Beyond Static Calibration

The Impact of User Preference Dynamics on Calibrated Recommendation

Lin, Kun; Mansoury, Masoud; Eskandanian, Farzad; Sabouri, Milad; Mobasher, Bamshad

DOI
10.1145/3631700.3664869

Publication date
2024

Document Version
Final published version

Published in
UMAP Adjunct '24

Beyond Static Calibration: The Impact of User Preference Dynamics on Calibrated Recommendation

Kun Lin
DePaul University
Chicago, USA
klin13@depaul.edu

Masoud Mansoury
Delft University of Technology
Delft, Netherlands
m.mansoury@tudelft.nl

Farzad Eskandanian
DePaul University
Chicago, USA
feskanda@depaul.edu

Milad Sabouri
DePaul University
Chicago, USA
msabouri@depaul.edu

Bamshad Mobasher
DePaul University
Chicago, USA
mobasher@cs.depaul.edu

ABSTRACT

Calibration in recommender systems is an important performance criterion that ensures consistency between the distribution of user preference categories and that of recommendations generated by the system. Standard methods for mitigating miscalibration typically assume that user preference profiles are static, and they measure calibration relative to the full history of user's interactions, including possibly outdated and stale preference categories. We conjecture that this approach can lead to recommendations that, while appearing calibrated, in fact, distort users' true preferences. In this paper, we conduct a preliminary investigation of recommendation calibration at a more granular level, taking into account evolving user preferences. By analyzing differently sized training time windows from the most recent interactions to the oldest, we identify the most relevant segment of user's preferences that optimizes the calibration metric. We perform an exploratory analysis with datasets from different domains with distinctive user-interaction characteristics. We demonstrate how the evolving nature of user preferences affects recommendation calibration, and how this effect is manifested differently depending on the characteristics of the data in a given domain. Datasets, codes, and more detailed experimental results are available at: https://github.com/nicolelin13/DynamicCalibrationUMAP.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommender systems, calibration, preference dynamics

ACM Reference Format:

Kun Lin, Masoud Mansoury, Farzad Eskandanian, Milad Sabouri, and Bamshad Mobasher. 2024. Beyond Static Calibration: The Impact of User Preference Dynamics on Calibrated Recommendation. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 01–04, 2024, Cagliari, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3631700.3664869

UMAP Adjunct '24, July 01–04, 2024, Cagliari, Italy

ACM ISBN 979-8-4007-0466-6/24/07
https://doi.org/10.1145/3631700.3664869

1 INTRODUCTION

Recommender systems learn from past user preferences in order to predict future user interests and provide users with personalized suggestions. Traditionally, these systems are evaluated using accuracy metrics like precision, recall, and normalized discounted cumulative gain (NDCG), measuring the relevancy of the recommended items delivered to the users. Optimizing recommender systems for accuracy, however, may result in an imbalance in the distribution of preference categories in recommendations: the recommendations may only reflect a user's dominant preferences in certain categories and ignore less dominant interests while still being considered accurate. This type of imbalance, in the long term, can lead to undesirable effects such as the "rabbit hole" effect [9, 17] or filter bubbles [8].

To address this phenomenon, the notion of *Calibration* in machine learning has been extended to the realm of recommender systems [12]. The calibration metric measures the degree to which a user's prior preference are reflected in the recommendations generated by the system. For example, in a hypothetical movie recommender system, a user whose past preferences suggest 70% interest in Action movies and 30% in Drama movies, should ideally see the same genre ratios in her set of recommended movies. Calibration is typically measured as the correlation between the distribution of preference categories in a user's profile and that of preference categories in generated recommendations.

To improve calibration (or reduce miscalibration), most existing approaches assume that users' preferences distribution is fundamentally static. To measure calibrations, these approaches typically utilize the full history of items in users' profiles to identify preference distributions. This may result in presumably calibrated recommendations that do not truly reflect users' current preferences. The problem is that user preferences evolve over time. If a user's profile is collected over a long period of time, only certain segments of that profile may represent the relevant interests of users.

For example, consider a user interacting with a book platform for a period of 10 years. For the first 7 years, the user was intensely interested in books in the spy-thriller category. But 3 years ago, her interests shifted primarily to historical novels and then, in the last two years, to science fiction. Using the standard approach

to measuring calibration, a recommender that recommends items from all three categories based on the user's full history might be considered *highly calibrated*. However, the fact is that the user may not perceive these recommendations as reflective of her current preferences. On the other hand, if a recommender only recommends items from the historical and science fiction categories (i.e., more current user's preferences), then the generated recommendations may be considered to be *miscalibrated*.

