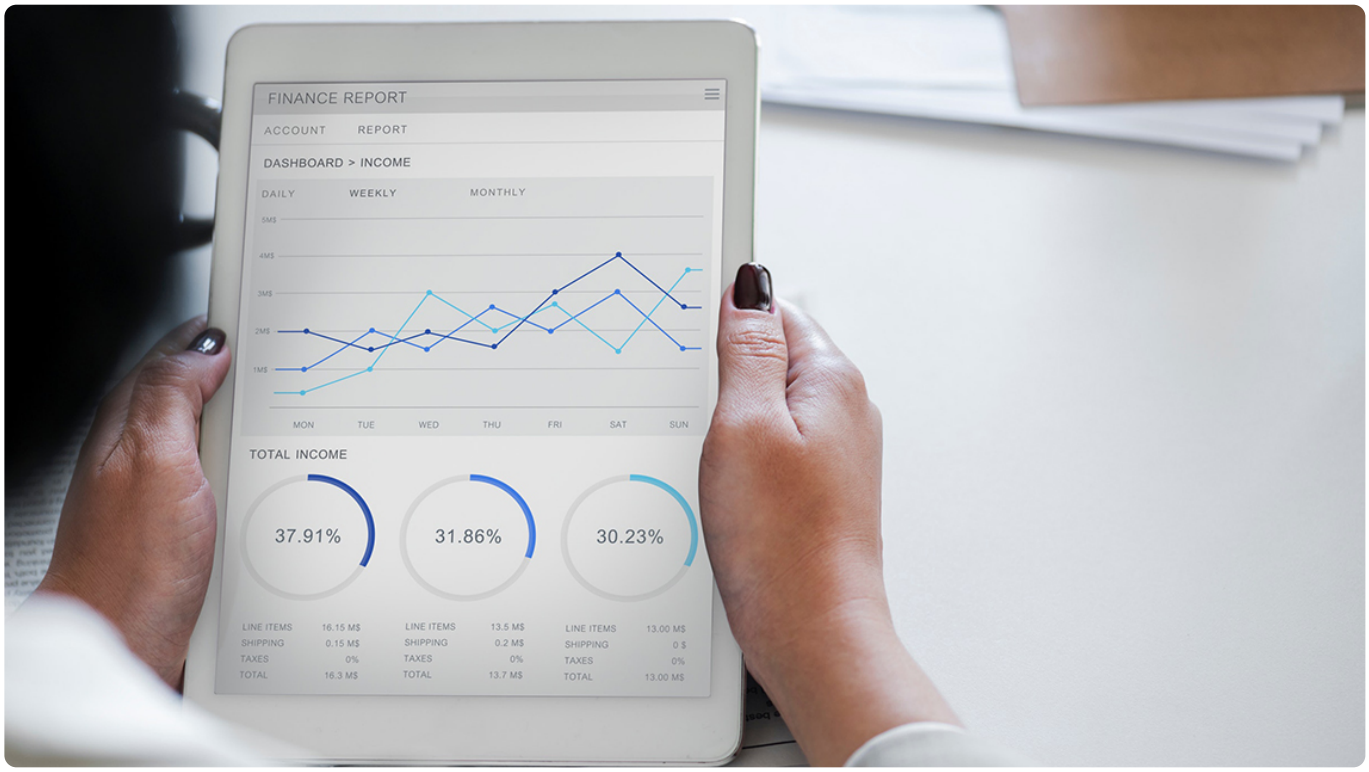
【效果保证】推荐系统的测试方法及常用指标介绍

2018-05-14 刑无刀

推荐系统三十六式 进入课程 >



讲述：黄洲君
时长 16:47 大小 7.69M

当我们刚开始学习推荐系统的时候，我就希望你想清楚为什么要做推荐系统。在逐渐深入的过程中，我开始唠叨推荐系统的林林总总。

到了今天，假如你已经有了自己的推荐系统，这个系统已经上线，代替了以前绝大多数人工的工作，夜以继日地工作，为电商网站创造销售额，为信息流创造阅读时间和互动，为社交网站创造社交关系。

为什么要关注指标

然而，这样你就可以安心睡大觉了吗？显然你想错了，它成功上线时，也是你失业的时候，我们暂且不说是否真的有这一天。就算是一切正常运作，你还是需要每天把这个系统捧在手心，教它在刁钻的用户面前如何长大，既要小心它学坏，也要小心它偷懒不学无术。

