The Demographics of Cool

Popularity and Recommender Performance for Different Groups of Users

Michael D. Ekstrand and Maria Soledad Pera
People and Information Research Team
Dept. of Computer Science, Boise State University, Boise, Idaho, USA
{michaelekstrand,solepera}@boisestate.edu

ABSTRACT

Typical recommender evaluations treat users as an homogeneous unit. However, user subgroups often differ in their tastes, which can result more broadly in diverse recommender needs. Thus, these groups may have different degrees of satisfaction with the provided recommendations. We explore the offline top-N performance of collaborative filtering algorithms across two domains. We find that several strategies achieve higher accuracy for dominant demographic groups, thus increasing the overall performance for the strategy, without providing increased benefits for other users.

CCS CONCEPTS

• **Information systems → Recommender systems**;

KEYWORDS

collaborative filtering, evaluation popularity bias

1 INTRODUCTION

Recommender system evaluation—offline and online —typically focuses on the system's effectiveness, in aggregate over the entire user population. While individual user characteristics are sometimes taken into account, as in demographic-informed recommendation, evaluations typically still aggregate over all users [8]. In this work, we connect recent work leveraging user demographics to deepen understanding of different users' satisfaction with search engines [7], with the work of Bellogin et al. [1] measuring recommenders' performance for different *items* to examine recommender system accuracy for users in different demographic groups in an offline setting. This attention is necessary because, by default, the largest subgroup of users will dominate overall statistics; if other subgroups have different needs, their satisfaction will carry less weight in the final analysis. This can result in an incomplete picture of the performance of the system and and obscure the need to identify how to better serve specific demographic groups. To the well-known problems of *popularity bias* [2] and *misclassified decoys* [3, 5] (a good item recommendation counted as a error given that the user has yet to interact with the item in available data), we add a third consideration: *demographic bias*, where the satisfaction (approximated in offline settings by top-N accuracy) of some demographic groups is weighted more heavily than others. Demographic bias also has a complex expected interaction with popularity bias: the most active and numerous users will have a greater impact on popularity than other users, so popularity bias in evaluation will further encourage the selection of algorithms that perform well on the largest subgroup's tastes.

Our central research question is this: what changes about our assessment of relative or absolute recommender effectiveness when we consider performance for different subgroups of users– basically when we consider all subgroups' satisfaction to be equally important? Does popularity bias exacerbate demographic bias effects? How do popularity bias mitigations affect the demographic bias?

2 INITIAL ANALYSIS

We answer these questions with an offline analysis using LensKit [4] [1] and two **datasets** that provide user demographics of some form. *MovieLens-1M*[2] [6] contains 1M 5-star ratings of 3,900 movies by 6,040 users who joined MovieLens through 2000. Each user has self-reported age, gender, occupation, and zip code. *LastFM* contains data of 359,347 users who played 294,015 unique artists. The main record set consists of 17,559,530 tuples of the form $\langle user, artist, playCount\rangle$. For most users, gender, age, country, and sign-up date are provided. We employed several classical and widely-used collaborative filtering **algorithms**: (1) *Popular* (Pop), recommending the most frequently rated or played items; (2) *Item-Item* (II), an item-based collaborative filter using 20 neighbors and cosine similarity; (3) *User-User* (UU), a user-based collaborative filter configured to use 30 neighbors and cosine similarity; and (4) *FunkSVD* (MF), which is based on gradient descent matrix factorization technique with 40 latent features and 150 training iterations per feature. Each algorithm is tagged with its variant: '-E' are explicit-feedback recommenders (applicable only to MovieLens); '-B' are implicit-feedback recommenders that only consider *whether* an item was rated or played, disregarding its rating value or play count; '-C' are implicit-feedback recommenders that consider the number of times an artist was played as repeated implicit feedback (LastFM only). We applied 5-fold **cross-validation**, using two methods: (1) LensKit's default strategy and (2) Bellogin's UAR method [1] for neutralizing popularity bias; this works like the default, except it picks test sets of items instead of users. An initial experiment revealed that regardless of the **metric**, i.e., Recall, Mean Reciprocal Rank (MRR), and Mean Average Precision, the algorithms exhibit similar behavior, thus we report our results using MRR.

Demographic distribution and its impact on evaluation. Figure 1 shows user gender distribution; with the majority of users reporting as male. The age distribution reveals some differences: the largest block of MovieLens users belong to the [25-35] group, whereas a plurality of LastFM users belong to the [18-24] group.[3]

RecSys 2017 Poster Proceedings, August 27–31, Como, Italy.

[1] Code and scripts are available at https://doi.org/10.18122/B2ND8P
[2] Later MovieLens dataset do not include demographic information.
[3] For consistency, we binned LastFM users into the same groups used in MovieLens-1M.

