推荐系统中的召回

小小挖掘机 2020-07-26

以下文章来源于CS的陋室 ，作者机智的叉烧

CS的陋室

陋室，用知识装点。房主主要谈论与数学和计算机相关的知识，不定时推送和个人学习...

召回的重要性，在最近工作的逐步了解中慢慢有了更多的理解吧，召回其实对丰富度的要求很高，后面排序建设的再完美，好东西没有被召回，排序模型将没有任何意义，考虑这个角度，其实我们需要在召回层一定准确程度（要求很低）的情况下，尽可能多召回用户可能喜欢的东西。

在此基础上，我还看到了很多方法，我的想法是给大家拓宽一下思路，看看召回会有哪些比较简单有用的方法。在这里先列举出来，后面逐一细说：

- CB，content base。基于物料的召回。
- CF。大家常说的协同过滤。
- 基于统计信息。这个其实就是所谓的热门召回的具体实现版本。
- 运营手段。
- 向量和模型召回。这其实就是大家常看到的一些召回方式了。
- 物料冷启动手段。
- 兴趣试探。

CB

基于物料信息的召回其实非常简单，实现起来成本不高且收益非常直接，应该是推荐系统前期的就能快速建设的东西了，成本不高收益不低，且是白盒，可控性还是挺高的。但是其实简单的归类为CB可能有些不合适，但很多简单的规则其实就是以物料为中心去做的。

首先聊的是最简单的规则，“捆绑”，电商场景，乒乓球拍配乒乓球，游戏机配游戏卡，手机配手机壳，这都是一些用户可能高频出发的高关联项，这个其实可以通过关联规则之类的挖掘出来，例如“啤酒和尿布”的经典例子。

然后就是“用户画像”，用户在对一些行为后，通过分析用户点击的物料的规律，可以发现一些可以参考的标签喜好，例如恐怖小说看得多，那我们其实可以给用户推恐怖小说，玩射击类游戏多，那就可以继续推射击类游戏，看体育新闻多那就可以推体育新闻。

再者，对于一些冷启动或者比较低频的用户，可以通过人群画像来进行召回，例如北京人喜欢看相声，那新来一个北京用户，就可以尝试和给他推相声，当然人群的划分方式很多，可以根据实际场景和标签来圈取人群推荐。这个虽然不是一个千人千面，但也算是一些千人百面的情况，能提升冷启用户的体验，让用户向活跃用户转化。

总结起来，我之所以想分一个CB出来，其实是因为这一路全都要求对物料有非常高的理解能力，球拍和球是相关的，小说是恐怖的，新闻是体育的，这都要求我们对物料有很高程度的理解。

CF

协同过滤这一路其实不太想聊，因为最基础的推荐算法应该就是协同过滤了，推荐系统国内比较早的书应该要数项亮的《推荐系统实践》了，里面花费不少精力在讲这个东西，没有理解的话希望还是能去看看书尽快理解好。我之前也自己手写实现了一遍，连续写了几篇文章讨论这个东西，大家可以看看：

- 【RS】协同过滤-基础篇
- 【RS】协同过滤-user_based
- 【RS】协同过滤-item_based
- 【RS】协同过滤-进阶篇

都说了要拓宽大家思路嘛，那我就谈谈这个协同过滤有什么比较有用的玩法了，其实协同过滤就是找相似的人推这个人点击的东西，或者找点击过的物料的相似物料来推荐，那么这里的核心就是这个“相似”怎么定义了，常规的协同过滤或者是协同过滤的例子都是用重合度来衡量相似的，但实际上只有这个方法吗？显然不是的，举个例子，我们矩阵分解得到的用户向量物料向量，是不是可以用做相似度判断的判据？这个相似度的衡量是不是会更加精准？大家可以试试。

统计信息

这个可能是推荐系统里最不费脑子的方案了，在线根据一些统计指标为用户推荐内容，例如1小时点击率，1小时点击量飙升等，根据某一些指标为物料做类似榜单的东西，这些东西推荐给用户，其实就是我们常言的“热门”，仔细想类似1小时内高点击率这不就是热门了吗。

