arXiv:2304.00969v1 [cs.AI] 3 Apr 2023

越多是否总是更好？个人特征和详细程度对推荐系统中解释感知的影响

MOHAMED AMINE CHATTI，杜伊斯堡-埃森大学，德国
MOUADH GUESMI，杜伊斯堡-埃森大学，德国LAURA VORGERD，杜伊斯堡-埃森大学，德国THAO NGO，杜伊斯堡-埃森大学，德国SHOEB JOARDER，杜伊斯堡-埃森大学，德国QURAT UL AIN，杜伊斯堡-埃森大学，德国ARHAM MUSLIM，国立科技大学，巴基斯坦

尽管人们已经认识到，不同终端用户对解释的感知可能存在很大差异，但可解释推荐系统（RS）传统上一直采用一刀切模型，即向每位用户提供相同详细程度的解释，而不考虑用户的个体情境，也就是其目标和个人特征。为填补这一研究空白，本文旨在推动可解释推荐从一刀切模式转向个性化方法，让用户能够自主选择想要查看的解释。我们开发了透明的推荐与兴趣建模应用程序（RIMA），可按需提供推荐的个性化解释，并设置基础、中级和高级三个详细程度层级，以满足不同类型终端用户的需求。我们开展了一项被试内研究（N=31），旨在探究用户个人特征与解释详细程度之间的关系，以及这两个变量对用户针对不同解释目标感知可解释推荐系统的影响。结果表明，解释目标和用户类型对用户感知不同详细程度的可解释推荐系统所产生的影响各不相同。因此，我们提出了一些理论和设计指南，以支持根据用户情境系统地设计推荐系统中的解释型界面。

CCS 概念：• **以人为中心的计算** → **交互系统与工具**；• **计算方法** → **人工智能**。

补充关键词与短语：智能解释界面；推荐系统；可解释推荐；个性化解释；个人特征

ACM 参考文献格式：
Mohamed Amine Chatti、Mouadh Guesmi、Laura Vorgerd、Thao Ngo、Shoeb Joarder、Qurat Ul Ain 和 Arham Muslim。2022年。越多是否总是更好？个人特征和详细程度对推荐系统中解释感知的影响。载于第*30*届 *ACM* 用户建模、自适应与个性化会议论文集（*UMAP '22*），*2022*年*7*月*4*日至*7*日，西班牙巴塞罗那。ACM，纽约，纽约州，美国，16页。https://doi.org/10.1145/3503252.3531304

特此授权，允许为个人或课堂使用而免费制作本作品的全部或部分内容的数字副本或纸质副本，前提是这些副本不得用于营利或商业用途，也不得为此制作或传播，并且副本首页须包含本声明和完整引文。对于本作品中归其他方面而非 ACM 所有的组件，必须尊重其版权。允许在注明出处的情况下摘录。若要以其他方式复制、重新出版、发布到服务器或重新分发至列表，须事先获得特定许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2022年 计算机协会。稿件提交至 ACM

1 引言

近年来，推荐系统（RS）的可解释性受到广泛关注。可解释推荐是指个性化推荐算法不仅向用户提供推荐项目，还提供解释，让用户了解为何推荐这些项目 [36]。解释旨在回答问题，这些问题也称为可理解性查询或类型，例如是什么、为什么、如何、如果.....会怎样、为什么不、如何做以及还有什么 [21, 26]。可解释推荐研究聚焦于不同维度和设计选择。除了可理解性类型之外，这些维度和选择还包括：(a) 解释目标（例如透明度、可审查性、信任、有效性、说服力、效率、满意度），(b) 解释风格（例如基于内容、基于协作、社交、混合），(c) 解释范围（即输入：用户模型；过程：算法；输出：推荐项目），以及 (d) 解释形式（例如文本、视觉） [2, 27, 33, 36]。可解释推荐中另一项至关重要的设计选择，涉及应向最终用户提供多详细的解释。不同的解释设计选择，例如解释风格、范围、格式和详细程度，会受到解释目标和用户类型的影响 [26]。用户可能对解释能够提供的所有信息都不感兴趣 [25]。不同用户对解释有不同需求；如果解释难以理解，可能会造成负面影响（例如认知负荷过高、困惑、缺乏信任） [9, 14, 17, 35, 37]。因此，重要的是提供足够详细的解释，使用户能够准确构建关于推荐系统如何运行的心理模型，同时又不至于让他们不堪重负。

个体用户差异和人为因素对用户使用可解释推荐系统时行为的影响，直到最近才开始受到研究 [13, 16, 24, 29]。这些研究表明，个人特征可能会影响用户对解释的感知，并促使研究者根据不同用户的情境（即目标和个人特征），为其选择不同的解释信息（即内容）。然而，就设计选择（即可理解性类型、解释风格、范围、格式或详细程度）而言，目前大多数可解释推荐系统的设计仍遵循一刀切方法，没有尝试识别和满足不同用户的需求与偏好。由于可解释推荐中的设计选择会受到用户情境的影响，一个顺理成章的方向是改进现有解释技术，以满足不同类型终端用户的需求。与推荐本身一样，解释也应针对不同终端用户进行个性化 [24]。因此，预期可解释推荐系统能够为相应的用户群体提供恰当的解释 [26]。这要求可解释推荐从一刀切方法转向个性化方法，以满足不同用户的需求与偏好。

本文尤其关注解释详细程度这一可解释推荐中的重要设计选择。我们认识到，单独考量解释详细程度和用户的个人特征通常是不够的，因此开展了一项用户研究，探究这两个因素之间的依赖关系，以及它们对用户对不同解释目标（透明度、可审查性、信任、有效性、说服力、效率和满意度）的感知所产生的影响。基于研究结果，我们总结了一些设计指南，供设计具有不同详细程度的解释时参考，以确保其与解释目标和用户的个人特征相契合。

为开展这项研究，我们开发了一款透明的推荐与兴趣建模应用程序（RIMA），可按需提供推荐的个性化解释，并设有三种不同的详细程度（基础、中级、高级），以满足不同用户的需求和偏好。本研究旨在回答以下研究问题：**个人特征如何影响用户对不同解释目标下解释详细程度的感知?**

研究结果表明：（1）提供不同详细程度的解释对于满足不同用户的需求十分重要；（2）解释应针对特定的解释目标和特定的用户类型进行设计；（3）解释内容应根据用户数据量身定制。

本文的主要贡献有两点：首先，我们将推荐系统中的个性化解释提升至设计选择层面，通过提供详细程度不同的按需解释来实现这一点。其次，我们为解释目标、用户类型与解释详细程度之间存在依赖关系提供了证据。

