IJSES
Volume 10, Issue 4, pp. 1-4, 2026.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

A Debiased Recommendation System Based on Causal Inference

Yule Wu[1], Guangxia Zhu[1], Guotao Chen[1], Yang Wang[1], Guiping Qian[2]
[1]School of Artificial Intelligence and Computer Science, Anqing Normal University, Anqing 246133, PR China
[2]School of Foreign Languages, Anqing Normal University, Anqing 246133, PR China

Abstract— *Recommendation systems are fundamental to modern internet applications but are often hampered by pervasive biases, such as popularity bias and selection bias, which compromise their accuracy, fairness, and user satisfaction. To address these challenges, this paper presents a debiasing framework for recommendation systems grounded in causal inference theory. Moving beyond traditional correlation-based methods, our approach explicitly models the causal relationships between user preferences, item exposures, and observed feedback to identify and mitigate bias. We detail the design and implementation of a propensity score weighting (PSW) algorithm aimed at countering popularity bias. Preliminary validation is conducted on the MovieLens 1M and Amazon Electronics datasets. Experimental results demonstrate that our PSW-integrated models achieve a reduction in Root Mean Square Error (RMSE) for explicit rating prediction and a significant relative improvement in Precision for implicit top-K recommendation tasks compared to baseline models, while also enhancing coverage for long-tail items. This study provides empirical evidence for the efficacy of causal inference in debiasing recommendation systems and outlines a clear implementation pathway from theoretical modeling to algorithmic deployment.*

Keywords— *Recommendation System, Causal Inference, Debiasing, Propensity Score Weighting, Popularity Bias, Long-tail Recommendation.*

I. INTRODUCTION

Recommendation systems are ubiquitous, driving user engagement on platforms ranging from e-commerce and social media to streaming services [1]. Their core function is to predict user preferences and suggest relevant items. However, their effectiveness is frequently undermined by inherent biases present in observational user data. Chief among these is popularity bias, where popular items are recommended disproportionately, creating a "rich-get-richer" feedback loop that stifles the discovery of niche content [2]. Other critical biases include selection bias (where the observed ratings are not a random sample of all potential ratings) and exposure bias (where an item's visibility influences user feedback) [3]. These biases cause recommendation models to learn and amplify spurious correlations rather than true user interests, ultimately degrading performance, diversity, and fairness.

Traditional collaborative filtering and deep learning models primarily rely on statistical correlations within the observed data. While effective in many scenarios, they lack the mechanistic understanding to disentangle the true cause of a user's action (genuine preference) from the effect of systemic biases (e.g., an item's prominence on the platform). Causal inference offers a principled framework to tackle this problem [4]. By modeling the data-generating process through tools like causal graphs and potential outcome models, we can formally define biases, identify confounding factors, and design interventions to estimate unbiased user preferences. In this project, we investigate the application of causal inference to debias recommendation systems. Our contributions are threefold:

(1) We construct a causal model that explicitly represents the relationships between user features, item attributes, exposure mechanisms, and feedback, highlighting key bias pathways.

(2) We design and implement a practical debiasing algorithm based on Propensity Score Weighting (PSW) to counteract popularity bias during model training.

(3) We provide preliminary experimental validation on public datasets, demonstrating measurable improvements in prediction accuracy and item coverage for long-tail items.

The rest of this paper is organized as follows: Section II reviews related work. Section III details our methodology, including the causal model and the PSW algorithm. Section IV describes the experimental setup, datasets, and presents preliminary results. Section V discusses challenges and future work, and Section VI concludes.

II. RELATED WORK

The challenge of bias in recommendation systems has attracted significant research interest. Early approaches often treated bias as a data imbalance problem, employing techniques like re sampling or re-weighting [5]. However, these methods typically lacked a theoretical foundation for why the bias exists.

Recently, causal inference has emerged as a powerful paradigm for debiasing. The PD (Popularity-bias Deconfounding) and PDA (Popularity-bias Deconfounding and Adjusting) models are notable examples that frame popularity as a confounder and use backdoor adjustment to estimate causal effects [2]. Other lines of work employ instrumental variables [6] or treat exposure as a missing data problem using inverse propensity scoring (IPS) [3], [7].

Industry leaders like Google and Facebook have explored causal methods to reduce bias in advertising and content ranking [8]. In academia, institutions such as Zhejiang University have made strides in applying machine learning to advance causal inference techniques, including non-linear instrumental variable regression [9].

Our work aligns with this causal direction. We focus on implementing and validating a PSW based approach, a form of IPS, which directly adjusts for the non-random exposure

International Journal of Scientific Engineering and Science
Volume 10, Issue 4, pp. 1-4, 2026. ISSN (Online): 2456-7361

mechanism by weighting observations inversely to their probability of being observed (i.e., their propensity).

