Consumer Fairness Benchmark in Recommendation

Discussion Paper

Ludovico Boratto, Gianni Fenu, Mirko Marras and Giacomo Medda*

Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

Abstract

Several mitigation procedures have emerged to address consumer unfairness in personalized rankings. However, evaluating their performance is difficult due to variations in experimental protocols, such as differing fairness definitions, data sets, evaluation metrics, and sensitive attributes. This makes it challenging for scientists to choose a suitable procedure for their practical setting. In this paper, we summarize our previous work on investigating the properties a given mitigation procedure against consumer unfairness should be evaluated on. To this end, we defined eight technical properties and leveraged two public datasets to evaluate the extent to which existing mitigation procedures against consumer unfairness met these properties. Source code and data: https://github.com/jackmedda/Perspective-C-Fairness-RecSys.

Keywords

Recommender Systems, Consumer Fairness, Mitigation Procedure, Reproducibility, Evaluation Protocol

1. Introduction

With the large adoption of decision-support systems, governments are establishing regulations to account for their trustworthiness. Indeed, it is fundamental to highlight and administer the harmful impacts of artificial intelligence (AI) systems. Recommender systems denote a notable example of systems where trustworthiness and safety are key aspects to be concerned about. In such systems, people are provided with personalized suggestions generated by a certain model [1, 2]. Prior studies have however shown that recommender systems often lead to discriminatory outcomes [3, 4, 5], affecting the entity being ranked or the users the recommendations are targeted to (consumers) [6, 7, 8]. Despite the growing interest in providing fair recommendations to consumers, diverging definitions of consumer fairness have led to unfairness mitigation procedures built on top of heterogeneous evaluation protocols. It is then crucial to discussing which properties a mitigation procedure against consumer unfairness should be evaluated on.

In this paper, we summarize our prior work [9] on building a common ground that can act as a basis for the evaluation of consumer unfairness mitigation procedures. To this end, we defined eight technical properties a given mitigation procedure against consumer unfairness should meet for being effective in practice. We then benchmarked the extent to which existing mitigation procedures meet the defined properties, qualitatively and quantitatively (when possible), on two public data sets. Finally, we gathered the evaluation performance of the mitigation procedures under each property and highlighted the extent to which each procedure meets these properties.

IIR2023: 13th Italian Information Retrieval Workshop, June 8th - 9th, 2023, Pisa, Italy

*Corresponding author.

giacomo.medda@unica.it (G. Medda)

2. Perception of the State of the Art

In this section, we describe the process followed for collecting papers about consumer unfairness mitigation procedures and then categorizing them based on how unfairness was defined, mitigated, and assessed (Table 1). Please refer to the original work [9] for detailed information.
Paper Collection Process. Mitigation procedures against consumer unfairness proposed so far in the literature were collected by scanning Information Retrieval conferences and workshops proceedings as well journals with high impact. Relying on the framework shared by [10] when possible, we reproduced the mitigation procedures proposed in the collected papers.
Fairness Definition Perception. There is no consensus on how to perceive unfairness from a consumer perspective in recommendation. Studies often consider different viewpoints to analyze, mitigate, and evaluate unfairness. Generally, these studies explored fairness notions that mainly address two principles: equity of certain metric scores between demographic groups (EQ); independence of a certain outcome from the sensitive attribute (IND).
Unfairness Mitigation Perception. Studies focusing on fairness from an EQ perspective usually perform a mitigation by balancing the representation of groups in the training set (e.g., [11]), reducing the error across groups (e.g., [6, 16]) or re-ranking items (e.g., [12, 14]). From an IND perspective, unfairness is usually countered by decoupling the user and item latent representations from sensitive attribute information (e.g., [15, 17]) or introducing independence guarantees between the sensitive attribute and the predicted relevance score (e.g., [13]).

3. Research Methodology

In this section, we describe the data sets, sensitive attributes, recommendation models, and the unified evaluation protocol [10] used to benchmark the collected mitigation procedures.
Experimental Data Sets. We selected the two public data sets reported in Table 2, namely ML1M (movies) and LFM1K (music). We considered binarized protected attribute labels (if not already binary, the groups were binarized to have the most similar representation possible).
Recommendation Models. The range of recommendation models evaluated in prior work

Table 1
Consumer unfairness mitigation procedures examined in our study under a top-*n* recommendation task.

