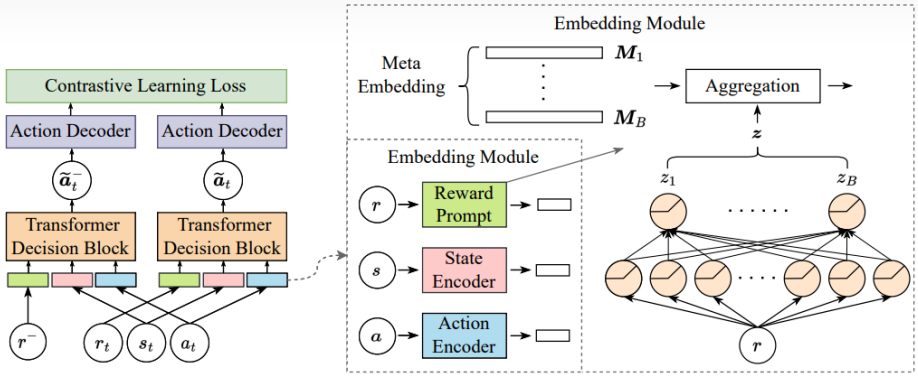
知乎



百度-2023：使用Decision Transformer，攻克用户留存难题的推荐策略

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

SRS模型通过历史交互和潜在有趣内容的推荐受到关注，用户平台的参与度反映了用户的即时反馈和喜好，但不足以揭示其所有偏好。优化用户长期参与度（即用户留存）是提升用户满意度的关键。尽管现有的RL-based SRS能够探索和模拟用户的动态兴趣，但由于离线学习的挑战，它们还有许多不足之处。

与游戏场景不同，在线SRS中从头开始训练RL代理是不可行的，因为推荐不合适的物品可能会导致用户流失的风险。因此，整个社区最近对离线RL-based SRS给予了关注。
然而，将离线RL应用到实践中在价值基和政策基方法中都是令人沮丧的。对于价值基方法，著名的不稳定问题（即"致命三角形"）推动了模型基方法的发展。但是，由于推荐场景中的巨大状态空间，估计转换概率是一个问题，并进一步导致性能不佳。对于策略基方法，反事实策略评估的无界方差驱动着社区剪切或丢弃反事实权重，这可能导致权重不准确并阻碍性能。

（1）缺乏奖励建模。作为DT最关键的部分，奖励直接影响了推荐的质量。然而，在DT中，将奖励转换为嵌入忽略了其部分顺序，导致模型训练中的缺陷。
（2）推荐生成的不一致性。在DT下，推荐是根据最大奖励生成的，这与训练阶段遇到的多样化奖励不符。因此，该模型无法利用数据中小型回报的知识（图例所示）。
（3）性能评估不可靠 。尽管DT解决了离线学习的问题，但我们仍然需要重要加权的离线评估来衡量学习策略的效果，这导致方差无界并导致评估结果不可靠。

为了解决这些问题，我们提出了一个新颖的框架DT4Rec，该框架首先部署了Decision Transformer用于推荐。具体来说， 我们通过自动权重加总从离散化奖励值生成的元嵌入来生成奖励嵌入，以保持奖励之间的部分顺序关系。

然后，我们引入加权对比学习来弥补推理和训练之间的差异，它通过对比大型和小型奖励样本来利用较小回报样本。 此外，我们还提出两种新的可靠指标，即基于模型的和基于相似性的用户保留分数，以全面评估策略。与无策略评估方法相比，它实现了更低的方差（即更大的稳定性），从而在性能评估方面达到更低的方差（图例所示）。

主要贡献如下：

- 提出了一种新型的Decision Transformer基础的SRS模型，DT4Rec。该模型消除了使用基于RL的推荐系统进行离线学习的困难。
- 提出了自动离散化奖励提示和对比性监督策略学习的方法，来解决奖励建模的缺陷和训练-推理间的差距。
- 提出了一种基于模型和相似性的用户保留分数评估方法。该方法相比于离线策略更公正地评估模型性能。
- 实验验证了我们提出的模型在两个基准数据集上的优势。

