REASONER：一个包含多方面真实用户标注真实值的可解释推荐数据集

迈向更可度量的可解释推荐

Xu Chen[1,2], Jingsen Zhang[1, 2, *], Lei Wang[1, 2, *], Quanyu Dai[3], Zhenhua Dong[3], Ruiming Tang[3], Rui Zhang[4], Li Chen[5], Ji-Rong Wen[1,2]中国北京大数据管理与分析方法北京市重点实验室[2]中国人民大学高瓴人工智能学院，中国北京[3]华为诺亚方舟实验室，中国，[4]www.ruizhang.info[5]香港浸会大学计算机科学系，香港

arXiv:2303.00168v1 [cs.IR] 1 Mar 2023

摘要

可解释推荐受到产业界和学术界的广泛关注。它在提升推荐说服力、信息量和用户满意度方面展现出巨大潜力。尽管过去几年提出了许多颇具前景的可解释推荐模型，但这些模型的评估策略仍存在若干局限。例如，解释的真实标注并非由真实用户标注，解释大多仅从一个方面进行评估，而且评估策略难以统一。为缓解上述问题，我们提出构建一个包含多方面真实用户标注真实值的可解释推荐数据集。具体而言，我们首先开发了一个视频推荐平台，并围绕推荐可解释性精心设计了一系列问题。随后，我们招募了约 3000 名背景各异的用户使用该系统，并收集他们的行为以及对问题的反馈。本文详细介绍了数据集的构建过程，并对其特征进行了全面分析。此外，我们还开发了一个库，在统一框架中实现了十种知名的可解释推荐模型。基于该库，我们为不同的可解释推荐任务构建了若干基准。最后，我们介绍了该数据集带来的诸多新机遇，期望为可解释推荐领域带来新的启发。数据集、库及相关文档已发布于 https://reasoner2023.github.io/。

ACM 引用格式：
Xu Chen[1,2], Jingsen Zhang[1, 2, *], Lei Wang[1, 2, *], Quanyu Dai[3], Zhenhua Dong[3], Ruiming Tang[3], Rui Zhang[4], Li Chen[5], Ji-Rong Wen[1,2]。2023。REASONER：具有多方面真实用户标注的真实值的可解释推荐数据集：迈向更可衡量的可解释推荐。载于*ACM*会议论文集（*Conference'17*）。ACM，纽约，纽约州，美国，10页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

[*] 这些作者对本作品的贡献相同。

本文依据知识共享署名 4.0 国际版（CC-BY 4.0）许可协议发布。作者保留在其个人和企业网站上传播本作品的权利，但须注明适当的出处。*Conference'17*, *2017*年7月，美国华盛顿特区© 2023 IW3C2（国际万维网大会委员会），依据知识共享 CC-BY 4.0 许可协议发布。ACM 国际标准书号 978-x-xxxx-xxxx-x/YY/MM。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

推荐系统已成为大量现实世界应用中不可或缺的组成部分，涵盖电子商务 [28] 和视频分享网站 [40]，以及教育 [26]和医疗保健 [33]平台。基本上，推荐系统连接内容提供者（例如电子商务网站上的卖家）与用户。对于内容提供者而言，更好的推荐能够有效展示用户更可能点击或购买的物品，从而更好地实现业务目标。对于用户而言，高效的推荐模型可以帮助他们快速发现感兴趣的物品，从而节省时间并缓解信息过载问题。

近年来，为推荐系统提供解释已被视为一个重要的研究方向 [32]。本质上，解释可以进一步提升内容提供者和用户从推荐中获得的效用。对于内容提供者而言，恰当的解释可以更好地说服用户与推荐物品互动，从而提高收益 [9]。对于用户而言，合理的解释可以揭示更多有用的物品特征，帮助用户做出更明智的决策 [15]。

尽管现有的可解释推荐模型取得了显著成果，但这些模型的评估策略可能存在若干局限。例如，许多论文通过可视化来评估解释 [7], ，但这种方式无法用于对不同模型进行定量比较。许多人将用户评论视为解释真实值 [6]。然而，评论更多反映的是用户购买物品后的感受，而这可能不同于促使他们作出购买决定的原因。例如，用户可能因为品牌而购买 iPhone，却在评论中提到它价格高昂。为了克服上述评估策略带来的缺陷，许多人借助众包收集更可靠的人工反馈 [22]。然而，在大多数这类方法中，评估数据是基于现有公开数据集生成的，但真实标注者并非这些数据集中的真实用户，因此可能无法准确评估解释。此外，以往的工作大多只从一个方面评估解释，例如“解释是否能更好地说服用户 [29]”或“解释是否能揭示更多有价值的物品特征 [2]”，因此可能无法全面评估模型的可解释性。近年来，虽然已有少数研究从多个方面评估解释 [35], ，但其采用的策略和已标注数据集

都是针对各自的模型量身定制的，无法用于更通用的场景 [30]。

为了解决上述问题，本文提出构建一个由真实用户标注的通用数据集，可从多个方面评估推荐解释。我们将该数据集命名为**REASONER**，其中**RE** 表示真实值由真实用户标注。**ASON** 表示该数据集能够评估多方面解释。**ER** 是可解释推荐的缩写。尽管该数据集对可解释推荐领域十分重要，但构建起来并不容易。首先，如果通过扩展现有数据集来构建我们的数据集，就很难找到原始数据集中的真实用户来标注解释的真实标注。如果不依赖现有数据集，如何获取用户-物品交互也会成为一项挑战。此外，与其他领域不同，在其他领域中，标注任务通常是客观的（例如，识别图像内容）；而我们必须收集用户主观数据，这类数据对标注任务的设计更加敏感，例如“如何提出每个问题”以及“问题的顺序是什么”。如果设计不当，标注人员可能会误解问题，或受到外部因素干扰，从而降低标注准确率。最后，如何控制数据集质量也是一项挑战。我们的数据集旨在收集用户主观偏好，由于不同用户可能有不同观点，因此很难仅凭判断标注结果对错来净化数据。我们需要找到有效的规则，根据结果的合理性对其进行筛选。

为克服上述挑战，我们提出分两步构建数据集。首先，我们开发了一个视频推荐平台，并围绕推荐可解释性精心设计了一系列问题。例如：“哪些特征是你想观看这段视频的原因？”以及“哪些特征最能说明这段视频的特点？”。随后，我们招募了约 3000 名标注者使用该平台并回答上述问题。这使得解释的真实标注的标注者恰好就是产生用户-物品交互的真实用户[1]。为控制数据集的质量，我们制定了许多规则来检查并筛选标注结果。例如：“标注时间是否过短？”以及“同一用户的回答是否相互矛盾？”。除上述数据集外，我们还开发了一个库，其中包含十种知名的可解释推荐模型。我们为该库提供了详尽的文档和全面的示例。通过仔细调整库中的模型，我们基于数据集为不同的可解释推荐任务构建了若干基准测试。最后，我们指出了该数据集带来的若干新机遇。我们已在 https://reasoner2023.github.io/ 发布上述数据集、库和文档。

