Explainability in music recommender systems

如今，收听录制音乐的最常见方式是通过流媒体平台，该平台提供对数千万首曲目的访问。为了帮助用户有效地浏览这些大型目录，音乐推荐系统（MRS）的集成变得至关重要。当前现实世界的MRS通常非常复杂，并且针对推荐准确性进行了优化。它们结合了基于协作过滤和基于内容的推荐的多个构建块。这种复杂性可能会妨碍向最终用户解释建议的能力，这对于被视为意外或不适当的建议尤其重要。虽然纯粹的推荐性能通常与用户满意度相关，但可解释性对信任和宽恕等其他因素有积极影响，这些因素对于保持用户忠诚度至关重要。

在本文中，我们将讨论如何在MRS的上下文中解决可解释性问题。我们提供了关于可解释性如何改进音乐推荐算法和增强用户体验的观点。首先，我们回顾了推荐者可解释性以及可解释人工智能（XAI）的一般维度和目标，并详细说明了这些维度和目标在多大程度上适用于或需要适应音乐消费和推荐的特定特征。然后，我们展示了如何将可解释性组件集成到MRS中，以及可以提供哪种形式的解释。由于解释质量的评估与纯粹基于准确性的评估标准脱钩，因此我们还讨论了评估音乐推荐解释的要求和策略。最后，我们描述了在大规模工业MRS中引入可解释性的当前挑战，并提供了研究观点。

推荐系统（RS）技术渗透到我们的日常生活中。与其他领域的推荐系统（RS）（Ricci，Rokach和Shapira 2015）一样，音乐推荐系统（MRS）（Schedl等人，2015）的核心是信息过滤算法，它从通常庞大的音乐项目目录（例如，艺术家，专辑或歌曲）中选择那些与目标用户最相关的音乐项目。因此，音乐推荐系统（MRSs）引导用户在当今触手可及的音乐数量庞大。1

对机器学习 （ML） 算法（包括 RS 中使用的算法）透明度的认识不断提高和持续讨论，导致用户大量要求获得解释，说明为什么向他们推荐某些项目（Zhang and Chen 2020）。同样从 RS 提供商的角度来看，这些方面对于建立和维护用户对系统的信任非常重要。因此，使MRS具备向用户提供解释的能力是双方都感兴趣的。

虽然音乐推荐与其他媒体推荐任务（如视频或电影）共享某些属性，但也存在明显的差异。在文献中确定的特征（例如，Schedl et al. 2018）中，以下特征与MRS的可解释性相关，我们将在后续部分中详细说明：

物品消费的持续时间通常比其他领域短得多，也就是说，歌曲的典型长度为几分钟，而电影甚至书籍需要更长的时间来消费。

音乐数据有多种表示形式，包括音频、MIDI 和文本元数据（例如，编辑元数据，还有用户生成的标签）。此外，可以在MRS中使用的音乐相关数据是高度多模态的，包括图像（例如，专辑封面）和视频（例如，音乐视频剪辑），旁边是音频和文本元数据。最后，从各种活动（例如，喜欢、收藏、跳过歌曲）中收集用户反馈。

聆听环境强烈影响音乐偏好。例如，听众的情绪、位置和社交状况（例如，独自一人与与朋友）已被证明会影响音乐需求和要求（Ferwerda， Schedl， and Tkalcic 2015;Rentfrow， Goldberg， and Zilca 2011）。

音乐通常按顺序使用，即收听会话或播放列表中的曲目。因此，对于音乐，我们经常专注于顺序推荐任务，例如自动创建或延续播放列表（Bonnin and Jannach 2014;Zamani 等人，2019 年），利用了长期和短期用户偏好。

MRS 存在各种用例，以不同的任务为中心。其中，最重要的是首页推荐（推荐在平台用户界面首页上向用户展示的主题音乐收藏的内容（也称为书架或频道）（Bendada 等人，2020 年）、音乐探索/发现（例如，基于旋律、节奏或歌词方面的项目相似性）（Goto 和 Dannenberg 2019; Knees， Schedl， and Goto 2020）， 自动播放列表生成（通常基于用户配置文件，但可能仅基于种子描述，例如“放松音乐”）和自动播放列表延续（基于种子曲目序列）（Jannach， Lerche， and Kamehkhosh 2015;扎马尼等人，2019 年）。