知乎

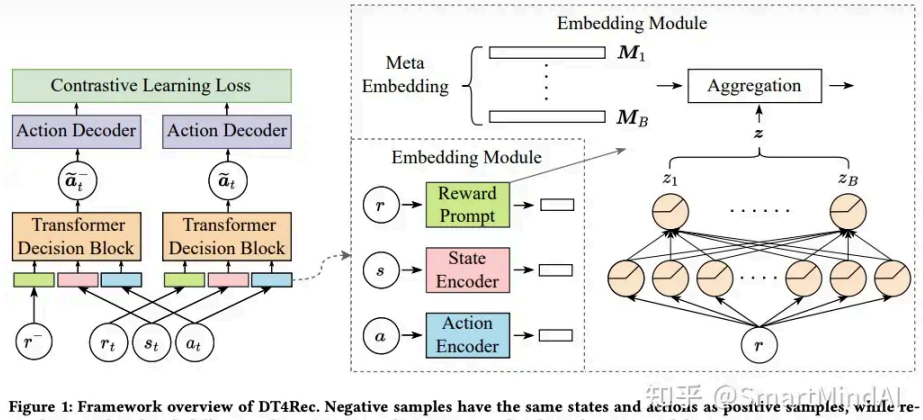


Figure 1: Framework overview of DT4Rec. Negative samples have the same states and actions as positive samples, while rewards are replaced with different values. Positive and negative samples share the same model parameters.

Problem Formulation

给定t步后的推荐序列U_t和N个用户可能感兴趣的物品集合a_t，求解最优化问题$\max_a \Theta(a)$，以生成最优推荐序列a_t^*。

$$a_t^* = \arg\max_{a_t} \Theta(a_t|U_t).$$

推荐目标可设为提高用户即时反馈或长期用户参与度，其中后者定义为未来K个时间周期内的登录次数。

$$e_t = \sum_{k=1}^{K} 1[\text{log-in}_{t+k}],$$

在强化学习优化指标Θ中，我们引入了MDP框架下的序列推荐。状态$s_t = U_t$表示历史交互，动作a_t由推荐策略π生成，每次推荐前会将用户感兴趣的项添加到末尾，形成
$s_t = s_{t-1} \oplus \{U_t - U_{t-1}\}$. 根据$\Theta$评估指标，奖励通常是函数形式。特别地，对于用户留存，奖励设置为与e_t相同。最后，用户的交互轨迹由s_t, a_t和r_t定义。

$$\tau = [s_1, a_1, r_1, \ldots, s_t, a_t, r_t].$$

Decision Transformer based Recommendation

- 嵌入模块
  $\tau_t' = [\hat{r}_1, s_1, a_1, \ldots, \hat{r}_t, s_t, a_t]$,
  $\hat{r}_t$表示累计回报，即回退到前往。T表示总推荐轮数。然后，嵌入模块将原始特征序列τ_t'转换为特征向量序列$\boldsymbol{\tau}_t'$，其中
  $\boldsymbol{\tau}_t' = [\hat{\boldsymbol{r}}_1, \boldsymbol{s}_1, \boldsymbol{a}_1, \ldots, \hat{\boldsymbol{r}}_t, \boldsymbol{s}_t, \boldsymbol{a}_t]$,
  其中$\hat{\boldsymbol{r}}_t$是奖励编码器模型（第几章）。特别地，由于奖励的重要性，我们设计了一个奖励提示模块来引导DT生成期望的推荐列表。
- 决策块： 决策块是模型的中心，它将稠密的上下文信息
  $\tau_t' - \{a_t\} = [\hat{r}_1, s_1, a_1, \ldots, \hat{r}_t, s_t]$
  转化为目标时间步的上下文特征$\boldsymbol{A}_t$用于生成所需的响应，其中$\hat{\boldsymbol{r}}_t$是生成的奖励提示。
- 在给定上下文信息$\boldsymbol{A}_t$的情况下，期望动作解码器能够生成与地面真实动作a_t匹配的一系列动作$\hat{a}_t$。
- 监督策略学习
  通过指定特定损失函数*来最小化生成动作$\hat{a}_t$和真实动作a_t之间的差异。因此，它将普通的RL任务转化为一个监督学习问题。通过在推理中指定最优奖励作为提示，DT旨在推荐能够最大化用户保留的商品。

