Advanced Topics

November 17, 2024

Breaking the Filter Bubble: Migrating Bias with Machine Learning Fairness Approaches

SICEN LIU

Abstract

In today's era of widespread mobile device usage, personalized recommendation systems continuously collect user information to create "filter bubbles", confining individuals within homogeneous information environments. This can distort people's overall understanding of issues and potentially exacerbate societal biases. Current mainstream approaches focus on optimizing recommendation strategies, utilizing diversified algorithms, or controlling recommendations based on diversity metrics. While these methods can alleviate some of the homogeneity in recommendations, they do not entirely prevent existing filter bubbles from contaminating the database used in real-time recommendation algorithms.

This paper proposes drawing from approaches used in machine learning fairness to mitigate bias, aiming to reduce the impact of unevenly distributed sensitive features on recommendation data sets. By addressing these biases, limiting the reinforcement of filter bubbles in personalized recommendations may be possible.

1 Industry trends and needs

With the widespread adoption of mobile devices, the use of social media has grown exponentially, and personalized content recommendations have followed suit. In 2009, Google began personalizing its search results. Rather than simply offering the most popular results, Google predicts what users will most likely click on, tailoring the information to individual preferences.

Eli Pariser, in his book The Filter Bubble: What the Internet Is Hiding from You [1], introduced the "filter bubble" hypothesis. He explains that data companies track your personal information to sell to advertisers, ranging from your political views to the products you browse online. In a personalized world, Pariser warns, we are increasingly categorized and shown only content that is familiar, comforting, and aligned with our beliefs. Because these filters operate invisibly, we often have no idea what is being excluded from our view. As a result, our past interests determine our future exposure, leaving less room for unexpected encounters that foster creativity, innovation, and the democratic exchange of ideas.

Filter bubbles have been blamed for several major societal issues, such as the spread of misinformation during the Brexit referendum and the 2016 U.S. presidential election [2]. They have also been linked to protests against immigration in Europe [3] and even health crises, such as the measles outbreaks in 2014 and 2015 [4]. In these cases, instead of fostering connections between different groups, social media platforms, through their personalized algorithms, have exacerbated divisions, reinforcing ideological differences and isolating groups further apart from one another.

The powerful role of filter bubbles in shaping public opinion and dividing society has made it one of the most pressing issues in today's social media landscape. Alleviating this problem has become a key focus in the research of recommendation algorithms.

2 Current Solutions and Critical Analysis

Existing methods to combat filter bubbles focus on improving users' awareness of diverse social opinions [5,6], enhancing models' diversity [7–9] and serendipity [10,11], correcting model behavior by watching debunking content [12], optimizing long-term user satisfaction [13], and making recommendation system controllable [14,15].

Tim Donkers *et al.* [5] suggested the use of fundamentally different diversification strategies to counteract two different types of echo chambers that occur in social media contexts. Mingkun Gao *et al.* [6] designed an intelligent system that improves awareness of diverse social opinions by providing visual hints and recommendations of opinions on different sides with different indicators. Guy Aridor *et al.* [7] designed a model that highlights the importance of collecting data on user beliefs and their evolution over time. Yong Liu *et al.* [8] proposed a novel diversified recommendation model DC^2B for interactive recommendation with users' implicit feedback, which employs a determinantal point process in the recommendation procedure to promote diversity of the recommendation results. Antonela Tommasel *et al.* [9] devised FRediECH, an echo chamber-aware friend recommendation approach that learns users and echo chamber representations from the shared content and past users' and communities' interactions to recommend potentially relevant and diverse friends from outside the network of influence of the users' echo chamber. Zachary A. Pardos *et al.* [10] showed a dramatic lack of novelty in RNN recommendations and depicted the characteristic trade-offs that make serendipity difficult to achieve. Yuanbo Xu *et al.* [11] developed NSR that can achieve superior serendipity by a 12% improvement in average while maintaining stable accuracy compared with state-of-the-art methods. Matus Tomlein *et al.* [12] attempted to burst the filter bubble by having the pre-programmed agents watch misinformation-debunking content. Chongming Gao *et al.* [13] proposed CIRS that augments offline reinforcement learning with causal inference, which first learns a causal user model on historical data to capture the overexposure effect of items on user satisfaction, then uses the learned causal user model to help the planning of the RL policy. Wenjie Wang *et al.* [14] proposed a new recommender prototype called the User-Controllable Recommender System, which enables users to actively control the mitigation of filter bubbles. Zhenyang Li *et al.* [15] proposed a general and easy-to-use reinforcement learning-based method, which can adaptively select few but effective connections between nodes from different communities as the exposure list and proved that it can serve as an effective solution to alleviate the filter bubble and the separated communities induced by it.