Recognizing that a key contributor to miscalibration may be the dynamic nature of users' preferences, we hypothesize that selectively utilizing the most relevant segments from users' profiles for training the recommendation model can result in recommendations that better represent users' current interests. Our work attempts to shift the focus from predominantly post-processing methods for recalibrating recommendations to incorporating calibration directly into the training phase of the recommendation process, promising a more fundamental solution to the challenge of maintaining relevance in the face of evolving user preferences.

In this preliminary work, we conduct an exploratory analysis of calibration in recommendation systems at a more granular level, taking into account the dynamics of user preferences. We empirically identify relevant segments of users' profiles through a data-driven study of two different datasets with different user interaction characteristics. Our goal is to show how the existing approach of measuring calibration that is based on static user profiles can be misleading and fails to precisely measure the degree to which recommendations reflect current user interests. To do this, we propose a preprocessing simulation process that identifies the most representative segment of the users' profile by measuring changes in the calibration metric as user preferences shift over time.

In this simulation process, we first split the users' profiles into a number of segments, with the size of segments determined empirically based on the frequency of user interactions with items. Then, from the most recent segments to the oldest, we iteratively combine the segments to obtain subsamples of the data. Finally, on each subsample, we train a recommendation model and measure recommendation calibration. This process returns the most relevant segment of a user's profile that results in the highest calibration with respect to the current user preference distribution.

Our initial experiments with two datasets from distinctive domains show that our proposed simulation process enables measuring calibration more precisely based on users' changing preferences. We hope that this preliminary work will lead to more extensive studies of how the dynamics of user preferences in different domains should be reflected in the design of calibration mechanisms in recommender systems.

2 BACKGROUND AND RELATED WORK

We denote $\mathcal{U} = \{u_1, ..., u_N\}$ and $\mathcal{I} = \{i_1, ..., i_M\}$ as the set of N users and M items in the system. We show a user u's profile as I_u which contains all interacted items by u. R_u is the recommendation list delivered to u which contains all items recommended to u. Each item is assigned one or more categories $C = \{c_1, ..., c_W\}$, with W categories in total.

2.1 Calibration Metric in Recommendation

The user's preference distribution over item categories is represented as a numerical vector of size W that shows the preference ratio of a user toward each item category. We denote this vector for a user u as $\mathcal{P}_u = \{p(c|u)|c \in C\}$ where $p(c|u)$ is defined as follows:

$$p(c|u) = \frac{\sum_{i \in I_u} p(c|i)}{|I_u|} \tag{1}$$

where $p(c|i)$ is the fraction that category c represents item i (e.g., if i is assigned c_1 and c_2, then $p(c_1|i) = 0.5$ and $p(c_2|i) = 0.5$). Analogously, we define category distribution on the recommendation list delivered to user u, denoted by $\mathcal{Q}_u\{q(c|u)|c \in C\}$, where $q(c|u)$ is defined as follows:

$$q(c|u) = \frac{\sum_{i \in R_u} p(c|i)}{|R_u|} \tag{2}$$

Given the category distribution over the user's profile and recommendation list, miscalibration is computed by measuring the distance between these two distributions. There are various distance measures to consider for this computation, but following [12], we use Kullback-Leibler (KL) divergence in this paper:

$$MC(\mathcal{P}, \mathcal{Q}) = KL(\mathcal{P}||\mathcal{Q}) = \sum_{c \in C} p(c|u) \log \frac{p(c|u)}{q(c|u)} \tag{3}$$

2.2 Methods for Calibrated Recommendation

Steck [12] initially defined calibration emphasizing its role in achieving personalization that mirrors a user's past behavior. This concept has evolved, intersecting significantly with fairness and diversity [1, 15].

Current calibration methods [2, 3, 11], such as those critiqued by [6, 7], often separate personalization from calibration, leading to a potential reduction in recommendation accuracy. Additionally, these methods typically assume static user preferences, neglecting the dynamic nature of user interests, as highlighted by Zhao et al. [17].

