课题组已核实初稿及修改稿中所有建议，确认已无其他技术细节未补充并按以下文本提交专利局，接受专利局的审查决定。
项目负责人签字
日期

说 明 书

端上实时的视频流物品优化推荐实现方法

【组合多个现有技术于一体的发明创造或将其实现的过程，当任一部分均按现有方式工作且全文未揭示任何从未被公开的技术手段时，专利局将评价该申请不具有创造性。简单替换、拼凑现有技术或者实质上由不同发明创造特征、要素简单组合形成将被评价为非正常申请】

【专利制度针对自然科学范畴，涉及社会科学范畴（买卖、社会福利、金融、价格）等无法客观界定效果的技术方案如仅涉及信息含义内容不同，对信息载体的处理手段均为现有技术或实质上均为现有技术，则申请将不符合专利法规定无法通过审查】

技术领域

[0001] 本发明涉及的是一种神经网络领域的技术，具体是一种端上实时的视频流物品优化推荐实现方法。

背景技术

[0002] 现有推荐技术中的重排序模型已经广泛应用于移动设备中的推荐系统和在线物品。这些模型以用户在设备上的反馈作为输入，并根据用户实时兴趣动态地重新排序项目。然而，基于预测值贪婪地重排序项目而不考虑其对未来的影响，即项目的外部性，可能会导致预期奖励的次优。此外，直接采用常用的广义第二价格(GSP)和 Vickrey-Clarke-Groves(VCG)机制算法的展示效能计算单元不能保证经济属性，并可能导致在线物品市场的不稳定。

发明内容

[0003] 本发明针对现有技术存在的上述不足，提出一种端上实时的视频流物品优化推荐实现方法，适用于资源有限的设备，以实现实时的设备上拍卖，经过在公共数据集上的实验结果证明本发明在期望社会福利，即在推荐场景下公式化定义的平台效能与物品主效用的总和方面的有效性以及具有主导策略激励兼容(DSIC)和个体理性(IR)的经济学性质。

[0004] 本发明是通过以下技术方案实现的：

[0005] 本发明涉及一种端上实时的视频流物品优化推荐实现方法，基于强化学习的端上重排序模型对用户端上实时行为进行建模，并得到预估物品的点击概率、下翻概率及带来的期望社会福利后，挑选并展示最大化社会福利的物品，再基于所预测的不同物品的点击概率、下翻概率及带来的期望社会福利，计算物品对社会福利的边际贡献作为展示效能。

技术效果

[0006] 本发明通过改进的端上重排模块实现了预估更准确的推荐系统；通过改进的机制算法实现了最大化平台社会福利的展示分配，保障了 DISC 和 IR 性质，减少了物品主策略性行为，使物品市场的平稳运行。

附图说明

[0007] 图 1 为本发明系统示意图；

[0008] 图 2 为本发明流程图；

[0009] 图 3 为端上重排模块示意图；

[0010] 图 4 为实施例效果示意图。

具体实施方式

[0011] 如图 1 所示，为本实施例涉及一种端上实时的视频流物品优化推荐实现系统，包括：基于 DDQN 的端上重排模块以及机制算法模块，其中：端上重排模块根据静态用户属性、动态环境特征、候选物品集合和目标物品信息进行推荐模型推理，得到物品点击率、下翻率和期望社会福利的预估值；机制算法模块根据端上重排模块输出的预估值信息，进行物品排序、展示物品选取及展示效能计算，得到最终展示位分配结果。

[0012] 所述的端上重排模块包括：特征映射单元、长序列特征抽取单元、反馈预估单元以及社会福利预估单元，其中：特征映射单元将静态用户属性、动态环境特特征、候选物品集合、目标物品等特征信息，进行特征映射处理，得到映射后的低维嵌入，长序列特征抽取单元将经过特征映射的动态环境信息，进行序列建模处理，得到序列特征低维嵌入，反馈预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，进行神经网络前向传播处理，得到点击率和下翻率预测值，社会福利预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，进行神经网络前向传播处理，得到期望社会福利预估值结果。

[0013] 所述的机制算法模块包括：物品打分单元、物品排序与展示选取单元以及展示效能计算单元，其中：物品打分单元根据端上重排模块输出的点击率、下翻率和期望社会福利预估值$\hat{\theta}_i, \hat{\gamma}_i, \hat{w}_i$信息计算累计期望社会福利并得到物品打分结果；物品排序与展示选取单元根据物品打分结果，进行排序并选择打分最高物品展示，得到选择进行展示的物品信息；展示效能计算单元根据展示的物品信息、物品打分、端上重排模块输出的点击率、下翻率和期望社会福利预估值信息计算边际贡献，得到该展示物品的展示效能结果。

[0014] 所述的累计期望社会福利，具体为：$\hat{\theta}_i b_i + \hat{\gamma}_i \hat{w}_i$，其中：$\hat{\theta}_i$为端上重排模块为物品$i$预估的点击率；$b_i$为物品主$i$根据自身点击价值提交给平台的展示估值；$\hat{\gamma}_i$为端上重排模块为物品$i$预估的下翻率；$\hat{w}_i$为端上重排模块为物品主$i$预估的期望社会福利。

[0015] 该打分公式考虑了当前与未来的社会福利，从而能从整体优化期望社会福利。

[0016] 所述的排序与展示选取，具体为：$j = \text{argmax}_i \hat{\theta}_i b_i + \hat{\gamma}_i \hat{w}_i$，其中：$\hat{\theta}_i$为端上重排模块为物品$i$预估的点击率；$b_i$为物品主$i$根据自身点击价值提交给平台的展示估值；$\hat{\gamma}_i$为端上重排模块为物品主$i$预估的下翻率；$\hat{w}_i$为端上重排模块为物品$i$预估的期望社会福利；$j$为所选出进行展示的物品。表示对待展示物品根据物品打分进行排序，并选择打分最高的物品j进行展示。