本文组织如下。首先，我们概述本研究的背景并讨论相关工作。随后，介绍 RIMA 中不同解释的实现。之后，我们描述用户研究、呈现结果，并讨论研究发现的意义。基于这些发现，我们进一步提出在推荐系统中提供个性化解释的设计指南。最后，我们指出局限性，总结本文工作，并概述未来的研究计划。

2 相关工作

2.1 不同详细程度的解释

解释的详细程度是可解释推荐系统设计过程中的一个重要因素。本文中，详细程度指解释中呈现的信息量。总体而言，在可解释人工智能（XAI）领域，Mohseni 等人 [26]认为，不同用户群体在使用 XAI 系统时会有不同的目标。同样，Miller [25] 认为，提供生成特定推荐的确切算法未必是最好的解释。人们往往不会根据解释的生成过程来判断其质量，而是看它是否有用。

除了用户的目标之外，另一个影响他们理解解释的重要因素是其认知能力。以往关于可解释人工智能的研究结果表明，对于特定用户或用户群体而言，提供详细解释并不一定会带来更高的信任和用户满意度，因为额外解释的提供会增加认知努力 [14, 17, 35, 37]。Kulesza 等人 [17] 概述了一套用于设计解释的原则，以实现交互式机器学习的个性化。这些原则包括“可靠”“完整”和“不要让人应接不暇”，这意味着解释中的信息量与用户在与人工智能系统交互时所形成的感知透明度和信任之间存在权衡。*可靠性* 是指只陈述事实。它指的是解释忠实度，即“解释内容的每个组成部分在描述底层系统时的真实性程度”。评估可靠性需要将解释与学习系统的数学模型进行比较，“解释越能反映底层模型，其可靠性就越高”。*完整性* 是指陈述全部事实。它指的是“解释对底层系统的描述程度”。完整的解释会告知用户学习系统掌握的全部信息，以及系统如何使用这些信息。作者指出，评估完整性的一种方法是采用 Lim 和 Dey 的可理解性类型（例如输入、模型、原因、假设情境和确定性） [21], ；更完整的解释包含的可理解性类型也更多 [17]。在一项较早的研究中，Kulesza 等人 [18] 探讨了智能代理应如何向终端用户解释自身，尤其关注解释的可靠性和完整性对终端用户心理模型的影响。作者发现，提高完整性有助于参与者形成心理模型，也改善了他们对理解解释所需投入的成本效益权衡的感知。另一方面，当可靠性很低时，参与者会感到更大的心理负荷，并对解释失去信任。这项研究表明，解释需要具备足够的可靠性和完整性，才能帮助

用户建立对系统工作方式的准确心理模型，同时又不至于让他们应接不暇。另一项研究中，Kizilcec[14] 考察了同伴评估情境下，系统透明度的三个层级对用户在算法界面中信任的影响，并得出结论：旨在建立信任的设计需要保持适度的界面透明度，即"既不能太少，也不能太多"。在他们的研究中，Yang 等人 [35] 也观察到，参与者对可视化解释的理解程度与其信任相关。作者表明，不同的可视化解释会带来不同程度的信任；如果解释难以理解，还可能导致不恰当的信任。总而言之，这些研究表明：（1）不同用户对解释信息的需求程度不同；（2）恰当的解释详细程度取决于用户的情境，即其目标和个人特征；（3）提供不恰当的解释详细程度可能会产生负面影响。

尽管在可解释人工智能研究中，提供不同详细程度的解释日益普遍，但在可解释推荐领域的文献中，这种做法仍然少见。只有 [24] 中介绍的研究提供了详细程度各异的解释。根据该研究的研究发现，作者建议：（1）用户应能选择是否查看解释；（2）解释组件应具有足够的灵活性，能够根据用户的偏好呈现不同详细程度的内容。遵循这些设计指南，作者开发了一款音乐推荐系统：它不仅允许用户通过点击"Why?"按钮选择是否查看解释，还允许用户通过点击"More/Hide"按钮选择详细程度。

2.2 个人特征的影响

近期关于可解释推荐的研究表明，个人特征会影响用户对解释的感知，因此在设计解释时考虑个人特征十分重要[13, 16, 24, 29]。这些研究考察了大五人格特质、认知需求和可视化熟悉度等人为因素的影响，并证实，具有特定个人特征的用户会以不同方式感知可解释推荐系统并与之互动。具体而言，既有研究考察了个人特征对不同解释风格（如基于用户、基于项目、基于内容、社交） [16] 以及不同解释形式（文本、视觉） [13, 29] 的感知所产生的影响。然而，在可解释推荐研究中，个人特征对不同详细程度解释的感知影响仍未得到充分探索。

3 RIMA

我们开发了透明的推荐与兴趣建模应用程序（RIMA），旨在通过不同详细程度的解释来说明推荐结果。RIMA 是一种基于内容的推荐系统，可生成基于内容的解释。它采用用户驱动的个性化解释方式，提供详细程度不同的解释，并让用户能够按照自己的意愿引导解释过程 [10, 11]。该应用程序提供按需解释，也就是说，用户可以决定是否查看解释，也可以选择想要查看的解释详细程度 [12]。在这项工作中，我们专注于推荐推文，并借助解释性可视化，让用户深入了解推荐过程。

3.1 兴趣模型推断

RIMA 中的兴趣模型是根据用户的出版物推断得出的。该应用使用用户提供的 Semantic Scholar ID 来收集其出版物，并使用 [7] 中介绍的语义兴趣建模工具包（SIMT），根据用户的出版物推断其兴趣模型。它对收集到的

出版物应用无监督关键短语提取算法，以生成基于关键短语的兴趣。为了解决语义问题，应用程序利用维基百科作为知识库，将关键短语映射到维基百科页面，并生成基于维基百科的兴趣。

3.2 推荐生成

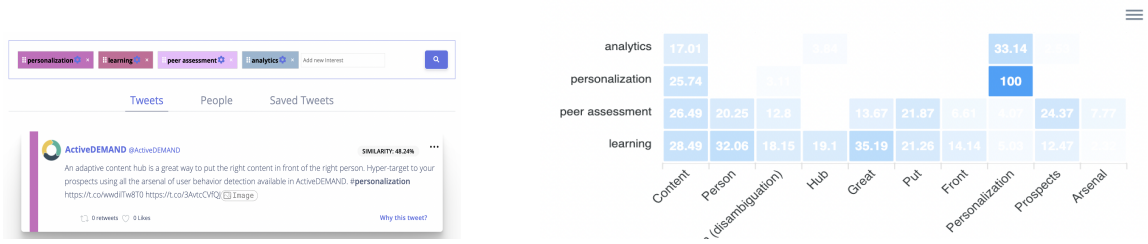
本应用部分旨在根据推断出的兴趣模型提供推文推荐。为获取候选推文，我们使用 Twitter API 获取包含一个或多个用户兴趣的推文，并将这些兴趣作为推荐的输入。随后，我们对获取的推文应用无监督关键短语提取算法，从推文文本中提取关键词。为了比较用户兴趣与候选推文之间的相似度，我们使用词嵌入技术生成兴趣模型和推文的向量表示。获得两种嵌入表示（即兴趣模型嵌入和推文嵌入）后，我们计算它们之间的余弦相似度，以获得语义相似度分数。语义相似度分数高于 40% 阈值的推文将显示给用户。