Auto-Discretized Reward Prompt

为了使奖励间的顺序关系得以保留，我们提出了使用自动离散化方法*生成更有效的提示。这个方法包括数值离散化和使用MLP学习的元嵌入自动加权聚合。这种方法可以共享相似奖励的嵌入，并且可以根据奖励值将其转换为一个权重分数。该权重分数用于在批处理中对所有可学习嵌入$\boldsymbol{M}_b$进行加权聚合，其中b∈$1, B$。

给出加权分数z和奖励嵌入$\hat{r}$被设置为元嵌入的聚合，可以表示为：

$$\hat{r} = \sum_{i=1}^{B} \frac{\exp(z_i)}{\sum_{j=1}^{B} \exp(z_j)} \cdot e_i$$

$$\hat{r} = \sum_{b=1}^{B} z_b M_b,$$

r的值直接影响神经网络的输入，从而保证奖励间的偏序关系。$\hat{r}$相似则嵌入也相似，条件为神经网络平滑。

State-Action Encoder

$$\begin{aligned} H_n &= GRU_e(v_n, H_{n-1}) \\ a_t &= H_N, \end{aligned}$$

GRU_e为循环层，N为最大序列长度，H_n为隐藏状态。假设a_t的嵌入为最后一个时间步的隐藏状态，即a_t。

Transformer Decision Block

Transformer决策块可以通过学习用户的兴趣从交互轨迹中获取，从而充分利用Transformer和RL的优点。具体而言，单向Transformer层用于建模复杂特征交互，而跳过连接则用于防止过拟合，前馈神经层则用于线性映射特征。因此，可以将推荐决定的上下文信息表示为Transformer决策块的结果。

$$\widetilde{A} = FFN\left(MultiHeadAttention\left(\tau' - \{a_t\}\right)\right),$$

$\widetilde{A}$是动作嵌入矩阵，包含时间步数为T的行为向量$\tilde{a}_t$。$FFN(x)$是一个带有跳连接的前馈神经层，其输入为x与权重矩阵W_1, W_2和偏置项b_1, b_2的乘积，经过GELU激活函数输出。$MultiHeadAttention$展示了基于定义在 中的多头自注意力机制的有效性，该机制可以从不同的表示子空间学习信息并将其应用于推荐系统的大规模部署中。

Action Decoder

考虑到每一步用户与物品的交互信息，动作解码器旨在使用GRU来解码用户感兴趣的物品序列。当前的交互物品对于解码未知。我们只使用之前的交互和上下文信息来预测它。

$$\begin{aligned} \hat{v}_n &= v_n \oplus \tilde{a}_t \\ \tilde{v}_{n+1} &= GRU_d\left(\tilde{v}_n, \hat{v}_n\right), \end{aligned}$$

其中，v_n是第n个位置的嵌入向量，$\tilde{v}_{n+1}$是对第$n+1$个位置项的预测，$\oplus$表示连接操作。由于没有信息用于预测第一个位置，我们使用'bos'标记作为开始标记，并随机初始化$\tilde{v}_0$。 整个序列可以被解码为

$$\begin{aligned} \widetilde{V} &= \text{decoder}\left(\text{bos}, \hat{v}_1, \ldots, \hat{v}_n, \ldots, \hat{v}_{N-1}\right) \\ &= \left[\tilde{v}_1, \ldots, \tilde{v}_n, \ldots, \tilde{v}_N\right], \end{aligned}$$

在预测过程中，我们不知道解码器的序列长度，因此我们在每个序列的末尾添加一个'eos'标记作为结束标志，并将其填充为长度为N的零。解码器不会在预测'eos'时继续预测。

Contrastive Supervised Policy Learning

在决策变换器中，只考虑最大奖励进行推理，因为其设计旨在生成最高奖励的动作。然而，小奖励的样本可能未被充分利用（已在第节进行了验证）。为充分利用样本知识，我们提出了加权对比学习法，将小奖励动作视为负样本以避免推荐小奖励的动作。因此，我们的目标函数由两部分构成：CE损失和加权对比学习损失。