2 相关数据集

据我们所知，专门用于构建可解释推荐数据集的研究非常少，[16] 是其中具有代表性的一项。具体而言，[16] 将用户评论视为解释的真实标注，并将相似的评论句子归入同一簇，以便进行模型优化。在以往的大多数研究中，可解释推荐数据集旨在评估特定模型或开展用户研究。这些论文中的数据集可分为两类。第一类中，解释数据集基于传统推荐数据集构建，作者会招募标注者，让他们回答有关其设计模型生成的解释的问题。例如，[4] 要求标注者设想自己是真实用户，并标注解释的真实标注。[20] 要求标注者比较设计模型与基线模型生成的解释。在 [18, 36],中，标注者必须评估解释在帮助用户做出更好决策方面的有用性。[31, 34] 让标注者表达满意度，并评定解释的信息量。在这类数据集中，标注者通常并非真实用户，因此标注结果可能无法准确反映真实用户偏好。第二类中，解释数据集完全由标注者构建。例如，[24] 中的作者招募了286名标注者，用于评估一些给定的可解释推荐模型的透明度和有效性。在 [25], 中，作者聘请了20名标注人员来研究基于特征的解释的有效性。在 [13],中，研究者招募了152名标注人员，旨在研究用户对解释类型的感知。尽管这些数据集中的结果能够很好地代表真实用户偏好，但其中的标注问题高度依赖具体模型，因而很难用于其他研究。

表1：不同可解释推荐数据集的比较。“真实用户标注”表示真实标注是否由真实用户标注。“多方面解释”表示数据集是否包含多方面解释。“跨研究适用性”表示数据集是否可用于不同研究。“多模态解释”表示数据集是否包含多模态解释。此外，我们还列出了每个数据集中的标注者数量。

数据集	真实用户标注	多方面解释	跨研究适用性	多模态解释	# 标注者
[16]	✓	✗	✓	✗	<100
[4]	✗	✗	✓	✗	<100
[20]	✗	✗	✗	✗	<100
[18, 36]	✗	✗	✗	✗	<100
[31, 34]	✗	✓	✗	✗	<100
[24]	✓	✓	✗	✗	100-500
[25]	✓	✓	✗	✗	<100
[13]	✓	✓	✗	✓	100-500
REASONER	✓	✓	✓	✓	≈ 3000

与上述数据集相比，REASONER 具有许多显著优势。例如，它由真实用户标注，可用于评估不同模型。此外，它包含多方面的解释真实值，用户反馈中既有文本解释，也有视觉解释。为清楚了解 REASONER 的优势，我们在表1中将其与以往的可解释推荐数据集进行了比较。

3 数据集构建

为使标注者与真实用户保持一致，我们的数据集基于两个步骤构建。第一步是搭建推荐

[1]在下文中，我们不区分用户和标注者。

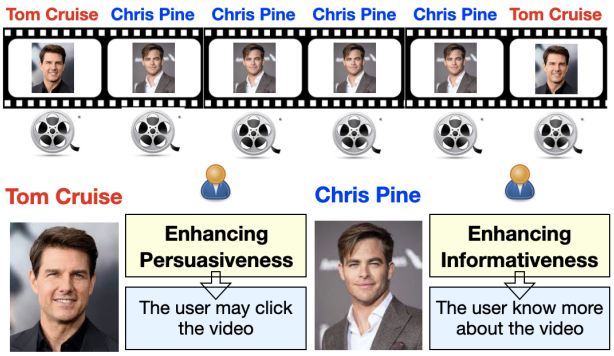

图1：不同用途的推荐解释所对应的不同真实标注。

平台，并设计许多关于推荐可解释性的问题。第二步是招募标注者使用上述平台，并收集他们的行为以及对上述问题的回答。下面将详细介绍这两个步骤。

3.1 构建推荐平台

与传统推荐系统相比，我们的平台专为收集用户可解释行为而设计。本节首先介绍平台上的物品，以及标注者可以从中选择的解释候选项。然后，我们描述用于收集解释真实标注的问题。最后，我们介绍每位标注者在平台上完成标注的完整流程。

3.1.1 物品。我们选择视频作为推荐项目，因为视频内容丰富，能够提供充足的解释候选项。考虑到标注时间过长可能会分散用户注意力，我们将视频时长限制在三分钟以内。所选视频涵盖不同类别，例如科技、音乐等，旨在全面体现用户偏好。所有视频均来自一家商业短视频平台。

3.1.2 解释候选项。 我们通过要求标注者从众多视频特征中进行选择来收集解释真实标注，例如，例如"哪些特征是你想观看该视频的原因？"。最重要的特征包括：

• **标签**。这些是文本特征，收集自三个来源：(1) 视频作者提供的原始标签。(2) 观众观看视频后发布的总体评论。(3) 观众以实时方式观看视频时提供的实时同步评论 [21]。上述信息来自与视频相同的网站。由于总体评论和实时同步评论原本是原始文本，我们使用神经语言处理工具 [38] 提取候选标签。我们合并来自不同来源的所有标签，并手动删除重复、无意义或包含脏话的标签。

• **预览图**。这些是从视频中提取的视觉特征。为了让预览图更具代表性，我们使用一种知名的关键帧提取工具[2] 来生成预览图。我们手动检查不同预览图之间的相似度，并删除重复项。最终，我们为每个视频生成五张预览图，供标注者使用。

除了标签和预览图之外，我们还向标注者展示视频标题和简要介绍。此外，如果标注者发现上述所有特征都无法表达自己的想法，也可以在预留的文本框中填写答案。

3.1.3 标注问题。如前所述，推荐解释可以服务于不同目的。对于内容提供方而言，解释可用于增强推荐说服力。对于用户而言，解释可用于提供更丰富的物品特征，帮助用户更快地做出决策。因此，解释可能存在多个真实标注。例如，在图 1 中，用户是汤姆·克鲁斯的粉丝。如果我们用"汤姆·克鲁斯"来解释该视频，用户可能会点击观看，从而达到内容提供方的目的。从这个角度来看，"汤姆·克鲁斯"是提升推荐说服力的真实标注。然而，虽然汤姆·克鲁斯出现在视频中，但他出现的时间很短，真正的主演是克里斯·派恩。如果我们用"克里斯·派恩"来解释该视频，用户就能做出更准确、更明智的决策。从这个角度来看，"克里斯·派恩"是增强推荐信息性的真实标注。为了从多个方面获取解释真实值，我们设计了一系列问题供标注者作答，并在下文介绍其中最重要的问题：

Q1. 哪些特征是您想观看这段视频的原因？ 此问题在用户观看视频之前提出。用户选择的特征即为用户期望，揭示了促使用户观看视频的原因。此问题旨在获取解释的真实标注，以提升推荐说服力。