Paper	Year	Mitigation			Evaluation		
		Notion [1]	Groups [2]	Type [3]	Data Sets	Utility Metrics	Fairness Metrics
Burke et al. [6]	2018	EQ	G	IN	ML1M	NDCG	CES
Ekstrand et al. [11]	2018	EQ	G	PRE	ML1M-LFM1K	NDCG-MRR	DP
Tsintzou et al. [12]	2019	EQ	G	POST	ML1M-SY	-	BD
Frisch et al. [13]	2021	IND	G-A	IN	ML1M	NDCG	EPS-CHI
Li et al. (a) [14]	2021	EQ	B	POST	AM	NDCG-F1	DP
Li et al. (b) [15]	2021	IND	G-A-O-MS	IN	ML1M-INS	NDCG-HIT	AUC
Wu et al. (b) [16]	2021	EQ	G-A	IN	ML1M-LFM1K	RECALL-NDCG	DP
Wu et al. (a) [17]	2021	IND	G	IN	NEWS	AUC-NDCG-MRR	AUC-F1

[1] **Notion**: Equity (EQ), Independence (IND); [2] **Groups**: Gender (G), Age (A), Occupation (O), Country (C), Marital Status (MS).
[3] **Type**: Pre-Processing (PRE), In-Processing (IN), Post-Processing (POST).
[4] **Data Sets**: MovieLens 1M (ML1M), LastFM 1K (LFM1K), Amazon (AM), Synthetics (SY), [18] (NEWS).

Table 2
The data sets with consumer's sensitive attributes included in our study.

Data Set	#Users	#Items	#Ratings	Sensitive Attributes
ML1M [21]	6,040	3,952	1,000,209	Gender (M : 71.7%; F : 28.3%) Age ($< 35 : 56.6\%; \geq 35 : 43.4\%$)
LFM1K [22]	268	51,609	200,586	Gender (M : 57.8%; F : 42.2%) Age ($< 25 : 57.8\%; \geq 25 : 42.2\%$)

in terms of consumer unfairness was heterogeneous, since no common protocol existed. Our study in this paper focuses on recommendation models considered in at least one prior work.
Evaluation Protocol. For each setup, we obtained the predicted relevance scores and monitored the utility of top-*n* recommendations through NDCG. Unfairness between consumer groups was monitored from an equity (EQ) perspective in terms of NDCG Demographic Parity (DP) [19], computed as the difference in NDCG between the majority group and the minority group (w.r.t. their representation in the data set), and from an independence (IND) perspective by means of a Kolmogorov-Smirnov test (KS) on the predicted relevance scores, as also proposed by [20].

4. Mitigation Procedures Benchmark

In this section, we propose eight key properties to consider while evaluating a mitigation procedure offline, before moving it into practice. Table 3 reports the performance of the recommender systems, before and after unfairness was mitigated, in terms of recommendation utility and fairness between gender groups. Other results can be found in our original study [9].
Applicability. *Indicates the extent to which a mitigation procedure can be technically run on a wide range of different recommendation models without requiring any substantial change to the fundamental steps it is based on.* Pre-processing approaches potentially have a very high *applicability*, while the *applicability* of in-processing and post-processing approaches could be affected by aspects related to the implementation or to the adopted fairness notion.
Coherence. *Indicates the extent to which a mitigation procedure tends to reduce the biased outcomes for the originally disadvantaged group, without reversing the disparate outcome towards the other group(s).* In Table 3, low *coherence* was reported by SLIM-U since applying the mitigation of [6] led to male users being advantaged instead of female users.
Consistency. *Indicates the ability of a mitigation procedure to substantially reduce the model's unfairness according to the pursued fairness notion, given any data set and any consumer grouping method.* Overall, Li et al. [14] was the only *consistent* mitigation procedure across data sets and sensitive attributes under our unified evaluation protocol. Instead, under the papers' original evaluation protocols, no procedure was *consistent* according to our definition.
Data Robustness. *Indicates the ability of a mitigation procedure to reduce unfairness also in challenging cases related to data distribution (e.g., imbalances) and relationships between unfairness and other features.* Our analysis uncovered how leveraging data characteristics causally-related to unfairness, e.g., popularity bias [11], to reduce it could provide better insights on the problem.
Reproducibility. *Indicates the ability of taking the original source code that implements a mitigation procedure and being able to execute it under the same or a different evaluation protocol, with respect to the one used in the original paper.* Our analysis showed that 2 out of 8 papers were