Building upon the insights from previous studies, this research proposes a novel approach to address miscalibration in recommender systems. Traditional calibration methods often rely on entire historical data, where older interactions may no longer reflect the user's latest preferences, leading to recommendations that are out of sync with current interests. By selectively utilizing the most relevant segments of users' profiles for training the model, our approach aims to incorporate calibration directly into the training phase, taking into account the evolving nature of user preferences.

3 SIMULATION PROCESS FOR DYNAMIC CALIBRATION

In this section, we describe our pre-processing simulation process for identifying the most representative segments of users' profiles in training the recommendation model that would yield more calibrated recommendations. Figure 1 illustrates this simulation process as follows:

(1) Given the user-item interaction matrix containing the profile of all users and their interacted items, we sort each user u's profile P_u chronologically in descending order and split P_u into n subprofiles as $\{P_u^1, P_u^2, ..., P_u^n\}$ where P_u^1 contains

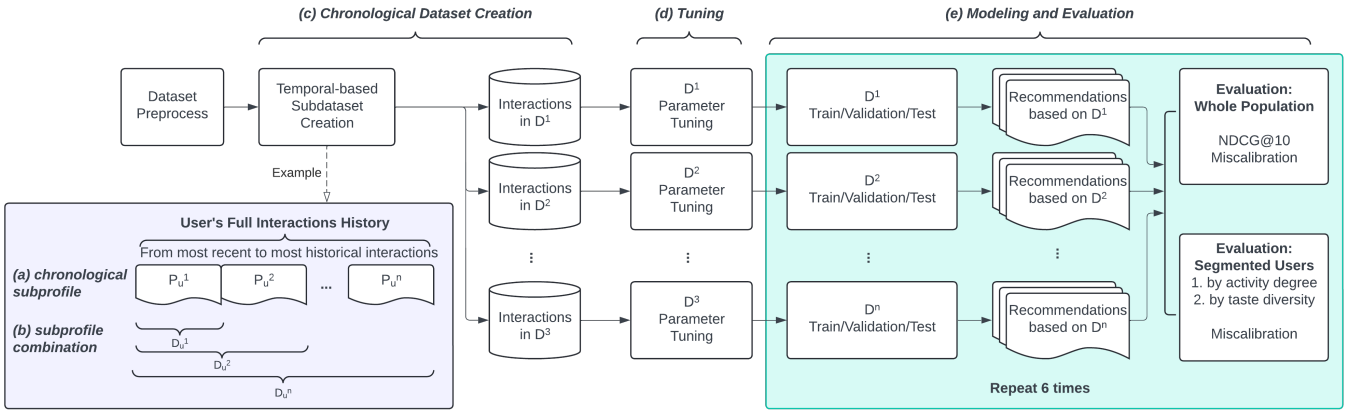


Figure 1: Experiment workflow for dynamic calibration

the most recent u's interactions and P_u^n contains the oldest interactions. This process is shown in part (a) of Figure 1. The choice of n (number of subprofiles) is a domain-specific parameter. The time *window size* for creating the subprofiles should be set based on how frequently users interact with items. For example, in a short video streaming platform where users interact with items very frequently, a time window of shorter length, like daily, may be appropriate. However, in a book recommendation platform, a time window of a longer period, such as one or more years, might better capture users' evolving tastes. We discuss and analyze the choice of the time window and n on different recommendation domains in the methodology section.

(2) As shown in part (b) of Figure 1, we create samples of the dataset D by iteratively and chronologically combining the subprofiles of users from different time windows as follows:

$$D^l = \{P_u^1 \cup P_u^2 \cup ... \cup P_u^l | \forall u \in \mathcal{U}\}, \quad where \;\; l \leq n \tag{4}$$

This will create samples of the dataset (i.e., $D^1, D^2, ..., D^n$) as shown in part (c) of Figure 1.

(3) We iteratively pick each sample created in the previous step and evaluate the recommendation performance built on that sample. As shown in part (d) of Figure 1, we first tune the recommendation model on each sample of the data for the best model. Then, in part (e) of Figure 1, with the best model identified in the previous step, we evaluate the recommendation performance (i.e., miscalibration and accuracy) on each sample of the dataset.

When considering the different segments of the user's profile, the most recent segment of the users' profiles (i.e., D^1, D^2, ..., or D^n) that yields the lowest miscalibration is the one that is most representative of users' current preferences.

