27 | 评估体系：如何解决A/B测试资源紧张的窘境？

你好，我是王喆。

我们在进行推荐系统评估时经常会遇到两类问题。

一类是在做线上 A/B 测试的时候，流量经常不够用，要排队等别人先做完测试之后才能进行自己的测试。线上 A/B 测试资源紧张的窘境，会大大拖慢我们试验的新思路，以及迭代优化模型的进度。

另一类是，离线评估加上在线评估有那么多种测试方法，在实际工作中，我们到底应该选择哪一种用来测试，还是都要覆盖到呢？

其实，这两个问题的答案是有深刻联系的，并不是孤立的。我认为最好的解决办法就是，建立起一套**推荐系统的评估体系，用它来解决不同评估方法的配合问题，以及线上 A/B 测试资源紧张的问题。这节课，我就带你一起来厘清如何建立起一整套推荐系统评估体系。**

什么是推荐系统的评估体系？

首先，什么是评估体系呢？我先给它下一个定义，**推荐系统的评估体系指的是，由多种不同的评估方式组成的、兼顾效率和正确性的，一套用于评估推荐系统的解决方案**。一个成熟的推荐系统评估体系应该综合考虑评估效率和正确性，可以利用很少的资源，快速地筛选出效果更好的模型。

那对一个商业公司来说，最公正也是最合理的评估方法就是进行线上测试，来评估模型是否能够更好地达成公司或者团队的商业目标。但是，正如我们开头所说，线上 A/B 测试要占用宝贵的线上流量资源，

这些有限的线上测试机会远远不能满足算法工程师改进模型的需求。所以如何有效地把线上和离线测试结合起来，提高测试的效率，就是我们迫切的需求。

那我们该怎么去构建起一整套评估体系呢？图 1 就是一个典型的评估体系示意图。从图中我们可以看到，处于最底层的是传统的离线评估方法，比如 Holdout 检验、交叉检验等，往上是离线 Replay 评估方法，再往上是一种叫 Interleaving 的线上测试方法，我们等会还会详细介绍，最后是线上 A/B 测试。