[0017] 所述的边际贡献，具体为：$p_i = (\hat{\theta}_i \cdot b_i - \hat{w}_i + \hat{w}_{-i})/\theta_i$，其中：$\hat{\theta}_i$为端上重排模块为物品$i$预估的点击率；$b_i$为物品主$i$根据自身点击价值提交给平台的展示估值；$\hat{w}_i$为端上重排模块为物品$i$预估的期望社会福利；$\hat{w}_{-i}$为将物品主$i$从当前候选物品集合中去除，用端上重排模块预估次优物品主j在去除了物品主i时的期望社会福利。本发明在展示效能计算单元的设计保障了 DISC 和 IR 性质，减少了物品主策略性行为，使物品市场的平稳运行。

[0018] 如图 2 所示，为本实施例基于上述系统的端上实时的视频流物品优化推荐实现方法，包括：

[0019] 步骤一：云侧服务器训练端上重排模块，具体包括：

[0020] 1.1 云侧服务器采集端设备用户日志数据。

[0021] 所述的用户日志数据包括现有推荐技术所需的特征，如用户侧特征：年龄、性别、地域；物品侧特征：物品类别、物品名称；动态环境特征：用户最近浏览物品的物品侧特征。

[0022] 1.2 云侧服务器采集用户请求日志。

[0023] 所述的用户请求日志包括用户发起的请求，记录了用户发起请求的时间、请求的上下文信息如历史记录等。

[0024] 1.3 云侧服务器根据采集的端设备用户日志数据训练环境模型。

[0025] 所述的环境模型采用任意的最先进的上下文感知精排模型，例如 DIEN、SIM 等，本实施例采用了 DIEN 精排模型。

[0001] 1.4 云侧服务器建立如图 3 所示的端上重排模型。

[0026] 所述的端上重排模型包括：特征映射单元、长序列特征抽取单元、反馈预估单元以及社会福利预估单元，其中：特征映射单元对离散特征采用嵌入查找方法，对连续特征进行非线性转换并拼接，得到映射后的低维嵌入；长序列特征抽取单元将经过特征映射的动态环境信息，使用多头注意力模块进行序列建模处理，得到序列特征低维嵌入；反馈预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，采用门控多专家混合模块(MMoE)，得到点击率和下翻率预测值；社会福利预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，采用多层感知机(MLP)进行神经网络前向传播处理，得到期望社会福利预估值结果。

[0002] 所述的端上重排模型中长序列特征提取单元包含两个多头注意力模块，用于提取动态环境和候选物品集的上下文信息，具体为：对于动态环境建模，将动态环境的特征低维嵌入作为键K和值V，将目标物品的特征低维嵌入作为查询Q，假设采用 h 头注意力模块以及动态环境特征向量组中低维嵌入维度为d^k，得到动态环境向量表征为$\text{Multihead}(Q, K, V) =$

$\text{Concat}(head_1, \ldots, head_h)W^O$，其中：$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d^k}}\right)V$待$head_i = \text{Attention}\left(QW_i^Q, KW_i^V, VW_i^V\right)$；同理，得到候选物品集的低维嵌入表征。

[0003] 所述的端上重排模型中反馈预估单元包括学习用户对物品的点击率和下翻浏览下一条物品的概率，是一项监督学习任务。采用了广泛使用的多门控混合专家(MMoE)结构进行多任务学习，将静态用户属性特征低维嵌入、动态环境特征低维嵌入以及目标物品低维嵌入作为输入，用于点击率和下翻率的预测。

[0004] 所述的端上重排模型中期望社会福利预测单元包含一个强化学习任务，使用多层感知机(MLP)来进行期望社会福利的预测。它以静态用户属性特征低维嵌入、动态环境低维嵌入、候选物品集低维嵌入以及目标物品低维嵌入为输入，以此来预测目标物品在给定状态下可能带来的最大期望社会福利。

[0027] 1.5 云侧服务器随机模拟用户请求，并用云侧推荐模型生成物品候选集合，即从用户请求日志记录中随机采样回放用户的请求。

[0028] 1.6 对于一条模拟的用户请求，用端上重排模块决策展示的物品，并用环境模型模拟真实用户的点击、下翻行为。端上重排模块的输入为端设备用户日志数据包括用户静态属性，如用户性别、年龄、地域等，动态环境特征为至少一条用户最近浏览的物品信息，物品包含物品侧特征，如物品的 ID、物品投放者 ID、物品的类别、物品所属的物品组、物品所属地域等，以及用户对物品的交互行为，及描述用户或端设备状态的额外特征，如电池电量、网络状况等。端上重排模块对候选集合内的物品输出预估的点击率、下翻率、期望社会福利$\hat{\theta}_i, \hat{\gamma}_i, \hat{w}_i$，它根据机制算法物品打分单元及物品排序与展示选取单元，选取$\hat{\theta}_i b_i + \hat{\gamma}_i \hat{w}_i$最大的物品进行展示。环境模型可以根据用户请求特征预估点击率、下翻率，根据概率采样得到模拟的真实用户的点击、下翻行为Θ, Γ；

[0029] 1.7 根据模拟的用户行为更新动态环境特征、候选物品集合，同时将这一次经验记录到重放缓冲区供端上重排模型训练。

[0030] 所述的动态环境特征、候选物品集合具体更新方法为：动态环境特征中加入新的浏览记录，并将刚曝光的物品从候选物品集合中去除。

[0031] 所述的训练使用一个重放缓冲区$\mathcal{D} = \{(s_i, s_i', a_i, \Theta_i, \Gamma_i)\}$来存储来自学习过程的经验。一则经验是一个由当前状态s_i、下一个状态s_i'、动作a_i以及用户的反馈Θ_i和Γ_i组成的元组，其中：状态、动作、奖励均为强化学习领域中的概念，分别表示智能体面对的环境、面对该环境选择的动作以及该动作所获得的奖励，在本实施例中，状态为静态用户属性、动态环境特特征、候选物品集合的组合，动作为选择展示的物品，奖励为用户点击价值$\Theta_i b_i$。

