

通过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，尤其涉及通过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结果，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，可以同时推荐营销业务中的多个目标进行学习，以综合全面地促进业务效果。

20 [06] 根据第一方面，提供了一种通过深度强化学习系统进行推荐营销的方法，所述深度强化学习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，针对所述执行环境的状态信息确定营销行为，所述方法包括：

[07] 从所述执行环境中获取，用户响应于所述营销行为而产生的多项执行结果，所述多项执行结果分别对应于营销效果链路上的多个效果目标；

[08] 根据所述多项执行结果，确定强化学习的奖励分数；

25 [09] 将所述奖励分数返回给所述智能体，用于所述智能体更新其营销策略。

[10]根据一个实施例，上述执行环境的状态信息可以包括，备选营销信息和当前用户的用户信息。

[11]在一个实施例中，通过以下方式获取用户的多项执行结果：从服务端获取用户的交互行为，通过所述交互行为获取所述多项执行结果；或者，通过客户端页面埋点捕获用户的操作行为，通过用户的操作行为，获取所述多项执行结果。

[12]在一个实施例中，通过分数评估函数确定奖励分数，所述分数评估函数以所述多项执行结果为变量，并被设计为，正相关于上述多项执行结果与对应的多个效果目标之间的匹配度。

[13]在另一实施例中，确定强化学习的奖励分数包括，获取针对所述多项执行结果中至少部分执行结果对应构造的独立函数，至少基于所述独立函数的加权求和确定奖励分数。

[14]进一步地，在一个实施例中，上述独立函数可以被构造为，正相关于作为变量的执行结果与对应效果目标之间的匹配度。

[15]在一个可能的实施例中，所述多项执行结果包括第一执行结果，该第一执行结果具有离散结果；所述独立函数包括第一函数，所述第一函数根据所述第一执行结果的离散结果输出离散值。

[16]更进一步地，根据一种实施方式，上述第一函数在第一执行结果取第一结果时，输出第一离散值，在第一执行结果取第二结果时输出第二离散值；其中所述第一结果比第二结果具有与所述第一执行结果对应的第一效果目标更高的匹配度，所述第一离散值大于第二离散值。

[17]根据一种实施方式，还可以通过以下方式确定奖励分数：基于上述多项执行结果确定至少一个结果组合；获取针对该至少一个结果组合而构造的至少一个组合函数；至少基于所述至少一个组合函数的加权求和确定奖励分数。

[18]根据另一种实施方式，还可以通过以下方式确定奖励分数：确定所述多项执行结果所匹配的效果目标或效果目标组合；获取预先建立的效果目标或效果目标组合与奖励分数的映射关系；根据所述映射关系，确定所匹配的效果目标或效果目标组合对应的奖励分数。

[19]根据第二方面，提供一种通过深度强化学习系统进行推荐营销的装置，所述深度强化学习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，针对所述执行环境的状态信息确定营销行为，所述装置包括：

[20] 获取单元，配置为从所述执行环境中获取，用户响应于所述营销行为而产生的多项执行结果，所述多项执行结果分别对应于营销效果链路上的多个效果目标；

[21] 确定单元，配置为根据所述多项执行结果，确定强化学习的奖励分数；

[22] 返回单元，配置为将所述奖励分数返回给所述智能体，用于所述智能体更新其营销策略。

5

[23] 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。

[24] 通过本说明书实施例提供的方法和装置，利用深度强化学习系统来学习推荐营销业务，其中智能体在进行营销策略的更新时，同时考虑营销效果链路上的多个效果目标，也就是深度强化学习系统同时对营销链路上多个目标进行学习，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。

10

附图说明

[25] 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。

15

[26] 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；

[27] 图 2A 示出典型的深度强化学习系统；

[28] 图 2B 示出根据一个实施例的深度强化学习推荐营销系统；

[29] 图 3 示出根据一个实施例通过深度强化学习系统进行推荐营销的方法；

20

[30] 图 4 示出根据一个实施例的装置的示意性框图。

具体实施方式

[31] 下面结合附图，对本说明书提供的方案进行描述。

[32] 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。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，营销的过程可以包括以下多个环节。首先，系统根据一些推荐算法，选择一些

