arXiv:2303.00400v1 [cs.IR] 1 Mar 2023

关于推荐中准确性、校准失准和流行度偏差的研究

Dominik Kowald(✉)[1,2⋆]、Gregor Mayr[2⋆]、Markus Schedl[3,4]和Elisabeth Lex[2]

[1] Know-Center GmbH，奥地利格拉茨 dkowald@know-center.at[2] 格拉茨技术大学，奥地利格拉茨 gregor.mayr@student.tugraz.at dominik.kowald、elisabeth.lex@tugraz.at[3] 约翰内斯·开普勒大学和林茨技术研究所，奥地利林茨 markus.schedl@jku.at

摘要。 近期研究提出了多种用于衡量推荐性能不一致性的指标，包括用户群体之间的准确性差异、校准失准和流行度提升。然而，目前仍缺乏一项研究，探讨不同用户群体中校准失准和流行度提升与推荐准确率之间的关系。此外，某些特定类型是否促成了不同用户群体之间推荐性能不一致性的产生，目前尚不明确。本文针对五种广为人知的推荐算法，对偏好热门内容程度不同的用户群体，从上述三个方面进行了分析。此外，我们还研究了不同类型如何影响推荐性能的不一致性，以及这种影响与各类型受欢迎程度之间的一致性。通过使用来自 Last.fm、MovieLens 和 MyAnimeList 的数据，我们得出了两个关键发现。首先，我们发现，对热门内容兴趣较低的用户获得的推荐准确率最差，而这一结果与校准失准和流行度提升相一致。其次，我们的实验表明，特定类型在导致推荐性能不一致性方面的影响程度不同，尤其是在 MyAnimeList 数据集的校准失准方面。

关键词： 推荐系统 · 流行度偏差 · 校准失准 · 准确性 · 推荐不一致性 · 流行度提升

1 引言

推荐系统通过为用户提供电影或音乐等内容的个性化建议，使用户受益。然而，我们也从以往研究中了解到，推荐系统在不同用户群体之间的推荐表现存在不一致性 [2,9]。这种不一致性的一个例子是，不同用户群体的推荐准确率存在差异，这可能

⋆ 两位作者对本研究的贡献相同。

导致偏好不属于社区主流的用户受到不公平待遇 [19,18]。另一个例子是，推荐系统的输入数据与生成的推荐之间存在不一致，这可能导致推荐内容过于受欢迎，和/或与特定用户群体的兴趣不匹配 [2,9]。因此，流行度偏差可以被视为推荐不一致性的一种具体表现。

除了测量不同用户群体之间的推荐准确度差异外，相关研究 [2] 还建议通过两个指标来量化推荐性能的不一致性，即校准失准和流行度提升。校准失准用于量化用户画像中的类型谱与实际推荐之间的偏差 [29,24]。例如，若用户聆听的歌曲中，流行音乐占 45%、摇滚占 35%、说唱占 20%，那么经过校准的推荐列表也应包含相同的流派分布。

相关研究还提出了流行度提升指标，用于考察推荐算法在多大程度上加剧了流行度偏差方面的不一致性 [3,4]。该流行度提升指标量化了系统推荐的热门项目所占比例过高的程度。例如，流行度提升为正，表明推荐项目的平均受欢迎程度高于用户档案中的项目。因此，在本文余下部分，我们将流行度提升称为衡量推荐算法流行度偏差的指标。

然而，目前仍缺乏一项研究，探讨不同用户群体中的校准失准和流行度提升与推荐准确性之间的关系。我们认为，这类研究的结果有助于为每个用户群体选择最合适的推荐去偏方法。此外，特定类型是否会导致不同用户群体之间的推荐表现不一致，这一点尚不明确。了解这一点可能会有所帮助，例如改进基于校准的推荐去偏方法。

本文工作。本文研究五种知名推荐算法在音乐（Last.fm）、电影（MovieLens）和动漫（MyAnimeList）领域中的准确性、校准失准和流行度偏差。这些算法用于预测用户对项目的偏好，即 UserItemAvg、UserKNN、UserKNNAvg [12], 非负矩阵分解 [25],和协同聚类 [10] 。我们根据每个数据集中用户对热门内容的偏好程度，将用户划分为低、中、高三个用户群体，分别称为 LowPop、MedPop 和 HighPop。通过这项研究，我们旨在深入探讨推荐中的准确性、校准失准和流行度偏差之间的联系。

此外，本文还探讨用户群体中的哪些类型会特别影响不同算法和领域之间的推荐不一致性。借此，我们旨在了解特定类型是否会导致推荐性能不一致性，并探究这是否与这些类型的受欢迎程度相一致。

研究发现与贡献。我们发现，低流行度消费者始终获得最低的推荐准确率；在所有调查的数据集中，

这一用户组的校准失准程度最高。就流行度提升而言，我们观察到，所有算法都会加剧流行度偏差。

在对流派层面的分析中，我们发现确实有些流派对不一致性影响很大，尤其是在 MyAnimeList 数据集中的校准偏差方面。总之，我们的论文有四方面贡献：

1. 我们为推荐系统领域的三个知名数据集补充了流派信息，以研究推荐性能的不一致性。2. 我们评估了五种知名推荐算法在准确性、校准偏差和流行度提升方面的表现。3. 我们从流派层面考察推荐不一致性，并表明不同流派对推荐性能不一致性的形成有不同影响。4. 为促进研究的可复现性，我们与研究社区分享了扩展后的数据集以及研究中使用的源代码。

