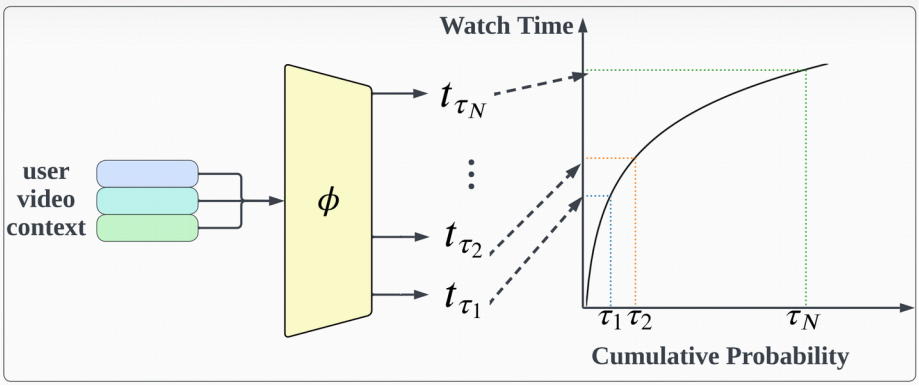


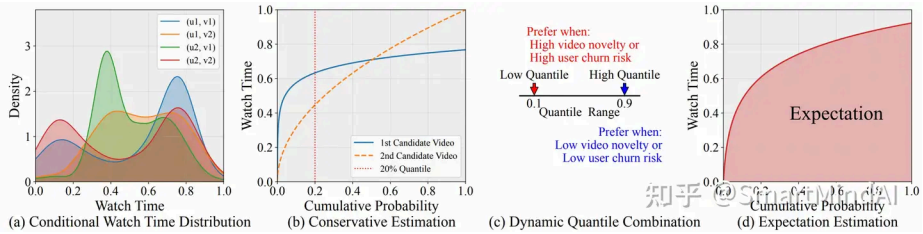
快手2024：CQE —— 短视频推荐中的创新时长分位数预测框架

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

在线视频平台的快速增长已经彻底改变了用户消费数字内容的方式，与传统推荐问题（例如电子商务和新闻推荐）不同，衡量用户兴趣和参与度的关键指标是观看时间，它全面反映了用户的选择和投入程度。因此，准确预测观看时间对于优化推荐策略和提升用户体验至关重要。然而，预测观看时间仍然是一个具有挑战性的任务，因为用户行为的不确定性以及行为的异质性。在现实场景中，很难在相同条件下为同一用户-视频对获取多次观察结果，因为用户很少在完全相同的情境下多次观看同一视频。这一限制使得我们无法直接从数据中估计真实的观看时间分布。

现有方法往往侧重于预测观看时间的单一估计值（例如均值或中位数），忽略了观看时间分布的复杂性和多样性。这些方法无法充分捕捉不同用户-视频对之间的行为差异，导致推荐性能受限。使用单一平均值来描述这些复杂分布模式的不足凸显了我们全面建模观看时间的条件分布的需求。



为了应对这些挑战，我们提出了条件分位数估计（CQE）框架，该框架学习在用户-视频配对及其相关上下文给定时预测观看时间的条件分布。如图所示，CQE利用分位数回归技术来估计观看时间分布的多个分位数⁺，提供了用户参与模式的全面视图。如图（a）所示，CQE模型为不同的用户-视频配对预测的观看时间条件分布显示出在形状、峰值位置和分散程度上的显著多样性。这种异质性反映了用户在不同上下文中的偏好和参与度的内在不确定性。

条件观看时间分布的建模对于理解用户参与模式和设计有效的推荐策略至关重要。通过考虑观看时间分布的详细特征，我们可以深入了解不同用户群体的多样化观看行为。这种精细的理解使我们能够为各种场景和用户偏好定制推荐策略。 基于CQE模型，我们设计了三种主要的推荐策略。保守估计策略（图（b））在预期观看时间相似的情况下，优先选择具有更高下分位数的视频，以降低用户脱机的风险。动态量化组合策略（图（c））根据用户流失风险和视频新颖性等因素调整选择的分位数。对于高流失风险的用户或新颖视频，它赋予较低分位数更多的权重，确保满意的体验；对于低流失风险的用户或熟悉视频，它赋予较高分位数更多的权重，可能提供更具吸引力的推荐。最后，期望估计策略（图（d））提供了一种全局优化视角，旨在考虑整个观看时间分布，最大化整体用户参与度。

