第14周周报 覃海旭

一、论文阅读

1.Cross-domain Recommendation with Behavioral Importance Perception

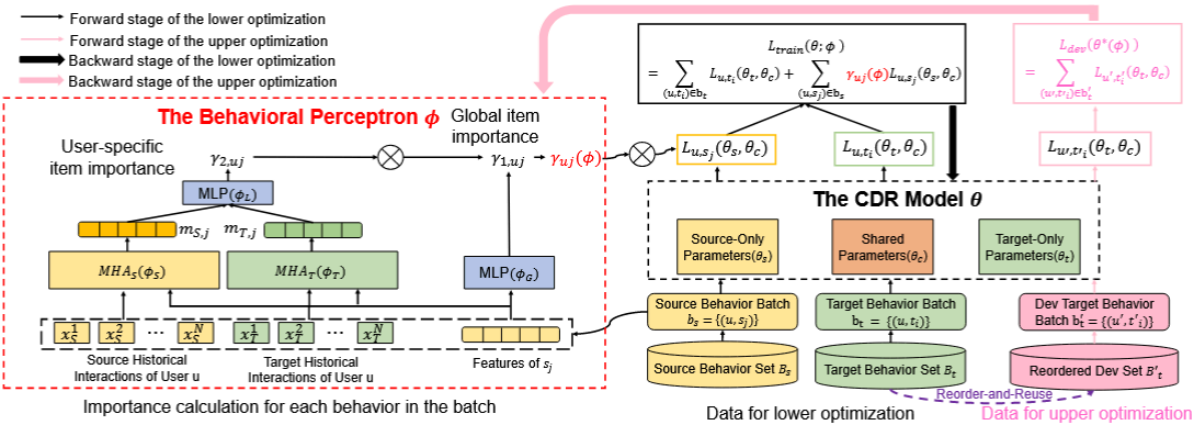
> 期刊会议：WWW2023
> 方向分类：跨域推荐

- **解决的问题：**

  现有方法主要集中于模型结构设计以实现源域知识的迁移，但忽略了中源域不同行为对模型优化的不同重要性。例如，如果想利用用户在“图书”域的行为来帮助“玩具”域的推荐，与“儿童”或“玩具故事”相关的图书交互行为可以为目标域提供有用信息，而与“爱情故事”相关的图书交互行为则几乎没有帮助。

- **创新点：**

  1. 设计了行为感知器，可以预测每个源域行为的重要性，结合全局影响和用户特定的局部影响进行评估。
  2. 引入双层优化方法，通过加权源域行为损失和目标域损失的联合优化提升模型性能。
  3. 提出重新排序与复用策略，有效利用目标域数据以降低偏差并减少信息损失。



- **模型架构：**BIAO框架由以下模块构成：

  1. **行为感知器**：用于确定每个源域行为的重要性，从以下两个方面判断：
     - **全局重要性**：通过MLP对源域物品的特征（如物品ID、类别等）进行编码得到全局重要性权重，用于衡量源域物品对目标域的整体影响。
     - **局部重要性**：利用自注意力机制和MLP提取用户历史交互数据中与当前源域行为相关的兴趣。用于衡量源域物品对交互用户的个性化影响。
     - 结合全局和局部权重，计算行为的最终重要性。
  2. **双层优化**：联合优化跨域推荐模型参数θ和行为感知器参数ϕ
     - **下层优化**：固定行为感知器参数ϕ，通过目标域损失和加权源域损失的联合优化，更新推荐模型的参数θ，得到最优的$\theta^*(\phi)$。
     - **上层优化**：当$\theta^*(\phi)$达到最优时，利用隐式梯度优化行为感知器的权重参数ϕ。
  3. **实用优化算法：**设计高效策略来进行双层优化：
     - **批量优化：**从整个数据集中取出一个个batch进行损失和参数梯度的计算。
     - **固定间隔的下层优化：**理论上需要下层优化收敛到最优值再进行上层优化，过于耗时，为了提高效率，仅在每次上层优化前进行固定轮次的下层优化。

        - **重新排序与复用策略：**
            - 先前的研究通常从目标域数据集中拆分一个小数据集用于上层优化，并使用剩余的用于下层优化。然而，这种数据划分容易导致上层优化的偏差和下层优化中的信息丢失。
            - 为了解决该问题，通过对目标域数据集进行重新排序来获得一个开发数据集，并使用这个重新排序的数据集进行下层优化。由于采用批量优化，在下层优化中提取的批次数据与上层优化中提取的批次数据不同。
- **数据集：**
    - Amazon(Books→Movies)
    - Amazon(Books→CDs)
    - Amazon(Books→Elec)
    - Amazon(Books→Toys)
    - Amazon(CDs→Cloth)
    - Amazon(CDs→Kitchen)
    - Amazon(Elec→Cloth)

