问题的引出：客户想使用DRL深度强化学习对短视频主播推荐业务问题建模

客户想把DRL用在首页个性化推荐业务的排序阶段（DRL可以用在首页推荐的不同阶段，比如微软新闻推荐DRN模型以及阿里淘宝主搜的锦囊推荐DRL用在了排序阶段；youtube的DRL直接做了一路召回，京东的DRL模型也可以作为一路召回，其实youtube和京东的这两个模型也可以作为一个单独链路同时完成召回和排序的任务，因为他们是对整个推荐list建模的；除了上面两种思路，美团在他们的实践中，使用DRL对单独的两个CTR模型和CVR模型最后的打分的加权的权重系数进行建模），所以下面建模的DRL的各种因素都是按照微软DRN模型的建模思路来定义的。

针对这个业务来说，DRL中的MDP马尔科夫决策过程建模参考如下（环境就是用户，智能体agent就是推荐系统排序引擎）：

episode： 把用户在首页推荐中的多次翻页或者刷屏看成多轮交互。episode的开始就是用户刚登录进入首页；episode的结束可以认为该用户开始长时间观看某个主播视频（这个结束的定义是参考电商网站中用户下单购买动作作为一个episode的结束。当然要从业务角度思考这个结束的定义是否合适，比如也可以考虑episode结束的定义为用户长时间比如半个小时没有在首页推荐页面有任务动作---这个经常是被看作一个用户session的结束，常常用在用户行为的序列化建模中。另外需要从业务侧定义好长时间观看的标准是什么，是必须看完还是看了%70就算）

**episode不一定是必须建模的，但是通过建模episode，能更好的理解DRL中的任务的最终目标是使得当前动作的长期累积回报最高**。那么所谓的长期累积回报，和这个episode发生的后续行为是有关系的。而且对于目标Q值的设计，有的时候是需要根据当前episode是否结束来设置不同的长期累积回报的。

状态： 用户侧的特征（包括静态特征，动态特征（比如多个周期统计特征，用户近期的以itemid表示的行为序列）），物品侧的特征（包括静态特征，动态统计特征），上下文特征（当前用户访问session的一些上下文信息），甚至可以有用户和物品的交叉特征（微软的DRN强化学习模型就用到了这些特征）

---------id类的特征需要做embedding，embedding最好端到端在DRL模型中做，当然也看到了有的公司比如京东把pretrain的ID的embedding拿来直接用在DRL模型作为状态的一部分。（这个两种思路都可以尝试，一般建议优先尝试第一种。）

动作： 要推荐给用户的listwise的推荐列表（这个在实际落地中，看到微软新闻推荐DRN和阿里精囊推荐DRL这两个项目，他们并没有真正直接把动作用一个推荐列表来建模，而是把单个候选item和其他特征都作为状态并且作为模型的输入，输出该候选动作的Q值-----也就是说对于微软这个建模方式并没有明显的定义动作。这里需要自己实现一个逻辑，**把召回的物品依次和其他特征联合一起送入模型**，得到了所有候选item的Q值，对Q值排序就可以获得topK的列表。**这里也看到其实在做DRL建模的过程中，不是每个RL的因素都是那么的清晰的定义**。）

奖励： 根据用户反馈给予Agent相应的奖励，为业务目标直接负责。agent根据当前的state，采取相应的action即推荐K个物品列表给用户之后，根据用户对推荐列表的反馈(忽略、点击或长时间观看/看完)来得到当前state-action的即时奖励reward。

最终目标： 最大化用户点击率，完播率和总的播放时长

模型的选择：

可以先尝试value-based的方法，比如用DQN来做尝试（微软新闻推荐的DRN和阿里淘宝主搜的锦囊推荐的DRL都用了DQN来做）；之后可以考虑联合使用value-based和policy-based的方法的模型比如A3C,SAC-Discrete（即离散版SAC），PPO等。