2 相关工作

信息检索和推荐系统中的偏差是一项新兴研究主题，相关研究已展示了量化系统中不同偏差的多种方法 [7,22]。其中一种偏差是流行度偏差，即受欢迎程度较高的项目比受欢迎程度较低的项目更常被推荐。研究 [9] 发现，并非所有用户受到的影响都相同，有些用户群体收到的推荐比其他群体更不一致。Ekstrand 等人 [9,17], 例如，发现不同人口统计群体之间的推荐准确率存在不一致性；在多个数据集中，性别和年龄不同的群体在有效性方面呈现出具有统计显著性的差异。作者评估了不同的推荐算法，并确定了效用影响程度的差异。

Abdollahpouri 等 [3,4,2] 也对这一研究方向作出了贡献，并提出了两个指标，用于从用户视角量化推荐性能的不一致性。第一个是校准失准指标，用于量化用户档案中的类型谱与该用户所获推荐中的类型谱之间的不匹配程度。第二个是流行度提升指标，用于衡量用户受流行度偏差影响的程度，也就是用户档案和该用户所获推荐中热门项目分布不均的程度。他们在电影领域的数据集上发现，受流行度偏差影响较大的用户，也会得到校准失准程度更高的结果。同样，Kowald 等 [19]分析了音乐领域不同用户群体之间的流行度偏差和准确性差异。作者发现，由于用户音乐收听行为中普遍存在重复消费模式，流行度提升指标在音乐领域得出的结果与电影领域有所不同。

表1：数据集统计，包括用户数 $|U|$、项目数 $|I|$、评分数$|R|$、不同类型数 $|C|$，以及稀疏度和评分范围 R-范围。

数据集	$\|U\|$	$\|I\|$	$\|R\|$	$\|C\|$	$\|R\|/\|U\|$	$\|R\|/\|I\|$	稀疏度	R-范围
LFM	3,000	131,188	1,417,791	20	473	11	0.996	$[1-1,000]$
ML	3,000	3,667	675,610	18	225	184	0.938	$[1-5]$
MAL	3,000	9,450	649,814	44	216	69	0.977	$[1-10]$

本文在这些研究的基础上，将校准失准和流行度提升与不同用户群体的推荐准确率联系起来。此外，我们还考察特定类型是否会导致不同用户群体和数据集之间出现推荐不一致性。我们希望借此为流行度偏差缓解方法的研究提供参考。例如，[5] 提出了基于校准的处理中方法，用于推荐去偏。我们认为，关于哪些类型会导致结果校准失准的发现，可用于改进这些方法。此外，相关研究还提出了对推荐列表进行重排序的后处理方法 [1,6]。我们认为，对于不同用户群体中准确性与流行度提升之间关系的发现，有助于确定应对哪些用户进行此类重排序。

3 方法

本节介绍本研究所使用的数据集、实验设置和评估指标。

3.1 数据集

我们使用了音乐、电影和动漫领域的三个不同数据集。具体而言，我们使用了先前工作 [17][4] 提供的 Last.fm (LFM)、MovieLens (ML) 和 MyAnimeList (MAL) 数据集样本。每个数据集恰好包含 3,000 名用户，并分为三个规模相同的群体，每组 1,000 名用户。我们每个用户组采用 1,000 名用户，以便与同样使用这一规模群体的先前研究进行比较。我们根据用户对热门项目的倾向来划分群体。根据 [17], 中给出的定义，我们将用户 u 对热门项目的倾向定义为热门项目在 u 的用户档案中所占的比例。如果某个项目的项目流行度分数位于前 20%（即与 i 互动的用户相对数量），我们就将其定义为热门项目。我们分别将对热门项目倾向最低、中等和最高的群体称为低流行度、中流行度和高流行度。图 1 展示了三个数据集中这三组用户的用户画像中热门项目占比的箱线图。

[4] 由于缺乏流派信息，我们不使用 BookCrossing 数据集。

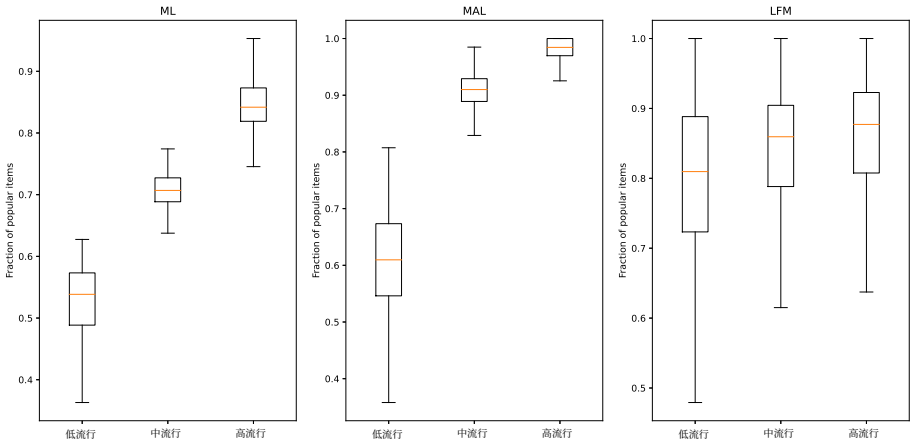

图1：箱线图展示了三个用户组在各数据集中用户档案中热门项目所占的比例。与中流行度组和高流行度组相比，低流行度组中热门项目的比例最低。在 LFM 数据集中，由于音乐聆听行为中的重复消费模式，这种差异不像其他数据集那么明显。

