

利用深度强化学习进行推荐营销的方法及装置

技术领域

[01] 本说明书一个或多个实施例涉及机器学习领域，尤其涉及利用深度强化学习进行推荐营销的方法及装置。

5 背景技术

[02] 机器学习的迅猛发展使得各种机器学习的模型已经应用在各种各样的业务场景中，例如用于评估用户的信用风险，用于向用户推荐合适的内容，或向用户推送合适的营销信息，用于预测广告的投放效果等。

10 [03] 实际上，一项业务场景的完整业务流程经常包含多个阶段或多个环节，其中涉及诸多因素。例如，在推荐营销的业务场景中，业务至少可以包括以下环节：营销信息的选择和推送，推送之后用户的反馈和追踪。在这些环节中，又涉及诸多因素和参数，例如营销的渠道，推送方式，触达率，转化率等等。因此，在试图将机器学习应用到业务场景中时，面对复杂的业务场景中的诸多环节和诸多因素，如何设计和选择适用的模型，引入何种因素衡量业务的执行状况，都是需要考虑的问题。

15 [04] 因此，希望能有改进的方案，更加高效地用机器学习提高业务执行效果。

发明内容

[05] 本说明书一个或多个实施例描述了一种利用深度强化学习进行推荐营销的方法和装置，可以综合全面地对推荐营销业务中的多个环节多个因素进行学习，以综合全面地促进业务效果。

20 [06] 根据第一方面，提供了一种利用深度强化学习进行推荐营销的方法，所述方法包括：

[07] 获取执行环境的状态信息，所述状态信息至少包括当前用户的用户信息；

[08] 根据营销策略，针对所述状态信息确定相应的营销行为，所述营销行为包括营销渠道、营销内容和营销时段的组合；

[09] 获取执行环境针对所述营销行为的当前奖励分数；

25 [10] 至少根据所述当前奖励分数更新所述营销策略。

- [11]根据一种实施方式，所述用户信息包括，用户基本属性信息，以及用户历史行为信息。
- [12]进一步地，在一个实施例中，用户历史行为信息包括以下中的至少一项：历史操作序列、最近浏览页面列表、最近接受的营销信息列表、消费记录、交易记录、贷款记录、
- 5 出行记录、保险记录。
- [13]在一种可能的设计中，用户信息包括用户画像信息，所述用户画像信息基于用户基本属性信息和/或用户历史行为信息预先确定。
- [14]在一个实施例中，所述状态信息还包括，所述当前用户所处场景的场景信息。
- [15]进一步地，在一个例子中，场景信息包括，位置服务场景信息，支付场景信息，钱
- 10 包操作事件信息。
- [16]在一种可能的设计中，营销内容包括营销消息和利益包，所述利益包包括以下中的一种或多种：红包、优惠券、权益。
- [17]根据一种实施方式，营销渠道包括以下中的一种或多种：消息推送，卡片展示，角标提醒，红点提醒。
- 15 [18]在一个实施例中，方法还包括，至少基于所述当前奖励分数确定回报价值，所述回报价值包括所述当前奖励分数和至少一项未来奖励分数的加权求和；并且更新营销策略的步骤包括，确定后续营销策略，使得所述回报价值最大化。
- [19]根据第二方面，提供一种利用深度强化学习进行推荐营销的装置，所述装置包括：
- [20]状态获取单元，配置为获取执行环境的状态信息，所述状态信息至少包括当前用户
- 20 的用户信息；
- [21]行为确定单元，配置为根据营销策略，针对所述状态信息确定相应的营销行为，所述营销行为包括营销渠道、营销内容和营销时段的组合；
- [22]分数获取单元，配置为获取执行环境针对所述营销行为的当前奖励分数；
- [23]更新单元，配置为至少根据所述当前奖励分数更新所述营销策略。
- 25 [24]根据第三方面，提供了一种计算设备，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，实现第一方面的方法。
- [25]通过本说明书实施例提供的方法和装置，全面获取营销环境的当前状态，不仅观察