难点：

**样本的稀疏度问题**：

这个比较麻烦（尤其对于DRL来说更是问题，因为DRL本质就是想要过拟合他看到的数据），有一些workaround：

京东通过一个仿真器来仿真没有出现过的state-action的reward值，用于训练和评估线下模型；

阿里通过使用一个模拟器利用生成对抗模仿学习模型的判别器来输出奖励（同时利用真实的用户行为以及模拟的用户行为来训练），也就是说不需人为的设计奖励函数，这个有点像逆强化学习自动建模奖励函数的意思。

**状态的定义问题**：

用户侧特征（用户侧特征包括比如最近一段时间点击过的item id序列）和每个候选item特征。

智能体（推荐系统）给用户曝光的item list，这个可以简单认为是智能体的动作。之后这个动作会导致用户产生不同的行为，用户可能点击/转化，或者忽略。那下一个状态很可能会变化即所谓的状态转移，也就是说我们需要修改用户侧特征中比如最近点击过的item id序列的值。

**问题来了**：注意这里下一个状态中的候选item特征一般会和当前状态中的候选item不一样（**因为两次不同时刻召回的item不同，这个很常见**）。也就是说在数据准备的时候，要严格根据时间戳并且按照点击事件/转化事件来整理数据。

微软新闻推荐DRN和阿里精囊推荐DRL并没有给出如何定义状态的落地细节，所以这里需要我们自己思考如何来做。我觉得按照用户ID来整理数据比较方便，也就是mini-batch中的每个样本需要包括当前状态，下一个状态。所以准备的样本可能是类似如下的样子：

**当前状态uid\_1**的特征（包括统计特征以及最近发生行为的item id序列特征），**候选item\_1**特征，是否放生了点击/转化行为，**下一个状态uid\_1**的特征（统计特征以及序列特征这些动态特征可能有变化），**下一个状态候选item\_x1**的特征

**当前状态uid\_1**的特征（包括统计特征以及最近发生行为的item id序列特征），**候选item\_1**特征，是否放生了点击/转化行为，**下一个状态uid\_1**的特征（统计特征以及序列特征这些动态特征可能有变化），**下一个状态候选item\_x2**的特征

........

**当前状态uid\_1**的特征（包括统计特征以及最近发生行为的item id序列特征），**候选item\_2**特征，是否放生了点击/转化行为，**下一个状态uid\_1**的特征（统计特征以及序列特征这些动态特征可能有变化），**下一个状态候选item\_x1**的特征

**当前状态uid\_1**的特征（包括统计特征以及最近发生行为的item id序列特征），**候选item\_2**特征，是否放生了点击/转化行为，**下一个状态uid\_1**的特征（统计特征以及序列特征这些动态特征可能有变化），**下一个状态候选item\_x2**的特征

........

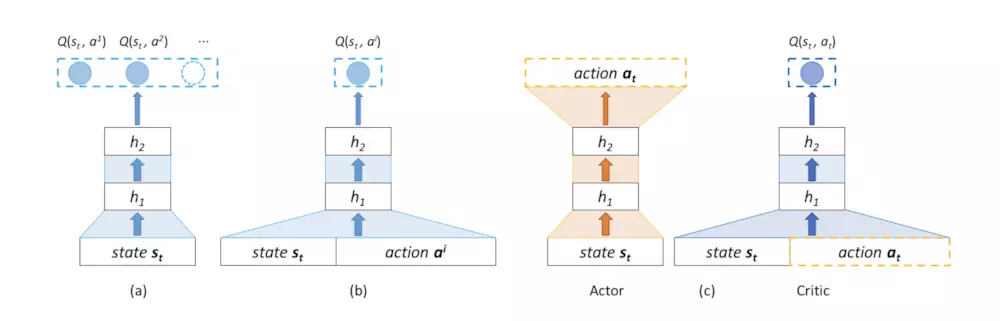
以此类推，每个uid的样本都这样组织好（有点类似笛卡尔积的意思）。

这样，在每次迭代中，直接能取到当前state，曝光后的下一个state，当前state的即时奖励根据当前state是否发生点击/转化行为来计算。接着就可以把（current\_state, reward, next\_state）放入replay buffer pool中（注意这里并没有也不需要显示把action放入经验池中）。

