ChatGPT

用户公平性定义与衡量指标

用户公平性（User Fairness）关注推荐系统是否对不同用户群体给出一致或平衡的结果。一般可从群体公平（不同敏感属性组的用户应获得相似的推荐质量或曝光）和个体公平（相似用户应受相似对待）的角度刻画[1][2]。例如，Ekstrand等人指出，应保证不同性别、年龄等群体用户的推荐准确率相近[1]；Leonhardt等则引入“得分差异”（Score Disparity）和“推荐差异”（Recommendation Disparity）来量化用户公平性，前者用基尼系数衡量用户群体满意度分布不平等，后者衡量推荐列表质量差异[2][3]。常用的公平性指标包括群体之间**准确率或满足度差异**、**曝光率差异**、**Calibrated Fairness**（按用户兴趣比例推荐的匹配度）、统计测试（如KS检验、χ^2检验）等。一般会计算不同敏感组用户的平均命中率、评分偏差或曝光占比，并通过最大差异、比率或信息论度量（如互信息）来评估不公平程度[1][2]。需要注意的是，不同定义下指标各异，且目前尚无统一标准[4]。

主要方法分类

研究中提出了多种技术手段来提升用户公平性，可大致分为**数据层面**、**模型层面**和**结果层面**处理三类方法[5]：

- **数据处理（Pre-processing）**：通过调整训练数据来减轻偏见。例如，对低曝光或受歧视用户群体进行过采样/重采样（resampling），或注入“公平示例”（antidote data）[6]。Ekstrand等（FAT* 2018）调整保护组比例；Rastegarpanah等（WSDM 2019）往训练集中加入人为标注的反歧视数据。

- **模型训练（In-processing）**：在推荐模型训练目标中加入公平性约束。常见做法包括**正则化**和**对抗训练**。正则化方法将公平指标作为损失项或约束，例如Yao&Huang（NeurIPS 2017）直接将价值公平性指标加入优化[7]；Kamishima等（FAT* 2018）添加分布匹配和互信息正则[8]；Wan等（WSDM 2020）用方差分析（ANOVA F-statistic）作为正则[9]。对抗训练方法则利用对抗网络去除用户表征中的敏感信息，如Bose&Hamilton（ICML 2019）在图嵌入中加入公平性对抗约束[10]，Wu等（WWW 2021）学习去关联的用户表示[11]，Li等（WSDM 2021/WWW 2022）采用GAN或文本重构正则学习公平表示[12]。

- **重排序与后处理（Re-ranking / Post-processing）**：在生成初始推荐结果后再进行调整，使结果满足公平性要求。方法包括**Slot-wise排序**（逐个位置优化，如MMR最大边际相关性）和**全局排序**（如Integer Programming全局最优重排）[13][14]。典型工作如Zehlike等（CIKM 2017）、Karako&Manngala（UMAP 2018）、Steck（RecSys 2018）使用MMR保证群体公平；Serbos等（WWW 2017）、Biega等（SIGIR 2018）利用贪心或线性规划保证累积公平[15][16]。还有**基于用户视角的重排序**，如Xiao等（RecSys 2017）从帕累托效率角度优化每个用户的推荐列表[14]。此外，强化学习也被用于长期公平性优化（如Ge等WSDM 2022将公平视作奖励目标）[17]。

- **其他方法**：如**因果/去偏**策略，考虑用户行为的选择偏差。如Tang等（RecSys 2023）提出使用逆倾向评分（IPS）校正交互选择偏差[16]；Zhang等（NeurIPS 2023）在敏感属性不全时通过对抗学习提升模型鲁棒公平[18]。还有采用GAN、VAE等深度生成模型进行数据或表征去偏的尝试。