[0032] 1.8 重复 1.6 直到用户不再下翻。

[0033] 1.9 随机抽取一批数据$\{(s_i, s_i', a_i, \Theta_i, \Gamma_i)\}_{i=1}^{N}$进行学习，其中：$N$为抽取的数据样本量。具体的学习任务包括：监督学习任务、强化学习任务和辅助任务。监督学习任务：用户点击率和下翻率预测为监督学习任务。在本实施例中，采用了二元交叉熵作为监督学习的损失函数，假设对抽取样本中第i条样本，模型预测的点击率、下翻率分别为$\hat{\theta}_{a_i}, \hat{\gamma}_{a_i}$，则定义如下公式表述的损失：

$$\mathcal{L}_\theta = -\frac{1}{N}\sum_i \left[\Theta_i \log(\hat{\theta}_{a_i}) + (1-\Theta_i)\log(1-\hat{\theta}_{a_i})\right],$$

$$\mathcal{L}_\gamma = -\frac{1}{N}\sum_i \left[\Gamma_i \log(\hat{\gamma}_{a_i}) + (1-\Gamma_i)\log(1-\hat{\gamma}_{a_i})\right],$$

，其中：$\mathcal{L}_\theta$为点击率损失；$\mathcal{L}_\gamma$为下翻率损失；N为抽取的数据样本量；i为抽取缓冲区记录样本中的第i条；a_i为该条缓冲区记录的动作(选择展示的物品)；Θ_i为该条缓冲区记录的用户点击反馈；Γ_i为该条缓冲区记录的用户下翻反馈；$\hat{\theta}_{a_i}$为端上重排模型为物品a_i预估的点击率；$\hat{\gamma}_{a_i}$为端上重排模型为物品a_i预估的下翻率。

[0034] 强化学习任务：期望社会福利预测是强化学习任务。首先，根据 DDQN 算法，先计算出期望社会福利的目标值为：$W_i = \Theta_i \cdot b_{a_i} + \Gamma_i \cdot Q'^{w}(s_i', \arg\max_a Q^w(s_i', a)),$，其中：$W_i$为计算的期望社会福利的目标值；$i$为抽取缓冲区记录样本中的第$i$条；$a_i$为该条缓冲区记录的动作(选择展示的物品)；$\Theta_i$为该条缓冲区记录的用户点击反馈；$\Gamma_i$为该条缓冲区记录的用户下翻反馈；$s_i'$为该条缓冲区记录的下一状态；$b_{a_i}$为物品主$a_i$根据自身点击价值提交给平台的展示估值；，$Q^w, Q'^w$分别为 DDQN 算法中的策略网络和目标网络，他们以状态、动作作为输入，输出预估的社会福利。

[0035] 根据目标期望社会福利W_i和模型预测的期望社会福利$\hat{w}_i$计算最小平方误差，$\mathcal{L}_w = \frac{1}{N}\sum_i (W_i - \hat{w}_i)^2.$，其中：$\mathcal{L}_w$为社会福利损失；$N$为抽取的数据样本量；$i$为抽取缓冲区记录样本中的第$i$条；$W_i$为前述计算的期望社会福利的目标值；$\hat{w}_i$为端上重排模型为物品$a_i$预估的期望社会福利。

[0036] 辅助任务：用于减少点击率、下翻率预估和期望社会福利预估之间的独立性，提高模型效果。由于监督学习和强化学习使用了独立的结构，对这两项任务进行单独优化可能会导致预测值存在蹊跷板效应。因此，在本实施例中，引入了一个额外的辅助学习任务来减少这种差距

$$\mathcal{L}_{aux} = \frac{1}{N}\sum_i [\underbrace{(\hat{w}_{a_i} - \hat{\theta}_{a_i}\cdot b_{a_i})}_{\text{future welfare in } s_i} - \hat{\gamma}_{a_i} \cdot \underbrace{\max_a \hat{w}'_a}_{\text{max welfare in } s_i'}]^2,$$

，其中：$\mathcal{L}_{aux}$为辅助任务损失；N为抽取的数据样本量；i为抽取缓冲区记录样本中的第i条；a_i为该条缓冲区记录的动作(选择展示的物品)；$\hat{\theta}_{a_i}$为端上重排模型为物品a_i预估的点击率；$\hat{\gamma}_{a_i}$为端上重排模型为物品a_i预估的下翻率；$\hat{\theta}_{a_i}$为端上重排模型为物品a_i预估的点击率；b_{a_i}为物品主a_i根据自身点击价值提交给平台的展示估

值；$\widehat{w}_{a_i}$为状态s_i时模型所预测展示物品a_i的期望社会福利，$\widehat{w}'_a$为状态s'_i时模型所预测的展示物品a的期望社会福利。模型最终的损失为这四个损失的加权求和：$\mathcal{L}_{total} = \mathcal{L}_\theta + a\mathcal{L}_\gamma + b\mathcal{L}_w + c\mathcal{L}_{aux}$，其中：$\mathcal{L}_{total}$为最终损失总和；$a, b, c$分别为各损失的权重，本实施例采取了 uncertaintyweight 的方法动态设定权重。通过神经网络的反向传播最小化模型预测值与实际目标值之间的差距。