Method

Problem Formulation

知乎

$\mathbf{x} \in \mathbb{R}^n$，这个向量包含了用户特征、视频属性、上下文信息以及历史交互数据。

设W为一个随机变量，代表观看时间。我们的目标是根据输入特征估计W的条件概率分布：

$$P(W|\mathbf{x}) = P(W|\psi(u,v,c))$$

不同于传统方法专注于估算单一数值（例如，预期观看时间$\mathbb{E}[W|\mathbf{x}]$），我们的目标是描述整个条件分布。这使我们能够捕捉在用户参与中内在的不确定性与变异性，提供更全面的可能的用户行为理解，从而更精确地刻画整个条件分布，提供更深入的洞察力，以增强对潜在用户行为的理解。

Conditional Quantile Estimation Model

为了全面捕捉观看时间的分布情况，我们提出了一种条件分位数估计（CQE）模型。如图所示的左侧部分，这种方法允许我们同时估计观看时间分布的多个分位数，从而提供对用户参与度更全面的洞察。 设 $\{\tau_1, \tau_2, \ldots, \tau_N\}$ 是N个预定义的分位数级别集合，其中$\tau_i = \frac{i}{N+1}$。

我们的CQE模型旨在给定输入特征$\mathbf{x}$的情况下，估计每个分位数级别的相应观看时间值 $\{t_{\tau_1}, t_{\tau_2}, \ldots, t_{\tau_N}\}$ 以提供对用户参与度更全面的了解：

$$\{t_{\tau_1}, t_{\tau_2}, \ldots, t_{\tau_N}\} = \phi(\mathbf{x}; \theta)$$

其中$\phi(\cdot)$是由参数θ参数化的神经网络。为了确保分位数估计的单调性，我们采用一下架构：

$$\begin{aligned}
\mathbf{h} &= f(\mathbf{x}; \theta_f) \\
\mathbf{d} &= \mathrm{ReLU}(g(\mathbf{h}; \theta_g)) \\
t_{\tau_i} &= \sum_{j=1}^{i} d_j, \quad \text{for } i = 1, \ldots, N
\end{aligned}$$

这里的 $f(\cdot)$ 和 $g(\cdot)$ 是神经网络的组成部分 $\mathbf{h}$ 是一个中间隐藏表示$\mathbf{d}$是一个非负元素的矢量。最终的分位数估计t_{τ_i}通过累积求和获得，则排序： $t_{\tau_1} \leq t_{\tau_2} \leq \ldots \leq t_{\tau_N}$

这种形式让我们的模型能够捕捉输入特征与观看时间的分位数之间的复杂的非线性关联，同时保持分位数函数的单调性特性，从而能够捕捉复杂非线性关系，同时维持必要的单调性属性。

知乎

Function `Train`(D, N, α):
  Input: Training data $D = \{(\mathbf{x}_i, y_i)\}$, number of quantiles N, learning rate α
  Output: Trained CQE model ϕ
  Initialize model parameters θ;
  Define quantile levels $\tau_1, ..., \tau_N$;
  for *each epoch* **do**
    for *each mini-batch* $B \subset D$ **do**
      Compute quantile estimates $\{t_{\tau_1}, ..., t_{\tau_N}\} = \phi(\mathbf{x}; \theta)$;
      Compute loss L_{QR} using Eq. 5 ;
      Update $\theta \leftarrow \theta - \alpha \nabla L_{QR}$;
    end
  end
  return ϕ
Function `Infer`$(\phi, \mathbf{x}, S)$:
  Input: Trained CQE model ϕ, feature vector $\mathbf{x}$, inference strategy S
  Output: Predicted watch time $\hat{y}$
  Compute quantile estimates $\{t_{\tau_1}, ..., t_{\tau_N}\} = \phi(\mathbf{x})$;
  Apply inference strategy S to $\{t_{\tau_1}, ..., t_{\tau_N}\}$ to get $\hat{y}$;
  return $\hat{y}$

