推荐系统评估

原创 gongyouliu 大数据与人工智能 2019-02-20

本文大约1万字，阅读需要45分钟左右，建议先收藏

作者在上篇文章《推荐系统的工程实现》中提到推荐系统要很好地落地到业务中，需要搭建支撑模块，其中效果评估模块就是其中非常重要的一个。本篇文章作者来详细说明怎么评估(Evaluating)推荐系统的效果，有哪些评估手段,在推荐业务中的哪些阶段进行评估，具体的评估方法是什么。借此希望更好地帮助大家在实际业务中实施推荐系统评估模块。

本篇作者主要从“什么是一个好的推荐系统”、“在推荐系统业务的各个阶段怎么评估推荐系统”，“推荐系统怎么更好地满足用户的诉求”的角度来讲解。至于推荐算法怎么帮助企业更好地服务于用户，并利用算法产生更多的商业价值，作者会单独写一篇文章《推荐系统的商业价值》来详细介绍，敬请期待。现在，我们从评估的目的、评估的常用指标、评估方法、评估需要关注的问题四个维度来详细说明。

1 推荐评估的目的

推荐系统评估是跟推荐系统的产品定位息息相关的,推荐系统是解决信息高效分发的手段，希望通过推荐更快更好地满足用户的不确定性需求。那么推荐系统的精准度、惊喜度、多样性等都是需要达到的目标，同时，推荐系统的稳定性，是否支撑大规模用户访问等方面也对推荐系统发挥价值起关键作用。当然，不是所有目标都能达到，评估后怎么做到各个目标的平衡，权衡得失，是推荐系统设计需要注意的问题。

推荐系统评估的目的就是从上面说的这么多的维度来评估出推荐系统的实际效果及表现,从中发现可能的优化点，通过优化推荐系统，期望更好地满足用户的诉求，为用户提供更优质的推荐服务，同时通过推荐获取更多的商业利益。

2 评估推荐的常用指标

怎么评估推荐系统呢？从哪些维度来评估推荐系统呢？这要从推荐系统解决的商业问题来思考，作者在《推荐系统介绍》这篇文章对推荐系统做了比较系统详细的介绍，推荐系统可以很好地解决“标的物”提供方、平台方(提供产品服务的公司)、用户三方的需求(见下面图1)，推荐系统作为嵌入产品的服务模块，它的评估可以从以下四个维度来衡量。

图1：推荐系统通过整合到产品中，为用户提供“标的物”推荐

1 用户的维度

用户最重要的诉求永远是更方便快捷地发现自己想要(喜欢)的“标的物”。推荐系统多好地满足了用户的这个诉求，用户就会多依赖推荐系统。一般来说，从用户维度有如下几类指标可以衡量推荐系统对用户的价值。

01 准确度

准确度评估的是推荐的“标的物“是不是用户喜欢的。拿视频推荐来说，如果推荐的电影用户点击观看了，说明用户喜欢，看的时间长短可以衡量用户的喜好程度。但是要注意，用户没看不代表用户不喜欢，也可能是这个电影用户刚在院线看过。这里所说的准确度更多的是用户使用的主观体验感觉。

02 惊喜度(serendipity)

所谓惊喜度，就是让用户有耳目一新的感觉，无意中给用户带来惊喜。举个例子，比如作者的朋友春节给我推荐了一部新上映的很不错的电影，但是作者忘记电影名字了，怎么也想不起来，但是突然有一天电视猫给我推荐了这部电影，这时作者会非常惊喜。这种推荐超出了用户的预期，推荐的不一定跟用户的历史兴趣相似，可能是用户不熟悉的，但是用户感觉很满意。

03 新颖性(novelty)

新颖性就是推荐用户之前没有了解过的“标的物”。人都是“喜新厌旧”的，推荐用户没接触过的东西，可以提升用户的好奇心和探索欲。

04 信任度(Confidence& Trust)

在现实生活中，如果你信任一个人，他给你推荐的东西往往你会关注或者购买。对推荐系统来说也是类似的，如果推荐系统能够满足用户的需求，用户就会信任推荐系统，会持续使用推荐系统来获取自己喜欢的“标的物”。

05 多样性

用户的兴趣往往是多样的，在做推荐时需要给用户提供多“品类”的“标的物”，以挖掘用户新的兴趣点，拓展用户的兴趣范围，提升用户体验。

06 体验流畅度

推荐系统是一个软件产品，用户的体验是否好，是否卡顿，响应是否及时，对用户的行为决策非常关键。
流畅的用户体验，是推荐服务的基本要求。但只要服务不稳定，响应慢，会极大影响用户体验，甚至导致用户卸载产品。
上面这些指标，有些是可以量化的(比如精准度、流畅度)，有些是较难量化的(比如惊喜度、新颖性)，所有这些指标汇聚成用户对推荐模块的满意度。

