

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 11 月 5 日 (05.11.2020)



(10) 国际公布号
WO 2020/220757 A1

- (51) 国际专利分类号:
G06F 16/9535 (2019.01)
- (21) 国际申请号: PCT/CN2020/071699
- (22) 国际申请日: 2020 年 1 月 13 日 (13.01.2020)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910355868.6 2019年4月29日 (29.04.2019) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 陈岑(CHEN, Cen); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 胡旭(HU, Xu); 中国

浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 傅驰林(FU, Chilin); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 张晓露(ZHANG, Xiaolu); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街9号嘉华大厦B座409, Beijing 100085 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: METHOD AND DEVICE FOR PUSHING OBJECT TO USER BASED ON REINFORCEMENT LEARNING MODEL

(54) 发明名称: 基于强化学习模型向用户推送对象的方法和装置

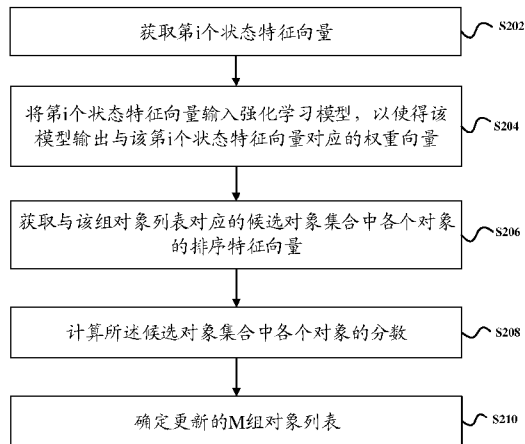


图 2

S202 Obtain the ith state feature vector
S204 Input the ith state feature vector into a reinforcement learning model to enable the reinforcement learning model to output a weight vector corresponding to the ith state feature vector
S206 Obtain sorting feature vectors of objects in a candidate object set corresponding to the group of object lists
S208 Calculate scores of the objects in the candidate object set
S210 Determine updated M groups of object lists

(57) Abstract: A method and device for determining a push object list for a user based on a reinforcement learning model. The method comprises: for each group of object lists, obtaining the ith state feature vector (S202); inputting the ith state feature vector into the reinforcement learning model to enable the reinforcement learning model to output a weight vector corresponding to the ith state feature vector (S204); obtaining sorting feature vectors of objects in a candidate object set corresponding to the group of object lists (S206); calculating scores of the objects in the candidate object set on the basis of a point product of the sorting feature vectors of the objects in the candidate object set and the weight vector (S208); and for the M groups of object lists, determining updated M groups of object lists on the basis of the scores of the objects in M candidate object sets corresponding to the M groups of object lists (S210), wherein each group of object lists in the updated M groups of object lists comprises i objects.

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种基于强化学习模型确定针对用户的推送对象列表的方法和装置, 所述方法包括: 对于每组对象列表, 获取第*i*个状态特征向量 (S202); 将所述第*i*个状态特征向量输入所述强化学习模型, 以使得所述强化学习模型输出与该第*i*个状态特征向量对应的权重向量 (S204); 获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量 (S206); 以及基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积, 计算所述候选对象集合中各个对象的分数 (S208); 以及对于所述*M*组对象列表, 基于与所述*M*组对象列表分别对应的*M*个候选对象集合中各个对象的分数, 确定更新的*M*组对象列表 (S210), 其中, 所述更新的*M*组对象列表中的每组对象列表包括*i*个对象。

基于强化学习模型向用户推送对象的方法和装置

技术领域

[01]本说明书实施例涉及机器学习领域，更具体地，涉及一种基于强化学习模型确定针对用户的推送对象列表的方法和装置。

5 背景技术

[02]传统的客户服务是人力/资源密集型和耗时的，因此，构建能够自动回答用户面临问题的智能助手非常重要。最近，人们越来越关注如何用机器学习来更好地构建这样的智能助手。作为客户服务机器人的核心功能，用户意图预测旨在自动预测用户可能想要询问的问题，并向用户呈现候选问题以供其选择以减轻用户的认知负担。更具体地说，用户意图预测任务可以被视为前 N 项（Top N）推荐的任务，其中每个预定好的问题是一个意图类（class）。目前的现有方法将该任务视为一个多分类问题，在给定当前用户状态下预测用户最可能感兴趣的对象（item）列表，即问题列表。这些方法旨在最大化即时奖励，例如点击，而忽略了推荐列表中在前的推荐对象对在后的推荐对象的影响。

[03]因此，需要一种更有效的向用户推送一组对象列表的方案。

15 发明内容

[04]本说明书实施例旨在提供一种更有效的基于强化学习模型确定针对用户的推送对象列表的方案，以解决现有技术中的不足。

[05]为实现上述目的，本说明书一个方面提供一种基于强化学习模型确定针对用户的推送对象列表的方法，其中，对于第一用户，已预先通过所述方法确定有 M 组对象列表，每组对象列表中当前包括 i-1 个对象，其中，M、i 都为大于等于 1 的整数，其中，i 小于等于预定整数 N，所述方法包括：

[06]对于每组对象列表，

[07]获取第 i 个状态特征向量，所述第 i 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括该组对象列表所述 i-1 个对象各自的属性特征；

[08]将所述第 i 个状态特征向量输入所述强化学习模型，以使得所述强化学习模型输出

与该第 i 个状态特征向量对应的权重向量，所述权重向量包括预定数目的排序特征各自的权重；

[09] 获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量，所述排序特征向量包括所述预定数目的排序特征各自的特征值；以及

5 [10] 基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积，计算所述候选对象集合中各个对象的分数；以及

[11] 对于所述 M 组对象列表，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表，其中，所述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

10 [12] 在一个实施例中，所述动态特征至少包括所述 $i-1$ 个对象各自的以下属性特征：当前热度、对象标识、对象所属类型。

[13] 在一个实施例中，所述 M 组对象列表中包括第一组对象列表，与该第一组对象列表对应的候选对象集合中包括第一对象，与该第一对象对应的排序特征向量至少包括以下排序特征的值：所述第一用户对该第一对象的预估点击率、该第一对象的当前热度、该
15 第一对象相对于所述第一组对象列表中的 $i-1$ 个对象的多样性。

[14] 在一个实施例中，已预先通过所述方法确定有 M 组对象列表包括，已预先通过所述方法确定有一组对象列表，其中，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表包括，基于与该组对象列表对应的候选对象集合中各个对象的分数，以所述候选对象集合中分数最高的对象作为该组对象
20 列表的第 i 个对象，并将该组对象列表作为更新的一组对象列表。

[15] 在一个实施例中， M 大于等于 2，其中，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表包括，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，通过集束搜索算法确定更新的 M 组对象列表。

25 [16] 在一个实施例中， i 等于 N ，所述方法还包括，通过集束搜索算法，从所述更新的 M 组对象列表中确定针对所述第一用户的推送对象列表。

[17] 在一个实施例中，所述方法还包括，

[18] 以所述推送对象列表中各个对象的排列顺序，向所述第一用户推送所述各个对象，

以获取所述第一用户的反馈;

[19]基于所述排列顺序和所述反馈获取 N 个回报值,所述 N 个回报值与对所述方法的从 $i=1$ 至 N 的 N 次循环分别对应;

5 [20]获取第 $N+1$ 个状态特征向量,所述第 $N+1$ 个状态特征向量包括静态特征和动态特征,其中,所述静态特征包括所述第一用户的属性特征,所述动态特征包括所述推送对象列表中 N 个对象各自的属性特征;以及

10 [21]基于与所述 N 次循环分别对应的 N 组数据训练所述强化学习模型,以优化所述强化学习模型,其中,所述 N 组数据包括第 1 至第 N 组数据,其中,第 i 组数据包括:与所述推送对象列表对应的第 i 个状态特征向量、与该第 i 个状态特征向量对应的权重向量、与
所述推送对象列表对应的第 $i+1$ 个状态特征向量、以及与第 i 次循环对应的回报值。

