在现今的推荐技术和算法中，最被大家广泛认可和采用的就是基于协同过滤的推荐方法。本文将带你深入了解协同过滤的秘密。下面直接进入正题

1 什么是协同过滤

协同过滤是利用集体智慧的一个典型方法。要理解什么是协同过滤 (Collaborative Filtering, 简称 CF)，首先想一个简单的问题，如果你现在想看个电影，但你不知道具体看哪部，你会怎么做？大部分的人会问问周围的朋友，看看最近有什么好看的电影推荐，而我们一般更倾向于从口味比较类似的朋友那里得到推荐。这就是协同过滤的核心思想。

换句话说，就是借鉴和你相关人群的观点来进行推荐，很好理解。

2 协同过滤的实现

要实现协同过滤的推荐算法，要进行以下三个步骤：

收集数据——找到相似用户和物品——进行推荐

**收集数据**

这里的数据指的都是用户的历史行为数据，比如用户的购买历史，关注，收藏行为，或者发表了某些评论，给某个物品打了多少分等等，这些都可以用来作为数据供推荐算法使用，服务于推荐算法。需要特别指出的在于，不同的数据准确性不同，粒度也不同，在使用时需要考虑到噪音所带来的影响。

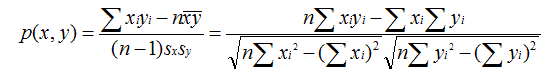
**找到相似用户和物品**

这一步也很简单，其实就是计算用户间以及物品间的相似度。以下是几种计算相似度的方法：

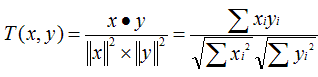
　　欧几里德距离

IMG_256 IMG_257

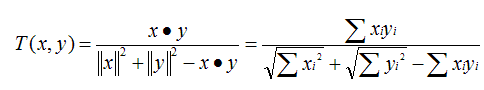
　　皮尔逊相关系数



　　Cosine 相似度



　　Tanimoto 系数

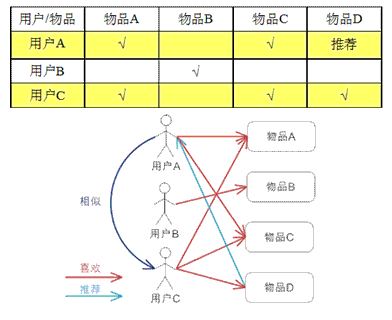


**进行推荐**

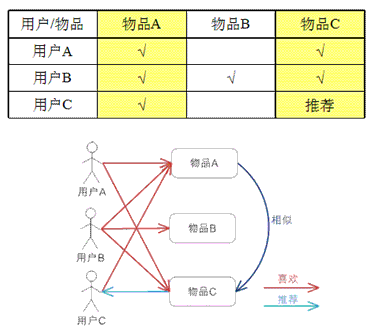
在知道了如何计算相似度后，就可以进行推荐了。

在协同过滤中，有两种主流方法：基于用户的协同过滤，和基于物品的协同过滤。具体怎么来阐述他们的原理呢，看个图大家就明白了

基于用户的 CF 的基本思想相当简单，基于用户对物品的偏好找到相邻邻居用户，然后将邻居用户喜欢的推荐给当前用户。计算上，就是将一个用户对所有物品的偏好作为一个向量来计算用户之间的相似度，找到 K 邻居后，根据邻居的相似度权重以及他们对物品的偏好，预测当前用户没有偏好的未涉及物品，计算得到一个排序的物品列表作为推荐。 下图给出了一个例子，对于用户 A，根据用户的历史偏好，这里只计算得到一个邻居 - 用户 C，然后将用户 C 喜欢的物品 D 推荐给用户 A。



基于物品的 CF 的原理和基于用户的 CF 类似，只是在计算邻居时采用物品本身，而不是从用户的角度，即基于用户对物品的偏好找到相似的物品，然后根据用户的历史偏好，推荐相似的物品给他。从计算的角度看，就是将所有用户对某个物品的偏好作为一个向量来计算物品之间的相似度，得到物品的相似物品后，根据用户历史的偏好预测当前用户还没有表示偏好的物品，计算得到一个排序的物品列表作为推荐。下图给出了一个例子，对于物品 A，根据所有用户的历史偏好，喜欢物品 A 的用户都喜欢物品 C，得出物品 A 和物品 C 比较相似，而用户 C 喜欢物品 A，那么可以推断出用户 C 可能也喜欢物品 C。



总结

　　以上两个方法都能很好的给出推荐，并可以达到不错的效果。但是他们之间还是有不同之处的，而且适用性也有区别。下面进行一下对比

　　计算复杂度

　　Item CF 和 User CF 是基于协同过滤推荐的两个最基本的算法，User CF 是很早以前就提出来了，Item CF 是从 Amazon 的论文和专利发表之后（2001 年左右）开始流行，大家都觉得 Item CF 从性能和复杂度上比 User CF 更优，其中的一个主要原因就是对于一个在线网站，用户的数量往往大大超过物品的数量，同时物品的数据相对稳定，因此计算物品的相似度不但计算量较小，同时也不必频繁更新。但我们往往忽略了这种情况只适应于提供商品的电子商务网站，对于新闻，博客或者微内容的推荐系统，情况往往是相反的，物品的数量是海量的，同时也是更新频繁的，所以单从复杂度的角度，这两个算法在不同的系统中各有优势，推荐引擎的设计者需要根据自己应用的特点选择更加合适的算法。

　　适用场景

　　在非社交网络的网站中，内容内在的联系是很重要的推荐原则，它比基于相似用户的推荐原则更加有效。比如在购书网站上，当你看一本书的时候，推荐引擎会给你推荐相关的书籍，这个推荐的重要性远远超过了网站首页对该用户的综合推荐。可以看到，在这种情况下，Item CF 的推荐成为了引导用户浏览的重要手段。同时 Item CF 便于为推荐做出解释，在一个非社交网络的网站中，给某个用户推荐一本书，同时给出的解释是某某和你有相似兴趣的人也看了这本书，这很难让用户信服，因为用户可能根本不认识那个人；但如果解释说是因为这本书和你以前看的某本书相似，用户可能就觉得合理而采纳了此推荐。

　　相反的，在现今很流行的社交网络站点中，User CF 是一个更不错的选择，User CF 加上社会网络信息，可以增加用户对推荐解释的信服程度。