
实验一 第 2 阶段 豆瓣数据的个性化检索与推荐

实验背景

豆瓣 (www.douban.com) 是一个中国知名的社区网站，以书影音起家，用户可以在豆瓣上查看感兴趣的电影、书籍、音乐等内容，还可以关注自己感兴趣的豆友。

在索引实验的基础上，结合用户的评价信息及用户间社交关系，进行个性化电影、书籍推荐。

实验要求

本次实验要求分组完成，每组最多 3 人（可以少于 3 人，但无优惠政策）。

推荐任务：基于豆瓣 Movie&Book 的 tag 信息、我们提供的豆瓣电影与书籍的评分记录以及用户间的社交关系，判断用户的偏好。在这个阶段中，你们需要对用户交互过的 item（电影、书籍）进行（基于得分预测的）排序。

实验内容

（1）数据集说明

在第 2 阶段中，我们在阶段 1 数据的基础上提供了社交网络信息和用户评分信息，有关数据的详细说明如下：

1) **“contacts.txt”**，为社交网络信息。

例如，一条记录为：A: B, C, D，则意味着 A 与 B、C、D 三位用户之间存在社交关系，这里的社交关系是双向的（或无向的）。

因为实验数据进行了筛选，而社交网络数据没有做筛选，所以其中可能包含若干未在评分记录中出现的用户 ID，需要自行过滤。至于是否需要利用社交网络信息，如何利用这部分数据，请同学自定。

2) **“Movie_score.csv”** 与 **“Book_score.csv”** 为用户的评分信息，具体内容格式如下：

User ID, Item (Movie/Book) ID, Rating (0-5), Timestamp[, Tag 1, Tag 2, ...]

例如：1000001, 1293510, 3, 2005-06-26T20:41:22+08:00, black humor 表明，ID 为 1000001 的用户给电影 1293510 打了 3 分，时间为 2005-06-26T20:41:22+08:00，同时留下了 black humor 的标签。

具体文件地址如下（数据集以及样例代码）：

链接：<https://rec.ustc.edu.cn/share/cbc5cc30-6bcf-11ee-8cfe-95b0b178faec>

（2）任务说明

在这次实验中，我们会给出训练集与测试集的划分代码，在测试集上为用户对书籍和电影的评分进行个性化排序，并结合真实评分采用 NDCG 对自己的预测结果进行评分和进一步分析，同时也可以借助 MSE 等指标进行辅助分析，比较不同指标下模型的表现情况。助教会给出全部流程的样例代码，可以进行参考或者部分采纳，但严禁照搬全抄。

要求完成实验：

- A. 采用协同过滤方式，仅利用用户-项目（电影或书籍）的评分矩阵进行评分预测。可以在基本协同过滤的基础上自行添加部分策略，如考虑评分的情境或周期性等因素，并讨论这些因素会对排序精度产生何种影响。
- B. 根据提供的 tag 等文本信息进行辅助预测，辅助形式自行选择（如：使用 tag 补充书籍的信息）。模型不必复杂，可以使用 tf-idf 或 word2vec 等课上讲过的方法，或 chinese-bert 等预训练模型等表示文本信息，添加到用户/book 的 embedding 中来补全信息（样例代码中仅仅合并 tag 来聚合信息，因此效果有较大提升空间，需要思考如何使用文本信息，以及对于用户或者 Book 等数据，分别该使用哪部分信息）。同时，详细分析引入文本信息作为辅助后，对于实验效果的影响作用。

选做实验：

基于社交网络关系的推荐，可以自行选择方法，利用社交网络关系辅助推荐，如：基于用户聚类、基于社交传播（口碑影响）、基于图神经网络等，并简要分析社会网络在该方案中的作用和效果。采用的方案数量、种类均不限，量力而行。

关于实验的所有部分（输入、输出、评测、模型），包括附加实验，我们均会给出样例代码，但严禁完全抄袭！

根据徐老师的最高指示以及往年惯例，为了将反卷贯彻到底，本次实验不以最终结果为指标，无论结果好坏，只需针对结果给出分析（针对效果变好/变差的例子给出解释）即可。

具体而言，任务流程大致如下：

① 数据划分

我们提供了划分用户数据的代码，部分打分分值数据被抹去，以作为实验中用于预测的数据，即：用户与这些电影/书籍交互过，但（假装）不知道得分。

有一些用户的评分数据过少（其实数据已经够稠密了），你们可以自行决定是否使用这些数据进行分析或预测。因此，可以不必完全按照示例代码，根据个人需要进行适当修改。

② 评分排序

你们需要对上面抹去分值的对象进行顺序位置预测，即：若以升/降序排序用户的所有评价，那这些数据应该放在第几位。将你们预测出的对象顺序与实际的顺序进行比较，并用 NDCG（全部数据或 Top k）评估你们的预测效果。

同学们可能注意到了，在这里我们的用词是“顺序”，即不一定要预测用户的实际评分，给出合理的顺序即可（当然，也可以先预测评分再排序）。如果同学需要预测评分，可以参考课件使用 kNN 或 SVD 等方法，使用 MSE 等指标进行评价。

我们给出的数据除了评分本身，还有社交关系/tag/时间戳等，若有需要同学可以自行取用。

③ 结果分析

你们需要根据上面的得分对自己的方法和结果进行一定分析，若采用了不同的方法，也可以比较不同方法的结果。同时，你们需要保留预测结果和过程代码以备助教查验。

在实验报告中你们需要对以上几步里你们的分析、采用的方法、取得的效果进行举例和阐释。同时你们需要保留本次实验的预测结果供助教查验，这些数据不用提交。

(3) 样例代码说明

- Data 文件夹存储所有实验数据

-
- Score 文件存储用户的交互记录，包含 item, time, rate, tag 等，自行取用其中 tag 等信息
 - Tag 文件是对 item 的所有 tag 进行聚合后的结果，实际实验中未必有效，自行取用
 - Contacts 文件包含用户的所有社交关系
 - 代码
 - Text_embedding 文件是 mf 与 text embedding 的实现方法
 - Graphrec 文件是一种基于社交网络的推荐方法（其余的文件均是他的配置文件）

提交说明

请于截止日期（**2024 年 11 月 20 日晚 23:59**）以前提交到课程邮箱 ustcweb2022@163.com，具体要求如下：

1. 邮件标题以及压缩包命名为"组长学号-组长姓名-实验 1"格式。邮件正文中请列出小组所有成员的姓名、学号。
2. 因未署名造成统计遗漏责任自行承担，你可以将邮件抄送你的队友。
3. 实验报告请务必独立完成，如果发现抄袭按 0 分处理。
4. 迟交实验将不被接收。
5. **整个实验一只需提交一份实验报告，请在实验一两个阶段的全部内容完成后，将内容合在一个实验报告内统一提交，请不要分开提交以免影响评分。**