当前用户的用户信息，可选地还观察用户所处的场景，基于全面的营销环境状态，确定营销行为。营销行为的确定也综合考虑多种因素，包括营销渠道的选择，营销内容的选择，营销时段的确定，至少基于渠道、营销内容、营销时段这三者的组合，确定营销行为。如此使得，智能体确定的营销行为同时考虑到营销过程中的多个因素，也就是深度强化学习系统同时对营销过程中多个环节多个因素进行学习，从而综合全面地学习推荐营销的整个业务流程和业务目标，更好地促进业务执行效果。

附图说明

[26]为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

- [27]图 1 本说明书披露的一个实施例的推荐营销的执行示意图；
- [28]图 2 示出深度强化学习系统的示意图；
- [29]图 3 示出根据一个实施例利用深度强化学习进行推荐营销的方法；
- [30]图 4 示出多种营销渠道的例子；
- [31]图 5 示出根据一个实施例的深度强化学习推荐营销的总体示意图；
- [32]图 6 示出根据一个实施例的推荐营销系统架构示意图；
- [33]图 7 示出根据一个实施例的装置的示意性框图。

具体实施方式

- [34]根据本说明书的一个或多个实施例，采用深度强化学习的方式，来促进推荐营销的业务效果。下面结合附图，对本说明书提供的方案进行描述。
- [35]图 1 为本说明书披露的一个实施例的推荐营销的执行示意图。在推荐营销的场景中，营销最终目标的实现可能需要通过多次营销，变换各种营销方式来完成。例如，对于某个用户 A，首先，为其推送一条营销消息，但是用户没有响应。接着，通过渠道 1 为其发送一种利益包，例如优惠券，用户仍然没有响应。然后，通过另一种渠道，例如渠道 2，为其发送另一种利益包，例如红包，用户对营销内容进行了点击，但是没有签约转

化。接下来，在特定场景下，用特定渠道发出类似的红包，用户进行了签约转化。

[36]通过以上过程可以看到，推荐营销是一个复杂的过程，其完整业务流程包含多个环节多次交互，例如，至少包括以下环节：营销信息的选择和推送，推送之后用户的反馈和追踪。在这些环节中，又涉及诸多因素和参数，例如营销的渠道，推送方式，推送内容，推送场景，推送时间等等。

[37]在一种实施方案中，考虑到交互过程中多个环节多个因素对最终营销目标的影响，针对各个环节的各个因素分别进行建模学习。例如，为了考虑渠道差异化对用户转化的影响，建立并训练渠道选择模型，来发掘用户对不同渠道的感知度。为了考虑营销内容（又称为 offer，即各种提供给用户的营销信息）对用户转化的影响，建立并训练 offer 差异选择模型，来发掘不同用户对不同 offer 种类的兴趣度。考虑到不同用户使用 app 的习惯不同，建立并训练时段选择模型，以进行时段差异化营销。考虑到营销目标往往通过多个环节多次交互来达成，建立时序化的系列营销模型（例如通过长短期记忆 LSTM 网络），以追踪多次营销对用户的系列影响。利用这样的方案，可以针对各个环节、各个因素都进行有针对性的学习。

[38]然而，在这样的方案中，需要对各个模型分别进行设计、训练、评估，成本很高。如果要估计测算业务的总体执行效果，就需要利用多个模型逐一预测，再对预测结果进行融合，预测过程的复杂度和运算代价也会比较高。如果更换业务内容，通常需要对各个模型都重新进行训练。

[39]另一方面，在这样的方案中，任何一个模型出现问题都可能会影响最终营销效果。而实际上，多个单模型中常常会有一部分模型的预测效果并不尽如人意。例如，在一个例子中，利用深度神经网络 DNN 建立的 offer 差异选择模型，将 offer 划分为 27 类，进行多分类学习。从针对该模型测得的宏平均和混淆矩阵来看，预测准确度有待进一步提高。

