攻击预训练推荐系统

Yiqing Wu*†中国科学院计算技术研究所，中国北京
wuyiqing20s@ict.ac.cn

Ruobing Xie†腾讯微信，北京，中国
ruobingxie@tencent.com

Zhao Zhang‡Yongchun Zhu*中国科学院计算技术研究所，中国北京
{zhangzhao2021,zhuyongchun18s}@ict.ac.cn

Fuzhen Zhuang§ 北京航空航天大学人工智能研究院，中国北京
zhuangfuzhen@buaa.edu.cn

Jie Zhou 腾讯微信，北京，中国
withtomzhou@tencent.com

Yongjun Xu Qing He*中国科学院计算技术研究所 北京，中国 {xyj,heqing}@ict.ac.cn

arXiv:2305.03995v1 [cs.IR] 6 May 2023

摘要

最近，一系列开创性研究表明，预训练模型在序列推荐中具有强大潜力，为构建面向不同下游推荐任务的全知统一预训练推荐模型指明了方向。尽管取得了这些进展，经典推荐系统中的漏洞也以一种新的形式存在于预训练推荐中，而预训练推荐模型的安全性仍未得到探索，这可能会威胁其广泛的实际应用。在本研究中，我们提出了一种用于预训练推荐后门攻击的新型框架。我们证明，预训练模型的提供者可以轻易地在预训练过程中植入后门，从而提高目标物品对目标用户组的曝光率。具体而言，在各种推荐预训练使用场景下，我们设计了两种新颖且有效的后门攻击：基础替换和提示词增强。真实世界数据集上的实验结果表明，与干净模型相比，我们提出的攻击策略将目标物品对目标用户的曝光率提高了数百倍。源代码已发布于 https://github.com/wyqing20/APRec。

CCS 概念

• **计算机系统组织** → **嵌入式系统**。

关键词

数据集、神经网络、视线检测、文本标注

ACM 参考格式：

Yiqing Wu、Ruobing Xie、Zhao Zhang、Yongchun Zhu、FuzhenZhuang、Jie Zhou、Yongjun Xu 和 Qing He。2018年。《攻击预训练推荐》。载于请从版权确认邮件中填写正确的会议标题（会议简称*"XX"*）。ACM，纽约州纽约市，5 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

推荐系统旨在根据用户的偏好，为其提供合适的物品，而这些偏好主要隐含在其历史行为序列中。近年来，序列推荐（SR）取得了巨大成功，并已广泛应用于实践[4, 28]，；它以用户的历史行为序列为输入，并输出预测的物品。许多有效的序列建模方法已在 SR 中得到验证 [7, 16]。近年来，受大规模预训练模型强大能力的启发 [1]，，一些先驱工作将预训练引入序列推荐，并取得了巨大成功 [17, 19, 21, 25]。一些雄心勃勃的研究甚至探索为各种下游任务构建统一的大型预训练推荐模型 [2, 5, 20]。推荐预训练的广泛应用前景可期。

鉴于推荐系统具有重大的社会影响和商业价值，其安全性至关重要。例如，在电子商务中，攻击者可以通过虚假用户和欺诈行为误导模型，使其推荐目标物品。已有研究揭示了传统推荐 [8, 9, 12, 23]中的漏洞和安全威胁。在这些研究中，攻击通常涉及两个角色：平台和用户。恶意用户可以通过自然交互污染日志数据（即用户行为历史），从而攻击平台并操纵系统，使其输出攻击者期望的结果 [18, 24, 27]。遗憾的是，尽管预训练推荐取得了进展，预训练推荐模型的安全性仍未得到考量，而推荐模型在预训练环境下面临新的威胁。与传统推荐不同，在新的预训练推荐范式中，除了平台 和用户之外，还存在另一个角色

*中国科学院（CAS）智能信息处理重点实验室。并任职于中国科学院大学，中国。

†Yiqing Wu 和 Ruobing Xie 对本研究贡献相同。
‡通讯作者。§Fuzhen Zhuang 也任职于中国北京的中关村实验室

允许免费制作本作品全部或部分内容的数字或纸质副本，供个人或课堂使用，但不得以营利或商业利益为目的制作或分发副本，且副本必须附有本声明及首页上的完整引用信息。对于本作品中归他人所有的组成部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或重新分发至列表，须事先获得特别许可和/或支付费用。请通过 permissions@acm.org 申请许可。会议简称*"XX"*，*2018*年*6*月*3*日至*5*日，纽约州伍德斯托克© 2018年计算机协会。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX

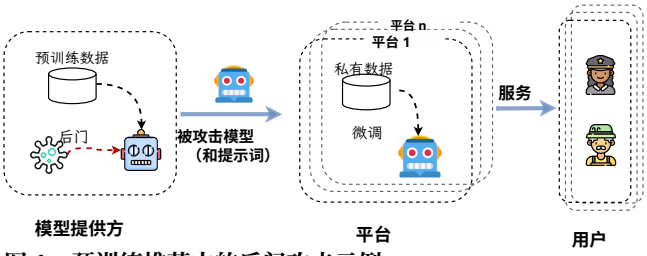

图 1：预训练推荐中的后门攻击示例。

：预训练模型提供方。预训练模型可能遭受某些攻击者主动（例如植入后门）或被动（例如虚假用户）投毒，从而损害下游平台。在这种情况下，针对公开预训练推荐模型的攻击更加隐蔽，难以被平台明确检测和控制，因为预训练模型经过黑盒训练后便固定不变。

在本研究中，我们首次对预训练推荐中的新型攻击范式进行了探索。我们设计了两种后门攻击，以提高特定目标物品对目标用户（即全体用户或某一用户群体）的曝光率。具体而言，我们首先设计了一种简单且有效的基础替换策略，用于生成包含额外目标物品的虚假用户行为序列。通过这种基础攻击，目标物品的曝光率可提高数百倍。此外，为了增强用户群体攻击的效果，我们提出了一种提示增强攻击策略，针对主流微调范式“提示调优”。具体而言，我们设计了一种三步训练框架，用于模拟提示调优的过程，从而能够在不同设置下，对目标用户群体实施更加精准且更难以检测的攻击，同时对其他用户群体的影响较小。此外，我们还提出了一种初步的预训练模型异常检测方法，以应对这类新型预训练后门攻击。在实验中，我们针对不同数据设置开展了多种攻击策略的实验，证明了推荐预训练中攻击的存在及其危害，同时验证了潜在检测方法和防御方法的有效性。

总体而言，我们首次通过所提出的后门攻击方法，系统地展示了针对预训练推荐的后门攻击威胁。我们探索了预训练推荐中的这一新型攻击范式，以提醒开发者和研究人员注意那些一直被忽视但极其重要的安全问题，敲响警钟。

2 预备知识与相关工作

推荐预训练。推荐预训练主要包含三个部分：模型提供方、平台和用户。令 S和 P分别表示模型提供方和平台。我们将模型提供方和平台中未对齐的用户集合分别记为 U^S和 U^P。在本文中，我们假设物品集合是共享的（在电影、图书和音乐推荐中广泛存在类似场景。未来的大规模预训练推荐模型也可能涉及几乎所有物品[5]），记作 V。每个用户 $u \in (U^P \cup U^S)$都有一段历史行为序列 $s_u = \{v_1^u, v_2^u, \ldots\ldots v_{|s_u|}^u\}$（长度为 $|s_u|$），并按时间排序。此外，每个用户还有 m个属性 $A_u = \{a_1^u, a_2^u, \ldots\ldots a_m^u\}$（即用户画像，例如年龄和性别）。模型提供方在 U^S的全部预训练数据上预训练一个大型序列模型 $f_{seq}(s_{u^s}|\Theta)$，然后由下游平台采用，为其用户提供服务。通常，平台可以通过三种方式使用预训练模型：直接使用、微调和提示调优。直接使用时，平台部署预训练模型来执行下游任务，不作任何更改。微调时，平台使用其私有数据对预训练模型进行微调。提示调优时，平台采用提示 [10, 20, 26] 3 2 进行参数高效的微调范式（提示调优的详情见第 . 节）。提示调优的详情见第3.2节。

威胁模型。攻击者的目标可分为提升/降级攻击和可用性攻击[15]。可用性攻击旨在使推荐系统无法提供服务，这类攻击容易被检测到，且对我们的预训练攻击场景毫无意义。提升/降级攻击旨在提高/降低目标物品的推荐频率 [18, 24, 27]。在本研究中，我们聚焦于提升攻击。攻击者的目标是在 Top-K 推荐系统中，提高目标物品对目标用户的曝光率。此处的目标用户可以是平台上的所有用户，也可以是特定用户群体。**私有数据**：由于隐私保护的要求，用户信息和行为不能公开，因此模型提供方和平台各自的行为数据均为私有数据，双方无法观察到对方的数据。

