063 | 简单推荐模型之一：基于流行度的推荐模型

2018-02-26 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 08:49 大小 4.04M

今天，我们正式进入专栏的另一个比较大的模块，那就是**推荐系统**。之前我们详细且全面地介绍了搜索系统的各个组成部分。在接下来的几周时间里，我们一起来看推荐系统的技术要点又有哪些。

我们还是从简单推荐系统聊起，由易到难，逐步为你讲述一些经典的推荐模型。

推荐系统目前已经深入到了互联网的各类产品中。不管是到电子商务网站购物，还是到新闻阅读网站获取信息，甚至是在出行的时候希望听到不同的音乐，不同种类的推荐系统都在我们的生活中发挥着举足轻重的作用。

那么，搭建一个最简单的推荐系统，应该如何入手呢？今天我们就来聊一个最基本的推荐模型：**基于流行度的推荐模型。**

最简单的流行度估计

什么是基于流行度（Popularity-based）？通俗地说，就是什么内容吸引用户，就给用户推荐什么内容。

这里面其实有一个隐含的假设，那就是物品本身的质量好坏和流行度有一定的正比关系。什么意思呢？就是说好的东西，关注的人自然就多，自然就会有更多的谈论。当然，这是一个主观的假设，并不是所有质量高的物品都会有很高的流行度。然而，在不需要过多其他信息和假设的情况下，流行度可以算是衡量物品质量好坏的一个最简单的测度。

那么，如果我们能够在每一个时间点上准确地估计到一个物品的流行度，就只需要按照流行度的数值从高到低排序显示所有的物品就可以了。

然而，这里牵涉到一个问题，那就是如何判断一个物品在任何时间点上的流行度呢？有两个重要的因素影响着物品流行度的估计，那就是**时间和位置**。

我们先来说一下时间因素。很显然，用户访问每一个应用或者服务都有一定的规律，这种规律导致每一个应用的**流量规律**也不一样。比如，人们可能更倾向于在早上或者傍晚打开新闻网站，看一看一天都发生了什么事情。因此，任何文章投放到这两个时段自然就会有比较高的关注度。这并不代表这些文章就要好于其他的文章，可能仅仅是由于时间的关系。因此，我们在对流行度建模的时候就需要考虑时间的因素。

另外一个重要的因素是位置。这个“位置”并不是真正的地理位置，而是在一个服务或者网站的什么位置显示你的物品。因为用户心理对于不同位置的感受，在很多类型的服务中常常都有隐含的“**位置偏差**”（Position Bias）。

这些偏差给我们估计某个物品的流行度带来了很大的困难。比如说，在绝大多数的搜索引擎服务中，排名第一的物品所受到的关注度很可能大大高于排名第二和之后的物品。因此，一个物品只要放到第一的位置，关注度自然就会升高。当然，这并不能完全代表这个物品本身的属性。

因此，我们在估计物品的流行度时就需要考虑上面所说的这两个重要因素。

要解决刚才说的两个问题，我们就**不能使用绝对数值来对流行度建模**。比如我们使用在单位时间内点击的数目，购买的数目，或者点赞的数目，都会受到刚才所说的两种偏差的影响。假设一篇文章在 9 点到 10 点这个时段被点击了 100 次，在 10 点到 11 点这个时段被点击

了 50 次，这并不能代表这个文章在 10 点到 11 点这个时段就变得不受欢迎了，很可能是这个时段的总的用户量比较多。

因此，**对于流行度的衡量，我们往往使用的是一个“比值”（Ratio），或者是计算某种“可能性”（Probability）**。也就是说，我们计算在总的用户数是 N 的情况下，点击了某个文章的人数。这个比值，取决于不同的含义，如果是点击，往往叫作点击率；如果是购买，叫作购买率。为了方便讨论，我们在下面的例子中都使用点击率。

然而，点击率本身虽然解决了一部分时间和位置偏差所带来的影响，但是点击率的估计所需要的数据依然会受到偏差的影响。因此，我们往往希望能够建立无偏差的数据。

关于如何能够无偏差地估计，这是一个研究课题，我们今天不详细展开。不过，有一种比较经济的方法可以收集没有偏差的数据，那就是把服务的流量分成两个部分。

一个部分，利用现在已有的对物品流行度的估计来显示推荐结果。另外一个部分，则随机显示物品。这种方法是一种特殊的**EE 算法**（Exploitation & Exploration），叫“**epsilon 贪心**”（epsilon-Greedy）。

我们之后还会聊到这个话题。根据这样的方式搜集的数据可以认为是没有位置偏差的。我们从随机显示物品的这部分流量中去估计流行度，然后在另外一个部分的流量里去显示物品。

如果从数学上对点击率建模，其实可以把一个物品在显示之后是否被点击看成是一个“**伯努利随机变量**”，于是对点击率的估计，就变成了对一个伯努利分布参数估计的过程。