Q2. 对于这段视频，哪些特征的信息量最大？ 此问题在用户观看视频之后提出。此时，用户已充分了解视频，因此所选特征能够客观地体现视频。通过此问题，我们希望获取解释的真实标注，以提升推荐信息性。

Q3. 您对哪些特征最满意？ 此问题同样在用户观看视频之后提出，旨在获取解释的真实标注，以提升用户满意度。与 Q1 相比，此真实标注是在用户充分了解视频内容之后获取的。与 Q2 相比，此真实标注包含用户的主观因素。

Q4. 请根据您的偏好为该视频评分（1 到 5 分）。此问题旨在收集用户对该视频的整体偏好。

Q5. 您如何评价这段视频？ 此问题旨在收集用户对该视频更详细的偏好，用户可以在文本框中表达自己的观点。

通过上述问题，我们从三个方面收集解释的真实标注数据，即说服力（Q1）、信息量（Q2）和用户满意度（Q3）。

[2]https://github.com/FFmpeg/FFmpeg

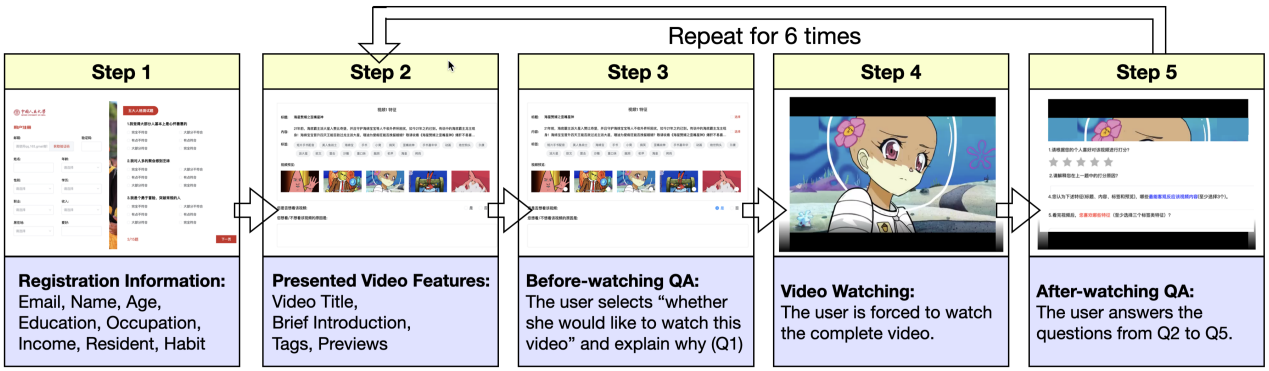

图2：为收集我们的数据集，每位用户的完整标注处理流程。

3.1.4 标注任务的工作流程。基于上述问题，本节介绍针对每位用户的完整标注流程（见图 2）。具体而言，该流程包含五个重要步骤：

步骤1：用户注册。首先，用户必须提供其基本个人信息，在我们的平台上注册。随后，我们还要求用户完成一项大五人格测试。该测试包含15个问题，这些问题均为标准问题，并在全球范围内得到广泛使用[3]。根据这些问题的答案，我们可以从五个维度了解用户的人格，即开放性、尽责性、外向性、宜人性和神经质。关于人格测试的更详细介绍，请参阅 [10, 11]。大五人格测试为我们研究不同用户人格对推荐解释选择倾向的影响提供了机会。对于上述个人信息，我们已获得用户许可，可将其用于研究目的。为保护用户隐私，我们已根据用户ID对数据集进行了脱敏处理。

步骤 2：平台推荐。用户登录平台后，系统会向其推荐三段视频。对于每段视频，用户还会看到第 3.1.2 节中介绍的特征。视频完全随机推荐，也就是说，每段视频被展示给用户的概率相同。此设置旨在收集更多样化的用户偏好，避免信息茧房效应。此外，我们的数据集也有助于去偏推荐领域的研究 [3], ，因为在现实场景中，收集均匀的物品曝光流量可能成本过高 [39]。

步骤 3：用户选择及观看前问答。收到推荐后，用户必须仔细查看每段视频的特征，并决定是否想观看该视频。如果用户决定观看视频，则需要回答第 3.1.3 节中的 Q1，即"哪些特征是你想观看这段视频的原因？"。否则，用户需要选择那些让自己不想观看该视频的特征。

步骤 4：观看视频。在这一步中，用户需要观看视频。由于视频时长很短（<3分钟），用户可以轻松快速地了解视频内容。对于用户在上一步中不想观看的视频，我们也会强制要求她观看这些视频，并在后续步骤中回答问题，这有助于研究用户感兴趣的视频是否会因不恰当的解释而被错过。

步骤 5：观看后答题。观看完视频后，我们要求用户回答一系列问题。首先，用户需要对视频进行评分并发表评论（对应 Q4 和 Q5）。接着，我们要求用户选择对视频最有信息量的特征（对应 Q2）。最后，用户需要选择自己最满意的特征（对应 Q3）。

对于每位用户，步骤 1 只需执行一次，步骤 2 至步骤 5 则至少重复六次。由于系统每次推荐三个视频，因此在整个标注流程中，每位用户会收到超过 3×6 （=18）条推荐。

3.2 数据集收集

基于上述平台，我们随后招募标注者使用该平台并完成上述标注流程。为保证数据集质量，我们制定了许多规则来清理标注结果。

3.2.1 标注者。 为保证数据集质量，我们与一家专业公司合作招募标注者。对于每位标注者，我们要求其此前曾受雇于该公司，并且具有良好的工作记录。此外，我们还要求标注者的背景足够多样化。我们总共招募了约 3,000 名标注者，并向每位完成标注任务的标注者支付 8 美元。我们的数据集总成本约为 28,000 美元（包括支付给该公司的费用）。

3.2.2 数据集质量控制。 在传统的标注任务中，例如图像识别或实体标注，真实标注是客观的，因此很容易判断标注结果是对还是错，这使得数据集质量控制并不困难。然而，我们的数据集旨在收集用户主观且个性化的偏好，不存在严格意义上的对错，这给数据集质量控制带来了困难。为解决这一问题，我们首先制定一系列规则，用于判断标注结果是否合理，然后移除不合理的结果。具体而言，这些规则主要围绕以下方面设计：