总之，养过孩子的人会懂的。面对推荐系统这样一个有诸多复杂因素联动起作用的系统，要时时刻刻知道它好不好，健不健康，你同样需要掌握一些测试方法和检测指标。

推荐系统的测试方法

在最开始几篇中，我说过你需要有不确定性思维，但是这绝不是帮你在老板那里开脱的说辞。

推荐系统也需要测试，只是它不同于传统的功能测试。传统软件的功能测试，功能的响应是有预期的，点击一个加关注按钮，应该有什么响应，是被产品文档明确规定的；也因此在开发功能的时候，可以同步写出测试用例来。

这非常明白，在功能开发时，你做了任何改动，只要跑一下测试用例，逻辑对不对就一目了然了。反观推荐系统就没那么容易了，你什么都没动，可能两次推荐的结果都有可能不一样，而且很可能这个不一样也是你自己或者你老板要求的。

那么推荐系统要怎么测试呢？与其说推荐系统没有确定性的预期响应，不如说推荐系统的响应维度更高。

因为确定性的功能响应像是一个点，而推荐系统的响应则是高维空间中的一个区域，而不是一个点。那么是不是推荐系统不需要单元测试了呢？显然也不是。

归纳起来，推荐系统的测试方法有四种： 业务规则扫描、离线模拟测试、在线对比测试、用户访谈。

1. 业务规则扫描

首先，业务规则扫描本质上就是传统软件的功能测试。确定的业务规则会对应有确定的规则，这些规则就可以翻译成单元测试，像是运行单元测试那样，对推荐系统逐一扫描业务规则。

通常这些业务规则对测试的要求也有“软的”和“硬的”两种。前者会对业务规则违反情况做一个基线规定，比如触发几率小于万分之一，在扫描测试时统计触发次数，只要统计触发几率不超过基线，就算是合格。

而硬的规则，就是一票否决，例如一些业务黑名单，简直就是高压线，测试时碰不得，碰了就是 Bug，就要想办法修正。

除了业务规则，还有一些容易被人忽视的地方，比如绝大多数推荐模型都涉及了数学计算，而数学计算中也有一些潜在的规则不能违反。

比如除数不能为 0，比如计算机的浮点数精度有限，经过一些指数运算后可能就出现预期之外的结果，还可能一些连续相乘的计算要防止出现 0 的乘数，类似这些在计算中的潜在业务规则，也需要扫描测试。

2. 离线模拟测试

其次，就是在离线模拟测试。这是一种军事演习式的测试。模拟测试当然无法代替真实数据，但是也能暴露一些问题。通常做法是先收集业务数据，也就是根据业务场景特点，构造用户访问推荐接口的参数。

这些参数要尽量还原当时场景，然后拿这些参数数据去实时访问推荐推荐，产生推荐结果日志，收集这些结果日志并计算评测指标，就是离线模拟测试。

显然，离线模拟测试是失真的测试，并且评测指标也有限，因为并不能得到用户真实及时的反馈。但是仍然有参考意义。

这些模拟得到的日志可以统称为曝光日志，它可以评测一些非效果类指标，例如推荐覆盖率，推荐失效率，推荐多样性等。关于这些指标具体含义，稍后再讲。那是不是离线模拟测试就对效果一无所知、无法模拟呢？

也并不是，有一种办法是，利用历史真实日志构造用户访问参数，得到带评测接口的结果日志后，结合对应的真实反馈，可以定性评测效果对比。

比如，可以评测推荐结果的 TopK 的准确率，或者排序效果 AUC。这些模型效果类指标，虽然不能代表最终关注的商业指标，但是两者之间一般存在一定的相关性。

通常来说 TopK 准确率高，或者 AUC 高于 0.5 越多，对应的商业指标就会越好，这是一个基本假设。通过离线模拟评测每一天的模型效果指标，同时计算当天真实的商业指标，可以绘制出两者之间的散点图，从而回归出一个简单的模型，用离线模型效果预估上线后真实商业指标。

3. 在线对比测试

第三种测试方法就是真正的实战了，那就是 ABTest，即在线对比测试，分流量做真实的评测。这需要一个支持流量正交切分的 ABTest 框架，在前面的文中已经讲到过。ABTest 在样本充分的前提下，基本上可以定性新的推荐系统是否比老的推荐系统更加优秀。