4 METHODOLOGY

4.1 Datasets and Preprocessing

We use two datasets from distinct domains. The short video platform is characterized by rapid user engagement and high interaction frequencies, where user preferences evolve quickly. For the book domain, users tend to transition between different book genres more gradually. To better investigate the calibration dynamics with the shifts in preference patterns, we preprocess datasets to focus on active users across time windows of different sizes. Detailed statistics of processed datasets are provided in Table 1.

KuaiRec dataset [4] is collected on the short video mobile app Kuaishou[1] from July to September 2020. Each video is manually tagged with video categories. With watch_ratio as the user's preference signal, we refer to its mean (0.9) as the threshold defining the user's interest. Positive interactions ($watch\_ratio \geq 0.9$) are labeled as "1" with others as "0". Users in this dataset are very active. 86.4% of users are active for 60 out of the total 63 days, while 99.9% are active for more than 50 days. In our future work, we will include an analysis of the larger matrix from KuaiRec with higher sparsity which reflects a more realistic interaction scenario.

The book recommendation dataset [13, 14] from GoodReads[2] includes user-book interactions. Each book has its assigned book genres. To reduce dataset size and sparsity, we focused on interactions associated with more engaged users and frequently accessed items. The dataset was filtered to include only active users, defined as those engaging in 20 to 50 annual interactions from year 2010 to 2017. Subsequently, items with fewer than 1,000 interactions were excluded. The final step filtered in users with a consistent activity level of more than 4 interactions per year.

4.2 Tuning Time Windows Representing User Preference Segments

The time window selection over user profile is critical when creating users' subprofiles in Figure 1 part (a). In this exploratory work, we are interested in observing changes in calibration when extending the training time window sizes. Given that goal, we look for the smallest window size, leading to sufficient interactions for training and representing user preferences. Due to differences in user interaction frequency and evolving preferences across domains, time window sizes vary across domains. Users in the book reading domain have less frequent interactions and more stable

[1] https://www.kuaishou.com/cn
[2] https://mengtingwan.github.io/data/goodreads.html

Table 1: Statistics of the Preprocessed Datasets.

Dataset	#Users	#Items	Category type	#Categories	#Interactions	Sparsity
KuaiRec	1,411	3,327	content tag	31	1,799,403	61.6%
GoodReads	865	1,662	book genre	7	104,325	92.7%

preferences than users in the short video domain. After tuning the time window sizes for the two datasets (for KuaiRec, 1/7/14 days; for GoodReads, 3/6/12 months), we picked 1 day for KuaiRec and 0.5 years (6 months) for GoodReads.

4.3 Recommendation Models

Recommendations are generated by well-tuned Bayesian Personalized Ranking (BPR) [10][3], implemented using the library RecBole [16]. With NDCG@10 as objective, hyperparameters are tuned via an exhaustive search for learning rate ranging from 0.0001 to 0.01 and embedding size between 64 to 128. Models for different time periods are independently tuned as shown in Figure 1 part (d). Also, we repeated the experiments six times to ensure the stability and accountability of the results. The results for analysis are based on the aggregated results from the repeated experiments.

4.4 User Segment Analysis

Given that users' varied behavioral patterns in the same dataset may be related to different degrees of calibration across individuals, we extend our analysis of calibration dynamics to user segments with different characteristics. We explore two factors for segmenting users: user activity frequency (number of user interactions) and category-wise entropy in the user profile. We simply used a percentile-based approach, segmenting users into three groups based on each factor. A higher entropy indicates that user preferences are more evenly distributed across categories, while users with lower entropy have more intense preferences over a few specific categories. [5] showed user profile entropy is an important user profile factor positively associated with the miscalibration of the received recommendation.

5 RESULTS AND DISCUSSION

5.1 Calibration Dynamics Due to Time Window Selections

Fig. 2 depicts the miscalibration dynamics with the time window changing. Regardless of domains, we observe that miscalibration can vary based on the time window selections. In KuaiRec, the effect is more pronounced due to the highly dynamic nature of user preferences, with the optimal calibration achieved with a window size of 5 days. It is noteworthy that this window size also corresponds to the optimal NDCG values. In GoodReads, the calibration peaked at time window 7, indicating the most recent 3.5 years. In this domain, user preferences tend to be more stable over time, so there is less variance in calibration values. We note, however, that in this case, recommendation accuracy increases as larger profile segments are used in training.

