Computational Versus Perceived Popularity Miscalibration in Recommender Systems

Oleg Lesota
oleg.lesota@jku.at
Johannes Kepler University Linz and Linz Institute of Technology
Linz, Austria

Gustavo Escobedo
gustavo.escobedo@jku.at
Johannes Kepler University Linz
Linz, Austria

Yashar Deldjoo
deldjooy@acm.org
Polytechnic University of Bari
Bari, Italy

Bruce Ferwerda
bruce.ferwerda@ju.se
Jönköping University
Jönköping, Sweden

Simone Kopeinik
skopeinik@know-center.at
Know-Center GmbH
Graz, Austria

Elisabeth Lex
elisabeth.lex@tugraz.at
Graz University of Technology
Graz, Austria

Navid Rekabsaz
navid.rekabsaz@jku.at
Johannes Kepler University Linz and Linz Institute of Technology
Linz, Austria

Markus Schedl
markus.schedl@jku.at
Johannes Kepler University Linz and Linz Institute of Technology
Linz, Austria

ABSTRACT

Popularity bias in recommendation lists refers to over-representation of popular content and is a challenge for many recommendation algorithms. Previous research has suggested several offline metrics to quantify popularity bias, which commonly relate the popularity of items in users' recommendation lists to the popularity of items in their interaction history. Discrepancies between these two factors are referred to as popularity miscalibration. While popularity metrics provide a straightforward and well-defined means to measure popularity bias, it is unknown whether they actually reflect users' perception of popularity bias.

To address this research gap, we conduct a crowd-sourced user study on Prolific, involving 56 participants, to (1) investigate whether the level of perceived popularity miscalibration differs between common recommendation algorithms, (2) assess the correlation between perceived popularity miscalibration and its corresponding quantification according to a common offline metric. We conduct our study in a well-defined and important domain, namely music recommendation using the standardized LFM-2b dataset, and quantify popularity miscalibration of five recommendation algorithms by utilizing Jensen–Shannon distance (JSD). Challenging the findings of previous studies, we observe that users generally *do* perceive significant differences in terms of popularity bias between algorithms if this bias is framed as popularity miscalibration. In addition, JSD correlates moderately with users' perception of popularity, but not with their perception of unpopularity.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591964

KEYWORDS

recommender systems, music recommendation, user study, popularity bias, popularity calibration, miscalibration, metrics, ecological validity

ACM Reference Format:
Oleg Lesota, Gustavo Escobedo, Yashar Deldjoo, Bruce Ferwerda, Simone Kopeinik, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2023. Computational Versus Perceived Popularity Miscalibration in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591964

1 INTRODUCTION

Popularity bias in recommender systems refers to over-exposure of popular content in recommendation lists, which can be introduced or amplified by common recommendation algorithms [16, 29]. This bias can be harmful to both content creators (lack of exposure for lesser known producers) and consumers (lack of diversity) [1, 5, 13].

A common way to investigate popularity bias — and the one we follow in the work at hand — is to relate the popularity of the items in the recommendation list of a user to the popularity of the items that the user interacted with in the past. This method adopts a user- or consumer-centric perspective, *calibration*, originally proposed for genre [28], and later adapted to popularity [16]. The main assumption of calibration is that users prefer recommendations of similar popularity as in their interaction history. The discrepancy between item popularity in previous interactions and recommendation lists is considered a *popularity miscalibration* of the recommendation algorithm or model. Previous research (e.g., [1, 13, 16]), has examined the extent of popularity bias in different recommendation algorithms and domains via well-defined offline metrics. However, the fundamental question of the ecological validity of those metrics has largely been neglected so far. In the present work, we turn the

viewpoint and investigate users' *perception* of popularity miscalibration in recommendation lists and its concordance with a commonly used offline metric, i.e., Jensen–Shannon distance. Conducting a user study, we obtain through questionnaires human assessments of popularity miscalibrations and compare them to offline measures, and strive to answer the following research questions: **[RQ1]** Do users perceive differences in popularity miscalibration in recommendation lists created by different algorithms? **[RQ2]** Does users' perception of popularity miscalibration correlate with a corresponding common bias metric?