Weighted Contrastive Loss

$-\sum_i (r_i + c)\log(\hat{p}(y_i|x_i))$,

其中r_i是positive reward，c是cost term，y_i是true label，$\hat{p}(y_i|x_i)$是模型预测的概率。

$$\mathcal{L}_{\mathrm{CL}} = -\sum_{\widetilde{V}^- \in \Upsilon} \kappa\left(\widetilde{V}^-\right) f_s\left(\widetilde{V}, \widetilde{V}^-\right),$$

Υ为负样本集，$f_s(\cdot)$通过计算矩阵$\widetilde{V}$和$\widetilde{V}^-$中每行嵌入的点积平均来衡量两个序列间的相似度。$\kappa(\widetilde{V}^-)$基于负样本来设定权重超参数集，即权重与奖励值成反比。这样做的目的是为了防止过高的奖励导致用户留存率下降，所以我们希望在权值上与其他不同。此外，我们的模型还优化了使用原始交叉熵损失的DT（决策树）。

$$\begin{aligned} \widehat{Y} &= \psi\left(W_v \widetilde{V} + b_v\right) \\ \mathcal{L}_{CE} &= CE\left(\widehat{Y}, Y\right), \end{aligned}$$

标签矩阵Y中的每一行都包含one-hot标签，而预测标签分布则为$\widehat{Y}$。这里的CE表示交叉熵损失函数，ψ代表softmax函数。可学习的参数有W_v和b_v。

$\mathcal{L} = \mathcal{L}_{CE} + \beta\mathcal{L}_{CL}$, where β is a hyper-parameters.

Experiments

Datasets

我们对两个实际数据集，即爱奇艺用户留存数据（IQiYi）和ML-1m进行了实验。在表中，我们提供了这两个数据集的详细信息。具体来说， 具体来说, IQiYi dataset 我们收集了包含700万用户和4百万视频的交互记录的数据集。该数据集稀疏且存在噪音。为了保持一般性，我们在实验中仅考虑有20条及以上交互记录的用户。 我们未对不同类型交互进行区分，将其视为同等的交互。 用户保留度定义为用户在接下来一周内登录的平均天数，即K=7(WSDM杯2022年)的设置。 ML-1m数据集是常用基准数据集，在SRSs中常被使用，其记录了6,014名用户的对3,417部电影的评分。考虑到ML-1m记录跨越了较长的时间范围，我们以月为单位计算用户保留度。

Table 1: Statistics of datasets. UR is the short for average user retention.

Datasets	Users	Items	Interactions	UR	Density
IQiYi	3,000,000	4,000,000	71,046,026	4.80	5.8e-6
ML-1m	6,014	3,417	1,000,000	4.13	4.84%

Evaluation

Prediction Accuracy.

本文评估预测准确度时，采用了BLEU、ROUGE、NDCG和HR等四项常见的指标。

- BLEU评估了动态长度建议在序列到序列推荐中的精度，这是NLP领域常用的评价指标之一。
- ROUGE衡量的是动态长度推荐的召回率。
- HR\@K衡量真实物品在前K个推荐中的概率。
- NDCG\@K衡量前K个推荐结果的累积增益分数，考虑了位置对推荐结果的影响。其中CG表示项与真实值之间的相似性。

User Retention.

- MB-URS是基于模型的用户返回分数，用于评估推荐项序列的有效性。它直接返回用户的保留率。我们的模型是一个监督模型，用于预测特定状态动作对的奖励。我们在验证数据集的30%上进行独立测试，该模型通过最小化预测奖励与真实奖励之间的MSE损失来进行优化。
- 权重求和评估推荐有效性，权重基于真实用户保留分数。将样本分为8类，计算每个类的真实奖励与预测序列之间的BLEU-1得分作为相似度，计算SB-URS：
  $\text{SB-URS} = \sum_{k=0}^{K} s_k \cdot \left(g_k - \frac{K}{2}\right) \cdot N_k$,
  求解相似度与真实奖励的关系：$\min(s_k)$使得N_k尽可能大。
- 改进用户留存率衡量用户平均留存提升与离线数据平均留存的百分比，反映用户参与度。大比例表示推荐列表让更多用户活跃于系统。
- NRC衡量了推荐列表对于用户在系统中的参与度的影响。

