029 | 如何评测搜索系统的在线表现?

2017-12-08 洪亮劼

讲述：初明明
时长 07:54 大小 3.62M

我在本周前面的两篇文章中为你讲解了基于“二元相关”和基于“多程度相关”原理的线下评测指标。利用这些指标，研发人员在半个世纪的时间里开发了一代又一代的搜索系统，这些指标和系统也都在不断演化。

虽然我们这周讲过的这些指标都很有指导意义，但大多数指标被提出来的时候都是基于线下的静态数据集，并不是真正去检测用户和系统的互动（虽然后期也有研发人员直接使用这些评测工具用于在线评测，但在使用上就产生了问题）。那有什么样的方法来评测搜索系统的在线表现呢?

为了回答这个问题，我们今天就来探讨一下进行在线评测的几个话题。

在线可控实验

我们先回到整个评测指标的初衷，为什么要进行线下测试呢？

第一个原因是在信息检索系统（也就是最早的搜索系统）的开发时期，还很难做在线可控实验（Controlled Experiments），研发人员还没有开发出值得依赖的手段来判断用户的行为。因此，在那个年代，比较可靠的方法就是调查问卷和后来开发出来的线下静态评测。可以说，这些手段都是真正了解用户行为的一个“代理”（Proxy）。

要进行评测，不管是线下还是线上，另外一个原因就是我们需要某种手段来分辨两个系统的好坏，从而能够不断地通过这种手段来改进系统，做到数据驱动。

那么，能够正确观测两个系统不同的工具，就是“在线可控实验”，有时候又称作“在线实验”，或者也叫作“在线 A/B 实验”。

在线可控实验其实是建立“因果联系”（Causal Relationship）的重要工具，也可以说是唯一完全可靠的工具。这里面的基础是统计的假设检验。

具体来说，就是我们针对访问网站或者应用的人群，进行某种划分，一般情况下是平均随机划分，百分之五十的用户进入划定的一个群组，叫作“控制组”（Control Bucket），而另外百分之五十的用户进入另外一个群组，叫作“对照组”（Treatment Bucket）。“控制组”和“对照组”的唯一区别在于所面对的系统。

假设有一个搜索系统，我们想对其中的某个部分进行改进，那么，我们可以保持住其他的部分，让这个希望得到改进的部分成为唯一的“独立变量”（Independent Variable），也就是在整个实验设置中的变量。这样，我们就希望看到，能否通过在线实验以及假设检验的工具，来认定这个“独立变量”是否会带来系统性能上的提高，亦或是降低。

这里面还有一个需要提前确定的，那就是需要评测的指标，特别是用户指标，比如网站的点击率、搜索的数量等等。这些指标我们称之为“依赖变量”（Dependent Variable）。说白了，**我们就是希望能够在“独立变量”和“依赖变量”之间通过假设检验建立联系。**

虽然在概念上很容易理解在线可控实验，但在实际操作中会面临很多挑战。

虽然在理想状态下，我们可以把用户五五对分，让用户分别进入“控制组”和“对照组”。然而现实中，经过随机算法分流的用户群在这两个群组中很可能并不呈现完全一样的状态。什么意思呢？

举个例子，比如，在“控制组”中，相比于“对照组”而言，可能存在更多的女性用户；或者是在“对照组”中，可能存在更多来自北京的用户。在这样的情况下，“依赖变量”，比如网站点击率，在“控制组”和“对照组”的差别，就很难完全解释为“独立变量”之间的差别。

也就是说，如果“控制组”下的点击率比“对照组”高，是因为我们更改了系统的某部分产生了差别呢，还是因为这多出来的女性用户呢，又或者是因为女性用户和系统的某些部分的交互，产生了一定复杂的综合结果导致的呢？这就比较难说清楚了。对于刚才说的有更多来自北京的用户这个例子也是一样的。

当然，在现实中，如果说我们依然可以比较容易地通过算法来控制一两个额外的变量，使得在“控制组”和“对照组”里面这些变量的分布相当，那么，面对十几种（例如，年龄、性别、地域、收入层级等等）重要变量，要想完全做到两边的分布相当，难度很大。