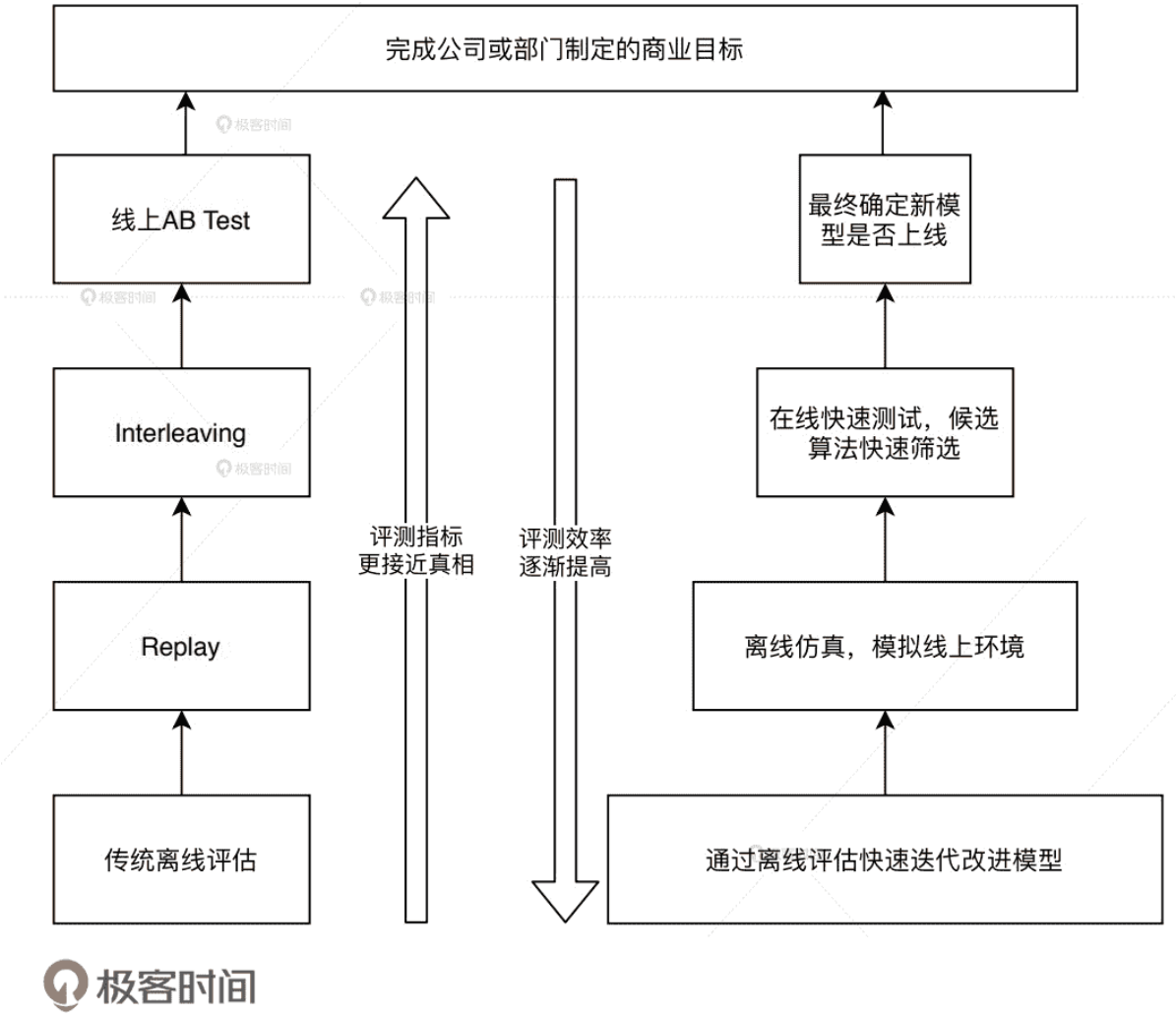

图1 推荐系统的评测体系
（出自《深度学习推荐系统》）

这四层结构共同构成完整的评估体系，做到了评估效率和评估正确性之间的平衡，越是底层的方法就会承担越多筛选掉改进思路的任务，这时候“评估效率”就成了更关键的考虑因素，那对于“正确性”的评估，我们反而没有多么苛刻的要求了。

总的来说，离线评估由于有着更多可供利用的计算资源，可以更高效、快速地筛选掉那些“不靠谱”的模型来改进思路，所以被放在了第一层的位置。

随着候选模型被一层层筛选出来，越接近正式上线的阶段，评估方法对评估“正确性”的要求就越严格。因此，在模型正式上线前，我们应该以最接近真实产品体验的 A/B 测试，来做最后的模型评估，产生最具说服力的在线指标之后，才能够进行最终的模型上线，完成模型改进的迭代过程。

讲了这么多，你可能会觉得，道理没问题，但工作中真的是这样吗？不如，我们来看个例子。下图就是一个很形象的工作中的模型筛选过程。

假设，现在有 30 个待筛选的模型，如果所有模型都直接进入线上 A/B 测试的阶段进行测试，所需的测试样本是海量的，由于线上流量有限，测试的时间会非常长。但如果我们把测试分成两个阶段，第一个阶段先进行初筛，把 30 个模型筛选出可能胜出的 5 个，再只对这 5 个模型做线上 A/B 测试，所需的测试流量规模和测试时间长度都会大大减少。这里的初筛方法，就是我们在评估体系中提到的离线评估、离线 Replay 和在线 Interleaving 等方法。