The goal of these methods is to improve the output of recommendation systems in various aspects, and they have achieved certain successes within their respective areas of optimization. However, they fail to address a fundamental issue: the recommendation system's database is inherently biased, and this bias is likely to worsen over time. Recommendation systems continuously track user behavior to update their recommendation databases. However, the behavior of users already trapped in filter bubbles is inherently constrained. If a certain type of information is overwhelmingly more frequent than others in what a user sees, the likelihood of them clicking on similar content will inevitably be much higher unless they intentionally manipulate their interactions. These behaviors are recorded by the recommendation system, causing it to fall deeper into bias. Therefore, we need methods to prevent the continued contamination of recommendation system databases by uneven data distributions caused by existing filter bubbles.

3 New Solution: Utilizing Machine Learning Fairness Approaches

To my knowledge, the field of machine learning fairness has been working on such efforts: researchers are focused on developing methods that enable machine learning models built on datasets with uneven feature distributions to produce results that balance both accuracy and fairness, thereby reducing model bias. This aligns with the current needs of recommendation systems. Therefore, applying these methods to recommendation models would undoubtedly help break the filter bubble.

The key distinction between research in machine learning fairness and traditional machine learning approaches lies in its shift away from treating accuracy as the sole performance metric and from treating all features equally. Instead, it focuses on the special optimization of sensitive features. These sensitive features, such as race or gender, typically involve socially marginalized groups, and the goal of fairness approaches is to prevent machine learning models from making biased or unfair decisions about distinct groups because of imbalanced training datasets.

3.1 Fairness Problem Transformation

Typically, machine learning fairness problems define sensitive features along with favorable and unfavorable labels. Fairness metrics are then applied to optimize the machine learning models with respect to these sensitive features.

Fairness metrics

For convenience, encode the terminologies used in measuring model fairness and evaluating the approach. Supposing a given sensitive feature **A**, encode the privileged group to **1**, encode the unprivileged group to **0** as well as encode the favorable label to **1**, unfavorable label to **0**. Then according to the literature [16,17], the fairness metrics can be calculated as follow, where $\hat{Y}$ demotes predicted label and Y denotes real label.

- **EOD** (Equal Opportunity Difference) indicates the TP rate difference between privileged and unprivileged groups.

$$EOD = P[\hat{Y} = 1|A = 0, Y = 1] - P[\hat{Y} = 1|A = 1, Y = 1] \tag{1}$$

- **AOD** (Average Odds Difference) indicates the average of the FP rate difference and the TP rate between privileged and unprivileged groups.

$$\begin{aligned} AOD =& 0.5[(P[\hat{Y} = 1|A = 0, Y = 0] - P[\hat{Y} = 1|A = 1, Y = 0]) \\ &+ (P[\hat{Y} = 1|A = 0, Y = 1] - P[\hat{Y} = 1|A = 1, Y = 1])] \end{aligned} \tag{2}$$

- **SPD** (Statistical Parity Difference) indicates the difference of probabilities of favorable labels gained by privileged and unprivileged groups.

$$SPD = P[\hat{Y} = 1|A = 0] - P[\hat{Y} = 1|A = 1] \tag{3}$$

Applying Fairness Metrics to Recommendation Systems

For a recommendation system that contains a specific filter bubble, sensitive features refer to two or more opposing content where exposure levels differ. The privileged group consists of the content that receives more recommendations, while the unprivileged group is the content that receives fewer recommendations. The favorable label and the unfavorable label are whether the user is interested in (or is likely to click on) the content or not.