[0037] 步骤二：向端设备传输已训练好的端上重排模块，并部署机制算法。

[0038] 步骤三：在端设备上进行重排服务，具体包括：

[0039] 3.1 端设备向云侧服务器请求新的物品分页，云侧推荐系统通过召回、粗排、精排模型向端设备返回候选物品集合。

[0040] 3.2 部署在端设备上的重排模块对所有候选物品预估点击率、下翻率、期望社会福利。

[0041] 3.3 机制算法模块根据端上重排模块的输出值，选择当前展示位的展示物品，同时计算平台能获取的该物品的展示效能。

[0042] 3.4 若用户点击当前展示的物品，则平台获取相应数量的展示效能。

[0043] 3.5 端设备维护端上重排模块所需的特征，包括静态用户属性、动态环境特特征、候选物品集合，维护方法同前文。

[0044] 3.6 当云侧下发的候选物品集合均已展示，则重复 3.1；否则重复 3.2，直到用户不再下翻，表明用户已退出 app。

[0045] 经过具体实际实验，在模拟环境设置下，本发明在期望社会福利上有 1.74%的提升，证明了本公开提出端上实时的视频流物品优化推荐实现方法的有效性。具体模拟环境设置如下：本发明在公开数据集 Mobile 上进行模拟实验。Mobile 是一个针对栏目推荐场景的数据集。它记录用户连续八天的会话数据，包含 24.6 万个会话、1.2 万次点击、3.5 万名用户和 8 个栏目。本发明采用随机生成方法来创建缺失的物品主价值数据。具体来说，假设物品商i的价值b_i遵循一个正态分布$\mathcal{N}(\mu_i, \sigma_i^2)$，其中：$\mu_i \sim \mathcal{N}(\mu, \sigma^2)$且方差$\sigma_i$遵循均匀分布$\sigma_i \sim \mathcal{U}(0, \tau)$。通过调整参数$\mu, \sigma, \tau$，可以控制物品主价值的分布。本发明选择的参数为$\mu = 1, \sigma = 0.3, \tau = 0.1$。

[0046] 具体的对比基线方法为：设置以下三个基线方法 1.贪婪 GSP：根据物品的千次展示效能(eCPM)排序，采用广义二价机制进行展示效能计算(GSP)；2.uGSP：采用和本发明相同的分配方法，但是展示效能计算方法不同；3.AdaAuc-2:在本发明中，物品打分单元调整为$\hat{\theta}_i b_i + 2\hat{\gamma}_i \widehat{w}_i$。具体的效果指标为：归一化的期望社会福利(SSW)、会话长度(SL)和点击次数(NC)，分别反映不同方法社会福利效果，以及满足用户兴趣的能力。具体实验结果如表格 1。

[0047] 表 1 模拟实验结果

方法	归一化期望社会福利（SSW）	会话长度（SL）	点击次数（NC）
贪婪 GSP	1(-)	2.2580(-)	1.1081(-)
AdaAuc-2	1.0075(+0.75%)	**2.3359(+3.45%)**	**1.1499(+3.77%)**
AdaAuc（本发明）	**1.0174(+1.74%)**	2.3127(+2.42%)	1.1432(+3.17%)

[0005] 表 1 中可以发现本发明方法可以达到最高的归一化期望社会福利。由于 AdaAuc-2 在分配打分单元中未来期望社会福利的权重调整为两倍，因此该方法倾向于选择未来收益更高的物品进行展示，因此可以达到更长的会话长度，以及得到更多的点击。但 AdaAuc-2 并没有因此获得更高的归一化期望社会福利，这反映 AdaAuc-2 高估物品对未来期望社会福利的影响，而本发明提出的方法很好地平衡当前与未来社会福利的关系。

[0048] 此外，模拟实验中还考察了端上重排模块的不准确性对激励兼容性质的影响。在的实验中，随机选择一个物品主，并将其价值调整为其真实估值的α倍，其中：$\alpha\in[0,2]$，同时所有其他物品商都不变。然后，分析在不同调整策略α下物品主的平均效用。如图 4 所示，贪婪 GSP、uGSP、AdaAuc-2 和本发明都存在违反 DSIC 性质的问题，因为物品商获得最大效用的点不在诚实调整策略点α=1 上。然而，在这些方法中，使用本发明提供的方法时，达到物品商最优效用的策略最接近诚实调整策略，这表明本发明在 DSIC 性质上表现更好。此外，AdaAuc 的曲线比其他基线方法的曲线更陡，这意味着本发明对不诚实价值行为施加更严重的惩罚。总之，尽管不准确的物品展示效果预估模型损害本发明的 DSIC 性质，但实验证据表明，与贪婪 GSP 相比，本发明面对具有策略性的物品主时受到的影响较小。因此，它可以更好地激励物品主提交真实的价值。此外，本发明的性质有数学保障，因此提出更准确的模型可以保证 DSIC。

[0049] **上述具体实施可由本领域技术人员在不背离本发明原理和宗旨的前提下以不同的方式对其进行局部调整，本发明的保护范围以权利要求书为准且不由上述具体实施所限，在其范围内的各个实现方案均受本发明之约束。**

