推荐系统-召回、粗排、精排、重排

# 1.推荐系统介绍

从大体上来说，推荐系统需要包含这几种模块：索引/库存池，特征服务，排序模块，线上展示逻辑，日志系统，分析系统。如下图所示：

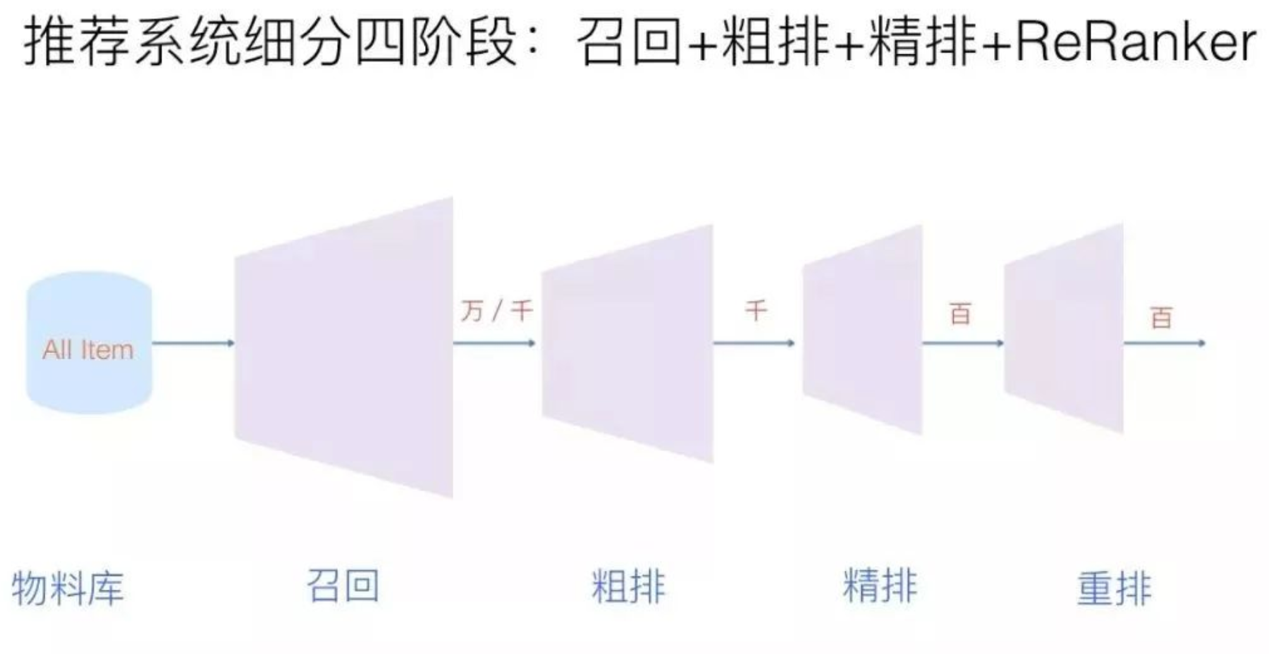
图示

描述已自动生成

* 索引池是对当前所有item的判定，并不是所有item都可以出现在推荐这整个大的逻辑下面。举个例子，广告主的某个计划，只设定了相应的预算，如果预算花完了，或者广告主已经不想投了，那就需要从索引池里面拿掉。另一种情况是可能有多种索引池，广告主不想投放20-30的人群的时候，索引池就等于是其他年龄段的索引池合并起来。
* 特征服务，用户发生请求（刷新，刚进入app等都会有请求发生）时，算出该用户信息所对应的特征，比如通过他的一些行为判定他的年龄性别等。另一个很重要的则是获取这个用户的一些历史行为。对于item当然也要提取特征，但是图上没有画出来，原因是item的大多数特征都相对固定，而用户的行为特征变化很快，需要专门的服务来处理
* 排序模块，这就是模型主导的部分了，下面要讲的召回精排粗排都在这里，下一讲要涉及的打压保送策略也包含在内。**排序模块的作用就是从很多候选的item中挑出最好的一个或者多个进入到展示逻辑中**
* 展示逻辑，这里有的同学可能会问，知道哪个最好不就直接放出去就好了吗？为啥还要有一个逻辑的区分？其实这里一个主要的点指的是广告和内容的混排，或者是视频和文章的混排。广告和文章，在各自的排序阶段是谁也看不见谁的，也就是说，上面的排序系统是双倍的。当双方都排好之后，需要对广告进行判定，要么是质量很高（预估的ctr，cvr都很高），要么是此刻之前展示的广告比较少，门槛下降了。如果符合这两种情况就可以对广告选择合适的位置投放。
* 日志系统：记录推送前后系统发生的一切事情。注意在图上有一个日志系统返回排序模块的箭头，这个箭头的含义是，用户的行为要落盘，形成新的训练数据来让排序模块继续训练
* 分析系统：这个系统依赖于日志，**此处主要指AB测试系统**。AB测试就是指，把用户随机进行划分，一部分用户应用对照组（A组，也就是原来的系统），另一部分用户应用实验组（B组，也就是我们想添加的改进点）。通过对比AB之间的差异，来展示我们所加的改进点是否有效。所以整个系统的迭代都是严重依赖AB实验的，想一个idea-做线下实验-上AB试试-有效就推广到全量，一般是这么个流程。当然AB实验也不是万能的，这个后面会讲到。另外要注意的是，虽然这里叫做AB测试，但是实际上对照组的用户不是都放在同一个桶里面，实践中往往是AABB实验，即对照组也有多组，实验组也有多组。这么做是为了观察组内的方差和组间的方差，假如两个对照组之间的观看时长差距有3%，对照组和实验组差距只有2%，那我们就无法说明这个实验是有正向的。

