Homework 7:
混合推荐系统设计

大数据原理与技术 (Spring 2025)

22336226 王泓沣
Lectured by: Changdong Wang
Sun Yat-sen University

1 问题描述

- 基于 MovieLens 数据集实现协同过滤（用户/物品相似度）与基于内容的推荐（电影标签），对比两者召回率。
- （可选）设计加权混合策略或其他策略提升推荐多样性。

2 数据描述

这个数据集（ml-32m）来自电影推荐服务网站 http://movielens.org，描述了用户的 5 星评分和自由文本标签行为。它包含了针对 87,585 部电影的 32,000,204 条评分和 2,000,072 条标签应用。这些数据由 200,948 位用户在 1995 年 1 月 9 日至 2023 年 10 月 12 日之间创建，并于 2023 年 10 月 13 日生成。
随机选取用户来组成这个数据集。所有被选中的用户都至少对 20 部电影进行了评分。数据集中没有包含任何人口统计信息。每个用户仅用一个 ID 表示，未提供其他信息。
数据存储在 `links.csv`、`movies.csv`、`ratings.csv` 和 `tags.csv` 文件中。
实验中使用评分超过 1000 条的用户和电影作为数据集，训练集和测试集划分为 4:1

3 Method

3.1 协同过滤

使用用户相似度进行协同过滤。首先构建用户-物品评分矩阵，并对每个用户的评分进行均值归一化，去除不同用户评分习惯的影响。接着，将归一化后的矩阵转换为稀疏矩阵，并利用余弦相似度计算用户之间的相似性，得到用户相似度矩阵。对于目标用户，从相似度矩阵中选取最相似的若干用户（例如前 50 个），并对这些相似用户对各电影的评分进行加权求和，排除目标用户已评分的电影后，输出加权得分最高的 10 个电影作为推荐结果。

3.2 基于内容推荐

基于内容推荐主要利用电影的标签（tags）信息进行实现。首先，对电影标签进行文本预处理，并采用 TF-IDF 方法将标签转换为向量表示，从而构建电影的特征矩阵。随后，利用余弦相似度计算所有电影之间的相似度，形成电影内容相似度矩阵。对于目标用户，选取其评分高于一定阈值（例如 4.0）的电影作为“喜欢”的电影，并计算这些电影与其他电影之间的平均内容相似度得分；如果目标用户没有评分较高的电影，则直接采用评分总和较高的热门电影作为推荐候选。最后，返回得分最高的 10 个电影作为推荐结果。

3.3 混合策略

混合推荐策略结合了协同过滤和基于内容推荐两种方法的优势。具体实现过程如下：首先分别计算目标用户的协同过滤得分和基于内容推荐得分。协同过滤部分采用目标用户与最相似用户的加权评分，而基于内容推荐部分则根据目标用户喜欢电影的标签相似度计算电影的得分。接着，对两部分得分分别进行归一化处理，以消除数值尺度的差异。最后，通过线性加权融合两种得分，即令

$$\text{混合得分} = \alpha \times \text{协同过滤得分} + (1-\alpha) \times \text{内容得分},$$

其中 α（默认值为 0.5）用于平衡两种方法的贡献。根据混合得分对未评分的电影进行排序，最终选取得分最高的 10 个电影作为推荐结果。

4 Result

4.1 协同过滤

选取最相似用户数	召回率
5	0.1082
10	0.1302
50	0.1730

Table 1: 选取最相似用户数超参分析

4.2 基于内容推荐

评分阈值	召回率
3.0	0.0052
4.0	0.0066
5.0	0.0122

Table 2: 评分阈值超参分析

4.3 加权混合推荐

协同过滤权重	召回率
0.5	0.09275
0.6	0.1088
0.7	0.12485

Table 3: 权重超参分析