要创建音乐推荐引擎，根据用例，采用了多种方法。这些包括潜在因子模型（例如，奇异值分解或分解机），图挖掘技术（例如，随机游走或图嵌入）和基于深度学习的技术（例如，卷积神经网络，循环神经网络或自动编码器）。此外，音频信号处理和自然语言处理技术通常用于创建音乐项目的矢量表示或使用相关标签注释音乐项目。

这些引入的特征、任务和方法如何为MRS的可解释性策略提供信息，将在本文的后续部分中详细介绍。工业背景下的例子和挑战将在最后一部分提供。

近年来，人们对可解释的建议的兴趣激增，尽管这个概念出现在2000年代（Herlocker，Konstan和Riedl 2000）。可解释 RS 的这种演变伴随着可解释人工智能 （XAI） 的日益普及，它与它共享根源、方法和术语。可解释人工智能（XAI）代表了许多研究学科的融合，包括计算机科学，人机交互，哲学和心理学。连贯和稳定的XAI定义和术语直到最近才开始出现（Arrieta等人，2020年;吉多蒂等人，2018年;立顿2018）。与此同时，RS的研究已经开发出了一般XAI所不知道的与解释相关的概念，然而，其中一些概念依赖于难以捉摸的描述。这种特殊性可能是由于 RS 本身的性质，它们的任务、输入和结果与 XAI 的一般趋势不同。将这两个可解释性领域联系起来，不仅会导致RS中更标准化的解释方法，而且还可以直接应用从XAI到MRS的方法。

在本节中，我们将回顾 RS 中可解释性的定义和概念。随后，我们将它们与 XAI 的进行比较和连接。请注意，这不是对 XAI 或可解释 RS 的调查，因为在这个问题上存在其他有价值的资源（Arrieta 等人，2020 年;吉多蒂等人，2018年;努内斯和詹纳赫 2017;张和陈2020）。

解释建议是什么意思？在RS领域，Tintarev和Masthoff（2015）用“通过给出详细的描述来明确”来解决这个问题，而Zhang和Chen（2020）用“一个可解释的建议旨在回答为什么的问题”。因此，我们可以看出解释作为建议的补充信息的作用。但这些定义是有限的;例如，确保公平的建议涉及追踪建议的“原因”，但仅限于某些关键方面（例如，潜在的性别偏见），并且没有说明如何根据这些方面采取行动。随着我们接下来的发展，“补充信息”和“公平建议”塑造了可解释性的许多方面中的两个。

借用最近XAI术语协调工作的一般思想，区分解释对象和目标更方便。解释指定解释系统的结果，它们形成“解释系统和目标受众之间的接口”（Guidotti 等人，2018 年）。与可解释性互换，我们将使用术语可解释性，具有更被动的特征：系统可以解释 - 例如，决策树通常是可解释的，神经网络不是。相反的概念是黑匣子。我们强调，自动得出结论认为树和线性回归是可解释的，而神经网络不是，这是值得怀疑的。2正如我们接下来将看到的，这取决于解释任务的精确表述，这些任务不允许一刀切的规则。

前面提到的“受众”是必不可少的，因为给定的解释类型可能只向特定的人传达有意义的信息。在RS研究中，解释的目标受众通常是最终用户，因为他们是推荐决策的目标，并且可能对此持怀疑态度。然而，其他利益相关者可能有兴趣接受解释，例如，系统设计人员和数据科学家可能会询问他们的系统根据数据的歧视性偏见做出决策。

我们将继续一个警示故事：可解释性的不同概念导致了对XAI的许多误用（Lipton 2018）。因为我们通常无法在野外获得地面真相的解释，实际上在工业环境中也不会，所以许多XAI作品都依赖于直觉的概念，即他们的目标解释应该是什么。这首先使评估变得困难。正如Doshi-Velez和Kim（2017）所强调的那样，解释的相关性通常以“当你看到它时你会知道它”的方式提出，这为许多确认偏见铺平了道路。其次，几个违反直觉的结果已经揭晓。例如，广泛同意的可解释模型比黑盒模型更可取的观点受到了挑战：产生的解释——类似于模型预测——可能具有误导性或偏见（Adebayo 等人，2018 年;迪努、比厄姆和科尔特 2020;考尔等人，2020;鲁丁2019）。而且，如果没有明确的解释任务，怎么能比较几个XAI系统呢？我们真的能量化可解释性和解释质量吗？是否可以评估拟议解释指标的相关性？我们如何发现基于虚假机制的误解和解释？所有这些问题都回到了解释的定义。