2 平台方的维度

平台方提供一个平台(产品)，对接“标的物”提供方和用户，通过服务好这两方来赚取商业利润。不同产品挣取利润的方式不同，有的主要从用户身上挣钱(比如视频网站，通过会员盈利)，有的从“标的物”提供方挣钱(比如淘宝，通过商家的提成及提供给商家的服务挣钱)，有的两者兼而有之，但大部分互联网产品都会通过广告挣钱(广告主买单，即所谓的”羊毛出在猪身上”)。不管哪种情况，平台方都要服务好用户和“标的物“提供方(有些产品平台方和”标的物”提供方是一样的，比如视频网站，是直接花钱购买视频版权的)。

对于平台方来说，商业目标是最重要的指标之一，平台方的盈利目的又需要借助用户来实现(不管是用户购买，还是广告，都需要有大量用户)，所以平台方除了关注绝对的收益外，还需要关注用户活跃、留存、转化、使用时长等用户使用维度的指标。

推荐系统怎么更好的促进收益增长，促进用户活跃、留存、转化等就是平台方最关注的商业指标。

同时，为第三方提供平台服务的平台方(如淘宝商城)，还需要考虑到商家生态的稳定发展。为好的商家提供获取更多收益的机会也是平台方的责任和义务。

所以，站在平台方角度看，最重要的指标主要有如下3类：

- 用户行为相关指标；
- 商业变现相关指标；
- 商家(即“标的物”提供方)相关指标；

我会在下一篇文章《推荐系统的商业价值》中详细探讨推荐系统的商业价值，本文不会过多讲解推荐系统的商业指标。

3 推荐系统自身维度

推荐系统是一套算法体系的闭环，通过该闭环为用户提供服务，从推荐系统自身来说，主要衡量指标包括如下：

01 准确度

作为推荐系统核心的推荐算法,本身是一种机器学习方法，不管是预测、分类、回归等机器学习问题都有自己的评估指标体系。推荐系统准确度的评估也可以自然而然的采用推荐算法所属的不同机器学习范式来度量，我们在第三部分会根据该方式来度量准确度指标。
关于准确度，第二部分会详细说明具体的评估方法，准确度也是学术界和业界最常用最容易量化的评估指标。

02 实时性

用户的兴趣是随着时间变化的，推荐系统怎么能够更好的反应用户兴趣变化，做到近实时推荐用户需要的“标的物”是特别重要的问题。特别像新闻资讯、短视频等满足用户碎片化时间需求的产品，做到近实时推荐更加重要。

03 鲁棒性

推荐系统一般依赖用户行为日志来构建算法模型，而用户行为日志中会包含很多开发过程中、系统、人为(比如黑客攻击)等产生的垃圾数据，推荐算法要具备鲁棒性，尽量少受“脏”的训练数据的影响，才能够为用户提供稳定一致的服务。

04 响应及时稳定性

用户通过触达推荐模块，触发推荐系统为用户提供推荐服务，推荐服务的响应时长，推荐服务是否稳定(服务正常可访问，不挂掉)也是非常关键的。

05 抗高并发能力

推荐系统是否能够承受高并发访问，在高并发用户访问下(比如双十一的淘宝推荐)，是否可以正常稳定的提供服务，也是推荐系统的重要能力。

除了上面说的这些指标外，推荐模型的可维护性、可拓展性、模型是否可并行训练、需要的计算存储资源、业务落地开发效率等也是推荐业务设计中需要考虑的重要指标。

4 标的物提供方的维度

“标的物”的提供方通过为用户提供“标的物”获取收益(如淘宝上的商家通过售卖物品获取收益)，怎么将自己更多的“标的物”更快的“卖出去”是“标的物”提供方的诉求。评估推荐系统为“标的物”提供方创造价值的指标除了下面的覆盖率和挖掘长尾能力，还有更多的商业化指标，这里不做过多说明，作者会在下篇文章《推荐系统的商业价值》中详细讲解。

01 覆盖率

从“标的物”提供方的角度来看，希望自己提供的“标的物”都能够被用户“相中”，不然这个“标的物”就没有任何价值。所以推荐系统需要将更多的“标的物”推荐(曝光)出去，只有曝光出去，才有被用户“消费”的可能。

02 挖掘长尾的能力

推荐系统的一个重要价值就是发现长尾(长尾理论是ChrisAnderson提出的,不熟悉该理论的读者可以自行百度或者看ChrisAnderson出的《长尾理论》一书)，将小众的“标的物”分发给喜欢该类“标的物”的用户。度量出推荐系统挖掘长尾的能力，对促进长尾“标的物”的“变现”及更好地满足用户的小众需求从而提升用户的惊喜度非常有价值。

