Fairness Testing for Recommender Systems

Huizhong Guo
Zhejiang University
Hangzhou, China
huiz_g@zju.edu.cn

ABSTRACT

The topic of fairness in recommender systems (RSs) is gaining significant attention. However, current fairness metrics and testing approaches primarily cater to classification systems and are not suitable for RSs. To bridge this gap, we aim to address the specific challenges involved in fairness testing for RSs. In this paper, we present a novel testing approach specifically designed for RSs, which enables us to achieve accurate results while maintaining high efficiency. Additionally, we suggest potential avenues for further research in the realm of fairness testing for RSs.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Software and its engineering** → **Software testing and debugging**.

KEYWORDS

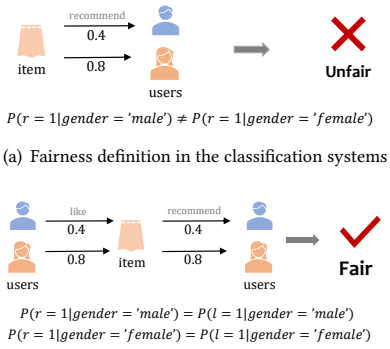
Recommender Systems, Fairness Testing, AI Ethics

ACM Reference Format:
Huizhong Guo. 2023. Fairness Testing for Recommender Systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '23), July 17–21, 2023, Seattle, WA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3597926.3605235

1 INTRODUCTION

Recommender systems (RSs) are extensively employed across various domains, including e-commerce, social networks, news media, and more. They offer a convenient means for users to discover relevant information while assisting content providers in reaching their intended audience. Nevertheless, during the collection of user feedback data, RSs can inadvertently introduce various biases, which can result in discrimination against certain groups. For instance, studies have revealed that Google's online ad recommendations exhibit a tendency to offer women fewer recommendations for high-paying job opportunities [3].

The current body of research on fairness testing primarily revolves around classification systems. Researchers have proposed diverse fairness testing strategies, including random sampling [5], probabilistic sampling [10], and gradient-based approaches [14], to uncover various forms of fairness issues. These approaches aim to reveal concerns like individual discrimination [6, 8] and group disparity [2, 7]. However, there is a need to extend these testing methodologies to accommodate the unique characteristics and challenges inherent in RSs.



(a) Fairness definition in the classification systems

(b) Fairness definition in the recommender systems

Figure 1: The difference between classification systems and recommender systems on the fairness definition.

Firstly, the fairness definition used in classification systems does not directly apply to personalized RSs. For example, as shown in Figure 1(a), a RS suggests items to male and female users with different probabilities. According to the fairness principle of demographic parity, which aims for equal favorable outcomes across user groups, this difference would be considered unfair. However, in practice, users of different genders often have distinct personalized preferences, as shown in Figure 1(b). In this context, the recommendation results are considered fair because the RS accurately captures and reflects each user's unique preferences. In addition, defining fairness in RSs is further complicated by users' expectations of factors like diversity, novelty, and popularity in the recommendations.

Second, existing testing methods do not match the testing requirements of RSs. As mentioned earlier, recommendation results in RSs are closely linked to a user's personalized preferences. Consequently, current individual fairness testing methods, which concentrate on generating discriminatory test cases [4, 14], are ill-suited for evaluating fairness in RSs. Moreover, the existing group fairness testing methods [5, 13] face challenges in meeting the high testing efficiency demands of RSs. RSs incorporate diverse sensitive attributes, including demographic factors like gender and age, as well as user behavioral attributes such as activity level and purchasing power. The combination of multiple attributes creates a vast search space, with significantly more potential candidate groups compared to traditional fairness testing domains. This expanded search space presents a substantial challenge for fairness testing in RSs. Additionally, RSs continually gather user feedback data and periodically update their models, necessitating an effective fairness testing method that can be regularly applied within a limited time frame.

ISSTA '23, July 17–21, 2023, Seattle, WA, USA

ACM ISBN 979-8-4007-0221-1/23/07.
https://doi.org/10.1145/3597926.3605235

2 RELATED WORK

Several studies have put forth definitions of fairness for RSs [9, 11, 12]. For instance, Yao et al. [12] categorized the population based on gender and proposed four group fairness metrics to assess disparities between male and female groups. They regarded the numerical magnitude of these metrics as indicators of RS fairness. Similarly, Li et al. [9] divided users into active and inactive groups according to their past behavioral history in an e-commerce RS. They used performance difference between user groups to measure RS fairness. However, these approaches primarily divide the population into two groups, limiting their ability to uncover deeply-hidden fairness issues in RSs. Additionally, they overlook the diverse user needs beyond recommendation performance and lack efficient testing methods.