[22]在一个实施例中,所述对象为询问问题,对于第 1 至 $N-1$ 次循环中的第 i 次循环,与
所述第 i 次循环对应的回报值基于所述第一用户的如下反馈获取:是否点击所述推送对象列表中的第 i 个问题。

15 [23]在一个实施例中,与所述第 N 次循环对应的回报值基于所述第一用户的如下反馈获取:是否点击所述推送对象列表中的第 N 个问题、以及提交的满意度信息。

[24]在一个实施例中,所述强化学习模型为基于深度确定策略梯度算法的模型。

20 [25]本说明书另一方面提供一种基于强化学习模型确定针对用户的推送对象列表的装置,其中,对于第一用户,已预先通过所述方法确定有 M 组对象列表,每组对象列表中当前包括 $i-1$ 个对象,其中, M 、 i 都为大于等于 1 的整数,其中, i 小于等于预定整数 N ,
所述装置包括:

[26]用于每组对象列表的,

[27]第一获取单元,配置为,获取第 i 个状态特征向量,所述第 i 个状态特征向量包括静态特征和动态特征,其中,所述静态特征包括所述第一用户的属性特征,所述动态特征包括该组对象列表中所述 $i-1$ 个对象各自的属性特征;

25 [28]输入单元,配置为,将所述第 i 个状态特征向量输入所述强化学习模型,以使得所述强化学习模型输出与该第 i 个状态特征向量对应的权重向量,所述权重向量包括预定数目的排序特征各自的权重;

[29]第二获取单元,配置为,获取与该组对象列表对应的候选对象集合中各个对象的排

序特征向量，所述排序特征向量包括所述预定数目的排序特征各自的特征值；以及

[30] 计算单元，配置为，基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积，计算所述候选对象集合中各个对象的分数；以及

5 [31] 第一确定单元，配置为，对于所述 M 组对象列表，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表，其中，所述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

10 [32] 在一个实施例中，已预先通过所述方法确定有 M 组对象列表包括，已预先通过所述方法确定有一组对象列表，其中，所述第一确定单元还配置为，基于与该组对象列表对应的候选对象集合中各个对象的分数，以所述候选对象集合中分数最高的对象作为该组对象列表的第 i 个对象，并将该组对象列表作为更新的一组对象列表。

[33] 在一个实施例中，其中，M 大于等于 2，其中，所述第一确定单元还配置为，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，通过集束搜索算法确定更新的 M 组对象列表。

15 [34] 在一个实施例中，i 等于 N，所述装置还包括，第二确定单元，配置为，通过集束搜索算法，从所述更新的 M 组对象列表中确定针对所述第一用户的推送对象列表。

[35] 在一个实施例中，所述装置还包括，

[36] 推送单元，配置为，以所述推送对象列表中各个对象的排列顺序，向所述第一用户推送所述各个对象，以获取所述第一用户的反馈；

20 [37] 第三获取单元，配置为，基于所述排列顺序和所述反馈获取 N 个回报值，所述 N 个回报值与对所述方法的从 $i=1$ 至 N 的 N 次循环分别对应；

[38] 第四获取单元，配置为，获取第 N+1 个状态特征向量，所述第 N+1 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括所述推送对象列表中 N 个对象各自的属性特征；以及

25 [39] 训练单元，配置为，基于与所述 N 次循环分别对应的 N 组数据训练所述强化学习模型，以优化所述强化学习模型，其中，所述 N 组数据包括第 1 至第 N 组数据，其中，第 i 组数据包括：与所述推送对象列表对应的第 i 个状态特征向量、与该第 i 个状态特征向量对应的权重向量、与所述推送对象列表对应的第 i+1 个状态特征向量、以及与第 i 次循环对应的回报值。

[40] 本说明书另一方面提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行上述任一项方法。

[41] 本说明书另一方面提供一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现上述任一项方法。

- 5 [42] 通过根据本说明书实施例的基于强化学习模型确定针对用户的推送对象列表的方案旨在优化长期累积的混合奖励，例如最后的回报值可基于用户点击和用户满意度等多个维度获取，另外，策略函数可以随着问题火热度和用户行为模式的变化而不断动态更新和调整，从而有助于提升点击率。

附图说明

- 10 [43] 通过结合附图描述本说明书实施例，可以使得本说明书实施例更加清楚：
- [44] 图 1 示出根据本说明书实施例的对象推送系统 100 的示意图；
- [45] 图 2 所示方法为例如图 1 所示的模型单元 11 的一次决策过程；
- [46] 图 3 示出根据本说明书实施例的模型的 N ($N=6$) 步决策过程；
- [47] 图 4 示意示出当采用贪婪搜索方式时在图 1 所示系统中确定推送对象列表的过程；
- 15 [48] 图 5 示意示出通过集束搜索方式确定两组对象列表的过程；
- [49] 图 6 示出根据本说明书实施例的一种基于强化学习模型确定针对用户的推送对象列表的装置 6000。

具体实施方式

[50] 下面将结合附图描述本说明书实施例。

- 20 [51] 图 1 示出根据本说明书实施例的对象推送系统 100 的示意图。所述对象推送系统例如为问题预测系统，其使得用户联系客服时，可自动预测该用户可能想要询问的问题的问题列表，并在客服页面显示该问题列表，以提高用户的使用体验，并节省人工客服成本。可以理解，根据本说明书实施例的对象推送系统 100 不限于进行询问问题列表的推送，而可以用于推送各种对象的列表，如商品、影视作品、新闻等等。如图 1 所示，系
- 25 统 100 包括模型单元 11、训练单元 12 和排序单元 13。所述模型单元 11 例如包括神经网络，以用于实施强化学习算法，在本说明书实施例中，可使用各种强化学习模型，如

基于如下任一算法 DDPG、DPG、Actor-critic 的模型等等，在此不一一列出，下文将以 DDPG 算法为例进行描述。

[52]在通过系统 100 进行例如问题列表推送的情况中，通过依次向模型单元 11 输入连续的 N 个状态 $(s_1, s_2 \dots s_N)$ ，而最终在排序单元 13 获取包括 N 个问题的推送问题列表。

5 其中，例如，在输入 s_1 的情况中，模型单元 11 基于 s_1 输出相应的行为 a_1 ，在排序单元 13 中，基于 a_1 和候选问题各自的排序特征对各个问题打分，并基于各个问题的分数，确定推送问题列表的第一个问题。这里可通过贪婪搜索的算法确定第一问题，可以理解，本说明书实施例不限于此，例如还可以采用集束搜索的算法确定。在确定了上述第一问题之后，也相应地确定了环境的第二个状态 s_2 ，即环境的当前状态与用户的特征和已经
10 确定的推送问题列表中的问题相关，而在确定第二个状态 s_2 之后，可相应地确定行为 a_2 ，及推送问题列表中的第二个问题。从而，在推送问题列表被预设为包括 N 个问题的情况中，可通过模型的 N 次决策过程，获取包括 N 个问题的推送问题列表。

[53]在获取上述推送问题列表之后，通过将该列表展示给用户，可获取用户的反馈，从而可基于该反馈获取模型的每次决策的回报值 r_i 。从而可在训练模型 12 中，基于上述各
15 个状态和行为以及各个回报值（即 N 组 (s_i, a_i, s_{i+1}, r_i) ），训练所述强化学习模型，并将更新的参数传送给模型单元 11，以对其进行更新。

[54]下面将对上述模型决策过程和模型训练过程进行详细描述。

[55]图 2 示出根据本说明书实施例的一种基于强化学习模型确定针对用户的推送对象列表的方法，其中，对于第一用户，已预先通过所述方法确定有 M 组对象列表，每组对象列表中当前包括 $i-1$ 个对象，其中， M 、 i 都为大于等于 1 的整数，其中， i 小于等于
20 预定整数 N ，所述方法包括：