课题组已核实初稿及修改稿中所有建议，确认已无其他技术细节未补充并按以下文本提交专利局，接受专利局的审查决定。
项目负责人签字
日期

说 明 书 附 图

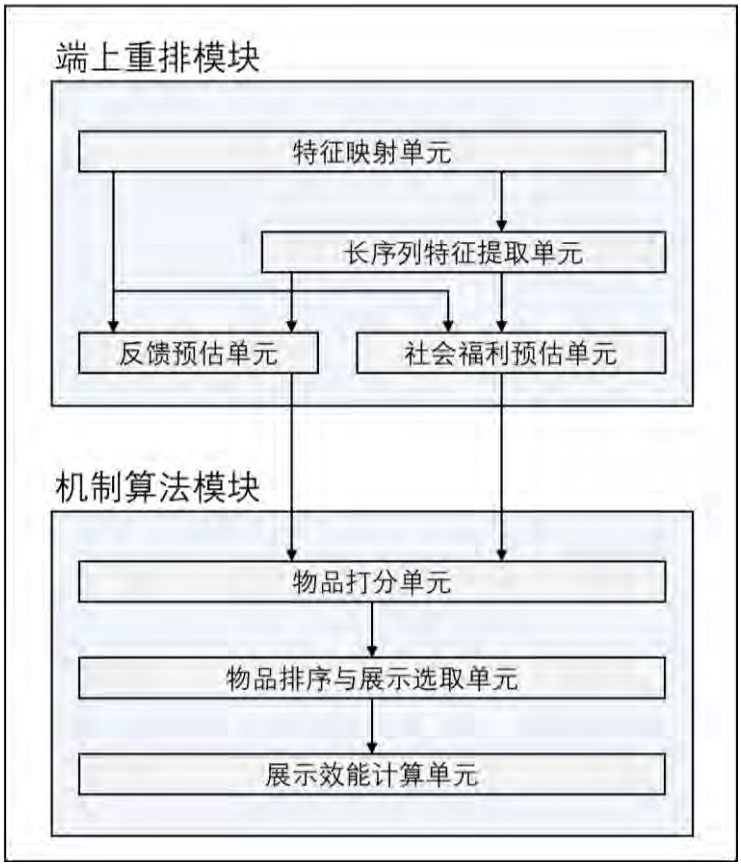

图 1

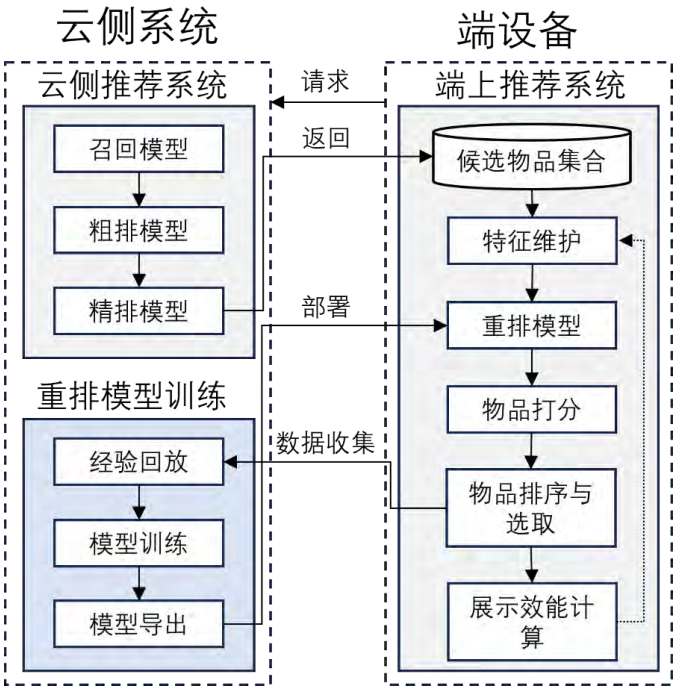

图 2

课题组已核实初稿及修改稿中所有建议，确认已无其他技术细节未补充并按以下文本提交专利局，接受专利局的审查决定。
项目负责人签字
日期

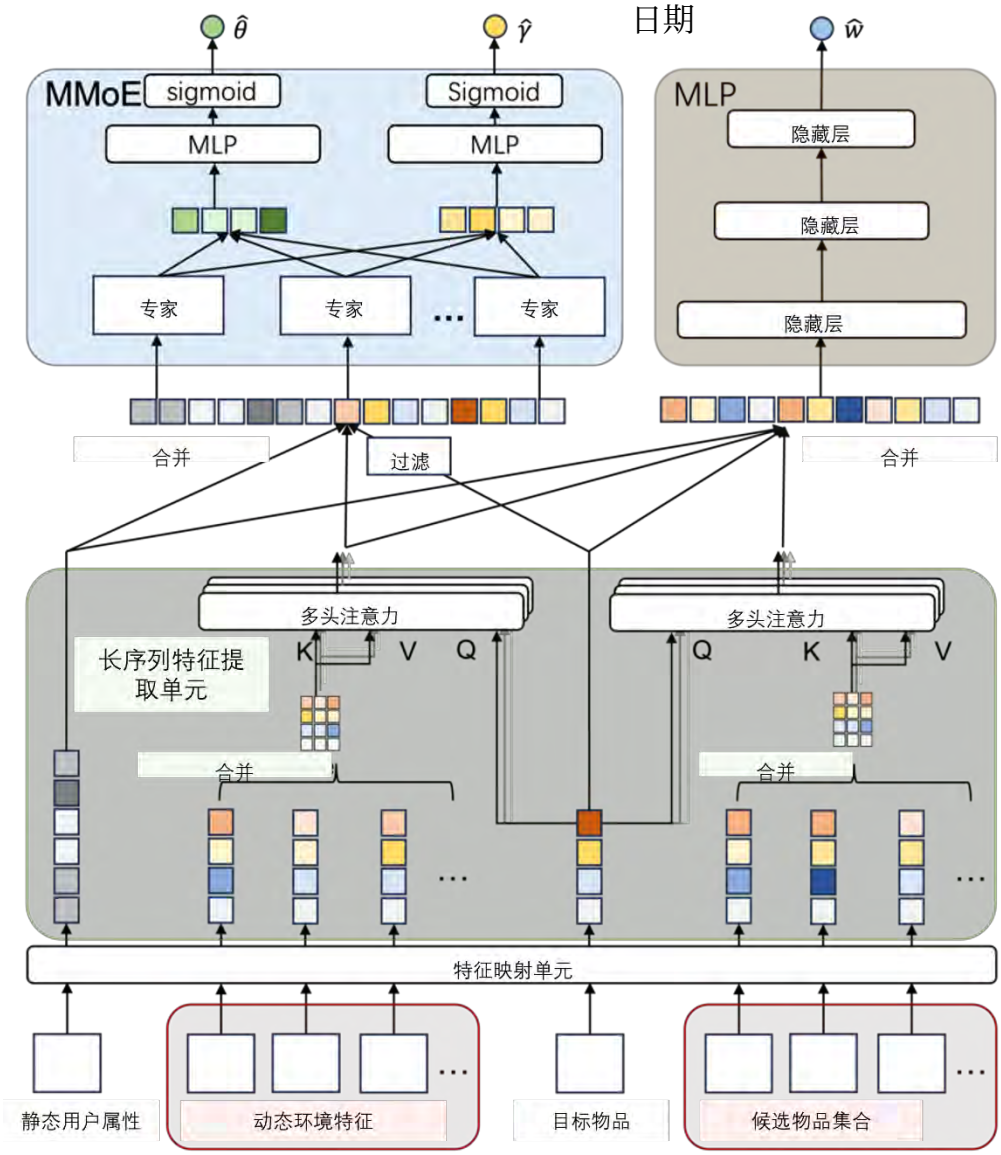

图 3

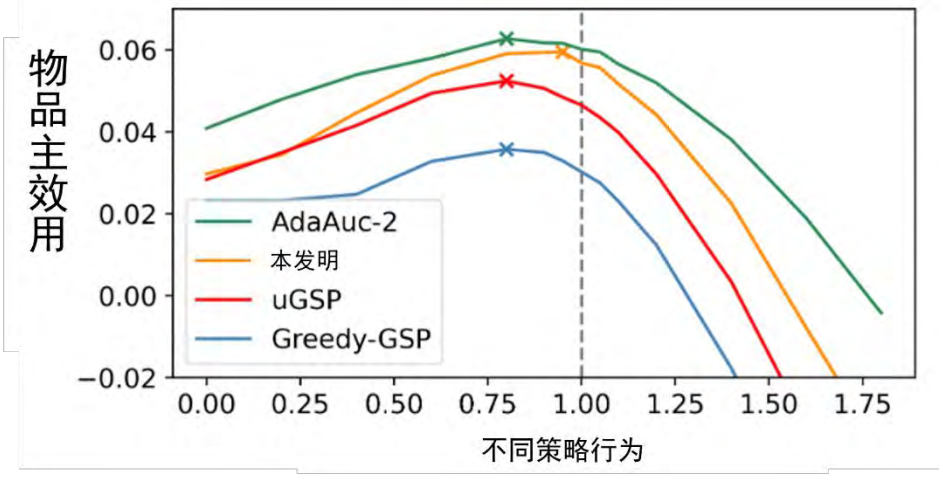

图 4

课题组已核实初稿及修改稿中所有建议，确认已无其他技术细节未补充并按以下文本提交专利局，接受专利局的审查决定.
项目负责人签字
日期

权 利 要 求 书

1、一种端上实时的视频流物品优化推荐实现系统，其特征在于，包括：基于 DDQN 的端上重排模块以及机制算法模块，其中：端上重排模块根据静态用户属性、动态环境特征、候选物品集合和目标物品信息进行推荐模型推理，得到物品点击率、下翻率和期望社会福利的预估值；机制算法模块根据端上重排模块输出的预估值信息，进行物品排序、展示物品选取及展示效能计算，得到最终展示位分配结果。

2、根据权利要求 1 所述的端上实时的视频流物品优化推荐实现系统，其特征是，所述的端上重排模块包括：特征映射单元、长序列特征抽取单元、反馈预估单元以及社会福利预估单元，其中：特征映射单元将静态用户属性、动态环境特特征、候选物品集合、目标物品等特征信息，进行特征映射处理，得到映射后的低维嵌入，长序列特征抽取单元将经过特征映射的动态环境信息，进行序列建模处理，得到序列特征低维嵌入，反馈预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，进行神经网络前向传播处理，得到点击率和下翻率预测值，社会福利预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，进行神经网络前向传播处理，得到期望社会福利预估值结果。