存在的挑战与争议

尽管方法繁多，当前研究也面临不少挑战与争议。一是**公平性本质模糊、场景依赖**。许多工作在抽象数据上提出算法，却没有明确公平的规范意涵或业务目标 19 4 。例如仅考虑推荐多样性可能无法自动提高用户公平度 19 ；另外，在特定场景下是否要区别对待用户（如付费会员与普通用户）也存在争论 20 。二是**评价方式的局限性**。大部分研究依赖离线实验和代理指标，缺少真实用户研究和在线测试 21 。如有工作通过用户调查发现，在某些偏见消除后用户反而更喜欢原始推荐 22 。三是**多目标权衡与冲突**：用户公平有时与效果指标或其他利益（如平台收益、物品公平）产生冲突 23 22 。例如，提高长尾商品曝光虽有益物品公平，但可能降低主流用户满意度；反之，最大化准确率可能加剧热门物品偏向，损害长尾用户体验。如何在准确率、用户公平和其他目标之间取得平衡是一个开放问题。四是**数据与评测缺乏**：现实中敏感属性标注不完备，且很少有数据集专门用于测试用户公平（如招聘、社交推荐等领域），导致研究往往局限于通用数据集 24 25 。

代表性论文与机构

近年，用户公平性研究涌现了多篇顶会论文。例如，Zhang等（USTC）在**NeurIPS 2023**提出FairLISA框架，通过有限敏感属性对抗学习提升用户建模公平性 18 ；杨浩等（人民大学/蚂蚁金服）在**RecSys 2023**提出基于分布鲁棒优化的公平推荐方法，对抗训练-测试偏移问题 26 ；Tang等（人民大学/腾讯）在**RecSys 2023**定义了“无偏公平Top-N推荐”任务，提出用IPS加权消除数据噪声对公平性的影响 16 。其他相关工作还包括Chenyang Wang等（清华/中国移动）在**RecSys 2023**提出的双面校准（用户层面和系统层面的兴趣分布校准） 27 ，以及Ionescu等（苏黎世大学）探讨内容创作者群体公平**（RecSys 2023）**。早前的会议上，如**SIGIR/WWW/AAAI**等也有多项贡献：例如Serbos等人（WWW 2017）提出组推荐用户公平包算法，Steck（RecSys 2018）研究推荐校准和多目标公平 13 ；Islam等（微信）采用预训练微调策略改善公平性（WWW 2021）；Wu等（阿里巴巴）在图神经推荐中学习公平用户表示（WWW 2021）等。这些研究主要集中在中国科学院、清华大学、浙江大学、瑞士苏黎世大学、腾讯、阿里巴巴、华为等机构，顶会包括RecSys、SIGIR、WSDM、WWW、NeurIPS、ICML等 28 26 16 。

总之，近两年学界对用户公平性的关注度持续升高，相关工作层出不穷，但仍需要更多从应用语境出发的规范讨论、多元化的实证评估，以及考虑多方利益和用户感知的新方法，以推动该领域取得更系统的进展 4 21 。

参考文献：以上内容依据相关文献整理 1 2 5 19 18 26 16 等。

---

1 5 6 7 8 9 10 11 12 13 14 15 17 20 23 28 [2206.03761] A Survey on the Fairness of Recommender Systems
https://ar5iv.labs.arxiv.org/html/2206.03761

2 3 User Fairness in Recommender Systems
https://arxiv.org/pdf/1807.06349

4 19 21 22 24 Fairness in recommender systems: research landscape and future directions | User Modeling and User-Adapted Interaction
https://link.springer.com/article/10.1007/s11257-023-09364-z

16 26 27 RecSys 2023 - Accepted Contributions - RecSys – RecSys
https://recsys.acm.org/recsys23/accepted-contributions/

18 proceedings.neurips.cc
https://proceedings.neurips.cc/paper_files/paper/2023/file/81a12aed87eb9c75dfdf91ed99d5519d-Paper-Conference.pdf

25 Consumer-side fairness in recommender systems: a systematic survey of methods and evaluation | Artificial Intelligence Review
https://link.springer.com/article/10.1007/s10462-023-10663-5