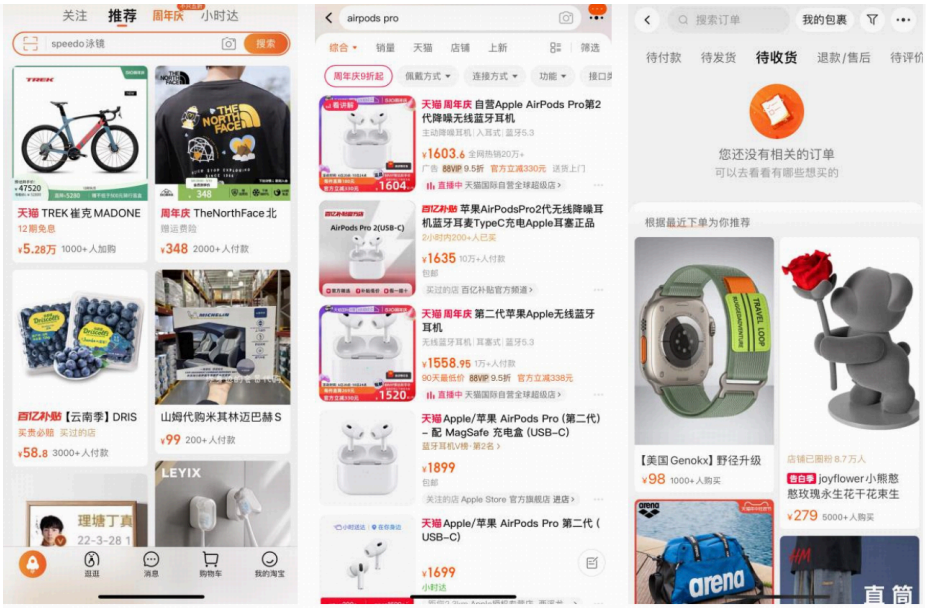
知乎



2024阿里：淘宝推荐引擎提速 —— 探索多场景高效召回框架

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

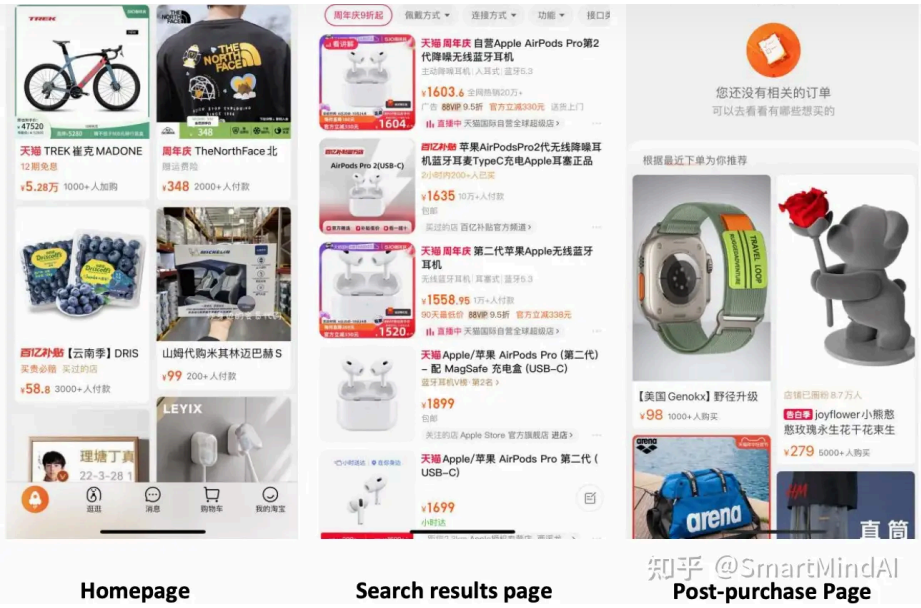
Introduction & Related Work

随着电子商务平台的快速发展，推荐系统已成为个性化内容推荐、优化用户体验和增加业务收入的关键。在大多数工业场景中，推荐系统面临大量候选集和严格的延迟挑战，因此包含了多个阶段（例如召回、粗排和精排）来平衡效率和准确性。

召回阶段旨在快速从庞大的候选池中筛选出相关内容，优先考虑低计算复杂度以实现可扩展性。这为后续的粗排和精排阶段奠定了基础，在这些阶段，基于初步筛选的结果，进行更细致的打分。

现有的推荐系统召回算法主要分为两大类：协同过滤方法和深度学习方法。协同过滤方法通过利用用户和内容的历史交互或潜在因素来理解用户偏好，但它们往往面临数据稀疏性的问题。另一方面，深度学习方法通常采用端到端的神经网络模型来推断用户对每个内容的兴趣。虽然这些方法效果不错，但它们依赖于简单的模型结构，可能无法完全理解用户偏好的复杂性。