即便我们能够做到通过随机算法使得已知变量在两个群组中的分布相当，我们依然不能对当前还未知的变量进行如此操作。因此，**如何处理因人群特性所带来的对结论的影响是现实中在线实验的难点之一。**

在线实验的难点之二是，我们有可能很难做到如设想中的那样，让系统的某个部分成为“控制组”和“对照组”中唯一的“独立变量”，即便是除去了刚才所提到的人群差异。

在现代网站或者应用中，有很多服务、子系统、页面、模块同时在为整个网站服务。而这些服务、子系统、页面和模块，都有不同的前端系统和后端系统，很可能属于不同的产品和工程团队。每个部分都希望能够做自己的可控实验，希望自己改进的部分是唯一变化的“独立变量”。然而，我们从宏观的角度去看，如果每个部分都在做自己的实验，而我们做实验的基本单元依旧是每个用户的话，那这就很难保证用户之间的可比性。

举个例子，如果用户 U1，进入了首页的“控制组”，然后访问了搜索页面的“对照组”继而离开了网站。而用户 U2，直接访问了帮助页面的“对照组”，然后访问了搜索页面的“控制组”。那 U1 和 U2 两个用户最终产生的点击率的差别，就很难从他们访问网站页面的过程中得出结论。即便是在有大量数据的情况下，我们也很难真正去平衡用户在所有这些页面的组别之间的关系。

实际上，如何能够有效地进行在线实验，包括实验设计、实验评测等，都是非常前沿的研究课题。每年在 KDD、WSDM、ICML 等学术会议上都有不少新的研究成果出炉。

利用因果推论对实验结果进行分析

今天的最后我想提一下因果推论（Causal Inference）。因果推论不是普通的统计教科书内容，也不在一般工程类学生接触到的统计内容之内。然而这个领域在最近几年受到了机器学习界越来越多的关注，因此了解因果推论对于学习机器学习的前沿知识来说很有必要。

像我们刚才提到的种种实验中产生的用户特征不平均、实验之间可能存在关系等，在这些方面我们都可以利用很多因果推论的工具进行分析。另外，对于工程产品而言，并不是所有的情况都能够通过 A/B 测试来对一个希望测试的内容、模型或者产品设计在一定时间内找到合理的结果，有很多情况下是不能进行测试的。因此，在不能进行测试的情况下还能通过数据研究得出期望的结果，也就是说，我们能否模拟在线实验，这就是因果推论的核心价值。

一般而言，在机器学习中需要因果推论的场景也很常见。比如，我们需要用数据来训练新的模型或者算法，这里面的数据采集自目前线上的系统。然而，现在的线上系统是有一定偏差的，那么，这个偏差就会被记录到数据里。在此，因果推论就为机器学习带来了一系列工具，使得在一个有偏差的数据中依然能够无偏差地进行训练以及评测模型和算法。

小结

今天我为你讲了在现代搜索技术中，如何利用在线实验，特别是可控实验来评价我们构建的系统。

一起来回顾下要点：第一，详细介绍了在线实验的一些因素，并分析了在线实验中可能产生的用户不平衡以及实验有相互作用的问题。第二，简短地提及了现在利用因果推论来进行在线实验数据分析以及“偏差”调整的一个思路。

最后，给你留一个思考题，如何建立在线实验评测结果和线下指标比如 nDCG 之间的关系呢？

欢迎你给我留言，和我一起讨论。

上一篇　028 | 搜索系统评测，有哪些高级指标?

下一篇　030 | 文档理解第一步：文档分类

精选留言 (1)

写留言

韩康
2018-04-04

请问，在使用LTR模型的系统里做A/B实验时，团队里怎么多人同时做A/B实验?
团队里多个人新增特征重新训练了LTR模型，同时做A/B实验，如果相比于线上基线模型效果都很好，这些特征要全量的话，还要把所有的新增特征重新训练一个新模型，再做A/B实验。
如果团队很多人同时做LTR优化，如何验证不同的特征并全量?

展开