의 근사 오차가 누적되어 성능이 저하된다. 이에 본 실험에서는 성능 수렴 이후 큰 향상이 없거나 하락하는 시점을 기준으로 CFM $N=S=10$, DDPM $N=S=200$을 각 모델의 대표 설정값으로 선정한다.

나. 속도-성능 트레이드오프 분석

표 3. CFM 추론 스텝에 따른 성능-속도 트레이드오프 (DDPM $N=S=200$ 기준)

데이터셋	추론 스텝	R@20	성능 유지율	추론 시간 (초)	속도 향상
ML1M	7	0.2305	71.31%	0.15	21.76x
	8	0.2566	79.39%	0.16	19.98x
	9	0.2895	89.57%	0.17	18.48x
	10	0.3356	103.83%	0.29	10.95x
CiteULike	7	0.2309	62.65%	0.61	24.17x
	8	0.2371	64.33%	0.67	21.91x
	9	0.2522	68.43%	0.73	20.20x
	10	0.3524	95.62%	0.77	19.09x

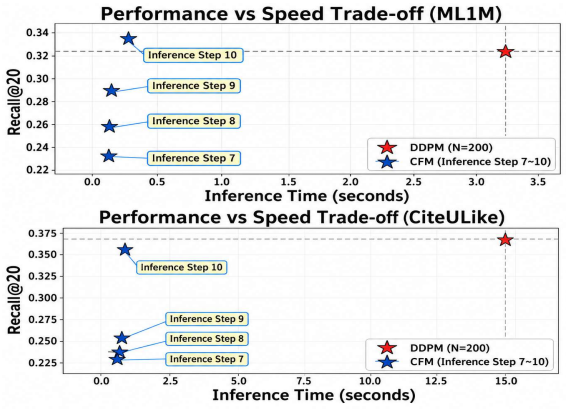

그림 3. CFM 추론 스텝에 따른 속도 비교 그래프

표 3과 그림 3은 CFM $N=10$ 조건에서 추론 스텝 S를 7~10으로 조절하며 측정한 성능-속도 트레이드오프를 보여준다. ML1M에서 CFM $S=10$은 DDPM $S=200$ 대비 $R@20$ 기준 103.83%의 성능을 달성하면서 약 10.95배 빠른 추론이 가능하다. 예를 들어, ML1M의 경우 추천 품질이 중요한 환경에서는 S=10으로 설정하여 DDPM 수준의 성능을 유지하고, 응답 속도가 우선시되는 실시간 환경에서는 S=9로 설정하여 약 18배 이상의 추론 가속과 함께 89% 이상의 성능 유지율을 확보할 수 있다. 이처럼 추론 스텝 수 S는 서비스 요구사항에 따라 유연하게 조정 가능한 실용적 파라미터로 활용될 수 있다.

다. 모델 간 추천 성능 비교

표 4. CFM($N=S=10$)과 DDPM($N=S=200$) 추천 성능 비교

데이터셋	모델	R		N	
		@10	@20	@10	@20
ML1M	CFM	**0.2246**	**0.3356**	**0.3958**	**0.3952**
	DDPM	0.2135	0.3232	0.3656	0.3724
CiteULike	CFM	0.2390	0.3524	0.1969	0.2334
	DDPM	**0.2447**	**0.3686**	**0.1980**	**0.2384**

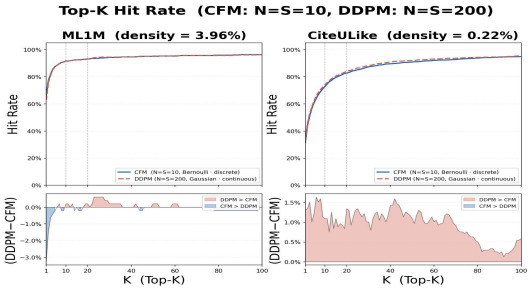

그림 4. CFM과 DDPM의 Top-K 적중률 비교 그래프

표 4는 CFM $N=S=10$, DDPM $N=S=200$ 조건에서 두 모델의 추천 성능을 비교한 결과이다. ML1M에서 CFM은 DDPM 대비 모든 지표에서 우세한 성능을 보이며, 평균적으로 약 5~8%의 성능 향상을 달성한다. 반면 CiteULike에서는 CFM이 DDPM 대비 전 지표에서 성능 감소를 보이며, 대부분의 지표에서 약 2~4% 낮은 성능을 기록한다. 그림 4는 K를 증가시키며 측정한 비교 결과이다. ML1M에서는 K=1~5 구간에서 CFM이 DDPM 대비 약 3%p 높은 적중률을 보이나 이후 두 모델은 유사한 성능을 나타낸다. CiteULike에서는 전 구간에서 DDPM이 CFM보다 약 0.5~1.5%p 높은 적중률을 유지한다. 이는 두 데이터셋의 희소성 차이와 각 모델의 생성 방식 특성에서 기인하며, 구체적으로 다음과 같이 분석된다. ML1M(밀도 3.96%)은 상호작용이 상대적으로 풍부하여 CFM의 베르누이 분포 기반 이진값 직접 예측 구조가 정답 위치 복원에 적합하게 작용한다. 반면 CiteULike(밀도 0.22%)는 극희소 환경으로, 전체 아이템 중 상호작

용이 존재하는 위치 자체가 희박하여 이진값으로 정답 위치를 직접 예측하는 CFM 구조가 불리하게 작용한다. 특히 CiteULike와 같이 양성 상호작용 비율이 매우 낮은 환경에서는 대부분의 차원이 0으로 구성되기 때문에, 이진값 복원 방식은 소수의 양성 위치를 안정적으로 구분하는 데 한계가 있을 수 있다. DDPM은 가우시안 분포 기반의 연속값을 출력하므로, 극희소 환경에서도 아이템별 점수가 연속적으로 분포하여 랭킹 구분이 가능하기 때문으로 분석된다. 이러한 한계는 향후 연구 과제로 남긴다.

