추천 시스템의 성능 편차에 따른 사용자 분류 및 언러닝 강건성 분석

정유진, 이상철*
대구경북과학기술원 기초학부, *대구과학경북기술원 나노기술연구부

eugene_jeong@dgist.ac.kr, *sangchul.lee@dgist.ac.kr

User Segmentation by Performance Deviation in Recommender Systems and Robustness Analysis of Unlearning

Eugene Jeong, Sang-chul Lee*
School of Undergraduate of Studies, DGIST, *Division of Nanotechnology, DGIST

요 약

본 논문은 추천 시스템 언러닝 기법인 SISA 가 갖는 구조적 한계를 분석하고, 이에 대한 대안으로 성능 편차 기반 언러닝 분석 프레임워크를 제안한다. 제안하는 프레임워크는 전체 평균 지표에 은폐될 수 있는 국소적 성능 붕괴를 포착하여 언러닝 기법의 실질적 강건성을 판별하는 기준을 제시한다. 실험 결과, SISA 는 평균 40% 이상의 성능 하락을 보였으며 특히 학습이 까다로운 이상치 사용자 그룹에서 그 취약성이 극대화됨을 확인하였다. 반면, PGU 는 대다수의 사용자 구간에서 Retrain 와 대등한 성능을 유지하며 우수한 강건성을 입증하였다.

1. 서 론

개인정보 보호 규제의 강화로 언러닝 연구가 필수 과제로 대두되었으나 기존 연구들은 추천 시스템의 고유한 구조적 특성을 충분히 반영하지 못하고 있다. 특히 협업 필터링은 사용자-아이템 간 연결성 전파가 성능의 핵심이므로, 데이터 분할 방식인 Sharded, Isolated, Sliced, and Aggregated training(SISA)는 연결 경로를 단절하여 추천 시스템의 학습 원리와 상충할 위험이 있다[1].

또한 본 논문은 평균 정확도 지표가 특정 유저 그룹에서 발생하는 치명적인 성능 붕괴를 은폐할 수 있다는 점을 주목한다. 상호작용의 희소성 등으로 인해 학습이 까다로운 취약 사용자 그룹은 구조적 교란에 매우 민감하게 반응한다. 따라서 전체를 하나로 묶는 단순 평균 평가는 언러닝 기법의 실질적인 강건성을 과대평가하는 오류를 범할 수 있다.

이러한 한계를 극복하기 위해 본 논문은 성능 편차에 따른 사용자 세분화 분석을 제안하여 그룹별 강건성을 실증한다. 제안하는 기법은 파라미터 공간에서 그레이디언트 투영을 수행하는 Projected-Gradient Unlearning(PGU)이다[2]. 실험 결과 SISA 는 취약 그룹에서 50.38%까지 폭락하며 한계를 드러냈으나, PGU 는 대부분의 사용자 그룹에서 Retrain 과 대등한 성능을 유지했다. 이에 본 논문은 효율성과 강건성을 겸비한 현실적 대안으로 PGU 를 제안한다. 본 논문은 대규모 추천 시스템에서 효율성과 강건성을 동시에 달성할 수 있는 현실적 대안으로 PGU 를 제안한다.

2. 성능 편차 기반 사용자 세분화

그림 1 은 본 논문에서 제안하는 성능 편차 기반 세분화를 설명한다. 본 논문에서는 언러닝 기법의 강건성을 정밀하게 분석하기 위해 성능 편차 기반의 사용자 세분화를 수행한다. 이는 사용자-아이템 상호작용 희소성과 연결 구조의 복잡성이 복합적으로 반영된 척도로, 전체 평균 대비 개별 사용자의 성능 지표인 노출 및 효용이 벗어난 정도인 성능 편차를 통해 계산된다. 전체 사용자 집합을 U, 개별 사용자 u

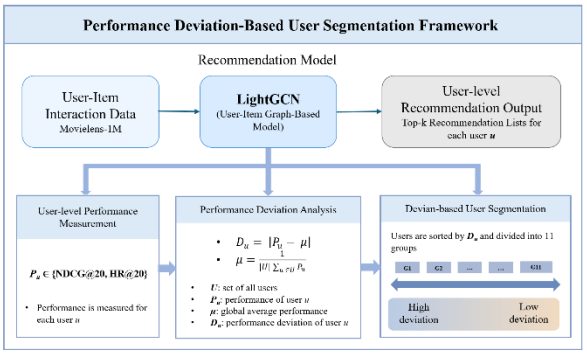

그림 1. 성능 편차 기반의 사용자 세분화 프레임워크

의 성능 지표를 P_u라고 할 때, 전체 평균 성능 μ 와 사용자 u의 성능 편차 D_u는 식 (1)과 같이 정의된다.

