**学习进度**

本周学习任务：

学习Part I General Recommendation Techniques：A Comprehensive Survey of Neighborhood-Based Methods for Recommender Systems.

下周学习任务：

学习Part I General Recommendation Techniques：Advances in Collaborative Filtering

1. 基本认知

**相较于基于内容的推荐方法**，协同过滤不依赖于质量未知的内容信息，而是使用用户评级（显式反馈）和使用模式（隐式反馈）等信息，用于生成针对特定用户的商品推荐。同时，协同过滤方法在不断添加新用户或新项目的在线环境中，具有更为显著的稳定性。

**协同过滤可以划分为基于邻域和模型两大类别**，前者直接使用用户历史行为数据预测用户对于未知物品的评级，后者则使用同类型数据训练而成的预测模型进行推荐服务。其中，基于邻域的方法：能够在离线阶段预先计算最近邻，并于在线阶段提供近乎瞬时的推荐；亦能提供关于推荐的合理解释，为交互系统打下坚实的基础。然而，基于邻域的方法存在数据稀疏性、局部化的相似性计算及冷启动问题。其中，局部化的相似性计算是指，在推荐过程中只考虑到了与目标对象非常相似的部分近邻，导致了推荐的覆盖范围受限。而相较于基于邻域的方法，基于模型的方法则更擅长使用潜在因素描述用户偏好及物品特征。然而，基于模型的方法具备更高预测性能的同时，也相对更缺乏意外发现珍奇事务的才能。

值得注意的是，**基于邻域与模型方法的区别**在于：基于邻域的方法直接利用计算而来的相似度信息进行推荐，而基于模型的方法则通过参数学习的方式来预测用户对于未知物品的评级。基于邻域的优化过程，主要涉及相似度计算方式、邻域的选择及预测规则等方面的改进。而基于模型的优化过程，类似于机器学习模型的学习过程，涉及模型结构、目标函数、优化算法及预测规则等方面的改进与调整。

值得补充的是，**推荐任务可以归结为评级预测及Top-N推荐两类问题**，前者预测用户对于未见物品的评分，后者预测用户最可能感兴趣的N个商品。当评级数据可用时，主要考虑评级预测问题，此类问题又可细分为回归与分类两种，使用的准确度度量主要有：平均绝对误差MAE及均方根误差RMSE，公式如下：



而当评级数据不可用时（比如：只知道用户购买过的物品列表，不知道衡量预测结果准确性的评级标签），主要考虑Top-N推荐问题。同评级推荐一样，Top-N问题也需要将用户交互过的物品列表划分为训练与测试集。前者用于训练模型，使模型在训练过程中尽可能地拟合训练集数据；后者用于测试模型，用于评价与调整模型关于预测用户对某物品喜好程度的能力。Top-N推荐问题的评价指标有：准确率、召回率及平均逆命中率（Average Reciprocal Hit-Rank, ARHR），公式如下：





其中，L(u)指用户u最感兴趣的N项物品，T(u)指测试物品中用户u认为相关的物品。由此，准确率描述了最终的推荐列表L(u)中有多少比例是发生过的用户-物品记录，召回率描述了有多少比例的用户-物品记录包含在最终的推荐列表中。而对于ARHR而言，L(u)是根据用户u感兴趣程度排列好的N项物品有序列表，rank()表示每个用户u实际交互过的每个物品在L(u)中的排名（若该物品不在L(u)中，则1/rank=0）。故此，ARHR通过对所有用户求加和平均的方式，得到整个数据集上的ARHR(L)值。其值越大，则表明系统给定的Top-N推荐列表中，实际交互过物品的排名越靠前，则推荐效果越好。

1. 基于邻域的一般方法

**在基于邻域的推荐方法中，基于用户或物品是较为基础的两种实现方法**。其中，基于用户通过借鉴同用户u最近邻的部分用户对于项目i的评分，预测用户u对于项目i的评分；基于物品则通过借鉴用户u评价过且与项目i最近邻的若干项目的评分，预测目标用户u对于目标项目i的评分。

对于**基于用户的评级预测**而言，需要分别讨论回归及分类两种情形。当评分刻度连续时，评级预测属于回归任务。此时，首先需要根据用户相似度矩阵，确定k个同目标用户u最相似且对目标物品i评过分的近邻用户；然后引入作为权重来衡量不同近邻用户对于物品i评分的重要程度，并权重归一化；最后还需通过定义尺度标准化函数的方式，从而缓解不同用户关于评分标准不同，对于评分预测结果造成的的负面影响。具体公式，如下左式所示。

 

