Bias Amplification in AI-Based Algorithms for Personalized Recommendation

PI: Bamshad Mobasher, Professor

School of Computing, CDM

URC – Summer Research Grant Program – 2023

Abstract:

With the growing influence of AI-based data-driven personalization algorithms our daily decisions across various domains, there has been a shift of focus in research from model accuracy to other socially-sensitive concerns such as diversity, fairness, and bias mitigation. Optimization that enhances overall accuracy in predictions may have the unwanted side effect of disadvantaging certain users by perpetuating or even amplifying existing biases in data used to train the predictive models. Our main goal in this research is to study how different machine learning algorithms used in personalized recommender systems propagate or amplify existing preference biases in the input data. We use the notion of *bias disparity* to measure bias propagated by different algorithms with respect to the preferences of specific user groups such as men or women towards specific item categories such as different news categories, movie genres, or job types. We will conduct extensive experiments using existing data sets used for training recommendation algorithms to measure the impact of bias propagation on groups of users and the system as a whole. If successful, our findings could be used by system designers in determining the choice of algorithms and parameter settings in critical domains where the output of the system must conform to legal and ethical standards or to prevent discriminatory behavior by the system, or in situations where there is a danger of amplifying existing intense preferences leading to the "rabbit-hole effect."

Project Description

1. Motivation and Project Goals

Intelligent personalized applications such as recommender systems have become essential online tools in many domains such as e-commerce, music and video streaming, online news, and social media marketing. These systems help alleviate information overload and assist users in decision making by tailoring their recommendations users' preferences. The prevalence of these AI-based data-driven algorithmic systems into many realms of society has raised concerns that such systems may exhibit bias, produce unfair results, and entrench problems of inequity [1, 2, 4]. This is particularly evident where high stakes decisions are made, ones that have significant impact on individuals' lives and livelihoods.

Due to these concerns, in recent years there has been a shift of focus from achieving the best accuracy in recommendation to other important measures such as diversity, novelty, as well as socially concerns such as fairness [3, 4]. One of the key issues with which to contend is that biases in the input data (used for training predictive models) are reflected, and in some cases amplified, in the results of recommender system algorithms. This is especially important in contexts where fairness and equity matter or are required by laws and regulations such as in lending, education, housing, and employment.

Personalized recommendation algorithms learn from existing patterns of bias in users' preferences (users' tendencies to choose one type of item over another). In and of itself, this type of bias is not necessarily a negative phenomenon. In fact, patterns in users' preference biases are the key ingredients that recommendation algorithms use to construct predictive models and provide users with relevant personalized content. However, in certain contexts the propagation of preference biases can be problematic, especially if algorithms amplify these biases in a way that diverge from users' interests. For example, in the news recommendation domain, as well as in social media, amplification of preference biases can cause filter bubbles [5] and limit the exposure of users to diverse perspectives and content. In job recommendation and lending domains, existing biases in the input data may reflect historical societal biases against protected groups, which must be accounted for by the system [6].

Our main goal in this research is to study how different machine learning algorithms used in personalized recommender systems might propagate or amplify existing preference biases in the input data and the different kinds of impacts such disparity between input and the output might have on users. For the purpose of this analysis, we use the notion of *bias disparity*, a recently introduced metric [7, 11] that considers biases with respect to the preferences of specific user groups such as men or women towards specific item categories such as different news categories, movie genres, or job types. This metric evaluates and compares the preference biases in both the input and the output data and measures the degree to which recommendation algorithms may propagate these biases, in some cases dampening them and in others amplifying them.

We are specifically interested in answering the following research questions:

- **RQ1:** How do different recommendation algorithms propagate existing preference biases in the input data to the generated recommendation lists?
- **RQ2:** How does the bias disparity between the input and the output differ for different

user groups (e.g., men versus women)?

- **RQ3:** How does bias disparity impact individual users with a high degree of preference intensity (positive or negative) with respect to different categories of items?

To address these research questions, we will conduct experiments on a movie rating dataset to study the behavior of different types of algorithms in the way in which they propagate preference biases in the input data. These findings maybe especially important for system designers in determining the choice of algorithms and parameter settings in critical domains where the output of the system must conform to legal and ethical standards or to prevent discriminatory behavior by the system, or in situations where there is a danger of amplifying existing intense preferences leading to the "rabbit-hole effect."

Relation to Broader Research and Future Funding

This research activity will be conducted under the auspices of the Center for Web Intelligence directed by Professor Mobasher. The center's work encompasses a variety of research projects related to recommender systems, Web personalization, and social computing. This effort over the summer of 2023 will be part of a larger research effort focusing on responsible and ethical artificial intelligence, and more specifically on mitigating bias in recommender systems. We hope that collectively these research efforts lead to more extensive proposals for external funding.

