Combining Mitigation Treatments against Biases in Personalized Rankings: Use Case on Item Popularity

Discussion Paper

Ludovico Boratto[1], Gianni Fenu[1] and Mirko Marras[2]

[1]University of Cagliari, Cagliari, Italy
[2]EPFL, Lausanne, Switzerland

Abstract

Historical interactions leveraged by recommender systems are often non-uniformly distributed across items. Though they are of interest for consumers, certain items end up therefore being biasedly under-recommended. Existing treatments for mitigating these biases act at a single step of the pipeline (either pre-, in-, or post-processing), and it remains unanswered whether simultaneously introducing treatments throughout the pipeline leads to a better mitigation. In this paper, we analyze the impact of bias treatments along the steps of the pipeline under a use case on popularity bias. Experiments show that, with small losses in accuracy, the combination of treatments leads to better trade-offs than treatments applied separately. Our findings call for treatments rooting out bias at different steps simultaneously.

Keywords

Recommender Systems, Bias, Fairness, Discrimination, Mitigation, Rankings

1. Introduction

Conventionally, recommender systems rank items in order of their decreasing relevance for a given consumer, estimated via machine learning. The literature thus focused on optimizing relevance for consumer's recommendation utility [1]. However, biases such as those against item popularity may emphasize the occurrence of filter bubbles, thus hampering the recommendation quality and several beyond-accuracy aspects [2, 3, 4]. Mitigating the impact of a bias becomes therefore fundamental. Existing treatments for their mitigation are often performed at a single step of the pipeline [5, 6, 7, 8]. Controlling biases by acting only at a single step might lead to sub-optimal trade-offs and, hence, bias-aware pipelines are urging more advanced solutions.

In this paper, we analyze the impact of introducing bias mitigation at different steps of the pipeline simultaneously under a popularity bias scenario, summarizing the findings of our recently published work [9]. Specifically, being two widely-adopted classes of recommendation algorithms highly biased against item popularity, we applied mitigation treatments in pre-processing, through an interaction sampling that balances the training examples where the observed item is more or less popular than the unobserved item, and in in-processing, through a regularization term that minimizes the biased correlation between relevance and item popularity.

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy
ludovico.boratto@acm.org (L. Boratto); fenu@unica.it (G. Fenu); mirko.marras@acm.org (M. Marras)
0000-0002-6053-3015 (L. Boratto); 0000-0003-4668-2476 (G. Fenu); 0000-0003-1989-6057 (M. Marras)

Experiments proved that combining the treatments at two different steps leads to better trade-offs among recommendation quality, beyond-accuracy objectives, and popularity bias reduction.

2. Datasets and Methods

In this section, we introduce the datasets, algorithms, protocols, evaluation metrics, and the pre- and in-processing mitigation treatments adopted in our analyses.

Datasets. Our analyses and experiments were run on two real-world datasets, namely Movielens1M and COCO600k. On one side, MovieLens1M (ML-1M) [10] contains 998,131 ratings applied to 3,705 movies by 6,040 users of the online service MovieLens. The sparsity of the user-item matrix is 0.95, and each user rated at least 20 movies. On the other hand, COCO600k (COCO) [11] includes 617,588 ratings provided by 37,704 learners to 30,399 courses of an online platform. The sparsity of the user-item matrix is 0.99, and each learner rated at least 10 courses.

Algorithms. Our study covers two personalized recommendation algorithms based on a *Point-Wise* [12] and a *Pair-Wise* [13] optimization strategy, respectively. They were chosen due to their performance and wide adoption as a basis of several state-of-the-art recommender systems [14, 15, 16]. Note that our methodology makes it easy to run this analysis on other algorithms.