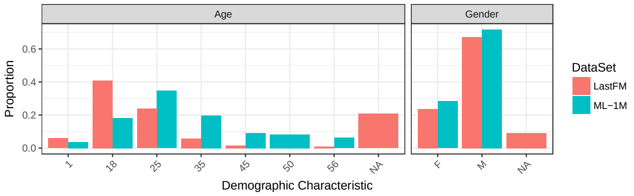

Figure 1: User distribution based on age and gender

Standard Results. Figure 2 shows the MRR achieved by each algorithm, grouped by demographic group. For each demographic characteristic, *All* is the accuracy achieved by averaging across all users, and *Bucketed* is the result of first averaging within each demographic group, and then averaging the groups' results (thus giving each group equal weight, instead of each user). The results across subgroups are broadly similar for both data sets, though the *All* analysis tracks most closely with the dominant group. However, if a decision is to be made based on "performs best", then the small differences become non-trivial, as they will affect the final decisions. One example case emerges from our analysis: on LastFM, II performs better using play counts ("-C") for some age groups, while the "-B" variant is more effective for other age groups.

While we cannot conclude, based on this ongoing study, *which* is the right decision, our preliminary analysis demonstrates the need for further exploration from a demographic perspective.

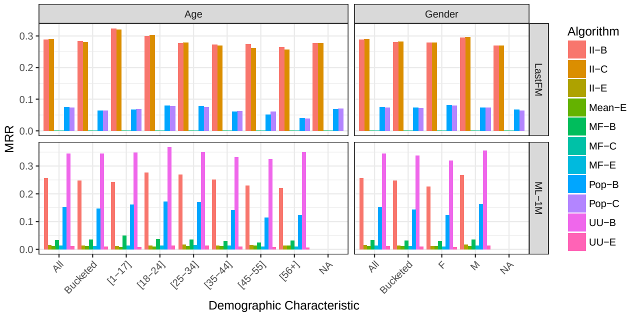

Figure 2: Results of basic run of results

Popularity Bias Mitigating Results. We also seek to understand how demographic bias interacts with mitigation techniques for other issues, such as popularity bias. To that end, we performed a version of our analysis using Bellogin's UAR technique [1]. We see (in Figure 3) that several of the smaller user groups have substantially *higher* accuracy measures than larger groups, particularly on age. An analysis using this method would find that the recommender is delivering better recommendations to these groups.

The differences obtained using UAR or traditional evaluations show that mitigating popularity bias comes with the cost of significantly changing the distribution of measured accuracy across user subgroups. (Analysis using 1R [1] did not produce results significantly different from Figure 2.) Which evaluation strategy better reflects actual user experience is still up for debate.

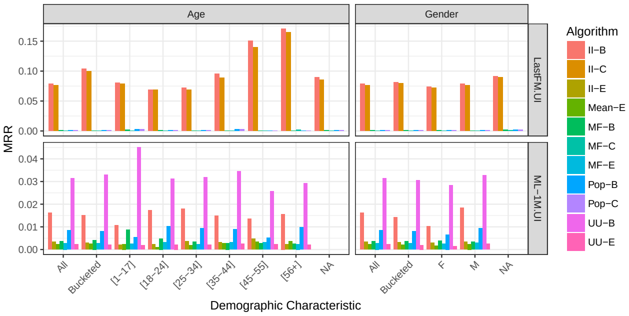

Figure 3: Results of UAR experiment

3 DISCUSSION AND FUTURE WORK

Our analysis showed that, unsurprisingly, a number of recommendation strategies achieve moderately higher accuracy metric values for dominant demographic groups. This can cause an algorithm's performance to increase without delivering benefit to smaller subgroups of the user population. In other words, the perceived satisfaction with a recommender may not be the same for the "cool" users—in the dominant group—as it is for those in smaller groups.

Demographic bias in accuracy metric results also has a complex interaction with mitigation strategies for other offline evaluation ailments such as popularity bias. A uniform item strategy results in disproportionately higher accuracy values for users in some smaller subgroups. Further work is needed to understand which paradigm maps most closely to actual user experience or response.

Our findings highlight the need for careful and multi-faceted consideration of recommender system behavior across a range of both users and items. As prior work has found that recommenders are not equally good at recommending for all items, we find that recommenders are not equally good for all users in predictable and socially-relevant ways. While the full social and business ramifications of our findings have yet to be explored, we encourage researchers and practitioners to pay attention to which users receive how much benefit from a particular recommender.

ACKNOWLEDGMENTS

We thank Ion Madrazo for helping with analysis, and the People and Information Research Team (PIReT) for their support.

REFERENCES

[1] A. Bellogin. *Performance prediction and evaluation in Recommender Systems: an Information Retrieval perspective.* PhD thesis, UAM, 2012.
[2] A. Bellogin, P. Castells, and I. Cantador. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proc. ACM RecSys '11*, 2011.
[3] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *ACM RecSys*, pages 39–46, 2010.
[4] M. D. Ekstrand, M. Ludwig, J. A. Konstan, and J. T. Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and lenskit. In *Proc. ACM RecSys '11*, 2011.
[5] M. D. Ekstrand and V. Mahant. Sturgeon and the cool kids: Problems with Top-N recommender evaluation. In *Proc. FLAIRS 30.* AAAI Press, 22 May 2017.
[6] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Trans. Interact. Intel. Sys.*, 5(4):19, 2016.
[7] R. Mehrotra, A. Anderson, F. Diaz, A. Sharma, H. Wallach, and E. Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proc. WWW '17 Companion*, 2017.
[8] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.