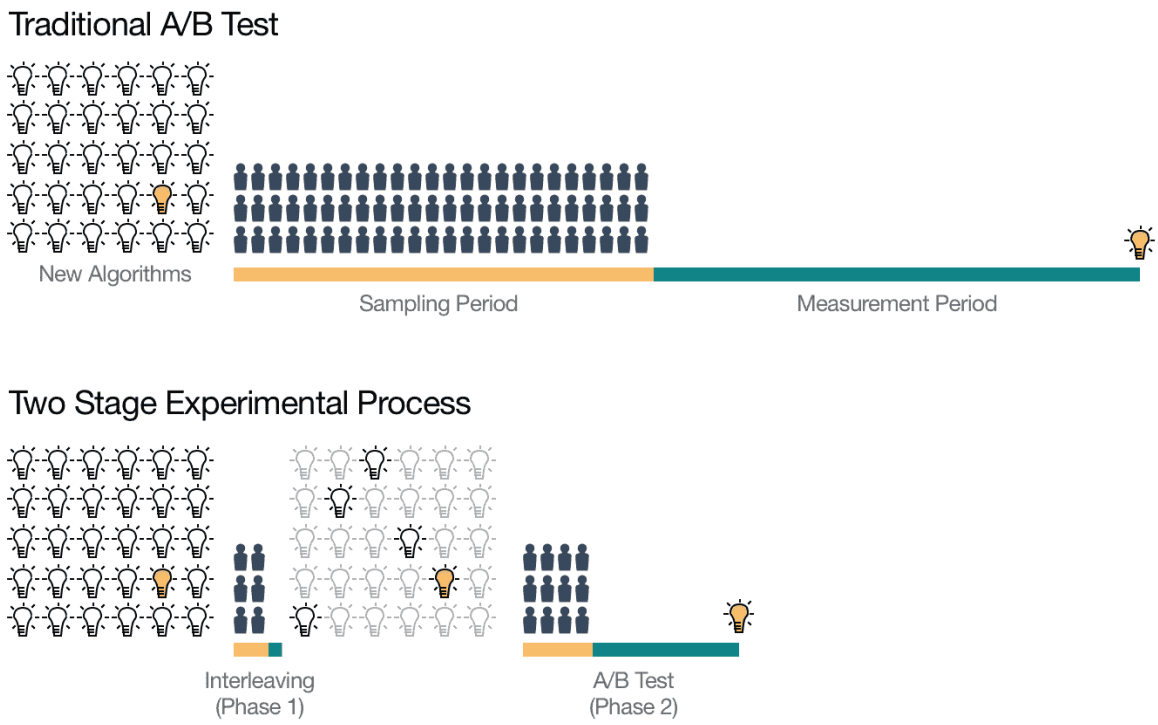

图2 模型的筛选过程

（图片出自The Netflix Tech Blog）

到这里，我想你已经清楚了什么是推荐系统的评估体系，以及评估体系是有哪些方法组成的。但在这些组成方法中，我们还有两点要重点注意：**一个是离线 Replay 这个方法，虽然我们之前讲过离线 Replay 的原理，但是对于它的相关工程架构还没有讲过；第二个是上面提到过的线上 Interleaving 方法。** 下面，我就借着流媒体巨头 Netflix 的实践方案，来讲解一下离线 Replay 和在线 Interleaving 的细节。

Netflix 的 Replay 评估方法实践

借着下图 3，我们来回顾一下，第 24 课学过的离线 Replay 方法的原理：离线 Replay 通过动态的改变测试时间点，来模拟模型的在线更新过程，让测试过程更接近真实线上环境。