2.PPGenCDR: A Stable and Robust Framework for Privacy-Preserving Cross-Domain Recommendation

> 期刊会议：AAAI2023
> 方向分类：隐私保护跨域推荐

- **解决的问题：**

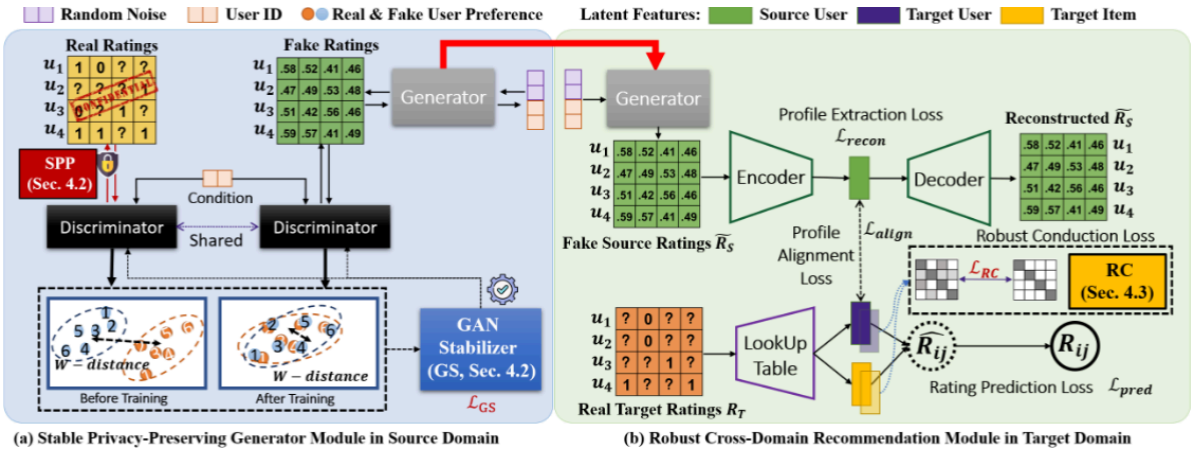
  跨域推荐中的隐私保护问题。现有的将**隐私保护数据发布**应用于跨域推荐的方法存在以下不足：

  1. 在隐私保护中加入大量噪声可能显著降低模型的推荐性能。
  2. 在源域模型的梯度中加入扰动后会导致GAN模型在生成器与判别器的对抗过程中很难收敛。
  3. 目标域模型在适应和使用经过隐私保护扰动的源域数据时鲁棒性较低，可能造成负迁移而降低模型性能。

  > **隐私保护数据发布:**其核心是**数据扰动和数据脱敏**。它通过对原始数据进行某种形式的修改，使得数据集的隐私性得以保持，同时仍然能提供有价值的统计信息或分析结果。

- **创新点：**

  1. 设计**选择性隐私保护器(SPP)**，仅对直接访问原始数据的模型层的梯度使用Ŕenyi差分隐私(RDP)技术对梯度进行选择性噪声扰动（针对问题1 ）
  2. 设计了一个**GAN稳定器(GS)**，通过从控制理论中派生的额外正则化项来稳定判别器的目标。（针对问题2 ）
  3. 引入了一个**鲁棒性协调器(RC)**，通过对批量中冗余表示的每个维度进行解耦，提升了目标域模型的性能。（针对问题3）