图 1 展示了三个用户群体和各数据集中用户画像里热门项目占比的箱线图。可以看出，与中流行度和高流行度用户组相比，低流行度用户组中热门项目的比例最低。在 LFM 数据集中，这种差异不像其他数据集那样明显，这是由音乐聆听行为中的重复消费模式造成的。

数据集的基本统计信息见表1，我们通过 Zenodo[5]分享了数据集样本。下文将进一步介绍这些数据集，以及我们如何利用流派信息对其进行扩展。此外，我们还会分析各数据集中评分和用户层面的流行度分布，为后续关于流派层面的研究提供背景。

Last.fm (LFM)。本研究使用的 LFM 数据集样本基于 LFM-1b 数据集 [28] 以及 [17]中使用的子集。它包含来自音乐流媒体平台 Last.fm 的聆听记录。我们仅纳入包含流派信息的音乐艺人聆听记录。流派信息通过将 Last.fm 的用户生成标签（分配给艺人）与 AllMusic 数据库中的 20 个主要流派进行匹配而获得（排名前三的是摇滚、另类摇滚和流行音乐）。将表1中的 LFM 数据集样本与 [17], 中的样本进行比较后，我们注意到艺人数目 $|I|$ 从 352,805 减少到 131,188，这意味着 LFM 中有大量长尾艺人没有可用的流派信息。不过，就评分而言，评分数量仅从 1,755,361 相对小幅地减少到 1,417,791。按照我们先前的研究 [17]，我们将次数

[5] https://doi.org/10.5281/zenodo.7428435

用户聆听某位艺人的次数解释为评分分数，并使用最小-最大归一化将其缩放至 [1; 1,000] 。我们在单个用户层面进行归一化，以确保所有用户的评分范围一致：用户聆听次数最多的艺人评分为 1,000，聆听次数最少的艺人评分为 1。

MovieLens（ML）。我们的 ML 数据集样本基于明尼苏达大学提供的 ML-1M 数据集 [11]。在此，我们直接从原始数据集中收集电影的流派信息[6]，该数据集为所有电影提供流派信息，并包含 18 种不同类型（排名前三：喜剧、剧情、动作）。就稀疏度而言，ML 是我们最稠密的数据集样本，而 LFM 是最稀疏的样本。

MyAnimeList（MAL）。我们研究中使用的 MAL 数据集基于 Kaggle 提供的推荐系统挑战赛数据集。与 ML 的情况一样，原始数据集[7] 已经为每个项目提供了类型信息，因此包含 44 种不同类型（排名前三：喜剧、动作、爱情）。不过，MAL 的一个特殊特征是，该数据集还包含隐式反馈（即用户将某部动漫加入书签时产生的反馈）。按照 [17]，我们将隐式反馈设为显式评分 5。在评分数量方面，MAL 是我们研究中使用的最小数据集，而 LFM 是最大的一个。

类型流行度分布。 为了更好地了解三个用户群体中各个流派的受欢迎程度，我们在图 2 中从评分数和用户数两个层面绘制了类型流行度分布。各流派按照其在所有三个用户群体中的总体评分受欢迎程度排序，即最受欢迎的流派位于最左侧。从评分数层面（左侧图表）来看，所有用户群体的流行度分布相似。有趣的是，在机器学习和 MAL 中，低流行度组在所有流派中的评分数都最多；而在 LFM 中，评分数最多的则是中流行度组。

从用户数层面来看，LFM 和机器学习中所有用户群体的流行度分布相似。然而，在 MAL 中，考察中流行度和高流行度用户群体时，我们发现“成人”流派的用户数显著下降。低流行度用户群体中并非如此，因此，低流行度用户对这些流派的独特偏好可能会导致 MAL 数据集中的低流行度组出现推荐性能不一致的情况。将这些结果与左侧的评分分布进行比较，我们发现，在“成人”流派中，中流行度和高流行度用户群体的评分数并未下降。不过，低流行度用户对该流派的评分数有所增加。这再次表明，低流行度用户对与“成人”流派相关的动漫兴趣浓厚。

最后，我们还考察了不同流派和用户群体中的物品流行度分布；在流派层面比较各用户群体时，我们未发现任何显著差异。

[6] https://grouplens.org/datasets/movielens/1m/[7] https://www.kaggle.com/CooperUnion/anime-recommendations-database