2 RELATED WORK

Investigating and mitigating algorithmic popularity bias in recommender systems has been the target of several research works, e.g. [1, 2, 4–7, 9, 11, 13, 16, 18, 29]. To debias recommendations, algorithms or recommendation lists are often optimized to increase user or producer fairness based on an (offline) popularity bias metric, assuming that such offline metrics accurately reflect human perceptions of bias and fairness. However, several recent studies argue that what computational bias and fairness metrics quantify may significantly differ from users' perceptions of bias and fairness [21, 24, 27]. Most closely related to the present work, Ferwerda et al. [8] conduct an online study with $N = 170$ participants recruited through Amazon Mechanical Turk and investigate the concordance of mathematical notions of popularity bias and perceived popularity. They find that the participants rarely notice popularity bias in the recommendation lists produced by various collaborative filtering-based music recommender systems. In a broader sense, Smith et al. [25] ask $N = 30$ participants of an exploratory interview study about their ideas and understanding of the meaning of fair treatment in recommender systems. Among several overarching themes, they identify that the respondents strive for more provider fairness and less accuracy when confronted with use cases with harmful personalization effects on particular stakeholders. Similarly, Sonboli et al. [26] perform a face-to-face, semi-structured interview study to investigate user perceptions of recommender systems, as well as users' opinions about fairness and fairness-aware objectives of recommender systems. Their study aims to provide an explanation method design for fairness-aware recommender systems. In the context of information retrieval, Krieg et al. [14, 15] study the effect of societal biases on the users' perception in respect to relevance judgment, while Kopeinik et al. [12] investigate how such biases are reflected in the way the users of a search engine formulate search queries. The work at hand differs from previous research by studying *popularity miscalibration* and the perception of users in this regard, in contrast to approaching a general notion of popularity as, for instance, done in Ferwerda et al. [8].

3 METHODOLOGY

To answer the research questions, we conduct a user study implemented as a web survey on the crowd-sourcing platform Prolific.[1] In the following, we describe the user study design, recommendation algorithms, bias metric, and statistical methodology.

User study design: The study uses a within-subject design with algorithm (five variations) and user demographics as independent variables. As dependent variables, we consider the participant's assessment of music items and lists. The study procedure is further structured in two parts. First, participants provide their demographic information (age, gender, country) and answer questions regarding their music expertise (e.g., playing an instrument, writing in a music-related blog) and experience with music recommender systems. Second, each participant receives five personalized music recommendation lists of ten items each (i.e., track title and artist/band name are shown, no listening required). Each list is produced by a different recommendation algorithm, where the order of algorithms is selected randomly. Participants are asked to evaluate each recommendation list with respect to popularity bias in two ways: (1) on a per-item basis and (2) considering the entire list as one entity. On the per-item level, users need to indicate if they recognize the track (yes/no), and in case they do, specify whether they consider the item too popular, too unpopular, or within their usual taste. On the list level, participants answer the following questions on a 5-point Likert scale ranging from "Disagree strongly" (coded as 1) to "Agree strongly" (coded as 5): **(q1)** "The list of recommendations matches my preferences.", **(q2)** "The list contains a lot of popular items.", **(q3)** "The list contains a lot of unpopular items.".

Generating recommendations: To create personalized recommendation lists for each user, we first pretrain the investigated models (see below) on a subsample of the public LFM-2b dataset [23] containing listening events of Last.fm users. The subsample comprises approximately 8M unique interactions of 18K users with 200K items sampled uniformly at random over all tracks interacted within 2018-2021. For each participant we create five recommendation lists of 10 items (one per model) by (1) mapping the tracks in the participant's Last.fm listening history to the LFM-2b dataset, (2) feeding the resulting user's interaction history into each of the trained models.

