How Sensitive is Recommendation Systems' Offline Evaluation to Popularity?

Amir H. Jadidinejad, Craig Macdonald, Iadh Ounis
University of Glasgow

ABSTRACT

Datasets used for the offline evaluation of recommender systems are collected through user interactions with an already deployed recommender system. However, such datasets can be subject to different types of biases including a system's popularity bias. In this paper, we focus on assessing the influence of popularity on the offline evaluation of recommendation systems. Our insights from a deeper analysis based on popularity-stratified sampling reveal that the current offline evaluation of recommendation systems are sensitive to popular items, raising questions about conclusions driven from the offline comparison of recommendation models.

1 INTRODUCTION

The offline evaluation of recommendation systems includes (1) gathering a collection of user's interactions from a deployed system and (2) the use of these interactions to evaluate and compare different recommendation models. The first stage can be subject to popularity bias (or any other type of biases), either through popular item selection by the users or through the actions of the deployed system [2] (i.e. popular items are over-represented by the deployed systems from which feedback datasets were collected).

User's interactions in recommendation systems are reminiscent of clickthrough data in Information Retrieval (IR); there is a strong relationship between the top ranked documents presented to the user, and those for which the system receives clicks (a.k.a. presentation bias). As a result, while the users' feedback can provide *relative* preferences among the displayed documents, it does not necessarily reflect retrieval quality [3]. While researchers in IR proposed several approaches to compensate for presentation bias such as pooling when forming test collections, the evaluation of recommendation systems continues to be often conducted using a held-out test set [1] obtained through random sampling from feedback datasets, which cannot capture the actual utility of models [2, 4].

In this paper, we focus on investigating the extent to which the offline evaluation of recommendation systems is sensitive to popularity bias in the observed user's feedback which was collected from a deployed system. We propose a method based on *popularity-stratified sampling* to evaluate the effectiveness of a given recommendation model across a range of items with various popularity levels.

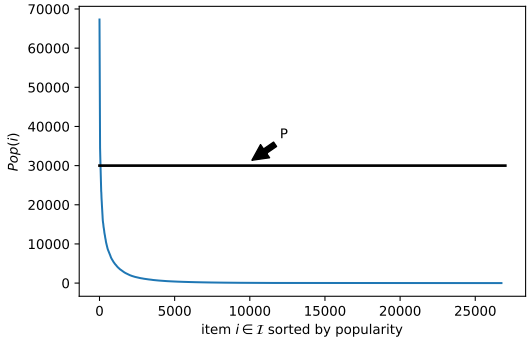

Figure 1: Popularity-stratified sampling on the MovieLens dataset.

2 POPULARITY-STRATIFIED SAMPLING

We formalize our strategy to evaluate a set of recommendation models based on popularity strata. Our aim is to sample different sub-populations of user-item interactions based on various levels of popularity. For a specific item $i \in \mathcal{I}$, we define $Pop(i)$ as the number of times that item i has been interacted with by different users $u \in \mathcal{U}$. Figure 1 shows the expected Power Law distribution of popularity values $Pop(i)$ for different items $i \in \mathcal{I}$ in the MovieLens dataset.

Definition 1. For a specific threshold P we can randomly sample a fixed number of user-item interactions $\mathcal{T}_P$ with $Pop(i)$ lower than P. This sub-population contains user's interactions with less popular items depending on the threshold P, i.e. $\mathcal{T}_P \leftarrow \{\mathcal{U} \times \mathcal{I} : Pop(i) < P\}$.

By varying the threshold P we can sample different sub-populations $\mathcal{T}_P$ from the whole user-item interactions with different levels of popularity. The popularity-stratified sampling method allows to neglect the interactions of a very few most popular items concerning each P threshold. For example, by setting the P threshold to 30k as shown in Figure 1, we neglect only the 63 most popular items among $26,745$ total items (0.24%) and randomly sample $\mathcal{T}_{30k}$ from the remaining items' interactions (99.76%).

In our experiments, we evaluate the following 5 models using nDCG@20 on the MovieLens-20m dataset: *Random* is a simple baseline model that recommends a random item among all available items; *Popularity* is a stronger baseline model that suggests the top-20 most popular items to each

Table 1: The offline evaluation of different models based on 20% held-out samples and nDCG@20. All differences are significant according to the paired t-test ($p < 0.001$).