Baselines

我们对比了多种推荐方法，包括：

- BERT4Rec使用双向Transformer学习序列信息，并通过掩码语言模型提高任务难度和训练能力。
- 使用一维Transformer捕捉用户顺序偏好。
- 该方法结合了Policy Gradient和Value-based DQN，并使用了AC框架。在推荐过程中，会同时推荐多个项目，并由模拟器自动学习最优策略以实现用户长期收益的最大化。
- TopK: 它采用一种基于策略的RL方法进行离线学习，并解决了分布不匹配与重要权重的问题。它假设一组非重复物品的奖励等于每个物品的奖励之和，并在每次时间步骤向用户推荐多个项目。在这篇论文中，我们将这种模型称为TopK。
- 无奖励的DT模型：DT4Rec-R

Hyperparameter Setting

在prompt生成器中，我们将B设为10，α设为0.1。在动作编码器和解码器中，我们设N为20，RNN层数为1。在决策变换器中，我们将最大轨迹长度设为从10到50的五个值，Transformer层数量为2，头部数量为8，嵌入大小为128。我们使用AdamW作为优化器，学习率为0.01。我们仅保存最近30次交互的物品作为状态。对于其他基线，我们将其超参数设置为作者推荐的最佳值或与我们的模型在同一范围内的值。每个模型的结果都在其最优超参数设置下报告。

RQ1: Overall Comparison

Prediction Accuracy

Table 2: Overall performance comparison in prediction accuracy. The best performance is marked in bold font.

Dataset	Model	Metric			
		BLEU↑	ROUGE↑	NDCG↑	HR↑
IQiYi	BERT4Rec	0.7964	0.7693	0.7384	0.7673
	SASRec	0.8009	0.7906	0.7827	0.7815
	DT4Rec	**0.8249***	**0.8172***	**0.8139***	**0.8044***
ML-1m	BERT4Rec	0.3817	0.3806	0.5286	0.2769
	SASRec	0.4052	0.3983	0.5409	0.3123
	DT4Rec	**0.4331***	**0.4185***	**0.5679***	**0.3342***

- 我们的模型在多个数据集上的表现优于所有基准。这证明了我们模型利用即时用户反馈进行优化的有效性。与现有的方法不同，我们的模型使用强化学习模拟每个时间步骤的用户奖励和状态。这使我们的模型能动态地模拟用户的兴趣特性，以满足用户需求的变化。因此，在预测准确性任务中，我们的模型表现出色。

知乎

User Retention

Table 3: Overall performance comparison in user retention. The best performance is marked in bold font.

Dataset	Model	Metric			
		MB-URS↑	SB-URS↑	IUR↑	NRC↓
IQiYi	DT4Rec-R	5.16	52131	7.5%	2.6%
	TopK	5.41	61045	12.7%	2.0%
	LIRD	5.63	63572	17.3%	1.9%
	DT4Rec	**6.05***	**72270***	**26.0%***	**1.4%***
ML-1m	DT4Rec-R	5.42	13050	31.2%	9.7%
	TopK	5.71	13964	38.3%	6.9%
	LIRD	5.86	14627	41.9%	4.8%
	DT4Rec	**5.93***	**15562***	**43.6%***	**3.6%***

- 本文比较了多个模型在不同基准上的表现，结果表明在MB-URS和SB-URS上，本模型优于其他基线；在IQiYi和ML-1m上，DT4Rec优于最佳的LIRD基线。
- 该文指出传统离线方法如TopK和LIRD存在 死贸易 问题及折扣因子导致的短视性问题，其性能不及所提出的方法。
- 策略梯度TopK模型易陷入次优解，确定合适学习率有难度；SRS动作空间大，LIRD估Q值更难。
- 我们的模型使用序列建模来学习动作-状态的奖励映射，无需bootstrap和"死贸易"问题。与传统RL策略不同，我们不估计Q值，解决了SRS中动作空间大问题。与DT4Rec-R对比显示奖励对于学习用户长期偏好的重要性。

RQ2: Effectiveness of Auto-Discretized Reward Prompt

Table 4: Ablation study on IQiYi dataset.