Table 3
[Consistency - Gender Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Orig*) and after mitigating (*Mit*) for gender groups, for key representative mitigation procedures.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
Burke et al.	SLIM-U	0.084	0.084	ˆ0.022	ˆ0.028	ˆ0.032	ˆ0.115	0.320	0.301	*-0.060	ˆ0.072	ˆ0.006	ˆ0.142
Frisch et al.	LBM	0.044	0.021	ˆ0.006	ˆ0.004	ˆ0.013	ˆ0.025	0.144	0.212	*-0.035	*-0.058	ˆ0.120	ˆ0.126
Li et al. (a)	BiasedMF	0.112	0.051	ˆ0.017	**ˆ0.001**	ˆ0.035	**ˆ0.006**	0.287	0.114	ˆ-0.095	ˆ-0.060	ˆ0.012	**ˆ0.001**
	NCF	0.117	0.057	ˆ0.016	**ˆ-0.001**	ˆ0.022	**ˆ0.006**	0.250	0.138	*-0.073	-0.026	ˆ0.033	**ˆ0.001**
	PMF	0.119	0.056	*0.013	ˆ-0.002	ˆ0.023	**ˆ0.006**	0.200	0.071	*-0.062	-0.027	ˆ0.010	**ˆ0.001**
	STAMP	0.022	0.020	***0.003**	ˆ0.003	**ˆ0.006**	**ˆ0.006**	0.160	0.113	-0.021	**0.002**	**ˆ0.001**	**ˆ0.001**
Ekstrand et al.	FunkSVD	0.018	0.015	ˆ0.004	0.002	ˆ0.027	ˆ0.018	0.010	0.013	**-0.006**	-0.003	ˆ0.107	ˆ0.119
	ItemKNN	**0.140**	**0.134**	ˆ0.038	ˆ0.030	ˆ0.030	ˆ0.031	0.287	0.286	ˆ-0.127	*-0.116	ˆ0.019	ˆ0.022
	UserKNN	0.137	0.131	ˆ0.031	ˆ0.024	ˆ0.074	ˆ0.052	**0.406**	**0.411**	ˆ-0.110	ˆ-0.106	ˆ0.067	ˆ0.067

Configurations that resulted in a statistically significant difference in NDCG (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "ˆ" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

not *reproducible*, which limited our work and it remarks the need to sharing the source code.
Scalability. *Indicates the ability of a mitigation procedure to scale well when the number of interactions, users, items, and sensitive attributes, and other relevant features increases consistently.* On data sets with a higher number of entities (e.g., users, interactions), some mitigation procedures (Li et al. and Burke et al. [6, 14]) would lead to unmanageable time and memory requirements.
Trade-off Management. *Indicates the ability of a mitigation procedure to preserve the performance estimate achieved by the target recommendation model originally (before the mitigation was applied).* Overall, Ekstrand et al. [11] reported the best *trade-off* across all the data sets and sensitive attributes. It reduced unfairness, while minimally affecting utility.
Transferability. *Indicates the ability of a mitigation procedure to be effective (and not only applicable) on a wide range of recommendations models, even those it was not originally designed for or tested on.* We applied the mitigation procedures of Ekstrand et al. [11] and Li et al. [14] on the models used by the other papers. Both methods do not hold a good transferability level.
Discussion. As a summary, for each property and mitigation procedure, we assigned one of the two following labels: `Higher` when the corresponding work was better than the others on average for the selected property, `Lower` otherwise. The mitigation procedures proposed by [14, 11] reported the highest number of above-average properties.

5. Conclusions and Future Work

In this work, we collected and reproduced relevant papers addressing consumer unfairness mitigation and categorized them according to the definition, mitigation, and assessment strategy. Then, we defined a unified experimental protocol, including eight technical properties a mitigation procedure should meet, and evaluated the reproduced mitigation procedures on two public data sets on the basis of the defined evaluation properties. Our work allows to have a better understanding of the aspects that could increase the mitigation effectiveness and what can be done to avoid the phenomena outlined by our experiments. Future work will consider novel mitigation procedures able to satisfy all the properties introduced in our paper.

References

[1] M. G. Armentano, A. Monteserin, F. Berdun, E. Bongiorno, L. M. Coussirat, User recommendation in low degree networks with a learning-based approach, in: Mexican International Conference on Artificial Intelligence, Springer, 2018, pp. 286–298.

[2] N. Mauro, L. Ardissono, S. Cocomazzi, F. Cena, Using consumer feedback from location-based services in poi recommender systems for people with autism, Expert Systems with Applications 199 (2022) 116972.

[3] K. Dinnissen, C. Bauer, Fairness in music recommender systems: A stakeholder-centered mini review, Frontiers in Big Data (????) 63.

