推荐策略产品经理必读系列—第二讲推荐系统的架构

产品名称	推荐策略产品经理必读系列—第二讲推荐系统的架构
公司名称	东莞市数云网络科技有限公司
价格	.00/个
规格参数	
公司地址	东莞市
联系电话	18665158422 18665158422

产品详情

编辑导语：推荐系统是如何做到向用户推荐他感兴趣的物料的，这个取决于我们的推荐系统架构。作者对推荐系统的架构进行了详细的阐析，总结了7大搭建架构环节，希望对你有所帮助。

上一篇为大家介绍了作为一个推荐策略产品必须了解的行业里发生的大事以及行业整体未来的趋势，本篇就为大家详细介绍推荐系统的架构，推荐系统是如何把我们感兴趣的物料推荐到我们面前的。

一、整体架构

将推荐系统架构里面主要的部分整体如上图，用户打开APP看到为自己推荐的物料，正常都是需要经过5-6个环节。常见的流程从建立物料索引，再到召回、粗排、精排、重排到过滤层，最终展示在了APP前段，为用户推荐了感兴趣的物料。

二、物料索引

APP里有哪些物料是可以推荐给用户，会有一个总的物料池，物料池本身存储在数据库中。同时为了查询时更加方便快捷，我们需要提前构建好物料索引。尤其是很多召回路是基于一些标签进行召回的，更加需要提前构建好倒排索引。物料的清洗和索引的构建是推荐系统的第一步基础工作。

三、召回层

构建完物料索引后，推荐系统是如何挑选出用户感兴趣的物料了。推荐算法发展到现在，我们可以称之为4.0时代。4.0时代的推荐算法都是以预估用户的点击率（CTR）为核心，基于用户对于每个物料的预估

点击率（Predicted-CTR）来进行排序，按照Predicted-CTR值进行倒排。以上介绍的是最理想的方式，但是实际却无法执行也没必要。

原因一：无法实现

物料库的量级太大，比如像淘宝这种平台物料几十亿商品，如果用户的每次请求都去预估几十亿商品的Predicted-CTR，需要大量的机器资源，同时时延会非常高。大家打开淘宝首页可能需要几十个小时，且不一定能加载出来。

原因二：没有必要

几十亿的商品，有很多商品是非常冷门商品，99%以上的商品和用户毫无关联，完全没有必要如此精细化去预估用户对于几十亿商品的兴趣度。

所以推荐系统演变成先通过召回层初步筛选出用户可能感兴趣的一些物料，比如800个。然后再进入粗排和精排，其他几十亿的物料直接在召回阶段就被过滤了。这样的架构设计大大降低了推荐系统的计算压

力，同时也不影响推荐系统的整体效果。

而召回阶段，市面上最先进的模式就是针对不同的用户进行分层，然后不同层级的用户使用的召回路数不一样，核心都是多路召回，每一路召回返回的物料个数以及对应的权重都是和用户本身的分层有很大关系，实现非常精细化的用户和召回路数的管理。

常见的召回方法有基于热销商品的召回、基于历史高点击商品的召回、用户历史看过商品的召回，还有一些常见的协同过滤算法比如Item-CF、User-CF。关于召回阶段常用的策略和算法后面会有专门的文章进行讲解。

四、粗排层

大家可以看到召回阶段会有非常多的召回策略，那我们如何将这些各路召回的商品汇总在一起进行一个统一排序了。比如热销路的召回，召回了Top 100的商品；历史高点击的召回，同样召回Top 100的商品。这两路召回的商品可能还会有重叠的。

在召回层里，每一路的召回都需要将物料的分数进行归一化。比如说热销路的召回，如何召回Top
100的商品，首先需要计算每一个商品的热销分数，然后取Top 100的商品。历史高点击的召回路也是一样，并且每一路的分数都需要归到【0，1】之间，这样各个路之间才能比较。