[40]在此基础上，本说明书实施例还提出一种方案，利用深度强化学习系统来进行推荐营销，以进一步提升业务效果。

[41]图 2 示出深度强化学习系统的示意图。一般地，深度强化学习系统包括智能体和执行环境，智能体通过与执行环境的交互和反馈，不断进行学习，优化其策略。具体而言，智能体观察并获得执行环境的状态（state），根据一定策略，针对当前执行环境的状态确定要采取的行为或动作（action）。这样的行为作用于执行环境，会改变执行环境的

状态，同时产生一个反馈给智能体，该反馈又称为奖励分数（reward）。智能体根据获得的奖励分数来判断，之前的行为是否正确，策略是否需要调整，进而更新其策略。通过反复不断地观察状态、确定行为、收到反馈，使得智能体可以不断更新策略，最终目标是能够学习到一个策略，使得获得的奖励分数累积最大化。这就是典型的强化学习过程。在学习和调整策略过程中，如果智能体采取的是包括神经网络在内的一些深度学习的算法，那么这样的系统则称为深度强化学习系统。

[42]根据本说明书的实施例，将深度强化学习应用到推荐营销系统中。在推荐营销的场景下，执行环境即为推荐营销环境，相应地，智能体观察营销环境的状态，基于一些营销策略，针对营销环境的当前状态确定适合的营销行为。该营销行为进而改变营销环境的状态，并产生一个奖励分数反馈到智能体。智能体进而可以根据奖励分数调整、更新其营销策略。根据本说明书一个或多个实施例，全面获取营销环境的当前状态，不仅观察当前用户的用户信息，可选地还观察用户所处的场景，基于全面的营销环境状态，确定营销行为。营销行为的确定也综合考虑多种因素，包括营销渠道的选择，营销内容的选择，营销时段的确定，至少基于渠道、营销内容、营销时段这三者的组合，确定营销行为。如此使得，智能体确定的营销行为同时考虑到营销过程中的多个因素，也就是深度强化学习系统同时对营销过程中多个环节多个因素进行学习，从而综合全面地学习推荐营销的整个业务流程和业务目标，更好地促进业务执行效果。下面描述以上构思的实现方式。

[43]图3示出根据一个实施例利用深度强化学习进行推荐营销的方法。可以理解，该方法可以通过如图2所示的深度强化学习推荐营销系统来实现，该系统包括智能体和执行环境，所述智能体用于根据营销策略，针对执行环境的状态确定营销行为，并根据针对该营销行为反馈得到的奖励分数更新营销策略。更具体地，该方法可以由深度强化学习系统中的智能体来执行。从物理实体的角度来说，上述智能体可以由任何具有数据计算、处理能力的装置、设备、平台、设备集群等来实现或执行。

[44]如图3所示，该方法至少包括：步骤31，获取执行环境的状态信息，所述状态信息至少包括当前用户的用户信息；步骤33，根据营销策略，针对所述状态信息确定相应的营销行为，所述营销行为包括营销渠道、营销内容和营销时段的组合；步骤35，获取执行环境针对所述营销行为的当前奖励分数；以及步骤37，至少根据所述当前奖励分数更新所述营销策略。下面描述以上各个步骤的具体执行方式。

[45]首先，在步骤31，获取执行环境的状态信息。如前所述，在利用深度强化学习的推

荐营销系统中，智能体不断观察和获取推荐营销环境的当前状态，用以确定应该采取的营销行为。在一种实施方式下，智能体获取的营销环境的状态信息可以包括，当前用户的用户信息。在一个实施例，用户信息可以包括用户基本属性特征，例如注册时长、年龄、性别、收入、资产等等。在一个实施例，用户信息还包括用户历史行为信息，
5 例如历史操作序列、最近浏览页面列表、最近接受的营销信息列表、最近的消费记录、最近的交易记录、最近的贷款记录、最近的出行记录、最近的保险记录等等。