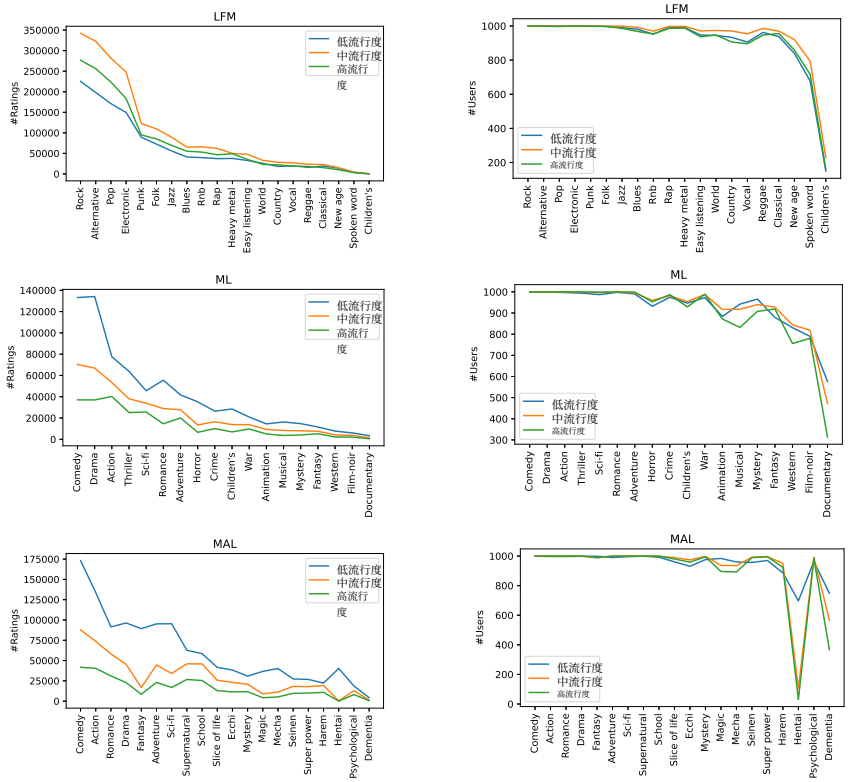

图 2：三个数据集和用户群体中，评分层面（左）和用户层面（右）的类型流行度分布。

3.2 实验设置

接下来，我们将介绍本研究中采用的五种推荐算法及评估协议。

推荐算法。 沿用我们此前的研究 [19,17]，，我们利用基于 Python 的 Surprise 框架 [12]，将推荐任务建模为评分预测问题。具体而言，我们采用了 [17] 中研究的四种协同过滤（CF）推荐算法。由于我们此前的工作 [17] 使用的数据集样本与本研究相同，因此我们沿用相同的超参数设置。有关确切的参数设置，请参阅我们通过 GitHub[8] 分享的源代码。我们没有进行任何额外的超参数优化，因为我们主要关注的是评估低流行度、中流行度和高流行度这三个用户群体之间评估指标的（相对）差异，而不是将一种新颖的推荐方法与最先进算法进行比较。这也是我们

[8] https://github.com/domkowald/FairRecSys

关注相关研究中采用的五种传统且易于理解的推荐算法，而非分析近期深度学习架构性能的原因；后者还会导致计算复杂度大幅提高。

本研究采用的推荐算法包括两种基于 KNN 的算法 UserKNN 和 UserKNNAvg，后者纳入了目标用户和物品的平均评分。我们还研究了协同聚类，这是一种可扩展的基于协同聚类的协同过滤方法 [10]，以及 NMF，即非负矩阵分解 [25]。此外，我们还加入了 [19]，中采用的一种非协同过滤方法，即 UserItemAvg。该方法使用数据集中的总体平均评分进行基线估计，并加入目标用户和物品的偏好偏差，例如，如果某个用户倾向于比平均用户给出更高的评分 [14]。

评估方案。关于评估方案，我们再次沿用先前的研究 [19,17] ，采用随机 80/20 训练/测试集划分，并以五折交叉验证的方式进行。因此，我们使用训练集训练算法，并通过比较实际评分与预测评分来衡量算法在测试集上的准确性。通过采用五折交叉验证，我们确保评估方案的稳健性，并控制随机训练/测试集划分可能导致推荐计算中的类型比例或离群值出现的波动。

在计算校准失准和流行度提升时，我们为目标用户采用一个包含前 10 个项目的推荐列表，即预测评分分数最高的 10 个项目。由于我们先前的研究 [19,17] 表明，LopPop 用户组通常在所有用户群体中获得最差的推荐准确率，因此我们尤其关注这一用户组。因此，我们使用 t 检验来检验统计显著性，分别比较 LowPop 与 MedPop，以及 LowPop 与 HighPop。我们报告所有指标在 5 折中的平均值，并且仅当统计显著性在全部 5 折中均成立时才予以标示。

3.3 评估指标

我们使用以下指标量化推荐性能的不一致性：(i) 用户群体之间的准确性差异，(ii) 校准失准，以及 (iii) 流行度提升：

准确性（平均绝对误差）。我们使用广为人知的平均绝对误差（平均绝对误差）指标来衡量准确性。用户 u 的平均绝对误差为：

$$MAE(u) = \frac{1}{|R_u^{test}|} \sum_{r_{u,i} \in R_u^{test}} |r_{u,i} - R_{u,i}| \quad (1)$$

此处，将用户 u 对物品 i 的预测评分$R_{u,\ i}$与 u的测试集 R_u^{test}中的真实评分$r_{u,\ i}$进行比较。鉴于均方根误差（RMSE）存在若干缺点，尤其是在比较观测数量不同的群体时，我们更倾向于使用平均绝对误差，而非通常使用的均方根误差（RMSE）指标

（在我们的案例中，即评分） [30]。我们通过对 g中所有用户的平均绝对误差值取平均，报告用户组g 的平均绝对误差。

为了在前-n 推荐评估设置中也通过平均绝对误差验证准确性结果，我们还报告了广为人知的精确率和召回率指标。为此，如果测试集中的某个项目的评分高于训练集中的平均评分，我们就将其归类为相关项目。