A1. 标注时间。在我们的平台上，用户回答每个问题所花费的时间都已记录在我们的数据库中。对于每个问题，我们收集所有用户的答题耗时，并检查

[3]https://www.psytoolkit.org/survey-library/big5-bfi-s.html

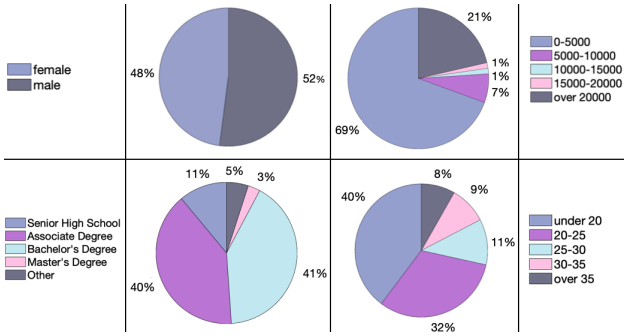

图3：用户统计信息。

是否存在异常值。对于所有问题，如果同一用户被认定为异常值超过五次，则会将该用户及其标注的所有结果从数据集中移除。

A2. 重复答案。对于这一方面，我们检查用户对不同问题的回答是否高度相似。例如，一些标注者可能会在所有视频的Q1中始终选择前三个标签。一些标注者可能会对不同视频使用相同的评论。通常，如果某个用户有超过三分之一的相似答案，我们就会移除该用户及其标注结果。

A3. 对同一问题的冲突答案。为了检查标注者是否认真完成了我们的任务，我们有意插入了许多标注者之前已经回答过的问题。对于同一名标注者，我们检查其对同一问题的回答是否一致。如果某个用户有超过三分之一的答案不一致，则会移除该用户。

A4. 不同问题之间的冲突答案。直观而言，如果标注者认真完成了我们的任务，那么不同答案之间应当存在一些合理的关系。例如，如果用户给出的评分非常低，那么评论大多也应当是负面的。起初，我们计划总结所有不合理的关系，并开发一个自动化工具来识别这些关系。然而，我们发现不同答案之间的关系模式过于复杂，而构建一个足够可靠的工具可能需要耗费太多时间。为了解决这些问题，我们招募了实验室中的五名学生，手动检查数据集中的每条记录。首先，我们要求每名学生独立检查数据集，并找出不合理的关系。然后，我们合并所有不合理的结果，让学生对每一项进行投票，以确定其是否合理。对于不合理的结果，我们会从最终数据集中移除相应的记录。

4 统计分析

本节介绍数据集的基本统计信息，重点探讨以下问题：（1）用户和视频的基本统计信息是什么？（2）用户-视频、用户-标签和视频-标签交互的统计信息是什么？（3）不同解释方面的真实标注之间有何相似之处？（4）倾向于选择单模态或多模态解释的用户有哪些统计特征？（5）用户评论的长度有哪些统计特征?

我们的数据集总计包含2997位用户、4672个视频和6115个标签。在图3中，我们根据用户的性别、收入、教育程度和年龄展示了用户统计数据。我们可以看到，用户

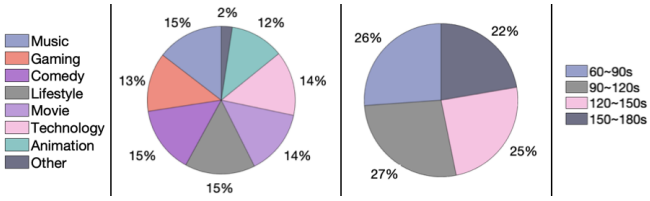

图4：视频的统计信息。

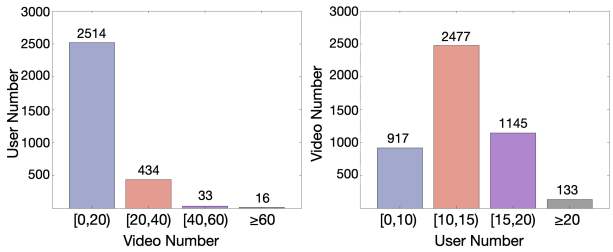

图5：用户-视频交互的分布。

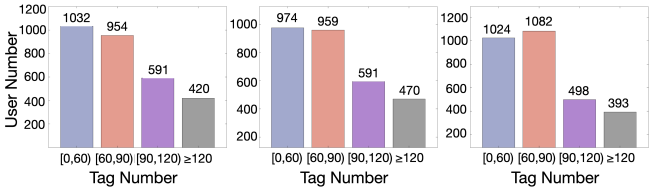

图6：用户-标签交互的分布。

覆盖了不同的人群，这使得我们的数据集更具普适性和代表性。在图4中，我们根据视频的类别和长度展示了视频统计数据。我们可以看到，视频的类别多种多样，并且对于不同类别，视频数量几乎相同，这有助于探索更加全面且无偏的用户偏好。

图5展示了用户-视频交互的分布。每个点（x，y）表示有 y 个用户（或视频）与 x 个视频（或用户）发生交互。为便于展示，我们将视频（或用户）的数量划分为四组。图6展示了用户-标签交互的分布。如第3.1.3节所述，Q1、Q2和Q3各有三种类型的标签，因此我们分别展示。图中的每个点（x，y）表示有 y 个用户在不同视频中总共选择了 x 个标签，其中 x 轴也分为四组。图7展示了视频-标签交互的分布。与图6类似，我们分别展示 Q1、Q2 和 Q3 的结果。每个点（x，y）表示有 y 个视频，其标签被不同用户总共选中了 x 次。

在我们的数据集中，我们为多方面解释标注了真实标注。一个有趣的问题是，用户针对同一视频，在不同解释方面选择的标签有何不同。为回答这个问题，我们基于杰卡德距离度量标签相似度。对于每两个解释方面，我们首先计算每个用户-物品对的标签相似度，然后对表 2 中所有用户-视频对的结果取平均并进行汇报。我们可以看到，对于不同的解释目的，重叠部分

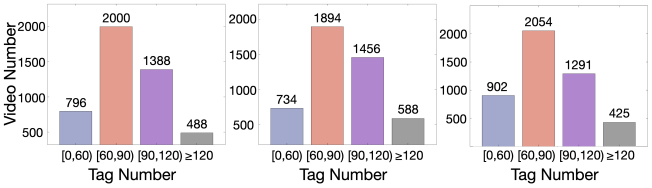

图7：视频-标签交互的分布。

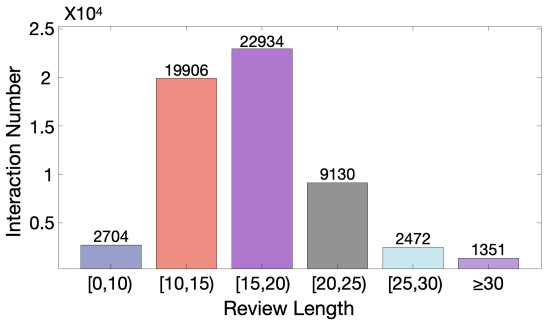

图8：用户评论长度的分布。

表2：不同解释方面的平均杰卡德相似度。

解释方面	Q1-Q2	Q1-Q3	Q2-Q3
杰卡德相似度	0.423	0.364	0.410

表3：选择多模态解释的用户数与其他用户数之比。

解释方面	Q1	Q2	Q3
比例	0.389	0.390	0.359

所选标签之间的重叠通常并不大（约为 0.35 到 0.45）。这表明，为了实现不同的解释目的，可能需要使用不同的标签。

在我们的数据集中，用户可以选择多模态解释。在此，我们计算选择多模态解释的用户数与所有用户数之比。我们在表3中分别展示了 Q1、Q2 和 Q3 的结果，可以看到，约有 38% 的用户倾向于同时选择文本解释和视觉解释，而其他用户则希望选择单模态解释。最后，我们在图8中展示了用户评论长度的分布。可以看到，大多数评论都超过十个词，能够全面揭示用户偏好。