知乎

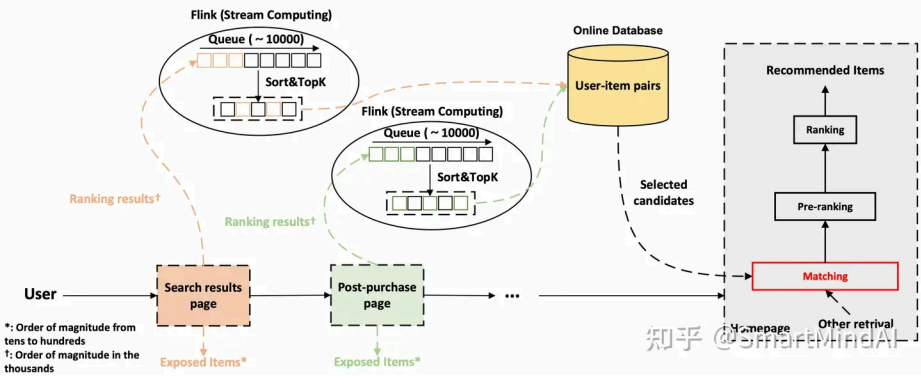


最近，多场景建模的概念在推荐系统中被引入。以淘宝为例，图展示了包括首页、搜索结果页面和购后页面在内的各种个性化推荐场景。在一种场景中观察到的用户行为可能有助于在另一种场景中改进推荐。然而，当前大多数多场景建模方法都采用复杂的框架，导致计算成本⁺显著增加，使得它们在召回阶段变得不切实际。实际上，在每个场景中，精排阶段会评估成千上万个候选内容，但受限于页面容量，通常只有数十个内容最终被展示给用户。因此，大量在精排阶段得分高的内容，尽管它们有潜在的高价值，但用户实际上无法看到它们。

为了应对上述挑战，本文提出了一种名为"多场景近线检索"（MNR）的创新方法，用于召回阶段。我们通过综合各种场景的排序结果和实时打分模块的优化，提高了推荐的效率和准确性。 在淘宝这样的中国领先电子商务平台上实施，我们的方法在实际环境中取得了显著的性能提升, 本文主要贡献如下：

1. 提出了一种多场景近线检索方法⁺，用于召回阶段，通过综合各种场景的排序结果和实时打分模块的优化，提高了推荐的效率和准确性。
2. 实现了在淘宝平台上的应用，验证了方法的有效性和实用性，展示了在实际推荐系统中的潜力和价值。
3. 解决了多场景建模所导致的计算成本高和实时响应延迟的问题，提供了一种更高效、更快速的召回解决方案。
4. 提供了一种整合多场景信息的策略，有助于提高个性化推荐的质量，满足用户在不同场景下的需求。

Method



Problem Formalization

工业推荐系统中召回阶段的目标是在浩瀚的商品池 $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{|E|}\}$ 中，为每一位用户$\mathbf{u}$进行筛选。这一过程旨在将商品缩小到一个远小于商品池的个性化子集$\mathcal{E}_u$，保证子集中每一件商品都与用户的兴趣相召回：

知乎

Multi-Scenario Nearline Retrieval

我们利用其他场景的排序结果来增强当前场景召回阶段。具体来说，用户可能在平台上参与多个场景（s1，s2，..，sl）。在每个场景s中，精排阶段会根据排序模型提供打分内容：

$$\mathbf{R}_s^{t1} = (\mathbf{e}_1^{t1}, \mathbf{e}_2^{t1}, \cdots, \mathbf{e}_n^{t1}),$$
$$\mathbf{R}_s^{t2} = (\mathbf{e}_1^{t2}, \mathbf{e}_2^{t2}, \cdots, \mathbf{e}_n^{t2}),$$

其中$t1$和$t2$表示用户的不同访问时间n表示精排阶段的打分候选集的数量。对于每个场景向量$\mathbf{s}$，MNR的目标可以表述为：

$$\mathcal{E}_s = \mathcal{F}_{NMR}(\mathbf{R}_s^{t1}, \mathbf{R}_s^{t2}, \cdots, \mathbf{R}_s^{tN})$$

其中$\mathcal{F}_{NMR}$表示一个函数，用于场景感知下从各种细化排序候选集中识别当前场景下最重要的前m个关键内容。

$\mathcal{E}_s = \{e_1, e_2, \cdots, e_m\}$ 是基于场景$\mathbf{s}$的最终选择，这些内容是通过在精排阶段根据细化排序候选集的结果进行检索后得到的。

Scenario-aware Ranking Results Collecting

在每个场景中，通过粗排阶段筛选的内容在精排阶段获得打分。尽管在精排阶段为所有内容分配了打分，但这些打分并不一定符合用户的所有偏好。因此，我们首先从每次访问中获取精排阶段的排序结果：

$$\mathcal{R}_s^{t1} = Truncate_s(\mathbf{R}_s^{t1})$$

然后将不同访问时刻的多个排序列表（即$\mathcal{R}_s^{t1}$、$\mathcal{R}_s^{t2}$、$\cdots$、$\mathcal{R}_s^{tl}$）合并到一个统一的集合中是不平等的。鉴于在线存储是制约条件，对于每个场景$\mathbf{s}$，我们使用优先队列*来记录每个用户的历史排序。

$$\mathcal{C}_s = Queue(\mathcal{R}_s^{t1}, \mathcal{R}_s^{t2}, \cdots, \mathcal{R}_s^{tl})$$