这路很简单，但也非常重要，除了考虑热点本身的意义，还承担着多样性的重要职责，在项目早期没有兴趣试探的时候，破除所谓的“信息茧房”很可能只能靠这种东西了，现在问问自己，当提到召回的时候，自己的视角里有没有这一路召回？

运营手段

一般来说，会给运营专门整一个池子，运营可以根据他们的分析情况进行配置，配置完以后这些内容会给特定用户稳定召回，可以后续直接参与排序，也可以设置置顶之类的强插项，当然这个升级版还有广告等，这个玩法可以更加多。

模型和向量召回

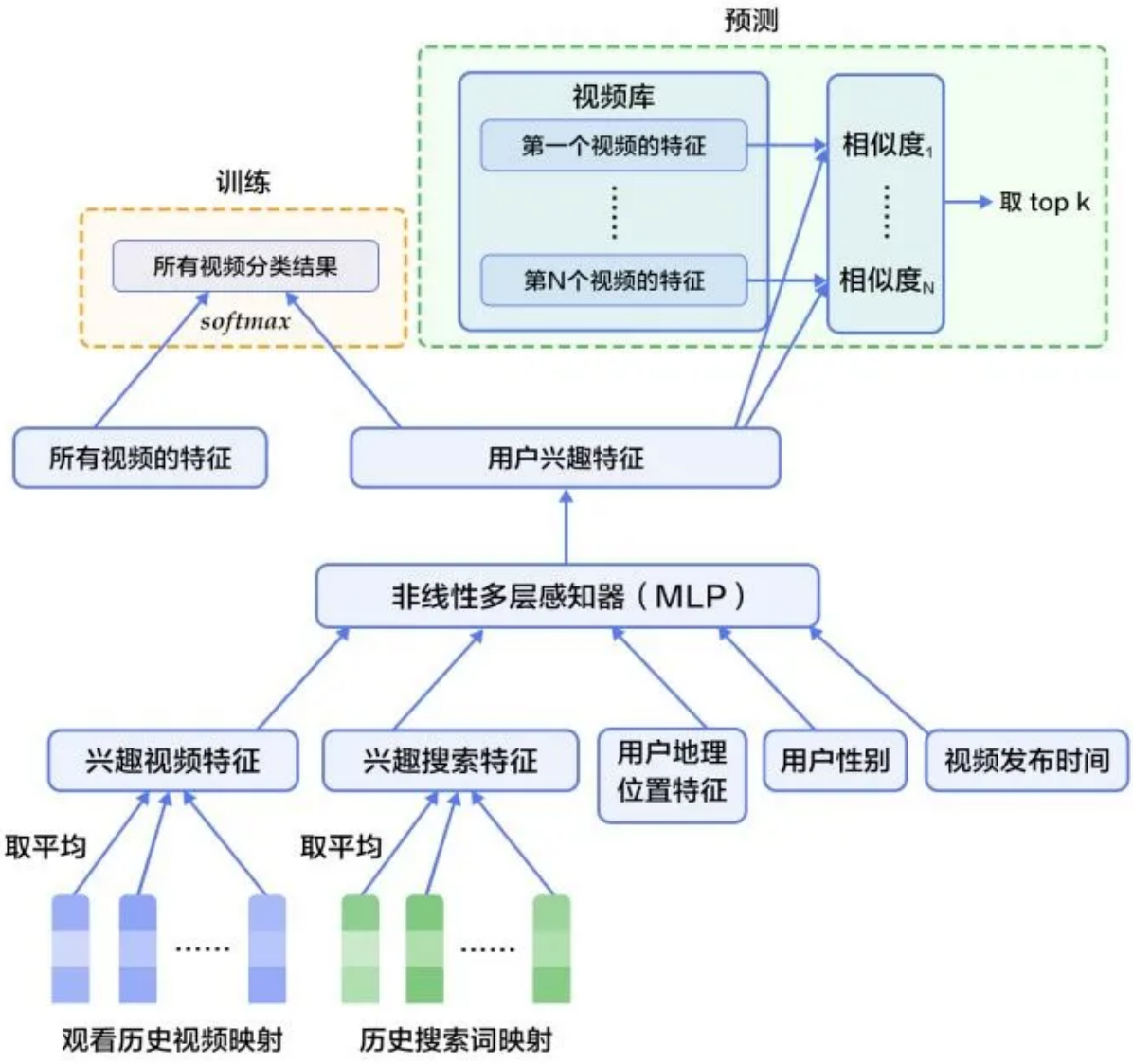
慢慢开始接近一些热门的操作了，向量化、embedding的思想最初来源于NLP，这个很多人应该都听说过，对物料也好，对用户也好，有一个比较抽象的向量表示后，其实可以做很多操作了。