另外，在使用DQN建模的时候，通过目标网络获得下一个状态的Q值，然后在累积计算价值长期回报的时候，在这里不适合只用一个状态来计算累积长期回报，所以可能的方式是对相关的下一个state的所有候选item的Q值做聚合然后在计算长期回报，并把这个作为评估网络针对当前state的label。

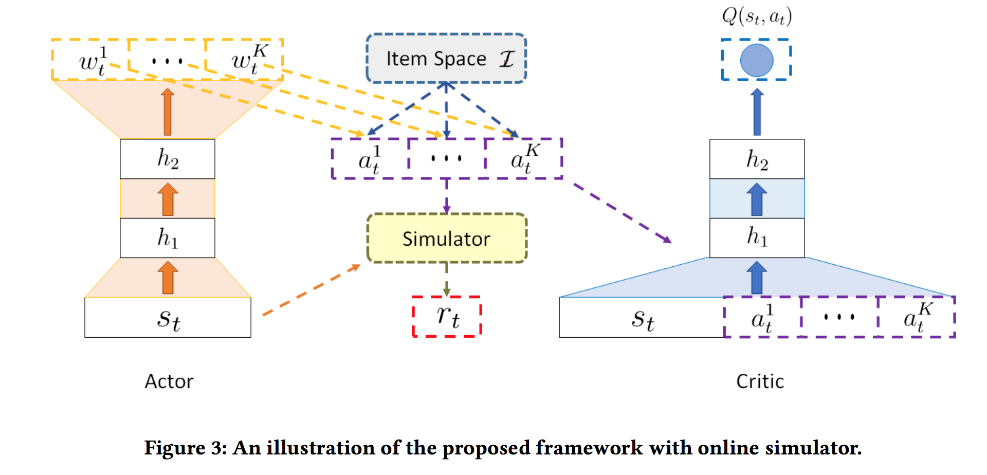
这个方式本质对于当前state还是point-wise的处理方式，并没有对当前state计算整个推荐列表list-wise的Q值（因为当前这样的建模假设不方便计算整个list-wise的Q值）。

这里一直讨论的是微软DRN模型（他其实就是用DQN来建模的），阿里的锦囊推荐也是类似的状态-动作建模思路。就是把用户特征和候选item作为网络的输入，Q值作为输出。而强化学习常见的三种状态-动作建模思路如下图所示：



也就是说微软DRN和阿里的锦囊推荐都用的是上图中的b图来建模的。

另外一种建模方式是类似京东电商推荐的方式，他是真正的list-wise的action，就是直接推荐一个list给用户。他是采用actor-critic的思路来建模状态-动作的，critic的输入也是state和一个推荐list（推荐list是有actor和他的模拟器一起计算得到的），输出的是整个list的Q值。京东的这个建模思路是对上图c的扩展（**注意图c中的critic价值网络的输入中的action是其actor策略网络的输入的action**）。京东电商推荐的DRL模型的结构如下：



**现状**：**网上基本找不到DRL落地推荐系统的细节以及参考代码；有些客户刚好也在做一些调研。基本上来说，DRL落地推荐系统是在一些大型互联网公司内部在做一些早期尝试（看到的相关的论文或者blog讲解也主要是2020年之前的，很多2018/2017年的），具体效果如何也不得知。至少从目前来看，还不是主流的趋势**。

**参考文章：**

1. 微软在2018年的论文，强化学习推荐模型DRN的详细解读可以参考（用于新闻的推荐）

<https://www.i4k.xyz/article/qq_35812205/121201008>

2. 阿里强化学习的应用实践中，参考第四章：强化学习在淘宝锦囊推荐系统中的应⽤。这个和我们讨论的这个场景比较相似。

3. 京东电商推荐系统中强化学习的实践：<https://www.6aiq.com/article/1547826520120>

4. 强化学习在美团首页推荐中“猜你喜欢”板块的实践：<https://tech.meituan.com/2018/11/15/reinforcement-learning-in-mt-recommend-system.html>

5. youtube推荐强化学习介绍：<http://d0evi1.com/youtube-recommend4/>