# 2.排序模块

整体上，排序是一个漏斗，如下图所示（这里少了一个重排）：

图形用户界面, 应用程序, 电子邮件

描述已自动生成

四个环节分别是：**召回、粗排、精排和重排**。召回目的如上所述；有时候因为每个用户召回环节返回的物品数量还是太多，怕排序环节速度跟不上，所以可以在召回和精排之间加入一个粗排环节，通过少量用户和物品特征，简单模型，来对召回的结果进行个粗略的排序，在保证一定精准的前提下，进一步减少往后传送的物品数量，粗排往往是可选的，可用可不同，跟场景有关。之后，是精排环节，使用你能想到的任何特征，可以上你能承受速度极限的复杂模型，尽量精准地对物品进行个性化排序。排序完成后，传给重排环节，传统地看，这里往往会上各种技术及业务策略，比如去已读、去重、打散、多样性保证、固定类型物品插入等等，主要是技术产品策略主导或者为了改进用户体验的。

简要介绍：一开始我们可能有成千上万的item，首先要由召回（也叫触发，recall）来挖掘出原则上任何用户有可能感兴趣的东西。这个环节是入口。有时候，单独的召回可能难以做到照顾所有方面，这个时候就需要多路召回。很好懂，就是多个召回路共同决定进入粗排的候选。比如可以由一个召回专门根据用户过往的兴趣来筛选候选，那可以再加一个召回专门输出近一段时间的热门视频。他们的出发点都不一样，共同组成下一级的输入。粗排（pre-rank）一般接受几千个输入，这时候压力就大大降低了。粗排就可以放一些较为复杂的结构，比如mlp就可以放上去了。粗排的输出一般是小于1000的，那这个压力就更小了，精排（rank）就可以变得很复杂，什么transformer啊放进去也不过分。

## 2.1召回

由于召回所要面对的item量是最大的，因此召回也是时延压力最大的。简单来说，就是要快。意味着他的模型结构最简单，甚至有时候都不是模型，而是规则。

数量级：几亿item->几千

时机：用户发起请求时，系统从不同通道召回item

召回通道（召回方式）包括：

* 协同过滤：Item CF、Swing、User CF
* 向量召回：双塔模型、Deep Retrieval（字节发布）等
* 其他：GeoHash、作者召回、相似作者召回、缓存召回等

**Item cf**：Item-based Collaborative Filtering，基于物品的协同过滤，通过计算物品之间的相似度来推荐用户可能感兴趣的物品。例如，如果用户A喜欢电影X和电影Y，而用户B喜欢电影X，那么可以推荐电影Y给用户B。