3 方法

3.1 基本替换攻击

我们首先介绍本研究的预训练部分。我们采用下一物品预测作为预训练任务，这是一种自然且广泛应用于推荐中的预训练任务 [2, 5, 22, 29]。给定用户行为序列 $s_u = \{v_1^u, v_2^u, \ldots\ldots, v_{|s_u|}^u\}$，下一物品预测的目标是预测用户可能会互动的物品。我们将其表述为：

$$u_s^t = f_{seq}(\{v_1^u, ..., v_t^u\}|\Theta),$$
$$L = \min \sum_{u \in U^S} \sum_{t=1}^{|s_u|} l(u_s^n, v_{t+1}^u), \quad (1)$$

其中，f_{seq}是预训练模型，Θ是参数。l是损失函数，我们采用经典的BPR损失作为我们的损失 [14]。

先前研究表明，生成虚假用户行为序列是推广目标物品的有效方式[6, 24, 27]。由于推荐模型是在用户行为数据上训练的。在推荐预训练中，直观地说，我们希望生成的虚假用户行为序列具备以下两个关键特征：(1) 它们应尽可能类似于真实用户的行为序列，以便被攻击的预训练模型能够在下游平台上使用，其中的用户是真实用户。(2) 它们应尽可能少地影响正常的推荐性能。基于此，我们设计了一种简单但有效且高效的随机替换策略。具体而言，给定一个行为序列 $s_u = \{v_1^u, v_2^u, ..., v_{|s_u|}^u\}$ 及其对应的真实物品 $v_{|s_u|+1}$，我们首先找出与真实物品相似的目标物品。然后，我们从其中随机选择一个物品来替换该真实物品，来自

以概率 r生成相似物品。通过这种方式，我们可以生成自然的虚假用户行为序列。实际上，除了被替换的物品外，这些序列均由真实用户产生。此外，通过调整替换率 r和相似度的衡量方式，我们可以平衡攻击效果与推荐效果。本文将同一类别下的物品视为相似物品。

目标用户可以是系统中的全体用户，也可以是某个特定的用户群体（例如，低年级学生）。对于用户群体攻击，攻击者旨在提高目标物品在目标用户组中的曝光率，同时避免影响非目标用户群体。因此，对于用户群体攻击，我们仅在目标用户组上应用基本替换攻击。

3.2 提示增强攻击

提示调优正逐渐成为一种利用预训练模型执行下游任务的新兴范式 [10, 11, 20, 26],，兼具有效性与高效性。提示词也展现出操纵序列模型输出的能力[1, 13]。鉴于提示词的优势，我们为用户群体攻击设计了一种提示增强攻击。提示通常是插入原始序列中的一小段硬文本或软嵌入，有助于从预训练模型中高效提取知识，以用于下游任务。在微调过程中，仅更新提示（其参数数量少得多），而整个预训练模型保持不变。在推荐中，由于重点在于个性化，提示通常由个性化信息（例如用户画像）生成。我们将提示调优形式化如下：

$$\begin{aligned} \boldsymbol{p}_1^u, \boldsymbol{p}_2^u, ..., \boldsymbol{p}_n^u &= Generator(A_u) \\ \hat{s_u^t} &= \{\boldsymbol{p}_1^u, \boldsymbol{p}_2^u, ..., \boldsymbol{p}_n^u, \boldsymbol{v}_1^u, \boldsymbol{v}_2^u, ..., \boldsymbol{v}_t^u\}. \end{aligned} \tag{2}$$

其中，$\{\boldsymbol{p}_1^u, \boldsymbol{p}_2^u, ..., \boldsymbol{p}_n^u\}$ 是生成的提示词，ϑ 是提示生成器的参数，我们在微调提示词时冻结 Θ。