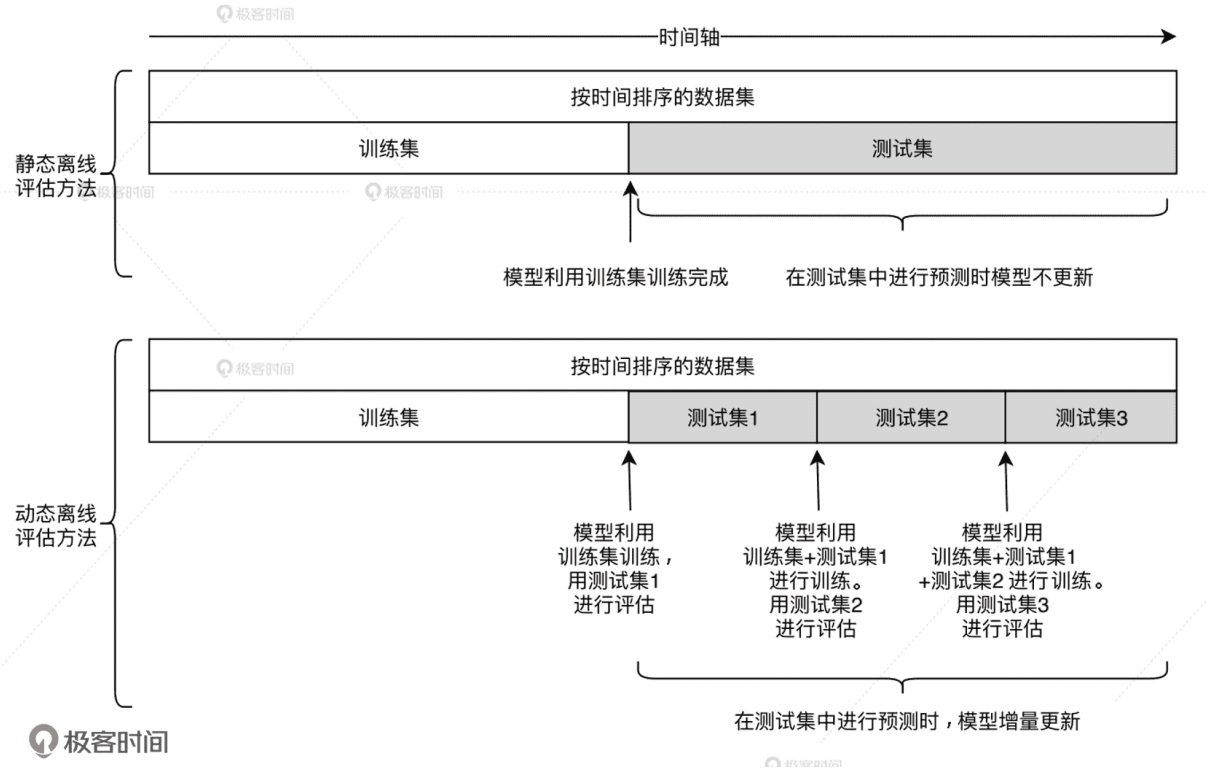

图3 静态时间分割评估与动态Replay评估
（出自《深度学习推荐系统》）

但是在 Replay 方法的实现过程中，存在一个很棘手的工程问题，就是我们总提到的“未来信息”问题，或者叫做“特征穿越”问题。因此在 Replay 过程中，每次模型更新的时候，我们都需要用历史上“彼时彼刻”的特征进行训练，否则训练和评估的结果肯定是不准确的。

我来举个例子，假设 Replay 方法要使用 8 月 1 日到 8 月 31 日的样本数据进行重放，这些样本中包含一个特征，叫做“历史 CTR”，这个特征只能通过历史数据来计算生成。

比如说，8 月 20 日的样本就只能够使用 8 月 1 日到 8 月 19 日的数据来生成“历史 CTR”这个特征，绝不能使用 8 月 20 日以后的数据来生成这个特征。在评估过程中，如果我们为了工程上的方便，使用了 8 月 1 日到 8 月 31 日所有的样本数据生成这个特征，供所有样本使

用，之后再使用 Replay 的方法进行评估，那我们得到的结论必然是错误的。

那么问题来了，在工程上，为了方便按照 Replay 方法进行模型评估，我们应该怎么去建立一套数据处理的架构，支持这种历史特征的复现呢？接下来，我们就看一看 Netflix 是怎么解决这个问题的。

Netflix 为了进行离线 Replay 的实验，建立了一整套从数据生成到数据处理再到数据存储的数据处理架构，并给它起了一个很漂亮的名字，叫做时光机（Time Machine）。

下图 4 就是时光机的架构，图中最主要的就是 Snapshot Jobs（数据快照）模块。它是一个每天执行的 Spark 程序，它做的主要任务就是把当天的各类日志、特征、数据整合起来，形成当天的、供模型训练和评估使用的样本数据。它还会以日期为目录名称，将样本快照数据保存在分布式文件系统 S3 中（Snapshots），再对外统一提供 API（Batch APIs），供其他模型在训练和评估的时候按照时间范围方便地获取。