3.3 解释设计

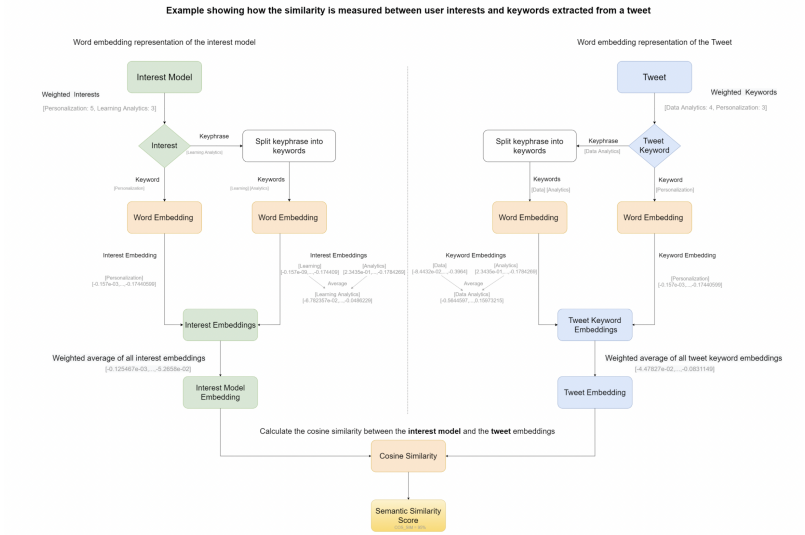
当前的解释设计主要源于作者与当地大学学生共同开展的多次头脑风暴会议。为了系统地设计不同详细程度的解释，我们以 Kulesza 等人的解释性调试原则 [17] 和 Lim 和 Dey 的可理解性类型框架[21]为基础。解释的视觉设计受到可解释推荐系统相关文献中常见可视化方式的启发，例如热力图和节点链接图 [9]。RIMA 旨在通过改变解释的可靠性和完整性，提供不同详细程度的解释。要可靠地解释推荐，就需要以详细的方式准确说明底层推荐算法的各个组成部分。要提供完整的解释，则需要告知用户推荐系统掌握的有关用户的所有信息，以及系统如何使用这些信息。Kulesza 等人 [17] 指出，完整的解释应包含 Lim 和 Dey 所说的输入（系统知晓的信息）、原因（特定决策背后的理由）和方式（系统决策过程的概述）这几种可理解性类型 [21]。RIMA 提供三种分层解释（即基础、中级、高级），用户可根据自己希望获取的信息多少进行选择。中级和高级解释默认隐藏，但用户可以按需查看。通过调整和组合可靠性与完整性的三个层级（即低、中、高），即可改变解释的详细程度。

3.3.1 基础解释。 我们设计了基础解释，用于以较低可靠性 和中等完整性解释推荐系统。它为推荐系统提供了一种简单、宽泛且低保真度的解释，不提供有关底层推荐算法的具体细节。就完整性而言，此解释包含输入 和原因可理解性类型。它以抽象的方式向用户介绍可作为推荐系统输入的兴趣模型，以及推荐系统做出特定推荐的原因。如图 1a 所示，搜索框最初填充了用户排名前五的兴趣，这些兴趣按照系统生成的权重排序。用户还可以在搜索框中添加新兴趣或删除已有兴趣。系统将使用这些兴趣作为推荐过程的输入。基础解释通过使用颜色带将推文映射到相关兴趣来实现。此外，推文文本中的兴趣也会被突出显示，以表明该推文包含这个特定词语（兴趣）。除了这两个视觉元素外，我们还在推文的右上角显示相似度分数，以表示用户兴趣与推荐推文之间的相似程度。在此层级，对“原因”可理解性类型的回答是

“因为推文文本包含您的兴趣 X，且该推文与您的兴趣档案具有 Y% 的相似度”的视觉呈现。



(a) 基础解释 (b) 中级解释



(c) 高级解释

图1. 通过三个详细程度解释推文推荐。

3.3.2 中级解释。 如需了解更多细节，用户可以点击推文右下角的“为什么推荐这条推文？”，选择中级解释 级别。中级解释是一种具有中等可靠性 和中等完整性的解释实例。它通过展示用户兴趣与从推荐推文中提取的关键词之间的计算相似度结果，为底层推荐算法提供更具体、保真度适中的解释。然而，它不会向用户准确披露推荐过程的详细步骤。在完整性方面，这种解释仅包含输入 和原因 两种可理解性类型。中级解释会告知用户推荐系统可用的信息来源（即用户兴趣和推文），并说明用户的各项兴趣与推文关键词之间的相似度对每项推荐都发挥了作用。如图 1b 所示，我们使用热力图来展示用户兴趣与从推文文本中提取的关键词之间的语义相似度。x 轴表示从推文中提取的关键词，y 轴表示用于生成推荐的用户兴趣。单元格显示计算得到的语义相似度

分数，即各项兴趣与各个关键词之间的相似度。在此级别，对“原因”这一可理解性类型的回答，是以视觉方式呈现所有作为推荐输入的用户兴趣与从推文中提取的所有关键词之间的相似度。

3.3.3 高级解释。 若要进入高级解释 层级，用户必须点击中级解释窗口右下角的“更多”按钮。高级解释进一步提升了可靠性和完整性。此解释旨在通过高保真地说明底层推荐过程，并准确详述用于计算用户兴趣与从推荐推文中提取的关键词之间相似度的所有算法步骤，从而实现高可靠性 。为帮助确保高完整性，我们添加了如何 可理解性类型，并与输入 和为何 可理解性类型一同呈现，从而提供更多信息。按照示例解释的方法，高级解释说明，推荐系统算法将用户兴趣模型和推文的词嵌入表示作为输入，计算兴趣模型与推文嵌入之间的语义相似度分数。随后，系统会推荐相似度分数较高的推文（见图1c）。