3 推荐系统的评估方法

上一节列举了很多评估推荐系统的指标,并对指标的含义做了简要说明。本节我们具体讲解怎么度量(量化)这些指标。

推荐算法本质上就是一个机器学习问题。我们需要构建推荐算法模型，选择认为合适的(效果好的)的算法模型，将算法模型部署到线上推荐业务中，利用算法模型来预测用户对“标的物”的偏好，通过用户的真实反馈(是否点击、是否购买、是否收藏等)来评估算法效果。同时,在必要(不一定必须)的时候，需要跟你的用户沟通，收集用户对推荐系统的真实评价,整个过程可以用如下的图2来说明。我们可以根据推荐业务流的时间线按照先后顺序将推荐系统评估分为三个阶段：离线评估、在线评估、主观评估。在下面我们会按照这三个阶段来讲解上一节的评估指标是怎么嵌入到这3个阶段当中的,并说明具体的评估方法。

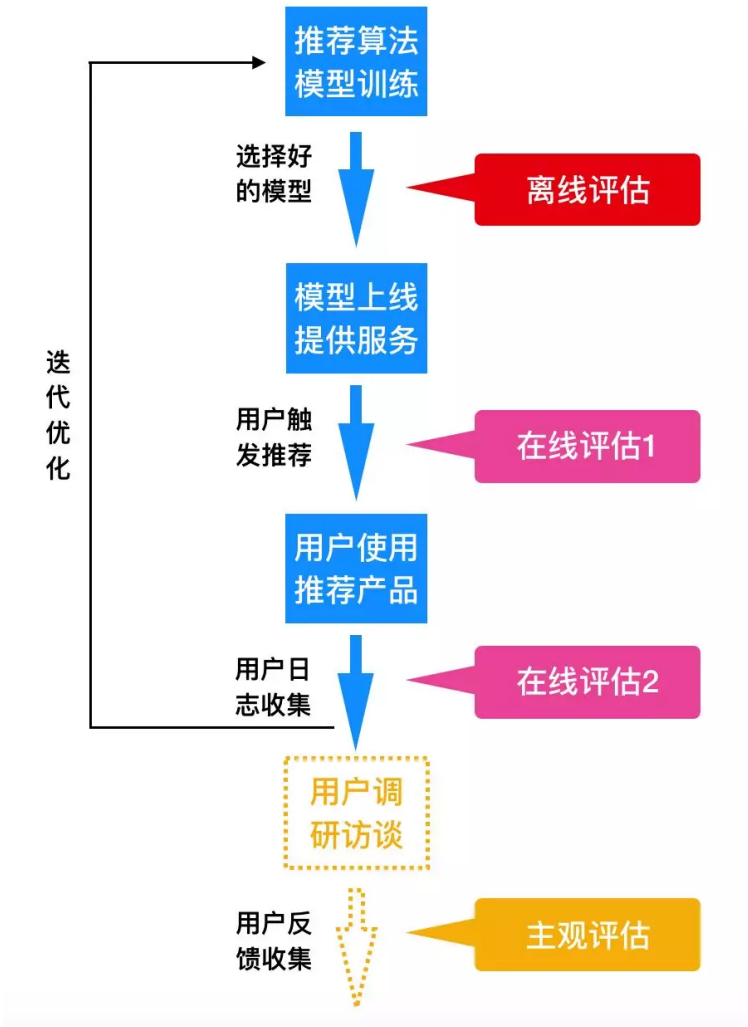

图2：根据推荐业务流，将推荐评估分为3个阶段

1 离线评估

离线评估是在推荐算法模型开发与选型的过程中对推荐算法做评估,通过评估具体指标来选择合适的推荐算法,将算法部署上线为用户提供推荐服务。具体可以评估的指标有：

01 准确度指标

准确度评估的主要目的是事先评估出推荐算法模型的好坏(是否精准),为选择合适的模型上线服务提供决策依据。我们期望精准的模型上线后产生好的效果。这个过程评估的是推荐算法是否可以准确预测用户的兴趣偏好。

准确度评估是学术界和业界最重要和最常用的评估指标，可以在模型训练过程中做评估，因此实现简单，可操作性强，方便学术交流与各类竞赛作为评比指标，同时通过评估可以对比不同模型的效果。

推荐算法是机器学习的分支,所以准确度评估一般会采用机器学习效果评估一样的策略。一般是将训练数据分为训练集和测试集,用训练集训练模型,用测试集评估模型的预测误差，这个过程可以参见下面的图3。