[56]对于每组对象列表，

[57]步骤 S202，获取第 i 个状态特征向量，所述第 i 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括该组对象列表中所述 $i-1$ 个对象各自的属性特征；
25

[58]步骤 S204，将所述第 i 个状态特征向量输入所述强化学习模型，以使得所述强化学习模型输出与该第 i 个状态特征向量对应的权重向量，所述权重向量包括预定数目的排序特征各自的权重；

[59]步骤 S206，获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量，

所述排序特征向量包括所述预定数目的排序特征各自的特征值；以及

[60] 步骤 S208, 基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积, 计算所述候选对象集合中各个对象的分数; 以及

5 [61] 步骤 S210, 对于所述 M 组对象列表, 基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数, 确定更新的 M 组对象列表, 其中, 所述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

10 [62] 图 2 所示方法为例如图 1 所示的模型单元 11 的一次决策过程, 即向强化学习模型输入 $s_1, s_2 \dots s_N$ 中任一状态以在排序问题列表中增加一个问题的过程。例如, 将要向模型输入状态 s_i , 其中 $1 \leq i \leq N$ 。如上文所述, 在通过贪婪搜索算法对问题排序的情况中, 在模型的分别基于 $s_1, s_2 \dots s_{i-1}$ 的决策过程中, 已确定有一组对象列表, 该组对象列表中当前包括 i-1 个对象。在通过集束搜索的算法排序的情况中, 例如将集束宽度预设为 2, 即 $M=2$, 从而, 在模型的分别基于 $s_1, s_2 \dots s_{i-1}$ 的决策过程中, 已确定有两组对象列表, 并且每组对象列表中当前包括 i-1 个对象。

[63] 下面详细描述该方法中的每个步骤。

15 [64] 其中, 步骤 S202-步骤 S208 是针对上述已有的 M 组对象列表中每组对象列表的步骤, 即针对 M 组对象列表中的每组对象列表分别实施步骤 S202-步骤 S208。

[65] 首先, 在步骤 S202, 获取第 i 个状态特征向量, 所述第 i 个状态特征向量包括静态特征和动态特征, 其中, 所述静态特征包括所述第一用户的属性特征, 所述动态特征包括该组对象列表中所述 i-1 个对象各自的属性特征。

20 [66] 所述第 i 个状态特征向量即为上述状态 s_i , 如上文所述, 在将要实施该方法的当前, 在预先确定的每组对象列表中当前包括 i-1 个对象, 在本说明书实施例中, 将 s_i 设定为不仅与用户的静态特征相关, 还与所述已经确定的 i-1 个对象相关, 从而可在确定第 i 个对象的过程中, 考虑列表中已有的对象的属性。其中, 所述用户的静态特征例如为该用户的年龄、教育背景、地理位置等等。所述动态特征例如为所述 i-1 个对象各自的当前热度、对象标识 (例如问题序号)、对象类型等。例如, 所述对象为用户的询问问题, 在对模型输入 s_1 以执行模型的第一次决策之前, 可预设预定数目的问题作为该次决策的候选问题集合。可根据多个用户在预定时段内对每个候选问题的提问次数, 确定每个候选问题的热度。可预先对上述预定数目的问题进行分类, 从而确定每个问题的类型, 例如, 在支付宝的客服系统中, 问题的类型例如包括: 关于花呗的问题、关于购物的问

25

题、热点问题等等。

[67]图 3 示出根据本说明书实施例的模型的 N ($N=6$) 步决策过程, 其中包括各步的输入状态 s_1-s_6 。如图 3 中所示, 在各个状态中, 下部的数据条对应于静态特征, 上部的数据条示意出动态特征的一部分。在动态特征部分中, 每个方形表示动态特征部分中的一个维度, 每个方形对应的数值表示前面各步决策中确定的各个问题的属性, 例如问题类型。如图中所示, 在输入 s_1 之前, 还未确定问题列表中的问题, 因此, 各个方框中的数值为 0, 在输入 s_2 之前, 模型已经基于输入的 s_1 进行了第一次决策从而确定了问题列表中的第一个问题, 因此, 可基于该第一个问题确定 s_2 的动态特征, 如图中所示, s_2 的动态特征的第一个方框对应于数值 5, 其例如表示第一个问题的类型标识, 类似地, s_3 的动态特征的第一个方框中的数值 5 和第二个方框中的数值 2 分别对应于相应问题列表中第一个问题和第二个问题的类型。

[68]步骤 S204, 将所述第 i 个状态特征向量输入所述强化学习模型, 以使得所述强化学习模型输出与该第 i 个状态特征向量对应的权重向量, 所述权重向量包括预定数目的排序特征各自的权重。仍然如图 3 中所示, 在每步决策中, 在确定状态 s_i 之后, 通过将 s_i 输入强化学习模型, 从而可使得模型输出相应的行为 (即权重向量) $a_i=\{w_{i0}, w_{i1}, \dots, w_{im}\}$, 其中, $i=1, 2, \dots, 6$, 其中 w_{ij} 表示排序特征 f_{ij} 的权重, 如图中所示, 每个权重向量 a_i 中的圆形表示该向量的一个维度, 即对应于一个 w_{ij} 的值, 其中, 三个圆形表示 $j=3$, 所述排序特征 f_{ij} 为各个对象的用于获取排序分数的特征, 将在下文对其详细描述。

[69]如上文所述, 所述强化学习模型例如为 DDPG 模型, 该模型通过基于神经网络进行学习而获取。所述神经网络中包括策略网络和价值网络, 在本说明书实施例中, 策略网络例如包括两层全连接层, 在策略网络中通过如下公式 (1) 和 (2) 基于 s_i 计算 a_i :

$$a_i = \mu(s_i) = \tanh (W_2 H_i + b_2) \quad (1)$$

$$H_i = \tanh (W_1 S_i + b_1) \quad (2)$$

[70]其中, W_1 、 W_2 、 b_1 和 b_2 为该策略网络中的参数, 通过激活函数 $\tanh()$, 将 a_i 各个元素 w_{ij} 的值限制在 $[-1,1]$ 之间。可以理解, 上述描述仅仅是示意性的, 所述强化学习模型不限于 DDPG 模型, 从而不限于通过策略网络来基于 s_i 获得 a_i , 另外, 所述策略网络的结构不限于使用激活函数 $\tanh$, 从而 w_{ij} 的值不必需限制在 $[-1,1]$ 之间。

[71]步骤 S206, 获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量, 所述排序特征向量包括所述预定数目的排序特征各自的特征值。

[72]如上文所述,例如,所述对象为询问问题,在对模型输入 s_1 以执行模型的第一次决策之前,可预设预定数目的问题作为该次决策的候选问题集合。在输入 s_1 之后,将根据模型决策结果确定至少一组问题列表,对于其中的一组问题列表,其中已经包括该问题列表的第一个问题,从而,在对模型输入例如 s_2 ,以进行第2步决策的过程中,与该组问题列表对应的候选问题集合为从上述预定数目的问题中移除上述第一个问题所获取的候选问题集合。在后续次的决策过程中,可同样地确定与该组问题列表对应的候选问题集合,即,所述候选问题集合为通过从初始预设的问题集合中移除该组问题列表中包括的问题所获取的问题集合。