4 实证研究

4.1 参与者

本研究的目标群体是至少发表过一篇科学出版物的研究人员和学生。参与者通过电子邮件、口耳相传以及社交媒体网络群组招募，并须满足两项参与条件：至少发表过一篇科学出版物，并拥有 Semantic Scholar ID，这是进行兴趣模型推断步骤所必需的。为获得多样化的样本，研究纳入了来自不同国家、教育程度和学习背景的参与者。共有36名参与者完成了研究。我们通过检查冗余的作答模式（例如始终只选择一个作答选项）和注意力检验（即“请在此问题中选择‘不同意’”）来确保数据质量。因此，排除了五名参与者。最终样本由$N = 31$ 名参与者组成（14名男性、17名女性），平均年龄为32岁。在31名参与者中，19人（61.3%）表示居住在德国，其中12人（38.7%）是来自八个不同国家的国际用户。所有参与者的英语能力均足以参与本研究。大多数参与者报告的最高学历为硕士学位 （61.3%）。多数参与者（38.7%）具有计算机科学学习背景。17名参与者（54.8%）被视为*Twitter* 用户 ，他们表示每周至少使用 Twitter 1 小时；另有14人（45.2%）表示通常一周内从不使用 Twitter。

4.2 研究流程

尽管这项研究最初计划以实验室实验的形式开展，但由于COVID-19疫情及其相关限制，我们决定开展在线研究。每场研究均有一名研究助理提供技术支持。工程学院计算机科学与应用认知科学系伦理委员会于2021年2月10日批准了开展该用户研究的伦理申请。所有参与者均同意参与研究并签署知情同意。参与者首先在SosciSurvey[1] 中填写问卷，提供其Semantic Scholar ID，并回答有关人口统计学特征和个人特征的问题。接着，参与者观看了一段简短的演示视频，了解如何使用RIMA应用程序。随后，参与者需要（1）使用自己的Semantic Scholar ID创建账户，（2）探索系统并找到

[1]https://www.soscisurvey.de

符合其兴趣的推荐，并且（3）仔细查看系统提供的每项解释。之后，参与者需要根据七个解释目标评估六项解释，分别是透明度、可审查性、信任、有效性、说服力、效率和满意度 [33]。所有参与者均采用迭代方式评估这些解释，即针对每项解释回答同一组问题。为避免顺序偏差，参与者评价解释的顺序经过随机安排。参与者平均花费48.09分钟完成问卷（标准差 = 9.40，范围为 = 24.08–65.23）。研究结束时，参与者接受了事后说明，并获得抽取五张亚马逊礼品券的机会作为补偿。

4.3 测量

4.3.1 个人特征。 本研究测量了六种个人特征，分别为：认知需求、视觉熟悉度、个人创新性、信任倾向、领域知识和技术专长。对于每种个人特征，受试者均使用 5 点李克特量表作答，范围从 1（“非常不同意”）到 5（“非常同意”）。表 1 展示了这六种个人特征的定义及示例条目。

认知需求： 认知需求（NFC）指个体倾向于参与并享受需要付出努力的认知活动 [6]。为测量认知需求，本研究采用了 Lins de Holanda Coelho 等人 [22] 编制的 NCS-6。该量表是 Cacioppo 等人 [6] 编制的原始 18 项认知需求量表（NCS-18）的精简版，共包含 6 个条目。NCS-6 能显著节省时间，有助于减轻参与者的疲劳，并提高较长问卷的数据质量 [22]。NCS-6 的信度和效度与原始 NCS-18 相当，且具有出色的内部一致性（Cronbach's α 介于 .90 和 .94 之间）。

可视化熟悉度： 可视化熟悉度（VF）指用户在分析和绘制数据可视化方面的经验程度。为测量 VF，本研究采用了 Kouki 等人提出的量表 [16]。该量表的内部一致性极佳（Cronbach's α = .92）。

个人创新性： 个人创新性（PI）是一种人格特质，体现了个人对采用新技术或新理念的信心或乐观态度。个人创新性程度不同，其采用新技术的意愿也会有所不同 [1]。为测量 PI，本研究采用了 McKnight 等人提出的量表 [23]。该量表的内部一致性良好（Cronbach's α = .89）。

信任倾向： 信任倾向（TP）描述了个人通常信任他人的自然倾向强度 [15] ，并与信任分布相关 [23]。为测量 TP，本研究采用了 Lee 和 Turban 的量表[20] 。该量表的内部一致性极佳（Cronbach's α = .90）。

领域知识： 领域知识（DK）指用户对推荐项目（即推文）的了解或相关经验。为衡量领域知识，本研究采用了 Al-Natour 等人[3]研究中使用的量表。研究将量表中的“computer”一词替换为“Twitter”，以使其适用于本研究的情境。该量表的内部一致性极佳（Cronbach's α = .95）。

技术专长： 在本研究中，技术专长（TE）指用户对人工智能和推荐系统的了解。为衡量技术专长，本研究采用了 Kunkel 等人 [19]研究中使用的量表。

4.3.2 解释目标。七个解释目标的测量方法采用自不同既有研究的测量方式 [4, 16, 30–32, 34, 37]。前六个解释目标采用5点李克特量表进行测量，而满意度采用7点李克特量表进行测量。表2概述了所使用的问卷条目。除了对解释目标进行定量测量外，参与者还可以通过回答一组开放式问题，对每项解释和整个推荐系统提供定性反馈。

个人（个人特征） 特征	定义	示例项	来源
认知需求 (认知需求)	个体参与并享受以下活动的倾向 需要投入努力的认知活动 [6]	相比简单问题，我更喜欢复杂问题- 题目。	[22]
可视化熟悉度 (VF)	用户对分析的经验程度- 和绘制数据可视化的经验程度	我经常分析数据可视化。	[16]
个人创新性 (PI)	对采用新 技术 [1]	我喜欢探索新网站。	[23]
信任倾向（TP）	个体天生倾向于信任他人的程度 即普遍信任他人的倾向 [15]	我很容易信任一个人/事物。	[20]
领域知识（DK）	关于以下类型的知识或经验 推荐项目	我对 Twitter 很了解。	[3]
技术专长（TE）	关于人工智能和推荐 系统的知识	过去我曾了解推荐 系统的运作方式。	[19]

表1. 个人特征的测量。

指标	陈述 此解释.....	来源
透明度	帮助我理解这些推荐的依据。	[4]
可审查性	让我能够反馈系统对我的偏好理解得有多准确。	[4]
信任（能力）	让我看到系统具备理解我的需求和偏好的专业知识。	[34]
信任（善意）	让我看到系统会考虑我的兴趣。	[34]
信任（诚信）	让我知道该系统是诚实的。	[34]
有效性	帮助我判断推荐是否符合我的兴趣。	[31]
说服力	令人信服。	[16]
效率	帮助我更快地判断这些推荐与我的兴趣有多契合。	[31]
满意度	问题 你认为这个解释有多好?	[30, 32]