计算复杂性方面，CQE模型与传统的点估计方法相当，仅因为需要估计多个分位数而略有轻微增加。在大规模推荐系统中，用户和项目数量经常达到数亿甚至数十亿级别。这些用户和项目通常通过各自的ID检索到高维嵌入表示。相比之下，为了有效估计所需的分位数，通常大约需要100个。因此，CQE的额外计算成本相对于处理大量特征所需的大量计算来说，微不足道。

Training Objective

为了使我们的CQE模型得到有效训练，我们采用了针对量化回归任务的pinball损失函数。对于单一分位数阈值τ，pinball损失函数的定义如下：

$$\mathcal{L}_\tau(y, t_\tau) = \begin{cases} \tau(y - t_\tau) & \text{if } y \geq t_\tau \\ (1 - \tau)(t_\tau - y) & \text{otherwise} \end{cases}$$

其中y代表实际观看时间，而t_τ是预测的第τ个分位数。 如图所示，指针损失函数有其几个关键特性：

1. 非对称性：损失的分布围绕真实值y是非对称的，非对称程度由τ决定。
2. 线性损失与预测值和实际值之间的距离成线性关系，其斜率在y轴的正方向和负方向上有所不同。
3. 分位数特定罚则：当τ大于0.5时，过估计的惩罚比过低估的惩罚更重；当τ小于0.5时，则反之。

这些特性使得pinball损失特别适合于分位数回归。对于我们的多分位数模型，我们对所有分位数水平上的pinball损失进行了聚合，以获得全面的估计结果。

$$\mathcal{L}_{\text{QR}} = \sum_{i=1}^{N} \mathcal{L}_{\tau_i}(y, t_{\tau_i})$$

这个聚合损失函数促使模型在整个分布范围内，为每个用户-视频组合提供精确的分位数预测，捕捉每对用户-视频可能观看时间的全范围。

Inference Strategies

于特定推荐情境。

Conservative Estimation

在成本有限的情况下，如果用户满意度是首要考虑因素，我们采用保守估计（CSE）策略。此策略侧重于观看时间分布的下位分位数，以确保提供满意的用户体验。

如在图（b）所示，当预期观看时间相似时，我们优先选择具有更高下位分位数的视频以提升用户满意度。此策略有助于降低用户脱机的风险。形式上，我们选择一个下位分位数τ_{low}（例如$\tau_{\text{low}} = 0.25$）来作为决策依据，并使用其对应的观看时间预测值⁺。

例如$\tau_{\text{low}} = 0.25$的情况下，我们选择的策略是通过使用此值对应的观看时间预测，以确保在成本高昂的环境下，用户能够获得满意的体验： $\hat{y}_{\text{CSE}} = t_{\tau_{\text{low}}}$

这个策略有助于减少推荐过于乐观导致的用户失望风险，因为实际观看时间很可能超过保守估计。

Dynamic Quantile Combination

为了适应不同用户偏好和内容特性变化，我们提出了一种动态分位数组合（DQC）策略。这种方法根据上下文因素结合不同分位数的预测。如图（c）所示，DQC策略动态适应选择的分位数，根据用户流失风险和视频的新颖性进行调整。对于高流失风险的用户或新颖的视频，系统会给予低分位数更多的权重，以确保提供满意的体验，确保用户获得良好的感受；而对于低流失风险的用户或熟悉的视频，则给予高分位数更多的权重，可能提供更具吸引力的推荐。这种动态方法允许系统在用户当前状态和内容熟悉度的基础上平衡安全推荐与可能更具奖励性的推荐。

令$k \in [0, 1]$为一个基于上下文的融合参数。我们计算最终预测为：
$\hat{y}_{\text{DQC}} = k \cdot t_{\tau_{\text{low}}} + (1 - k) \cdot t_{\tau_{\text{high}}}$

