

**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 1](#_Toc167890341)

[**3.1 实验内容** 1](#_Toc167890342)

[**3.2 实验过程** 2](#_Toc167890343)

[3.2.1 编程思路 2](#_Toc167890344)

[3.2.2 遇到的问题及解决方式 2](#_Toc167890345)

[3.2.3 实验测试与结果分析 2](#_Toc167890346)

[**3.3 实验总结** 2](#_Toc167890347)

# 实验五 推荐系统

## **3.1 实验内容**

**必做：**

给定Anime数据集，包含用户对动漫评分、动漫标签等文件，其中动漫评分文件分为训练集train\_set和测试集test\_set两部分

**基础版必做一：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-动漫效用矩阵，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动漫进行评分预测，选取评分最高的n个动漫进行推荐。预测评分按照以下方式计算：**

在测试集中包含100条用户-动漫评分记录，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-动漫需要**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**。

**基础版必做二：基于内容的推荐算法**

将数据集anime.csv中的动漫类别作为特征值，**计算这些特征值的tf-idf值**，得到关于动漫与特征值的n（动漫个数）\*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**动漫**之间的相似度矩阵**。

对某个用户-动漫进行预测评分时，获取当前用户的已经完成的所有动漫的打分，通过动漫相似度矩阵获得已打分动漫与当前预测动漫的相似度，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，如果已打分动漫与当前预测用户-动漫相似度大于零，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个动漫，score为我们预测的计算结果，score’(i)为计算集合中第i个动漫的分数，sim(i)为第i个动漫与当前用户-动漫的相似度。如果n为零，则score为该用户所有已打分动漫的平均值。

**要求能够对指定的userID用户进行动漫推荐，推荐动漫为预测评分排名前k的动漫。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-动漫进行预测评分，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**。

**选做（进阶）部分：**

进阶部分的主要内容是使用**最小哈希（minhash）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**。同学可以把最小哈希的模块作为一种近似度的计算方式。

注意minhash采用jarcard方法计算相似度，需要对矩阵进行01处理。

对于**协同过滤推荐算法**，可以**将效用矩阵中动漫0 - 5的评分置为0，5 - 10的评分置为1**。

对于**基于内容推荐算法，可以判断tf-idf矩阵中是否存在某特征值，如果存在则特征值为1，反之则为0。**

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**。

**对基于用户的协同过滤推荐算法和基于内容的推荐算法进行推荐效果对比和分析，选做的完成后再进行一次对比分析。**

## **3.2 实验过程**

### 3.2.1 编程思路

### 3.2.2 遇到的问题及解决方式

### 3.2.3 实验测试与结果分析

## **3.3 实验总结**