Capturing user perception: We derive and analyze four indicators of participant perception of the recommendation lists. Three are based on questions participants answer about the recommendation lists, as introduced above. Question 1 means to retrieve a general attitude towards each list and acts as a handle for accuracy. Questions 2 and 3 respectively are tailored to the evaluation of popularity bias in the recommendation lists. The fourth indicator is derived from per-track evaluations. The value of the indicator is calculated as the number of tracks, which the user marked as "too popular" or "too unpopular" for their taste. In this way, each list gets an inferred score between 0 and 10 (in this study, we investigate user perception of popularity miscalibration regardless of its actual nature; hence, both too popular and too unpopular tracks are treated similarly). For all four indicators, we consider mean raw score over all users to infer algorithm ranking in RQ1 (and raw scores themselves for significance tests). Analyzing correlations, in RQ2, we convert raw scores given to the lists by each user into rankings from 1 to 5 (giving identical ranks to lists tied by the score).

Recommendation algorithms: We create item lists using three popular recommendation algorithms and two algorithms we specifically devised to address the task at hand: (1) item-based k-nearest neighbors (`ItemKNN`) [22] is a collaborative filtering approach that considers an item as relevant for the target user u if it is similar to the items interacted with by u (in terms of items' interaction history over users); (2) sparse linear method (`SLIM`) [20] factorizes the item–item co-occurrence matrix and uses the learned item coefficients to sparsely aggregate past user interactions and recommend

[1] https://www.prolific.co/

new items; (3) multinomial variational autoencoders (`MultVAE`) [17] learns latent user representations from their item interactions optimized for reconstructing each user's original interactions, and recommends new items based on the output probability distribution over all items for a given target user; (4) `Genre-MostPop` and (5) `Genre-LeastPop` are two personalized variants of popularity-based algorithms. The former selects the items of the target user's five most frequent genre in their listening history and recommends, respectively, the most or least popular items that the user has not yet interacted with. To assign genre(s) to each track, we retrieve the list of user-generated tags from Last.fm[2] and index them using the Discogs genre taxonomy of the AcousticBrainz Genre Dataset.[3]

Popularity miscalibration measurement: To approach RQ2, we adopt a popular metric for popularity miscalibration [3, 16] on the user side, i.e., Jenson-Shannon distance (*JSD*), according to Equation 1 where H_u denotes the popularity (probability) distribution of a user u's consumption history and R_u the popularity (probability) distribution of u's personalized recommendation list. There $H_u(c)$ is the proportion of items of popularity category c in the consumption history of user u. *JSD* can be seen as symmetrical version of Kullback–Leibler divergence. Note that using log_2 we ensure that the value of *JSD* is bound between 0 and 1 [19]. *JSD* reflects the degree of mismatch (miscalibration) between the two distributions, with higher values meaning stronger mismatch.

$$JSD(H_u, R_u) = \frac{1}{2}\sum_c H_u(c) log_2 \frac{2H_u(c)}{H_u(c)+R_u(c)} + \frac{1}{2}\sum_c R_u(c) log_2 \frac{2R_u(c)}{H_u(c)+R_u(c)} \quad (1)$$

Following established practice [3, 16], we define the popularity category of each item based on the number of interactions with it. We distinguish popular, unpopular, and average categories of items. Popular items are the ones users most interacted with, which jointly attract 20% of all user-item interactions. Similarly, unpopular items are the least interacted with, receiving 20% of accumulated user-item interactions. The rest of items falls into the category average. This way, considering miscalibration of popularity distributions, we look at the discrepancy between two three-bin distributions, with each bin corresponding to one of the popularity categories.

Statistical methodology to address the research questions:

RQ1: We rank the five algorithms based on mean values of the four user perception indicators (Likert scale answers to the three per-list questions and aggregated per-track answers) computed over all participants (mean opinion score of each indicator separately). We then test significance of the ranking by performing paired t-tests between the highest scoring algorithm and the four lower ranking models. We interpret the results taking into account Bonferroni correction, therefore treating results at $p <= 0.0125$ as significant (as we perform four comparisons we divide the threshold α by 4). Assessing algorithm ranking based on per-track evaluations, we filter the test set, assuring that every participant is familiar with at least 5 of 10 recommended items in each list.

RQ2: We again limit the scope of the analysis to users familiar with at least half of recommended tracks. In order to minimize the influence of each user's evaluation bias, we convert scores they assign within each question to ranks, resolving ties in a dense manner. The offline evaluation (*JSD*) is converted to ranks as well (separately for each user). We then analyze Spearman's correlation between each indicator and *JSD* separately, as well as between the four list level indicators.