校准失准（匹配系数）。Steck [29] 提出的校准指标用于量化用户画像 p 的类型谱与实际推荐 q之间的相似度。Lin 等人 [24] 将该指标重新诠释为校准失准，即 p 与 q之间的偏差。我们沿用这一界定，并使用库尔巴克-莱布勒（KL）散度计算偏差，即计算 p中的类型分布（ $p(c|u)$）与 q中的类型分布（$q(c|u)$）之间的差异。公式如下：

$$KL(p||q) = \sum_{c \in C} p(c|u) \log \frac{p(c|u)}{q(c|u)} \tag{2}$$

这里， C 是数据集中所有类型的集合。因此， $KL = 0$ 表示完全校准，而较高的 KL值（即接近 1）表示推荐存在校准失准。与平均绝对误差的情况一样，我们报告某个群组 g中所有用户的校准失准平均值。

流行度提升（PL）。流行度提升指标考察推荐算法在多大程度上放大了用户画像 [3,4]中固有的流行度偏差。因此，它量化了针对特定用户组 g （即低流行度、中流行度、高流行度）过度推荐热门项目的程度。我们将组平均流行度 GAP_p（g）定义为该组 g用户画像 p 中项目受欢迎程度的平均值。同样， GAP_q（g）表示向该组 g所有用户推荐的项目的平均受欢迎程度。流行度提升的计算公式如下：

$$PL(g) = \frac{GAP_q(g) - GAP_p(g)}{GAP_p(g)} \tag{3}$$

这里， PL（g） > 0 表示向 g 的推荐过于热门，PL（g） < 0 表示向 g 的推荐不够热门，而PL（g） $= 0$ 则是理想值。

4 结果

本节将介绍并讨论本研究的结果：首先从更一般的层面展开，然后讨论类型层面的结果。

准确性、校准失准与流行度偏差之间的联系。表 2 总结了我们在三项指标（平均绝对误差、匹配系数、流行度损失）上的结果，涵盖三种

表 2：低流行度、中流行度和高流行度用户群体的平均绝对误差、匹配系数和流行度损失结果。最高（即最差）的结果以**粗体**标示。根据t检验，低流行度与中流行度以及低流行度与高流行度之间的统计显著性以 * 表示，$p < 0.05$。评分范围以方括号标示。

	Data	**LFM** [1-1,000]			**ML** [1-5]			**MAL** [1-10]		
算法	指标	平均绝对误差	匹配系数	流行度损失	平均绝对误差	匹配系数	流行度损失	平均绝对误差	匹配系数	流行度损失
UserItemAvg	低流行度	**48.02***	**0.52***	1.28	**0.74***	**0.78***	**0.70***	**0.99***	**0.95***	**1.12***
	中流行度	38.48	0.48	**1.61**	0.71	0.71	0.42	0.96	0.73	0.42
	高流行度	45.24	0.42	1.35	0.69	0.63	0.24	0.97	0.64	0.15
UserKNN	低流行度	**54.32***	**0.51***	0.52	**0.80***	**0.75***	**0.64***	**1.37***	**0.92***	**0.74***
	中流行度	46.76	0.50	**0.82**	0.75	0.69	0.37	1.34	0.72	0.22
	高流行度	49.75	0.45	0.80	0.72	0.62	0.20	1.31	0.63	0.08
UserKNNAvg	低流行度	**50.12***	**0.49***	0.35	**0.76***	**0.78***	**0.49***	**1.00***	**0.90***	**0.54***
	中流行度	40.30	0.47	0.61	0.73	0.70	0.33	0.95	0.73	0.24
	高流行度	46.39	0.42	**0.64**	0.70	0.61	0.20	0.95	0.64	0.11
NMF	低流行度	**42.47***	**0.54***	0.10	**0.75***	**0.78***	**0.57***	**1.01***	**0.91***	**0.87***
	中流行度	34.03	0.52	0.17	0.72	0.71	0.37	0.97	0.72	0.35
	高流行度	41.14	0.48	**0.33**	0.70	0.63	0.22	0.95	0.63	0.13
协同聚类	低流行度	**52.60***	**0.52***	0.68	**0.74***	**0.77***	**0.70***	**1.00***	**0.90***	**1.10***
	中流行度	40.83	0.51	**1.04**	0.71	0.70	0.43	0.96	0.72	0.42
	高流行度	47.03	0.45	0.99	0.68	0.62	0.25	0.98	0.63	0.16

用户群体（低流行度、中流行度、高流行度）、三个数据集（LFM、ML、MAL）和五种算法（UserItemAvg、UserKNN、UserKNNAvg、NFM、协同聚类）。所呈现的结果是对所有用户和所有折的结果取平均后得出的。我们可以看到，在 ML 和 MAL 数据集上，低流行度用户组在平均绝对误差、匹配系数和流行度损失方面的结果最差。根据显著性水平为 $p < 0.05$ 的 t检验，这些结果也具有统计显著性。对于 LFM 数据集，低流行度用户组在平均绝对误差和匹配系数指标上的结果也最差。