Architecture	Metric		
	SB-URS	MB-URS	IUR
DT4Rec	72270	6.05	26.0%
w/o contrastive	65175 (-9.82%)	5.68 (-6.11%)	18.3% (-29.6%)
w/o weight	69316 (-4.09%)	5.79 (-4.30%)	20.6% (-20.8%)
w/o auto-dis	70953 (-1.82%)	5.96 (-1.49%)	24.2% (-6.9%)

比较DT4Rec和DT4Rec-R结果，我们证明了奖励指导在模型训练中的重要性。为了衡量奖励顺序有效性，我们通过实验进行了改进。具体地，我们将DT4Rec与使用单层前馈神经网络[+]实现的"w/o auto-dis"（在表格中的位置）进行比较。结果显示，自动离散化奖励提示是有效的，没有它，MB-URS和SB-URS的性能分别下降了1.82%和1.49%，这说明自动离散化的奖励提示可以保持生成嵌入对应的奖励值之间部分顺序关系。

RQ3: Validity of Contrastive Supervised Policy Learning

我们的研究揭示了DT模型的OOD问题，并通过实验验证了改进的有效性。具体来说，我们评估了无对比学习的DT模型，移除了奖励小于4的样本。结果表明模型几乎不使用大型奖励样本的知识，支持我们对OOD问题的主张。接下来，我们验证了权重对比学习方法的有效性，通过对比DT4Rec与权重对比学习方法的模型，表1的性能下降证实了我们方法的有效性。最后，我们比较了

项序列。

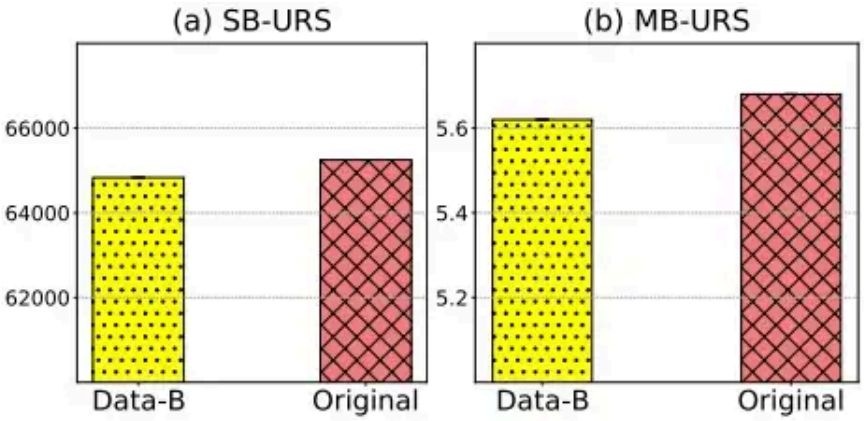


Figure 2: Comparison between the model trained on the original IQiYi dataset and the date removing the smaller reward parts, i.e., the Data-B.

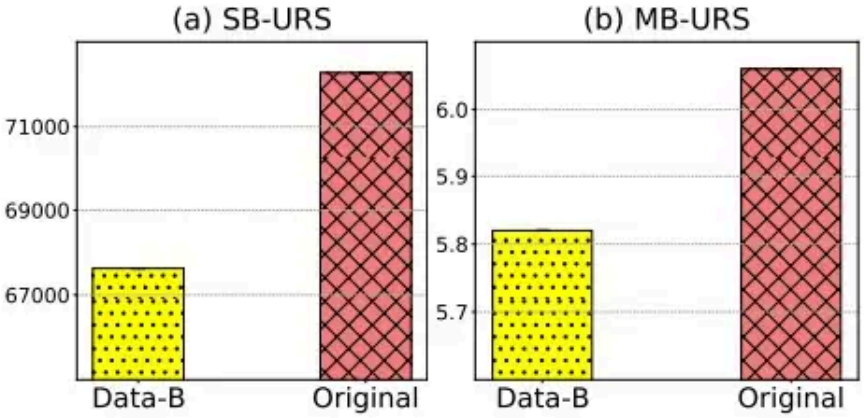


Figure 3: Comparison between the DT4Rec trained on the original IQiYi dataset and the date removing the smaller reward, i.e., the Data-B.