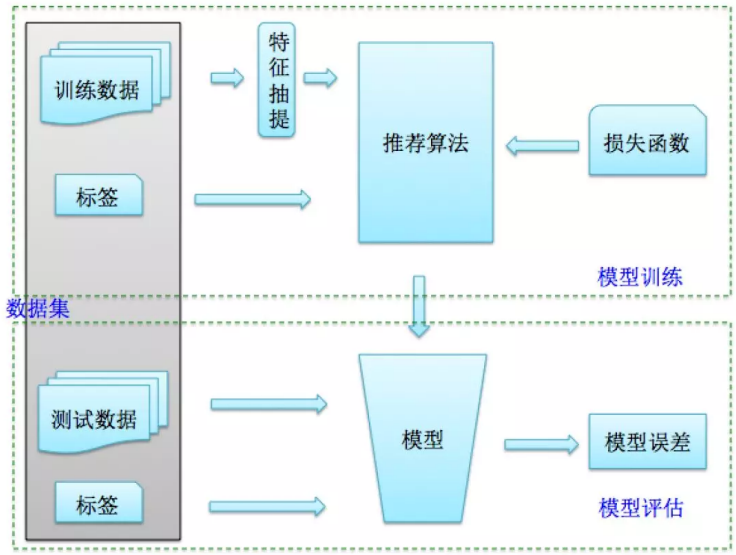

图3：推荐算法的模型训练与离线评估

具体怎么计算推荐算法模型误差(准确度),可以根据推荐算法模型的范式来决定采用不同的评估方法，这里我们主要根据三种不同范式来评估准确度。

第一种是将推荐算法看成预测(回归)问题。预测用户对“标的物”的评分(比如0~10分)。
第二种是将推荐算法看成是分类问题。可以是二分类，将“标的物“分为喜欢和不喜欢两类；也可以是多分类，每个“标的物”就是一个类，根据用户过去行为预测下一个行为的类别(如YouTube在2016发表的深度学习推荐论文DeepNeural Networks for YouTube Recommendations就是采用多分类的思路来做的)。
第三种是将推荐系统算法看成一个排序学习问题，利用排序学习(Learningto rank)的思路来做推荐。

推荐系统的目的是为用户推荐一系列“标的物”，击中用户的兴奋点，让用户“消费”“标的物”。所以，在实际推荐产品中，我们一般都是为用户提供N个候选集，称为TopN推荐，尽可能的召回用户感兴趣的“标的物”。上面这三类推荐算法范式都可以转化为TopN推荐。第一种思路预测出用户对所有没有行为的“标的物”的评分，按照评分从高到低排序，前面N个就可以当做TopN推荐(得分可以看成是用户对“标的物”的偏好程度，所以这样降序排列取前N个的做法是合理的)。第二种思路一般会学习出在某个类的概率，根据概率值也可以类似第一种思路来排序形成TopN推荐。第三种思路本身就是学习一个有序列表。

下面来详细讲解怎么按照推荐算法的上述3种范式来评估算法的准确度。

1、推荐算法作为评分预测模型

针对评分预测模型，可以评估的准确度指标主要有：RMSE(均方根误差)、MAE(平均绝对误差)。他们的计算公式分别是：

$$\mathrm{RMSE}=\frac{\sqrt{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}}{|T|}$$

$$\mathrm{MAE}=\frac{\sum_{u,i\in T}|r_{ui}-\hat{r}_{ui}|}{|T|}$$

其中，u代表用户，i代表“标的物”，T是所有有过评分的用户。r_{ui}是用户u对“标的物”i的真实评分，$\hat{r}_{ui}$是推荐算法模型预测的评分。其中RMSE就是Netflix在2006年举办的“NetflixPrize”大赛的评估指标。
常用的矩阵分解推荐算法(及矩阵分解算法的推广FM、FFM等)就是一种评分预测模型。

2、推荐算法作为分类模型

针对分类模型，评估推荐准确度的主要指标有：准确率(Precision)、召回率(Recall)。
假设给用户u推荐的候选集为$R_u(N)$(通过算法模型为用户推荐的候选集),用户真正喜欢的“标的物”集是A_u(在测试集上用户真正喜欢的“标的物”)，总共可通过模型推荐的用户数为集合U。其中N是推荐的数量。准确率是指为用户推荐的候选集中有多少比例是用户真正感兴趣的(“消费”过“标的物”)，召回率是指用户真正感兴趣的“标的物”中有多少比例是推荐系统推荐的。针对用户u，准确率(P_u)、召回率(R_u)的计算公式分别如下：