4. 用户访谈

最后一种测试方法就是用户访谈，或者说用户调查。前面三种测试方法，背后的思想是数据驱动。

然而正如我在本文开篇时所说，数据测量的是系统外在表现，并不反映系统原理，而且数据指标是人设计的，是存在主观性和片面性的，人的认知广度和深度各有不同。

因此，除了要紧紧团结在“数据驱动”这个核心思想周围，还需要深入用户，对用户做最直接的交流，对用户访谈，更重要的意义不是评测推荐系统，而是评测推荐系统的指标，设计是否合理，是否其高低反映了你预先的设定。

除此之外，通过前面三种测试方法如果得知系统表现不好，那么结合直接真实的用户调查和访谈，可以为系统优化找到真实原因。这种方法类比一下就是：维修下水道时，你需要钻进下水道。

常用指标

推荐系统有很多指标。你之前如果阅读过一些介绍推荐系统指标的文献或书籍，想必会对繁多的指标望而却步，总之就是各种率。实际上所有指标就是在回答两个问题：系统有多好，还能好多久？

这两个问题恰恰就是推荐系统里面一个老大难问题的反映：探索利用问题。

系统有多好？这就是想问问：对数据利用得彻底吗？还能好多久？这个问题就是想问问：能探索出用户新的兴趣吗？这样就能继续开采利用了。也好比在职场中看一个人，除了看他现在的经验和解决问题能力有多强，还要看他学习能力有多强，毕竟世界是变化的，朝阳也会变成夕阳。

下面我分别说说这两类指标有哪些。

1. 系统有多好?

检测系统到底有多好，其实，也有两类，一类是深度类，一类是广度类。

把数据看做是一座矿山，推荐系统是一个开采这座矿山的器械，“系统有多好”这个问题就是在关心开采得好不好，所以其实就看现有矿山上开采得深不深，开采得到不到位。广度类指标就是指在矿山上打满了钻井，而不仅仅盯着一处打钻井。

深度类指标，就是看推荐系统在它的本职工作上做得如何。还记得推荐系统的本职工作是什么吗？就是预测用户和物品之间的连接，预测的方法又有评分预测和行为预测。

因此深度类指标就指在检测系统在这两个工作上是否做得到位，有针对离线模型的指标，也有在线的指标，下面我分别说一说。

1. 评分准确度。通常就是均方根误差 RMSE，或者其他误差类指标，反映预测评分效果的好坏。在讲协同过滤时已经详细说过这个指标。这里不再赘述。

2. 排序。检测推荐系统排序能力非常重要，因为把用户偏爱的物品放在前面是推荐系统的天职。

由于推荐系统输出结果是非常个人化的，除了用户本人，其他人都很难替他回答哪个好哪个不好，所以通常评价推荐系统排序效果很少采用搜索引擎排序指标，例如 MAP，MRR，NDCG。

搜索引擎评价搜索结果和查询相关性，具有很强的客观属性，可以他人代替评价。推荐系统评价排序通常采用 AUC。也在前面介绍 BPR 模型时，专门讲到过。

3. 分类准确率。这个指标也是针对行为预测的，而行为预测就是分类问题，所以评价准确度就很自然。

在推荐系统中，评价准确度略微特殊，一般评价 TopK 准确率，与之对应还有 TopK 召回率，这里的 K 和实际推荐系统场景有关，就是实际每次推荐系统需要输出几个结果。

TopK 准确度计算方式如下：

如果日志中用户有 A、B 两个物品有正反馈行为，推荐系统推出一个物品列表，长度为 K，这个列表中就有可能包含 A、B 两个物品中的一个或多个，下面这个表格就说明了 TopK 准确率和 TopK 召回率的含义。

K	推荐输出	包含用户反馈物品数	TopK准确率	TopK召回率
1	A	1	100%	100%
2	A, C	1	50%	50%
3	A, C, D	1	33%	50%
4	A, C, D, E	1	25%	50%
4	A, B, C, D	2	50%	100%

这三个指标，比较直观地反映了推荐系统在“预测”这件事上对数据开采的深度，实际上由于模型不同，还可以有不同的指标，也可以自己设计指标，这里不再赘述。但这三个指标也属于比较初期的指标，距离最终商业指标还有一定的距离。