然而，在考察流行度提升指标时，我们观察到了不同的结果，即中流行度或高流行度项目的流行度提升最高。这与我们此前的研究 [19], 一致；该研究表明，流行度提升指标在 LFM 和机器学习上的结果有所不同。音乐与电影（以及动漫）的一个潜在区别是，音乐通常会被重复消费（即用户会多次收听同一位艺术家的作品），而电影大多只会观看一次。流行度提升指标的定义 [24]并未考虑重复消费模式 [15], ，因为无论项目的消费频率如何，它们都被赋予同等的重要性。这意味着，特定用户消费了 1,000 次的项目，与该用户只消费过一次的项目同等重要。

最后，在表3中，我们还使用广为人知的精确率和召回率指标，在 top-n 推荐评估设置下，以平均绝对误差验证了我们的准确性结果。为了在测试集中对相关项目进行分类，我们计算训练集中的平均评分，并将测试项目的评分超过

表3：以精确率和召回率衡量的准确性结果。我们使用 t 检验，对低流行度组与中流行度组用户、低流行度组与高流行度组用户进行统计显著性检验；当 $p < 0.05$ 时，以 * 标示。最佳（即最高）结果以 **粗体**标出。结果与平均绝对误差的结果一致，这意味着低流行度组的准确性结果最差，而高流行度组的准确性结果最佳。

	Data	**LFM**		**ML**		**MAL**	
算法	指标	精确率	召回率	精确率	召回率	精确率	召回率
UserItemAvg	低流行度	0.30	0.11	0.78*	0.19*	0.71*	0.15*
	中流行度	0.28	0.08	0.82	0.26	0.80	0.21
	高流行度	**0.39**	**0.14**	**0.83**	**0.36**	**0.80**	**0.33**
UserKNN	低流行度	0.33*	0.16	0.78*	0.18*	0.71*	0.15*
	中流行度	0.38	0.14	0.83	0.25	0.80	0.22
	高流行度	**0.53**	**0.22**	**0.83**	**0.35**	**0.81**	**0.34**
UserKNNAvg	低流行度	0.34	0.16	0.80*	0.20*	0.73*	0.16
	中流行度	0.34	0.12	0.83	0.27	0.80	0.23
	高流行度	**0.47**	**0.19**	**0.83**	**0.36**	**0.81**	**0.36**
NMF	低流行度	0.34	0.16	0.70*	0.14*	0.67*	0.13*
	中流行度	0.34	0.12	0.79	0.23	0.79	0.21
	高流行度	**0.46**	**0.19**	**0.82**	**0.34**	**0.81**	**0.33**
协同聚类	低流行度	0.33	0.16*	0.76*	0.17*	0.69*	0.14*
	中流行度	0.33	0.12	0.83	0.25	0.80	0.22
	高流行度	**0.46**	**0.20**	**0.84**	**0.35**	**0.81**	**0.34**

该训练平均评分的项目视为相关项目。我们发现，结果与平均绝对误差指标的情况非常相似。这意味着在几乎所有情况下，低流行度都得到最差的结果（即最低），而高流行度则得到最好的结果（即最高）。

流派对推荐不一致性的影响。此外，图3展示了我们针对三个数据集所做的调查结果，探究用户群体中的哪些流派尤其会影响推荐性能在校准失准方面的一致性。我们仅针对校准失准指标开展这项研究，因为在平均绝对误差和流行度提升指标方面，我们没有观察到不同流派之间存在明显差异。为了将用户的校准失准分数映射到流派 g，我们将用户 u 的 MC 分数赋予其收听过的所有流派 u。随后，对于每种流派 g，我们计算特定用户群体中所有收听过该流派的用户的平均 MC 分数 g。然后，我们在图3中绘制了非负矩阵分解算法和协同聚类算法的这些数值。为提高可读性，我们将数值按 0 到 1 的范围进行最小-最大归一化。

与图 2 一样，流派按受欢迎程度排序。为节省篇幅，我们只展示非负矩阵分解和协同聚类的结果；总体而言，这些结果与其他算法所得结果一致。不过，我们的 GitHub 仓库也支持查看其他