Protocols. We performed a temporal train-test split with the most recent 20% of ratings per user in the test set and the remaining 80% ones in the training set. User and item matrices are initialized with values uniformly distributed in the range [0, 1]. Each model is served with training batches of 256 examples. For the Point-Wise model, for each user u, we created $t = 4$ unobserved-item examples $((u, j), 0)$ for each observed item $((u, i), 1)$. For the Pair-Wise model, we created $t = 4$ triplets (u, i, j) per observed item i; the unobserved item j is randomly selected. Such parameters were tuned to find a balance between effectiveness and efficiency.

Evaluation Metrics. To assess the quality of the recommended lists, we consider the *Normalized Discounted Cumulative Gain* (NDCG) score [17]. The higher the NDCG score achieved by the recommender system is, the higher the quality of the generated recommendations is for consumers. Conversely, to assess beyond-accuracy objectives, we consider the *Novelty* of the recommendations, computed as the inverse of the average popularity of a recommended item, and the *Coverage* of the catalog, computed as the ratio of items appearing at least once across recommender lists [18]. For both, the higher the score is, the better the objective is met.

To assess popularity bias, we consider the *Item Equal Opportunity* (IEO) score, that encourages the true positive rates of different items to be the same [19]. Under this definition, platform owners may care more about preserving and retaining a certain degree of item popularity, while checking that no further distortions are emphasized by algorithmic bias on recommendation distributions. Therefore, a less biased algorithm tends to recommend each item proportionally to its representation in the ground-truth user preference. If there is a perfect equality of being recommended when items are known to be of interest, then the IEO score is 1. The IEO score decreases towards 0 when the probability of being recommended is high for only few items of interest. This case occurs when most of the niche items never appear in the recommended lists, even if they are of interest (i.e., bias emphasized the popularity phenomenon). Thus, the IEO score ranges between 0 and 1, and the greater it is, the less the popularity bias is emphasized.

Treatment	NDCG				IEO			
	ML-1M		COCO		ML-1M		COCO	
	Pair-Wise	Point-Wise	Pair-Wise	Point-Wise	Pair-Wise	Point-Wise	Pair-Wise	Point-Wise
None	**0.12**	0.10	0.03	0.04	0.07	0.21	0.01	0.05
Sam	0.07	0.06	0.02	0.03	0.16	0.25	0.03	0.12
Reg	0.11	**0.11**	0.01	0.03	0.12	0.19	0.01	0.12
Sam+Reg	0.09	0.08	**0.04**	**0.06**	**0.19**	**0.28**	**0.06**	**0.15**

Table 1
Impact of individual and combined treatments on recommendation quality and item equal opportunity.

Mitigation Treatments. We consider a combination of both pre- and in-processing operations, performing training examples mining and regularized optimization.

Training Examples Mining (sam). Under a point-wise optimization setting, t unobserved-item pairs $((u, j), 0)$ are created for each observed user-item interaction $((u, i), 1)$. The observed interaction $((u, i), 1)$ is replicated t times to ensure that the regularized optimization can work. On the other hand, under a pair-wise optimization setting, for each user u, t triplets (u, i, j) per observed user-item interaction (u, i) are generated. In both settings, the unobserved item j is selected among the items less popular than i for $t/2$ training examples, and among the items more popular than i for the other $t/2$ examples. These operations enable our regularization, as the training examples equally represent both popularity sides associated with the popularity of the observed item, during optimization. We denote training examples as D.

Regularized Optimization (reg). The training examples in D are fed into the original recommendation model in batches $D_{batch} \subset D$ of size m to perform an iterated stochastic gradient descent. Regardless of the family of the algorithm, the optimization approach follows a regularized paradigm derived from the original point- and pair-wise optimization functions. Specifically, the regularized loss function is formalized as a λ-weighted combination of the original accuracy term and a bias mitigation term that aims at minimizing the correlation between (i) the predicted relevance and (ii) the observed-item popularity. The model is penalized if its ability to predict a higher relevance directly depends on the popularity of the item.