[46]在一个实施例，用户信息还可以包括用户画像信息。可以理解，在一些实施方式中，已经根据一些算法，基于用户的基本属性信息和/或历史行为信息，为用户赋予了对应的用户画像。用户画像信息例如可以包括，基于基本属性信息将用户划分到一定人群
10 的人群标签，对用户进行聚类的类簇标签，基于用户历史行为构建的用户行为习惯标签，等等。上述各种用户信息可以构成当前用户的用户特征。

[47]在一个实施例，上述深度强化学习系统实时或在线地进行训练和决策。在这样的情况下，步骤 31 获取的执行环境的状态信息还可以包括，当前用户所处场景的场景信息。在一个例子中，场景信息包括以下中的至少一种，位置服务场景信息，支付场景信息，
15 钱包操作事件信息等。

[48]具体地，位置服务场景信息例如是基于位置的服务 LBS（location based service）所提供的场景信息。LBS 服务是通过无线电通讯网络或外部定位方式，获取移动终端用户的位置信息的服务。基于位置信息可以进一步确定用户所处的场景。例如，在基于 LBS 确定用户位于电影院的情况下，位置服务场景信息可以是电影场景的信息，在基于 LBS
20 确定用户位于商场的情况下，位置服务场景信息可以是购物场景的信息，在基于 LBS 确定用户位于美食街的情况下，位置服务场景信息可以是美食场景的信息。

[49]支付场景是另一种用户常用的场景。支付场景信息可以进一步包括，支付所涉及的交易双方的信息，交易对象（例如商品）的信息等。

[50]钱包操作事件可以包括用户在电子钱包中进行的各种操作，例如查看余额，绑定银行卡，转账，购买理财等等。
25

[51]以上这些场景信息都可以作为推荐营销环境的状态信息被提取，提供给智能体。

[52]接着，在步骤 33，智能体根据营销策略，针对以上获取的状态信息确定相应的营销行为，所述营销行为包括营销渠道、营销内容和营销时段的组合。

[53]可以理解，对于强化学习系统来说，智能体根据策略将环境的状态映射为接下来的

行为。在图 2 所示的深度强化学习推荐营销系统中，智能体是进行决策的实体，根据营销策略，将当前营销环境状态映射为接下来采取的营销行为。其中的营销策略通过复杂的深度学习算法来实现，例如 Q-learning, DQN 等，这些算法中通常包括一些深度神经网络的结构。如上所述，营销环境的当前状态包括了当前用户的用户特征，可选的，还包括用户所处的场景信息，一旦获取到这样的当前状态信息，智能体就通过深度学习算法确定应该针对当前用户采取怎样的营销行为。营销行为的确定至少包括选择和确定营销渠道、营销内容和营销时段，至少基于这三个因素的组合形成营销行为。换言之，智能体在获取到用户信息（以及可选的场景信息）的情况下，根据营销策略确定，针对（该场景下的）当前用户，应该在什么样时机，用什么渠道，向用户推荐什么内容的营销信息。

[54] 在一个实施例中，确定营销行为包括，确定营销渠道。可以理解，存在多种向用户传达营销信息的方式，这些方式都可以称为营销渠道。典型地，营销渠道可以包括消息推送，卡片展示，角标提醒，红点提醒等等。

[55] 图 4 示出多种营销渠道的例子。在图 4 的示例中，在进入 app 之后，页面中 101 部分以一整个区块作为卡片的样式进行营销展示，即为卡片展示，又可以称为广告 banner，或者腰封广告；102 指示出在图标 的 右 上 角 添 加 角 标 作 为 营 销 提 醒；103 指示出在图标的 右 上 角 添 加 红 点 作 为 营 销 提 醒。图 4 仅仅示例了几种常见的营销渠道，营销渠道并不限于此，例如还包括消息推送，消息推送可以是 通过手机短信给用户发送营销内容的短消息，也可以包括由 app 给安装 app 的终端推送提醒消息。