[73]可将模型的第 i 次决策中对象 k 的排序特征向量表示为 $f_i^k = \{f_{i0}^k, f_{i1}^k, \dots, f_{im}^k\}$,该排序特征向量的维度与上述模型输出的行为向量 a_i 的维度相同,与对象的各个排序特征分别对应。所述各个排序特征可基于具体场景中对对象排序的影响因素确定,例如,在所述对象为客服场景中的询问问题的情况中,所述排序特征例如包括:该场景中用户的预估点击率、该问题的当前热度和问题多样性。所述预估点击率可通过现有的点击率预估模型(CTR模型)基于例如用户的点击历史行为 and 用户特征等获取。其中,预估点击率用于体现用户的偏好,问题热度用于体现实时问题流行度,问题多样性用于体现推荐问题的多样性。例如,在模型进行第 i 步决策之前,当前已确定有第一组问题列表,与该第一组问题列表对应的候选问题集合中包括第一问题,从而该第一问题的问题多样性特征值基于该组问题列表中已有的 $i-1$ 个问题的类型确定,例如,在所述 $i-1$ 个问题的类型中不包括第一问题的类型的情况中,可将第一问题的多样性特征值确定为 1,在所述 $i-1$ 个问题的类型中已经包括第一问题的类型的情况中,可将第一问题的多样性特征值确定为 0。

[74]步骤 S208,基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积,计算所述候选对象集合中各个对象的分数。

[75]在通过上述步骤获取了在第 i 次决策中的权重向量和候选对象集合中各个对象的排序特征向量之后,对于候选对象集合中的问题 k ,可通过例如如下公式(3)计算问题 k 在第 i 次决策中的排序分 c_i^k :

$$c_i^k = \sum_{j=0}^m f_{ij}^k \times w_{ij} \quad (3)$$

[76]可以理解,所述公式(3)仅仅是可选的一种计算方法,所述分数的计算不限于此,例如,还可以通过分别对排序特征向量和权重向量进行归一化,然后再通过对其进行点积获取相应的分数,等等。

[77]步骤 S210, 对于所述 M 组对象列表, 基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数, 确定更新的 M 组对象列表, 其中, 所述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

[78]如上文所述, 在根据各个对象的分数确定对象列表中, 可通过贪婪搜索的方式或集束搜索的方式进行确定。

[79]在采用贪婪搜索方式的情况中, 在模型的每次决策中, 仅选出候选对象集合中分数最高的对象作为推送对象列表的第一个对象。图 4 示意示出当采用贪婪搜索方式时在图 1 所示系统中确定推送对象列表的过程。如图 4 所示, 图中包括图 1 中的模型单元 11 和排序单元 13。在初始, 排序单元中还未确定对象列表, 此时, 可认为对象列表中包括 0 个对象。基于该包括 0 个对象的对象列表, 确定第一次决策的状态 s_1 , 并将该状态 s_1 输入模型单元, 模型单元中的强化学习模型基于该状态 s_1 获取行为 a_1 , 在排序单元 13 中, 基于 a_1 获取候选对象集合中各个对象的分数, 从而将其中分数最高的对象确定为该对象列表中的第一个对象。在确定第一对象之后, 可基于该第一个对象, 确定模型的第二次决策的状态 s_2 , 类似地, 通过将 s_2 输入模型单元, 从而获取行为 a_2 , 继而基于 a_2 获取候选对象集合中各个对象的分数, 并基于分数确定该对象列表中的第二个对象, 从而可基于该对象列表中的第一个对象和第二个对象, 确定第三次决策的状态 s_3 , 可以理解, 在第二次决策中的候选对象集合与第一次决策中的候选对象集合已经不同, 其中已经不包括所述第一个对象。类似地, 后续的每次决策过程都可与前述决策过程类似地进行, 例如, 在模型的第 5 次决策过程之后, 确定行为 a_5 , 从而可计算相应的候选对象集合中每个对象的分数, 从而确定该对象列表中的第 5 个对象, 并继而基于该对象列表中已有的 5 个对象, 确定第 6 次决策的状态 s_6 , 通过将状态 s_6 输入模型, 获取行为 a_6 , 并基于 a_6 确定该对象列表中的第 6 个对象。从而可通过模型的六次决策过程确定包括 6 个对象的对象列表, 并且可将该对象列表作为推送对象列表推送给相应的用户, 如第一用户。

[80]在采用集束搜索的方式的情况中, 例如集束宽度为 2, 也就是说, 在模型的每次决策中, 将确定两组对象列表。图 5 示意示出通过集束搜索方式确定两组对象列表的过程。如图中左部所示, 在进行模型的第一次决策中, 在对模型输入 s_1 之后, 可与上述贪婪搜索方式中相同地, 计算候选对象集合中各个对象的分数, 从而可获取分数排在前两位的两个对象 (例如对象 1 和对象 2) 分别作为两组对象列表中的第一个对象, 该两组对象列表左侧的“ s_1 ”用于指示其都是基于状态 s_1 获取的。如图中右部所示, 在获取图中左部两

个对象列表之后，基于每个对象列表可确定新的状态 s_{21} 和 s_{22} ，类似地，可分别基于状态 s_{21} 和状态 s_{22} 进行模型的第二次决策，从而可分别确定相应的两个对象列表，即共确定图中右部的四个对象列表中，如图中所示，图中右部上面两个列表与状态 s_{21} 对应，即，其中第一个对象都为对象 1，下面两个列表与状态 s_{22} 对应，即其中第一个对象都为对象 2。在该四个对象列表中，可分别计算其中第一个对象和第二个对象的分数和，并取分数和排在前两位的两个对象列表作为在第二次决策中确定的两个对象列表，例如图中两个虚线框中的两个对象列表。例如将要通过模型的 6 次决策过程确定推送对象列表，即， $N=6$ ，在该情况中，在第 6 次决策中，在如上所述获取两个对象列表（该每个对象列表包括 6 个已确定的对象）之后，可将该两个对象列表中各个对象的分数和最高的对象列表作为推送对象列表推送给相应的用户。

[81]在如上所述获取用户（例如第一用户）的推送对象列表之后，可以以该列表中的对象的排列顺序向第一用户推送该列表中的 N 个对象。例如在客服界面中展示顺序排列的 N 个问题，或者顺序展示所述 N 个问题等等。在对第一用户进行上述推送之后，可获取第一用户的反馈，如第一用户对 N 个问题中任一问题的点击、用户提交的满意度信息等等。在一个实例中，客服界面中显示有满意度按钮，以通过用户的点击体现用户的满意，在用户点击对象列表中推送对象列表中第 p 个对象的情况中，可通过如下公式（4）确定模型的第 i 次决策对应的回报值 r_i ：

$$r_i = \begin{cases} 1^{p=i} \alpha^p & i \neq N \\ 1^{p=i} \alpha^p + r' & i = N \end{cases} \quad (4)$$

[82]其中，当用户点击满意度按钮时，可将 r' 设置为 1，否则将 r' 设置为 0。也就是说，当用户点击第 i 个问题，当 $i \neq N$ 时， r_i 等于 α^p ，当 $i = N$ （即模型的最后一次决策）时， r_i 等于 $\alpha^p + r'$ ；当用户未点击第 i 个问题，当 $i \neq N$ 时， r_i 等于 0，当 $i = N$ 时， r_i 等于 r' 。

[83]在通过模型的 N 次决策过程获取推送对象列表、并基于该推送对象列表获取模型的从 $i=1$ 至 $i=N$ 的各次决策的回报值 r_i 之后，可获取与该组推送对象列表对应的 N 组 (s_i, a_i, s_{i+1}, r_i) ，其中第 N 组数据中的 s_{N+1} 可基于该对象列表中的 N 个对象确定。从而，可基于上述 N 组数据训练上述强化学习模型。例如，在所述强化学习模型为 DDPG 模型的情况中，实施模型计算的神经网络除了包括上文所述的策略网络之外，还包括价值网络，从而可根据例如梯度下降法分别对所述策略网络和价值网络进行调参。例如通过 B 表示上述 N 组 (s_i, a_i, s_{i+1}, r_i) 的集合，通过 Ω 表示价值网络的参数，则可通过如下的公式（5）更新 Ω ：

$$\Omega \leftarrow \Omega - \nabla_{\Omega} \frac{1}{|B|} \sum_{(s_i, a_i, s_{i+1}, r_i) \in B} (Q_{\Omega}(s_i, a_i) - y(r_i, s_{i+1}))^2 \quad (5)$$