(a) Stable Privacy-Preserving Generator Module in Source Domain

(b) Robust Cross-Domain Recommendation Module in Target Domain

- **模型架构：**PPGenCDR框架如下：
    1. **稳定隐私保护生成器(SPPG)**
        - **目标：** 在源域中生成符合隐私保护要求的用户伪交互数据。
        - **主要组件：**
            - **GAN**
                - 使用Wasserstein距离来衡量真实用户偏好与伪用户偏好的差异
                - **生成器：**通过用户emb和随机噪声生成伪用户偏好数据。通过最小化Wasserstein距离来欺骗判别器。
                - **判别器：**最大化Wasserstein距离，从而区分真实和伪造的用户偏好。
            - **选择性隐私保护(SPP)：**仅对直接访问原始数据的判别器的第一层梯度进行RDP扰动，从而在保护源域中的私有数据的同时避免在目标域中造成严重的性能下降。
            - **GAN稳定器(GS)：** 为判别器的目标函数添加一个额外的正则化项，从而稳定判别器的学习过程。
    2. **鲁棒跨域推荐模块(RCDR)**
        - **目标：** 在目标域中使用源域生成的伪数据提升推荐性能。
        - **主要组件：**
            - **特征提取模块：** 把生成器生成的伪用户偏好重构为源域用户潜在特征。
            - **推荐预测模块：** 基于目标域数据预测用户偏好。
            - **跨域特征对齐模块：**最小化源域和目标域中的用户特征之间的差异来将知识从源域传递到目标域。
            - **鲁棒性协调器：** 解耦用户和物品特征的相关性，通过优化用户和物品特征的交叉相关矩阵，将非对角线元素的值调整为接近0，使得不同特征维度之间的关系尽可能独立，从而减少噪声影响。
- **数据集：**
    - Douban(book→music)
    - Douban(movie→book)
    - Douban(movie→music)
    - Amazon(music→book)
    - Amazon(movie→book)
    - Amazon(movie→music)

3.Privacy-Preserving Cross-Domain Sequential Recommendation

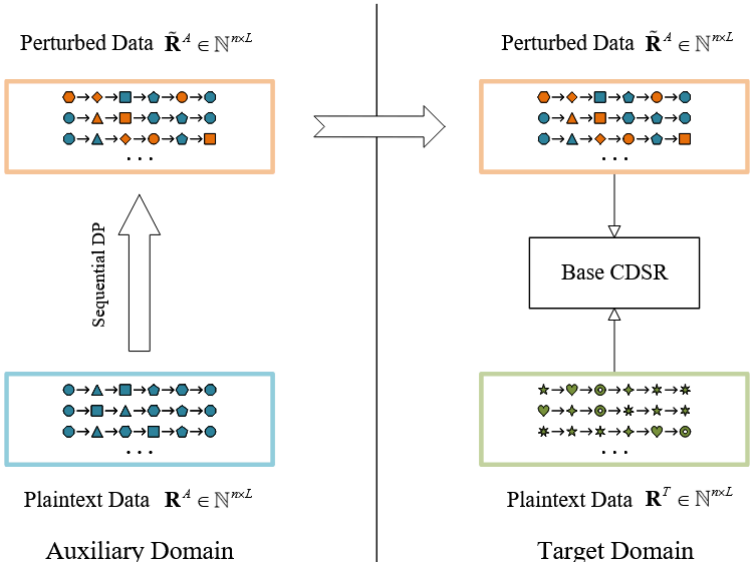
> 期刊会议：ICDM2023
> 方向分类：隐私保护跨域推荐

- **解决的问题：**

  跨域序列推荐系统中隐私保护的需求。

- **创新点：**

  1. **序列差分隐私（SDP）**：
     - 提出了新的隐私定义，保护用户交互数据的**ID信息**和**序列顺序信息**。
     - 与传统差分隐私只针对单点数据不同，SDP扩展到序列数据，保护了用户交互的动态特征。
  2. **设计了一种满足SDP的随机机制**：
     - 该机制通过随机扰动用户交互序列中的每一项，引入噪声以模糊真实数据。
     - 机制允许序列中的物品交换、插入、替换或删除，保证隐私的同时保留足够的推荐信息。
     - 提供了理论证明，证明此机制满足SDP。
  3. **非侵入式方法**：
     - 提出的PriCDSR框架无需修改现有的跨域序列推荐模型，只需在数据预处理阶段引入随机机制。



- **模型架构：**PriCDSR架构如下：

  1. **两阶段流程**：
     - **第一阶段**：在辅助域数据上引入随机机制添加噪声，生成一个扰动后的序列数据矩阵，并将其传输至目标域。
     - **第二阶段**：目标域使用现有的跨域序列推荐算法作为基础模型，结合扰动后的辅助域数据进行推荐。
  2. **随机机制细节**：
     - **扰动方式**：
       - 对于序列中的每个物品都通过随机机制进行一次采样，采样范围是源域中所有物品ID+0(填充项)。