提示词可以由模型提供方提供。在这种情况下，提示词可以与预训练模型一同训练。我们的实验表明，提示词是有效的，尤其是在用户群体攻击中。然而，通常情况下，平台会使用自己的私有数据训练提示词。在这种情况下，实施提示增强攻击颇具挑战性，因为模型提供方（攻击者）不知道提示词的参数和平台的私有数据。实验也表明，联合训练效果不佳。为解决这一挑战，我们提出了一种新颖的三步训练框架：(1) 步骤 *1*：使用提供方的数据预训练序列模型。此步骤旨在构建一个干净模型，以供后续微调。(2) 步骤 *2*：冻结预训练模型的参数，并使用模型提供方的私有数据对模型进行提示调优。目标是模拟平台的提示调优并获取伪提示。(3)步骤 *3*：冻结提示词的参数，并使用我们的基本替换攻击策略微调序列模型。通过这种方式，遭到攻击的预训练模型将对伪提示作出响应，从而实现操纵推荐系统的目标。

经过三步训练后，我们将后门植入预训练模型中。平台在私有数据上进行提示调优后，后门便会被触发。

4 实验

4.1 数据集

我们在两个真实世界的开放数据集上评估攻击方法，分别是 CIKM 和 AliEC。我们假设平台的数据较少、用户活跃度较低，因此需要使用预训练模型。为模拟模型提供方 和下游平台的角色，我们将数据集划分为两个子集。我们将交互次数少于十次的用户视为平台用户，其他用户视为模型提供方的用户。请注意，平台和模型提供方不共享用户和训练数据。**CIKM**。CIKM 数据集是由阿里巴巴发布的电子商务推荐数据集[1]。该数据集包含 13 万件物品。模型提供方拥有 6 万名用户和 200 万个点击实例。平台拥有 3.1 万名用户和 20 万次点击。在用户组攻击中，具有某一共同属性（即性别）的用户被视为目标用户组。**AliEC**。AliEC 是一个电子商务推荐数据集，包含 10.9 万件物品。模型提供方有 9.8 万名用户和 8.9 百万个点击实例，平台则有 6 千多名用户和 3 万个点击实例。年龄范围相同的用户被视为目标用户组。

4.2 实验设置

在本研究中，我们采用基于 Transformer 的序列模型 SASRec[7] 作为基础模型。如前所述，**由于场景和设置不同于以往的攻击方法**，因此不适合将其与其他方法进行比较，所以我们对所提出的方法进行内部比较。**(1)本地模型**（*Local*）：本地模型是在平台的本地私有数据上训练的模型。**(2) 预训练干净模型**（*Clean*）： *Clean* 是在模型提供方的私有数据上训练的干净模型。**(3) 基本替换攻击**（*BRA*）：*BRA* 是经我们的基本替换攻击处理的预训练模型，并直接部署在平台上。**(4) 基本替换攻击 + 微调**（BRA^{FT}）： BRA^{FT} 是基于 *BRA*，在平台的私有数据上微调得到的模型。**(5)提示增强攻击 + 直接使用**（PEA^{D}）： PEA^{D} 采用提示增强方法，直接在平台上使用模型提供方提供的预训练模型和提示。**(6) 提示增强攻击 + 提示调优**（PEA^{PT}）： PEA^{PT}采用专为提示增强攻击设计的三步训练框架。平台使用其私有数据，对提示增强预训练模型进行提示调优。

参数设置与评估对于参数，所有方法的嵌入维度均设为 64。所有方法的替换率 r 均为 0.5。我们对超参数进行网格搜索。L2归一化系数设为 $1e-6$。推荐预训练（即模型提供方）的学习率设为 $1e-4$，推荐调优（即平台）的学习率设为 $1e-5$。我们随机选择 300 个物品作为攻击中的目标物品。对于 CIKM，目标用户为男性。对于 AliEC，目标用户的年龄在 20 至 30 岁之间。**评估**方面，我们采用经典的留一法策略来评估性能 [3, 7, 22]。我们使用广泛认可的 HIT@N 和 NDCG@N 来衡量推荐准确率。

[1]https://tianchi.aliyun.com/competition/entrance/231719/introduction

表 1：用户群体攻击评估。λ 表示目标用户与非目标用户在 HIT@N 和 NDCG@N 上的平均攻击成功率差距。所有改进相较于基线方法均具有统计显著性（t检验，$p < 0.05$）。