**Swing**：Swing是一种改进的协同过滤算法，结合了物品之间的共现关系以及用户行为数据，能够更准确地捕捉用户对物品的兴趣。

**User CF (User-based Collaborative Filtering)**：基于用户的协同过滤，通过计算用户之间的相似度来推荐物品。例如，如果用户A和用户B有相似的兴趣爱好，那么可以推荐用户A喜欢的物品给用户B。

**双塔模型 (Two-Tower Model)：**双塔模型是一种神经网络结构，用于将用户和物品映射到同一个向量空间，从而通过计算向量之间的相似度来进行推荐。两个塔分别用于处理用户和物品的特征。由用户塔和物品塔组成，用户塔的主要作用是**将用户的离散特征**（如用户ID、用户的历史行为等）转化为一个固定维度的嵌入向量，这个嵌入向量能够捕捉用户的偏好和行为特征。物品塔的主要作用是**将物品的离散特征（**如物品ID、类别、描述等）转化为一个固定维度的嵌入向量，这个嵌入向量能够捕捉物品的属性和特征。

图示

描述已自动生成

**GeoHash**：GeoHash是一种地理编码技术，将地理位置信息转换为字符串，便于进行地理位置相关的召回。例如，可以根据用户当前位置推荐附近的餐厅或商店。

**作者召回 (Author-based Retrieval)**：根据用户对某些作者的兴趣进行召回，例如，如果用户喜欢某个作者的多篇文章，可以推荐该作者的其他作品。

**相似作者召回 (Similar Author Retrieval)**：通过计算作者之间的相似度，推荐用户可能感兴趣的相似作者的作品。

**缓存召回 (Cached Retrieval)：**利用缓存技术，快速召回之前已经计算好的推荐结果，提高推荐系统的响应速度。

*这些技术和方法结合使用，可以在召回阶段高效地筛选出一批可能感兴趣的物品，为后续的粗排、精排和重排提供基础。*

## 2.2 粗排

简要介绍：粗排通常为较小模型，主要目标是「快速」给物品打分，取分数较高的物品喂给

精排模型

粗排是一个非常容易照本宣科的地方，因为粗排不是必需的环节。如果你的候选数量非常少，那连召回都不需要了；如果你的精排能吃的下召回的输出，那可以考虑实验对比是不是需要粗排。但是假如不加粗排，总感觉欠缺点什么。有的地方甚至出现过粗排输出候选变少，整个系统反而涨点的情况。像这样的情况出现，就说明整个链路设计存在不合理的地方。