3、根据权利要求 1 所述的端上实时的视频流物品优化推荐实现系统，其特征是，所述的机制算法模块包括：物品打分单元、物品排序与展示选取单元以及展示效能计算单元，其中：物品打分单元根据端上重排模块输出的点击率、下翻率和期望社会福利预估值$\hat{\theta}_i, \hat{\gamma}_i, \hat{w}_i$信息计算累计期望社会福利并得到物品打分结果；物品排序与展示选取单元根据物品打分结果，进行排序并选择打分最高物品展示，得到选择进行展示的物品信息；展示效能计算单元根据展示的物品信息、物品打分、端上重排模块输出的点击率、下翻率和期望社会福利预估值信息计算边际贡献，得到该展示物品的展示效能结果。

4、根据权利要求 3 所述的端上实时的视频流物品优化推荐实现系统，其特征是，所述的累计期望社会福利，具体为：$\hat{\theta}_i b_i + \hat{\gamma}_i \hat{w}_i$，其中：$\hat{\theta}_i$为端上重排模块为物品$i$预估的点击率；$b_i$为物品主$i$根据自身点击价值提交给平台的展示估值；$\hat{\gamma}_i$为端上重排模块为物品$i$预估的下翻率；$\hat{w}_i$为端上重排模块为物品主$i$预估的期望社会福利；

所述的排序与展示选取，具体为：$j = \text{argmax}_i \hat{\theta}_i b_i + \hat{\gamma}_i \hat{w}_i$，其中：$\hat{\theta}_i$为端上重排模块为物品$i$预估的点击率；$b_i$为物品主$i$根据自身点击价值提交给平台的展示估值；$\hat{\gamma}_i$为端上重排模块为物品主$i$预估的下翻率；$\hat{w}_i$为端上重排模块为物品$i$预估的期望社会福利；$j$为所选出进行展