		目标用户				非目标用户				隐蔽性
数据集	模型	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	λ
CIKM	*Clean*	0.002	0.001	0.004	0.002	0.002	0.002	0.004	0.002	**0.9x**
	BRA	0.755	0.388	1.798	0.72	0.234	0.116	0.610	0.236	**3.1x**
	PEA^D	1.437	0.786	3.056	1.302	0.072	0.037	0.017	0.0	**44.6x**
	BRA^{FT}	0.033	0.016	0.106	0.040	0.016	0.008	0.047	0.018	**2.1x**
	PEA^{PT}	0.725	0.393	1.586	0.667	0.064	0.034	0.151	0.061	**11.1x**
AliEC	*Clean*	0.009	0.006	0.016	0.008	0.007	0.004	0.014	0.006	**1.3x**
	BRA	0.272	0.148	0.671	0.269	0.114	0.057	0.288	0.113	**2.4x**
	PEA^D	1.649	0.972	3.303	1.499	0.033	0.018	0.067	0.029	**51.2x**
	BRA^{FT}	0.272	0.144	0.660	0.267	0.112	0.051	0.284	0.111	**2.5x**
	PEA^{PT}	0.597	0.335	1.267	0.548	0.055	0.031	0.116	0.050	**10.8倍**

表 2：全局攻击评估结果。所有提升均显著优于基线方法（t检验，$p < 0.05$）。

数据集	模型	H@5	N@5	H@10	N@10
CIKM	*Local*	0.0002	0.0001	0.0004	0.0002
	Clean	0.002	0.001	0.004	0.002
	BRA	0.649	0.338	1.323	0.554
	BRA^{FT}	0.456	0.239	0.901	0.381
AliEC	*Local*	0.002	0.001	0.004	0.002
	Clean	0.010	0.006	0.016	0.008
	BRA	0.789	0.444	1.656	0.721
	BRA^{FT}	0.639	0.357	1.353	0.585

4.3 全局攻击评估

在全局攻击评估中，我们试图提升目标物品在平台所有用户中的曝光率。表 2 展示了我们全局攻击的效果。我们发现：（1） *BRA* 成功在预训练模型中植入了后门。与 *Local*相比，*BRA*的目标物品 HIT@5（表示攻击成功率）提高了数百倍以上。（2）与 *BRA* 和 BRA^{FT} 相比，通过对预训练模型进行微调，攻击效果有所减弱，但仍提高了数百倍。这不仅证明了我们攻击的有效性，也表明微调是针对我们后门攻击的一种较好的防御机制。（3）在推荐准确率方面，即使基于遭受攻击的预训练模型，我们 BRA 模型的结果也显著优于 Local 模型。在各项指标上，BRA 的准确率约为本地模型的 6 至 8 倍。需要注意的是，攻击通常不可避免地会导致准确率下降，而这并非本研究的重点。总体而言，遭受攻击的预训练模型带来的准确率大幅提升，证实了我们的攻击在推荐预训练中的隐蔽性。

4.4 用户组攻击评估

在用户群体攻击中，我们认为优秀的攻击者应具备两个特点：首先，攻击者能够**提升特定目标用户**中目标物品的曝光率。其次，攻击者应**尽可能不影响非目标用户**中目标物品的曝光率。因此，目标用户与非目标用户之间的曝光率**差距**越大，这种攻击就越隐蔽。如表 1 所示，我们可以观察到：(1) 所有攻击方法都成功实施了用户群体攻击。与 *Clean* 模型相比，HIT@N 和 NDCG@N 指标提升了数百倍以上。(2) 我们的提示增强方法在攻击效果上使目标用户与非目标用户之间的**差距达到 10 至 50 倍**，而基础替换方法的**差距仅为 2 至 3 倍**。更高的 λ 表明预训练攻击在隐蔽性方面表现更好，这意味着在提示词上经过特殊设计的 PEA方法能够实现更有效、更隐蔽的攻击。(3) 我们的 PEA^{PT} 使目标物品在目标用户中的曝光率提升了 300 倍以上，且攻击效果在目标用户与非目标用户之间的**差距超过 10 倍**。请注意，PEA^{PT} 仅通过预训练模型实施攻击，并不了解平台的微调设置和数据。这证明了所提出的三步式提示增强攻击框架的强大能力。

