

推荐算法作业描述

数据集

数据集从以下链接选择, 下载 `small` 版本的数据集.

<https://grouplens.org/datasets/movielens/latest/>

MovieLens Latest Datasets

These datasets will change over time, and are not appropriate for reporting research results. We will keep the download links stable for automated downloads. We will not archive or make available previously released versions.

Small: 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users. Last updated 9/2018.






- [README.html](#)
- [ml-latest-small.zip](#) (size: 1 MB)

Full: approximately 33,000,000 ratings and 2,000,000 tag applications applied to 86,000 movies by 330,975 users. Includes tag genome data with 14 million relevance scores across 1,100 tags. Last updated 9/2018.

- [README.html](#)
- [ml-latest.zip](#) (size: 335 MB)

Permalink: <https://grouplens.org/datasets/movielens/latest/>

数据集包含如下文件, 详细介绍见README.

 <code>README.txt</code>	2018/9/27 4:50	文本文档	9 KB
 <code>links.csv</code>	2018/9/27 4:50	Comma Separat...	194 KB
 <code>movies.csv</code>	2018/9/27 4:49	Comma Separat...	483 KB
 <code>tags.csv</code>	2018/9/27 4:49	Comma Separat...	116 KB
 <code>ratings.csv</code>	2018/9/27 4:49	Comma Separat...	2,426 KB

数据集划分: 在进行训练和测试时, 需要将 `ratings.csv` 中的数据划分为**训练集(70%)**和**测试集(30%)**

任务描述

原始数据集划分为训练集 (`trainset`) 和测试集 (`testset`), 其均包含 `ratings.csv` 中的字段.

目标: 根据用户的历史行为信息的信息 (即 `trainset`) 向目标用户推荐没有见过的电影 (即 `testset`)

完整流程包括

1. 使用 `trainset` 的数据训练推荐模型

2. 对 `trainset` 中的用户推荐电影

1. 从 `trainset` 中选择 `user_id` 作为输入
2. 利用训练得到的推荐模型, 输入 `user_id`, 输出推荐电影列表 `movies_list`
3. 对推荐结果进行评价

评价方法

输入: `user_id`, 目标用户id

输出: `movies_list`, 推荐电影列表

真实值: `target_movie_list`, 即 `testset` 中对应用户打过分的电影 (即用户可能会看的电影)

推荐命中次数: `hit=(movies_list==target_movie_list)`, 即推荐的电影等于真实值的次数

评价标准1, 精确率: `precision = hit/(len(movies_list))`

评价标准2, 召回率: `recall = hit/(len(target_movie_list))`