表2. 用于解释评估的问卷条目概览。

4.4 研究设计

RIMA应用程序以三种不同的详细程度（基础、中级、高级）解释推荐内容。按照被试内设计，参与者根据上述七个解释目标对这三种解释进行评分。我们将各项个人特征的测量得分计算为相应条目所报告数值的平均值。此外，我们将信任的评估得分计算为三种信任信念（即能力、善意和正直）各自报告数值的平均值。

5 结果

5.1 交互效应

为回答我们的研究问题，我们进行了七次重复测量协方差分析，并将七个解释目标的评价得分作为因变量（DV）、解释层级（基础、中级、高级）作为自变量（IV），将个人特征得分作为协变量。需要强调的是，我们并未使用协方差分析来评估自变量（详细程度）对因变量（对解释目标的感知）的总体效应，同时控制协变量（个人特征）；这种做法会假设自变量与协变量之间不存在交互作用（即回归斜率同质性）。我们进行分析的目的，是寻找详细程度与个人特征之间可能存在的交互作用。

为了直观呈现显著的交互效应，我们对每项个人特征分别按中位数进行划分，将参与者分为低分组和高分组。

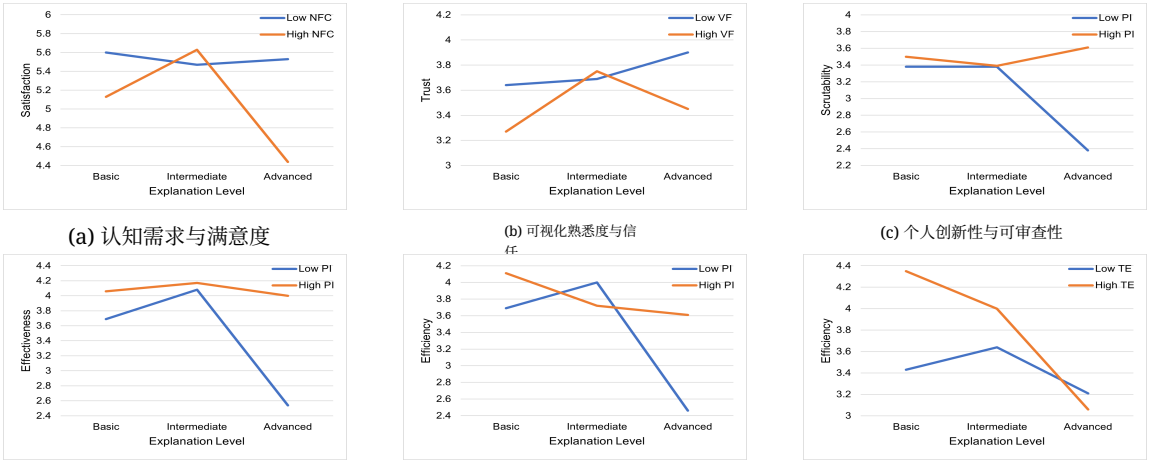
认知需求： 发现认知需求与解释层级在满意度方面存在显著交互作用（$F(2,48) = 3.654$，$p = .033$，$f = .39$）。该效应量对应中等效应 [8]。图2a显示，对于认知需求较高的用户，满意度先上升，而高级解释的平均满意度最低。认知需求与解释层级在其他解释目标方面均未发现显著交互作用。

可视化熟悉度： 发现可视化熟悉度与解释层级在信任方面存在显著交互作用（$F(2,48) = 3.639$，$p = .034$，$f = .39$）。该效应量对应中等效应 [8]。图2b显示，对于可视化熟悉度较低的用户，信任度随着解释层级的提高而上升；而可视化熟悉度较高的用户对中级解释的平均信任度最高。可视化熟悉度与解释层级在其他解释目标方面均未发现显著交互作用。

个人创新性： 在可审查性 ($F(2,48) = 3.478$, $p = .039$, $f = .38$)、有效性 ($F(2,48) = 4.231$, $p = .030$, $f = .42$)和效率 ($F(2,48) = 3.237$, $p = .048$, $f = .37$)方面，个人创新性（PI）与解释层级之间存在显著交互作用。可审查性和效率的效应为中等，有效性 的效应较强 [8]。图 2c、2d 和 2e 的交互作用图显示，个人创新性较低的用户在高级解释下，对可审查性、有效性和效率的平均感知最低。在透明度、信任、说服力或满意度方面，个人创新性与解释层级之间均不存在显著交互作用。

技术专长： 在效率($F(2,48) = 3.262$, $p = .047$, $f = .37$)方面，技术专长（TE）与解释层级之间存在显著交互作用。该效应量对应中等效应 [8]。图 2f 显示，在基础和中级解释条件下，技术专长较高的用户对效率的平均感知高于技术专长较低的用户。在其他解释目标方面，技术专长与解释层级之间均未发现显著交互作用。

最后，在这七个解释目标方面，**信任倾向**和**领域知识**与解释层级之间均不存在显著交互作用。



(a) 认知需求与满意度

(b) 可视化熟悉度与信任

(c) 个人创新性与可审查性

(d) 个人创新性与有效性

(e) 个人创新性与效率

(f) 技术专长与效率

图 2. 关于解释目标的个人特征与解释层级之间的交互效应。

5.2 定性分析

除了定量分析外，我们还对开放式问题进行了定性分析，以进一步了解对解释的感知中个体差异背后的原因。我们遵循了 Braun 和 Clarke [5] 提出的方法。为此，我们首先熟悉定性数据的深度和广度。接着，我们系统地审阅数据集，并对每个回答进行编码，以识别数据集中的模式。随后，我们将这些编码归纳为有意义的组别。此次分析主要采用演绎法，因为我们的目标是为定量分析的研究发现寻找补充解释。

5.2.1 不同详细程度的解释。 在开放式问题中，26位参与者提到的一个主要主题，是系统提供不同详细程度解释这一总体特征。具体而言，大多数参与者（26人中的23人）表示，能够选择“对于相应的推荐，多少信息才足够”（P24）很有帮助，而且这“让用户有机会按照自己的意愿，了解或探索尽可能多或少的信息”（P12）。九位参与者认同，高级解释“对普通用户来说可能有点太多了”（P17）。因此，参与者认为，最好提供不同详细程度的解释，以“满足不同人的需求”（P2），这样“每个人都能使用这个系统”（P4），而不是“一次性提供所有信息”（P5）。除了个体差异外，一些参与者还认为，所需的信息量也取决于特定情境和具体情况：用户会“根据自己的好奇心”（P7）选择所需的详细程度，也会视“当下什么有帮助”（P3）而定；或者“出了问题时，我可能会想要更详细的解释，以便充分理解错误”（P1）。只有五位参与者表示，他们不希望看到不同详细程度的解释，因为他们只需要一个“系统认为适合我”的解释（P21），而不想“一下子面对这么多内容”（P16）。