Existing research in the field of fairness testing has predominantly focused on classification systems rather than RSs. For instance, Galhotra et al. proposed Themis [1, 5], a method that measures group discrimination in software by grouping users based on various sensitive attribute values. They employed brute-force enumeration to calculate group discrimination scores and causal discrimination scores for each user group. Zhang et al. proposed TestSGD [13], which automatically generates an interpretable rule set. They divide the population based on each rule, distinguishing those who conform to the rule from those who do not. By comparing the fairness scores of each group, they identify the groups experiencing the most disparate treatment. While these methods have proven effective for addressing fairness concerns in classification systems, they are not entirely applicable to the fairness definition within RSs. Furthermore, the efficiency of these testing methods falls short of the requirements of recommender systems.

3 FRAMEWORK

In this work, we propose a novel unified fairness testing framework called FairRec, which is specifically designed for RSs. We will introduce FairRec from two aspects of fairness definition and fairness testing.

Fairness definition for RSs. In a personalized RS, it is not sufficient to solely rely on evaluating fairness based on the model's output. To address this, we take into consideration the actual needs of users in real-life scenarios through the following three dimensions: *1) Recommendation performance.* It evaluates the accuracy of the recommendations for individual users. By analyzing the performance disparities, we can identify instances of potential unfairness where certain user groups may be disadvantaged. *2) Diversity.* The diversity dimension measures the variety of recommended content that users receive. It enables us to evaluate whether users consistently receive homogeneous recommendations, which may indicate the presence of an *echo chamber effect. 3) Popularity.* The popularity dimension evaluates the tendency of RSs to recommend predominantly popular items to users. By examining whether RSs consistently recommend highly popular items, regardless of user preferences, we can assess the potential impact of popularity bias.

In summary, we evaluate the recommendation results obtained by users using the three aforementioned metrics. Following that, we employ fairness testing methods to identify the most advantaged and disadvantaged user groups within the RS. The fairness of the RS can be defined based on the differences observed in these evaluation metrics between the identified user groups.

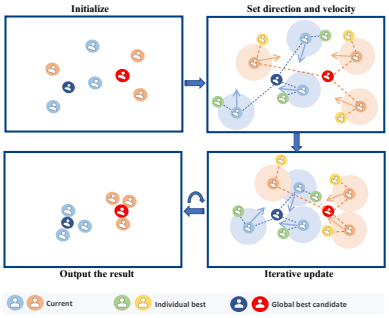


Figure 2: Double-ended discrete particle swarm optimization (DPSO) algorithm.

Fairness testing for RSs. In this work, our focus lies in conducting fairness testing among user groups that are segmented based on multiple sensitive attributes. To address the challenging search problem associated with RSs, we propose a novel algorithm called double-ended discrete particle swarm optimization (DPSO).

In the DPSO algorithm, we model each user group as a particle within the search space. For instance, consider a group comprised of 20-year-old male teachers, which can be represented as the point (0, 20, 5) in a 3-dimensional search space, where the dimensions correspond to gender (0 for male), age (20), and occupation (5 for teacher). To initiate the testing process, we initialize two particle swarms, each with the objective of finding the most advantaged and disadvantaged user groups, respectively. During the testing iteration, each particle navigates through the search space based on the guidance provided by its individual best position and the global best position, which represents the best results among all particles. The DPSO algorithm iteratively evaluates the testing objective and updates the particle swarms until the allocated testing budget is exhausted. Ultimately, DPSO generates two target groups along with their respective fairness scores, providing valuable insights into the fairness of the RS based on the identified advantaged and disadvantaged user groups.

4 FUTURE WORK

The field of fairness testing for recommender systems (RSs) is still in its early stages, and there is ample opportunity for future research to enrich this area. In this section, we discuss potential directions for further investigation:

Multiple stakeholders. In our work, the primary focus is on examining the fairness of users in RSs. However, it is important to acknowledge that item providers and platforms are also key stakeholders in the RS ecosystem. The fairness requirements and interests of these different parties can vary significantly, creating a complex challenge in balancing their rights and interests.

Implicit Attributes. Indeed, the consideration of implicit attributes in RSs is a significant aspect that can contribute to fairness testing research. Implicit attributes, such as activity level or purchasing power level, are often challenging to acquire directly from user profile data. However, they can still be utilized in the training process of RS models and potentially introduce biases that result in unfairness.

REFERENCES

[1] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 871–875.

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[3] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491* (2014).

[4] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. 2022. Explanation-Guided Fairness Testing through Genetic Algorithm. *arXiv preprint arXiv:2205.08335* (2022).

[5] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. 498–510.

[6] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, Vol. 1. Barcelona, Spain, 2.

[7] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[8] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[9] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*. 624–632.

[10] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 98–108.

[11] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware news recommendation with decomposed adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4462–4469.

[12] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems* 30 (2017).

[13] Mengdi Zhang, Jun Sun, Jingyi Wang, and Bing Sun. 2023. TestSGD: Interpretable Testing of Neural Networks Against Subtle Group Discrimination. *ACM Trans. Softw. Eng. Methodol.* (apr 2023). https://doi.org/10.1145/3591869 Just Accepted.

[14] Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 2020. White-box fairness testing through adversarial sampling. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 949–960.

Received 2023-05-24; accepted 2023-06-07