示的物品，表示对待展示物品根据物品打分进行排序，并选择打分最高的物品j进行展示；

所述的边际贡献，具体为：$p_i = (\hat{\theta}_i \cdot b_i - \hat{w}_i + \hat{w}_{-i})/\hat{\theta}_i$，其中：$\hat{\theta}_i$为端上重排模块为物品$i$预估的点击率；$b_i$为物品主$i$根据自身点击价值提交给平台的展示估值；$\hat{w}_i$为端上重排模块为物品$i$预估的期望社会福利；$\hat{w}_{-i}$为将物品主$i$从当前候选物品集合中去除，用端上重排模块预估次优物品主j在去除了物品主i时的期望社会福利。

5、一种基于权利要求 1-4 中任一所述系统的端上实时的视频流物品优化推荐实现方法，其特征在于，基于强化学习的端上重排序模型对用户端上实时行为进行建模，并得到预估物品的点击概率、下翻概率及带来的期望社会福利后，挑选并展示最大化社会福利的物品，再基于所预测的不同物品的点击概率、下翻概率及带来的期望社会福利，计算物品对社会福利的边际贡献作为展示效能。

6、根据权利要求 5 所述的端上实时的视频流物品优化推荐实现方法，其特征是，具体包括：

步骤一：云侧服务器训练端上重排模块，具体包括：

1.1 云侧服务器采集端设备用户日志数据；

1.2 云侧服务器采集用户请求日志；

1.3 云侧服务器根据采集的端设备用户日志数据训练环境模型；

1.4 云侧服务器建立端上重排模型，包括：特征映射单元、长序列特征抽取单元、反馈预估单元以及社会福利预估单元，其中：特征映射单元对离散特征采用嵌入查找方法，对连续特征进行非线性转换并拼接，得到映射后的低维嵌入；长序列特征抽取单元将经过特征映射的动态环境信息，使用多头注意力模块进行序列建模处理，得到序列特征低维嵌入；反馈预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，采用门控多专家混合模块(MMoE)，得到点击率和下翻率预测值；社会福利预估单元根据特征映射单元及长序列特征抽取单元输出的低维嵌入信息，采用多层感知机(MLP)进行神经网络前向传播处理，得到期望社会福利预估值结果；

1.5 云侧服务器随机模拟用户请求，并用云侧推荐模型生成物品候选集合，即从用户请求日志记录中随机采样回放用户的请求；

1.6 对于一条模拟的用户请求，用端上重排模块决策展示的物品，并用环境模型模拟真实用户的点击、下翻行为；

1.7 根据模拟的用户行为更新动态环境特征、候选物品集合，同时将这一次经验记录到重

放缓冲区供端上重排模型训练；

1.8 重复 1.6 直到用户不再下翻；

1.9 随机抽取一批数据$\{(s_i, s_i', a_i, \Theta_i, \Gamma_i)\}_{i=1}^{N}$进行学习，其中：$N$为抽取的数据样本量；

所述的学习包括：监督学习任务、强化学习任务和辅助任务；

步骤二：向端设备传输已训练好的端上重排模块，并部署机制算法；

步骤三：在端设备上进行重排服务，具体包括：

3.1 端设备向云侧服务器请求新的物品分页，云侧推荐系统通过召回、粗排、精排模型向端设备返回候选物品集合；

3.2 部署在端设备上的重排模块对所有候选物品预估点击率、下翻率、期望社会福利；

3.3 机制算法模块根据端上重排模块的输出值，选择当前展示位的展示物品，同时计算平台能获取的该物品的展示效能；

3.4 若用户点击当前展示的物品，则平台获取相应数量的展示效能；

3.5 端设备维护端上重排模块所需的特征，包括静态用户属性、动态环境特特征、候选物品集合，维护方法同前文；

3.6 当云侧下发的候选物品集合均已展示，则重复 3.1；否则重复 3.2，直到用户不再下翻，表明用户已退出 app。

7、根据权利要求 5 所述的端上实时的视频流物品优化推荐实现方法，其特征是，所述的端上重排模型中长序列特征提取单元包含两个多头注意力模块，用于提取动态环境和候选物品集的上下文信息，具体为：对于动态环境建模，将动态环境的特征低维嵌入作为键K和值V，将目标物品的特征低维嵌入作为查询Q，假设采用 h 头注意力模块以及动态环境特征向量组中低维嵌入维度为d^k，得到动态环境向量表征为$\text{Multihead}(Q,K,V) = \text{Concat}(head_1, \dots, head_h)W^O$，其中：$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d^k}}\right)V$待$head_i = \text{Attention}\left(QW_i^Q, KW_i^V, VW_i^V\right)$；同理，得到候选物品集的低维嵌入表征；

所述的端上重排模型中反馈预估单元包括学习用户对物品的点击率和下翻浏览下一条物品的概率，是一项监督学习任务，采用了广泛使用的多门控混合专家(MMoE)结构进行多任务学习，将静态用户属性特征低维嵌入、动态环境特征低维嵌入以及目标物品低维嵌入作为输入，用于点击率和下翻率的预测；

所述的端上重排模型中期望社会福利预测单元包含一个强化学习任务，使用多层感知机(MLP)来进行期望社会福利的预测，它以静态用户属性特征低维嵌入、动态环境低维嵌入、候选物品集低维嵌入以及目标物品低维嵌入为输入，以此来预测目标物品在给定状态下可能带

来的最大期望社会福利。

8、根据权利要求 5 所述的端上实时的视频流物品优化推荐实现方法，其特征是，所述的动态环境特征、候选物品集合具体更新方法为：动态环境特征中加入新的浏览记录，并将刚曝光的物品从候选物品集合中去除；

所述的训练，使用一个重放缓冲区$\mathcal{D}=\{(s_i,s_i',a_i,\Theta_i,\Gamma_i)\}$来存储来自学习过程的经验，一则经验是一个由当前状态s_i、下一个状态s_i'、动作a_i以及用户的反馈Θ_i和Γ_i组成的元组，其中：状态、动作、奖励均为强化学习领域中的概念，分别表示智能体面对的环境、面对该环境选择的动作以及该动作所获得的奖励。