在这里，我们采用先进先出（FIFO）策略来保留用户的最新偏好。这样，我们可以在不增加额外计算开销的情况下，保留用户在其他场景中的兴趣。

Streaming Candidate Scoring

在汇总了场景$\mathbf{s}$中用户偏好的综合排序结果$\mathcal{C}s$后，关键在于细化并管理检索集的大小，以优化用户体验。我们需要关注两个方面：初始排序顺序和用户交互的时间。精排阶段通常使用复杂模型，包含强大的特征交互，这意味着得分较高的内容更符合用户的兴趣。然而，由于用户品味可能会变化，接近当前时刻的排序往往与当前偏好有更好的相关性。为了在排序和访问时间之间找到平衡，我们可以定义打分函数*如下：

$$finalScore = (\frac{\alpha}{\alpha + rank\_index}) \times (\frac{\beta}{\beta + time\_index}),$$

其中$rank_index$表示内容在初始排序结果中的倒序位置，即从高到低排列。$time_index$指的是用户访问的序列，最近一次访问的索引为0。α和β作为平衡超参数*，分别调整内容排序和访问时间的权重。通过增加α的值，打分公式会更重视用户的访问时间。相反，提升β的值，焦点将转移到内容原始排序分数上。根据等式，通过从$\mathcal{C}s$中选择前k个内容，最终得到检索结果$\mathcal{E}_s$。

Implementation Details

本文所提出的框架使用Flink平台订阅来自不同场景的排序日志（即排序过程中候选内容的得分记录）。进行队列管理、聚合和前k计算等接近实时的操作。为了促进类别多样性，我们在生成TopK结果时，内容被分散到不同的类别中，为每个类别设置了预定义的内容数量上限。

Experiment

知乎

Table 1: The performance of evaluated methods on Taoba Homepage.*

	Truncation	PVR	Hitrate	CTCVR
$Online_{ALL}$	/	100%	0.62%	base
$SWING$	/	-	0.43%	+21.2%
$MIND$	/	-	1.21%	+7.5%
$Glare$	/	-	0.66%	+0.2%
$MNR_{New-Detail}$	500	15.54%	1.94%	+35.0%
$MNR_{Post-Purchase}$	500	15.29%	1.08%	+11.3%
$MNR_{MainSearch}$	500	6.99%	2.02%	+57.0%
$MNR_{In-shop}$	20	0.71%	0.90%	-33.0%
$MNR_{PhotoSearch}$	20	0.36%	1.66%	+103.7%

表展示了不同方法相关的性能结果，其中$Online_{ALL}$表示召回阶段的总体性能。表中的亮点是$MNR_{MiniDetail}$和$MNR_{MainSearch}$在提升性能方面表现出的稳健性，分别在CTCVR指标上实现了35.0%和57.0%的改进。这些指标突出了MNR在跨场景洞察策略上的效率，提升了其在预测和召回用户偏好的精确度。

在Hitrate指标上 $MNR_{MiniDetail}$ 和 $MNR_{MainSearch}$ 实现了显著的提升，分别达到 1.94%和2.02%，远超基础的0.62%的$Online_{ALL}$。此外，表中的结果进一步表明，MNR在超越其他竞争方法如SWING、MIND和Glare方面也表现出色。 另外，像$MNR_{In-Shop}$和$MNR_{PhotoSearch}$这样的场景由于缺乏足够的精细排序分数而面临挑战，因此采用了较低的截断阈值。这导致进入召回池的内容数量减少，从而降低了PVR。$MNR_{In-Shop}$的不佳表现可以归因于店内场景的特性，店内场景的特性是导致用户仅限于看到该商店内的几到大约一百件商品。因此，MNR框架捕获的召回内容大多是尾部商品。

Online performance.

我们通过在一个月内部署MNR，并利用淘宝首页的多组精选场景数据处理实际流量，进行了在线实验。淘宝首页的实际流量结果表明，相对于基准，我们提出的方法在交易量上实现了持续稳定的5%增长。此外，我们提出的算法具有高效的计算性能和低能耗，使其成为推荐系统中实时个性化推荐场景的有潜力的解决方案。

原文《Simple but Efficient: A Multi-Scenario Nearline Retrieval Framework for Recommendation on Taobao》

编辑于 2024-09-04 11:13 · IP 属地北京

推荐系统　工业级推荐系统　阿里巴巴最新技术

理性发言，友善互动

2 条评论　默认　最新

水中木

贡献1召回阶段怎们实现的好像没说啊，一直在说怎么打分🤔

08-29 · 山东　回复　喜欢

David

定义个场景相似度矩阵，按相似度融合呗

09-12 · 江苏　回复　喜欢