其中，保守预测由$t_{\tau_{\text{low}}}$表示，而乐观预测由$t_{\tau_{\text{high}}}$表示。

融合参数k可以根据用户风险偏好、视频的新颖性或平台目标等因素进行调整。例如，对于新用户或新颖内容，我们可能会使用更高的k，倾向于保守估计；而对于已有用户或熟悉的内容类型，则可能使用更低的k。

Conditional Expectation

在我们旨在优化预期观看时间的场景中，我们采用条件期望⁺策略。这种方法通过在预测的分位数之间进行插值⁺来估计平均观看时间。 如图所示，条件期望估计（CDE）策略提供了一个全局优化视角，旨在通过考虑整个观看时间分布来最大化整体用户参与度。如图的左侧部分所示，这些输出观看时间值体现了观看时间的分布情况。在任何两个连续的分位数之间，我们面临的问题是没有输出值 $\tau \in (\tau_i, \tau_{i+1})$

为了克服这一信息缺失的问题，我们使用插值方法来近似条件分布。我们采用线性插值⁺在连续的边界点之间，因此在τ_i和τ_{i+1}之间的预期观看时间变为 $(t_{\tau_i} + t_{\tau_{i+1}})/2(N+1)$

对于两个边界点，我们假设$t_0 = t_{\tau_1}$和$t_1 = t_{\tau_N}$。然后，我们可以近似计算整体观看时间的期望值为：

$$\begin{aligned}\hat{y}_{\text{CDE}} =& \frac{1}{2(N+1)}[(t_{\tau_1} + t_{\tau_1}) + (t_{\tau_1} + t_{\tau_2}) + (t_{\tau_2} + t_{\tau_3}) \dots \\ &+ (t_{\tau_{N-2}} + t_{\tau_{N-1}}) + (t_{\tau_{N-1}} + t_{\tau_N}) + (t_{\tau_N} + t_{\tau_N})] \\ =& \frac{1}{N+1} \sum_{i=1}^{N} t_{\tau_i} + \frac{t_{\tau_1} + t_{\tau_N}}{2(N+1)}\end{aligned}$$

理论上，这种期望提供了最精确的一般预测，并且当$N \to \infty$时，它将实现最佳预测。实际上，我们将在实验分析部分⁺验证其优越性。然而，这种策略可能不适合那些用户无法容忍不良推荐或推荐系统需要动态调控的场景。

为了验证条件分位数估计（CQE）框架在现实世界中的效果，我们在一个拥有众多数亿用户短视频平台进行了大量在线A/B测试。这些实验在拥有大量用户的实际环境中，使我们能够评估CQE框架的三种策略的有效性。

Experiment Setup

用户被随机分配到控制组和实验组，每天用户流量的最小百分比为10%，分配给实验组以确保统计显著性。每项在线A/B测试持续超过一周，提供了充足的时间进行数据收集⁺和可靠的结果分析。推荐系统采用两阶段过程：候选检索随后排序。在排序阶段，我们整合了CQE模型，用于预测观看时间，这是推荐过程中的关键组成部分。

我们使用了四个关键指标来评估推荐系统的性能：

- 每个用户的平均观看时间：这个关键指标直接通过量化用户观看推荐视频的平均时间来评估用户参与度。
- 总播放次数：这个指标表示了所有用户累计播放推荐内容的数量，反映了用户与推荐内容的交互频率。
- 用户活跃日数: 这个指标衡量用户与平台互动的天数，反映用户留存率。
- 每日活跃用户量：这个指标值反映了与平台互动的唯一用户量，体现系统保持和发展用户基础的能力。

Experiment Results

Table 1: Conservative Estimation's enhancements in active days and active users. A boldface means a statistically significant result (p-value < 5%).

Active dyas	Active Users
+0.033%	**+0.031%**

Table 2: Comparison between CQE Strategies and baseline online. A boldface means a statistically significant result (p-value < 5%).