- 举个例子，之前说的协同过滤，找相似用户推荐相似用户的历史点击，这个相似用户可以通过用户embedding衡量。

当然了，升级版本就是一些模型。简单举几个例子吧：

- DSSM。来源于NLP的语义相似度模型，用的时候是把标签从是否相似换成是否点击。
- Youtobe。
- MF，矩阵分解。

很多人读了各种召回模型会觉得很奇怪，这么一个预测模型其实只是衡量了一个用户和一个物料之间的关系，但是在线时对一个用户，肯定不可能一个一个计算一遍再进行召回，我们来看看youtube具体是怎么操作的，参考文献见文末。



youtube这里，训练阶段其实是用多个用户特征结合视频特征去训练了分类器，这个分类器只是一个抓手，实际上是去训练用户的embedding和视频的embedding，有了这个embedding后，在线就可以通过向量召回的方式获取用户相关的视频。

这里提到了向量召回，这其实是一个搜索中使用的技术，通过建立索引、划分区域等方式来实现这种最近邻相似，这个最近邻相似的应用就在最简单的机器学习方法——KNN，里面的核心任务也是快速找到TOP K，具体方案有这几个：KD tree、ball tree、LSH等，当然，也可以比较粗暴的划分区域，如geohash等，精度会下降但是速度肯定有所提升。

物料冷启动

上面的内容基本都是基于用户兴趣给用户推荐内容，但现在问题来了，新来的物料，用户是否喜欢，我们是无从得知的，因此我们需要做的是给这些物料一些曝光机会，让用户有机会接触到，获取一些反馈，才能进行更加个性化的预测，这就有点死循环了，无曝光不展示-无反馈不推荐，这就和要工作经验才能找到工作一样，在这里，我们需要有一些在用户角度相对激进的方式来处理——物料试探。粗暴的，我们可以随机给少量用户推荐一些新物料，看用户是否点击，点击多的说明质量比较优质。

当然，这种试探也不会盲目地试，如果新物料有一些先验信息，那我们就可以应用，例如新闻是体育新闻，那我们可以给喜欢体育新闻的用户推送，这样能缩小口径，这种试探也会更加精准。

用户兴趣探索

说完物料冷启动，说说用户兴趣探索，这是一个冷启动用户很高效的方法，要快速探索新用户的兴趣，除了基本的人群画像外，还可以考虑更加多元化的兴趣探索。

这里就要介绍多臂老虎机了。这个比喻可以说是非常灵活了，有很多个老虎机，开始阶段我们会去试探那个老虎机赚的钱多，后续阶段我们自然就开始去玩赚钱多的那个，推荐系统里面，其实就是给用户很多类目选择，例如新闻里面有娱乐新闻、体育新闻、科技新闻，那么我们都给用户推，后面用户点击率高的日后就多给用户推一些，从而快速提升用户粘性。

更加详情用户可以参考，入门看这个会比较好懂：https://zhuanlan.zhihu.com/p/84140092

小结

推荐系统接触了也有段时间了，结合干了这么久搜索，越发感觉到一个问题，已经入行的都在深耕，做了一些深度比较高的操作并且做了总结，营销的媒体则是什么容易火就蹭什么，隔三差五一个神坛，重新定义，这其实很误导初学者，让初学者的视角出现很多偏差，例如我发现很多初学者视角里完全没有统计召

回、甚至协同过滤之类的东西，这个很可怕。我其实蛮想和大家分享这些东西，让大家视角能更加全面一些，别一天到晚模型模型的，有一说一，大部分场景不用模型就能解决。

最后，告诫大家，用模型之前先问问自己，用模型的原因是自己只会模型，还是现在这个问题非模型不可。