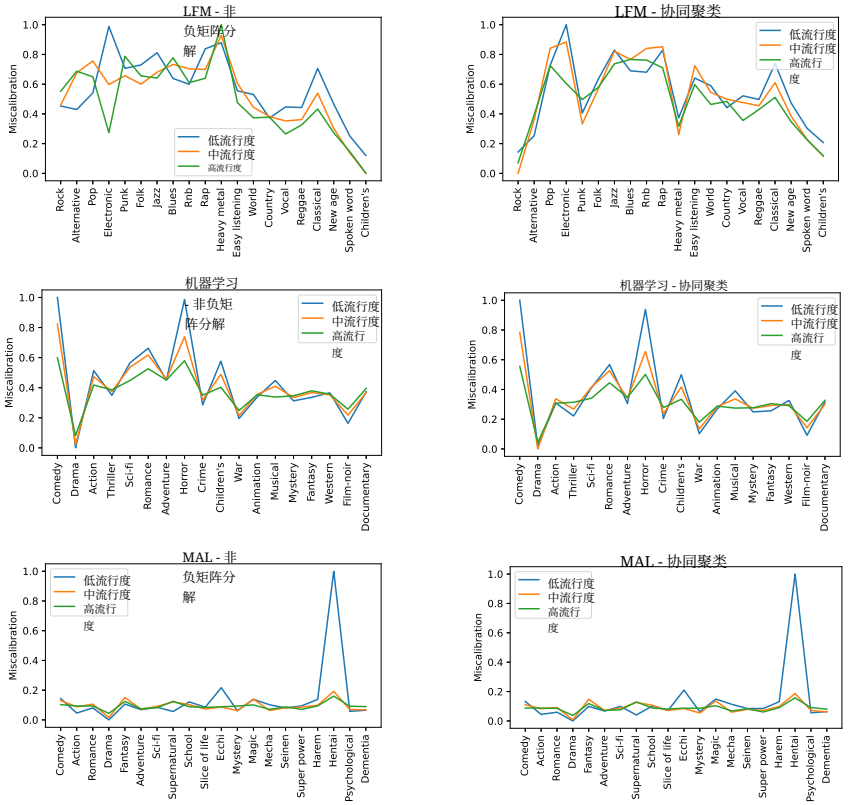

图 3：不同类型对非负矩阵分解算法（左）和协同聚类（右）的匹配系数的影响。我们发现，某些类型会显著加剧不一致性，尤其是在动漫（MAL）中。

算法的结果。此外，对于 MAL，我们排除了 24 个未观察到明显波动的流派。因此，如同 LFM 一样，图中展示了 20 个流派。

对于 MAL 数据集和低流行度组，我们观察到“Hentai”流派的结果存在严重的校准偏差。具体而言，从其位置可以看出，“Hentai”对大多数 MAL 用户来说是一个不流行的流派。然而，如图 2 所示，对于低流行度组中的用户（且仅限这一用户组）而言，它却是一个相关流派，但在他们的推荐列表中所占比例不足。这表明，确实存在某些特定流派，会在很大程度上导致特定用户群体的推荐不一致性。

5 结论与未来工作

本文研究了三个不同领域中不同用户群体的准确性、校准失准与流行度偏差之间的相互关系。

在此，我们以流行度提升来衡量流行度偏差，该指标比较推荐列表中项目的受欢迎程度与用户画像中项目的受欢迎程度。此外，我们还研究了校准失准，该指标比较用户画像中的类型分布与推荐列表中的类型分布。我们发现，总体而言，推荐在校准失准和流行度提升方面的不一致性与较低的准确性表现相一致。

其中一个例外是音乐推荐中的流行度提升指标；不过，这一结果与我们此前的研究一致 [19], ，在那项研究中，我们考察了重复消费场景。此外，我们发现，不同类型对校准失准和流行度提升的影响各不相同。这一发现尤其体现在面向低流行度用户的动漫推荐中，以及其他用户群体中不受欢迎的类型上。我们工作的另一项贡献是，向研究社区公开分享本研究中使用的数据集和源代码。

局限性与未来工作。 我们工作的一个局限是，仅关注了多媒体/娱乐领域的数据集，即音乐、电影和动漫。尽管我们研究了具有和不具有重复消费模式的领域，未来我们还计划研究其他领域中的准确性、校准失准和流行度偏差。这些领域可以包括在线市场中的推荐 [20] ，或社交网络中的推荐 [16]。这将有助于提升研究发现的普适性。为了进一步增强研究的普适性，我们还计划使用采用基于深度学习方法的新型推荐算法开展更多实验 [22]。

我们工作的另一个局限是，仅使用平均绝对误差、精确率和召回率作为衡量推荐准确性的指标。未来，我们计划进一步研究基于排序的指标，例如归一化折损累计增益 [13,21] ，以及衡量推荐新颖性和多样性的指标 [8]。在这方面，我们还计划改进评估方案，从随机训练/测试集划分转向时间训练/测试集划分 [27]。最后，我们还计划开展实验，采用更多用户组，并减少每组的用户数量（例如，设置 10 个用户组，每组 300 名用户）。借此，我们希望解决一个潜在局限：同一组内可能存在不同的流行度倾向。

作为未来工作的总体方向，我们计划基于本文的研究结果，制定策略来克服不同用户群体之间推荐效果的不一致。例如，对于我们发现校准失准严重的特定类型，我们计划研究基于校准的去偏方法 [5]。解决推荐系统中流行度偏差的另一种可能途径，是基于心理学概念构建模型 [23]。最后，我们计划研究新的指标，以衡量重复消费场景（例如音乐推荐）中的流行度提升。为此，我们计划引入该指标的加权变体，或研究将隐式反馈（例如播放次数）转换为显式评分的替代方法 [26]。

致谢。本研究由“DDAI” COMET 模块资助，该模块隶属于 COMET 卓越技术中心计划。该计划由奥地利联邦交通、创新与技术部（bmvit）、奥地利联邦数字与经济事务部（bmdw）、奥地利研究促进机构（FFG）、施蒂利亚州（SFG）以及来自工业界和学术界的合作伙伴共同资助。此外，本研究还获得了奥地利科学基金（FWF）的资助：P33526 和 DFH-23。