[84] 其中，通过如下公式（6）获取公式（5）中的 $y(r_i, s_{i+1})$ ：

$$y(r_i, s_{i+1}) = \begin{cases} r_i & i=N \\ r_i + \gamma Q_{\Omega_{tgt}}(s_{i+1}, \mu_{\Theta_{tgt}}(s_{i+1})) & i \neq N \end{cases} \quad (6)$$

[85] 其中， Ω_{tgt} 为价值网络的目标参数， Θ_{tgt} 为策略网络的目标参数， $\mu_{\Theta_{tgt}}(s_{i+1})$ 为例如

5 公式（1）所示的函数，所述 Ω_{tgt} 、 Θ_{tgt} 为基于软更新所获取的值，在上述通过公式（5）更新 Ω 之后，可通过软更新基于 Ω 更新 Ω_{tgt} 。对策略网络的目标参数 Θ_{tgt} 的更新也可通过梯度下降法基于上述 N 组数据和价值网络的输出 Q 进行，在此不再详述。

[86] 图 6 示出根据本说明书实施例的一种基于强化学习模型确定针对用户的推送对象列表的装置 6000，其中，对于第一用户，已预先通过所述方法确定有 M 组对象列表，每
10 组对象列表中当前包括 i-1 个对象，其中，M、i 都为大于等于 1 的整数，其中，i 小于等于预定整数 N，所述装置包括：

[87] 用于每组对象列表的，

[88] 第一获取单元 601，配置为，获取第 i 个状态特征向量，所述第 i 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态
15 特征包括该组对象列表中所述 i-1 个对象各自的属性特征；

[89] 输入单元 602，配置为，将所述第 i 个状态特征向量输入所述强化学习模型，以使得所述强化学习模型输出与该第 i 个状态特征向量对应的权重向量，所述权重向量包括预定数目的排序特征各自的权重；

[90] 第二获取单元 603，配置为，获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量，所述排序特征向量包括所述预定数目的排序特征各自的特征值；以及

[91] 计算单元 604，配置为，基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积，计算所述候选对象集合中各个对象的分数；以及

[92] 第一确定单元 605，配置为，对于所述 M 组对象列表，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表，其中，所
25 述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

[93] 在一个实施例中，已预先通过所述方法确定有 M 组对象列表包括，已预先通过所述方法确定有一组对象列表，其中，所述第一确定单元还配置为，基于与该组对象列表对

应的候选对象集合中各个对象的分数，以所述候选对象集合中分数最高的对象作为该组对象列表的第 i 个对象，并将该组对象列表作为更新的一组对象列表。

[94] 在一个实施例中，其中， M 大于等于 2，其中，所述第一确定单元还配置为，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，通过集束搜索算法确定更新的 M 组对象列表。

[95] 在一个实施例中， i 等于 N ，所述装置还包括，第二确定单元 606，配置为，通过集束搜索算法，从所述更新的 M 组对象列表中确定针对所述第一用户的推送对象列表。

[96] 在一个实施例中，所述装置 6000 还包括，

[97] 推送单元 607，配置为，以所述推送对象列表中各个对象的排列顺序，向所述第一用户推送所述各个对象，以获取所述第一用户的反馈；

[98] 第三获取单元 608，配置为，基于所述排列顺序和所述反馈获取 N 个回报值，所述 N 个回报值与对所述方法的从 $i=1$ 至 N 的 N 次循环分别对应；

[99] 第四获取单元 609，配置为，获取第 $N+1$ 个状态特征向量，所述第 $N+1$ 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括所述推送对象列表中 N 个对象各自的属性特征；以及

[100] 训练单元 610，配置为，基于与所述 N 次循环分别对应的 N 组数据训练所述强化学习模型，以优化所述强化学习模型，其中，所述 N 组数据包括第 1 至第 N 组数据，其中，第 i 组数据包括：与所述推送对象列表对应的第 i 个状态特征向量、与该第 i 个状态特征向量对应的权重向量、与所述推送对象列表对应的第 $i+1$ 个状态特征向量、以及第 i 次循环对应的回报值。

[101] 本说明书另一方面提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行上述任一项方法。

[102] 本说明书另一方面提供一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现上述任一项方法。

[103] 通过根据本说明书实施例的基于强化学习模型确定针对用户的推送对象列表的方案，相比于现有的点击预测分类模型具有如下优势：首先，本说明书实施例的方案在考虑用户点击率之外，还考虑了用户点击对象的位置和用户的其它反馈（如用户是否满

意等)，这些额外的信息体现在模型的回报值中；其次，根据本说明书实施例的强化学习模型以 CTR 模型打分及一些实时特征作为输入，特征空间小，模型的迭代更新可以很快，辅助不同滑动时间窗口的实时数据进行综合打分，在充分利用 ctr 模型的前提下，能及时应用环境的实时变化；最后在本说明书实施例中，模型状态中包含用户、场景及层次化的信息，从而可控制对象推送的多样性和探索性，另外，根据本说明书实施例的模型参数可根据在收集数据、用户体验和保证效果各方面的需要进行干预调节。

[104] 需要理解，本文中的“第一”，“第二”等描述，仅仅为了描述的简单而对相似概念进行区分，并不具有其他限定作用。

[105] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[106] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[107] 本领域普通技术人员应该还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执轨道，取决于技术方案的特定应用和设计约束条件。本领域普通技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[108] 结合本文中所公开的实施例描述的方法或算法的步骤可以用硬件、处理器执轨道的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器 (RAM)、内存、只读存储器 (ROM)、电可编程 ROM、电可擦除可编程 ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[109] 以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定

本发明的保护范围，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权利要求书

1. 一种基于强化学习模型确定针对用户的推送对象列表的方法，其中，对于第一用户，已预先通过所述方法确定有 M 组对象列表，每组对象列表中当前包括 $i-1$ 个对象，其中， M 、 i 都为大于等于 1 的整数，其中， i 小于等于预定整数 N ，所述方法包括：

5 对于每组对象列表，

 获取第 i 个状态特征向量，所述第 i 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括该组对象列表所述 $i-1$ 个对象各自的属性特征；

 将所述第 i 个状态特征向量输入所述强化学习模型，以使得所述强化学习模型输出
10 与该第 i 个状态特征向量对应的权重向量，所述权重向量包括预定数目的排序特征各自的权重；

 获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量，所述排序特征向量包括所述预定数目的排序特征各自的特征值；以及

 基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积，计算所
15 述候选对象集合中各个对象的分数；以及

 对于所述 M 组对象列表，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表，其中，所述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

2. 根据权利要求 1 所述的方法，其中，所述动态特征至少包括所述 $i-1$ 个对象各自的
20 以下属性特征：当前热度、对象标识、对象所属类型。

3. 根据权利要求 1 所述的方法，其中，所述 M 组对象列表中包括第一组对象列表，与该第一组对象列表对应的候选对象集合中包括第一对象，与该第一对象对应的排序特征向量至少包括以下排序特征的值：所述第一用户对该第一对象的预估点击率、该第一对象的当前热度、该第一对象相对于所述第一组对象列表中的 $i-1$ 个对象的多样性。

25 4. 根据权利要求 1 所述的方法，其中，已预先通过所述方法确定有 M 组对象列表包括，已预先通过所述方法确定有一组对象列表，其中，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表包括，基于与该组对象列表对应的候选对象集合中各个对象的分数，以所述候选对象集合中分数最高的对象作为该组对象列表的第 i 个对象，并将该组对象列表作为更新的一组对象列表。

30 5. 根据权利要求 1 所述的方法，其中， M 大于等于 2，其中，基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，确定更新的 M 组对象列表包括，