2. Related Work

Various metrics have been introduced for detecting model biases. The metrics presented in [2], such as absolute unfairness, value unfairness, underestimation, and overestimation focus on the discrepancies between the predicted scores and the true scores across protected and unprotected groups and consider the results to be unfair if the model consistently deviates (overestimates or underestimates) from the true ratings for specific groups. These metrics can be used to measure unfairness towards users belonging to specific groups.

Steck [8] has proposed an approach for calibrating recommender systems to reflect the various interests of users relative to their initial preference proportions. The degree of calibration is quantified using the Kullback-Leibler (KL) divergence. This metric compares the preference distribution over a set of item categories preferred by a user to the distribution in a user's recommendation list generated by the system. A postprocessing re-ranking algorithm is then used to adjust the calibration degree in the recommendation list.

The authors in [9] have looked into the influence of algorithms on the output data; they tracked the extent to which the diversity in user profiles change in the output recommendations. The work in [10] has also looked into the author gender distribution in user profiles and has compared it with that of the output recommendations. These results have suggested that commonly used algorithms such those based on the k-nearest-neighbor methods tend to amplify and propagate existing biases.

Tsintzou et al. [11] sought to measure unfairness towards user groups by modeling the bias disparity by calculating the difference between the preferences of the user in the input data and the predicted preference of the user by the recommendation algorithm. Bias disparity metric looks at these differences in a fined-grained way, evaluating the preferences of specific user groups for specific item categories. KL divergence used in Steck's approach measures more generally the difference in preference distributions across categories. One of the limitations of the work of

Tsintzou et al. [11] is that they perform their analysis only for K-nearest-neighbor models. In our work, we build on this earlier research by considering a variety of recommendation algorithms and identifying specific characteristics of user preference profiles that may lead to bias propagation under different algorithms.

3. Proposed Research Approach and Activities

A key aspect of our analysis is the quantification of the notion of bias disparity, enabling the measurement of the degree to which a given recommendation algorithm may propagate or amplify existing biased in the initial user preference data used for model training. To this end, we first define the notion of preference ratio to capture the level of preference by a given user group for a category of items.

Let ***G*** be a group of users and let ***C*** be a category of items with which user in ***G*** have interacted. We define *preference ratio* of ***G*** on ***C*** as:

$$PR_S(G,C) = \frac{\sum_{u\in G}\sum_{i\in C} S(u,i)}{\sum_{u\in G}\sum_{i\in I} S(u,i)} \qquad (1)$$

where, ***S*(*u*, *i*)** represents a measure on the preference of user ***u*** on and item ***i*** (for example a rating value given to item ***i*** by user ***u*** in the course of interaction with the system). In essence, this ratio measures the conditional probability of selecting an item from category ***C*** given that this selection is done by a user in group ***G***.

Now, bias disparity can be defined as the relative difference of the preference bias ratios between the input ***S*** and output of a recommendation algorithm ***R***:

$$BD(G,C) = \frac{PR_R(G,C) - PR_S(G,C)}{PR_S(G,C)} \qquad (2)$$

We assume that a recommendation algorithm provides each user ***u*** with a ranked list of recommended items $\boldsymbol{R_u}$. Let ***R*** be the collection of all the recommendations to all the users represented as a binary matrix, where ***R*(*u*, *i*)** = 1 if item ***i*** is recommended to user ***u***, and zero otherwise. The overall bias disparity for a category ***C*** is obtained by averaging bias disparities across all users regardless of the group.

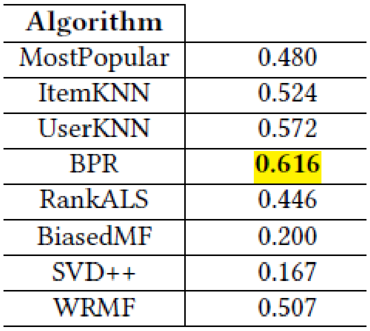
A bias disparity of zero or near zero means that the input and output of the algorithm are almost the same with respect to the prevalence of the chosen category: the algorithm reflects the users' preferences quite closely. A negative bias disparity means that the output preference bias is less than that of the input. In other words, the preference bias towards a given category is dampened. The extreme value, BD = −1, would indicate that a category important in a user's profile is completely missing from the system's recommendations ($\boldsymbol{PR_R = 0}$). If the bias disparity value is positive, the output preference bias towards an item category is higher than that of the input, indicating that the importance of the given category has been amplified by the algorithm.