RQ4: Evaluation Efficiency

我们将与@wu2017returning提出的评估方法比较并证明MB-URS和SB-URS的有效性，我们称其为'RiB'。我们将测试集*分为5个子集，并使用这5个子集上的性能方差作为评估方法稳定性的指标。为了公平比较，我们用ASB-URS替换SB-URS，ASB-URS根据式中的N_k对用户返回分数进行平均相似度度量。图展示了在IQiYi和ML-1m两个数据集上评估方法的方差。结果显示，我们的评估方法在两个数据集上的方差均较小，从而证实了我们提出的方法的有效性。

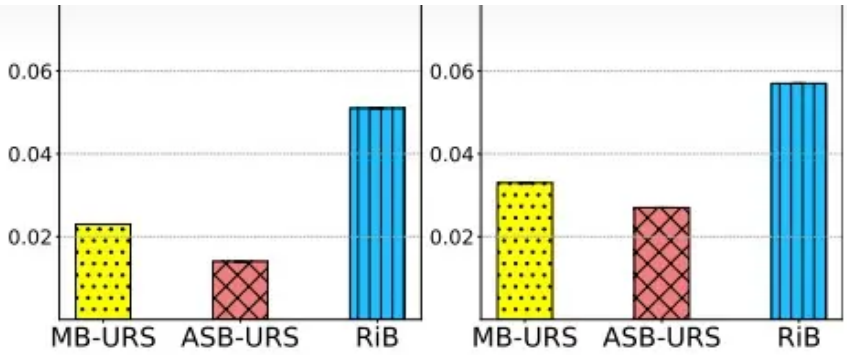


Figure 4: Comparison of different evaluation methods' variance on IQiYi and ML-1m datasets.

RQ5: Model Generalization Analysis

在新数据集上，我们的模型超越TopK，表明方法具有强泛化能力；使用仅10%高奖励样本，DT4Rec的SB-URS高于20%，证明方法有效性且能提升用户保留率；当高奖励样本较少时，策略评估会变得不准确和困难；但我们的模型通过监督方式训练，避免策略评估高方差并提升测试性能。

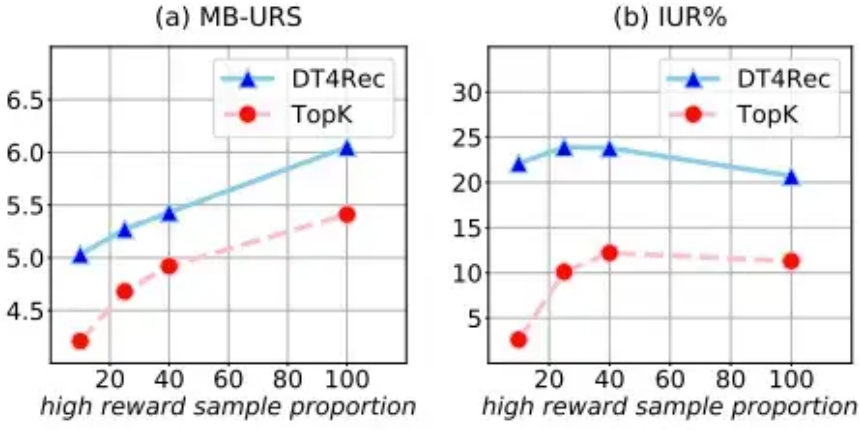


Figure 5: Behavior cloning analysis on IQiYi dataset.

Conclusion

本文提出了一种新型的基于强化学习的序列推荐系统------DT4Rec。该系统通过将RL问题视为自回归问题，以避免不稳定性和无界变异性。为此，我们设计了一种自动离散化奖励提示，以便模拟奖励的数值并引导模型进行长期用户参与的训练。另外，我们还提出了对比性监督策略学习，以减少决策变换器在推理和训练中的不一致。为了评估我们的模型，我们提出了MB-URS和SB-URS两种稳定度指标，这两种指标已被证实比现有方法更稳定。我们在基准数据集上进行了广泛实验，以验证所提出方法的有效性。

原文《User Retention-oriented Recommendation with Decision Transformer》

编辑于 2024-02-19 18:16 · IP 属地北京

百度 用户留存 Transformer