5.2.2 认知需求。 定量分析中一个令人意外的发现是，认知需求高的用户对推荐的高级解释的总体平均满意度显著低于认知需求低的用户。分析他们的回答后，我们发现，在批评推荐的高级解释的评论中，大多数（15 条中的 12 条）来自认知需求高的用户。例如，P1 表示 “有些字太小了，我没法完整阅读解释”，P2 表示 “ 我建议优化图的视图和格式设置 ”。此外，三位认知需求高的参与者不喜欢高级解释是 “ 静态的 ”（ P22 ），而且“只讲算法”（P26），因此“每条推文的解释都一样”（P9）。总体而言，认知需求高的用户似乎希望详细了解解释内容，但发现其中展示的示例值难以辨认后感到失望；而认知需求低的用户则不需要阅读每一处细节。

5.2.3 可视化熟悉度。同样，定量分析发现，与低 VF 用户相比，高 VF 用户认为推荐的高级解释可信度较低。定性分析发现，高 VF 用户不喜欢推荐高级解释的静态外观。例如，P18 表示“它看起来像是系统的标准解释”，P28 则表示“它们没有展示实际公式（即向量模型），也没有展示完整数据（特征向量的长度和值）”。此外，一位高 VF 用户认为高级解释“有点隐蔽”，并表示它“应该更加显眼”（P21）。信任度降低的一种可能解释是，对数据可视化了解更多的用户花了更多时间详细探索解释，并理解推荐输入与输出之间的相互作用。他们可能设想过，如果图表展示了其实际特征向量的值，图表会是什么样子。由于这些预期未能实现，他们可能因此认为系统没有如实说明其计算自身推荐相似度分数的方式，

进而对系统的信任度较低。不过，这一推测仍需进一步研究。

5.2.4 个人创新性。我们在定量分析中发现的另一种模式是，个人创新性较低的用户认为推荐的高级解释最缺乏可审查性、有效性和效率；他们对高级解释的评价也远低于个人创新性较高的用户。我们观察到，大多数（13名中的10名）个人创新性较低的参与者表示，他们更喜欢中级解释，因为它与自己的实际兴趣相关，这“能快速验证兴趣交集”（P5）。一位个人创新性较低的参与者认为，中级解释比高级解释更具可审查性：“中级解释确实帮助我理解了为什么会推荐一条陌生的推文”，而且“在推文视图中，高级解释的用处最小”（P23）。这似乎表明，通过示例进行解释会降低他们对可审查性、有效性和效率的评价。

5.2.5 技术专长。 我们发现，与低 TE 用户相比，高 TE 用户对基础和中级推荐解释效率的感知更高。比较他们的回答后，我们观察到，高 TE 用户确实认为基础和中级解释效率很高：“相似度分数足以快速查看推荐推文”（P20）以及“中级解释能快速、轻松地提供有用的数据”（P12）。相比之下，低 TE 用户似乎更难理解基础和中级解释：“我发现很难完全理解这种可视化”（P4）以及“一开始有点让人不知所措”（P18）。

6 讨论

我们围绕以下研究问题讨论本研究的研究发现：“个人特征如何影响用户对不同解释目标下解释详细程度的感知？”我们的结果表明，在提供不同详细程度的解释时，并不存在诸如“不能太少，也不能太多” [14]这样的通用规则。具体情况取决于多种因素。解释目标和用户类型对用户感知可解释推荐系统不同详细程度的影响各不相同。这些影响总结于表3，并将在下文讨论。

6.1 主要发现

满意度。 一项主要发现是，认知需求影响了用户对兴趣模型解释的满意度。低认知需求的用户对基础解释更满意。与 [6] 所指出的高 NFC 用户更偏好高度详细的高级解释这一研究发现不同，我们的分析显示，高 NFC 用户对推文推荐的高级解释并不满意，因为他们希望解释能够使用其实际兴趣和推文关键词来说明推荐背后的推理过程，而不是使用示例值。因此，该解释未能满足他们更强烈的理解推荐的需求。总体而言，结果表明，高 NFC 用户对根据自身数据进行个性化的详细解释更满意。

信任。 同样，我们发现，高级推荐解释降低了高 VF 用户的信任。定性分析显示，这些用户对高级解释的静态外观感到失望。此外，我们的分析表明，他们还希望解释内容更加醒目。因此，由于高级解释需要最多的交互步骤才能查看，这也可能使用户认为系统没有如实呈现其内部逻辑，进而降低了系统的可信度。总体而言，我们的研究发现证实，除了详细程度之外，整体设计或可用性等其他方面也会影响用户对解释的感知。

效率。另一项发现是，与具备推荐系统相关先验知识的用户相比，TE 较低的用户认为基础和中级推荐解释的效率较低。我们认为，已经了解推荐系统如何运作的用户，需要更少时间来理解这些解释（例如相似度分数和热力图），因此能够更快地判断推荐与其兴趣的匹配程度。相比之下，TE 较低的用户可能需要更多时间理解这些解释，因此认为它们效率较低。这表明，用户的专业知识是一个重要的影响因素，在设计解释以提高效率时应予以考虑。

可审查性。我们的分析还表明，中级解释（即热力图）提高了个人创新性（PI）较低用户对可审查性的感知，因为这种解释帮助他们验证自己的兴趣与推文推荐之间的匹配情况。因此，似乎需要根据实际数据为个人创新性（PI）较低的用户量身定制解释，让他们能够反馈推荐与其兴趣的关联程度。

Goal	**PC**	**详细程度**	**喜欢（+）/ 不喜欢（-）**
满意度	低认知需求	基础	简单（+），易于理解（+）
满意度	高 NFC	高级	设计（-），静态（-）
信任	高 VF	中级	与档案相关（+）
信任	高 VF	高级	静态（-），隐藏（-）
效率	高 TE	基础，中级	快速概览（+），易于理解（+）
效率	低技术专长	基础、中级	难以理解（-），信息过多（-）
可审查性	低个人创新性	中级	与档案相关（+）