$$D_u = |P_u - \mu|, \mu = \frac{1}{|U|}\sum_{u \in U} P_u \qquad (1)$$

이 편차가 큰 사용자는 모델의 주류 학습 분포에서 벗어난 이상치로 간주되며, 이는 언러닝과 같은 구조적 교란에서의 취약점으로 작용한다. 본 논문은 전체 사용자를 해당 편차 순으로 정렬하여 동일한 크기의 11 개 그룹(G1~G11)으로 분할하였다. 이 중 G1 은 편차가 가장 큰 이상치 그룹으로서 SISA 와 같은 데이터 분할 방식에 취약한 반면, G11 은 편차가 가장 작은 일반 사용자 그룹으로 전형적인 연결 구조를 통해 모델이 안정적으로 예측 가능한 특성을 가진다.

3. 실험 및 결과 분석

3.1. 실험 환경

모든 실험은 MovieLens-1M 데이터셋을 기반으로 동일한 데이터 분할과 모델 설정 하에서 수행되었으며, 추천 모델로는 사용자-아이템 간의 고차원 연결성을 효과적으로 포착하는 LightGCN 을 채택하여 비교하였다[3, 4]. 성능

평가는 상위 20 개 추천 결과를 기준으로 NDCG@20 과 HR@20 지표를 사용하였다. 사용자 그룹별 강건성 분석을 위해, 전체 사용자를 성능 편차 순으로 정렬하여 G1 부터 G11 까지 동일한 크기의 11 개 구간으로 세분화하였다.

3.2. 데이터 분할 기반 언러닝의 구조적 한계

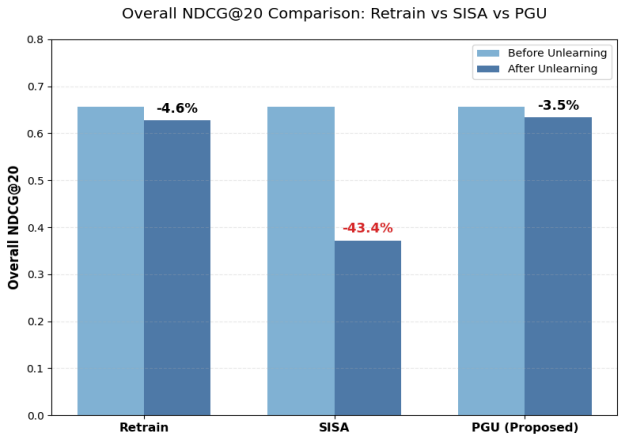

그림 2. 언러닝 기법 적용 전후의 전체 사용자 평균 NDCG@20 비교

$$\Delta = \frac{(P_{method} - P_{retrain})}{P_{retrain}} \times 100 \qquad (2)$$

각 언러닝 기법의 강건성을 정량적으로 평가하기 위해 Retrain 모델 대비 성능 변화율 Δ 를 식 (2)와 같이 정의하였다. 그림 2 은 언러닝 적용 전후의 전체 사용자 평균 NDCG@20 변화를 나타낸다. 실험 결과 파라미터 업데이트 방식인 Retrain 과 PGU 는 언러닝 이후에도 오차 범위 5% 내외로 원본 대비 성능 하락 폭이 미미하여 기존 랭킹 품질을 안정적으로 유지함을 알 수 있다. 반면, 데이터 분할을 수행하는 SISA 는 평균 40% 이상의 급격한 성능 붕괴를 보였다. 이러한 경향성은 HR@20 지표에서도 동일하게 관찰되었다. 이는 데이터 분할 방식이 협업 필터링의 핵심인 사용자 간 연결성을 파괴함을 시사한다.

3.3. 사용자 그룹별 강건성 분석

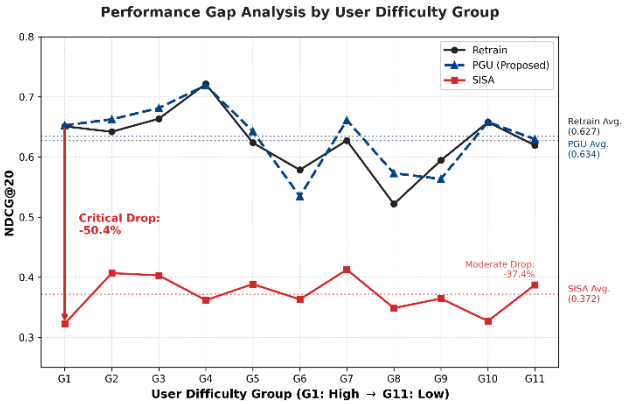

그림 3. 사용자 그룹별 NDCG@20 성능 경향성 비교

그림 3 은 사용자 그룹별 NDCG@20 변화 추이를 보여준다. 각 사용자 그룹의 특성은 단순히 모델의 성능을 예측하는 선형적 척도가 아닌, 상호작용의 희소성과 연결 패턴의 불확실성을 복합적으로 내포한 구조적 지표이다. 따라서 사용자 그룹 간의 성능 지표가 우상향을 보이지 않는 현상은 각 그룹 내에 존재하는 데이터의 분포적 다양성이 반영된 결과로 해석할 수 있다.