5 库

除了上述数据集之外，我们还开发了一个库，在统一框架中实现了十种知名的可解释推荐模型。本节首先介绍该库的架构，然后详细介绍已实现的模型。最后，我们展示一些如何使用该库的示例。

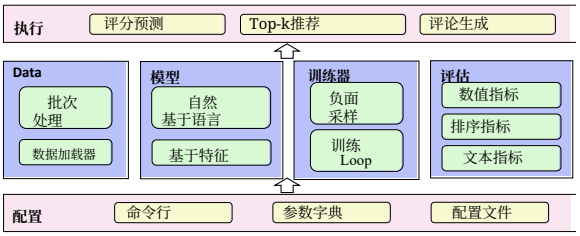

图9：我们的库的结构。该库包含六个部分：配置、数据、模型、训练器、评估和执行模块。

5.1 库的结构

我们的库的结构如图9所示。配置模块是库的基础部分，负责初始化所有参数。我们支持三种指定参数的方式，即命令行、参数字典和配置文件。基于配置模块，库还包含四个更高层的模块：

• **数据模块**：该模块旨在将原始数据转换为模型输入。它包含两个组件：第一个组件负责加载数据并为用户评论构建词汇表。第二个组件旨在将数据处理成模型输入所需的格式，并生成用于模型优化的样本批次。

• **模型模块**：该模块旨在实现可解释推荐模型。我们的库中包含两类方法。第一类包括基于特征的可解释推荐模型，第二类则包含带有自然语言解释的模型。下一节将详细介绍这些模型。

• **训练器模块**：该模块用于实现训练损失，例如贝叶斯个性化排序（BPR）和二元交叉熵（BCE）。此外，该模块还可以记录完整的模型训练过程。

• **评估模块**：该模块用于评估不同的模型，包含三种类型的评估任务，即评分预测、Top-k推荐和评论生成。

在上述四个模块之上，最上层设有一个执行模块。该模块负责针对不同任务优化推荐模型，例如评分预测、标签预测和评论生成。有关我们库架构的更详细介绍，请参阅我们的项目：https://reasoner2023.github.io/。

5.2 已实现的模型

在我们的库中，我们实现了两类在研究领域受到广泛研究的可解释推荐模型。第一类是基于特征的可解释推荐模型，其中的特征可以是标签、物品方面等。第二类是带有自然语言解释的模型。

具体而言，我们实现了以下具有代表性的基于特征的可解释推荐模型：

•**EFM** [37] 根据评论信息中的显式产品特征和用户意见预测用户偏好，并生成可解释的推荐。

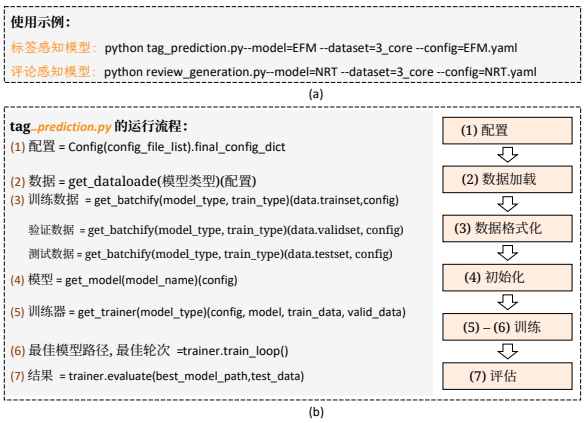

图 10：（a）基于特征和评论的模型示例。（b）我们的库的运行流程。

• **TriRank** [12] 基于用户评分和评论，将用户-物品-方面三元关系建模为异构三部图，并设计了一种用于推荐的顶点排序算法。

• **LRPPM** [5] 是一种张量矩阵分解算法，通过针对各种物品方面采用基于排序的优化目标来捕捉用户偏好。

• **SULM** [1] 通过方面级情感分析，不仅推荐物品，还推荐具体方面，从而改进推荐效果。

• **MTER** [34]是一种张量分解方法，使用用户、物品和特征构成的三阶张量来建模物品推荐任务。为适配我们的数据，我们在原始实现中省略了观点建模。

• **AMF** [14] 利用从用户评论方面提取的辅助信息来提高推荐准确率。

除了上述基于矩阵分解的浅层模型外，我们还实现了以下深度特征型可解释推荐模型（简称**DERM–**）：

• **DERM-MLP** 是一种用于联合预测评分和标签的深度推荐模型。两个任务共享一组用户/物品/标签嵌入。隐藏状态和标签嵌入分别输入对应不同任务的不同层。

• **DERM-MF** 首先基于用户/物品嵌入，通过矩阵分解获得隐藏状态，然后由神经网络计算输出。

• **DERM-C** 结合矩阵分解和多层感知器（MLP）来得到隐藏状态，并以拼接的方式合并输出。

• **DERM-H** 利用标签为用户和物品建立画像，然后使用与 DERM-MLP 相同的架构来预测评分和标签。

对于生成自然语言解释的模型，我们实现了以下几种代表性方法：

• **Att2Seq** [8] 是一种评论生成模型，它使用长短期记忆网络作为解码器，并直接基于用户/物品 ID 和评分信息生成文本。

• **NRT** [19] 根据输入的用户-物品对同时预测评论和评分，其中两个任务共享相同的嵌入和隐藏层。

表 4：基于特征的可解释推荐模型在预测说服力标签和评分方面的基准测试结果。

指标	说服力		评分预测	
	F1	NDCG	RMSE	MAE
EFM	38.96	24.02	2.291	1.876
TriRank	18.30	13.84	2.895	2.580
LRPPM	37.23	23.21	**1.223**	**0.961**
SULM	**42.57**	**25.80**	1.783	1.404
MTER	8.164	3.545	1.288	1.024
AMF	31.29	19.31	2.282	1.858
DERM-MLP	37.56	23.33	1.318	1.078
DERM-MF	36.80	22.81	1.317	1.136
DERM-C	37.01	23.12	1.323	1.110
DERM-H	37.03	23.15	1.288	1.045

• **PETER** [17] 利用 Transformer 生成用户评论，是一种最先进的评论生成模型。

5.3 使用我们的库示例

本节介绍如何使用我们的库。图 10(a) 展示了一个简单示例，用户可以直接运行 *tag_prediction.py* 或 *review_generate.py*，分别运行基于特征或基于评论的模型。在这些命令中，需要指定三个参数，分别表示模型、数据集和配置文件的名称。