$$P_u=\frac{|R_u(N)\bigcap A_u|}{|R_u(N)|}$$

$$R_u=\frac{|R_u(N)\bigcap A_u|}{|A_u|}$$

一般来说N越大(即推荐的“标的物“越多)，召回率越高，精确度越低，当N为所有“标的物”时，召回率为1，而精确度接近0(一般推荐系统“标的物”总量很大，而用户喜欢过的量有限，所以根据上面公式，精确度接近0)。
对推荐系统来说，当然这两个值都越大越好，最好是两个值都为1，但是实际情况是这两个值就类似量子力学中的测不准原理(你在同一时间无法知道粒子的位置和速度)，你无法保证两者的值同时都很大，实际构建模型时需要权衡，一般我们可以用两者的调和平均数($F1_u$)来衡量推荐效果，做到两者的均衡。

$$F1_u=\frac{2}{\frac{1}{P_u}+\frac{1}{R_u}}=\frac{2P_u\cdot R_u}{P_u+R_u}$$

上面只计算出了推荐算法对一个用户u的准确率、召回率、F1值。整个推荐算法的效果可以采用所有用户的加权平均得到，具体计算公式如下：

$$Precision=\frac{\sum_{u\in U}P_u}{|U|}$$

$$Recall=\frac{\sum_{u\in U}R_u}{|U|}$$

$$F1=\frac{\sum_{u\in U}F1_u}{|U|}$$

关于分类问题的评估方法，可以参考周志华的《机器学习》第二章“模型评估与选择”，里面有很多关于分类问题评估指标的介绍。作者这里就不详细介绍了。

3、推荐算法作为排序学习模型

上面两类评估指标都没有考虑推荐系统在实际做推荐时将“标的物”展示给用户的顺序，不同的排序用户的实际操作路径长度不一样，比如智能电视端一般通过遥控器操作，排在第二排推荐的电影，用户点击就要多操作遥控器按键几次。我们当然是希望将用户最可能会“消费”的“标的物”放在用户操作路径最短的地方(一般是最前面)。所以，推荐的“标的物“展示给用户的顺序对用户的决策和行为是有很大影响的,那怎么衡量这种不同排序产生的影响呢？这就需要借助排序指标,这类指标我们这里主要介绍MAP(MeanAverage Precision)，其他指标如NDCG(NormalizedDiscounted Cumulative Gain)，MRR(MeanReciprocal Rank)等读者可以自行了解学习，这里不再介绍。

MAP的计算公式如下：

$$MAP = \frac{1}{|U|} AP_u$$

其中，$AP_u = \sum_{u=1}^{|U|} \frac{1}{n_u} \sum_{i=1}^{n_u} \frac{i}{l_i}$，所以，

$$MAP = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{n_u} \sum_{i=1}^{n_u} \frac{i}{l_i}$$

其中，AP_u代表的是为用户u推荐的平均准确率，U是所有提供推荐服务的用户的集合；n_u是推荐给用户u，而用户u喜欢的“标的物”的数量(比如推荐20个视频给用户u，用户看了3个，那么n_u=3)；l_i是用户u喜欢的第i个“标的物”在推荐列表中的排序(比如给用户推荐20个视频,用户喜欢的第2个在这20个视频的推荐列表中排第8位，那么l_i=8)。

为了方便读者理解，这里举个搜索排序的例子(MAP大量用于搜索，推荐排序的效果评估中)。假设有两个搜索关键词，关键词1有3个相关网页，关键词2有6个相关网页。某搜索系统对于关键词1检索出3个相关网页(将所有相关的都检索出来了)，其在搜索结果中的排序分别为2,3,6；对于关键词2检索出2个相关网页(6个相关中只检索出了2个)，其在搜索列表中的排序分别为4,8。对于关键词1，平均准确率为(1/2+2/3+3/6)/3=0.56。对于关键词2，平均准确率为(1/4+2/8)/6=0.08。则MAP=(0.56+0.08)/2=0.32。

至此，关于离线评估的准确度指标已经介绍完了。下面介绍一下其他可以在离线阶段评估的指标。

02 覆盖率指标

对于任何推荐范式，覆盖率指标都可以直接计算出来。覆盖率(Coverage)的具体计算公式如下：

$$Coverage = \frac{\left|\bigcup_{u \in U} R_u\right|}{|I|}$$

其中U是所有提供推荐服务的用户的集合，I是所有“标的物”的集合，R_u是给用户u的推荐“标的物”构成的集合。

03 多样性指标

用户的兴趣往往是多样的，并且有些产品面对的用户也不止一个(比如智能电视前可能是一家人看电视)，同时人在不同的时间段可能兴趣也不一样(早上看新闻，晚上看电视剧)，个人兴趣也会受心情、天气、节日等多种因素影响。所以我们在给用户做推荐时需要尽量推荐多样的“标的物”，让用户从中找到自己感兴趣的，种类更多样的话，总有一款能够击中用户的兴趣点。