3. Experimental Results

In this section, we empirically evaluate the proposed treatments, answering to three key research questions: what are the effects of our mitigation elements separately and jointly (RQ1), what is the impact of our mitigation on internal mechanics (RQ2), and to what extent the treatments impact on beyond-accuracy objectives (RQ3). Experiments are organized below, accordingly.

Effects of Mitigation Elements (RQ1). First, we show an ablation study that aims to assess (i) the influence of the pre-processing sampling and the in-processing regularization on the model performance, and (ii) whether combining these two treatments can improve the trade-off between recommendation quality and popularity bias at a cut-off of 10 (Table 1). On ML-1M, combining our pre- and in-processing mitigation slightly decreased quality. However, the increase in item equal opportunity with respect to no or single treatment is large. On COCO, the combined treatment led to an improvement on both quality and item equal opportunity.

Impact on Internal Mechanics (RQ2). Then, we investigate if combining treatments can

Observed Item	Unobserved Item	ML-1M		COCO	
		Pair-Wise	Point-Wise	Pair-Wise	Point-Wise
Head	Any	0.88 (– 0.05)	0.91 (– 0.04)	0.92 (– 0.02)	0.84 (– 0.14)
Mid	Any	0.78 (+0.04)	0.85 (+0.01)	0.89 (+0.06)	0.82 (– 0.14)
Head	Head	0.77 (– 0.11)	0.87 (– 0.05)	0.89 (+0.00)	0.85 (– 0.11)
Head	Mid	0.93 (– 0.06)	0.95 (– 0.04)	0.95 (– 0.04)	0.83 (– 0.16)
Mid	Head	0.68 (+0.10)	0.80 (+0.06)	0.82 (+0.13)	0.82 (– 0.10)
Mid	Mid	0.89 (– 0.02)	0.90 (– 0.04)	0.94 (– 0.04)	0.81 (– 0.18)

Table 2
Pair-wise accuracy across user-item pairs, after applying our sam+reg approach. Numbers between brackets indicate the difference in percentage-points with respect to the accuracy of the original model.

Beyond-Accuracy Objective	ML-1M		COCO	
	Pair-Wise	Point-Wise	Pair-Wise	Point-Wise
Novelty	0.96 (+ 0.23)	0.96 (+ 0.14)	0.98 (+ 0.02)	0.99 (+ 0.05)
Coverage	0.36 (+ 0.12)	0.91 (+ 0.46)	0.07 (+ 0.56)	0.96 (+ 0.74)
Item Equal Opportunity	0.19 (+ 1.71)	0.28 (+ 0.33)	0.06 (+ 5.00)	0.15 (+ 2.00)

Table 3
Impact of the treatment combination on novelty, coverage, and item equal opportunity. Numbers in brackets indicate the difference in percentage-points w.r.t. the values achieved by the original model.

reduce the biased gap in pair-wise accuracy between head (highly popular) and mid (moderately popular) items. To answer this question, we computed the pair-wise accuracy for different combinations of observed-unobserved head and mid item pairs for the combined treatment in Table 2. The (mid, head) setup experienced a statistically significant improvement in pair-wise accuracy. Conversely, as far as the algorithms end up being well-performing for mid items, pair-wise accuracy on the setups involving observed head items slightly decreased. The improvement is generally higher under a pair-wise optimization (Pair-Wise) and less sparse datasets (ML-1M).

Linking Regularization Weight and Recommendation Quality (RQ3). We investigate finally how the combination of treatments influences beyond-accuracy objectives (novelty and coverage) and popularity bias (item equal opportunity). The results are reported in Table 3. Specifically, the combination ensured large gains in item equal opportunity, higher novelty and a wider coverage. Lower gains on coverage were experienced by the Pair-Wise strategy.

4. Conclusions

In this paper, we analyzed the impact of combining bias mitigation treatments at different steps of the recommendation pipeline under a popularity bias use case. Combining treatments resulted in lower popularity bias at the cost of a negligible decrease in recommendation quality, which confirms the trade-off experienced by other debiasing treatments [7, 20]. Our study also brings forth the discussion about the positive impact of bias mitigation treatments on beyond-accuracy objectives. The findings of this work call for treatments that root out bias at each stage of the recommendation pipeline simultaneously.