We design our experiments to address the aforementioned research questions by measuring bias disparity across all users and item categories for a variery of baseline algorihtms (RQ1), for contrasting user groups (e.g. male vs. female) across different item categories using a subset of algorithms (RQ2), and for speicfic users or groups of users who exhibit certain characteristics in their profiles indicating intese interest in a particular topic category (RQ3).

Data Set and Baseline Algorithms

For our initial experiments we will use the MovieLens (ML) dataset[1], a publicly available dataset for movie recommendation which is widely used in recommender systems experimentation. The ML data contains one million ratings (in the scale of 1 to 5) from 6,040 users on 3,702. For item groups we will use the pre-specified movie genres available as part of the data set. For identifying user groups, we either use demographic attributes associated with users in the dataset or use certain statistical characteristics of user profiles (e.g., the distribution of genres among the movies rated by the user indicating the diversity or intensity of user preferences).

For our experiments we use four groups of widely used recommendation algorithms: memory-based, model-based (ranking), model-based (rating) and a popularity-based baseline. We have selected both user-based and item-based k-nearest- neighbor methods from the memory-based category. BPR (Bayesian Personalized Ranking) [12], RankALS [13] were selected from the learning-to-rank category of algorithms. From the rating-oriented latent factor models, we chose Biased Matrix Factorization (BiasedMF) [14], SVD++ [15], and Weighted Regularized Matrix Factorization (WRMF) [16]. We chose a non-personalized most-popular recommender as a baseline as this algorithm would be expected to maximally amplify the popularity bias in the recommendation outputs [17]. For each of these algorithms, we have already tuned the parameters and picked the set of parameters that gives the best performance in terms of normalized Discounted Cumulative Gain (nDCG) of the top 10 listed items. The nDCG is a commonly used measure of ranking accuracy for recommendation results. The algorithms and their optimized accuracy (nDCG) values are depicted in Table 1. For our experiments on bias disparity discussed below, we plan to use this specific configurations of each of the specified algorithms.



Algorithm	
MostPopular	0.480
ItemKNN	0.524
UserKNN	0.572
BPR	**0.616**
RankALS	0.446
BiasedMF	0.200
SVD++	0.167
WRMF	0.507

Table 1: nDCG values with selected parameters

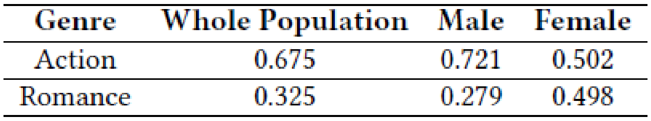
Experimental Design

In order to address RQ1, we will conduct a set of experiments using the full set of users as a single group but across multiple movie genres as item categories. The goal of these experiments is to test the hypothesis that certain classes of algorithms, such as neighborhood based models, tend to significantly amplify thr initial preference bias ratios in the data (regardless of user groups). In a secondary set of experiments, we will identify users with zero initial preference ratios (Eq. 1) on one of the selected genres to see the effects of different algorithms on bias disparity. Our goal is

[1] https://grouplens.org/datasets/movielens

to determine if input preference ratios are significantly different from the output preference ratios in the recommendations (i.e., if bias disparity was significantly different from 0, due to the dampening or amplification of preference biases). Understanding the behavior of algorithms in this context is essential to avoid systemic bias amplification (the type of bias amplification commonly seen on social media platforms).

To address RQ2, for our preliminary experiments we focus on bias disparity among male users versus female users relative to several movie genres. We hypothesize that neighborhood-based algorithms tend to amplify existing gender biases relative to specific movie genres. We also hypothesize that model-based algorithms tend to dampen such existing biases. In the first step, we separately calculate preference ratios (Eq. 1) of male users and female users (user groups) on the selected genres and then compute the corresponding bias disparity values (Eq. 2). In the second step, we will compute the preference ratios and bias disparities for our movie genres on the whole user data (without partitioning into separate user groups). This comparison will allow us to determine if the sex attribute is possibly a determining factor in the bias disparity, especially if it persists across different types of algorithms. While we will focus on different combinations of item categories and conduct the experiments across many algorithms, our initial analysis of the data suggests that there are indeed significant preference biases among the two groups for certain categories. For example, Table 2 indicates that there is strong preference bias towards action movies among male users and for romance movies among female users. Our experiments generating recommendation lists for these groups using different predictive algorithms will allows us to determine if certain algorithms tend to perpetuate and amplify these biases and if others tend to dampen existing biases.



Genre	Whole Population	Male	Female
Action	0.675	0.721	0.502
Romance	0.325	0.279	0.498

Table 2: Input preference ratio for Action and Romance

We performed a preliminary experiment only for the action movie genres showing that clearly some algorithms (such as item-based k-nearest-neighbor and the popularity-based algorithm) dramatically amplify bias disparities. These initial results (depicted in Figure 1) show that this line of investigation is likely to produce interesting results.