5.2 Calibration Dynamics Based on User Characteristics

For our analysis of the impact of user segment characteristics on calibration dynamics, we empirically obtain the optimal time window for calibration as before, but we do so for each user segment instead of the full set of users. Fig. 3 (a) and (b), depict the changes in miscalibration as window size increases for the three user segments (with (a) showing results for user segments based on interaction frequency and (b) depicting results for user segments based on profile entropy). These results suggest that in this domain there are no significant variances in calibration across different user segments, especially when considering smaller time windows. This is likely due to the highly dynamic and unstable nature of user preferences. As before, we still see an optimal window size of 5 days for all user segments.

Fig. 3 (c) and (d) show a very different picture for the GoodReads domain than that of KuaiRec. The optimal time window across user segments is not exactly consistent, but within the overall range that we saw for all users. Nevertheless, the results indicate that in this domain, it makes sense to consider different window sizes for distinct user segments to achieve optimal calibration. However, the more notable observation is that for user segments with a high activity level, a higher overall calibration is achieved regardless of window size. We conjecture that this is because users with higher activity in this domain (users who read books often) tend to have more stable profiles with little or no change in preference patterns over time. This is also the case for user segments with low profile entropy. These are users who tend to have very focused interests in a few genres, and so their preference patterns tend to be quite stable.

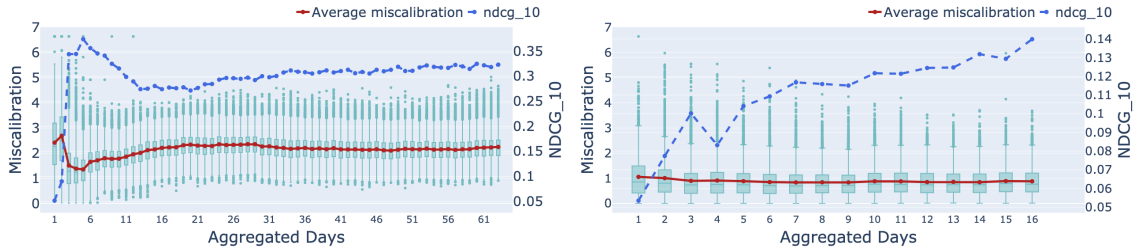
Overall, these results support the conjecture that the degree of calibration in recommendation depends on the dynamics of user preferences, with more stable preference profiles resulting in a higher degree of calibration.

6 CONCLUSION AND FUTURE WORK

In this exploratory analysis, we have shown that the calibration mechanism in recommender systems must take into account the dynamic nature of user preferences. Our simulation-based analysis shows that the degree of miscalibration decreases as the training time window size changes, focusing on the most relevant segments of the users' profiles. Our analysis also shows that user interaction characteristics associated with different domains lead to different optimal training window sizes to achieve better calibration.

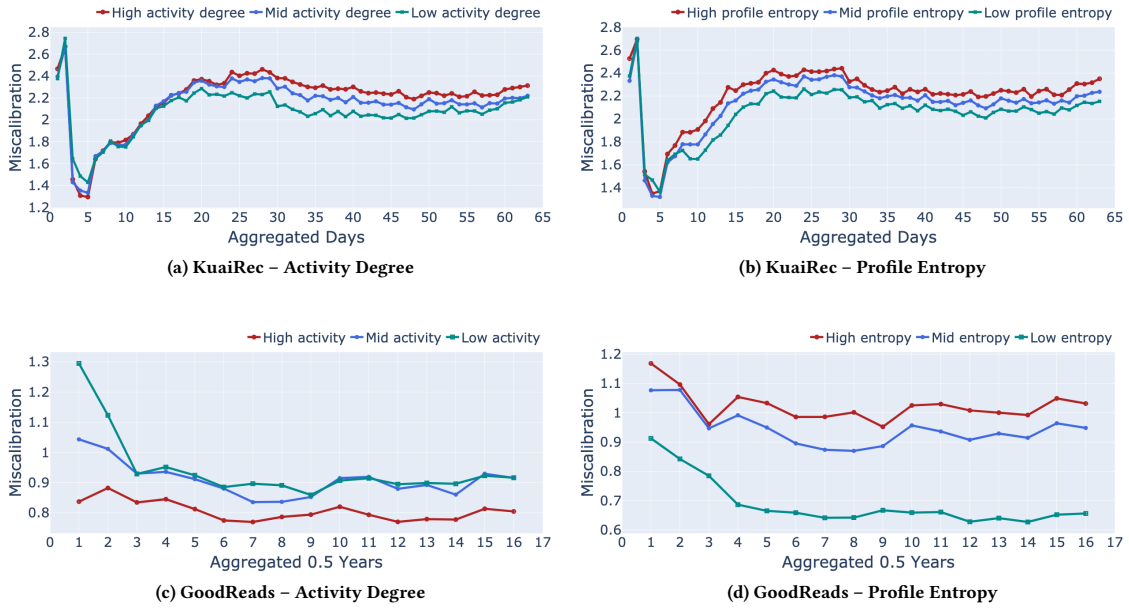
In future work, we extend this preliminary analysis to include comparison with post-processing calibration approaches, and also