表3. 目标、个人特征（PC）与详细程度之间的关系

6.2 设计指南

为总结本研究得出的见解，我们整理了一些建议，以有效设计推荐系统（RS）中的解释。

6.2.1 具有不同详细程度的解释。总体而言，我们的工作证实，不同用户对解释有不同的需求，并且需要不同详细程度的解释。因此，解释应针对用户类型进行定制，推荐系统也应提供具有不同详细程度 的解释，以满足不同用户的需求，这一点也与 [24] 中的建议一致。此外，我们的分析还表明，除用户特征外，特定情境和情况等其他方面也会影响用户希望看到的详细程度。例如，当用户发现系统作出了错误假设，或收到出乎意料的推荐时，他们可能会希望查看更多详细信息，以了解系统为何作出这一决策。这支持了先前的研究发现，即用户出于多种原因需要解释（例如，好奇心或系统错误 [28]）。同样，Gedikli 等人 [9]发现，当推荐存在疑问或出乎意料时，用户更需要了解该推荐。

6.2.2 目标导向、以人为中心的解释设计。由于解释风格、范围、格式或详细程度等不同设计选择会受到解释目标和用户类型的影响，因此解释设计必须采用目标导向、以人为中心 的方法：首先了解用户的目标和个人特征，再据此反向设计最能满足这些目标和个人特征的解释。为此，我们需要建立映射，以回答“哪种设计选择实例适用于哪种解释目标？”（映射：目标 → 设计选择）以及“哪种设计选择实例适用于哪种用户类型？”（映射：用户 → 设计选择）。由于设计选择与用户类型通常会对解释的感知产生交互效应，

我们需要探索形式为*(G, U, D)* 的三元组，以说明这些交互效应。其中，*G*表示解释目标（例如透明度、可审查性、信任、有效性、说服力、效率、满意度），*U*表示代表用户个人特征的一对值，例如（认知需求，高）或（可视化熟悉度，低）；*D* 表示代表某项设计选择实例的一对值，例如（详细程度，中级）或（解释形式，文本）。(G, U, D) 三元组可以解释为一条关联规则 G, U → D，例如，满意度，（认知需求，低） → （详细程度，基础），表示"为了提高满意度，应向认知需求高的用户提供详细程度较高的解释"。在我们的研究中，我们发现这类关联规则较少（见表3）。我们鼓励同行研究者开展更多用户研究，评估针对不同解释目标和用户类型设计的解释，并使用 (G, U, D) 三元组，对他们在不同应用领域的研究中可能发现的解释目标、个人特征与设计选择之间的关系进行形式化建模。

6.2.3 个性化解释。必须采用设计式个性化方法提供解释，使解释适应用户的情境，即目标和个人特征，Ain 等人也指出了这一点[2]，他们最近提出了一个用于推荐系统中个性化解释的多维概念化框架。个性化不应只体现在设计选择层面（即调整解释风格、范围、格式或详细程度），也应体现在内容层面（即根据用户数据调整解释内容）。我们可以考虑两种策略，以在推荐系统中实现个性化解释。第一种是通过提供工具来支持手动个性化，使用户能够根据自己的需求和偏好控制和调整解释（用户驱动的个性化解释）。例如，提供不同详细程度的解释，然后将选择权交给用户，让用户主动选择自己希望查看的详细程度。第二种策略是构建由可解释推荐系统控制的自动个性化（系统驱动的个性化解释）。(G, U, D) 三元组可用作关联规则，在考虑单个用户情境的基础上提供自动个性化解释。

7 局限性

尽管我们在用户研究的规划上投入了大量心力，但本研究仍存在一些局限性。首先，研究的样本量较小。因此，研究结果应谨慎解读，且不能推广到其他情境。更大的样本量可能会带来更显著、更可靠的结果。遗憾的是，由于研究要求参与者至少拥有一篇科学出版物和一个 Semantic Scholar ID，我们未能及时找到更多参与者。其次，个人特征的测量和对系统的评估完全基于自我报告。尽管数据经过了仔细收集，但这种方法仍会受到一些误差因素的影响，例如社会赞许性导致的不诚实，这可能造成偏差。第三，一些解释目标本可以采用客观而非主观的测量方式。例如，效率也可以通过日志数据来衡量，例如总交互时间。此外，一些参与者在研究过程中遇到了技术问题，这可能对他们感知 RIMA 应用程序及其解释产生了负面影响。最后，目前不同详细程度的设计主要源于作者与当地大学学生参与的头脑风暴会议。如果我们设计并呈现了其他解释（例如，个性化的高级解释，而不只是通过示例进行解释），本研究的结果可能会有所不同。

8 结论与未来工作

本文旨在阐明可解释推荐相关文献中一个研究不足的方面，即个人特征和详细程度对用户感知推荐系统（RS）中解释的影响。为此，我们开发并评估了一个透明的推荐与兴趣建模应用程序（RIMA），它以三种不同的详细程度（基础、中级、高级）解释推荐结果。研究结果表明，解释设计首先应根据用户的情境（即目标和个人特征）进行调整。根据研究发现，我们提出了一些设计指南，供设计与用户情境相契合的解释型界面时参考。本研究对推荐系统中个性化解释界面的理论研究具有启示意义。研究展示了个人特征与详细程度对用户感知可解释推荐系统的交互效应，并为解释目标、用户类型与设计选择之间存在依赖关系提供了证据。此外，本研究还通过提出有关如何恰当设计推荐系统中的个性化解释界面的建议，为实践作出了贡献。未来，我们将评估这些结果在不同应用领域中的普适性。我们还将探索其他可能的可视化方式，以提供三种详细程度的解释。特别是，我们将开发并评估根据用户数据量身定制的高级解释。此外，我们还将研究个人特征与其他设计选择（如解释风格、范围和格式）的交互效应对用户感知可解释推荐系统的影响。