25

当前用户适合的营销信息准备进行推荐。这些营销信息可以包括，广告，利益包（例如红包，购物券，折扣券等），营销通知消息等。这些推荐营销信息被发送到营销投放平台进行投放。营销投放平台根据预定投放规则和策略，从推荐营销信息中选择一些营销信息进行曝光。一旦某条营销信息得到曝光，就被认为触达到用户。例如，在图 1 中，

5 当用户打开客户端 App(例如支付宝)，看到某条曝光的营销信息(例如营销信息 110)，这条营销信息即被认为触达用户。

[33]接着，用户有可能对触达的营销信息进行点击，例如通过点击营销信息 110，到达新的页面 120。然后，用户可能在该新的页面 120 上停留一段时间，然后对于该页面 120 上感兴趣的元素进行了二次点击，即二跳。该二跳例如跳转到新的页面 130，通过该页

10 面 130，用户完成了营销内容的目标操作，例如注册、转发、使用营销信息中的利益宝进行消费等，这时，可以认为该用户得到了转化。

[34]以上示出了推荐营销中期望实现的典型用户操作序列，也就是典型的用户营销效果链路，包含期望依次达成的多个效果目标，例如，触达->点击->停留->二跳->转化，其中用户的转化为推荐营销的最终目标。

15 [35]在一种实施方案中，为了促进业务效果，针对最终目标进行建模和机器学习。例如，对于推荐营销而言，用户的转化为最终目标，那么针对用户转化率进行建模学习。然而，用户的最终转化往往是通过多个步骤多次交互实现的，交互过程中自然地存在流量漏斗现象，即越靠后流量越小。因此，如果只学习最终的业务指标，即转化率，那么可用于学习的正样本会比较少。

20 [36]在另一种实施方案中，考虑到交互过程中多个环节步骤对最终目标都有一定影响，针对各个环节的效果目标分别进行建模学习。例如，对于触达环节，点击环节等等分别进行建模学习，得到例如触达率预估模型，点击率预估模型，转化率预估模型等等。利用这样的方案，获得更多的用于学习的样本，并且针对各个环节都进行有针对性的学习。然而，在这样的方案中，各个模型分别单独训练，成本很高。如果要估计测算业务的总

25 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。

[37]在综合考虑以上方案的基础上，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。图 2A 示出典型的深度强化学习系统。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

30

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。

[38]图 2B 示出根据一个实施例的深度强化学习推荐营销系统，也就是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，可观测的环境状态包括，有待推荐的营销信息，当前用户信息等。智能体基于一些营销策略，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，对奖励分数的生成过程进行优化，基于与营销效果链路上多个效果目标分别对应的多项用户执行结果，例如是否进行点击，是否有二跳，停留时间长短等，来确定综合奖励分数，如此使得，智能体营销策略的更新同时考虑到营销效果链路上的多个效果目标，也就是深度强化学习系统同时对营销链路上多个目标进行学习，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。

[39]图 3 示出根据一个实施例通过深度强化学习系统进行推荐营销的方法。可以理解，上述深度强化学习系统可以是如图 2B 所示的深度强化学习推荐营销系统，其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。上述方法可以由任何具有数据计算、处理能力的装置、设备、平台、设备集群等来执行。如图 3 所示，该方法至少包括：步骤 31，从执行环境中获取，针对营销效果链路上的多个效果目标，用户分别对应的多项执行结果，所述多项执行结果是用户响应于所述智能体确定的营销行为而产生；步骤 33，根据所述多项执行结果，确定强化学习的奖励分数；步骤 35，将所述奖励分数返回给所述智能体，用于所述智能体更新其营销策略。下面描述以上各个步骤的具体执行方式。

[40]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，并根据营销策略确定当前状态下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,可选的营销信息和当前用户的用户信息。更具体地,可选的营销信息可以包括可选的营销内容,例如备选的广告、利益包(例如红包,购物券,折扣券等)、营销通知消息等。可选的营销信息还可以包括,可选的营销渠道,例如消息推送、广告 banner、角标提醒等等。另一方面,营销环境的状态信息还包括当前用户的用户信息,用户信息可以包括用户属性特征,例如注册时长、年龄、性别、收入等等。在一个实施例中,用户信息还包括用户历史行为信息,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等等。在一个实施例中,用户信息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,基于属性信息将用户划分到一定人群的人群标签,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。