基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数，通过集束搜索算法确定更新的 M 组对象列表。

6.根据权利要求 5 所述的方法，其中，i 等于 N，所述方法还包括，通过集束搜索算法，从所述更新的 M 组对象列表中确定针对所述第一用户的推送对象列表。

5 7.根据权利要求 6 所述的方法，还包括，

以所述推送对象列表中各个对象的排列顺序，向所述第一用户推送所述各个对象，以获取所述第一用户的反馈；

基于所述排列顺序和所述反馈获取 N 个回报值，所述 N 个回报值与对所述方法的从 i=1 至 N 的 N 次循环分别对应；

10 获取第 N+1 个状态特征向量，所述第 N+1 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括所述推送对象列表中 N 个对象各自的属性特征；以及

15 基于与所述 N 次循环分别对应的 N 组数据训练所述强化学习模型，以优化所述强化学习模型，其中，所述 N 组数据包括第 1 至第 N 组数据，其中，第 i 组数据包括：与
15 所述推送对象列表对应的第 i 个状态特征向量、与该第 i 个状态特征向量对应的权重向量、与
20 所述推送对象列表对应的第 i+1 个状态特征向量、以及与第 i 次循环对应的回报值。

8.根据权利要求 7 所述的方法，其中，所述对象为询问问题，对于第 1 至 N-1 次循环中的第 i 次循环，与
20 所述第 i 次循环对应的回报值基于所述第一用户的如下反馈获取：
20 是否点击所述推送对象列表中的第 i 个问题。

9.根据权利要求 8 所述的方法，与
25 所述第 N 次循环对应的回报值基于所述第一用户的如下反馈获取：是否点击所述推送对象列表中的第 N 个问题、以及提交的满意度信息。

10.根据权利要求 7 所述的方法，其中，所述强化学习模型为基于深度确定策略梯度
25 算法的模型。

11. 一种基于强化学习模型确定针对用户的推送对象列表的装置，其中，对于第一用户，已预先通过所述方法确定有 M 组对象列表，每组对象列表中当前包括 i-1 个对象，其中，M、i 都为大于等于 1 的整数，其中，i 小于等于预定整数 N，所述装置包括：

用于每组对象列表的，

30 第一获取单元，配置为，获取第 i 个状态特征向量，所述第 i 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特

征包括该组对象列表中所述 $i-1$ 个对象各自的属性特征;

输入单元, 配置为, 将所述第 i 个状态特征向量输入所述强化学习模型, 以使得所述强化学习模型输出与该第 i 个状态特征向量对应的权重向量, 所述权重向量包括预定数目的排序特征各自的权重;

5 第二获取单元, 配置为, 获取与该组对象列表对应的候选对象集合中各个对象的排序特征向量, 所述排序特征向量包括所述预定数目的排序特征各自的特征值; 以及

计算单元, 配置为, 基于所述候选对象集合中各个对象的排序特征向量与所述权重向量的点积, 计算所述候选对象集合中各个对象的分数; 以及

10 第一确定单元, 配置为, 对于所述 M 组对象列表, 基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数, 确定更新的 M 组对象列表, 其中, 所述更新的 M 组对象列表中的每组对象列表包括 i 个对象。

12. 根据权利要求 11 所述的装置, 其中, 所述动态特征至少包括所述 $i-1$ 个对象各自的以下属性特征: 当前热度、对象标识、对象所属类型。

15 13. 根据权利要求 11 所述的装置, 其中, 所述 M 组对象列表中包括第一组对象列表, 与该第一组对象列表对应的候选对象集合中包括第一对象, 与该第一对象对应的排序特征向量至少包括以下排序特征的值: 所述第一用户对该第一对象的预估点击率、该第一对象的当前热度、该第一对象相对于所述第一组对象列表中的 $i-1$ 个对象的多样性。

20 14. 根据权利要求 11 所述的装置, 其中, 已预先通过所述方法确定有 M 组对象列表包括, 已预先通过所述方法确定有一组对象列表, 其中, 所述第一确定单元还配置为, 基于与该组对象列表对应的候选对象集合中各个对象的分数, 以所述候选对象集合中分数最高的对象作为该组对象列表的第 i 个对象, 并将该组对象列表作为更新的一组对象列表。

25 15. 根据权利要求 11 所述的装置, 其中, M 大于等于 2, 其中, 所述第一确定单元还配置为, 基于与所述 M 组对象列表分别对应的 M 个候选对象集合中各个对象的分数, 通过集束搜索算法确定更新的 M 组对象列表。

16. 根据权利要求 15 所述的装置, 其中, i 等于 N , 所述装置还包括, 第二确定单元, 配置为, 通过集束搜索算法, 从所述更新的 M 组对象列表中确定针对所述第一用户的推送对象列表。

30 17. 根据权利要求 16 所述的装置, 还包括, 推送单元, 配置为, 以所述推送对象列表中各个对象的排列顺序, 向所述第一用户推送所述各个对象, 以获取所述第一用户的反馈;

第三获取单元，配置为，基于所述排列顺序和所述反馈获取 N 个回报值，所述 N 个回报值与对所述方法的从 $i=1$ 至 N 的 N 次循环分别对应；

第四获取单元，配置为，获取第 $N+1$ 个状态特征向量，所述第 $N+1$ 个状态特征向量包括静态特征和动态特征，其中，所述静态特征包括所述第一用户的属性特征，所述动态特征包括所述推送对象列表中 N 个对象各自的属性特征；以及

训练单元，配置为，基于与所述 N 次循环分别对应的 N 组数据训练所述强化学习模型，以优化所述强化学习模型，其中，所述 N 组数据包括第 1 至第 N 组数据，其中，第 i 组数据包括：与所述推送对象列表对应的第 i 个状态特征向量、与该第 i 个状态特征向量对应的权重向量、与所述推送对象列表对应的第 $i+1$ 个状态特征向量、以及与第 i 次循环对应的回报值。

18.根据权利要求 17 所述的装置，其中，所述对象为询问问题，对于第 1 至 $N-1$ 次循环中的第 i 次循环，与所述第 i 次循环对应的回报值基于所述第一用户的如下反馈获取：是否点击所述推送对象列表中的第 i 个问题。

19.根据权利要求 18 所述的装置，与所述第 N 次循环对应的回报值基于所述第一用户的如下反馈获取：是否点击所述推送对象列表中的第 N 个问题、以及提交的满意度信息。