通常检测推荐系统的商业指标有：点击率，转化率。其实把用户从打开你的应用或者网站开始，到最终完成一个消费，中间要经历数个步骤，也是大家常说的漏斗转化过程。

推荐系统如果在其中某个环节起作用，那么就要衡量那个环节的转化率，这个相比前面三个指标，更加接近真实效果。

除了比例类的商业指标，还要关注绝对量的商业指标，常见的有：社交关系数量，用户停留时长，GMV（成交金额），关注绝对数量，除了因为它才是真正商业目标，还有一个原因，是要看推荐系统是否和别的系统之间存在零和博弈情况。

假如推荐系统导流效果提升，搜索引擎导流下降，从整个平台来看，因为整个平台的商业目标并没有那么成绩喜人，也需要警惕。

讲完深度类指标，下面进入广度类指标。

4. 覆盖率。这项指标就是看推荐系统在多少用户身上开采成功了，覆盖率又细分为 UV 覆盖率和 PV 覆盖率。UV 覆盖率计算方法是。

$$COV_{uv} = \frac{N_{l>c}}{N_{uv}}$$

解释一下，首先要定义有效推荐，就是推荐结果列表长度保证在 C 个之上，独立访问的用户去重就是 UV，有效推荐覆盖的独立去重用户数除以独立用户数就是 UV 覆盖率。PV 覆盖率计算方法类似，唯一区别就是计算时分子分母不去重。

$$COV_{pv} = \frac{N^*_{l>c}}{N^*_{pv}}$$

5. 失效率。失效率指标衡量推荐不出结果的情况。也分为 UV 失效率和 PV 失效率。UV 失效率计算方法是。

$$LOST_{uv} = \frac{N_{l=0}}{N_{uv}}$$

分子是推荐结果列表长度为 0 覆盖的独立用户数，分母依然是去重后的独立访问用户数。PV 失效率也一样，区别是不去重。

$$LOST_{pv} = \frac{N^*_{l=0}}{N^*_{pv}}$$

6. 新颖性。对于用户来说，“总是看到你这张老脸”会让他们审美疲劳，所以对用户来说，推荐的物品要有一定的新颖性。直观理解就是用户没见过。

所以新颖性需要讲粒度，物品粒度、标签粒度、主题粒度、分类粒度等等。每个粒度上评价用户没见过的物品比例。对于物品级别的新颖性，更多是靠直接过滤保证，这在前面章节已经专门讲到了对应的过滤算法。

7. 更新率。检测推荐结果更新程度。如果推荐列表每天几乎一样，显然不可取，尤其是新闻资讯类，要求每次刷新都不一样，对更新率要求更高。更新率可以有很多衡量方式，有一种是衡量每个推荐周期和上个周期相比，推荐列表中不同物品的比例。这个周期，可以是每次刷新，也可以是每天。

$$UPDATE = \frac{\Delta N_{diff}}{N_{last}}$$

2. 还能好多久?

除了关注系统表现有多好外，你还需要忧虑另一件事，你的系统还能好多久？也就是系统是否健康。

在推荐系统中，需要数据不断更新，这样系统才是一个活系统，用户兴趣客观上会变迁，数据源客观上也是会用光的一天，所以推荐系统如果不能应对这两个变化，就好不了太久。

衡量推荐系统是否健康的指标常用的有三个。

1. 个性化。虽然说到推荐系统时，言必称个性化，但实际上能做到真正个性化很难，那要求用户每个人都独立思考、爱好明确、不受群体影响。但是个性化程度的确能够反映推荐系统的健康程度，按照我们在专栏第一篇“是否需要推荐系统”中提出的那个公式来看：

$$\frac{\Delta connection}{\Delta user \times \Delta item}$$

如果没有个性化，那么分子上增加的连接数，其实是不受分母上增加的物品数影响的，因为所有人都只消费那少数几个物品，那么你其实不需要推荐系统。

个性化如何检测呢？有一个直观的方法，取一天的日志，计算用户推荐列表的平均相似度，如果用户量较大，对用户抽样。

2. 基尼系数。基尼系数衡量推荐系统的马太效应，反向衡量推荐的个性化程度。把物品按照累计推荐次数排序，排在位置 i 的物品，其推荐次数占总次数为 p_i。那么基尼系数为：