        - 检查采样出的物品ID是否出现在用户的后续交互物品ID中。如果是，则将后续的物品ID与当前物品ID交换序列中的位置。否则用采样到的物品ID替换当前物品ID。
    - **计算复杂度**：机制时间复杂度为 $O(n \times L)$（其中 n 为用户数，L 为序列长度），空间开销较低。
- **数据集：**
    - Amazon(Movie←Book)
    - Amazon(Movie←CD)
    - Amazon(Book←Movie)
    - Amazon(Book←CD)
    - Amazon(CD←Movie)
    - Amazon(CD←Book)

4.Cross-Domain Causal Preference Learning for Out-of-Distribution Recommendation

> 期刊会议：DASFAA2024
> 方向分类：因果推断/跨域推荐

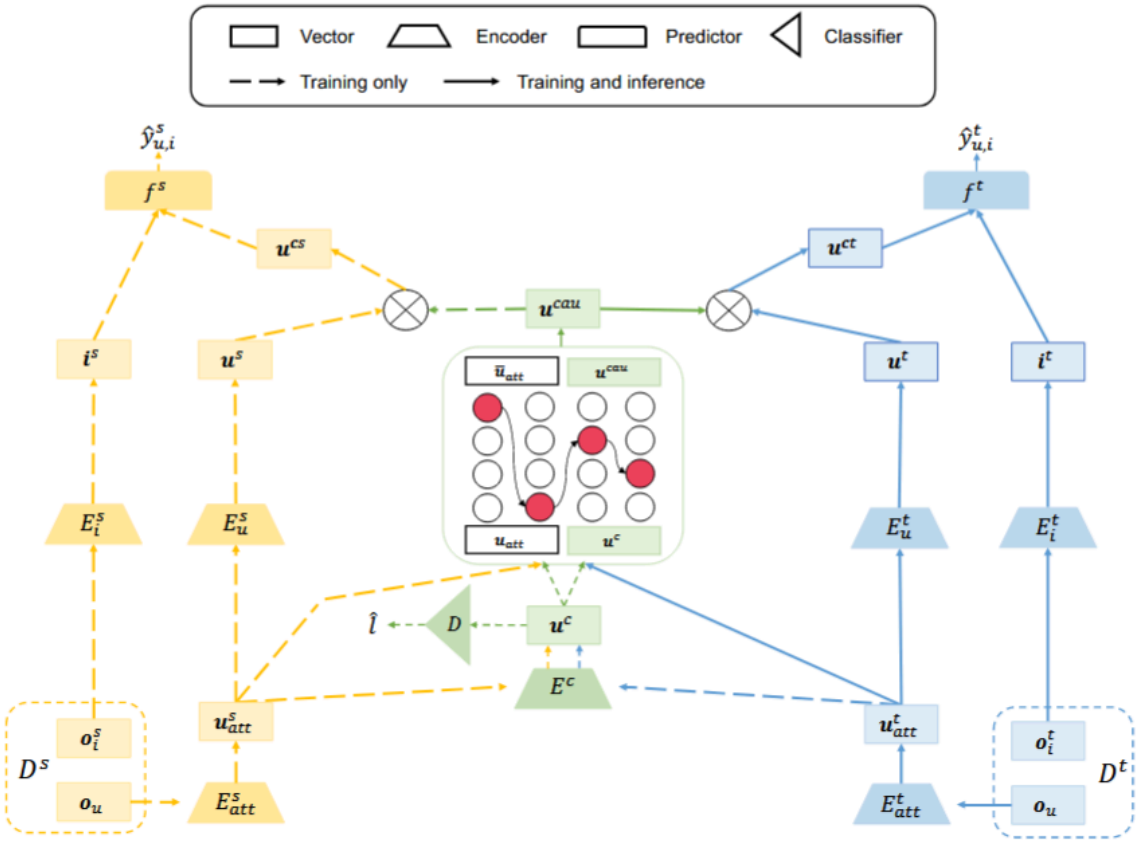
- **解决的问题：**

  现有推荐系统假设训练数据和测试数据分布一致，但在实际中用户的偏好会随时间变化，这种偏移会导致训练和测试数据的分布不一致，称为 **Out-of-Distribution (OOD)** 问题。

- **创新点：**

  利用在源域中学习到的因果关系改善目标域的推荐性能。

  1. **因果结构学习**：通过学习因果结构（用户属性与用户偏好之间的因果关系），从而使模型能够理解和捕捉用户偏好的生成过程，这样即便在数据分布变化的情况下也能进行有效推荐。
  2. **隐性属性建模**：扩展了因果推理模型的应用场景，从显性属性（如性别、年龄）的学习扩展到隐性属性（如物品 ID），减少对隐私敏感数据的依赖，并适应数据稀疏场景。