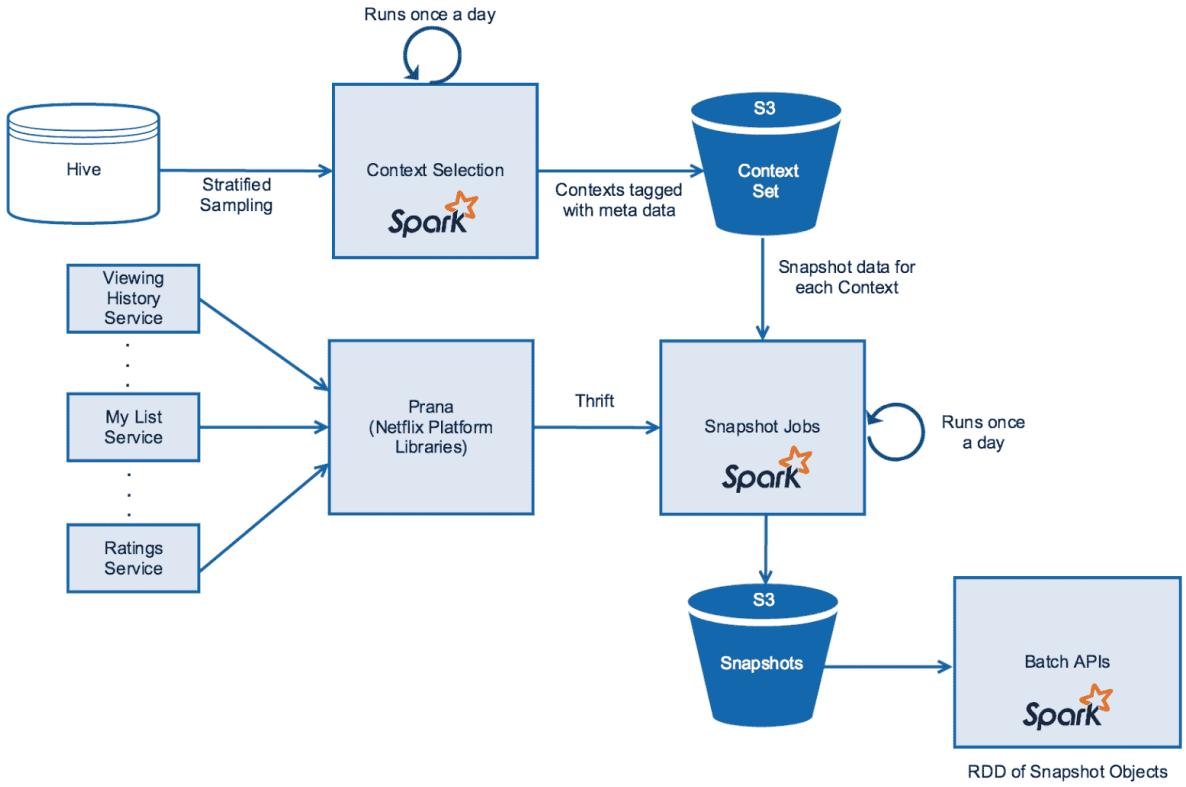

图4 Netflix的离线评估数据流架构——时光机
（出自The Netflix Tech Blog ）

这个 Snapshot Jobs 主任务的源数据是从哪来的呢？你可以重点关注它上方的 Context Set 模块和左边的 Prana 模块。接下来，我再详细和你说说这两个模块的任务。

Context Set 模块负责保存所有的历史当天的环境信息。 环境信息主要包括两类：一类是存储在 Hive 中的场景信息，比如用户的资料、设备信息、物品信息等数据；另一类是每天都会发生改变的一些统计类信息，包括物品的曝光量、点击量、播放时长等信息。

Prana 模块负责处理每天的系统日志流。 系统日志流指的是系统实时产生的日志，它包括用户的观看历史（Viewing History）、用户的推荐列表（My List）和用户的评价（Ratings）等。这些日志从各自的

服务（Service）中产生，由 Netflix 的统一数据接口 Prana 对外提供服务。

因此，Snapshot Jobs 这个核心模块每天的任务就是，通过 Context Set 获取场景信息，通过 Prana 获取日志信息，再经过整合处理、生成特征之后，保存当天的数据快照到 S3。