在处理这些问题时，提出了不完整的概念。其目的是描述“缺失的部分”，证明使用解释系统的合理性（Doshi-Velez和Kim 2017）。在这里，可解释的RS和一般XAI的文献存在分歧。在Tintarev和Masthoff（2015）中提出了RS解释目标的区别。我们可以通过Arrieta等人（2020）在一般XAI中确定的目标来丰富这一讨论。图 1 显示了这两组目标，并附有简短的定义。我们发现，这两者中的任何一个都不能单独解释所有MRS目的：可解释的RS目标大多属于信息性类别。这比RS的透明度范围更广，后者过于关注模型内部机制的分解。此外，已经发现RS目标是相互关联的（Balog和Radlinski 2020），特别是满意度可以说是任何解释方法的理想副产品。也就是说，说服力是 RS 的一个重要维度，在一般 XAI 中不存在（Ehrlich 等人，2011 年）;当以透明度为目标时，创建一个有说服力的系统可能看起来是矛盾的。

确定这些目标至关重要，因为如果不完全性不是问题，可能根本不需要解释。然后应对每个目标的不完整进行评估，以避免目标不匹配。最后，可理解性的概念只是弥合了所选的XAI系统与针对目标受众的目标/不完整的新概念之间的差距。所有这些概念都进行了说明，并相应地放置在图 2 中。我们将在后续段落中讨论 XAI 的其他分类轴。

最后，可解释性可以通过迈克尔·乔丹（Michael Jordan）对ML未来的有趣看法来构建。3XAI 的目标不仅是让决策者理解模型预测，而且允许两者之间来回交互。你为什么做出这个决定？如果这方面不同呢？那么如果这方面不同呢？“相应的决策涉及考虑从未放入原始数据的新事实，这些事实与当前情况相关。关于不完备性和可理解性的讨论的目的是超越将可解释性仅仅作为一种补充预测的观点，而是允许几个参与者之间相互获得知识。

由于MRS通常提供许多建议，因此在解释范围内有一个重点区别：本地全球（Doshi-Velez和Kim 2017）。本地或实例解释针对特定输入-推荐对的模型决策，例如，解释向最终用户推荐曲目是因为其某些功能匹配。必须根据每个预测量身定制局部解释。这种类型的解释符合欧洲通用数据保护条例（GDPR）的“解释权”（Parliament 2016），该条例使用户有权查询算法结果背后的推理，从而支持信息性作为可解释性目标。

相比之下，全局解释提供了模型逻辑的大局，涵盖了多个模型决策。例如，估计模型学习的用户嵌入集群可能有助于合理化几个一般社区中MRS的行为。模型的这种广泛视图对于检测模型的系统偏差（解决公平性目标）和检查模型适合部署的位置（解决信息性、可信度和置信度）是必要的。最后，请注意，这两种类型可能是有联系的：有时通过提供多个局部解释来制定全局解释是相关的（Ribeiro， Singh， and Guestrin 2016）。

我们还可以区分解释系统w.r.t.可解释性是否应该是RS内在可解释性的固有部分，还是应该作为已经工作的RS事后可解释性的补充提供。

内在可解释性是指 RS 提供足够信息以使其内部功能向特定受众清晰的能力（Arrieta 等人，2020 年）。在这种情况下，解释与模型一致。作为模型中固有的，内在可解释性必须提前规划，使其成为模型设计的组成部分。例如，Item-k-Nearest Neighbors模型推荐艺术家，因为他们与用户收听的艺术家相似，因此允许诸如“我们推荐您<艺术家>，因为它类似于<艺术家>”之类的解释。

事后或外在可解释性是指使用外部 XAI 从黑盒模型中产生知识。它可以被认为是对模型的逆向工程（Guidotti等人，2018）。例如，黑盒模型的建议可以通过事后选择导致建议的相关特征来解释;他们提供诸如“我们推荐您这样做，因为它具有您可能喜欢的<功能>之类的解释。内在观点和事后观点都与透明度的概念有关，从而支持信息性、因果关系和信心。然而，事后解释受到其外部性的阻碍，需要额外检查它们对所研究模型的忠实度。然而，与内在设计相比，它们将模型设计与解释设计分开，允许在后期考虑XAI系统，或将它们应用于已经工作的模型。