[41]在一个具体例子中,以营销信息对应的营销特征和当前用户对应的用户特征为基础,构建营销环境的当前状态的状态向量,输入到智能体。智能体基于一些营销策略,针对营销环境的当前状态,确定要采取的营销行为。

[42]可以理解,对于强化学习系统来说,智能体中的策略用于将环境的状态映射为接下来的行为。在图 2B 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反映了当前用户的用户特征和可选的营销信息的特征,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,例如,用什么样的渠道向其传送什么样的营销内容。

[43]在这样的营销行为作用于执行环境后,就会相应产生执行结果。因此,在步骤 31,可以从执行环境中获取,用户响应于上述营销行为而产生的、针对营销效果链路上的多个效果目标的多项执行结果。

[44]图 1 示出了一个具体例子中的营销效果链路,其中包含希望依次达成的多个效果目标,包括:触达->点击->停留->二跳->转化。相应地,用户针对这些效果目标的执行结果可以包括,营销对用户的触达状况(例如是否触达用户,触达或展示时间),用户点击状况(例如是否点击),用户停留时长,用户二次点击状况(例如是否进行第二点击,

即二跳)，用户转化状况（例如是否得到转化）。

[45]在一个实施例中，以上用户针对各项效果目标的执行结果体现为用户与执行环境的交互，特别是用户与营销服务端的交互。因此，可以通过从服务端获取用户的交互行为来获取上述执行结果。在另一实施例中，还可以通过客户端页面埋点的方式，捕获用户的操作行为，通过用户的操作行为，获取用户针对上述效果目标的执行结果。

[46]对于以上获取的，用户响应于营销行为而产生、针对多个效果目标的多项执行结果，在步骤 33，根据该多项执行结果，确定强化学习的奖励分数。

[47]下面结合触达->点击->停留->二跳->转化的营销效果链路以及对应的用户执行结果来描述奖励分数的确定。

10 [48]在一个实施例中，以全部多项执行结果为变量，构建分数评估函数，该分数评估函数被设计为，正相关于上述多项执行结果与对应的多个效果目标之间的匹配度，也就是，上述多项执行结果与对应的多个效果目标总体上越接近或越匹配，评估函数输出值越大，即奖励分数越高。

[49]即：

$$15 \quad R_t = f(X_1, X_2, \dots, X_n) \quad (1)$$

[50]其中 $X_1, X_2, \dots, X_n$ 为效果目标 $Y_1, Y_2, \dots, Y_n$ 分别对应的执行结果。并且，分数评估函数 f 被设计为，执行结果 $X_1, X_2, \dots, X_n$ 总体上与效果目标 $Y_1, Y_2, \dots, Y_n$ 匹配度越高， R_t 分数越高。

20 [51]在另一个实施例中，分别针对多项执行结果中至少部分执行结果构造对应的独立函数。在确定奖励分数时，获取预先构造的这些独立函数，至少基于这些独立函数的加权求和确定奖励分数。更具体地，针对第 i 项执行结果 X_i ，构造独立函数 $f_i(X_i)$ ，将最终的奖励分数确定为：

$$R_t = W_i * f_i(X_i) \quad (2)$$

[52]其中 W_i 为第 i 项执行结果对应的权重。

25 [53]例如，在以上营销效果链路的例子中，多项执行结果包括， $X_1 = \text{show}$ ，表示触达状况； $X_2 = \text{click}$ ，表示点击状况； $X_3 = \text{stay}$ ，表示用户停留时长； $X_4 = \text{click2}$ ，表示二次点击状况； $X_5 = \text{convert}$ ，表示转化状况，那么奖励分数可以确定为：

$$R_t = W_1 * f_1(\text{show}) + W_2 * f_2(\text{click}) + W_3 * f_3(\text{stay}) + W_4 * f_4(\text{click2}) + W_5 * f_5(\text{convert}) \quad (3)$$

[54] 在一个实施例中，至少一个独立函数 $f_i(X_i)$ 被确定为，正相关于执行结果 X_i 与对应效果目标 Y_i 之间的匹配度，即 X_i 越接近 Y_i ， $f_i(X_i)$ 值越大。