在生成每天的数据快照后，使用 Replay 方法进行离线评估就不再是一件困难的事情了，因为我们没有必要在 Replay 过程中进行烦琐的特征计算，直接使用当天的数据快照就可以了。

在时光机这个架构之上，使用某个时间段的样本进行一次 Replay 评估，就相当于直接穿越到了彼时彼刻，用当时的日志和特征进行模型训练和评估，就像进行了一次时光旅行（Time Travel）一样。

Interleaving 评估方法是什么

讲完了离线 Replay 的工程实现方法，我们再来聊一聊什么是 Interleaving 在线评估方法。

那 Interleaving 评估方法提出的意义是什么呢？主要有两方面：首先，它是和 A/B 测试一样的在线评估方法，能够得到在线评估指标；其次，它提出的目的是为了比传统的 A/B 测试用更少的资源，更快的速度得到在线评估的结果。

清楚了 Interleaving 评估方法提出的意义，我们就可以更好地理解 Interleaving 方法的具体细节了。下面，我们对比 A/B 测试，来看看 Interleaving 方法的具体实现过程。

在传统的 A/B 测试中，我们会把用户随机分成两组。一组接受当前的推荐模型 A 的推荐结果，这一组被称为对照组 。另一组接受新的推荐

模型 B 的推荐结果，这组被成为实验组。

在 Interleaving 方法中，不再需要两个不同组的用户，只需要一组用户，这些用户会收到模型 A 和模型 B 的混合结果。也就是说，用户会在一个推荐列表里同时看到模型 A 和模型 B 的推荐结果。在评估的过程中，Interleaving 方法通过分别累加模型 A 和模型 B 推荐物品的效果，来得到模型 A 和 B 最终的评估结果。

下图可以帮助我们更形象地对比 A/B 测试和 Interleaving 方法。

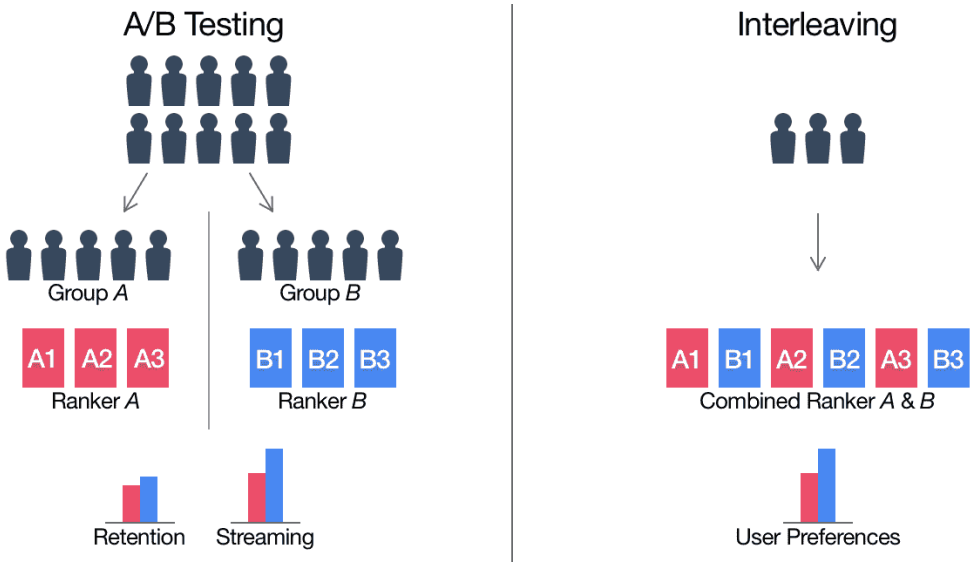

图5 传统A/B测试和Interleaving方法的比较
（出自The Netflix Tech Blog ）

那你可能想问了，在使用 Interleaving 方法进行测试的时候，我们该怎么保证对模型 A 和模型 B 的测试是公平的呢？如果有一个模型的结果总排在第一位，这对另一个模型不就不公平了吗？