在具体推荐系统工程实现中，可以通过对“标的物”聚类(可以用机器学习聚类或者根据标签等规则来分类)，在推荐列表中插入不同类别的“标的物”的方式来增加推荐系统推荐结果的多样性。

04 实时性指标

用户的兴趣是随着时间变化的，推荐怎么能尽快反应用户兴趣变化，捕捉用户新的兴趣点在日益竞争激烈的互联网时代对产品非常关键，特别是新闻、短视频这类APP，需要快速响应用户的兴趣变化。
一般来说，推荐系统的实时性分为如下四个级别T+1(每天更新用户推荐结果)、小时级、分钟级、秒级。越是响应时间短的对整个推荐系统的设计、开发、工程实现、维护、监控等要求越高。下面我们给大家提供一些选型的建议。

1. 对于“侵占”用户碎片化时间的产品，如今日头条、快手等。这些产品用户“消耗”“标的物”的时间很短，因而建议推荐算法做到分钟级响应用户兴趣变化；
2. 对于电影推荐、书推荐等用户需要消耗较长时间“消费”标的物的产品，可以采用小时级或者T+1策略；
3. 一般推荐系统不需要做到秒级，但是在广告算法中做到秒级是需要的。

上述这些建议不是绝对的，不同的产品形态，不同的场景可能对实时性级别的需求不相同。大家可以根据自己的产品特色、业务场景、公司所在的阶段、公司基础架构能力、人力资源等综合评估后做选择。但是最终的趋势肯定是趋向于越来越实时响应用户需求。

05 鲁棒性指标

推荐系统是否受脏数据影响，是否能够稳定的提供优质推荐服务非常关键。为了提升推荐系统的鲁棒性，这里提四个建议。

1. 尽量采用鲁棒性好的算法模型；
2. 做好特征工程，事先通过算法或者规则等策略剔除掉可能的脏数据；

3 在日志收集阶段，对日志进行加密，校验，避免人为攻击等垃圾数据引入；

4 在日志格式定义及日志打点阶段，要有完整的测试case，做好冒烟回归测试，避免开发失误或者bug引入垃圾数据。

06 其他指标

另外，像模型训练效率，是否可以分布式计算(可拓展性)，需要的计算存储资源等都可以根据所选择的模型及算法提前预知，这里不再细说。

2 在线评估

根据上面图2，推荐系统的在线评估可以分为两个阶段，其实这两个阶段是连接在一起的，这里这样划分主要是方便对相关的评估指标做细分讲解。下面分别来讲解每个阶段可以评估哪些指标及具体的评估方法。

在线评估第一阶段

第一阶段是推荐算法上线服务到用户使用推荐产品这个阶段,在这个阶段用户通过使用推荐产品触发推荐服务(平台通过推荐接口为用户提供服务)。这个阶段可以评估的指标有：

01 响应及时稳定性指标

该指标是指推荐接口可以在用户请求推荐服务时及时提供数据反馈,当然是响应时间越短越好，一般响应时间要控制在200ms之内，超过这个时间人肉眼就可以感受到慢了。

服务器响应会受到很多因素影响，比如网络、CDN、Web服务器、操作系统、数据库、硬件等，一般无法保证用户的每次请求都控制在一定时间内。我们一般采用百分之多少的请求控制在什么时间内这样的指标来评估接口的响应时间(比如99%的请求控制在50ms之内)。

那怎么量化服务器的响应情况呢？我们可以在web服务器(如Nginx)端对用户访问行为打点，记录用户每次请求的时长(需要在web服务器记录/配置接口请求响应时长)，将web服务器的日志上传到大数据平台，通过数据分析可以统计出每个接口的响应时长情况。一般公司会采用CDN服务来缓存、加速接口，上述从web服务器统计的时长，只能统计接口回源部分的流量，被CDN扛住的部分流量的响应时长是需要CDN厂商配合来统计的。另外，上面统计的web服务器响应时长只是web服务消耗的时长，用户从触发推荐到返回结果，除了web服务器的响应时长，还要加上web服务器到用户APP这中间的网络传输时长和APP处理请求渲染展示出来的时长，这部分时间消耗需要采用其他技术手段来计算统计，这里不再细说。

02 抗高并发能力指标

当用户规模很大时，或者在特定时间点有大量用户访问(比如双十一的淘宝)时，在同一时间点有大量用户调用推荐服务，推荐接口的压力会很大，推荐系统能否抗住高并发的压力是一个很大的挑战。