- 5 [55] 在又一实施例中，基于上述多项执行结果确定至少一个结果组合，各个结果组合包括两项或更多项执行结果，为该至少一个结果组合构造至少一个组合函数，并至少基于所述至少一个组合函数的加权求和确定奖励分数。

[56] 例如，在以上 X_1 - X_5 的多项执行结果的例子中，可以选择 X_1 和 X_2 作为第一结果组合，为其构造第一组合函数 F_1 ，选择 X_3 和 X_4 作为第二结果组合，为其构造第二组合函数 F_2 。各个结果组合可以包含不同数目的执行结果。

[57] 如此，在一个具体例子中，奖励分数可以确定为：

$$R_t = W_{F1} * F_1(\text{show}, \text{click}) + W_{F2} * F_2(\text{stay}, \text{click2}) \quad (4)$$

- [58] 该实施例可以与式 (2) 或式 (3) 对应的实施例相结合，例如对于未被选择作为结果组合的执行结果，为其构建对应独立函数，将这些独立函数与组合函数进行加权求和；
 15 或者，在各个执行结果对应函数的加权求和基础上，进一步引入组合函数的加权求和；
 或者，从多个执行结果中选择一部分执行结果构建相应的独立函数，将这些独立函数与组合函数进行加权求和。

[59] 例如，在一个具体例子中，可以将奖励分数确定为：

$$R_t = W_3 * f_3(\text{stay}) + W_4 * f_4(\text{click2}) + W_5 * f_5(\text{convert}) + W_{F1} * F_1(\text{show}, \text{click}) + W_{F2} * F_2(\text{stay}, \text{click2}) \quad (5)$$

[60] 其中，从 X_1 到 X_5 中选择了 X_3 , X_4 , X_5 ，为其构建独立函数；并且选择了 X_1 , X_2 作为第一组合，构建第一组合函数 F_1 ；以及选择了 X_3 , X_4 作为第二组合构建第二组合函数 F_2 ，将各个独立函数和组合函数进行加权求和，以此确定奖励分数。

- [61] 在一个实施例中，所述多项执行结果包括第一类执行结果，这类执行结果具有离散结果值，例如 X_2 表示点击状况，在一个例子中，可以将点击状况划分为点击或者未点击的二元离散结果值。

[62] 对于这样的第一类执行结果，可以为其构造离散函数作为其独立函数，该离散函数根据执行结果的离散结果值输出离散函数值。进一步地，在一个实施例中，对于针对某

个第一类执行结果 X_i 设置的离散函数 f_i ，可以将其离散输出值设置为，在执行结果 X_i 的离散结果值与对应的效果目标 Y_i 更接近的情况下， f_i 输出更大的离散值。例如，可以针对 X_2 构建离散函数 $f_2(X_2)$ ，当 X_2 对应于点击的时候， f_2 输出离散值 V_1 ；当 X_2 对应于未点击的时候， f_2 输出离散值 V_2 。更具体地，由于对应的效果目标为点击，因此 X_2 对应于点击时与效果目标更匹配，于是在一个例子中，可以将 V_1 设置为大于 V_2 ，例如 $V_1=1$ ， $V_2=0$ 。

[63]该实施例可以与前述式 (2) 到式 (5) 对应的任一实施例相结合，将其中针对第一类执行结果的独立函数设置为输出离散值的离散函数。例如，在与式 (3) 结合的情况下，对于 X_1 到 X_5 中具有二元结果的 X_1 ， X_2 ， X_4 和 X_5 ，可以将其对应独立函数改写为 I ，表示输出为 0 或 1 的函数，此时，式 (3) 可以改写为：

$$R_t = W_1 * I(\text{show}) + W_2 * I(\text{click}) + W_3 * f_3(\text{stay}) + W_4 * I(\text{click}_2) + W_5 * I(\text{convert})$$

(6)