Once the sensitive features and favorable/unfavorable labels of the recommendation system are defined, fairness metrics can be applied to the recommendation system, utilizing any machine learning fairness optimization approaches to improve the recommendation model.

3.2 Machine Learning Fairness Approaches

Overview of Machine Learning Fairness Approaches

According to different working mechanisms, existing machine learning fairness approaches can be divided into three classes [18]:

- Pre-processing methods [19,20], which take effects before decision-making algorithms via processing the dataset with data mutation or differentiated sampling;
- In-processing methods [21–23], which optimize the decision-making algorithms themselves via introducing the regular item, Constrained Markov Decision Process (CMDP), or group selection;
- Post-processing methods [24,25], which do not modify the data or decision-making algorithm but adjust the decision results with re-ranking, greedy algorithm, or bias disparity.

Among various mitigation strategies, pre-processing methods are preferred in most cases [26]. Most of the pre-processing techniques rely on balancing the samples in the dataset (using over-sampling, under-sampling, data points mutation, etc.) to alleviate biases [16,19,20,27].

An Example of Machine Learning Fairness Approaches

This section briefly introduces a straightforward machine learning fairness approach, MirrorFair [28].

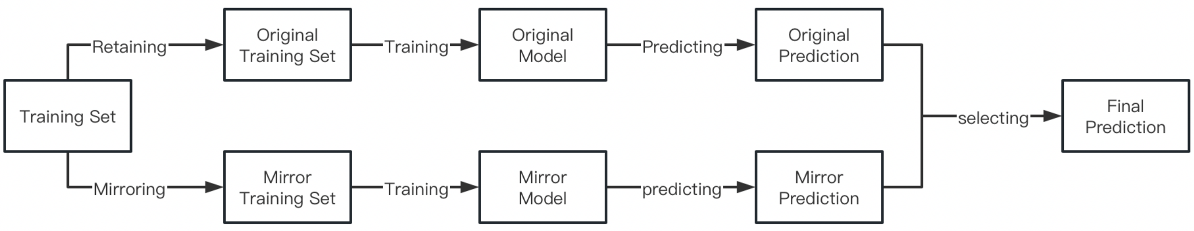

Figure 1: Workflow of MirrorFair.

Figure 1 shows the workflow of MirrorFair. Copy the training set data into two copies, one of which is retained as the original training set, and the other mirrors all the values of sensitive features (i.e., change 0 to 1) as the mirror training set. The original model and the mirror model are obtained by using these two training sets respectively. During the prediction task, the original model and the mirror model are respectively used to predict, and the prediction result with greater confidence probability is selected as the final prediction result.

MirrorFair has been proven to perform well in fairness tasks. This approach can be directly applied to recommendation systems to mitigate the impact of uneven data distribution caused by filter bubbles on the recommendation model.

References

[1] Eli Pariser. 2011. The filter bubble: What the Internet is hiding from you. Penguin.

[2] Jasper Jackson. 2017. Eli Pariser: activist whose filter bubble warnings presaged Trump and Brexit: Upworthy chief warned about dangers of the internet's echo chambers five years before 2016's votes. The Guardian (2017).

[3] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. British Journal of Social Psychology 58, 1 (2019), 129–149.

[4] Harald Holone. 2016. The filter bubble and its effect on online personal health information. Croatian medical journal 57, 3 (2016), 298.

[5] Tim Donkers and Jürgen Ziegler. 2021. The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending. In Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21). Association for Computing Machinery, New York, NY, USA, 12–22. https://doi.org/10.1145/3460231.3474261

[6] Mingkun Gao, Hyo Jin Do, and Wai-Tat Fu. 2018. Burst Your Bubble! An Intelligent System for Improving Awareness of Diverse Social Opinions. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18). Association for Computing Machinery, New York, NY, USA, 371–383. https://doi.org/10.1145/3172944.3172970

[7] Guy Aridor, Duarte Goncalves, and Shan Sikdar. 2020. Deconstructing the Filter Bubble: User Decision-Making and Recommender Systems. In Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20). Association for Computing Machinery, New York, NY, USA, 82–91. https://doi.org/10.1145/3383313.3412246

[8] Yong Liu, Yingtai Xiao, Qiong Wu, Chunyan Miao, Juyong Zhang, Binqiang Zhao, and Haihong Tang. 2020. Diversified Interactive Recommendation with Implicit Feedback. In AAAI '20. 4932–4939.