Strategy	Watch Time	Play Count
Conservative Estimation	+0.008%	**+0.346%**
Dynamic Quantile Combination	**+0.106%**	**+0.177%**
Conditional Expectation	**+0.165%**	-0.088%

- CSE 提供了一种均衡的方法，增强所有指标，特别是在提升平台互动的广泛性和用户留存率⁺方面，表现出色。
- DQC 提供了一种折衷方案，同时既加深了深度，又扩大了内容的范围，从而提升了参与度和内容多样性。
- CDE 在深化用户对单个内容件的参与度方面表现出色。

这些策略的选择取决于特定平台的目标，例如，为了优先考虑深度参与、广泛互动、用户留存或内容多样性。此外，这些策略可能根据用户群体或内容类型进行组合或动态应用，以优化整个系统的性能。

Offline Experiments (RQ2 and RQ3)

同提供了评价CQE框架在推荐系统中有效性的一个全面视角。 预测观看时长直接捕捉了用户对内容的参与时长，这是用户参与度的关键指标。然而，仅预测观看时长可能无法完全捕捉用户兴趣。因此，我们引入了预测用户兴趣任务，它将观看时长与视频时长相结合，提供了一个更细致的用户兴趣衡量标准。这两个任务相辅相成：预测观看时长提供了直接的行为预测，而预测用户兴趣帮助我们理解这些行为背后的动机。

Experimental Results

CQE~CDE~与其他方法的比较： 我们在观看时间预测任务中比较了不同方法的表现，并将结果列在表中。

Methods	Kuaishow		CIKM16	
	MAE	XAUC	MAE	XAUC
WLR	6.047	0.525	0.998	0.672
D2Q	5.426	0.565	0.899	0.661
OR	5.321	0.558	0.918	0.664
TPM	4.741	0.599	0.884	0.676
CQE_{CDE}(Ours)	**4.437**	**0.610**	**0.823**	**0.694**

在观看时间预测任务中，TPM和CQE~CDE~在MAE和XAUC指标上都超越了其他方法，这凸显了将不确定性纳入模型的重要性。此外，我们的方法在两个指标上都比TPM表现出更优性能，这强调了采用分位数建模技术的优势。 另外，MAE和XAUC指标的一致性行为也验证了观看时间估计可以作为排名指标的可行性。

对于用户兴趣预测任务，我们比较了基于不同主干模型（DeepFM和AutoInt）的多个框架，并在表中呈现了结果。在所有情况下，我们的提议CQE~CDE~始终超越替代方案，这表明CQE~CDE~具有稳健性和有效性。 关于优化框架，CE通常比MSE表现更好，这证明了将序列分类信息作为指导的正确性。 在用户兴趣指标优化方面，CQE~CDE~可以在所有设计中提升性能，这意味着提议的框架适用于不同的标签设置，具有普适性。 这证明了CQE~CDE~在不同任务和框架下的广泛适用性和优越性，强调了其在处理观看时间和用户兴趣预测任务中的重要性和有效性。

原文《Conditional Quantile Estimation for Uncertain Watch Time in Short-Video Recommendation》

发布于 2024-08-09 14:51 · IP 属地北京

快手　排序算法　工业级推荐系统

知乎

点t_tau？ 2. 我看论文里有一个t_tau与tau的累积函数分布映射图，你们是对于N个输出t_tau都对应一个不同的tau，这样每个输出t_tau与各自相应的tau计算pinball损失然后加和得到最终的损失函数？

09-12 · 北京 回复 喜欢

展开其他 1 条回复 ❯

太家

在t_i和t_{i+1}之间的预期观看时间变为 (t_{i} + t_{i+1}) / 2(N+1) 这一步是怎么计算的？

08-27 · 上海 回复 喜欢

justopit ‣ **太家**

看文章的顶部图片。曲线变直线。曲线面积然后变成很多个梯形的面积。

08-28 · 北京 回复 喜欢

太家 ‣ **justopit**

是假设任意两个分位点的距离都一样了？

08-28 · 上海 回复 喜欢

展开其他 1 条回复 ❯

理性发言，友善互动