我们可以在接口上线前对接口做打压测试，事先了解接口的抗并发能力。另外可以采用一些技术手段来避免对接口的高并发访问，比如增加缓存，web服务器具备横向拓展的能力，利用CDN资源，在特殊情况下对推荐服务进行分流、限流、降级等。

上述两个指标，作者只做了相对简单的介绍，作者会在后续文章《推荐系统的高可用高并发架构设计》中对这些点做详细讲解，敬请期待。

在线评估第二阶段

第二阶段是用户通过使用推荐算法产生行为(购买、点击、播放等)，我们通过收集分析用户行为日志来评估相关的指标。这一阶段我们主要站在平台方角度来思考指标，主要有用户行为相关指标、商业化指标、商家相关指标，这里我们只介绍用户行为相关指标。另外说下,像离线评估中所介绍的一些准确度指标(如准确率、召回率等)其实可以通过适当的日志打点来真实的统计出来，计算方式类似，这里也不再细说。

推荐模型上线提供推荐服务后,最重要的用户行为指标有转化率、购买率、点击率、人均停留时长、人均阅读次数等，一般用户的行为是一个漏斗(例如，推荐曝光给用户->用点击浏览->用户扫码->用户下单，参考下面的图4),我们需要知道从漏斗一层到下一层的转化率。漏斗模型可以非常直观形象的描述用户从前一个阶段到下一个阶段的转化,非常适合商业上定位问题，通过优化产品流程，提升用户在各个阶段的转化。

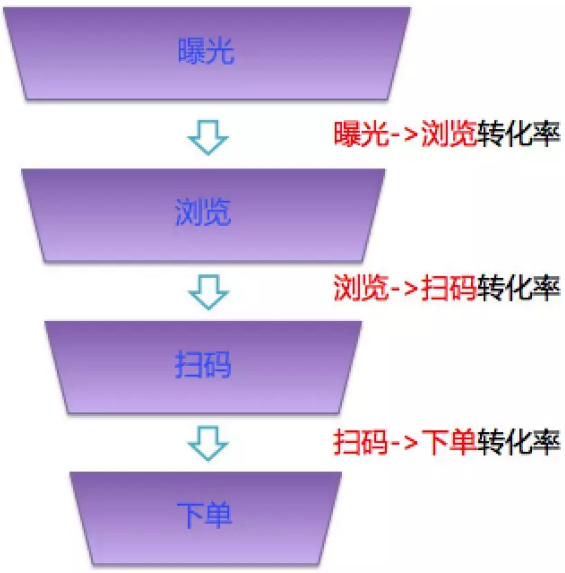

图4：用户行为的漏斗模型

线上评估一般会结合AB测试技术，当采用新算法或者有新的UI交互优化时，将用户分为AB两组，先放一部分流量给测试组(有算法或UI优化的组)，对比组是优化之前的组。如果测试组与对比组在相同指标上有更好的表现,显著(具备统计显著性)提升了点击或者转化，并且提升是稳定的，后续逐步将优化拓展到所有用户。这种借助AB测试小心求证的方法，可以避免直接一次性将新模型替换旧模型，但是上线后效果不好的情况发生（会严重影响用户体验和收益指标，造成无法挽回的损失）。

另外,针对用户行为指标,我们需要将推荐算法产生的指标与大盘指标(用户在整个产品的相关指标)对比,可以更好地体现推荐算法的优势(比如通过推荐系统产生的人均播放次数和人均播放时长比大盘高，就可以体现推荐的价值)，让推荐系统和推荐工程师的价值得到真正的体现，也可以让管理层从数据上了解推荐的价值。

最后，通过日志分析，我们可以知道哪些“标的物”是流行的，哪些是长尾。拿视频推荐来举例，我们可以根据二八定律，将电影播放量降序排列，播放量占总播放量80%的前面的电影，算作热门电影，后面的当做长尾(参考下面图5)。
在度量推荐系统长尾能力时，我们可以从如下三个维度来度量：

1. 所有长尾“标的物”中每天有多少比例被分发出去了；
2. 有多少比例的用户，推荐过了长尾“标的物”；
3. 长尾内容的转化情况和产生的商业价值；

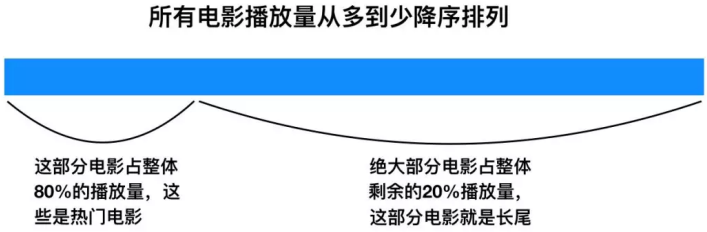

图5：对于电影推荐，长尾的定义

3 主观评估

第二节提到了很多用户维度的指标，如准确度、惊喜度、新颖性、信任度、体验流畅度等。这些指标有很多是用户的使用主观感受(如惊喜度)，有些指标也因人而异(如新颖性)，有些很难利用已知的数据来量化(如信任度)。