4.5 后门检测

此前的检测研究主要关注识别数据集中恶意用户行为。遗憾的是，在推荐预训练场景中，由于平台无法访问模型提供方的数据，以往的方法并不适用。在这项工作中，我们针对这一新场景提出了一种基于统计的检测方法。该方法包括三个步骤：(1) 在平台的私有数据上训练一个模型。随后，我们利用该模型估计系统中所有物品的平均 HIT@N（$HIT@N^L$）。(2) 利用预训练模型计算平均 HIT@N（$HIT@N^P$）。(3) 将两者的差异嵌入表示$Diff = HIT@N^P - HIT@N^L$输入 K 均值模型，该模型将 $Diff$聚类为两类，以检测异常物品。这里我们采用 N=5、10、50、100。检测结果如表 3 所示。我们可以看到：(1) 与用户群体攻击相比，全局攻击更容易被检测出来。(2) 与 BPR^{group} 相比，我们的提示增强攻击更难被检测出来。针对推荐预训练中此类攻击的更有效检测方法仍有待探索。

表3：我们的检测方法的结果。

数据集	模型	F1	召回率	精确率
CIKM	BRA	0.467	0.4	0.563
	BRA^{group}	0.151	0.100	0.260
	PEA^{PT}	0.060	0.040	0.120
AliEC	BRA	0.210	0.157	0.318
	BRA^{group}	0.017	0.035	0.023
	PEA^{PT}	0.0	0.0	0.0

5 结论与未来工作

在本研究中，我们首次系统地揭示了预训练推荐模型中存在的后门攻击威胁，并相应地提出了两种攻击方法。具体而言，我们提出了一种有效的基础替换策略，用于植入后门。此外，针对提示调优，我们提出了一种提示增强攻击，以增强隐蔽性更高的用户组攻击。实验结果表明，我们的攻击方法能够显著提高目标物品在目标用户（群体）中的曝光率。未来，我们将探索针对预训练模型中此类攻击的更优检测与防御方法，并研究模型提供商和平台可能面临的用户隐私问题。

6 致谢

本研究工作得到中国国家重点研发计划项目（编号：2021ZD0113602）的资助。本研究工作还得到中国国家自然科学基金项目（编号：62206266、62176014 和 61976204）的资助。Zhao Zhang 还获得了中国博士后科学基金会项目（编号：2021M703273）的资助。

