大数据分析实验
——推荐系统（大实验）

目录

一 任务背景

一 任务背景

- 1、了解推荐系统的多种推荐算法并理解其原理。
- 2、实现User-User的协同过滤算法并对用户进行推荐。
- 3、实现基于内容的推荐算法并对用户进行推荐。
- 4、对两个算法进行电影预测评分对比
- 5、两种推荐算法中，加入minhash算法对效用矩阵进行降维处理

二 任务描述

二 任务描述

◆ 任务一

数据集：给定Recommend-data.zip电影数据集，包含电影评分文件（ratings.csv，原始电影评分数据，仅供参考），训练集train_set.csv和测试集test_set.csv（原始电影评分拆分出来的两个数据集），movies.csv电影信息数据。

基于用户的协同过滤推荐算法。对训练集中的评分数据构造用户-电影效用矩阵，使用pearson相似度计算方法计算用户之间的相似度，也即相似度矩阵。（子任务1）对单个用户进行推荐时，找到与其最相似的k个用户，用这k个用户的评分情况对当前用户的所有未评分电影进行评分预测，选取评分最高的n个电影进行推荐。（子任务2）此外，在测试集中包含100条用户-电影评分记录，用于计算推荐算法中预测评分的准确性，对测试集中的每个用户-电影需要计算其预测评分，再和真实评分进行对比，误差计算使用SSE误差平方和。

二 任务描述

◆任务二

基于用户的协同过滤推荐优化方法。此方法采用minhash算法对效用矩阵进行降维处理，从而得到相似度矩阵，注意minhash采用jaccard方法计算相似度，需要对效用矩阵进行01处理，也即将0.5-2.5的评分置为0，3.0-5.0的评分置为1。

二 任务描述

◆ 任务三

将数据集movies.csv中的电影类别作为特征值，计算这些特征值的tf-idf值，得到关于电影与特征值的n（电影个数）*m（特征值个数）的tf-idf特征矩阵。根据得到的tf-idf特征矩阵，用余弦相似度的计算方法，得到电影之间的相似度矩阵。对某个用户-电影进行预测评分时，获取当前用户的已经完成的所有电影的打分，通过电影相似度矩阵获得已打分电影与当前预测电影的相似度，按照下列方式进行打分计算：

$$\text{score} = \frac{\sum_{i=1}^{n} score'(i) * sim(i)}{\sum_{i=1}^{n} sim(i)}$$

选取相似度大于零的值进行计算，如果已打分电影与当前预测用户-电影相似度大于零，加入计算集合，否则丢弃。（相似度为负数的，强制设置为0，表示无相关）假设计算集合中一共有n个电影，score为我们预测的计算结果，score’(i)为计算集合中第i个电影的分数，sim(i)为第i个电影与当前用户-电影的相似度。如果n为零，则score为该用户所有已打分电影的平均值。要求能够对指定的userID用户进行电影推荐，推荐电影为预测评分排名前k的电影。userID与k值可以根据需求做更改。推荐算法准确值的判断：对给出的测试集中对应的用户-电影进行预测评分，输出每一条预测评分，并与真实评分进行对比，误差计算使用SSE误差平方和。

二 任务描述

◆任务四

基于内容的协同过滤推荐优化方法。此方法使用minhash算法对基于内容推荐算法的相似度计算进行降维，把最小哈希的模块作为一种近似度的计算方式，从而得到相似度矩阵，注意minhash采用jaccard方法计算相似度，特征矩阵应为01矩阵。因此特征矩阵选取采用方式为，如果该电影存在某特征值，则特征值为1，不存在则为0，从而得到01特征矩阵。

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

三 算法流程

三 基于协同过滤的推荐系统算法流程

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

计算用户间的相似度 → 选取最相似的K个用户 → 对推荐对象计算预测值

pearson相似度

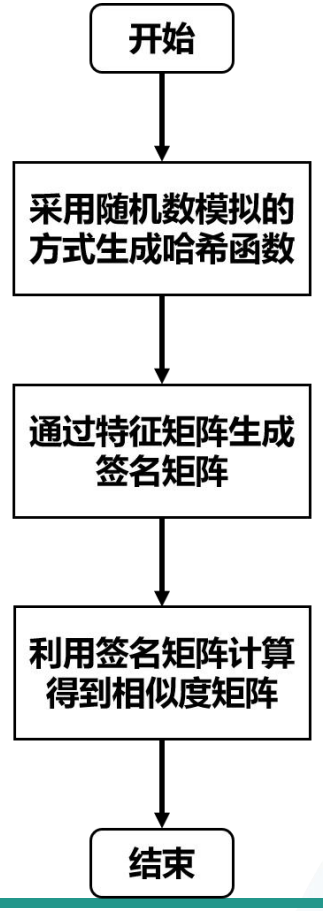
k可以取30

思考：对于K个用户，若某个电影只有部分用户评分，没有评分的部分怎么算？
（取0分？取平均值？数据填补方法 KNN？）

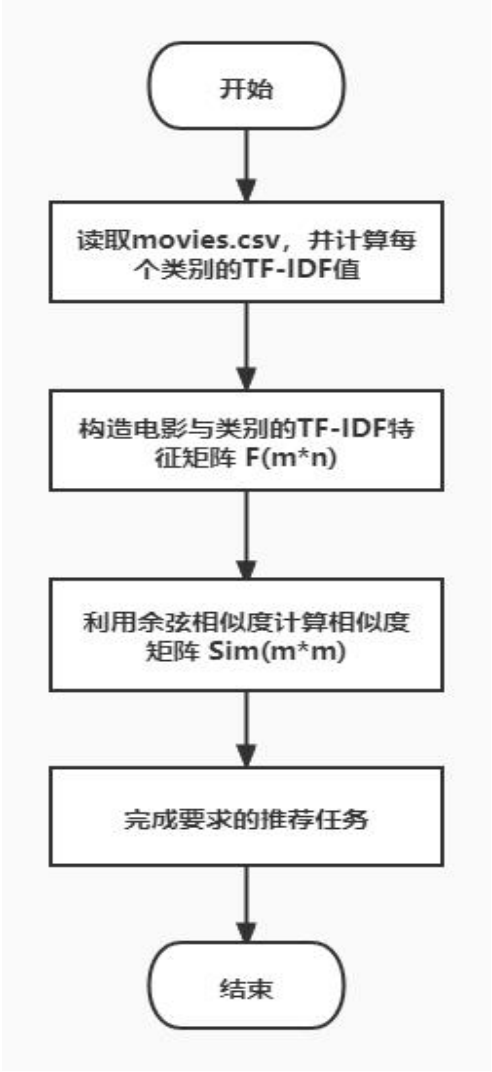
三 基于协同过滤的推荐系统算法-优化

使用minhash相似度计算用户之间的相似度

两个集合的随机的一个行排列的minhash值相等的概率和两个集合的Jaccard相似度相等



三 基于内容的推荐系统算法流程



sub_task1:

对userId=4的用户进行排名前5的推荐。

=> 预测该用户对所有电影的打分，然后输出预测打分最高的前5部电影，以及相应的分数。

sub_task2：

利用误差平方和公式评估推荐算法的准确性。

=> 测试集中的真实打分与预测分做差值，得出误差平方和SSE，以对算法进行评估。

$$S_E = \sum (x - \hat{x})^2$$

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

四 验收流程

四 验收事项

- 4个任务下的推荐算法的SSE值，针对给定的userID用户进行推荐前k个电影。
- 验收时对代码的大致解释；
- 验收时的提问与回答。