[4] O. Lesota, A. Melchiorre, N. Rekabsaz, S. Brandl, D. Kowald, E. Lex, M. Schedl, Analyzing item popularity bias of music recommender systems: Are different genders equally affected?, in: Fifteenth ACM Conference on Recommender Systems, 2021, pp. 601–606.

[5] J. Neidhardt, M. Sertkan, Towards an approach for analyzing dynamic aspects of bias and beyond-accuracy measures, in: International Workshop on Algorithmic Bias in Search and Recommendation, Springer, 2022, pp. 35–42.

[6] R. Burke, N. Sonboli, A. Ordonez-Gauger, Balanced neighborhoods for multi-sided fairness in recommendation, in: Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 202–214. URL: http://proceedings.mlr.press/v81/burke18a.html.

[7] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, CoRR abs/2010.03240 (2020). URL: https://arxiv.org/abs/2010.03240. arXiv:2010.03240.

[8] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. A. Pizzato, Multistakeholder recommendation: Survey and research directions, User Model. User Adapt. Interact. 30 (2020) 127–158. URL: https://doi.org/10.1007/s11257-019-09256-1. doi:10.1007/s11257-019-09256-1.

[9] L. Boratto, G. Fenu, M. Marras, G. Medda, Practical perspectives of consumer fairness in recommendation, Inf. Process. Manag. 60 (2023) 103208. URL: https://doi.org/10.1016/j.ipm.2022.103208. doi:10.1016/j.ipm.2022.103208.

[10] L. Boratto, G. Fenu, M. Marras, G. Medda, Consumer fairness in recommender systems: Contextualizing definitions and mitigations, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 552–566.

[11] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, M. S. Pera, All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness, in: Conference on Fairness, Accountability and Transparency, FAT 2018, volume 81, PMLR, 2018, pp. 172–186. URL: http://proceedings.mlr.press/v81/ekstrand18b.html.

[12] V. Tsintzou, E. Pitoura, P. Tsaparas, Bias disparity in recommendation systems, in: R. Burke, H. Abdollahpouri, E. C. Malthouse, K. P. Thai, Y. Zhang (Eds.), Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark,

September 20, 2019, volume 2440 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2440/short4.pdf.

[13] G. Frisch, J. Léger, Y. Grandvalet, Stereotype-aware collaborative filtering, in: M. Ganzha, L. A. Maciaszek, M. Paprzycki, D. Slezak (Eds.), Proceedings of the 16th Conference on Computer Science and Intelligence Systems, Online, September 2-5, 2021, 2021, pp. 69–79. URL: https://doi.org/10.15439/2021F117. doi:10.15439/2021F117.

[14] Y. Li, H. Chen, Z. Fu, Y. Ge, Y. Zhang, User-oriented fairness in recommendation, in: WWW '21: The Web Conference 2021, ACM / IW3C2, 2021, pp. 624–632. URL: https://doi.org/10.1145/3442381.3449866. doi:10.1145/3442381.3449866.

[15] Y. Li, H. Chen, S. Xu, Y. Ge, Y. Zhang, Towards personalized fairness based on causal notion, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 1054–1063. URL: https://doi.org/10.1145/3404835.3462966. doi:10.1145/3404835.3462966.

[16] H. Wu, C. Ma, B. Mitra, F. Diaz, X. Liu, Multi-fr: A multi-objective optimization method for achieving two-sided fairness in e-commerce recommendation, CoRR abs/2105.02951 (2021). URL: https://arxiv.org/abs/2105.02951. arXiv:2105.02951.

[17] C. Wu, F. Wu, X. Wang, Y. Huang, X. Xie, Fairness-aware news recommendation with decomposed adversarial learning, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 4462–4469. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16573.

[18] C. Wu, F. Wu, T. Qi, Y. Huang, X. Xie, Neural gender prediction from news browsing data, in: M. Sun, X. Huang, H. Ji, Z. Liu, Y. Liu (Eds.), Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings, volume 11856 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 664–676. URL: https://doi.org/10.1007/978-3-030-32381-3_53. doi:10.1007/978-3-030-32381-3_53.

[19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021) 115:1–115:35. URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[20] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Recommendation independence, in: Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 187–201. URL: http://proceedings.mlr.press/v81/kamishima18a.html.

[21] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2016) 19:1–19:19. URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.

[22] Ò. Celma, Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space, Springer, 2010. doi:10.1007/978-3-642-13287-2.