协同过滤算法

1. 算法介绍

协同过滤简单来说是利用某兴趣相投、拥有共同经验之群体的喜好来推荐用户感兴趣的信息，个人通过合作的机制给予信息相当程度的回应（如评分）并记录下来以达到过滤的目的进而帮助别人筛选信息，回应不一定局限于特别感兴趣的，特别不感兴趣信息的纪录也相当重要。其核心思想是如果目标用户和某一用户再某些物品评分上很相似，那么目标用户对新物品的评分与该用户对新物品的评分也是类似的。

协同过滤方法可以大致分为两类：基于邻域的方法和基于模型的方法。

1. 基于邻域

在基于领域的协同过滤方法中，用户对物品的历史评分数据可以用来预测用户对新物品的评分。基于领域的方法包括两种：基于用户的推荐和基于物品的推荐。在基于用户的协同过滤方法中，目标用户对某一未接触物品的感兴趣程度，是由和用户具有相似评分模式的其他用户（近邻用户）对该物品的评分来估计的。基于物品的协同过滤方法，是根据某一用户对目标物品相似物品（近邻物品）的评分来预测用户对该目标物品的评分。

与基于邻域的方法不同的是，基于模型的方法使用评分信息来学习预测模型。主要思想是使用属性构建用户和物品之间的联系，这里的属性表示用户和物品的潜在特征，比如用户喜欢的类别和物品所属的类别。

基于邻域的协同过滤有以下优点：

* 简单性：直接且容易实现
* 可解释性：推荐结果具有直观的解释性理由
* 高效：基于模型的协同过滤系统在训练阶段需要消耗大量资源
* 稳定：一旦计算完成，后续可进行增量计算。对于具有评分的新物品或新用户加入时，仅需要计算这些新对象和系统已有对象的相似度即可。

基于邻域的协同过滤的缺点：

* 覆盖受限：由于计算两个用户间的相似度是基于他们对相同物品的评分，而且只有对相同物品进行了评分的用户才可以作为近邻。这样仅仅被近邻用户评价过的物品才会被推荐，推荐方法的覆盖将受到限制
* 对稀疏数据的敏感：基于邻域的推荐方法的准确性会因为评分数目的缺少而受到影响。由于用户通常只是对一小部分物品进行了评分，导致评分数据具有稀疏性。一方面稀疏数据导致在推荐时只使用了非常有限的近邻，另外，相似度计算也只是依赖了少量的评分，导致相似度存在偏差，影响推荐结果

1. 基于模型