我们经常认为XAI方法是无监督的。特别是在最终用户方面，很难猜测哪些可能是用户的基本事实解释，因为他们对什么是好的解释的判断可能是有偏见的（Miller 2019）。尽管如此，有时还是可以获得目标解释（Balog and Radlinski 2020）。我们的目标不仅是做出解释预测，而且是解决不完整的问题;因此，目标预测的相关性必须受到质疑。这些真的能解决我们对不完整的需求吗？还是他们的代理？在后一种情况下，我们如何断言/评估它们在我们的目标下可理解性？我们提出了XAI的两个想法，用于图像域中的监督解释，可以应用于MRS。

在图像分类领域，一些数据集收集带有文本描述的图像。每组单词都可以与图像中相应的视觉方面进行匹配，从而能够通过RNN生成的文本为看不见的实例类预测生成视觉解释（Hendricks等人，2016）。根据保持分开的测试说明评估解释。在这里，解释的概念由两个前提驱动，首先是作为链接同一对象（图像和文本）的不同形式的方式，其次是作为通过产生特定于类的信息来传达有用信息的基本原理，这些信息将其与其他类区分开来。在无监督的环境中，获得这种信息性的区分质量是很棘手的。音乐数据的多模态（例如，音频、歌词、用户、播放列表）使其成为这种范式的良好候选者。

我们可以将另一行监督解释确定为链接不同的概念层次。例如，TCAV方法（Kim等人，2018年）允许根据人类可理解的概念检查预测，例如，斑马图像的模型预测对“条纹”的敏感程度。同样，与音乐有一个有趣的联系：在低级数据（即音频信号）及其与高级描述（例如，流派，情绪）的对应关系之间存在已知且未解决的语义差距（Celma，Herrera，and Serra 2006）。

我们以一个容易被忽视的重要而微妙的区别来结束本节：解释是否与 RS 模型处理或它所表示的数据有关？

一方面，模型解释侧重于模型学习的表示和参数，并旨在从中理解。稍微夸张一点，对于“为什么MRS根据我的历史推荐这首歌？”这个问题，RS的以模型为中心的答案可能是“考虑到所有其他用户的收听历史，它最大限度地提高了与您的历史记录共同收听的可能性。另一方面，数据解释更愿意关注“为什么这些项目首先被共同倾听？训练好的模型本身不如揭示“世界上的自然机制”的目标有趣（Chen et al. 2020）。在实践中，在第一种情况下，模型检查可能会暴露出不规则性，并导致调整其架构和正则化（例如，平衡公平权衡参数）;在第二种情况下，模型扮演数据代理表示的角色，检测到的错误将更合适地归因于输入数据的错误表示（例如，特征工程以获得更好的矩阵分解），最终目的是找到一个可靠的结构，给定问题的先验知识。

这些方面经常纠缠在一起。解释模型提供的嘈杂数据信息很少，如果模型假设没有抓住突出的方面（例如，相关性而不是因果关系），解释数据可能会产生误导。当目标是解释数据时，解释模型（通常更容易，特别是在使用可解释模型时）是一个普遍的谬误。

提高MRS的解释水平，同时为用户保持强大的简单性约束仍然是一个挑战。不过，最终用户并不是唯一受MRS影响的利益相关者。例如，音乐制作人的收入也受到影响。因此，就公平性和透明度而言，全球解释可能与这样的观众相关（解释不是关于单一推荐，而是关于解释为什么向特定人群推荐艺术家）。由于这种利益相关者没有简单性约束，因此解释可能会更加详细。

另一方面，保持系统的可解释性对于不断提高其性能非常重要。例如，只有当 RS 工程师能够理解这些错误建议的原因时，才能利用接收用户对不良建议的投诉或反馈来改进系统。依赖黑盒块的 RS 无法理解不良建议，从而阻碍了系统的改进。

最后，高级用户可能想要更多的控制，简单性约束对他们来说可能不那么重要：例如，Jin，Cardoso和Verbert（2017）认为，与少即是多的设计模式相反，让用户对RS的额外控制确实增加了认知负荷，但也增加了对它如何工作的深刻理解的用户的满意度。对 RS 的控制使正反馈循环成为可能：用户可以明确地利用解释来对 RS 进行操作并减少未来的虚假建议。

有趣的是，越来越多地使用语音控制设备来试点音乐流媒体服务，为部署可解释的MRS及其他服务创造了一个有前途的新游乐场，以创建完全交互式的体验，其中建议可以受到挑战，并最终得到改进。

这项工作得到了奥地利科学基金（FWF）的资助：P33526和DFH-23;以及上奥地利州和联邦教育、科学和研究部通过赠款 LIT-2020-9-SEE-113。