有一种参数估计的方法叫作“**最大似然估计法**”（Maximum Likelihood Estimation）。简而言之，就是说，希望找到参数的取值可以最大限度地解释当前的数据。我们利用最大似然法就可以求出在某一段时间内的点击率所代表的伯努利分布的参数估计。这个估计的数值就是某个物品当前的点击总数除以被显示的次数。通俗地讲，如果我们显示某个物品 10 次，被点击了 5 次，那么在最大似然估计的情况下，点击率的估计值就是 0.5。

很显然，这样的估计有一定的局限性。如果我们并没有显示当前的物品，那么，最大似然估计的分母就是 0；如果当前的物品没有被点击过，那么分子就是 0。在这两种情况下，最大似然估计都无法真正体现出物品的流行度。

高级流行度估计

我们从统计学的角度来讲了讲，如何利用最大似然估计法来对一个伯努利分布所代表的点击率的参数进行估计。

这里面的第一个问题就是刚才我们提到的分子或者分母为 0 的情况。显然，这种情况下并不能很好地反应这些物品的真实属性。

一种解决方案是对分子和分母设置“先验信息”。就是说，虽然我们现在没有显示这个物品或者这个物品没有被点击，但是，我们“主观”地认为，比如说在显示 100 次的情况下，会有 60 次的点击。注意，这些显示次数和点击次数都还没有发生。在这样的先验概率的影响下，点击率的估计，或者说得更加精确一些，点击率的后验概率分布的均值，就成为了实际的点击加上先验的点击，除以实际的显示次数加上先验的显示次数。你可以看到，在有先验分布的情况下，这个比值永远不可能为 0。当然，这也就避免了我们之前所说的用最大似然估计所带来的问题。

利用先验信息来“平滑”（Smooth）概率的估计，是贝叶斯统计（Bayesian Statistics）中经常使用的方法。如果用更加精准的数学语言来表述这个过程，我们其实是为这个伯努利分布加上了一个 Beta 分布的先验概率，并且推导出了后验概率也是一个 Beta 分布。这个 Beta 分布参数的均值，就是我们刚才所说的均值。

在实际操作中，并不是所有的分布都能够找到这样方便的先验分布，使得后验概率有一个解析解的形式。我们在这里就不展开讨论了。

另外一个可以扩展的地方就是，到目前为止，我们对于流行度的估计都是针对某一个特定的时段。很明显，每个时段的估计和前面的时间是有一定关联的。这也就提醒我们是不是可以用之前的点击信息，来更加准确地估计现在这个时段的点击率。

答案是可以的。当然，这里会有不同的方法。

一种最简单的方法还是利用我们刚才所说的先验概率的思想。那就是，当前 T 时刻的点击和显示的先验数值是 T-1 时刻的某种变换。什么意思呢？比如早上 9 点到 10 点，某个物品有 40 次点击，100 次显示。那么 10 点到 11 点，我们在还没有显示的情况下，就可以认为这个物品会有 20 次点击，50 次显示。注意，我们把 9 点到 10 点的真实数据乘以 0.5 用于 10 点到 11 点的先验数据，这种做法是一种主观的做法。而且是否乘以 0.5 还是其他数值需要取决于测试。但是这种思想，有时候叫作“**时间折扣**”（Temporal Discount），是一种非常普遍的时序信息处理的手法。

小结

今天我为你讲了基于流行度的推荐系统的基本原理。一起来回顾下要点：第一，我们简要介绍了为什么需要基于流行度进行推荐；第二，我们详细介绍了如何对流行度进行估计以及从统计角度看其含义；第三，我们简要地提及了一些更加高级的流行度估计的方法。

最后，给你留一个思考题，我们介绍了如何使用先验信息来对参数进行平滑，如何能够更加准确地确定先验概率中的数字呢？具体到我们的例子就是，如何来设置先验的点击和显示次数呢？

欢迎你给我留言，和我一起讨论。

上一篇　复盘 1 | 搜索核心技术模块

下一篇　064 | 简单推荐模型之二：基于相似信息的推荐模型

精选留言 (5)

写留言

rushui
2018-04-01

1

很可能是这个时段总的用户量比较多，是不是应该比较少？

展开

兔子ORZ
2018-04-10

超参的设置，在冷启动下，会考虑历史水平降权，但是对刚开始的更新很敏感。也是类似一个EE的取舍。另外如果资源允许，分批测试我觉得应该会更好。

展开

离忧
2018-03-02

比如求今天某商品的点击率和曝光率，如果当前没有显示率和点击率，就用之前的显示率除以现在的曝光率，在乘以时间折扣。，如果现在有点击率和曝光率，用贝叶斯平滑（用历史数据做的先验概率，分子加上点击率，分母加上曝光率）。

林彦
2018-02-26

或者是同一位置的次数之和

展开

林彦
2018-02-26

简单地猜想可以用整体的，或同一类别的，同一时间段的总显示次数和点击次数来作为先验概率中的数值。