[64]其中的函数 $I(X)$ 在 X 为真时输出 1，在 X 为假时输出 0。可以理解， X 为真对应于与效果目标一致的情况。

[65]在一个实施例中，多项执行结果还包括第二类执行结果，这类执行结果对应于连续变量，例如 X_3 表示停留时间，该停留时间可以表示为以秒为单位的连续变量。对于第二类执行结果，可以为其构造输出为连续值的第二类函数。在一个具体例子中，第二类函数可以是逻辑回归函数。在另一具体例子中，针对第二类执行结果构造线性回归函数。

[66]例如，对于连续变量 X_3 ，可以将对应的函数 f_3 构造为：

$$f_3(X_3) = A + B * X_3$$

(7)

[67]其中 A ， B 为预定线性参数。

[68]可以理解，以上式 (7) 对应的实施例可以与前述式 (2) 到式 (6) 对应的任一实施例相结合。

[69]根据另一种实施方式，预先建立达成的效果目标或效果目标组合与奖励分数的映射关系。在获取到与效果目标对应的执行结果之后，确定当前的执行结果所匹配的效果目标组合，进而根据上述映射关系，确定对应的奖励分数。

[70]在一个具体例子中，预先建立的映射关系可以如下表所示。

表 1

奖励分数	效果目标/目标组合
1	展示
2	展示+点击
3	停留
4	停留+二跳
5	转化

[71]通过表 1 可以看到，对于营销效果链路上期望依次达成的效果目标，如果仅达成展示或触达这一目标，那么奖励分数为 1；如果在展示基础上还达成了用户点击的效果目标，那么奖励分数为 2.；依次类推。达成的效果目标越接近营销效果链路的最终目标，奖励分数越高。

[72]基于这样的映射关系，在获取到各个效果目标分别对应的执行结果之后，将各个执行结果与对应效果目标进行比较和匹配，从而确定出当前的执行结果所匹配的效果目标或效果目标组合，进而可以通过表 1 的映射关系，查找确定出对应的奖励分数。

[73]需要理解的是，表 1 仅仅是一个示例。可以根据业务需要，设置更多、更少或不同的效果目标或效果目标组合，并相应地为奖励分数设置不同的分数取值。

[74]以上结合触达->点击->停留->二跳->转化的营销效果链路描述了奖励分数的多种确定方式。可以理解，营销效果链路并不限于以上描述中的例子，而是可以包含更多、更少或不同的效果目标。例如，在一些营销场景下，用户进行点击之后，要么转化要么退出，而不要求二跳；在另一些营销场景下，用户在二跳之后还可能进行进一步点击，例如发生三跳；在又一些营销场景下，用户在点击或二跳之后还可能进行其他交互，例如输入个人信息等等。这些操作都可以作为营销效果链路上的预定效果目标，进而作为强化学习系统学习的对象。

[75]以上根据用户针对营销效果链路上的多个效果目标的多项执行结果，确定了强化学习的奖励分数。于是，接下来在步骤 35，将该奖励分数返回给智能体，用于所述智能体更新其营销策略。

[76]如前所述，智能体不断从环境获得奖励分数的反馈，以此确定之前的营销策略和营

销行为是否恰当，据此更新营销策略。需要说明的是，此处的更新营销策略不仅包括在确定发出的营销行为不恰当时修改或调整之前的营销策略，还包括在确定营销行为效果良好时增强或者维持原有的营销策略。在本说明书的一个或多个实施例中，由于奖励分数基于用户对多个效果目标的多项执行结果来确定，因此，智能体在更新营销策略时，

5 会全面地考虑多个效果目标的执行结果，也就是，同时对营销效果链路上的多个效果目标进行了学习。如此使得，深度强化学习系统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。

[77]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 2B 所示包括智能体和执行环境，所述智

10 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，上述装置可以集成到深度强化学习系统的执行环境中；在另一实施例中，上述装置可以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；在又一实施例中，上述装置作为独立装置，从执行环境中提取数据进行处理，得到奖励分数，返回到智能体。

15 [78]图 4 示出根据一个实施例的装置的示意性框图。如图 4 所示，该装置 400 包括：获取单元 41，配置为从所述执行环境中获取，用户响应于所述营销行为而产生的多项执行结果，所述多项执行结果分别对应于营销效果链路上的多个效果目标；确定单元 43，配置为根据所述多项执行结果，确定强化学习的奖励分数；返回单元 45，配置为将所述奖励分数返回给所述智能体，用于所述智能体更新其营销策略。