- **模型架构：**CDCOR模型架构如下：

1. **领域特定模块**：
    - 编码源域和目标域的用户和物品嵌入
    - 编码用户域特定偏好。
2. **领域共享模块**：
    - 把源域和目标域的用户嵌入输入到域共享编码器(生成器)，得到源域和目标域的域共享偏好。
    - 判别器+梯度反转层判断域共享偏好来自源域还是目标域，使得生成器编码的共享偏好不依赖于具体领域，从而减少了领域偏移带来的影响。
3. **因果结构学习模块**：
    - 基于**有向无环图**建模用户属性与偏好的因果关系，利用优化约束确保模型捕获因果不变性。
    - 图中节点是用户嵌入和用户共享偏好，边A_{ij}表示节点 i 对节点 j 的因果影响。
    - 边的方向表示因果关系，无环是为了确保因果推理的原则（因不能是自己的果）。
    - 通过优化图的邻接矩阵和无环约束损失来进行学习。最终的输出是因果优化后的域共享嵌入。

- **数据集：**
    - Douban(movie→book)
    - Tenrec(video→article)

二、对比总结

对已阅读的跨域推荐相关论文进行对比总结**（加粗为本周论文）**：

对抗训练（GAN）思想的应用：

模型	方法	目的
DiCUR-CDR(第9周第1篇)	生成器生成反馈向量，判别器判别该反馈向量是生成器生成的还是用户真实的反馈	使得生成器生成更加真实的反馈向量作为推荐结果
CDAT(第9周第2篇)	生成器编码源域和目标域的领域不变的用户偏好分布（可以理解为域共享表示）,判别器判别该表示来自源域还是目标域	使得生成器编码的用户偏好表示具有领域不变性
C^2DR(第10周第4篇)	生成器编码域特定表示和域共享表示，域分类器用作判别器，判别输入的表示是域共享表示还是域特定表示。引入梯度反转层用于欺骗分类器，使得模型更难以区分域共享信息和域特定信息	使得生成器编码的共享表示真正对所有域都通用，而特定表示仅适用于相应的域
OmniMatch(第11周第2篇)	与C^2DR的方式基本相同，同样是采用具有梯度反转层的域分类器进行对抗训练	使用户特征保持领域无关，减少源域和目标域特征分布的差异
TPUF(第13周第1篇)	与上面2个基本相同，同样是采用具有梯度反转层的域分类器进行对抗训练	对齐来自两个领域的特征

模型	方法	目的
PPGenCDR(第14周第2篇)	**生成器生成伪用户偏好数据。通过最小化Wasserstein距离来欺骗判别器 判别器最大化Wasserstein距离，从而区分真实和伪造的用户偏好**	**在源域中生成符合隐私保护要求的用户伪交互数据**
CDCOR(第14周第4篇)	**与CDAT相同，生成器编码源域和目标域的域共享表示,判别器判别该表示来自源域还是目标域，也同样在判别器前加入了梯度反转层**	**使得生成器编码的用户偏好表示具有领域不变性**

域共享表示和域特定表示的解耦：

模型	方法	目的
DiCUR-CDR(第9周第1篇)	判别典型相关分析	在最大化域共享用户表示之间的相关性的同时添加额外的约束来学习域特定表示之间的差异
MITrans(第10周第2篇)	互信息约束	最大化不同域间物品嵌入的相似度以学习共享偏好，最小化域间物品嵌入的依赖关系以保留特定偏好
C^2DR(第10周第4篇)	因果表示解耦	确保域共享表示和域特定表示的向量正交性（余弦相似度最小化）和统计分布独立性（协方差为0）
FedDCSR(第13周第3篇)	互信息约束	(1)使域共享表示与域特定表示的互信息最小化，从而扩大它们的差异。 (2)使域共享表示与全局表示（域共享表示之和）的互信息最大化，促进领域共享特征与全局特征的一致性。 （3）使域特定表示与对应序列编码的互信息最大化，使领域专属表示能够重构其所在领域的序列数据。

负迁移问题的优化：

模型	方法	目的
PPA(第9周第5篇)	原型感知学习	通过构建偏好原型以量化用户偏好，从而减少源域冗余特征的影响
CUT(第10周第1篇)	相似性约束	通过限制损失函数让目标域中相似的用户对的嵌入在源域中仍然保持相似
CrossAug(第10周第3篇)	特征交叉重构	通过数据增强的方式来缓解负迁移
C^2DR(第10周第4篇)	正交化约束	强制域共享信息在两个域的损失函数梯度之间保持正交性。确保域特定信息在一个域内仅影响该域的推荐结果，不会对其他域的推荐产生干扰
M2GNN(第11周第4篇)	动态路由网络+自注意力+skip-gram正则化	增强标签语义相关性并去除无关标签，减少噪声对模型的影响