参考文献

1. Abdollahpouri, H., Burke, R., Mobasher, B.：通过个性化重排序管理推荐系统中的流行度偏差。载于：第三十二届 FLAIRS会议（2019） 2. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.：流行度偏差对推荐中的公平性和校准的影响。arXiv 预印本 arXiv:1910.05755（2019） 3. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.：推荐中流行度偏差所导致的不公平性。arXiv 预印本 arXiv:1907.13286（2019） 4. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.：推荐中流行度偏差、校准与公平性之间的联系。载于：第十四届 ACM推荐系统会议。第 726–731 页（2020） 5. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., Malthouse, E.：以用户为中心评估推荐系统中的流行度偏差。载于：第 29 届 ACM用户建模、自适应与个性化会议论文集。第 119–129 页（2021） 6. Adomavicius, G., Kwon, Y.：使用基于排序的技术提升总体推荐多样性。IEEE知识与数据工程汇刊**24**（5），896–911（2011） 7. Baeza-Yates, R.：搜索与推荐系统中的偏差。载于：第十四届 ACM推荐系统会议。第 2–2 页（2020） 8. Castells, P., Hurley, N., Vargas, S.：推荐系统中的新颖性与多样性。载于：推荐系统手册，第 603–646 页。Springer（2022） 9. Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., Pera, M.S.：那些酷孩子是如何融入其中的？推荐系统评估与有效性中的流行度和人口统计偏差。载于：公平性、问责制与透明度会议。第 172–186 页。PMLR（2018） 10. George, T., Merugu, S.：基于协同聚类的可扩展协同过滤框架。载于：第五届 IEEE国际数据挖掘会议（ICDM’05）。第 4 页。IEEE（2005） 11. Harper, F.M., Konstan, J.A.：MovieLens数据集：历史与背景。ACM交互式智能系统汇刊**5**，19:1–19:19（2015） 12. Hug, N.：Surprise：一个用于推荐系统的 Python库。《开源软件期刊》**5**（52），2174（2020） 13. Järvelin, K., Kekäläinen, J.：信息检索技术的基于累积增益的评估。ACM信息系统汇刊（TOIS）**20**（4），422–446（2002） 14. Koren, Y.：利用邻居进行因子分解：可扩展且准确的协同过滤。ACM数据知识发现汇刊（TKDD）**4**（1），1–24（2010）

15. Kotzias, D., Lichman, M., Smyth, P.：利用重复事件和新颖事件预测消费模式。IEEE知识与数据工程汇刊**31**（2），371–384（2018） 16. Kowald, D., Dennerlein, S., Theiler, D., Walk, S., Trattner, C.：社会语义服务器：一个为社会语义网络数据提供服务的框架。载于：第九届语义系统国际会议，I-SEMANTICS 2013。第 50–54 页。CEUR（2013） 17. Kowald, D., Lacic, E.：基于协同过滤的多媒体推荐系统中的流行度偏差。载于：Boratto, L., Faralli, S., Marras, M., Stilo, G.（编）《信息检索中的偏差与公平性进展》。第 1–11 页。Springer International Publishing，Cham（2022） 18. Kowald, D., Muellner, P., Zangerle, E., Bauer, C., Schedl, M., Lex, E.：支持地下音乐：非主流音乐听众的特征。EPJ数据科学**10**（1），14（2021） 19. Kowald, D., Schedl, M., Lex, E.：音乐推荐中流行度偏差所导致的不公平性：一项可复现性研究。载于：欧洲信息检索会议。第 35–42 页。Springer（2020） 20. Lacic, E., Kowald, D., Parra, D., Kahr, M., Trattner, C.：面向在线市场的可扩展社交推荐引擎：以 Apache Solr 为例。载于：第 23 届万维网国际会议论文集。第 817–822 页（2014） 21. Lacic, E., Kowald, D., Traub, M., Luzhnica, G., Simon, J.P., Lex, E.：利用室内定位系统应对推荐系统中的冷启动用户问题。载于：第九届 {ACM} 推荐系统会议海报论文集。ACM（2015） 22. Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., Schedl, M.：分析音乐推荐系统中的项目流行度偏差：不同性别受到的影响是否相同？载于：第 15 届 ACM推荐系统会议论文集。第 601–606 页（2021） 23. Lex, E., Kowald, D., Seitlinger, P., Tran, T.N.T., Felfernig, A., Schedl, M., 等：心理学指导的推荐系统。《信息检索基础与趋势》**15**（2），134–242（2021） 24. Lin, K., Sonboli, N., Mobasher, B., Burke, R.：协同过滤推荐系统中的校准：以用户为中心的分析。载于：第 31 届 ACM超文本与社交媒体会议论文集。第 197–206 页。HT '20，计算机协会，纽约，NY，美国（2020） 25. Luo, X., Zhou, M., Xia, Y., Zhu, Q.：一种用于推荐系统协同过滤的高效非负矩阵因子分解方法。IEEE工业信息学汇刊**10**（2），1273–1284（2014） 26. Pacula, M.：一种利用隐式用户反馈进行音乐推荐的矩阵因子分解算法。Maciej Pacula（2009） 27. Quadrana, M., Cremonesi, P., Jannach, D.：序列感知推荐系统。ACM计算评论（CSUR）**51**（4），1–36（2018） 28. Schedl, M.：用于音乐检索与推荐的 LFM-1b 数据集。载于：2016 年 ACM多媒体检索国际会议论文集。第 103–110 页（2016） 29. Steck, H.：校准推荐。载于：第 12 届 ACM推荐系统会议论文集。第 154–162 页。RecSys '18，计算机协会，纽约，NY，美国（2018） 30. Willmott, C.J., Matsuura, K.：评估平均模型性能时，平均绝对误差（MAE）优于均方根误差（RMSE）的优势。气候研究**30**（1），79–82（2005）