

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级：**

**学 号：**

**姓 名：**

**指导教师：**

**报告日期：**

**计算机科学与技术学院**

**目录**

[实验五 推荐系统算法及其实现 1](#_Toc198739600)

[**1实验目的** 1](#_Toc198739601)

[**2 实验内容** 1](#_Toc198739602)

[**3 实验过程** 1](#_Toc198739603)

[3.1 编程思路 1](#_Toc198739604)

[3.2 遇到的问题及解决方式 1](#_Toc198739605)

[3.3 实验测试与结果分析 1](#_Toc198739606)

[**4 实验总结** 2](#_Toc198739607)

# 实验五 推荐系统算法及其实现

## **1实验目的**

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、 加入minihash算法对效用矩阵进行降维处理。

## **2 实验内容**

给定MovieLens数据集，包含电影评分，电影标签等文件，其中电影评分文件分为训练集train\_set和测试集test\_set两部分。完成以下任务：

任务一：**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。**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。预测评分按照以下方式计算：**

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**。

任务二：**基于内容的推荐算法**

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：

**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**。

**任务三：使用迷你哈希进行降维**

主要内容是使用**迷你哈希（MiniHash）算法对协同过滤算法和基于内容推荐算法的相似度计算进行降维**，迷你哈希的模块作为一种近似度的计算方式。

协同过滤算法和基于内容推荐算法都会涉及到相似度的计算，迷你哈希算法在牺牲一定准确度的情况下对相似度进行计算，其能够有效的降低维数，尤其是对大规模稀疏01矩阵。可以**使用哈希函数或者随机数映射来计算哈希签名**。哈希签名可以计算物品之间的相似度。

对于任务一，minihash采用jarcard方法计算相似度，需要对效用矩阵进行01处理，也即将0-2.5的评分置为0，2.5-5.0的评分置为1。对于任务二，minihash采用jarcard方法计算相似度，特征矩阵应为01矩阵，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。

最终降维后的维数等于我们定义映射函数的数量，我们设置的映射函数越少，整体计算量就越少，但是准确率就越低。**大家可以分析不同映射函数数量下，最终结果的准确率有什么差别**。

## **3 实验过程**

### 3.1 编程思路

### 3.2 遇到的问题及解决方式

### 3.3 实验测试与结果分析

## **4 实验总结**