**大数据分析第六次作业**

电影推荐系统算法是基于人人相似的协同过滤算法和基于内容的推荐算法相结合的混合推荐算法

**混合推荐算法大致流程:**

首先对数据集使用人人相似的协同过滤算法, 计算每个人与其他人的cos相似度, 选择相似度最高的人,推荐他看过的电影. 若最高相似度小于0.2则认为没有人与之相似, 改为采用基于内容的推荐算法解决. 若相似度最高的用户没有评分过4分的电影可推荐, 也采用基于内容的推荐算法进行推荐.

**人人相似的协同过滤算法:**

1. 根据movieId合并movies.csv和ratings.csv,得到merged.csv
2. 读取merge.csv获取data字典, 格式为{userid:{movieid:rating, movieid:rating}}
3. 修正data字典, 具体为计算每部电影平均分, 将该电影的所有评分减去平均分, 使得该电影的未评价用户的默认评分为0, 以便于计算
4. 对于每个用户, 计算他与其他所有用户的cos相似度, 记录最高相似度的用户和两个用户的相似度, 最后获得最高相似度的用户列表
5. 将与该用户相似度最高的用户的电影进行排序, 并与该用户的电影进行比较, 选择该用户未看过且评分大于4的电影推荐给该用户
6. 如果出现最高相似度小于0.2, 或者最高相似度的用户没有电影可以推荐时, 采用基于内容的推荐算法

**基于内容的推荐算法**

1. 获取三种数据结构, 用户标签喜爱度列表, 不同标签的电影的评分排序以及每一部电影的种类
2. 用户标签喜爱度列表:

标签喜爱度计算公式=所在电影的原始权重\*所在电影的用户评分\*时间加权

其中, 所在电影原始权重=1/该电影的标签数

时间加权=时间戳的归一化

1. 不同标签的电影的评分排序: 评分是指每一部电影的平均分

数据结构: {tag:[ (movieId, rating), (movieId, rating) ]}

1. 每一部电影的种类:

数据结构: { movieId: { tag:ratio, tag:ratio }}

1. 获取每个用户标签喜爱度排名前两名的标签和它们的权重, 根据权重计算该标签推荐的电影数, 在不同标签的电影评分排序这个数据结构中找到该标签下排名前几名的电影进行推荐