数量级：几百

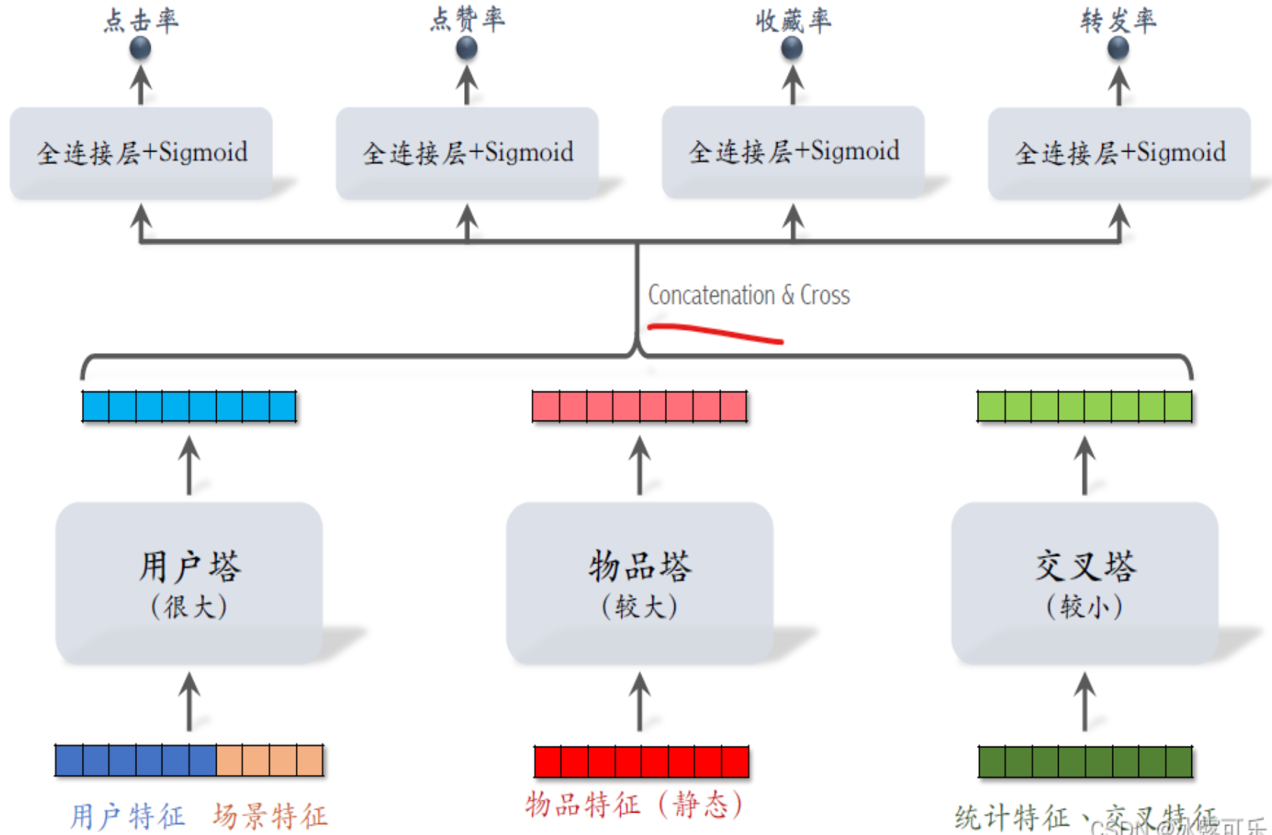
模型：三塔模型（在小红书2022），复杂度在召回双塔和精排模型之间，小红书的粗排是三塔模型，三塔模型，顾名思义有三个塔，分别是用户塔、物品塔、交叉塔。

**用户塔的输入是用户特征和场景特征。**

**物品塔的输入只有物品特征**

**交叉塔的输入包括统计特征和交叉特征**，交叉特征是指用户特征与物品特征做交叉。

三个塔分别输出三个向量，这三个向量做concat后交叉，得到一个向量，把这个向量送入多个头，他们输出点击率、点赞率等指标的预估



2.3 精排

精排用一个较大的神经网络给几百篇笔记打分，精排模型比粗牌模型大很多，用的特征也更多，所以精排模型打的分数更可靠，但是精排的计算量很大。这就是为什么我们先用粗排做筛选，然后才用精排，这样做可以比较好的平衡计算量和准确性。

模型：多目标排序模型（MMOE）目标排序模型（MMOE，Multi-gate Mixture-of-Experts）是一种用于多任务学习的模型结构，特别适用于需要同时优化多个目标（例如点击率、转化率等）的推荐系统。

2.4重排

此时的结果还存在一些不足，需要做一些调整。这一步叫做重排，重排主要是考虑多样性，

要根据多样性做随机抽样，从几百篇笔记中选出几十篇，然后还要用规则把内容相似的笔记打散。稍后我会解释重排，重排的结果就是最终展示给用户的物品，比如把前80的物品展示给用户，其中包括笔记和广告。

数量级：几百->几十

重拍方式：推荐系统链路上的最后一环是重排，重排最重要的功能是多样性抽样。

需要从几百篇笔记中选出几十篇笔记，常见的方法有MMR和DPP抽样的时候有两个依据，

**一个依据是精排分数的大小，另一个依据是多样性**。

做完抽样之后，会用规则打散相似内容。我们不能把内容过于相似的笔记排在相邻的位置上。

举个例子，根据金牌得到的分数，排前五的笔记全都是NBA的内容，这样就不太合适。

即使用户是个篮球迷，他也未必希望看到同质化的内容。如果排第一的是NBA的笔记，那么接下来几个位置就不能放NBA的内容，相似的笔记会往后挪。重排的另一个目的是插入广告和运营，推广的内容还要根据生态的要求调整排序，比如不能连接出很多美女图片。