20.根据权利要求 17 所述的装置，其中，所述强化学习模型为基于深度确定策略梯度算法的模型。

21. 一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-10 中任一项的所述的方法。

22. 一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-10 中任一项所述的方法。

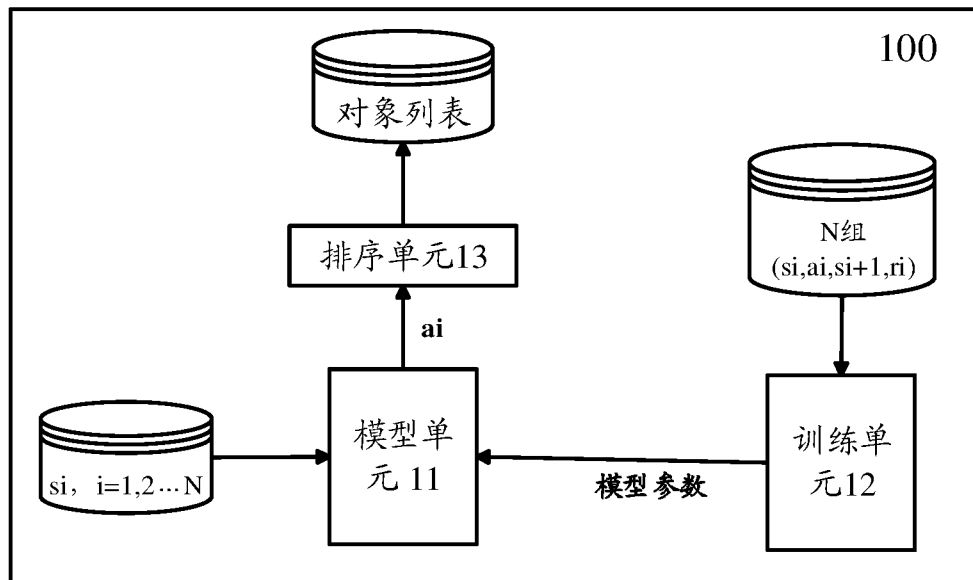


图 1

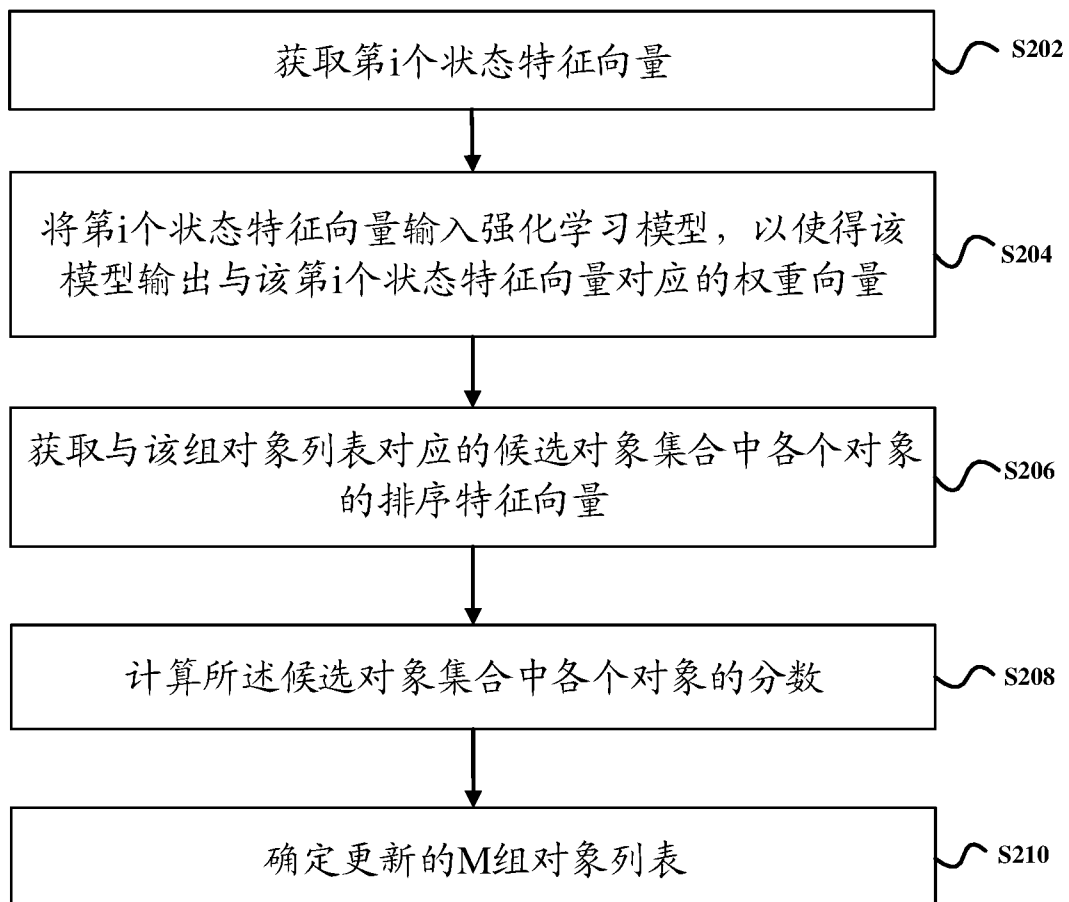


图 2

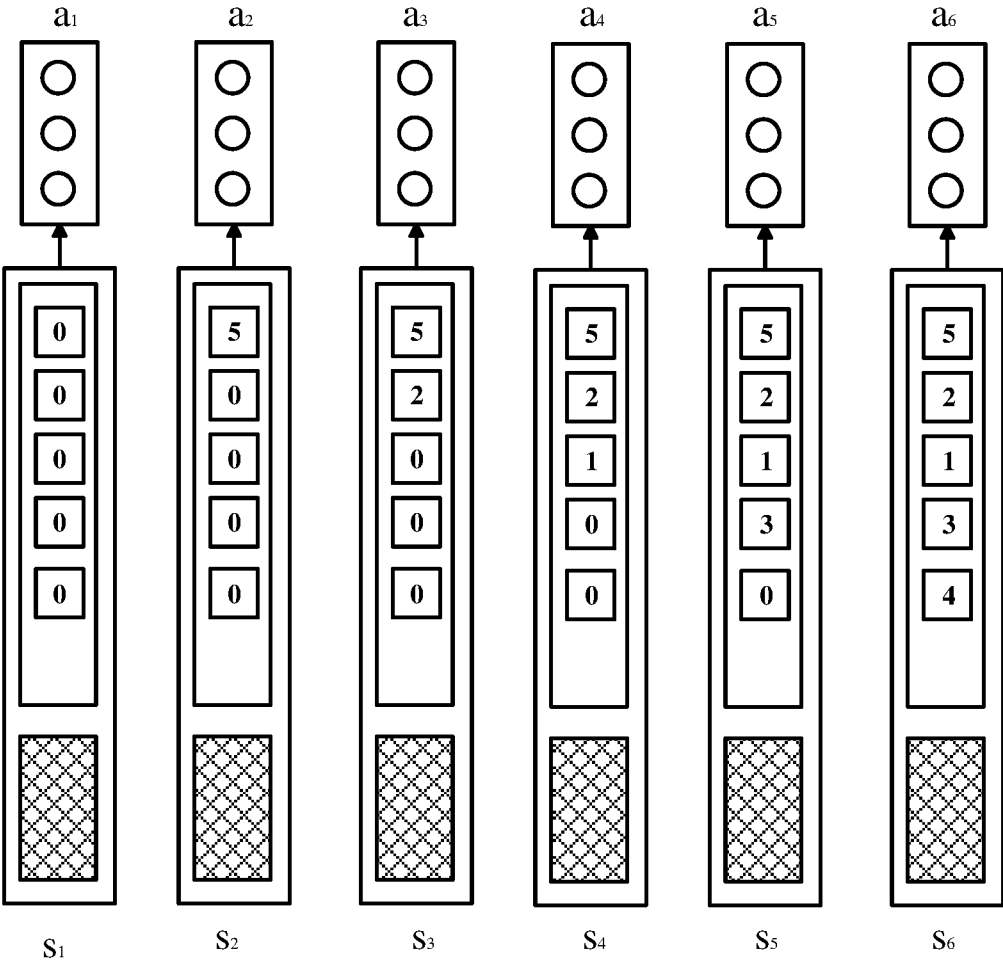


图 3

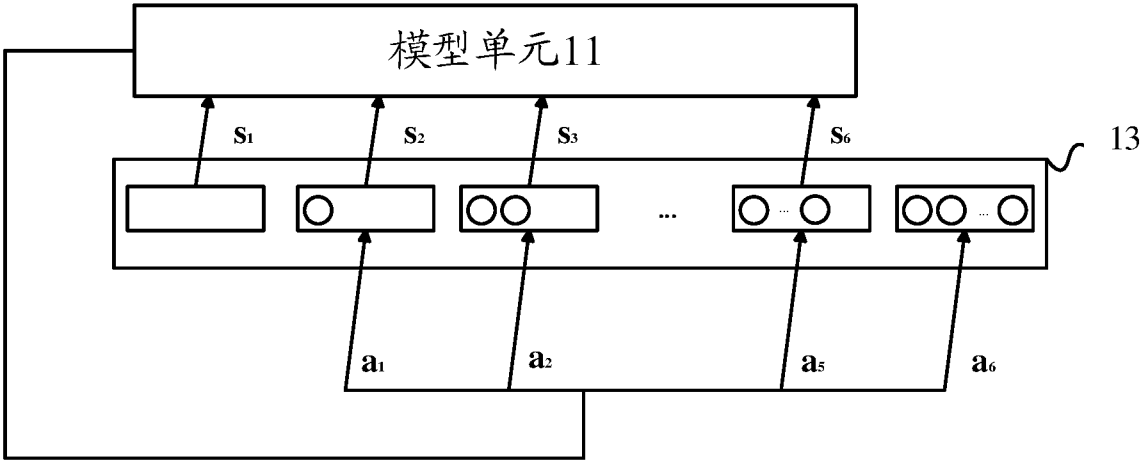


图 4

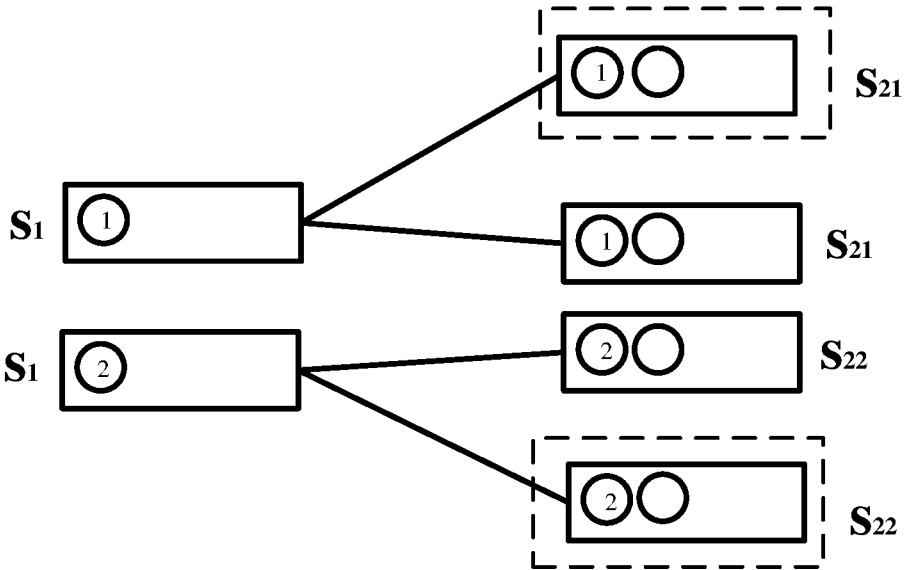


图 5

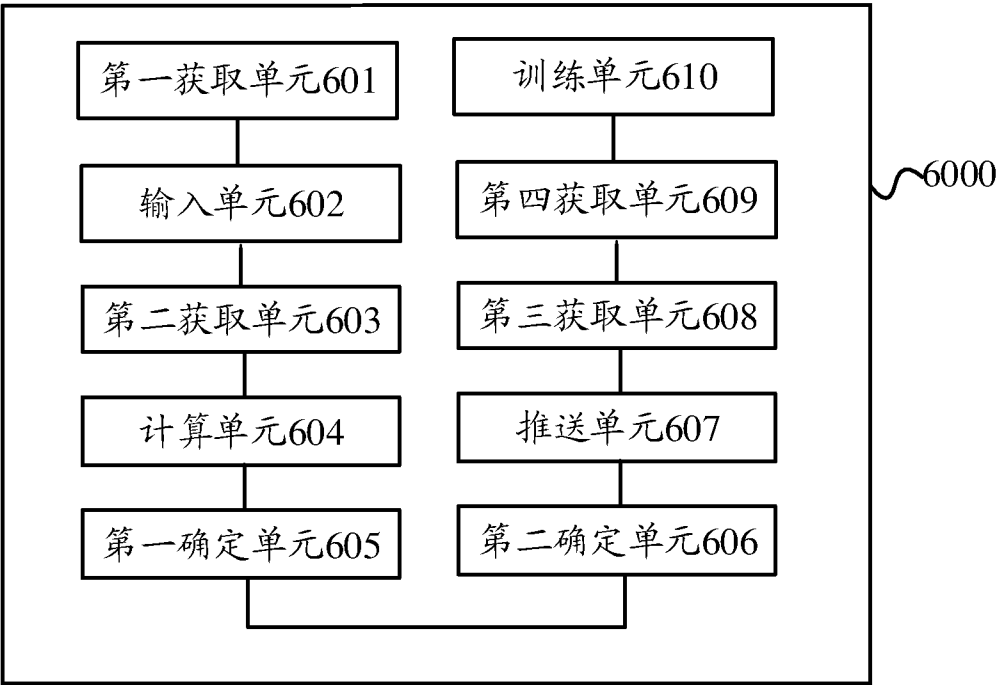


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/071699

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/9535(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI: 学习模型, 列表, 对象, 权重, 向量, 更新, learning w model, list, object, weight, vector, updat+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110263245 A (ALIBABA GROUP HOLDING LIMITED) 20 September 2019 (2019-09-20) claims 1-22	1-22
A	CN 108304440 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 20 July 2018 (2018-07-20) description, paragraphs 0003-0036	1-22
A	CN 104869464 A (HISENSE GROUP CO., LTD.) 26 August 2015 (2015-08-26) entire document	1-22
A	CN 108805594 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD., et al.) 13 November 2018 (2018-11-13) entire document	1-22
A	US 2014297476 A1 (ALIBABA GROUP HOLDING LIMITED) 02 October 2014 (2014-10-02) entire document	1-22



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

18 March 2020

Date of mailing of the international search report

13 April 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/071699

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110263245	A	20 September 2019	None			
CN	108304440	A	20 July 2018	None			
CN	104869464	A	26 August 2015	CN	104869464	B	18 September 2018