$$Gini = \frac{1}{n}\sum_{i=1}^{n} p_i * (2i - n - i)$$

看这个公式可以知道，如果推荐次数越不平均，那么基尼系数就越趋近于 1。

3. 多样性。多样性不但要在推荐服务吐出结果时需要做一定的保证，也要通过日志做监测。

多样性可能会损失一些效果指标，但是从长远上来看，对推荐系统所在平台是有利的。多样的推荐结果也会让产品显得生机勃勃，提升品牌形象。多样性衡量方式通常要依赖维度体系选择，例如常见的是在类别维度上衡量推荐结果的多样性。方法是下面这样的。

$$Div = \frac{\sum_{i=1}^{n} -p_i * log(p_i)}{n * log(n)}$$

多样性衡量实际上在衡量各个类别在推荐时的熵，一共有 n 个类别，分母是各个类别最均匀，都得到一样的推荐次数情况下对应的熵。

分子则是实际各个类别得到的推荐次数，p_i 是类别 i 被推荐次数占总推荐次数的比例，计算它的熵。两者求比值是为了在类别数增加和减少时，可以互相比较。

这种计算多样性是一个整体的评价，还可以具体评价每次推荐的多样性，每个用户的多样性，也就是 PV 多样性和 UV 多样性。

总结

推荐系统作为一种 AI 系统，其测试方法不完全相同于传统软件功能测试。对于推荐系统，也有一定的单元测试，扫描业务规则，对系统做一票否决制，因为这些业务规则定义明确。

除此之外，还要先经过离线模拟，再线上小范围实测，这部分测试就是在践行数据驱动。这部分指标主要在回答系统的两个问题。

1. 系统表现有多好?
2. 系统还能好多久?

只要系统现在表现好，并且系统生命力强，那么你的推荐系统就是好的推荐系统。这些指标就是在忠实反映这两个侧面的。

但是，光靠数据驱动，又容易走入歧途，还需要常常审视这些指标到底是否真实反映系统状态，所以还需要对用户做调查访谈，深入群众，听取最真实的感受，回来重新看看自己的指标是否合理，是否需要重新设计指标。

这里限于篇幅，没有完全列出推荐系统所用的指标，欢迎你留言，把你用过的指标分享出来，我们一起讨论。

感谢你的收听，我们下次再见。

上一篇 【关键模块】 推荐系统服务化、存储选型及API设计

下一篇 【效果保证】道高一尺魔高一丈：推荐系统的攻防

精选留言 (8)

写留言

惜心（伟祺...
2018-05-15
3

吉尼斯系数中 n是只推荐的次数吗 （2i-n-i）是什么含义

展开

小贝
2018-05-14
2

最好举举个例子吧，刑无刀老师讲的面覆盖广，但是偏泛泛，能不能举几个例子说明下？

shangqiu86
2019-05-09

惭愧惭愧，我的推荐系统的评价指标只有深度，基本上没有广度的，而且深度的也不全面，学习了，有具体的例子就好了，详细讲解实际中的操作就更好啦！自己摸索摸索吧

展开

走小調的凡...
2018-10-10

请问老师举的例子 TopK 准确率和 TopK 召回率 具体是怎么算的?

展开

张同学
2018-05-27

刑老师所讲的覆盖率这里，和以前我们看书及实际业务中的应用有些不一样，之前我们对覆盖率的定义是，推荐系统对长尾物品的发掘能力，即推荐出来的物品占总物品的比例。不知道这个差异是具体的产品业务的不同所导致的么？还是说这根本是两个不同的指标呢?

slvher
2018-05-27

总结的不错，但原文未对基尼系数（gini index）公式中出现的符号n做说明，且公式似乎有2个小问题，请确认：
1）p_i应该是按推荐次数统计排序后，第i个item的次数占总次数的比例吧？文中的解释是次数而非比例
2）求和项括号里是不是应该为(2i - n - 1)？...

展开

yzz
2018-05-16

同问：吉尼斯系数中 n是只推荐的次数吗 （2i-n-i）是什么含义

展开

hqzhao
2018-05-14

我觉得刑老师总结的挺全面，还讲到了很多实际应用时要考虑的指标，这是我在做科研时自己不曾考虑的