[9] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. 2021. I Want to Break Free! Recommending Friends from Outside the Echo Chamber. In Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21). Association for Computing Machinery, New York, NY, USA, 23–33. https://doi.org/10.1145/3460231.3474270

[10] Zachary A. Pardos and Weijie Jiang. 2020. Designing for serendipity in a university course recommendation system. In Proceedings of the Tenth International Conference on Learning Analytics and Knowledge (LAK '20). Association for Computing Machinery, New York, NY, USA, 350–359. https://doi.org/10.1145/3375462.3375524

[11] Yuanbo Xu, Yongjian Yang, En Wang, Jiayu Han, Fuzhen Zhuang, Zhiwen Yu, and Hui Xiong. 2020. Neural Serendipity Recommendation: Exploring the Balance between Accuracy and Novelty with Sparse Explicit Feedback. ACM Trans. Knowl. Discov. Data 14, 4, Article 50 (August 2020), 25 pages. https://doi.org/10.1145/3396607

[12] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3460231.3474241

[13] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2023. CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System. ACM Trans. Inf. Syst. 42, 1, Article 14 (January 2024), 27 pages. https://doi.org/10.1145/3594871

[14] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-controllable Recommendation Against Filter Bubbles. In Proceedings of the 45th International ACM SIGIR Conference on Research and

Development in Information Retrieval (SIGIR ’22). Association for Computing Machinery, New York, NY, USA, 1251–1261. https://doi.org/10.1145/3477495.3532075

[15] Zhenyang Li, Yancheng Dong, Chen Gao, Yizhou Zhao, Dong Li, Jianye Hao, Kai Zhang, Yong Li, and Zhi Wang. 2023. Breaking Filter Bubble: A Reinforcement Learning Framework of Controllable Recommender System. In Proceedings of the ACM Web Conference 2023 (WWW ’23). Association for Computing Machinery, New York, NY, USA, 4041–4049. https://doi.org/10.1145/3543507.3583856

[16] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do? In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2021). Association for Computing Machinery, New York, NY, USA, 429–440. https://doi.org/10.1145/3468264.3468537

[17] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. Artifact for "MAAT: A Novel Ensemble Approach to Addressing Fairness and Performance Bugs for Machine Learning Software" https://doi.org/10.5281/zenodo.7553144

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022), 35 pages. https://doi.org/10.1145/3457607

[19] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33, 1 (October 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[20] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML’13). JMLR.org, III–325–III–333.

[21] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’19). Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[22] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM ’19). Association for Computing Machinery, New York, NY, USA, 231–239. https://doi.org/10.1145/3289600.3291002

[23] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18). Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779

[24] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’20). Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/3397271.3401051

[25] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In Proceedings of the Web Conference 2021 (WWW ’21). Association for Computing Machinery, New York, NY, USA, 624–632. https://doi.org/10.1145/3442381.3449866

[26] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020). Association for Computing Machinery, New York, NY, USA, 642–653. https://doi.org/10.1145/3368089.3409704

[27] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020). Association for Computing Machinery, New York, NY, USA, 654–665. https://doi.org/10.1145/3368089.3409697

[28] Ying Xiao, Jie M. Zhang, Yepang Liu, Mohammad Reza Mousavi, Sicen Liu, and Dingyuan Xue. 2024. MirrorFair: Fixing Fairness Bugs in Machine Learning Software via Counterfactual Predictions. Proc. ACM Softw. Eng. 1, FSE, Article 94 (July 2024), 23 pages. https://doi.org/10.1145/3660801