20 [79]在一个实施例中，上述执行环境的状态信息包括，备选的营销信息和当前用户的用户信息。

[80]根据一个实施例，获取单元 41 通过以下方式获取用户的多项执行结果：从服务端获取用户的交互行为，通过所述交互行为获取所述多项执行结果；或者，通过客户端页面埋点捕获用户的操作行为，通过用户的操作行为，获取所述多项执行结果。

25 [81]在一个实施例中，确定单元 43 通过分数评估函数确定奖励分数，所述分数评估函数以所述多项执行结果为变量，并被设计为，正相关于上述多项执行结果与对应的多个效果目标之间的匹配度。

[82]在另一实施例中，确定单元 43 通过以下方式确定奖励分数：获取针对所述多项执行结果中至少部分执行结果对应构造的独立函数，至少基于所述独立函数的加权求和确定

奖励分数。

[83]进一步地，在一个例子中，上述独立函数被构造为，正相关于作为变量的执行结果与对应效果目标之间的匹配度。

[84]在一个例子中，多项执行结果包括第一执行结果，该第一执行结果具有离散结果；

- 5 所述独立函数包括第一函数，所述第一函数根据所述第一执行结果的离散结果输出离散值。

[85]在进一步的例子中，第一函数在第一执行结果取第一结果时，输出第一离散值，在第一执行结果取第二结果时输出第二离散值；其中所述第一结果比第二结果具有与所述第一执行结果对应的第一效果目标更高的匹配度，所述第一离散值大于第二离散值。

- 10 [86]根据一种实施方式，确定单元 43 还可以如下确定奖励分数：

[87]基于上述多项执行结果确定至少一个结果组合；

[88]获取针对该至少一个结果组合而构造的至少一个组合函数；

[89]至少基于所述至少一个组合函数的加权求和确定奖励分数。

- 15 [90]根据另一种实施方式，确定单元 43 如下确定奖励分数：确定所述多项执行结果所匹配的效果目标或效果目标组合；获取预先建立的效果目标或效果目标组合与奖励分数的映射关系；根据所述映射关系，确定所匹配的效果目标或效果目标组合对应的奖励分数。

[91]由于确定单元 43 基于用户对多个效果目标的多项执行结果来确定奖励分数，因此，智能体在更新营销策略时，会全面地考虑多个效果目标的执行结果，使得深度强化学习系统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。

- 20 [92]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。

[93]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
25 传输。

[94]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。

权利要求书

1. 一种通过深度强化学习系统进行推荐营销的方法,所述深度强化学习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,针对所述执行环境的状态信息确定营销行为,所述方法包括:
 - 5 从所述执行环境中获取,用户响应于所述营销行为而产生的多项执行结果,所述多项执行结果分别对应于营销效果链路上的多个效果目标;
根据所述多项执行结果,确定强化学习的奖励分数;
将所述奖励分数返回给所述智能体,用于所述智能体更新其营销策略。
 2. 根据权利要求1所述的方法,其中从所述执行环境中获取,用户响应于所述营销
10 行为而产生的多项执行结果包括:
从服务端获取用户的交互行为,通过所述交互行为获取所述多项执行结果;或者
通过客户端页面埋点捕获用户的操作行为,通过用户的操作行为,获取所述多项执行结果。
 3. 根据权利要求1所述的方法,其中所述营销效果链路上的多个效果目标包括以下
15 中的多个目标:营销信息的触达、首次点击、停留、进一步点击、转化。
 4. 根据权利要求1所述的方法,其中确定强化学习的奖励分数包括:通过分数评估函数确定奖励分数,所述分数评估函数以所述多项执行结果为变量,并被设计为,正相关于上述多项执行结果与对应的多个效果目标之间的匹配度。
 5. 根据权利要求1所述的方法,其中确定强化学习的奖励分数包括:
20 获取针对所述多项执行结果中至少部分执行结果各自构造的独立函数,至少基于所述独立函数的加权求和确定奖励分数。
 6. 根据权利要求5所述的方法,所述独立函数中的至少一个被构造为,正相关于对应的执行结果与效果目标之间的匹配度。
 7. 根据权利要求5所述的方法,其中所述多项执行结果包括第一执行结果,该第一
25 执行结果具有离散结果值;所述独立函数包括第一函数,所述第一函数根据所述第一执行结果的离散结果值输出离散函数值。
 8. 根据权利要求7所述的方法,其中所述第一函数在第一执行结果具有第一结果值时,输出第一离散值,在第一执行结果具有第二结果值时输出第二离散值;其中所述第一结果值比第二结果值具有与所述第一执行结果对应的第一效果目标更高的匹配度,所述
30 所述第一离散值大于第二离散值。
 9. 根据权利要求1或5-8中任一项所述的方法,其中确定强化学习的奖励分数包括:

基于上述多项执行结果确定至少一个结果组合;

获取针对该至少一个结果组合而构造的至少一个组合函数;

至少基于所述至少一个组合函数的加权求和确定奖励分数。

10. 根据权利要求 1 所述的方法, 其中确定强化学习的奖励分数包括:

5 确定所述多项执行结果所匹配的效果目标或效果目标组合;

获取预先建立的效果目标或效果目标组合与奖励分数的映射关系;

根据所述映射关系, 确定所匹配的效果目标或效果目标组合对应的奖励分数。

11. 一种通过深度强化学习系统进行推荐营销的装置, 所述深度强化学习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, 针对所述执行环境的状态信息确定
10 营销行为, 所述装置包括:

获取单元, 配置为从所述执行环境中获取, 用户响应于所述营销行为而产生的多项执行结果, 所述多项执行结果分别对应于营销效果链路上的多个效果目标;

确定单元, 配置为根据所述多项执行结果, 确定强化学习的奖励分数;

15 返回单元, 配置为将所述奖励分数返回给所述智能体, 用于所述智能体更新其营销策略。

12. 根据权利要求 11 所述的装置, 其中所述获取单元配置为:

从服务端获取用户的交互行为, 通过所述交互行为获取所述多项执行结果; 或者

通过客户端页面埋点捕获用户的操作行为, 通过用户的操作行为, 获取所述多项执行结果。

20 13. 根据权利要求 11 所述的装置, 其中所述营销效果链路上的多个效果目标包括以下中的多个目标: 营销信息的触达、首次点击、停留、进一步点击、转化。

14. 根据权利要求 11 所述的装置, 其中所述确定单元配置为: 通过分数评估函数确定奖励分数, 所述分数评估函数以所述多项执行结果为变量, 并被设计为, 正相关于上述多项执行结果与对应的多个效果目标之间的匹配度。

25 15. 根据权利要求 11 所述的装置, 其中所述确定单元配置为:

获取针对所述多项执行结果中至少部分执行结果各自构造的独立函数, 至少基于所述独立函数的加权求和确定奖励分数。

16. 根据权利要求 15 所述的装置, 所述独立函数中的至少一个被构造为, 正相关于对应的执行结果与效果目标之间的匹配度。

30 17. 根据权利要求 15 所述的装置, 其中所述多项执行结果包括第一执行结果, 该第一执行结果具有离散结果值; 所述独立函数包括第一函数, 所述第一函数根据所述第一

执行结果的离散结果值输出离散函数值。

18. 根据权利要求 17 所述的装置, 其中所述第一函数在第一执行结果具有第一结果值时, 输出第一离散值, 在第一执行结果具有第二结果值时输出第二离散值; 其中所述第一结果值比第二结果值具有与所述第一执行结果对应的第一效果目标更高的匹配度,

5 所述第一离散值大于第二离散值。

19. 根据权利要求 11 或 15-18 中任一项所述的装置, 其中所述确定单元配置为:

基于上述多项执行结果确定至少一个结果组合;

获取针对该至少一个结果组合而构造的至少一个组合函数;

至少基于所述至少一个组合函数的加权求和确定奖励分数。

10 20. 根据权利要求 11 所述的装置, 其中所述确定单元配置为:

确定所述多项执行结果所匹配的效果目标或效果目标组合;

获取预先建立的效果目标或效果目标组合与奖励分数的映射关系;

根据所述映射关系, 确定所匹配的效果目标或效果目标组合对应的奖励分数。

21. 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。

15