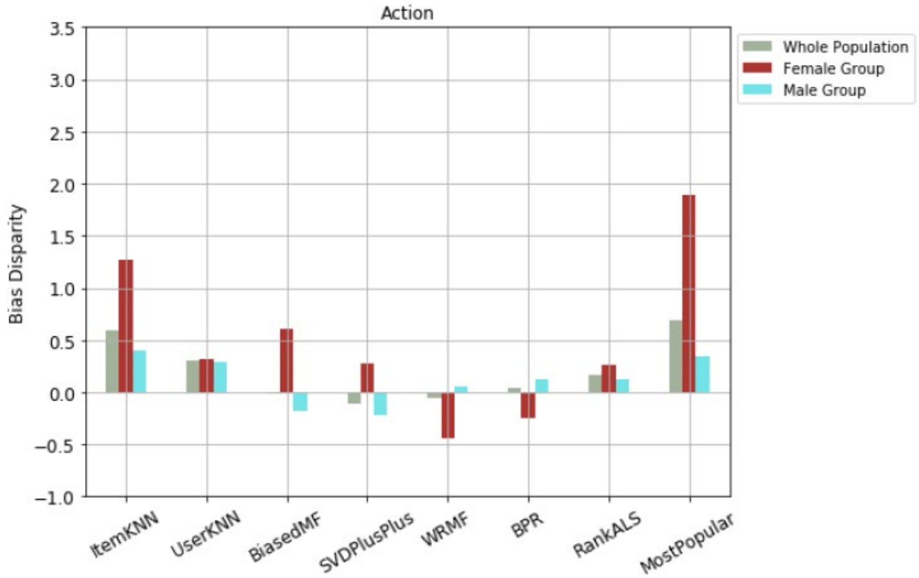


Figure 1. Degree of Bias Amplification Across Algorithms

To address RQ3, we will identify users with profiles that suggest intense interest in specific categories of items. These tend to be users whose preference ratios tend to be on the extremes along the bias scale with very few categories showing high preference ratio while other categories showing close to zero ratio. The reason for focusing on these "extreme" users is that, for certain classes of recommendation algorithms, they often exert more influence in terms of determining what items are recommended to other users. We hypothesize that algorithms that tend to amplify biases will also amplify this "bandwagon" effect resulting in skewing bias disparities across the whole system. For our experiments, we will identify the "extreme" users based on statistical analysis their rating profiles in terms of the distribution of genres (specifically using entropy as a measure of uniformity, or lack thereof, across movie genres). We will then conduct a set of experiments similar to those used for RQ1 to observe the behavior and the impact of each algorithm in terms of bias propagation.

While for the limited duration of this project we focus on a single (movie rating) data set. We intend to replicate these experiments across multiple datasets spanning other domains such as news recommendation, job recommendation, and music streaming.

References

[1] Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Recommendation and Retrieval. In Proceedings of the 13th ACM Conference on Recommender Systems (Copenhagen, Denmark) (RecSys ’19). Association for Computing Machinery, New York, NY, USA, 576–577.

[2] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In Advances in Neural Information Processing Systems. 2921–2930.

[3] Neil Hurley and Mi Zhang. 2011. Novelty and diversity in top-n recommendation–analysis and evaluation. ACM Transactions on Inter- net Technology (TOIT) 10, 4 (2011), 14.

[4] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Modeling and User-Adapted Interaction 25, 5 (2015), 427–491.

[5] Eli Pariser. 2011. The filter bubble: How the new personalized web is changing what we read and how we think. Penguin.

[6] Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism. nyu Press.

[7] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457 (2017).

[8] Harald Steck. 2018. Calibrated recommendations. In Proceedings of the 12th ACM conference on recommender systems. ACM, 154–162.

[9] Sushma Channamsetty and Michael D Ekstrand. 2017. Recommender response to diversity and popularity bias in user profiles. In The Thirtieth International Flairs Conference.

[10] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In Proceedings of the 12th ACM Conference on Recommender Systems. ACM, 242–250.

[11] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. 2018. Bias Disparity in Recommendation Systems. arXiv preprint arXiv:1811.01461 (2018).

[12] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 452–461.

[13] Gábor Takács and Domonkos Tikk. 2012. Alternating least squares for personalized ranking. In Proceedings of the sixth ACM conference on Recommender systems. ACM, 83–90.

[14] Arkadiusz Paterek. 2007. Improving regularized singular value decomposition for collaborative filtering. In Proceedings of KDD cup and workshop, Vol. 2007. 5–8.

[15] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 8 (2009), 30–37.

[16] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In ICDM, Vol. 8., 263–272.

[17] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality. In RecSys Posters, 2014.