[56] 智能体通过深度学习的算法，从例如以上所示的各种营销渠道中，选择当前环境状态所适用的营销渠道。

[57] 在一个实施例中，确定营销行为还包括，确定营销内容。可以理解，营销信息可以表现为多种营销内容，例如包括营销消息和利益包。营销消息可以是不含有利益成本的消息，例如一些通知消息，比如“蚂蚁森林新增好友交互种树功能”，“支付宝推出老用户专享服务”，等等。利益包则是含有一些利益成本的营销形式，例如包括红包、优惠券、权益等等。这其中，红包可以包括各种场景下可以使用或者抵扣的资金金额；优惠券又包括，例如立减券，折扣券，一定场景下才能使用的场景卡券（如电影券，美食券），免息券，等等；而权益可以包括其他利益相关的优惠，例如保险权益，理财增值权益等等。

[58] 智能体通过深度学习的算法，从例如以上所示的各种营销内容中，选择当前环境状

态所适用的营销内容。

[59]在一个实施例中，确定营销行为还包括，确定营销时段。可以理解，不同用户具有不同的 app 使用习惯，这些使用习惯可以体现为，在不同时段使用 app 的频次、保持注意力的时长等有所不同。同一用户在不同时段对于营销信息的敏感度和关注度也不同。

- 5 因此，根据一种实施方式，从多个维度对营销时段进行划分。例如，在一个例子中，从每周每天的时间尺度，根据工作时间将营销时段划分为工作日（例如周一到周五）和非工作日（例如周六和周日）。在一个例子中，在一天之内小时的尺度上，根据工作时间将一天的营销时段划分为工作时间（例如早 9 点到晚 6 点）和非工作时间。在另一例子中，根据平均的就餐时间，将一天划分为就餐时段和其他时段；进一步地，还可以将其
- 10 他时段划分为早上（早餐前）、上午（早餐到午餐之间）、下午（午餐和晚餐之间）和晚上（晚餐之后的时段）。可以理解，营销时段还可以采用其他方式进行划分。一般地，营销时段的划分可以基于用户与营销操作有关的行为在时间上的分布或关联统计而进行。

- [60]智能体通过深度学习的算法，从通过各种方式、在各种维度上划分的多个营销时段
- 15 中，选择当前环境状态所适用的营销时段。

[61]如此，智能体根据所学习和训练的营销策略，确定当前环境状态适用的营销渠道、营销内容和营销时段，基于这三者的组合，确定出接下来采取的营销行为。

[62]在这样的营销行为作用于营销环境之后，在步骤 35，可以获取执行环境针对所述营销行为的当前奖励分数。

- 20 [63]在一个实施例中，当前奖励分数根据用户对营销行为的响应结果而确定，例如，用户是否接收到该营销信息（即是否触达用户），用户是否点击，是否转化，是否将营销信息中推荐的内容设为首页或设为首页，等等。如此，通过奖励分数反映要优化的多个效果目标。

- [64]然后在步骤 37，至少根据当前奖励分数更新营销策略。如此，在各个环节综合考虑
- 25 多个因素，来学习、训练和优化营销策略。

[65]图 5 示出根据一个实施例的深度强化学习推荐营销的总体示意图。如图 5 所示，根据本说明书的实施例，深度强化学习系统综合观察和获取执行环境的状态信息，获取更加全面的状态特征，这些特征包括用户的全面用户特征，以及用户所处场景的场景特征。

[66]接着，利用深度学习和训练的营销策略，将以上的全面状态特征映射为营销行为，

营销行为涉及各种待决策的营销因素的叉乘组合，例如是营销渠道、营销内容和营销时段这三者的组合。营销渠道可以选自消息推送，卡片展示，角标提醒，红点提醒等等，营销内容可以选自营销消息、红包、优惠券、权益等等，营销时段可以选自工作时间、非工作时间、就餐时间等等。通过营销渠道、营销内容和营销时段的选择和组合，得出