표 1. 주요 사용자 그룹 및 전체 평균에 대한 Retrain 대비 성능 격차 비교

User Group	*Retrain*	*PGU*	*Gap (%)*	*SISA*	*Gap (%)*
G1 (High-risk)	0.6510	0.6527	+0.26%	0.3230	-50.38%
G6 (Median)	0.5785	0.5347	-7.57%	0.3632	-37.22%
G11 (Low-risk)	0.6195	0.6297	+1.65%	0.3875	-37.45%
Average	0.6274	0.6345	+1.13%	0.3716	-40.43%

표 1 에서 확인할 수 있듯, SISA 는 이러한 데이터 분포의 등락과 무관하게 구조적인 한계를 드러냈다. SISA 는 연결 정보가 풍부한 G11 에서도 성능 저하가 발생했으며 연결성이 희소한 G1 에서는 Retrain 대비 50.4%의 성능 붕괴를 기록하였다. 이는 데이터 분할 방식이 연결 구조가 취약한 사용자 그룹에게 더욱 심각한 피해를 입힘을 실증한다.

반면 PGU 는 G1 에서도 Retrain 과 대등한 수준의 성능 보존력을 보였다. 이는 PGU 가 데이터의 구조적 복잡도에 상대적으로 낮은 영향을 받으며 성능을 보존함을 입증한다.

4. 결론

본 논문에서는 추천 시스템의 언러닝 기법에서 기존의 전체 평균 정확도 지표가 모델의 구조적 강건성을 온전히 입증하지 못한다는 한계를 규명하였다. 특히 추천 시스템은 데이터 간의 고차원적 연결성 학습이 성능의 핵심이나, 단순 평균 지표는 SISA 와 같은 데이터 분할 방식이 야기하는 연결 구조의 물리적 훼손과 성능 붕괴를 포착하지 못한다. 반면 PGU 는 그레이디언트 투영을 통해 데이터 손실 없이 모델의 지식을 보존하는 기법으로, 두 방식의 구조적 차이는 단순 평균 지표만으로는 식별하기 어렵다.

따라서 본 논문은 데이터 분할이 초래하는 구조적 취약성을 정밀하게 진단하기 위해 성능 편차 기반의 사용자 세분화 분석 프레임워크를 제안한다. 데이터 분할로 인한 연결 단절은 상호작용이 희소한 사용자에게 집중적인 성능 붕괴를 초래한다. 제안하는 프레임워크는 전체 사용자를 예측 편차에 따라 세분화함으로써, 평균 지표 뒤에 은폐된 취약 사용자 그룹의 성능 붕괴를 국소적으로 정량화한다.

실험 결과, SISA 는 평균적으로 40% 이상의 성능 하락을 보였을 뿐만 아니라 사용자 그룹별 분석에서 그 구조적 한계를 더욱 명확히 드러냈다. 특히 연결성 의존도가 높고 모델 학습이 까다로운 G1 에서 SISA 는 최대 50.4%의 성능 붕괴를 보였다. 반면, PGU 기법은 대부분의 구간에서 Retrain 과 대등한 성능을 유지하여 데이터의 구조적 복잡도와 무관한 강건성을 입증하였다.

ACKNOWLEDGMENT

본 논문은 과학기술정보통신부의 재원으로 한국연구재단의 지원과(RS-2024-00345398) 과학기술정보통신부에서 지원하는 DGIST 기관고유사업의 지원을(26-ET-02) 받아 수행되었음.

참 고 문 헌

[1] Bourtoule L. et al." Machine Unlearning," IEEE Symposium on Security and Privacy (SP), pp. 141-159, 2021.

[2] Sekhari A. et al." Remember What You Want to Forget: Algorithms for Machine Unlearning," Advances in Neural Information Processing Systems (NeurIPS), pp. 18075-18086, 2021.

[3] He X. et al." LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation," Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 639-648, 2020.

[4] Harper F. M., Konstan J. A." The MovieLens Datasets: History and Context," ACM Transactions on Interactive Intelligent Systems, vol. 5, no. 4, pp. 1-19, 2015.