以 *tag_prediction.py* 为例，它会依次执行以下步骤：(1) 配置。在此步骤中，将来自不同来源（命令行、配置字典和文件）的模型架构和优化过程相关参数整合到一个字典中。(2) 数据加载。在此步骤中，根据模型类型选择数据加载器。对于评论感知模型，此步骤会读取所有记录并构建词汇表。(3) 数据格式化。以样本批次的形式，将训练集、验证集和测试集处理为模型输入所需的格式。(4) 初始化。根据配置中的参数值定义并初始化相应的模型类。(5)-(6) 训练。选择优化方法来训练模型。(7) 评估。衡量模型在不同推荐/解释任务上的表现。

我们的库具有很强的可扩展性，实现一个新模型只需三个步骤：（1）实现模型的基本功能，包括模型架构、偏好分数预测等；（2）在*train.py*中自定义训练方法；（3）在配置文件中指定参数。

6 基准测试

6.1 实验设置

考虑到解释有三种类型的真实标注，我们分别通过预测 Q1、Q2、Q3、Q1+Q2、Q2+Q3、Q1+Q3 和 Q1+Q2+Q3 的标签来评估模型性能。需要预测多种类型的真实标注时，我们通过共享嵌入参数，将原始模型扩展为多任务版本。我们按照 8:1:1 的比例随机将数据集划分为训练集、验证集和测试集。

表5：基于特征的可解释推荐模型在预测标签的信息量和评分方面的基准测试结果。

指标	信息量		评分预测	
	F1	NDCG	RMSE	MAE
EFM	9.120	8.030	1.914	1.406
TriRank	18.80	14.43	2.895	2.580
LRPPM	37.83	38.39	**1.225**	**0.963**
SULM	**42.79**	**42.76**	1.634	1.276
MTER	36.82	36.96	1.347	1.081
AMF	25.04	25.90	2.282	1.857
DERM-MLP	34.95	35.60	1.411	1.170
DERM-MF	37.41	35.89	1.317	1.138
DERM-C	38.02	38.30	1.319	1.105
DERM-H	36.01	35.79	1.267	1.036

表 6：基于特征的可解释推荐模型在预测满意度标签和评分方面的基准测试结果。

指标	满意度		评分预测	
	F1	NDCG	RMSE	MAE
EFM	6.021	2.295	1.911	1.403
TriRank	16.91	13.23	2.895	2.580
LRPPM	35.97	22.35	**1.223**	**0.960**
SULM	**40.57**	**24.77**	1.635	1.277
MTER	6.635	2.687	1.297	1.033
AMF	30.61	18.67	2.282	1.859
DERM-MLP	36.42	22.37	1.330	1.085
DERM-MF	35.76	21.76	1.317	1.137
DERM-C	35.74	21.89	1.308	1.098
DERM-H	34.11	19.39	1.319	1.080

表 7：基于特征的可解释推荐模型在联合预测说服力、信息量和评分标签方面的基准测试结果。

指标	说服力		信息量		评分	
	F1	NDCG	F1	NDCG	均方根误差	平均绝对误差
EFM	12.96	9.173	9.132	8.084	1.911	1.406
LRPPM	37.20	23.20	37.85	38.39	**1.223**	**0.962**
SULM	**41.53**	**25.74**	**42.73**	**42.77**	1.639	1.281
MTER	5.562	2.120	6.013	4.158	1.558	1.299
AMF	30.04	18.78	30.79	30.59	2.282	1.856
DERM-MLP	35.59	21.92	33.78	34.82	1.352	1.104
DERM-MF	36.54	22.56	37.53	37.31	1.318	1.139
DERM-C	38.09	23.60	38.83	39.05	1.311	1.102
DERM-H	34.27	20.97	37.91	38.17	1.266	1.034

对于评论生成任务，我们使用提及频率最高的 20,000 个词构建词汇表，并将生成句子的最大长度设为 17，这与数据集中评论的平均长度相同。对于所有模型，批量大小设为 256。我们通过网格搜索调整其他关键超参数。具体而言，我们分别在 $[0\ 1, 0\ 01, 0\ 001, 0\ 0001]$ 和 $[0\ 001, \ldots, 0\ldots0001, 0]$ 范围内调整学习率和 L2 正则化的权重。对于深度模型，我们在 $[32, 64, 128, 256]$ 和 $[1, 2, 3, 4]$ 范围内调整隐藏层大小和层数

表8：基于特征的可解释推荐模型在联合预测说服力、满意度和评分标签方面的基准测试结果。

指标	说服力		满意度		评分	
	F1	NDCG	F1	归一化折损累计增益	均方根误差	平均绝对误差
EFM	11.01	6.599	7.303	3.239	1.910	1.405
LRPPM	37.21	23.21	35.97	22.35	**1.223**	**0.961**
SULM	**41.50**	**25.74**	**40.49**	**24.76**	1.640	1.281
MTER	6.586	3.101	6.411	2.536	1.288	1.024
AMF	24.95	16.18	24.33	15.78	2.282	1.857
DERM-MLP	38.41	23.73	37.38	22.81	1.328	1.082
DERM-MF	36.62	22.57	35.60	21.37	1.317	1.137
DERM-C	38.37	23.67	37.19	22.74	1.321	1.116
DERM-H	37.84	23.53	36.81	22.65	1.264	1.025

表 9：基于特征的可解释推荐模型在联合预测信息量、满意度和评分标签方面的基准测试结果。

指标	信息量		满意度		评分	
	F1	NDCG	F1	NDCG	均方根误差	平均绝对误差
EFM	9.156	8.086	4.849	1.950	1.910	1.406
LRPPM	37.84	38.39	35.98	22.35	**1.223**	**0.961**
SULM	**42.64**	**42.72**	**40.42**	**24.74**	1.640	1.281
MTER	6.914	4.920	5.719	2.162	1.288	1.024
AMF	30.72	30.49	29.09	18.08	2.282	1.856
DERM-MLP	38.97	39.28	37.30	22.82	1.317	1.078
DERM-MF	37.53	37.25	35.57	21.35	1.318	1.139
DERM-C	38.83	39.11	37.00	22.68	1.309	1.106
DERM-H	38.61	38.93	36.88	22.69	1.261	1.022

具体而言，实验设置的更多细节请参见我们的项目，该项目已发布于 https://reasoner2023.github.io/。我们使用均方根误差和平均绝对误差作为指标，评估评分预测任务的性能。对于标签预测任务，我们选用 F1值和归一化折损累计增益来评估模型性能。为评估生成评论的质量，我们采用包括双语评估替补 [27] 和面向召回率的摘要评估 [23] 在内的指标进行模型比较。

6.2 实验结果

表4-10展示了基于特征的可解释推荐模型在标签预测和评分预测任务上的比较结果。表11展示了具有自然语言解释的模型在评论生成和评分预测任务上的比较结果。我们使用预测得分最高的前10个标签来计算 F1值和归一化折损累计增益，结果以百分比表示，并省略“%”符号。对于均方根误差和平均绝对误差，数值越低表示性能越好。对于每个评估指标，我们使用粗体标出最佳性能。由于我们实现的 TriRank 不支持同时预测多种类型的标签，因此在相应的表格中将其省略。