DataSet	Random	Popularity	MF	BPR	WARP
MovieLens	0.0007	0.067	0.06	0.068	**0.076**

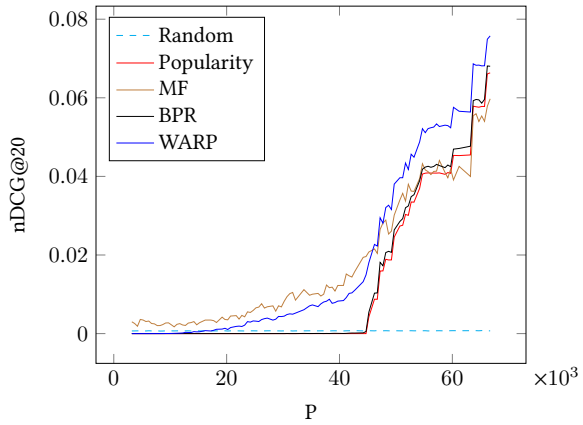

Figure 2: The performances of different models are evaluated based on held-out less popular samples $\mathcal{T}_P$ (Definition 1). By setting the P threshold to maximum, the evaluation of models based on $\mathcal{T}_{P_{max}}$ corresponds to the offline systems' evaluation presented in Table 1.

and every user regardless of their preferences; *Matrix Factorization (MF)* is a classical rating prediction model that represents both users and items as latent vectors; *Bayesian Personalized Ranking (BPR)* is a pair-wise ranking model that is trained based on uniform negative sampling; *Weighted Approximate-Rank Pairwise (WARP)* is a pair-wise ranking prediction model. Unlike BPR, the negative items are chosen among those negative items which would violate the desired ranking given the current state of the model. For reproducibility, our code and the used datasets are released at http://github.com/amirj/recsys_eval.

Typically, the offline evaluation of recommendation systems takes a set of *random* samples ($\mathcal{T}$) of *all* interactions between users and items ($\mathcal{U} \times \mathcal{I}$) as a held-out test set [1]. Table 1 shows the obtained performances of the 5 models on the held-out test set ($\mathcal{T}$) in terms of nDCG@20. We firstly note the low effectiveness of the well-established MF model compared to Popularity as a non-personalized baseline. On the other hand, WARP and BPR are shown to be the best performing systems. Therefore, based on these experiments we can conclude the following *relative* performances between these models:

$$\text{WARP} \overset{\mathcal{U}\times\mathcal{I}}{\gg} \text{BPR} \overset{\mathcal{U}\times\mathcal{I}}{\gg} \text{Popularity} \overset{\mathcal{U}\times\mathcal{I}}{\gg} \text{MF} \overset{\mathcal{U}\times\mathcal{I}}{\gg} \text{Random}.$$

On the other hand, Figure 2 shows the performance of the same 5 models on $\mathcal{T}_P$ by varying the P threshold (Definition 1). Comparing the *relative* performances of these models in Figure 2 reveals that the the relative performances of these models does not remain robust across different P thresholds, i.e. varying the popularity threshold leads to a different relative performance. For example, when $P < 45 \times 10^3$ the performance of MF is significantly *higher*[1] than WARP while for $P \approx 45 \times 10^3$, the performances of MF and WARP are *comparable*. Finally, for $P > 45 \times 10^3$, the performance of MF is significantly *lower* than WARP. On the other hand, the relative performance of some models remains robust across different popularity thresholds. For example, the performance of WARP is *always* better than BPR for all values of $\mathcal{T}_P$. The main role of a recommendation model is to recommend *personalized* items for each user. Because of the highly skewed distribution of user-item interactions towards popular items, any algorithm that favours popular items (e.g. Popularity), might be *over-represented* in the offline evaluation of recommendation systems. Therefore, popularity appears to play a key factor in the offline evaluation of recommendation models and appears to influence conclusions that could be drawn from the relative comparison of models.

3 CONCLUSIONS

We investigated the influence of popularity on the offline evaluation of recommendation systems. Our findings provided an experimental explanation of the general trend observed in the recent literature [1, 2, 4], showing that indeed the examined models are sensitive to popularity, i.e. random sampling from different popularity strata can considerably impact offline evaluation and subsequent conclusions. It is our hope that this work will help motivate researchers and practitioners to propose new approaches to assess recommendation systems by taking into account imbalanced data across a wide range of users and items. In addition, although popularity is a well-known feature in recommendation datasets, it is plausible that a deployed system might favour other groups of items (e.g. high profit items or suppliers). We further plan to measure the effect of other types of closed-loop feedback [2] caused by the deployed systems.

ACKNOWLEDGMENTS

The authors acknowledge support from EPSRC grant EP/R018634/1 entitled Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics.

[1] Significance tests are not shown in Figure 2 for clarity.

REFERENCES

[1] Rocío Cañamares and Pablo Castells. 2018. Characterization of Fair Experiments for Recommender System Evaluation – A Formal Analysis. In *Workshop on Offline Evaluation for Recommender Systems (REVEAL 2018) at the 12th ACM Conference on Recommender Systems*.

[2] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*.

[3] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2011. *Evaluating Search Engine Relevance with Click-Based Metrics*. Springer Berlin Heidelberg.

[4] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased Offline Recommender Evaluation for Missing-not-at-random Implicit Feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*.