同时不同场景下每一路召回的重要性也完全不一样，可以再设置一个对应的系数。比如热销路召回为0.5，历史高点击路召回为0.7，假设商品A在热销路召回分数为0.8，历史高点击路召回分数为0.3，那么最终这个商品的总召回分数为：0.5 X 0.8 + 0.7 X 0.3 = 0.61。粗排层就需要将每一个召回的商品进行上述计算方式进行处理后，汇总得到一个总的List，然后选择Top K的商品给到精排层。

五、精排层

精排层的核心任务就是预估用户对于召回层返回的Top K商品的Predicted-CTR。召回和粗排只是选择出了用户可能感兴趣的物料，但是每一个物料具体的预估CTR是多少并不知道。

精排层就需要基于用户历史点击过的物料作为正样本，曝光未点击的物料作为负样本，然后构建CTR预估模型，预估用户对于每一个物料的Predicted-CTR。在精排层核心要做的几件事情：数据清洗构建正负样本，选择合适的排序算法，构建特征工程、模型训练与效果评估。

最终基于精排模型预估出的CTR对于召回的物料再重新进行一次排序。精排模型的预估是整个推荐系统中耗时最多的，因为特征十分复杂，特征维度很多。如果针对几十亿物料全部进行CTR预估，系统直接崩溃，这也是需要先进行召回的原因。

六、重排层

那是不是精排过的物料，直接按照精排后的顺序直接展示在了APP前端了。很多时候推荐系统仍然有一些其他业务规则进行干预。比如在电商推荐系统里面，就会有以下的一些策略：

1. 类目打散

对于给用户推荐的商品如果类目集中度过高，会进行一定程度的打散。比如精排模型给用户返回的前10个商品全部都是鞋子，可能该用户偏好鞋子，但前10个商品全部是鞋子此种集中度还是过高了，重排层就会将后面其他类目的商品插入到这10个商品中。具体按照类目打散的规则每家不一样，核心是基于业务场景。打散不是目的，目的是为了提升推荐系统的点击率。

2. 不同类型物料混合

比如淘宝，淘宝推荐场景里面有的内容类型有：店铺、活动、直播、商品、视频等等。那这些不同类型内容之间如何进行混合。能否可以出现连续4个全部都是直播，或者4个全部都是视频。为了降低用户的审美疲劳，很多时候针对不同类型的内容推荐系统也会进行重新打散。但同样打散不是目的，目的是为了提升推荐系统的点击率。

3. 全局最优

重排层还有一个核心的逻辑就是实现全局最优。精排层是预估用户对于单个物料的CTR，这是一种局部最优的思想。但是用户在浏览时正

常都会一次性浏览多个，怎么样的物料组合可以实现全局最优而不是局部最优。同样4个坑位，有可能精排排序在1，3，5，7的四个物料组合比精排排序在1，2，3，4的四个物料组合整体CTR更高。

总的来说重排层是推荐系统最后一道策略和模型的调整了。

七、过滤层

重排层调整完的物料顺序还会再进行一些业务规则和策略的干预，比如电商领域会进行以下的过滤：

1. 未上架过滤

当前已经上架的商品不展示在APP前端。

2. 缺货过滤

当前已经缺货的商品不展示在APP前端。

包括还有同图过滤等等策略。过滤层很多时候我们会做在了粗排和精排之间，确保进入到精排的物料后续都是能够直接在APP前端展示的，这样后续的精排和重排层的价值才更高。原本重排挑选出的最优组合，结果全部在过滤层被过滤了，那么整体推荐系统的效果就会大打折扣。

八、APP前端

经过过滤层的物料顺序是不会再发生变化，但还是有最后一步工作要做，而很多推荐系统的文章都没有介绍。就是内容样式和创意。比如电商平台里面同样都是店铺的内容，到底应该展示哪一种样式。大家打开淘宝首页经常会觉得花里胡哨，就是因为内容的样式太多了。

APP前端选择最合适的一种内容样式进行展示，具体关于内容样式和创意的选择后续也会有专门文章进行介绍。

经过7个大的环节，推荐系统也就在APP推荐场景为用户推荐了他可能感兴趣的物料。以上就是关于推荐系统架构的一个完整介绍。下一期

为大家详细介绍推荐系统的召回策略，欢迎大家持续关注。

本文由 @King James