[2]https://www.last.fm/api/show/track.getTopTags
[3]https://mtg.github.io/acousticbrainz-genre-dataset/data_stats

Table 1: Algorithm ranking according to the accuracy-based indicator

q1: The list of recommendations matches my preferences			
Algorithm	q1-mean ↑	t-stat	p-value
`SLIM`	4.000000	0.000000	0.000000
`ItemKNN`	3.982143	-0.100633	0.920208
`MultVAE`	3.821429	-1.182700	0.242016
`Genre-MostPop`	3.767857	-1.753499	0.085088
`Genre-LeastPop`	2.535714	-7.483274	0.000000

Table 2: Algorithm ranking according to the indicator of over-representation of popular items

q2: The list contains a lot of popular items			
Algorithm	q2-mean ↑	t-stat	p-value
`MultVAE`	4.303571	0.000000	0.000000
`Genre-MostPop`	4.142857	-1.085190	0.282571
`SLIM`	3.803571	-2.924988	0.004996
`ItemKNN`	3.446429	-4.096298	0.000139
`Genre-LeastPop`	2.089286	-10.801745	0.000000

Table 3: Algorithm ranking according to the indicator of over-representation of unpopular items

q3: The list contains a lot of unpopular items			
Algorithm	q3-mean ↑	t-stat	p-value
`Genre-LeastPop`	3.821429	0.000000	0.000000
`ItemKNN`	2.714286	-5.131796	0.000004
`SLIM`	2.339286	-7.053822	0.000000
`Genre-MostPop`	2.214286	-7.021594	0.000000
`MultVAE`	1.946429	-9.574271	0.000000

4 RESULTS AND DISCUSSION

The data analysis is conducted for $N = 56$ accepted study participants (i.e. those who finished the study in at least 10 minutes and passed all attention checks) with a Male/Female/Diverse: 65%/31%/4% gender distribution and mean/median/SD: 26.9/25/8.19 values for age. We present main results regarding RQ1 in Tables 1, 2, and 3. Each table shows the ranking of the investigated recommender algorithms according to one of the three question-based indicators of user perception. We report the mean of the raw scores given to each algorithm by the users in the respective columns (such as q1-mean). Higher values of mean raw score denote an inclination towards "Agree strongly", lower values correspond to the inclination towards "Disagree strongly". The results of significance tests between the top-ranking algorithm and the rest are reported in columns "t-stat" and "p-value". Main results for RQ2 are reported

in Tables 4 and 5. Table 4 shows correlation between four different indicators of user perception of recommendation lists (three based on questions about the whole list and one inferred from aggregation of per-track evaluations) and investigated popularity bias metric *JSD*. In Table 5, we report the correlation between different user judgments (indicators), namely question-based indicators against aggregated track-wise evaluation. For RQ2, we limit our scope to users familiar with at least half of items in every recommended list. We also exclude Genre-LeastPop from this part of analysis as only 11 participants have shown sufficient familiarity with lists provided by it. This leaves us with 35 users and 4 algorithms giving 140 data points to calculate correlation over.

Addressing RQ1, we do not find a significant difference between SLIM and ItemKNN, MultVAE, Genre-MostPop in terms of general matching personal preference (q1), Table 1. This is expected as the former three are SOTA models, and the latter is a strong baseline enhanced with personalized filtering by genre. We find, however, that participants agree on Genre-LeastPop matching their preference worse than the others (in particular, significantly worse than SLIM). Analyzing Table 2, we see MultVAE as perceived to be prone to suggesting popular items to a similar degree as Genre-MostPop, and significantly more prone than SLIM, ItemKNN, and Genre-LeastPop. Taking into account the significant difference between SLIM and Genre-LeastPop (not reported in the table), we conclude that the users perceive different degrees to which recommendation lists are populated with mainstream items. Ranking algorithms with correspondence to q3 shown in Table 3, we see another confirmation of user perception of popularity miscalibration being gradual. Genre-LeastPop is perceived as significantly more prone to recommending unpopular items in comparison to the other algorithms. We also observe that the participants can agree on a ranking similar to Table 2 with regard to aggregated perceived per-track miscalibration (fourth indicator). Our observations show that the users perceive different degrees of popularity miscalibration and can agree on a ranking of algorithms with respect to certain evaluation criteria (q2 and q3).