这个问题很好，我们确实需要考虑推荐列表中位置偏差的问题，要想办法避免来自模型 A 或者模型 B 的物品总排在第一位。因此，我们需要以相等的概率让模型 A 和模型 B 产生的物品交替领先。这就像在野

球场打球的时候，两个队长会先通过扔硬币的方式决定谁先选人，再交替来选择队员。

理解了原理，我们再结合下面的图示，来进一步理解 Interleaving 方法混合模型 A 和 B 结果的过程。和刚才说的野球场选人的过程一样，我们先选模型 A 或者模型 B 的排名第一的物品作为最终推荐列表的第一个物品，然后再交替选择，直到填满整个推荐列表。所以，最后得到的列表会是 ABABAB，或者 BABABA 这样的顺序，而且这两种形式出现的概率应该是相等的，这样才能保证两个模型的公平性。

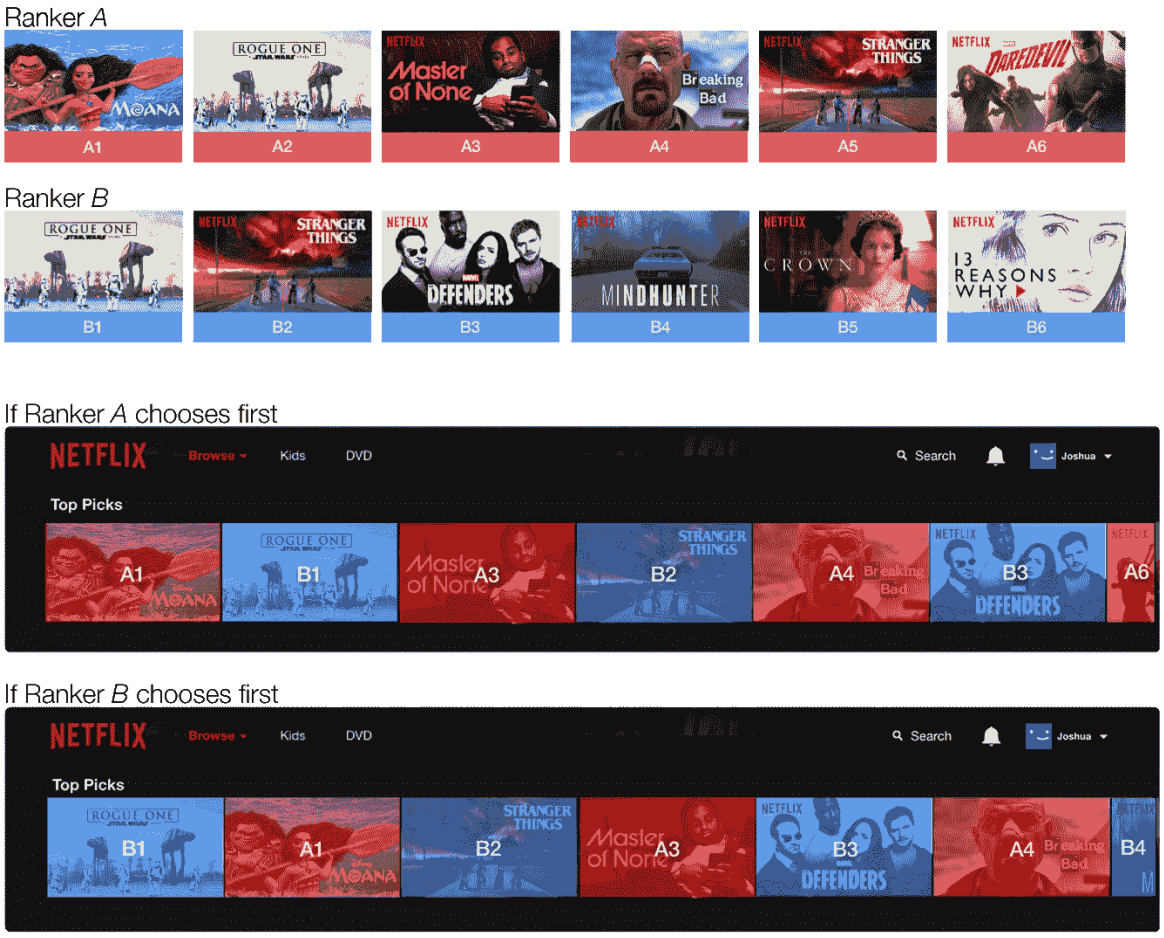

图6 Interleaving方法中推荐列表的生成方法

最后，我们要清楚推荐列表中的物品到底是由模型 A 生成的，还是由模型 B 生成的，然后统计出所有模型 A 物品的综合效果，以及模型 B 物品的综合效果，然后进行对比。这样，模型评估过程就完成了。