Ⅴ. 결론 및 향후 연구

본 연구는 콜드스타트 아이템 추천 환경에서 확산 모델의 추론 병목을 해소하기 위해, 아이템 부가 정보를 조건으로 활용하는 조건부 플로우 매칭 모델을 제안하였다. 이를 통해 추천 품질과 연산 효율성 간의 실용적 트레이드오프가 가능함을 확인하였으며, 제안 모델이 대규모 실시간 콜드스타트 추천 환경에 적합함을 입증하였다. 다만 극희소 환경에서는 성능이 저하되는 한계가 있으며, 이를 극복하기 위해 향후 연구에서는 희소 상호작용 데이터를 연속적인 잠재 공간으로 임베딩하는 연속형 플로우 매칭으로 확장하여 극희소 환경에서의 성능 및 일반화 능력을 개선할 계획이다.

REFERENCES

[1] C. A. Gomez-Uribe and N. Hunt, The Netflix Recommender System: Algorithms, Business Value, and Innovation, *ACM Transactions on Management Information Systems (TMIS)*, 6, 4, 1-19, 2016.01.

[2] Y. Koren, S. Rendle, and R. Bell, Advances in Collaborative Filtering, *Recommender Systems Handbook*, 91-142, 2021.

[3] H. Chen, Z. Wang, F. Huang, X. Huang, Y. Xu, Y. Lin, P. He, and Z. Li, Generative Adversarial Framework for Cold-Start Item Recommendation, *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2565-2571, 2022.07.

[4] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, and T.-S. Chua, Diffusion Recommender Model, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 832-841, Taipei, Taiwan, 2023.07.

[5] J. Han and S. Chun, Diffusion Model as a Base for Cold Item Recommendation, *Applied Sciences*, 15, 9, 4784, 2025.04.

[6] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, Flow Matching for Generative Modeling, *International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.05.

[7] C. Liu, Y. Zhang, J. Wang, R. Ying, and J. Caverlee, Flow Matching for Collaborative Filtering, *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 1765-1775, Toronto, Canada, 2025.08.

[8] F. M. Harper and J. A. Konstan, The MovieLens Datasets: History and Context, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 4, 1-19, 2015.12.

[9] C. Wang and D. M. Blei, Collaborative topic modeling for recommending scientific articles, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 448-456, San Diego, CA, USA, 2011.08.

[10] J. Ho, A. Jain, and P. Abbeel, Denoising Diffusion Probabilistic Models, *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6840-6851, 2020.12.

[11] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, Neural Ordinary Differential Equations, *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 6571-6583, 2018.12.

[12] J. Vig, S. Sen, and J. Riedl, The Tag Genome: Encoding Community Knowledge to Support Novel Interaction, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2, 3, Article 13, 2012.09.

[13] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, Neural Collaborative Filtering, *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 173-182, Perth, Australia, 2017.04.

[14] 이승주, 안석호, 이의종, 서영덕, "이종 데이터 간 관계 모델링을 통한 개인화 추천 시스템의 지식 그래프 확장 기법," *스마트미디어저널*, 제12권, 제4호, 27-40쪽, 2023년 5월

[15] 김태호, 장형준, 김상욱, "다중목표 대화형 추천시스템을 위한 사전 학습된 언어모델," *스마트미디어저널*, 제12권, 제6호, 35-40쪽, 2023년 7월

[16] 배홍균, 김지연, 김상욱, "익스플리싯 피드백 환경에서 추천 시스템을 위한 최신 지식증류기법들에 대한 성능 및 정확도 평가," *스마트미디어저널*, 제12권, 제9호, 89-94쪽, 2023년 10월

저 자 소 개

김건한(준회원)

2022년 3월 ~ 현재 동아대학교 재학 중

<주관심분야 : 추천시스템, 기계학습>

김대로(준회원)

2024년 3월 동아대학교 학사 졸업
2026년 3월 동아대학교 석사 졸업

<주관심분야 : 추천시스템, 정보 검색>

천세진(정회원)

2018년 10월 ~ 2020년 8월
미국 표준기술연구원(NIST) 정보기술연구소(ITL) 근무
2021년 3월 ~ 현재
동아대학교 소프트웨어대학 조교수.

<주관심분야: 지식그래프 추론, 온톨로지 엔지니어링>

한정규(종신회원)

2007년 7월 ~ 2014년 4월
NTT Software Innovation Center 근무.
2018년 6월 ~ 2020년 8월
NAVER AiRS 근무.
2020년 9월 ~ 현재 동아대학교
컴퓨터AI공학부 조교수.

<주관심분야: 추천 시스템, 정보 검색, 데이터 마이닝>