Approaching RQ2, we first note that the ranking of models inferred from q2, Table 2 is in line with previous work [16] showing that SLIM and ItemKNN demonstrate a similar degree of user-side popularity bias according to offline metrics, as well as MultVAE and Genre-MostPop. Analyzing correlations between user perception indicators and the offline metric *JSD*, Table 4, we do not find a significant correlation with utility-oriented q1, showing that recommendation lists (mis-)calibrated with user taste in terms of popularity do not necessarily satisfy their other preferences. We note a significant correlation with q2, which shows that *JSD* is receptive to cases when users perceive recommendation lists as dominated by popular items. This justifies the use of the offline metric for the evaluation of popularity bias on the user side. *JSD* also shows a significant correlation with aggregated perceived per-track miscalibration, serving as an additional confirmation of its validity. We do not observe significant correlation with q3, showing that *JSD* may be not the best choice for detecting cases when a recommendation list is overpopulated with unpopular items ('unpopularity' bias). Finally, we investigate agreement within indicators, in particular correlation of track-wise evaluation with question-based evaluations, Table 5. We observe significant correlation with q2 and absence of such correlation with q3, hinting that users are more receptive to over-representation of popular items.

Table 4: Spearman's correlation between four different user perception indicators and offline poularity bias metric *JSD*. Significant results in bold.

q1: The list of recommendations matches my preferences			
Corr. coeff.	0.029767	p-value	0.726993
q2: The list contains a lot of popular items			
Corr. coeff.	**0.4435**	p-value	0.000000
q3: The list contains a lot of unpopular items			
Corr. coeff.	-0.104542	p-value	0.218983
Aggregated perceived per-track miscalibration			
Corr. coeff.	**0.321041**	p-value	0.00011

Table 5: Spearman's correlation between question-based perceptual indicators and track-wise indicator. Significant results in bold.

q1: The list of recommendations matches my preferences			
Corr. coeff.	-0.06425	p-value	0.450739
q2: The list contains a lot of popular items			
Corr. coeff.	**0.327158**	p-value	0.000080
q3: The list contains a lot of unpopular items			
Corr. coeff.	-0.145327	p-value	0.086671

5 CONCLUSION AND FUTURE WORK

We conducted a user study to confront perceptual and computational indications of popularity miscalibration in recommendation lists. Through a series of experiments, we found that (1) users perceive various degrees of popularity miscalibration between different recommender algorithms; (2) users rank algorithms according to popularity miscalibration in line with previous offline evaluations [16], showing that MultVAE and Genre-MostPop, are perceived as being more biased than ItemKNN and SLIM, which in turn are perceived as more biased than Genre-LeastPop; (3) offline popularity bias metric Jensen-Shannon distance shows significant correlation with user perception of lists dominated by popular items, but not with perception of lists dominated by unpopular items. As opposed to [8], we obtain correlations between perceived popularity miscalibration and offline metrics, indicating that users can perceive popularity bias when using their own taste as a reference point.

Future work includes (1) studying the direction of miscalibration (positive versus negative popularity bias), (2) analyzing additional recommendation algorithms, (3) conducting experiments in domains other than music, (4) the influence of individual characteristics (e.g., [10]), and (5) investigating whether providing users with different explanations about popularity bias and miscalibration influences their perception of bias.

ACKNOWLEDGMENTS

This research is funded in whole, or in part, by the Austrian Science Fund (FWF): DFH-23 and P33526; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grants LIT-2020-9-SEE-113 and LIT-2021-YOU-215.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. In *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 (CEUR Workshop Proceedings, Vol. 2440)*, Robin Burke, Himan Abdollahpouri, Edward C. Malthouse, K. P. Thai, and Yongfeng Zhang (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2440/paper4.pdf

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura (Eds.). ACM, 726–731. https://doi.org/10.1145/3383313.3418487

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 119–129. https://doi.org/10.1145/3450613.3456821

[4] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2021. A Flexible Framework for Evaluating User and Item Fairness in Recommender Systems. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 457–511. https://doi.org/10.1007/s11257-020-09285-1

[5] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in Recommender Systems: Research Landscape and Future Directions. *User Modeling and User-Adapted Interaction* (2023). https://doi.org/10.1007/s11257-023-09364-z

[6] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.