针对上面这些指标,我们可以通过主观评估的方式来获得用户对推荐系统的真实评价。具体的方式可以是用户问卷调查、电话访谈、跟用户直接见面沟通等。这些方式可以很直接直观的知道

用户对推荐产品的反馈和想法，是很重要的一种评估推荐系统的补充方式。主观评估要想真实的发现推荐系统存在的问题，需要注意很多问题，下面针对主观评估做如下5点说明，作为主观评估有效执行的指导建议。

1. 主观评估是很消耗时间的，特别是电话沟通和见面访谈，即使是问卷调查，也需要很好地设计问卷的问题；
2. 让用户参与主观评估，往往需要给用户一定的好处，需要一定的资金支持；
3. 需要确保选择的样本有代表性，能够真实的代表产品的用户，所以选择的样本量不能太少，抽样方法也需要科学选择；
4. 设计问卷时，最好不要直接问“你觉得我们的推荐系统有惊喜度吗？”这样的问题，而要“我们的推荐系统给你推荐了哪些你特别想看，但是一直通过其他渠道没有发现的电影？”这样问，具体怎么设计问卷可以参考相关的专业书籍；
5. 用户访谈或者电话沟通时，用户的回答不一定是真实的想法，用户真实的想法可能不好意思表现出来，或者会选择讨好你的回答方式(毕竟参与调研的用户多少获取了一定的物质报酬)，调研者需要特别注意，采用一定的沟通技巧，尽量真实挖掘出用户的想法；

4 推荐系统评估需要关注的问题

推荐系统评估要想落地取得较好的效果，真实的反馈推荐系统的问题，为推荐系统提供优化的建议，必须要关注以下问题。

01 离线评估准确度高的模型，在线评估不一定高

离线评估会受到可用的数据及评估方法的影响，同时，模型上线会受到各种相关变量的干扰，导致线上评估跟离线评估结果不一致。所以有必要引入AB测试减少新算法上线对用户体验的影响；

02 推荐系统寻求的是一个全局最优化的方案(解)

在实际情况中，经常会有老板或者产品经理来找你，说某个推荐怎么怎么不准，虽然作为推荐算法工程师，需要排查是否真有问题，但是也要注意，推荐模型求解是满足整体最优的一个过程(推荐算法如矩阵分解就是将所有用户行为整合进来作为目标函数，再求解误差最小时用户对未知“标的物”的评分)，不能保证每个用户都是预测最准的。所以，遇到上述情况要做适当判断，不要总是怀疑算法。举个不太恰当的例子，推荐系统对部分用户可能推荐不准就像和谐社会的发展，虽然人民的整体生活是越来越好的，但是还是有人生活在水深火热中。

03 推荐系统是一个多目标优化问题

推荐系统需要平衡很多因素(商业、用户体验、技术实现、资金、人力等)，怎么做好平衡是一种哲学。在公司不同阶段，倾向性也不一样，创业前期可能以用户体验为主，需要大力发展用户，当用户量足够多后，可能会侧重商业变现(推荐更多的付费视频，在搜索列表中插入较多广告等)，尽快让公司开始盈利。

04 AB测试平台对推荐评估的巨大价值

推荐系统在线评估强烈依赖于AB测试来得出信服的结论，所以一套完善的推荐系统解决方案一定要保证搭建一套高效易用的AB测试框架，让推荐系统的优化有据可循，通过数据驱动来让推荐系统真正做到闭环。

05 重视线上用户行为及商业变现方面的评估

线上评估更能真实反映产品的情况，所以在实际推荐系统评估中，要更加重视线上效果评估，它能够很好的将用户的行为跟商业指标结合起来，它的价值一定大于线下评估，需要推荐开发人员及相关产品经理花费更多的时间和精力。

写在最后

至此，关于推荐系统评估的所有方面都讲完了，希望本文可以作为大家在实践推荐系统评估模块的参考指南。由于搜索、推荐及计算广告算法与本业务的相似性，本文也可以作为搜索、计算广告落地评估的参考。由于作者精力能力有限，不当之处，请批评指正！

♥ 精彩推荐

推荐系统介绍

推荐系统的工程实现

大数据营销之用户画像