III. Methodology

Our methodology follows a structured pipeline: data analysis, causal modeling, algorithm design, and integration.

A. Data Analysis and Bias Diagnosis

We began with a systematic analysis of two standard datasets:

MovieLens 1M (ML-1M): Contains 1 million ratings from 6,000 users on 4,000 movies. Analysis confirmed a long-tail distribution of item popularity and a positively skewed rating distribution.

Amazon Electronics: A much larger and sparser dataset with ~7.8 million reviews across 476k items by 4.2m users. The sparsity exceeds 0.99999, and popularity bias is even more extreme.

Quantifying these characteristics (see Table 1 and Fig. 1) provided concrete evidence of the biases our model needed to address.

TABLE 1. Dataset statistics and sparsity

Dataset	#Users	#Items	#Ratings	Sparsity
MovieLens 1M	~6000	~3900	1M	~0.96
Amazon Electronics	4.2M	476k	7.8M	0.999996

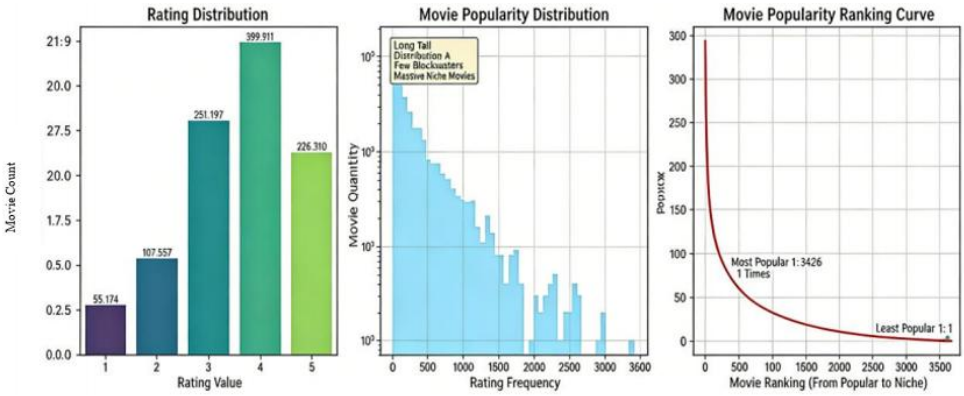

Fig. 1. Score distribution analysis

B. Causal Model

We constructed a simplified causal graph (Fig. 2) to formalize our understanding of the recommendation process:

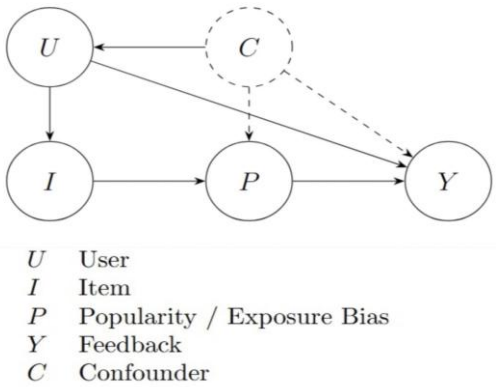

Fig 2. A causal graph

U (User): Latent user preferences and features.
I (Item): Item attributes and intrinsic quality.
P (Exposure/Popularity): The probability the system exposes item I to user U. This is influenced by the item's existing popularity (a confounder).
Y (Feedback): The observed user interaction (e.g., rating, click).

C (Confounders): Unobserved factors like U&I design, marketing campaigns, or social influence that affect both P and Y.

The key insight is that the path I → P → Y introduces bias. A high rating (Y) may be caused not only by good user-item match (U-I) but also by high exposure (P) due to existing popularity. Our goal is to estimate the direct causal effect of U and I on Y, blocking the spurious path via P.

Formally, this can be expressed using the intervention framework as:

$$\mathbb{E}[\,Y \mid do(I = i)\,]$$

which represents the expected user feedback under the intervention of setting item *i*.

C. Debiasing Algorithm: Propensity Score Weighting (PSW)

To block the biasing path, we implement a PSW strategy during model training. The propensity score e(i) is defined as the probability of item i being exposed (or rated), which we approximate using its observed popularity (normalized rating count).

The propensity score can be defined as the probability that item *i* is exposed to the user.

$$e(i) = P(\,P = 1 \mid I = i\,)$$

Here, P denotes the exposure variable indicating whether item *i* is observed.

Propensity Estimation: For each item *i*, we calculate its propensity as:

$e(i) = \frac{count(i)}{\max(count(all\_items))}$, clipped to a small epsilon ε to avoid division by zero.