参考文献

[1] Ritu Agarwal和Jayesh Prasad。1998年。信息技术领域个人创新性的概念与操作性定义。*Information systems research* 9，2（1998），204–215。[2] Qurat Ul Ain、Mohamed Amine Chatti、Mouadh Guesmi和Shoeb Joarder。2022年。推荐系统个性化解释的多维概念化框架。载于第*27*届智能用户界面国际会议伴随论文集。[3] Sameh Al-Natour、Izak Benbasat和Ronald T Cenfetelli。2008年。过程相似性与结果相似性对用户评价决策辅助工具的影响。*DecisionSciences* 39，2（2008），175–211。[4] Krisztian Balog和Filip Radlinski。2020年。衡量推荐解释质量：解释的相互冲突目标。载于第*43*届国际*ACM SIGIR*信息检索研究与发展会议论文集。329–338。[5] Virginia Braun和Victoria Clarke。2006年。在心理学中运用主题分析。心理学定性研究3，2（2006），77–101。[6] John T Cacioppo、Richard E Petty和Chuan Feng Kao。1984年。认知需求的高效评估。人格评估杂志48，3（1984），306–307。[7] Mohamed Amine Chatti、Fangzheng Ji、Mouadh Guesmi、Arham Muslim、Ravi Kumar Singh和Shoeb Ahmed Joarder。2021年。SIMT：语义兴趣建模工具包。载于第*29*届*ACM*用户建模、自适应与个性化会议附属论文集。75–78。[8] J. Cohen。1988年。行为科学统计功效分析。Lawrence Erlbaum Associates，美国新泽西州希尔斯代尔。[9] Fatih Gedikli、Dietmar Jannach和Mouzhi Ge。2014年。我该如何解释？推荐系统不同解释类型的比较。*International JournalofHuman-Computer Studies* 72，4（2014），367–382。[10] Mouadh Guesmi、Mohamed Amine Chatti、Yiqi Sun、Shadi Zumor、Fangzheng Ji、Arham Muslim、Laura Vorgerd和Shoeb Joarder。2021年。面向透明推荐的开放、可审查且可解释的兴趣模型。载于第*26*届智能用户界面国际会议伴随论文集。[11] Mouadh Guesmi、Mohamed Amine Chatti、Laura Vorgerd、Shoeb Joarder、Qurat Ul Ain、Thao Ngo、Shadi Zumor、Yiqi Sun、Fangzheng Ji和Arham Muslim。2021年。输入还是输出：解释焦点对不同详细程度下可解释推荐感知的影响。载于推荐系统界面与人类决策联合研讨会第*8*届论文集，第2948卷。55–72。[12] Mouadh Guesmi、Mohamed Amine Chatti、Laura Vorgerd、Shoeb Joarder、Shadi Zumor、Yiqi Sun、FangzhengJi和Arham Muslim。2021年。按需个性化解释助力透明推荐。载于第*29*届*ACM*用户建模、自适应与个性化会议附属论文集。246–252。[13] Diana C Hernandez-Bocanegra和Jürgen Ziegler。2020年。解释基于评论的推荐：档案透明度、呈现风格和用户特征的影响。*i-com* 19，3（2020），181–200。[14] René F Kizilcec。2016年。信息量多少才合适？透明度对算法界面信任的影响。载于*2016*年*CHI*计算系统中的人为因素会议论文集。2390–2395。[15] Sherrie Yi Xiao Komiak。2003年。内化和熟悉度对信任及推荐代理采纳的影响。博士论文。英属哥伦比亚大学。

[16] Pigi Kouki、James Schaffer、Jay Pujara、John O'Donovan和Lise Getoor。2019年。混合推荐系统的个性化解释。载于第24届智能用户界面国际会议论文集。379–390。[17] Todd Kulesza、Margaret Burnett、Weng-Keen Wong和Simone Stumpf。2015年。解释性调试原则：个性化交互式机器学习。载于第20届智能用户界面国际会议论文集。126–137。[18] Todd Kulesza、Simone Stumpf、Margaret Burnett、Sherry Yang、Irwin Kwan和Weng-Keen Wong。2013年。太多、太少，还是恰到好处？解释影响终端用户心理模型的方式。载于2013年IEEE可视化语言与以人为中心的计算研讨会。IEEE，3–10。[19] Johannes Kunkel、Thao Ngo、Jürgen Ziegler和Nicole Krämer。2021年。识别推荐系统的群体特定心智模型：一种新颖的定量方法（已投稿）。（2021年）。[20] Matthew KO Lee和Efraim Turban。2001年。消费者互联网购物信任模型。《国际电子商务杂志》6，1（2001），75–91。[21] Brian Y Lim和Anind K Dey。2009年。评估情境感知应用对可理解性的需求。载于第11届普适计算国际会议论文集。195–204。[22] Gabriel Lins de Holanda Coelho、Paul HP Hanel和Lukas J. Wolf。2020年。极为高效的认知需求评估：开发六题版本。《评估》27，8（2020），1870–1885。[23] D Harrison McKnight、Vivek Choudhury和Charles Kacmar。2002年。开发并验证电子商务信任测量指标：一种整合型分类体系。Information systems research 13，3（2002），334–359。[24] Martijn Millecamp、Nyi Nyi Htun、Cristina Conati和Katrien Verbert。2019年。要不要解释：解释音乐推荐时个人特征的影响。载于第24届智能用户界面国际会议论文集。397–407。[25] Tim Miller。2019年。人工智能中的解释：来自社会科学的见解。Artificial Intelligence 267（2019），1–38。[26] Sina Mohseni、Niloofar Zarei和Eric D Ragan。2018年。可解释人工智能系统的设计与评估：多学科综述与框架。arXiv（2018），arXiv–1811。[27] Ingrid Nunes和Dietmar Jannach。2017年。决策支持系统与推荐系统中的解释：系统综述与分类体系。《用户建模与用户适应性交互》27，3–5（2017），393–444。[28] Vanessa Putnam和Cristina Conati。2019年。探索智能辅导系统中对可解释人工智能（XAI）的需求。。载于IUI研讨会，第19卷。[29] Maxwell Szymanski、Martijn Millecamp和Katrien Verbert。2021年。视觉、文本还是混合：用户专业知识对不同解释方式的影响。载于第26届智能用户界面国际会议。109–119。[30] Nava Tintarev和Judith Masthoff。2008年。个性化电影解释的有效性：一项使用商业元数据的实验。载于自适应超媒体与自适应网络系统国际会议。施普林格，204–213。[31] Nava Tintarev和Judith Masthoff。2011年。推荐系统解释的设计与评估。载于《推荐系统手册》。施普林格，479–510。[32] Nava Tintarev和Judith Masthoff。2012年。评估推荐系统解释的有效性。《用户建模与用户适应性交互》22，4–5（2012），399–439。[33] Nava Tintarev和Judith Masthoff。2015年。解释推荐：设计与评估。载于《推荐系统手册》。施普林格，353–382。[34] Weiquan Wang和Izak Benbasat。2007年。电子商务推荐代理：解释功能对信任信念的影响。《管理信息系统杂志》23，4（2007），217–246。[35] Fumeng Yang、Zhuanyi Huang、Jean Scholtz和Dustin L Arendt。2020年。可视化解释如何促进终端用户对机器学习建立恰当的信任？载于第25届智能用户界面国际会议论文集。189–201。[36] Yongfeng Zhang和Xu Chen。2018年。可解释推荐：综述与新视角。arXiv预印本，arXiv:1804.11192（2018）。[37] Ruijing Zhao、Izak Benbasat和Hasan Cavusoglu。2019年。用户总是想了解更多吗？探究系统透明度与用户对建议提供系统的信任之间的关系。载于第27届欧洲信息系统会议论文集。