References

[1] F. Ricci, L. Rokach, B. Shapira, Recommender systems: Introduction and challenges, in: Recommender Systems Handbook, Springer, 2015, pp. 1–34. doi:10.1007/978-1-4899-7637-6_1.

[2] A. Nematzadeh, G. L. Ciampaglia, F. Menczer, A. Flammini, How algorithmic popularity bias hinders or promotes quality, CoRR abs/1707.00574 (2017). arXiv:1707.00574.

[3] R. Cañamares, P. Castells, Should I follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 415–424. doi:10.1145/3209978.3210014.

[4] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, F. Diaz, Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, ACM, 2018, pp. 2243–2251. doi:10.1145/3269206.3272027.

[5] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Correcting popularity bias by enhancing recommendation neutrality, in: Poster Proceedings of the 8th ACM Conference on Recommender Systems, RecSys 2014, Foster City, Silicon Valley, CA, USA, October 6-10, 2014, volume 1247 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014.

[6] L. Hou, X. Pan, K. Liu, Balancing popularity bias of object similarities for personalised recommendation, European Physical Journal 91 (2018) 47.

[7] H. Abdollahpouri, R. Burke, B. Mobasher, Managing popularity bias in recommender systems with personalized re-ranking, in: Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019, AAAI Press, 2019, pp. 413–418.

[8] H. Abdollahpouri, R. Burke, B. Mobasher, Controlling popularity bias in learning-to-rank recommendation, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017, ACM, 2017, pp. 42–46. doi:10.1145/3109859.3109912.

[9] L. Boratto, G. Fenu, M. Marras, Connecting user and item perspectives in popularity debiasing for collaborative recommendation, Information Processing & Management 58 (2021) 102387. doi:10.1016/j.ipm.2020.102387.

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

[11] D. Dessì, G. Fenu, M. Marras, D. R. Recupero, COCO: semantic-enriched collection of online courses at scale with experimental use cases, in: Trends and Advances in Information Systems and Technologies, volume 746, Springer, 2018, pp. 1386–1396. doi:10.1007/978-3-319-77712-2_133.

[12] X. He, T. Chua, Neural factorization machines for sparse predictive analytics, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, ACM, 2017, pp. 355–364. doi:10.1145/3077136.3080777.

[13] S. Rendle, C. Freudenthaler, Improving pairwise learning for item recommendation from

implicit feedback, in: Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014, ACM, 2014, pp. 273–282. doi:10.1145/2556195.2556248.

[14] Q. Zhang, L. Cao, C. Zhu, Z. Li, J. Sun, Coupledcf: Learning explicit and implicit user-item couplings in recommendation for deep collaborative filtering, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 3662–3668. doi:10.24963/ijcai.2018/509.

[15] Z. Deng, L. Huang, C. Wang, J. Lai, P. S. Yu, Deepcf: A unified framework of representation learning and matching function learning in recommender system, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 61–68. doi:10.1609/aaai.v33i01.330161.

[16] H. Xue, X. Dai, J. Zhang, S. Huang, J. Chen, Deep matrix factorization models for recommender systems, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, ijcai.org, 2017, pp. 3203–3209. doi:10.24963/ijcai.2017/447.

[17] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: http://doi.acm.org/10.1145/582415.582418. doi:10.1145/582415.582418.

[18] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems, ACM Trans. Interact. Intell. Syst. 7 (2017) 2:1–2:42. doi:10.1145/2926720.

[19] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 3315–3323.

[20] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, What recommenders recommend: an analysis of recommendation biases and possible countermeasures, User Model. User Adapt. Interact. 25 (2015) 427–491. doi:10.1007/s11257-015-9165-3.