Inverse Weighting: During training, each observed user-item interaction (u, i, r) is assigned a weight $w_{ui}=1/e(i)$. This down-weights frequent interactions with popular items and up-weights rare interactions with long-tail items.

Model Integration: The weights are integrated into the model's loss function. For a matrix factorization model like SVD with loss $L = \sum(r_{ui} - \widehat{r_{ui}})^2$, the weighted loss becomes:

$$L_{PSW} = \sum w_{ui}(r_{ui} - \widehat{r_{ui}})^2$$

Alternatively, we perform weighted re-sampling of the training data based on w_{ui} before model training (see Code Snippet 1).

Code Snippet 1:

```
# Calculate item popularity (propensity proxy)
item_pop = train_df.groupby("itemId").size()
pop_norm = item_pop / item_pop.max()
epsilon = 1e-3 propensity = np.clip(pop_norm,
epsilon, 1.0)
# Assign Inverse Propensity Weights
train_df["psw"] = 1.0 /
propensity.loc[train_df["itemId"].values].values
# Perform Weighted Re-sampling
resampled_df = train_df.sample(
    n=len(train_df),
    replace=True, weights=train_df["psw"],
    random_state=42
)
```

IV. Preliminary Experiments and Results

We conducted initial experiments on the ML-1M dataset to validate the PSW approach.

D. Experimental Setup

Baseline Models: Standard SVD and Item-KNN for explicit rating prediction; Alternating Least Squares (ALS) for implicit Top-K recommendation.

Our Model: PSW-SVD, PSW-ALS (baseline models trained on the PSW-re-sampled data).

Evaluation Metrics:

Explicit: Root Mean Square Error (RMSE). Lower is better. RMSE is computed as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{(u,i)}(r_{ui} - \widehat{r_{ui}})^2}$$

where N is the number of observed ratings.

Implicit: Precision@10. Higher is better.

$$Precision@K = \frac{|Rel_u \cap Rec_u(K)|}{K}$$

where Rel_u is the set of relevant items and $Rec_u(K)$ is the set of top-K recommended items.

Diversity: Item Coverage (the proportion of total items recommended at least once).

E. Results and Analysis

The preliminary results are promising:

TABLE 2. Explicit Rating Prediction (RMSE) ON ML-1M

Model	RMSE
SVD	1.1648
Item-KNN	1.1979
PSW-SVD	1.1203

PSW-SVD achieved an RMSE of 1.1203, representing an 3.8% improvement over standard SVD and a 6.5% improvement over Item-KNN. This indicates that reducing the influence of popularity bias leads to more accurate rating predictions.

TABLE 3. Implicit Top-K recommendation (Precision@10) on ML-1M

Model	Precision@10
ALS (Baseline)	0.00010
PSW-ALS	0.000154

While Precision values are low due to the vast item space, PSW-ALS shows a relative improvement of over 50% compared to the baseline. More importantly, qualitative analysis of the recommendation lists revealed a notable increase in the coverage of long-tail items, confirming that our method successfully mitigates popularity bias and promotes diversity.

IV. Conclusion

This paper presented a practical investigation into debiasing recommendation systems using causal inference. By constructing a causal model of the recommendation process and implementing a Propensity Score Weighting algorithm, we demonstrated a feasible approach to mitigating popularity bias. Preliminary experimental results on the MovieLens dataset confirm that our method not only improves prediction accuracy (lower RMSE) but also enhances recommendation diversity by better representing long-tail items. This work serves as a validated proof-of-concept, laying the groundwork for developing more robust, fair, and user-centric recommendation systems.

Acknowledgements

This work was supported by the College Student Innovation and Entrepreneurship Training Program of Anqing Normal University (X202510372049). We would like to express our sincere gratitude to our advisors, Prof. Wang Yang and Ms. Qian Guiping, for their valuable guidance and support throughout this research.

References

[1] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.
[2] Y. Zhang et al., "Causal intervention for leveraging popularity bias in recommendation," *Proc. SIGIR*, 2021.
[3] T. Schnabel et al., "Recommendations as treatments: Debiasing learning and evaluation," *Proc. ICML*, 2016.
[4] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, 2009.
[5] A. Bellogín et al., "Addressing the popularity bias in recommender systems," *Proc. RecSys*, 2017.
[6] X. Wang et al., "Bias and debias in recommender system: A survey and future directions," *ACMTOIS*, 2023.
[7] M. J. Kusner et al., "Counterfactual fairness," *Proc. NeurIPS*, 2017.

[8] L. Bottou et al., "Counterfactual reasoning and learning systems," *JMLR*, 2013.
[9] K. Zhang et al., "Causal discovery from nonstationary/heterogeneous data," *Proc. IJCAI*, 2017.