CN	108805594	A	13 November 2018	None			
US	2014297476	A1	02 October 2014	JP	6152173	B2	21 June 2017
				TW	201437933	A	01 October 2014
				WO	2014160648	A1	02 October 2014
				TW	I591556	B	11 July 2017
				JP	2016511906	A	21 April 2016
				US	9818142	B2	14 November 2017
				CN	104077306	A	01 October 2014
				CN	104077306	B	11 May 2018

国际检索报告

国际申请号

PCT/CN2020/071699

A. 主题的分类

G06F 16/9535 (2019. 01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

WPI, EPDOC, CNPAT, CNKI: 学习模型, 列表, 对象, 权重, 向量, 更新, learning w model, list, object, weight, vector, updat+

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 110263245 A (阿里巴巴集团控股有限公司) 2019年 9月 20日 (2019 - 09 - 20) 权利要求1-22	1-22
A	CN 108304440 A (腾讯科技深圳有限公司) 2018年 7月 20日 (2018 - 07 - 20) 说明书第0003-0036段	1-22
A	CN 104869464 A (海信集团有限公司) 2015年 8月 26日 (2015 - 08 - 26) 全文	1-22
A	CN 108805594 A (北京京东尚科信息技术有限公司 等) 2018年 11月 13日 (2018 - 11 - 13) 全文	1-22
A	US 2014297476 A1 (阿里巴巴集团控股有限公司) 2014年 10月 2日 (2014 - 10 - 02) 全文	1-22

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 3月 18日

国际检索报告邮寄日期

2020年 4月 13日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

焦永涵

电话号码 86-(10)-53961463

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/071699

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110263245	A	2019年 9月 20日	无			
CN	108304440	A	2018年 7月 20日	无			
CN	104869464	A	2015年 8月 26日	CN	104869464	B	2018年 9月 18日
CN	108805594	A	2018年 11月 13日	无			
US	2014297476	A1	2014年 10月 2日	JP	6152173	B2	2017年 6月 21日
				TW	201437933	A	2014年 10月 1日
				WO	2014160648	A1	2014年 10月 2日
				TW	1591556	B	2017年 7月 11日
				JP	2016511906	A	2016年 4月 21日
				US	9818142	B2	2017年 11月 14日
				CN	104077306	A	2014年 10月 1日
				CN	104077306	B	2018年 5月 11日