7 个新机遇

我们相信，我们的数据集能为可解释推荐领域带来以下新机遇：

• **多方面可解释推荐**。 对于几乎所有以往的可解释推荐模型，数据集中解释的真实标注都只涉及一个方面。因此，学到的模型可能无法在其他方面表现良好。对于

表10：基于特征的可解释推荐模型在联合预测说服力、信息量、满意度和评分标签方面的基准测试结果。

指标	说服力		信息量		满意度		评分预测	
	F1	NDCG	F1	NDCG	F1	NDCG	RMSE	MAE
EFM	38.96	24.02	5.433	3.993	6.323	3.101	2.291	1.876
LRPPM	38.65	23.67	39.26	39.35	37.39	22.89	1.484	1.217
SULM	**41.31**	**25.70**	**42.52**	**42.73**	**40.26**	**24.68**	1.691	1.333
MTER	35.00	21.84	35.78	36.09	33.94	20.66	1.889	1.554
AMF	31.72	19.42	32.65	31.98	31.20	18.85	2.282	1.858
DERM-MLP	38.65	23.87	39.33	39.55	37.61	23.02	1.324	1.083
DERM-MF	36.54	22.21	37.48	36.94	35.50	21.08	1.320	1.144
DERM-C	38.43	23.71	39.91	39.39	37.46	22.86	1.316	1.109
DERM-H	38.29	23.69	38.97	39.22	37.24	22.84	**1.263**	**1.029**

表11：我们的库中带有自然语言解释的模型的基准测试结果。对于双语评估替补和面向召回率的摘要评估，结果为百分比值，省略了"%"。"-"表示该模型不提供此评估指标。

指标	双语评估替补 (%)		ROUGE-1（%）			ROUGE-2（%）			评分预测	
	BLEU-1	BLEU-4	F1	召回率	精确率	F1	召回率	精确率	RMSE	MAE
Att2Seq	20.30	**3.616**	22.50	20.51	26.15	5.718	5.404	6.311	-	-
NRT	**20.89**	3.477	**23.10**	**21.23**	26.71	**5.818**	**5.528**	6.395	1.214	**0.917**
PETER	16.56	2.488	20.57	15.94	**29.80**	5.035	3.983	**7.028**	**1.200**	0.943

例如，如果模型仅针对提高推荐的说服力进行训练，那么生成的解释可能无法增强推荐的信息量（甚至会造成损害）。借助我们的数据集，人们可以同时考虑不同的解释方面，并学习更全面的可解释模型，以服务在线用户。我们认为这一方向仍有很大的探索空间，例如，不同的解释目的之间是否存在冲突；如果存在，应如何设计模型来更好地克服这些冲突。

• **多模态可解释推荐**。据我们所知，先前大多数可解释推荐模型仅考虑单模态解释。然而，在现实场景中，用户总是需要感知多模态信息。借助我们的数据集，人们可以围绕多模态解释开展许多研究，例如：（1）用户倾向于选择文本特征还是视觉特征？（2）如何构建一个统一模型，同时生成文本解释和视觉解释？（3）用户人格与解释模态之间有何关系？如何根据用户角色信息构建模型来预测解释模态?

• **包含丰富角色信息的可解释推荐**。在以往的大多数研究中，用于训练可解释推荐模型的数据集都不包含足够的角色信息。借助我们的数据集，人们可以获取经过脱敏处理的用户档案，这或许能推动许多研究方向。例如，可以研究解释公平性，其中敏感变量可以是年龄、教育程度、收入等。此外，人们还可以利用丰富的角色信息来提高解释的预测准确率。

除了上述有关可解释推荐的研究机会外，我们的数据集也能为其他研究领域作出贡献。例如，由于我们的数据集是在统一的物品曝光概率下收集的，人们可以使用我们的数据集研究去偏推荐。在我们的数据集中，我们收集了用户的大五人格信息，因此可以用它来研究心理学信息驱动的推荐。

8 结论与未来工作

本文介绍了一个新的可解释推荐数据集，详细阐述了其构建过程，并对其特征进行了深入分析。我们的数据集支持多模态解释，其中真实标注人员正是产生这些交互的真实用户。此外，我们的数据集还支持多方面解释，包含约 3,000 名用户。我们开发了一个可解释推荐工具包，并基于上述数据集针对不同的可解释推荐任务构建了多个基准测试。此外，我们还分析了该数据集带来的新机遇。事实上，我们的数据集只是迈向更具可测量性的可解释推荐的一小步，未来仍有很长的路要走。今后，我们计划将数据集扩展到其他领域，例如新闻、音乐等。此外，我们还希望聘用更多标注者，以扩大数据集规模。

9 致谢

我们衷心感谢来自中国人民大学的 Zhenlei Wang、Rui Zhou、Kun Lin、Zeyu Zhang、Jiakai Tang 和 Hao Yang，他们为检查数据集构建过程中不合理的结果作出了贡献。本研究部分得到了国家自然科学基金委员会（编号：62102420）、北京优秀青年科学家计划（编号：BJJWZYJH012019100020098）、智能社会治理平台以及中国人民大学"双一流"建设跨学科重大创新规划平台的支持。

