

互联网 | PM能看懂的个性化推荐（二）：常见召回通道（1）

原创 卿一 掌仙人儿 2019-05-08

犯规了噻（二） - （1）是怎么肥四？

文章太长不便阅读，强行掐断咩哈哈~~

算法在很多人看来深不可测，在与算法开发工程师沟通的过程中，会发现很多数据变化难以求得解释。

一方面这门功夫的确有很高的的壁垒，没有一定算法常识，可能很难理解。

另一方面算法不像我们定一个简单的策略（比如按照XX阈值捞取数据，按照XX排序，处理什么特殊情况）那么简单粗暴，而是多个引擎交叉影响的结果。这个过程是相对黑盒的，即使让工程师去解释也未必能有清晰的因果分析。我们只是尝试了这样的策略，分桶实验发现结果是好的，至于这个因子为什么能有这样的正向影响，是不一定有线性解释的。

就像经典的啤酒和尿布现象。沃尔玛发现啤酒和尿布放在一起会刺激两者的销量，解释是母亲在家带孩子，父亲出来买尿布，顺便买啤酒。这可能只是个传说，但却能粗略类比算法的因子和结果之间，有无数的**中介效应**。

但是这并不意味着算法推荐完全无章可循，虽然影响因子千千万，但是归宗的几大类还是可以简单了解一下。

用户行为

日常与人的沟通中，你要把你的观点装到别人的脑海中，首先要理解对方的想法，否则只是你一个人一厢情愿的强制灌输（歪个楼，强烈推荐《非暴力沟通》）。

做推荐系统就与日常交流一样，你要想让推荐内容吸引到用户，首先要做的是倾听用户的喜恶。

【信息收集】

我们在倾听用户时，有两种形式。

I 用户画像：用户在使用产品时，在每个角落会留下无数可以利用的痕迹，这些痕迹无不透露着用户心思，丰满用户的画像。我们通过这些数据，逐渐认识这名用户。

II 行为学习：我们的每套系统往往对应特定的推荐区域，用户在这个区域进行的交互，就像是在与我们进行有来有往的沟通，我们通过这些行为训练机器进行深度学习，校准针对性的推荐。

用户在特定区域的行为，包含常说的正负反馈。正反馈标识用户对该物品有明显的“兴趣”，负反馈则相反。正负反馈又分为显性反馈和隐性反馈。

- 正+显：购买、观看、收藏、打高分；
- 正+隐：点击、频繁浏览、高停留时长；
- 负+显：点击“不喜欢”、“减少此类推荐”；
- 负+隐：物品展示了多次但用户并未进行点击

e.g.: 红色：显性正反馈；蓝色：显性负反馈 ▼



【信息处理】

用户行为收集到之后，就要考虑怎样运用这些数据信息。一般利用用户行为进行的推荐算法，**协同过滤**流行很多年了。协同过滤也有很多算法模型可以运用，目前最广泛的是：

- I **基于用户**：A用户和B用户喜好相近，如果A对某物品表现强烈兴趣，也会把该物品推给B。
- II **基于物品**：A物品和B物品总是被同一拨人喜欢，如果某用户对A表现强烈兴趣，也会给该用户推荐B

看起来好像挺简单的~~但是具体过程细思极恐，比如在计算用户相似度时，如果用户量非常大，占用过大的计算资源，就要考虑建立物品到用户的倒排表，处理矩阵，得

出余弦相似度。再比如计算物品相似度时，如果某个物品特别热门，就会和很多物品“相似”，就要想办法用公式“惩罚”这种热门物品在相似度计算时的优势。

讲到这里再提一个有意思的小点，以证明算法推荐更多需要理性分析而非感性判断。对于高活跃用户，我们应该更“重视”他的行为，还是降权他的行为？我之前认为应该更重视，因为高频活跃代表用户的忠实度，如果更多参考他的行为，也会有利于提高整体用户活跃度。而事实上，越高频活跃的用户，有可能行为越离散，可参考性越低。比如一个每天买东西却不怎么用的购物狂，并不能提升他所买物品之间的关联度。

标签

标签是啥呢？是描述用户/物品的独立、没有明显层次组织的短词条。我认为它是我们认识一个对象（咳咳...）的最直观快速的方式，也是在推荐算法中建立喜恶关系的捷径之一。

【标签简述】

我又要给标签分类了：

I 用户标签：标识一个用户的基本面貌

- 社会自然：年龄、性别、受教育程度、城市、行业（比如斜杠青年的词条）
- 物品偏好：建立用户和物品的联系，根据产品特征决定，比如新闻用户的tag会有“社会”、“时尚”等；音乐用户的tag会有“雷鬼”、“jazz”等

II 物品标签：标识一个物品的基本属性

- 来源于PGC：比如上一篇中潘多拉的例子，由专业人员对物品打标签

- 来源于UGC：用户对物品打标签，比如豆瓣。UGC标签两个好处：第一，标签置信度更高，因为用户基数大，利用大数据可以校准少数人打标签的个人主观影响，而且可以计算出每个标签的关联度；第二，用户打出来的标签更客观地表现了用户和物品的联系，可以更好地反作用于推荐。

豆瓣打标签 ▼

添加收藏：修改 x

☐ 想看

☒ 看过

给个评价吧?(可选):

★★★★☆ 推荐

收起 ▲

标签(多个标签用空格分隔):

我的标签:

喜剧青春战争动作美国经典动画爱情电影

美剧传记科幻历史悬疑

常用标签:

超级英雄情怀动作2019漫威漫画改编美国科幻

感动史诗

简短评论:

245

☐ 仅自己可见

分享到 ☒ 豆瓣广播 [去绑定新浪微博](#)

保存

【标签利用】

标签在内容分发中的应用分为可感知和不可感知两种方式。

I 可感知：用户可见的标签。比如分类筛选页，或者某个内容的简介中展示标签。这种简短精炼的短词条可以降低用户对内容的理解门槛，辅助快速决策。

标签感知方式主要取决于产品形态是否能更好刺激用户，这里不多讲，多注意标签脏数据清理，比如无义词或者敏感词。

II 不可感知：主要指推荐系统中，对于用户和物品的标签关联。

基于标签推荐可以分为两步，第一步：用户打标签；第二步：利用标签给用户推荐物品。

- 打标签：如果你的平台有ugc基因那就太好了，你可以刺激鼓励用户在你这里对内容进行各种反馈，也包括打标签。通过用户打的标签，你可以得到宝贵的标签数据，也可以利用用户的打标签行为校准对用户喜好的判断。如果你的平台没有ugc基因也不必气馁，引入专家意见或者去同类的ugc网站抓取都是可以尝试的手段。
- 利用标签推荐：引入标签这个中介后，就可以针对用户&物品&标签建立三元组甚至多元组，简单来说最终希望计算出用户喜欢的物品标签，然后把这个标签关联度高的物品推荐给用户。这里也要注意脏数据清理，比如“不喜欢”这种否定词，用户给很多物品打了“不喜欢”标签，也不能给他推荐总被打“不喜欢”标签的物品。

利用标签推荐有个好处就是特别容易解释，一般给用户展示推荐理由会提升推荐置信度，提升转化。上文介绍的协同过滤有时候就不太好解释，“爱买尿布的用户也爱买啤酒”？这怎么说得出口.....但是“你可能喜欢酒类”就很容易接受。

小结

这篇讲了两个召回通道：利用用户行为和利用标签。下一篇会继续讲其他召回通道，欢迎追更哟~

END