5 针对环境状态信息的营销行为。

[67]这样的营销行为作用于执行环境之后，产生奖励分数。奖励分数可以基于用户对营销行为的多个要优化目标的响应结果而确定。

[68]如此，在利用深度强化学习进行推荐营销的过程中，考虑营销环境的全面特征，在进行决策时综合考虑营销涉及的多个因素，从而全面地学习、训练和优化营销策略，更

10 加全面有效地促进营销业务的业务效果。

[69]在一个实施例中，在获取到当前奖励分数的基础上，即在步骤 35 之后，在步骤 37 之前，基于所获取的当前奖励分数确定回报价值；从而在步骤 37，基于回报价值优化营销策略。可以理解，在深度强化学习系统中，智能体不断对策略进行更新和优化，其目的是使得，奖励分数的累积达到最大。也就是说，在序列学习的情况下，不仅要考虑当前奖励分数，还要考虑多次交互中奖励分数的累积值。为此，可以定义回报价值，来评

15 估奖励分数的预期累积。

[70]在以上实施例的深度强化学习推荐营销的场景下，可以将回报价值定义为，包括当前奖励分数和至少一项未来奖励分数的加权求和，其中各项奖励分数是执行环境针对当前营销行为的反馈。

20 [71]在一个具体例子中，回报价值定义为：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

[72]其中 G_t 为回报价值， R_{t+k+1} 是在 t 时刻执行一组行为 action 后，在 $t+k+1$ 时刻的奖励分数， γ 为折扣系数，可以起到权重的作用。如此， R_{t+1} 表示即时奖励，也就是前述的当前奖励分数，后续的 R_{t+k+1} (k 不为 0 时) 则表示预期未来奖励分数。当 γ 为 0 时，

25 表示只考虑当前奖励分数，而不考虑长期奖励； γ 不为 0 时，表示将长期奖励分数纳入考虑范围。一般地， $0 < \gamma < 1$ ， γ 值越大，表示对未来奖励分数越看中，另一方面，由于的指数作用，越是远期的奖励分数，对回报价值的影响越小。

[73]对于推荐营销的场景来说，仍然可以采用以上的回报价值的定义，来确定一项营销

行为对未来的影响。如前所述，所确定的营销行为至少包括营销渠道、营销内容和营销时段的组合，其中每个因素都会影响到当前奖励分数和未来奖励分数。

[74]对于以上的回报价值定义，通过求解贝尔曼 bellman 方程，来优化营销策略，优化目标是使得回报价值最大化。求解 bellman 方程的过程可以包括，例如值迭代、策略迭代、Q-learning、Sarsa 等。

5 [75]由于营销策略的每一步更新优化是基于回报价值，而回报价值中考虑了长期未来奖励分数，因此，如此训练得到的营销策略实际上考虑了营销的时序，从而实现了系列营销。例如，对于图 1 所示的多次营销，通过回报价值的学习和营销策略的优化，可以利用营销策略更快预测出用户可能接受签约的场景、渠道、营销内容，从而更快实现用户
10 转化，促进业务效果的提升。

[76]图 6 示出根据一个实施例的推荐营销系统架构示意图。如图 6 所示，该系统框架至少包括，环境数据库 601，流式计算模块 602，决策引擎 603，深度强化学习营销模型 604，投放平台 605。根据该架构，通过流式计算模块 602 从记录营销环境的数据库 601 读取营销环境的状态信息。记录营销环境的数据库 601 可以采用各种常见的数据库形式，
15 例如 ODPS，OSS，Hbase 等，流式计算模块 602 可以对接上述各种形式的数据库，实时读取营销环境状态数据。在离线训练和预测的情况下，流式计算模块 602 可以扫描数据库中各个用户的用户信息作为上述环境状态数据。在在线或实时训练和预测的情况下，流式计算模块 602 还可以获取用户所处场景的场景特征作为上述环境状态数据。