模型	方法	目的
CAT-ART(第12周第4篇)	注意力机制	通过注意力机制选择性集成相关领域的嵌入，同时弱化不相关或质量低的嵌入
BIAO(第14周第1篇)	**行为感知器**	**细化到行为粒度，衡量源域不同行为对目标域推荐的不同重要性**

无重叠user/item下的跨域推荐：

模型	方法
PrepRec(第9周第4篇)	通过建模物品流行度的变化来学习通用表示
MITrans(第10周第2篇)	通过互信息来学习不同域的共享偏好

特征信息补充/增强样本的生成:

模型	方法
MACD(第11周第1篇)	引入辅助行为序列（如点击）来丰富用户的兴趣信息
OmniMatch(第11周第2篇)	用辅助域相似用户的评论作为目标域冷启动用户的辅助评论
M2GNN(第11周第4篇)	引入标签信息
AutoTransfer(第11周第3篇)	显式选择源域中的实例迁移到目标域
CrossAug(第10周第3篇)	通过特征交叉重构的方式生成域内和跨域增强样本
MITrans(第10周第2篇)	通过预训练语言模型(BERT)，从项目文本数据中提取语义嵌入，用于后续的偏好学习
ARISEN(第13周第2篇)	引入用户的查询数据作为因果学习的工具变量
FedDCSR(第13周第3篇)	通过随机打乱序列顺序的方法对用户序列进行数据增强，生成新的训练样本
COAST(第13周第4篇)	从内容数据（如评论、标签、用户/物品档案）中提取特征作为图节点的初始嵌入表示

负采样优化：

模型	方法	目的
SCE(第7周第4篇)[普通序列推荐]	通过分桶计算相似度的方式找到难负样本，避免对所有可能item组合进行计算	减少查找难负样本的计算量
RealHNS(第12周第2篇)[跨域序列推荐]	（1）引入难负样本和伪难负样本的概念，排除那些与正样本过于相似的伪难负样本 （2）考虑到跨域推荐场景下的离群点用户，针对它们设计了两个不同的负样本集	找到更高质量的负样本

因果推断的使用：

模型	方法	目的
C^2DR(第10周第4篇)	将域共享信息和域特定信息视作因果变量，并设计了因果解耦正则化项，来确保域共享和域特定信息在表示空间中满足因果独立性（即向量正交性和统计分布独立性）	域共享表示和域特定表示的解耦
ARISEN(第13周第2篇)	基于工具变量的因果推断，引入用户查询记录作为工具变量，采用两阶段最小二乘法拟合用户交互（原因变量）与用户查询（工具变量）	将原因变量分解为因果关联（拟合部分）和非因果关联（残差部分）,从而去除混杂因素造成的偏差
CDCOR(第14周第4篇)	**有向无环图，节点是user emb和用户共享偏好，边A_{ij}表示节点 i 对节点 j 的因果影响，通过优化图的邻接矩阵来进行学习，得到因果优化后的域共享表示**	**使模型能够理解和捕捉共享偏好的生成过程，这样即便在数据分布变化的跨域场景下也能进行有效推荐**

跨域隐私保护：

模型	场景	方法
TPUF(第13周第1篇)	仅有源域预训练用户特征/序列推荐	把源域用户特征进行映射，再复制Q次（Q为序列长度），来模拟源域的用户序列
FedDCSR(第13周第3篇)	联邦推荐（域之间只传递用户全局表示和域共享编码器参数）/序列推荐	在每轮本地训练中利用全局表示和互信息原理来更好地学习本地域共享表示和域特定表示
PPGenCDR(第14周第2篇)	**目标域只有扰动后的源域用户交互数据**	**（1）源域中使用GAN和RDP扰动生成扰动伪样本给目标域 （2）目标域把伪样本重构为源域特征，并使用特征对齐和相关性解耦的方法来减小扰动的影响**
PriCDSR(第14周第3篇)	**目标域只有扰动后的源域用户交互数据**	**设计了一种满足序列差分隐私的随即机制，通过随机扰动用户交互序列中的每一项，引入噪声以模糊真实数据**