[3]Our full preliminary experiments include two algorithms: BPR and ItemKNN. Given the results between the two algorithms are similar, we focus on BPR in the main paper while making ItemKNN results available via https://github.com/nicolelin13/DynamicCalibrationUMAP.



(a) KuaiRec (b) GoodReads

Figure 2: Box plot of miscalibration distribution by time window



(a) KuaiRec – Activity Degree (b) KuaiRec – Profile Entropy

(c) GoodReads – Activity Degree (d) GoodReads – Profile Entropy

Figure 3: Miscalibration by user segments. The top row presents results based on KuaiRec, and the bottom row is based on GoodReads. For each row, the left is segmented by activity level, and the right is segmented by profile entropy.

to explore trade-offs between dynamic calibration and other beyond-accuracy metrics such as diversity and fairness. We will also conduct an analytical investigation on the impact of domain and user preference patterns on miscalibration dynamics. Currently, we have only considered two domains with a limited approach to segmenting users and provided preliminary conclusions based on these observations. In future work, we will do a more comprehensive investigation involving additional data sets with distinct characteristics. We will also perform an analysis of how different classes of recommendation algorithms, including those based on session-based, context-aware, and sequence-aware recommendation models, perform using the dynamic approach to calibration.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 726–731.

[2] Himan Abdollahpouri, Zahra Nazari, Alex Gain, Clay Gibson, Maria Dimakopoulou, Jesse Anderton, Benjamin Carterette, Mounia Lalmas, and Tony Jebara. 2023. Calibrated recommendations as a minimum-cost flow problem. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 571–579.

[3] Farzad Eskandanian and Bamshad Mobasher. 2020. Using stable matching to optimize the balance between accuracy and diversity in recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 71–79.

[4] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-Observed Dataset and Insights for Evaluating Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. 540–550. https://doi.org/10.1145/3511808.3557220

[5] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2020. Calibration in collaborative filtering recommender systems: a user-centered analysis. In

Proceedings of the 31st ACM Conference on Hypertext and Social Media. 197–206.

[6] Mohammadmehdi Naghiaei, Hossein A Rahmani, Mohammad Aliannejadi, and Nasim Sonboli. 2022. Towards confidence-aware calibrated recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4344–4348.

[7] Zahra Nazari, Praveen Chandar, Ghazal Fazelnia, Catherine M Edwards, Benjamin Carterette, and Mounia Lalmas. 2022. Choice of implicit signal matters: Accounting for user aspirations in podcast recommendations. In *Proceedings of the ACM Web Conference 2022*. 2433–2441.

[8] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. 677–686.

[9] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2015. Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review* 33, 4 (2015), 459–478.

[10] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[11] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic multi-aspect fairness through personalized re-ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 239–247.

[12] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.

[13] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*. 86–94.

[14] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. *arXiv preprint arXiv:1905.13416* (2019).

[15] Chenyang Wang, Yankai Liu, Yuanqing Yu, Weizhi Ma, Min Zhang, Yiqun Liu, Haitao Zeng, Junlan Feng, and Chao Deng. 2023. Two-sided Calibration for Quality-aware Responsible Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 223–233.

[16] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *proceedings of the 30th acm international conference on information & knowledge management*. 4653–4664.

[17] Xing Zhao, Ziwei Zhu, and James Caverlee. 2021. Rabbit holes and taste distortion: Distribution-aware recommendation with evolving interests. In *Proceedings of the Web Conference 2021*. 888–899.