参考文献

[1] Tom B Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell等。2020年。语言模型是少样本学习者。*NeurIPS*（2020年）。[2] Shijie Geng、Shuchang Liu、Zuohui Fu、Yingqiang Ge和Yongfeng Zhang。2022年。推荐即语言处理（RLP）：统一的预训练、个性化提示与预测范式（P5）。*RecSys*（2022年）。[3] Xiangnan He和Tat-Seng Chua。2017年。面向稀疏预测分析的神经因子分解机。载于*SIGIR*会议论文集。[4] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas和Domonkos Tikk。2016年。基于循环神经网络的会话推荐。载于*ICLR*。[5] Yupeng Hou、Shanlei Mu、Wayne Xin Zhao、Yaliang Li、Bolin Ding和Ji-Rong Wen。2022年。迈向推荐系统的通用序列表示学习。载于*KDD*。[6] Hai Huang、Jiaming Mu、Neil Zhenqiang Gong、Qi Li、Bin Liu和Mingwei Xu。2021年。针对基于深度学习的推荐系统的数据投毒攻击。*arXiv*预印本 *arXiv:2101.02644*（2021年）。[7] Wang-Cheng Kang和Julian McAuley。2018年。自注意力序列推荐。载于*ICDM*会议论文集。[8] Shyong K. Lam和John Riedl。2004年。为乐趣和收益刷量攻击推荐系统。万维网大会（2004年）。[9] Bo Li、Yining Wang、Aarti Singh和Yevgeniy Vorobeychik。2016年。针对基于因子分解的协同过滤的数据投毒攻击。神经信息处理系统（2016年）。[10] Xiang Lisa Li和Percy Liang。2021年。前缀调优：优化用于生成的连续提示。*ACL-IJNLP*（2021年）。[11] Xiao Liu、Kaixuan Ji、Yicheng Fu、Zhengxiao Du、Zhilin Yang和Jie Tang。2021年。P-tuning v2：提示调优在不同规模和任务中普遍可媲美微调。*arXiv*预印本 *arXiv:2110.07602*（2021年）。[12] Bamshad Mobasher、Robin Burke、Runa Bhaumik和Chad Williams。2007年。迈向可信推荐系统：攻击模型与算法鲁棒性分析。*ACM互联网技术汇刊*（2007年）。[13] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark等。2021年。通过自然语言监督学习可迁移视觉模型。载于国际机器学习大会。PMLR，8748–8763。[14] Steffen Rendle、Christoph Freudenthaler、Zeno Gantner和Lars Schmidt-Thieme。2009年。贝叶斯个性化排序：基于隐式反馈的贝叶斯个性化排序。不确定性人工智能会议（2009年）。[15] Mingdan Si和Qingshan Li。2020年。针对协同过滤推荐系统的刷量攻击：综述。《人工智能评论》53卷（2020年），291–319页。[16] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou和Peng Jiang。2019年。BERT4Rec：基于Transformer双向编码器表示的序列推荐。载于*CIKM*会议论文集。[17] Changxin Tian、Zihan Lin、Shuqing Bian、Jinpeng Wang和Wayne Xin Zhao。2022年。用于序列推荐的时间对比预训练。载于*CIKM*会议论文集。[18] Chenwang Wu、Defu Lian、Yong Ge、Zhihao Zhu和Enhong Chen。2021年。用于推荐系统中基于影响力的投毒攻击的三重对抗学习。载于第*27*届*ACM SIGKDD*知识发现与数据挖掘大会论文集。1830–1840页。[19] Yiqing Wu、Ruobing Xie、Yongchun Zhu、Fuzhen Zhuang、Ao Xiang、Xu Zhang、Leyu Lin和Qing He。2022年。通过提示实现推荐中的选择性公平。载于*SIGIR*会议论文集。[20] Yiqing Wu、Ruobing Xie、Yongchun Zhu、Fuzhen Zhuang、Xu Zhang、Leyu Lin和Qing He。2022年。用于序列推荐的个性化提示。*arXiv*预印本 *arXiv:2205.09666*（2022年）。[21] Chaojun Xiao、Ruobing Xie、Yuan Yao、Zhiyuan Liu、Maosong Sun、Xu Zhang和Leyu Lin。2021年。UPRec：面向推荐系统的用户感知预训练。*arXiv*预印本（2021年）。[22] Xu Xie、Fei Sun、Zhaoyang Liu、Shiwen Wu、Jinyang Gao、Jiandong Zhang、Bolin Ding和Bin Cui。2022年。用于序列推荐的对比学习。载于国际数据工程大会。电气电子工程师学会。[23] Xinyu Xing、Wei Meng、Dan Doozan、Alex C Snoeren、Nick Feamster和Wenke Lee。2013年。请认真对待：针对个性化服务的污染攻击。载于*USENIX*安全研讨会。671–686页。[24] Zhenrui Yue、Zhankui He、Huimin Zeng和Julian McAuley。2021年。通过无数据模型提取对序列推荐系统进行黑盒攻击。载于第*15*届*ACM*推荐系统大会论文集。44–54页。[25] Zheni Zeng、Chaojun Xiao、Yuan Yao、Ruobing Xie、Zhiyuan Liu、Fen Lin、Leyu Lin和Maosong Sun。2021年。通过预训练实现推荐中的知识迁移：综述与展望。《大数据前沿》（2021年）。[26] Chi Zhang、Rui Chen、Xiangyu Zhao、Qilong Han和Li Li。2023年。多行为推荐中的去噪与提示调优。*arXiv*预印本 *arXiv:2302.05862*（2023年）。[27] Hengtong Zhang、Yaliang Li、Bolin Ding和Jing Gao。2020年。针对下一物品推荐的实用数据投毒攻击。载于*2020*年万维网大会论文集。2458–2464页。[28] Guorui Zhou、Xiaoqiang Zhu、Chenru Song、Ying Fan、Han Zhu、Xiao Ma、Yanghui Yan、Junqi Jin、Han Li和Kun Gai。2018年。用于点击率预测的深度兴趣网络。载于*KDD*会议论文集。[29] Kun Zhou、Hui Wang、Wayne Xin Zhao、Yutao Zhu、Sirui Wang、Fuzheng Zhang、Zhongyuan Wang和Ji-Rong Wen。2020年。S3-rec：通过互信息最大化进行序列推荐的自监督学习。载于*CIKM*。