[77]流式计算模块 602 将读取的环境状态数据发送到决策引擎 603。决策引擎 603 利用
20 这样的环境状态数据，调用深度强化学习营销模型 604，也就是前述的深度强化学习系统中智能体所训练的营销策略模型。于是，营销策略模型针对环境状态数据，确定出应当采取的营销行为，该营销形式至少包括营销渠道、营销内容和营销时段的组合。决策引擎 603 接收到模型返回的营销行为，就将该营销行为数据发送到投放平台 605 进行投放。

25 [78]一般地，投放平台 605 预先设置有一些投放规则和策略，例如投放密度、疲劳度控制，营销竞争规则等等。在满足投放规则和策略的基础上，投放平台 605 根据营销行为中的营销渠道、营销内容和营销时段的具体设定，来投放相应营销内容。例如，如果营销行为中的营销渠道被选择为消息推送，且时段不限，那么投放平台可以立即进行相应营销内容的消息推送；如果营销渠道为卡片展示，则可以将卡片内容关联到相应用户，
30 使得该用户下次打开 app 时就可以看到该卡片展示。

[79]这样的营销行为作用于用户之后，产生营销效果。这一方面使得环境状态进一步更新，也就是用于更新环境数据库 601；另一方面，营销效果可以作为奖励分数或回报价值的评估基础，返回给深度强化学习营销模型 604，使得营销策略模型进一步优化和更新。

- 5 [80]如上所述，在本说明书的一个或多个实施例中，利用深度强化学习系统进行推荐营销。在此过程中，智能体获取营销环境的全面特征，在进行决策时综合考虑营销涉及的多个因素，从而全面地学习、训练和优化营销策略，更加全面有效地促进营销业务的业务效果。

- 10 [81]另一方面，本说明书实施例还提供一种装置，该装置应用于通过深度强化学习进行推荐营销的系统，其中深度强化学习系统如图 2 所示包括智能体和执行环境，所述智能体用于根据营销策略，针对所述执行环境的状态信息确定营销行为。上述装置可以由任何具有计算、处理能力的软件、硬件或其组合来实现。在一个实施例中，上述装置集成到深度强化学习系统中用于实现智能体的实体装置中。

- 15 [82]图 7 示出根据一个实施例的装置的示意性框图。如图 7 所示，该装置 700 包括：状态获取单元 71，配置为获取执行环境的状态信息，所述状态信息至少包括当前用户的用户信息；行为确定单元 73，配置为根据营销策略，针对所述状态信息确定相应的营销行为，所述营销行为包括营销渠道、营销内容和营销时段的组合；分数获取单元 75，配置为获取执行环境针对所述营销行为的当前奖励分数；更新单元 77，配置为至少根据所述当前奖励分数更新所述营销策略。

- 20 [83]根据一个实施例，用户信息可以包括，用户基本属性信息，以及用户历史行为信息。

[84]在一个例子中，用户历史行为信息包括以下中的至少一项：历史操作序列、最近浏览页面列表、最近接受的营销信息列表、消费记录、交易记录、贷款记录、出行记录、保险记录。

- 25 [85]在一个例子中，用户信息还可以包括，基于用户基本属性信息和/或用户历史行为信息确定的用户画像信息。

[86]根据一个实施例，执行环境的状态信息还包括，当前用户所处场景的场景信息。

[87]进一步地，场景信息可以包括，位置服务场景信息，支付场景信息，钱包操作事件信息。

[88]在一个实施例中，上述营销内容包括营销消息和利益包，所述利益包包括以下中的

一种或多种：红包、优惠券、权益。

[89]在一个实施例中，上述营销渠道包括以下中的一种或多种：消息推送，卡片展示，角标提醒，红点提醒。

[90]根据一种实施方式，所述装置还包括，回报价值确定单元（未示出），该单元至少
5 基于当前奖励分数确定回报价值，其中回报价值包括所述当前奖励分数和至少一项未来奖励分数的加权求和。相应地，更新单元 77 配置为，确定后续营销策略，使得所述回报价值最大化。

[91]根据又一方面的实施例，还提供一种计算设备，包括存储器和处理器，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，实现结合图 3 所述的方法。

10 [92]本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。

[93]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
15 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，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。