当评分刻度离散时，评级预测属于分类任务。此时，需要根据k个近邻用户对于目标物品i的评分完成投票，从而选择投票最多的离散评级，作为用户u对于目标物品i的评级预测值。具体公式，如上右式所示。

对于**基于物品的评级预测**而言，也需分别讨论回归及分类两种情形。当评分刻度连续时，评级预测属于回归任务。此时，首先需要获取用户u评价过的且与目标项目i最相似的若干项目；然后根据不同近邻项目与目标项目i的相似度，求用户u对于不同近邻物品评分的加权平均值；最后额外考虑用户不同的评分标准即可。具体公式，如下左式所示。而当评分刻度离散时，评级预测属于分类任务，具体公式如下右式所示，具体解释参上。

 

关键问题在于，**选择使用基于用户或物品邻域方法的准则是什么？**从解释性来看，由于预测中用到的近邻物品列表及相似度权重都可以作为推荐的解释提供给用户，基于物品更具解释性。从惊喜度来看，基于用户能够产生更新颖的推荐，尤其是参考的近邻用户很少时。从稳定性来看，需要考虑系统中用户和物品的改变频率和数量，如果用户改变频率更快，则基于物品的方法更稳定。从准确性和效率来看，需要考虑系统存储的用户数和物品数比例，当用户远多于物品时基于物品更准确且计算相似度所需内存和时间往往更少，反之亦然。

1. 相似度权重计算

用户或物品相似度，可以用于选择可信的近邻，也可以在预测规则中作为权重来衡量不同近邻评分的参考程度。常见的计算方法如下：

**皮尔逊相似度（Pearson Correlation, PC）**通过减去各自均值的方式完成了评分归一化，消除了评分量级的影响。而且PC的值域为[-1,1]，可以更直观地解释两者间线性相关性的强弱程度。此外，这种基于评分均值与方差的计算方式，对于异常值更具鲁棒性。计算用户或物品间相似度的公式，分别如下左右所示。

 

其中，为同时被用户u和v评价过的物品集合，为同时评价过物品i和j的用户集合。

**均方差（Mean Squared Difference, MSD）**则通过平方评分间平均差异的方式，更为强调关注评分差异较大的用户或物品。然而，MSD无法考虑用户或物品间的负关联。具体公式，如下所示。



**斯皮尔曼等级关联（Spearman Rank Correlation, SRC）**根据评分排名而非评分值完成计算，可以捕捉数据间的任意单调关系，特别适合于不需要线性关系假设的场景。同时，SRC使用的评分排名是一种非参数统计量，不依赖于数据的具体分布假设，绕开了标准化评分问题。此外，SRC适用于各种类型的数据（分类、顺序和连续数据等）且对异常值鲁棒。然而，当评分只有少量可选值时，SRC将会产生大量的并列排名，从而造成性能不佳。

具体公式，如下所示。其中，表示物品i在用户u所评分物品中的排名（并列评分则用它们的平均排名），表示用户所评物品的平均排名。



值得讨论的是，相较于用户数和项目数，评级数经常是稀疏的，且相似度也仅是通过很少一部分对相同物品的评分或同一用户所做的评分来计算获得的。故此，可以根据评分数量**度量相似度的可信程度**，主要有重要性权重和基于收缩两种方法。其中，重要性权重使用阈值γ（超参γ经验为50）来确定何时惩罚相似度，公式如下所示。

 

而相较于重要性权重，基于收缩的方法（超参β经验为100）更为连续，公式如下所示。

 

显而易见的是，用户对于普遍受欢迎物品的评分，所蕴含的信息量更少。而获得更少评价的物品，有时候更能反映评分用户的个性化差异。由此，基于逆用户频率（越少的人与项目i交互取值越大）改进的皮尔逊关联系数，应运而生。**频率加权皮尔逊关联系数（Frequency-Weighted Pearson Correlation, FWPC）**公式，如下所示。



此外，**在计算用户间相似度时，额外考虑用户评价过的物品与指定参照物品间的关系**，或许能够获得更可靠的用户间相似度值。然而，这种考虑需要额外计算，不适合应用于在线推荐系统。基于PC的改进公式，如下所示。



1. 邻域的选择

近邻数量及其选择标准，会对推荐系统的质量产生严重的影响，实验证明近邻数量和预测精度间的关系遵循凸函数。故此，关键在于如何存储更少相似权重的同时，保证推荐系统的预测性能。具体而言，对于**邻域的预过滤处理**，主要的方法有：阈值过滤（保留相似度大于给定阈值的权重）、负过滤（根据实际效果决定不考虑负评级）及Top-N过滤（每个用户或项目只适当地保留N个近邻及其各自的相似度权重列表）等。值得注意的是，前述三种过滤方法，可以结合起来使用。