参考文献

[1] Konstantin Bauman、Bing Liu 和 Alexander Tuzhilin。2017。基于方面的推荐：根据用户评论推荐具有最有价值方面的物品。载于第23届ACM SIGKDD知识发现与数据挖掘国际会议论文集。717–725。[2] Chong Chen、Min Zhang、Yiqun Liu 和 Shaoping Ma。2018。基于评论级解释的神经注意力评分回归。载于2018年万维网大会论文集。1583–1592。[3] Jiawei Chen、Hande Dong、Xiang Wang、Fuli Feng、Meng Wang 和 Xiangnan He。2020。推荐系统中的偏差与去偏：综述及未来方向。ACM信息系统汇刊（2020）。[4] Xu Chen、Hanxiong Chen、Hongteng Xu、Yongfeng Zhang、Yixin Cao、Zheng Qin 和 Hongyuan Zha。2019。基于多模态注意力网络、带视觉解释的个性化时尚推荐：迈向视觉可解释推荐。载于第42届国际ACM SIGIR信息检索研究与发展大会论文集。765–774。[5] Xu Chen、Zheng Qin、Yongfeng Zhang 和 Tao Xu。2016。学习对多个类别的推荐特征进行排序。载于第39届国际ACM SIGIR信息检索研究与发展会议论文集。305–314。[6] Xu Chen、Yongfeng Zhang 和 Zheng Qin。2019。基于神经注意力模型的动态可解释推荐。载于AAAI人工智能会议论文集，第33卷。53–60。[7] Xu Chen、Yongfeng Zhang、Hongteng Xu、Yixin Cao、Zheng Qin 和 Hongyuan Zha。2018。视觉可解释推荐。arXiv预印本arXiv:1801.10288（2018）。[8] Li Dong、Shaohan Huang、Furu Wei、Mirella Lapata、Ming Zhou 和 Ke Xu。2017。学习根据属性生成产品评论。载于第15届计算语言学协会欧洲分会会议论文集：第1卷，长篇论文。623–632。[9] Fatih Gedikli、Dietmar Jannach 和 Mouzhi Ge。2014。我该如何解释？推荐系统中不同解释类型的比较。《国际人机研究杂志》72，4（2014），367–382。[10] Lewis R Goldberg。1992。大五人格因素结构标记的发展。《心理评估》4，1（1992），26。[11] Samuel D Gosling、Peter J Rentfrow 和 William B Swann Jr.。2003。大五人格维度的极简测量。《人格研究杂志》37，6（2003），504–528。[12] Xiangnan He、Tao Chen、Min-Yen Kan 和 Xiao Chen。2015。TriRank：通过建模方面实现评论感知的可解释推荐。载于第24届ACM信息与知识管理国际会议论文集。1661–1670。[13] Diana C et al. Hernandez-Bocanegra。2020。论证式解释类型对基于评论的推荐感知的影响。载于UMAP。[14] Yunfeng Hou、Ning Yang、Yi Wu 和 Philip S Yu。2019。融合方面信息的可解释推荐。《万维网》22（2019），221–240。[15] Johannes Kunkel、Tim Donkers、Lisa Michael、Catalin-Mihai Barbu 和 Jürgen Ziegler。2019。让我来解释：个性化与非个性化解释对推荐系统信任度的影响。载于2019年CHI计算系统中的人为因素大会论文集。1–12。[16] Lei Li、Yongfeng Zhang 和 Li Chen。2021。Extra：用于可解释推荐的解释排序数据集。载于第44届国际ACM SIGIR信息检索研究与发展大会论文集。2463–2469。[17] Lei Li、Yongfeng Zhang 和 Li Chen。2021。用于可解释推荐的个性化Transformer。arXiv预印本arXiv:2105.11601（2021）。[18] Lei et al. Li。2020。通过神经模板实现推荐系统中的可控解释生成。载于TheWebConf。[19] Piji Li、Zihao Wang、Zhaochun Ren、Lidong Bing 和 Wai Lam。2017。用于推荐的神经评分回归与抽象式提示生成。载于第40届国际ACM SIGIR信息检索研究与发展大会论文集。345–354。[20] Yang et al. Li。2021。属性感知的可解释互补服饰推荐。《万维网》（2021）。[21] Zhenyu Liao、Yikun Xian、Xiao Yang、Qinpei Zhao、Chenxi Zhang 和 Jiangfeng Li。2018。TSCSet：用于探索改善用户体验的众包时间同步评论数据集。载于第23届智能用户界面国际会议。641–652。[22] Christopher Lin、Ece Kamar 和 Eric Horvitz。2014。沉默中的信号：众包推荐系统中的隐式反馈模型。载于AAAI人工智能会议论文集，第28卷。[23] Chin-Yew Lin。2004。面向召回率的摘要评估：一种摘要自动评估工具包。载于《文本摘要的新进展》。74–81。[24] Cataldo Musto、Pasquale Lops、Marco de Gemmis 和 Giovanni Semeraro。2019。通过对用户评论进行基于方面的情感分析来说明推荐理由。载于第27届ACM用户建模、自适应与个性化会议论文集。4–12。[25] Sidra 等人 Naveed。2020年。关于基于特征的协作式解释的运用：解释风格的实证比较。载于*UMAP*。[26] Charbel Obeid、Inaya Lahoud、Hicham El Khoury和Pierre-Antoine Champin。2018年。高等教育中的基于本体的推荐系统。载于*万维网会议2018*配套论文集。1031–1034。[27] Kishore Papineni、Salim Roukos、Todd Ward和Wei-Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集。311–318。[28] J Ben Schafer、Joseph Konstan和John Riedl。1999年。电子商务中的推荐系统。载于第*1*届*ACM*电子商务会议论文集。158–166。[29] Amit Sharma和Dan Cosley。2013年。社会化解释有效吗？研究并建模社会化解释对推荐系统的影响。载于第*22*届万维网国际会议论文集。1133–1144。[30] Juntao Tan、Shuyuan Xu、Yingqiang Ge、Yunqi Li、Xu Chen和Yongfeng Zhang。2021年。反事实可解释推荐。载于第*30*届*ACM*信息与知识管理国际会议论文集。1784–1793。[31] Yiyi Tao、Yiling Jia、Nan Wang和Hongning Wang。2019年。Fact：借助因子分解树驯服潜因子模型以实现可解释性。载于*SIGIR*。295–304。[32] Nava Tintarev和Judith Masthoff。2007年。推荐系统解释综述。载于*2007*年*IEEE*第*23*届数据工程国际会议研讨会。IEEE，801–810。[33] Thi Ngoc Trang Tran、Alexander Felfernig、Christoph Trattner和Andreas Holzinger。2021年。医疗保健领域的推荐系统：现状与研究问题。《智能信息系统杂志》57（2021），171–201。[34] Nan Wang、Hongning Wang、Yiling Jia和Yue Yin。2018年。通过多任务学习处理带观点文本数据，实现可解释推荐。载于第*41*届*ACM SIGIR*信息检索研究与发展国际会议。165–174。[35] Bingbing Wen、Yunhe Feng、Yongfeng Zhang和Chirag Shah。2022年。ExpScore：学习推荐解释指标。载于*ACM Web Conference 2022*论文集。3740–3744。[36] Yikun 等人 Xian。2021年。Ex3：可解释的属性感知物品集推荐。载于*RecSys*。[37] Yongfeng Zhang、Guokun Lai、Min Zhang、Yi Zhang、Yiqun Liu和Shaoping Ma。2014年。基于短语级情感分析的可解释推荐显式因子模型。载于第*37*届*ACM SIGIR*信息检索研究与发展国际会议论文集。83–92。[38] Yongfeng Zhang、Haochen Zhang、Min Zhang、Yiqun Liu和Shaoping Ma。2014年。用户是评分还是撰写评论？利用评论级情感分类提升短语级情感标注。载于第*37*届*ACM SIGIR*信息检索研究与发展国际会议论文集。1027–1030。[39] Chang Zhou、Jianxin Ma、Jianwei Zhang、Jingren Zhou和Hongxia Yang。2021年。面向大规模推荐系统的去偏候选生成对比学习。载于第*27*届*ACM SIGKDD*知识发现与数据挖掘会议论文集。3985–3995。[40] Renjie Zhou、Samamon Khemmarat和Lixin Gao。2010年。YouTube推荐系统对视频观看次数的影响。载于第*10*届*ACM SIGCOMM*互联网测量会议论文集。404–410。