总的来说，Interleaving 的方法由于不用进行用户分组，因此比传统 A/B 测试节约了一半的流量资源。但是 Interleaving 方法能彻底替代传统 A/B 测试吗？其实也不能，在测试一些用户级别而不是模型级别的在线指标时，我们就不能用 Interleaving 方法。

比如用户的留存率，用户从试用到付费的转化率等，由于 Interleaving 方法同时使用了对照模型和实验模型的结果，我们就不清楚到底是哪个模型对这些结果产生了贡献。但是在测试 CTR、播放量、播放时长这些指标时，Interleaving 就可以通过累加物品效果得到它们。这个时候，它就能很好地替代传统的 A/B 测试了。

到这里，我们就形成了一个完整、高效且准确的评估系统。希望你能从整体的角度重新审视一遍这个体系中的每个方法，如果有不清楚的，再好好回顾一下我讲的知识点。

小结

这节课，我们利用之前讲过的知识，总结出了推荐系统的评估体系。这个评估体系由传统离线评估、离线 Replay、线上 Interleaving，以及线上 A/B 测试四个层级组成。这四个层级由下到上评估效率逐渐降低，但是评估的准确性逐渐升高，它们共同组成一个能够高效筛选候选模型的评估体系。

针对这个评估体系中的两个要点，离线 Replay 实践和 Interleaving 方法，我们又深入学习了它们的工程架构和实现细节。

其中，离线 Replay 借鉴了 Netflix 时光机的经验，这个时光机的数据流体系通过融合日志流和场景信息数据，生成天级别的数据快照，并对外提供统一的 API，供模型训练和评估使用，使用时就像做了一次时光旅行。

对于 Interleaving 方法，我们应该清楚它实现的三个要点：

1. 它不进行用户分组；
2. 它的实验推荐列表是通过间隔地选择模型 A 和模型 B 的推荐物品得到的；
3. 为了保证它的公平性，我们要从模型 A 或者模型 B 中随机选择第一个物品，就像野球场选人一样完成推荐列表的生成。

还是老习惯，我把这节课的重要知识点总结在了下面的表格里，方便你及时回顾。

知识	知识描述
推荐系统评估体系组成	传统离线评估、离线Replay、线上Interleaving、线上A/B测试
推荐系统评估体系作用	高效且准确地完成推荐系统评估工作，进而加快推荐模型的迭代
Netflix的Replay方案	它叫时光机，它的数据流体系通过融合日志流和场景信息数据， 生成天级别的数据快照，并对外提供统一的API，供模型训练和评估使用
Interleaving评估方法的三个要点	1.它不进行用户分组； 2.它的实验推荐列表的生成过程， 是通过间隔地选择模型A和模型B的推荐物品得到的 3.为了保证公平性，它要从模型A和模型B中，随机选择第一个物品

极客时间

这节课也是我们模型评估篇的最后一节课，希望通过整个模型评估篇的学习，你不仅能够熟悉起每一种评估方法，而且能够清楚它们之间的区别和联系，形成一个高效的评估体系。相信它会加快你模型迭代的速度，对你的实际工作产生非常积极的影响！

课后思考

在 Interleaving 方法中，推荐列表是由模型 A 和模型 B 的结果共同组成的，那如果模型 A 和模型 B 的结果中有重叠怎么办？是保留模型 A 的结果还是模型 B 的结果呢？你有什么好的想法吗？

今天讲的评估体系，你知道怎么建立了吗？欢迎把你的思考和疑问写在留言区，不妨也把这节课分享给你的朋友们，我们下节课见！