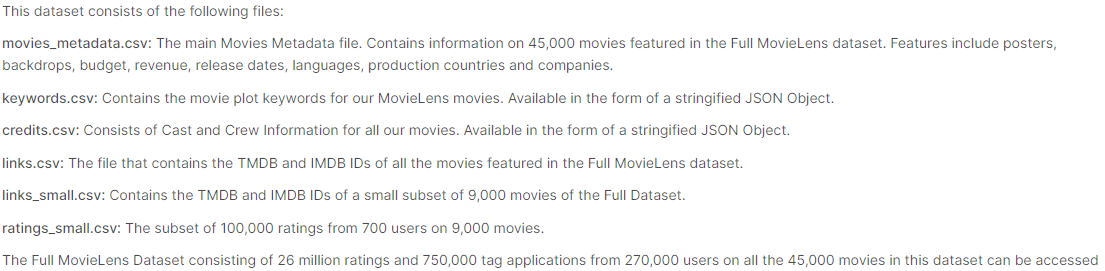
**开题报告**

MG20330022蒋申 MG20330046祁一帆 MF20330011陈雨铨

1. **研究题目和背景**

研究题目是基于Hadoop mapreduce对电影评分分析及推荐。我们计划从互联网中获取大约270000位用户对45000个电影26000000个评分数据。每个评分在1分到5分之间包括用户ID与电影ID，每部电影有自身的相关信息（如语言、制片商、国家、类型等）的标签。大致信息如下图所示：



我们希望将这些评分信息进行大数据分析，可以实现对每个观影用户进行电影个性化推荐。

1. **主要技术难点和拟解决的问题**

本次课程设计对电影评论的分析将实现对每个观影用户进行电影个性化推荐，以及web界面方便用户查看推荐内容（暂定是这些功能,后面实现时可能会有所更改）。由于样本数据高达27000000条，在单机上对数据进行清洗和分析将会需要较长处理时间。并且在单机系统上的开发，编程人员的dubug效率会因为运行速度过慢而变得过低，对于程序开发十分不利。因此本次实验将基于Hadoop mapreduce来实现本次课程设计，使用并行化技术来缩短处理数据的时间。本次课程设计需要解决的主要问题包含电影推荐算法的设计和对用户电影偏好的分析。主要技术难点在于对于一个电影而言，如何通过用户评分和自身标签评价其特点与其质量，对于一个用户而言，如何通过其历史评分记录判断用户偏好从而推荐相关电影。

1. **基本解决方法和设计思路**

本次课程设计的解决方法可以分为两个主要步骤：

1. 利用聚类算法将用户的偏好进行分类
2. 利用频繁项集挖掘对用户进行个性化推荐
3. **利用聚类算法将用户的偏好进行分类**

首先根据每个用户已有的打分信息，生成该用户的偏好向量。具体而言，比如用户Bob一共有两个评分信息，其中对电影A打分5，对电影B打分3，电影A的标签信息为美国、动作片，电影B的标签信息为美国、爱情片，那么Bob的偏好向量在标签“美国”的分值为5和3的平均值即为4，在标签“动作片”的分值为5，在标签“爱情片”的分值为3，对于其他没有涉及到的标签如“中国”、“喜剧片”等，分值取最高分5和最低分1的平均值即3。通过这样的方法，可以为每个用户生成一个远低于电影部数的偏好向量，并且对于后续处理更容易分析出数据中隐含的偏好信息。

在获得了每个用户的偏好向量后，通过聚类算法将用户分为K类，其中K的取值可能根据经验获得或者视后续推荐算法的效果而改变。

1. **利用频繁项集挖掘对用户进行个性化推荐**

对于一个需要个性化推荐的用户而言，将其电影评分记录输入系统中，系统自动计算出该用户的偏好向量，首先将该用户归入最相似的偏好类别中。系统预先对该偏好类别中所有电影进行频繁项集挖掘，寻找包含该用户看过的好评电影的项集，将这些项集中该用户未看过的电影推荐给该用户。

1. **课题分工**

蒋申负责web与系统的数据对接与界面的设计，祁一帆负责实现对用户的聚类算法，陈雨铨负责对同类用户观影记录的频繁项集挖掘。