080 | 现代推荐架构剖析之三：复杂现代推荐架构漫谈

2018-04-06 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 05:40 大小 2.60M

本周我们讨论现代推荐系统的架构体系。周一我们看了最简单的基于线下离线计算的推荐架构，周三我们聊了基于多层搜索架构的推荐系统。

今天，我们来谈一谈如何从这两种架构的思路出发，来满足更加复杂多变的实际情况。

推荐架构需要解决的问题

这周我反复强调推荐系统的几个基本需求点。第一，能够在一两百毫秒内给用户提供当前的推荐结果；第二，需要对用户和系统的交互结果做出响应；第三，需要考虑用户群体的覆盖率问题。

接下来我们就聊一些经常考虑的场景，起到一个抛砖引玉的作用，供你参考。

新用户的问题

如果你要搭建的系统面临的情况是新用户多，比如一个新上线的快速增长的产品，那么我们需要怎么考虑架构呢？

这里面有**两个基本思路**。第一，我们要更加快速地抓住这些新用户和系统的交互信息，从而更好地为他们推荐信息。第二，在我们还没有足够多的信息的时候，如何为这些用户提供推荐结果。

我们先从第一点说起，如果希望能够更加快速地抓住用户的交互信息，从而很好地为他们推荐内容，有两种做法：要么能够快速更新模型从而更新推荐结果，要么快速更新特性从而更新推荐结果。

如果我们整个产品只有一个全局的排序模型的话，不管是基于线下的静态架构还是基于搜索的架构，基本上都不可能很快地去更新这个全局模型。因此，在这种情况下就需要去思考如何更新特性。

对于搜索的框架，也许我们可以通过更新特性，从而达到在重排序的这个阶段，因特性改变而带来不同的结果。但是对于线下的静态架构，因为所有推荐结果都是事先处理好的，因此改变特性也不能改变结果，除非针对这个用户，对所有的推荐结果重新进行线下计算。这样做是可行的，但是计算成本还是相对比较高。因此，综合来看，如果在新用户比较多的情况下，并且我们还希望抓住用户的交互，静态架构可能就会显得有一些心有余而力不足了。

第二点则是新用户的交互信息一开始会比较少，如何处理冷启动呢？我们前面提到过，其实冷启动可以利用一些用户的其他信息，比如年龄、性别、地理信息来产生推荐结果。我们可以为用户显示当前比较流行的在某个年龄段、某个性别、某个地理区域的信息。

一个简单的思路是，这些年龄、性别、地域的信息，可以每个小时或者每天更新一次，单独存放在一个数据库里。当用户来网站的时候，我们可以尝试从搜索的架构里提取信息，也从这个单独的数据库里提取信息，然后在这个基础上进行全部重新排序。这样我们就能够保证架构的统一性，同时也解决了冷启动的问题。

新物品的问题

和大量新用户问题不同的是，大量新物品的问题则更加棘手一些。

在静态框架下，新物品意味着对于所有的用户，我们之前都没有考虑过这些物品，因此如果不进行特殊处理，我们是绝对没法把这些物品展示给用户看的。

这里有两种思路。**一种思路**，就是把新物品加入到内容池里，对于所有用户，全部重新生成推荐结果。这当然是最简单的想法，但是很显然，这样做是非常耗时的。**另外一种思路**，我们把当天产生的新物品单独存储在一个数据库，针对这些物品给出一些预估计的分数。这里当然可以针对物品的特性打分，也可以随机给定一些分数。然后我们在显示推荐的时候，可以混合之前线下已经产生的推荐结果和当天的新物品结果，这样从用户的角度来看，我们是可以对新物品进行推荐的。

在搜索的架构下，也有**两个类似的思路**。第一，那就是我们对索引进行重索引，但这个过程相对比较耗时。第二，那就是对新物品构建一个临时索引或者数据库，最后的结果是从索引和当天的临时存储中共同获取，然后进行重新排序。

在新物品比较多，并且很快就会过时的情况下，另外一个需要注意的棘手问题就是，推荐的模型一定不能仅仅抓住用户喜爱的某一个物品。比如新闻推荐，用户喜欢某一个新闻，但是这个新闻很快就会过时。这就和商品推荐有很大不同，对于商品来说，用户可以反复购买同一件日用品。

小结

今天我为你讲了利用推荐系统的一个重要问题，就是如何构架一个现代推荐系统。我们聊了两个场景下的一些更加细致的取舍，分别是新用户多的情况和新物品多的情况。

其实，所有的这些思路都不是“死规矩”，但是有一些基本的规则你可以去琢磨。

比如，我们尽可能把复杂的运算放在线下，因为毕竟需要在规定的时间内返回结果。在一切有可能的情况下，尽可能使用搜索引擎来减少需要对大量物品进行打分的步骤。再比如，对于活跃的用户，我们可以使用多层搜索架构；但是对于不活跃用户，我们可以依赖线下，提前产生所有的推荐结果。

一起来回顾下要点：第一，我们再次回顾了推荐架构的需求；第二，我们通过两个场景，新用户多和新物品多，分析了架构里面的一些取舍。

最后，给你留一个思考题，假设我们的推荐系统需要给一个手机客户端的产品进行推荐，有什么和桌面端不一样的，需要在架构上额外注意的地方呢？

欢迎你给我留言，和我一起讨论。

上一篇　079 | 现代推荐架构剖析之二：基于多层搜索架构的推荐系统

下一篇　081 | 基于深度学习的推荐模型之一：受限波兹曼机

精选留言 (4)

写留言

林彦　1
2018-04-06

网络性能，手机的性能，手机的交互空间和方式这些因素导致手机上的推荐结果需要更少，更精简，减少实时的计算量，信息的元素要选择更重要和更容易引起用户下一步简单交互操作的，也可以通过优化的并行框架来降低较多计算和信息处理的响应时间。用户的手机交互行为，我觉得有条件和桌面端区别开来作为运算的一种特性更好。

展开

微微一笑　1
2018-04-06

感觉讲的有点简单啊 能否一个实际案例结合，深入讨论下

展开

和平老三

2018-06-19

您能不能在留言区补充一些文章 这篇文章有些过于理论了

展开

Mr.Button

2018-05-21

洪老师，之前看您的博客。
感觉这个架构讲的有点简单。能不能提供一些新户冷启动算法上的建议。比如 雅虎之前的决策树方法之类的。

展开