[7] Mehdi Elahi, Himan Abdollahpouri, Masoud Mansoury, and Helma Torkamaan. 2021. Beyond algorithmic fairness in recommender systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 41–46.

[8] Bruce Ferwerda, Eveline Ingesson, Michaela Berndl, and Markus Schedl. 2023. I Don't Care How Popular You Are! Investigating Popularity Bias From a User's Perspective. In *Proceedings of the 8th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023)*. ACM, Austin, USA.

[9] Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. Unlearning Protected User Attributes in Recommendations with Adversarial Training. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2142–2147. https://doi.org/10.1145/3477495.3531820

[10] Mark P Graus and Bruce Ferwerda. 2021. The Moderating Effect of Active Engagement on Appreciation of Popularity in Song Recommendations. In *International Conference on Information*. Springer, 364–374.

[11] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. Mitigating Popularity Bias in Recommendation: Potential and Limits of Calibration Approaches. In *Advances in Bias and Fairness in Information Retrieval - Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers (Communications in Computer and Information Science, Vol. 1610)*, Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo (Eds.). Springer, 82–90. https://doi.org/10.1007/978-3-031-09316-6_8

[12] Simone Kopeinik, Martina Mara, Linda Ratz, Klara Krieg, Markus Schedl, and Navid Rekabsaz. 2023. Show me a "Male Nurse"! How Gender Bias is Reflected in the Query Formulation of Search Engine Users. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2023)*. ACM.

[13] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[14] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-BiasIR: A Dataset for Investigating Gender Representation-Bias in Information Retrieval Results. In *Proceedings of the 2023 ACM SIGIR Conference On Human Information Interaction And Retrieval (CHIIR 2023)*. ACM.

[15] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?. In *Proceedings of the European Conference on Information Retrieval, Workshop on Algorithmic Bias in Search and Recommendation (ECIR-BIAS 2022)*. Springer, Cham, 104–116.

[16] Oleg Lesota, Alessandro B. Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 601–606. https://doi.org/10.1145/3460231.3478843

[17] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France*, Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 689–698. https://doi.org/10.1145/3178876.3186150

[18] Allen Lin, Jianling Wang, Ziwei Zhu, and James Caverlee. 2022. Quantifying and Mitigating Popularity Bias in Conversational Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 1238–1247. https://doi.org/10.1145/3511808.3557423

[19] J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151. https://doi.org/10.1109/18.61115

[20] Xia Ning and George Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *Proceedings of the 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada*, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). IEEE Computer Society, 497–506. https://doi.org/10.1109/ICDM.2011.134

[21] Amifa Raj and Michael D Ekstrand. 2022. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 726–736.

[22] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (Hong Kong, Hong Kong) *(WWW '01)*. Association for Computing Machinery, New York, NY, USA, 285–295. https://doi.org/10.1145/371920.372071

[23] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, David Elsweiler (Ed.). ACM, 337–341. https://doi.org/10.1145/3498366.3505791

[24] Markus Schedl, Navid Rekabsaz, Elisabeth Lex, Tessa Grosz, and Elisabeth Greif. 2022. Multiperspective and Multidisciplinary Treatment of Fairness in Recommender Systems Research. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 90–94.

[25] Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. 2020. Exploring user opinions of fairness in recommender systems. In *CHI'20 Workshop on Human-Centered Approaches to Fair and Responsible AI*.

[26] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

[27] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

[28] Harald Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*. ACM, Vancouver, Canada, 154–162. https://doi.org/10.1145/3240323.3240372

[29] Emre Yalcin and Alper Bilge. 2022. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Inf. Process. Manag.* 59, 6 (2022), 103100. https://doi.org/10.1016/j.ipm.2022.103100