9、根据权利要求 5 所述的端上实时的视频流物品优化推荐实现方法，其特征是，所述的监督学习任务是指：用户点击率和下翻率预测为监督学习任务，采用了二元交叉熵作为监督学习的损失函数，假设对抽取样本中第i条样本，模型预测的点击率、下翻率分别为$\hat{\theta}_{a_i},\hat{\gamma}_{a_i}$，则定义如下公式表述的损失：

$$\mathcal{L}_\theta=-\frac{1}{N}\sum_i\left[\Theta_i\log(\hat{\theta}_{a_i})+(1-\Theta_i)\log(1-\hat{\theta}_{a_i})\right],$$

$$\mathcal{L}_\gamma=-\frac{1}{N}\sum_i\left[\Gamma_i\log(\hat{\gamma}_{a_i})+(1-\Gamma_i)\log(1-\hat{\gamma}_{a_i})\right],$$

，其中：$\mathcal{L}_\theta$为点击率损失；$\mathcal{L}_\gamma$为下翻率损失；N为抽取的数据样本量；i为抽取缓冲区记录样本中的第i条；a_i为该条缓冲区记录的动作(选择展示的物品)；Θ_i为该条缓冲区记录的用户点击反馈；Γ_i为该条缓冲区记录的用户下翻反馈；$\hat{\theta}_{a_i}$为端上重排模型为物品a_i预估的点击率；$\hat{\gamma}_{a_i}$为端上重排模型为物品a_i预估的下翻率；

强化学习任务：期望社会福利预测是强化学习任务，首先，根据 DDQN 算法，先计算出期望社会福利的目标值为：$W_i=\Theta_i\cdot b_{a_i}+\Gamma_i\cdot Q'^{w}(s_i',\arg\max_a Q^{w}(s_i',a)),$，其中：$W_i$为计算的期望社会福利的目标值；$i$为抽取缓冲区记录样本中的第$i$条；$a_i$为该条缓冲区记录的动作(选择展示的物品)；$\Theta_i$为该条缓冲区记录的用户点击反馈；$\Gamma_i$为该条缓冲区记录的用户下翻反馈；$s_i'$为该条缓冲区记录的下一状态；$b_{a_i}$为物品主$a_i$根据自身点击价值提交给平台的展示估值；，$Q^w,Q'^w$分别为 DDQN 算法中的策略网络和目标网络，他们以状态、动作作为输入，输出预估的社会福利；

根据目标期望社会福利W_i和模型预测的期望社会福利$\hat{w}_i$计算最小平方误差，$\mathcal{L}_w=\frac{1}{N}\sum_i(W_i-\hat{w}_i)^2.$，其中：$\mathcal{L}_w$为社会福利损失；$N$为抽取的数据样本量；$i$为抽取缓冲区记录样本中的第$i$条；$W_i$为前述计算的期望社会福利的目标值；$\hat{w}_i$为端上重排模型为物品$a_i$预估

的期望社会福利；

辅助任务：采用辅助学习任务来减少这种差距 $\mathcal{L}_{aux} = \frac{1}{N}\sum_i [\underbrace{(\hat{w}_{a_i} - \hat{\theta}_{a_i} \cdot b_{a_i})}_{\text{future welfare in } s_i} - \hat{\gamma}_{a_i} \cdot \underbrace{\max_a \hat{w}'_a}_{\text{max welfare in } s'_i}]^2$，其中：$\mathcal{L}_{aux}$为辅助任务损失；$N$为抽取的数据样本量；$i$为抽取缓冲区记录样本中的第$i$条；$a_i$为该条缓冲区记录的动作(选择展示的物品)；$\hat{\theta}_{a_i}$为端上重排模型为物品$a_i$预估的点击率；$\hat{\gamma}_{a_i}$为端上重排模型为物品$a_i$预估的下翻率；$\hat{\theta}_{a_i}$为端上重排模型为物品$a_i$预估的点击率；$b_{a_i}$为物品主$a_i$根据自身点击价值提交给平台的展示估值；$\hat{w}_{a_i}$为状态$s_i$时模型所预测展示物品$a_i$的期望社会福利；$\hat{w}'_a$为状态$s'_i$时模型所预测的展示物品$a$的期望社会福利，模型最终的损失为这四个损失的加权求和：$\mathcal{L}_{total} = \mathcal{L}_\theta + a\mathcal{L}_\gamma + b\mathcal{L}_w + c\mathcal{L}_{aux}$，其中：$\mathcal{L}_{total}$为最终损失总和；$a, b, c$分别为各损失的权重。

10、根据权利要求 9 所述的端上实时的视频流物品优化推荐实现方法，其特征是，所述的辅助任务中，采取 uncertaintyweight 的方法动态设定权重，通过神经网络的反向传播最小化模型预测值与实际目标值之间的差距。

课题组已核实初稿及修改稿中所有建议，
确认已无其他技术细节未补充并按以下
文本提交专利局,接受专利局的审查决定.
项目负责人签字
日期

说 明 书 摘 要

一种端上实时的视频流物品优化推荐实现方法，基于强化学习的端上重排序模型对用户端上实时行为进行建模，并得到预估物品的点击概率、下翻概率及带来的期望社会福利后，挑选并展示最大化社会福利的物品，再基于所预测的不同物品的点击概率、下翻概率及带来的期望社会福利，计算物品对社会福利的边际贡献作为展示效能。本发明适用于资源有限的设备，以实现实时的设备上拍卖，经过在公共数据集上的实验结果证明本发明在期望社会福利，即在推荐场景下公式化定义的平台效能与物品主效用的总和方面的有效性以及具有主导策略激励兼容(DSIC)和个体理性(IR)的经济学性质。

课题组已核实初稿及修改稿中所有建议，确认已无其他技术细节未补充并按以下文本提交专利局，接受专利局的审查决定.
项目负责人签字
日期

摘 要 附 图