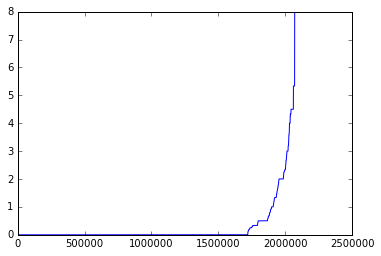
亚马逊电影推荐系统

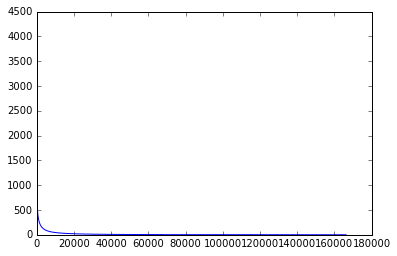
## 数据预处理

数据取自<https://snap.stanford.edu/data/web-Amazon.html>，总共有用户电影数据4607047条，对最初的数据进行了一下初步的分析:



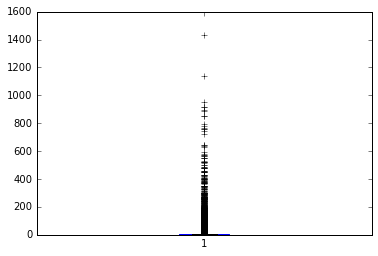
**用户评分方差图**

通过统计，在所有的200多万的用户中，有大约175万的用户要么只有一条评分数据，要么对所有电影的评分都一样，我们认为这样的数据不利于做预测，于是舍弃掉了这一部分数据。



**电影评分数量曲线**

如图，整个电影数据集当中总共有大约17万部电影，但是大部分的评论都集中在前2万部电影当中，在实验当中我们取了前5000部电影最为item集合。



**用户评价数目盒图**

通过统计每个用户所看的电影的数目，发现有一些用户的评论个数太多，我们怀疑存在有多个用户共享一个账号的情况，于是舍弃掉了这一部分用户，最终我们取了评论数目在10到100的用户。

最后我们取80%的数据做训练数据，20%的数据做测试数据，最终的数据集为：总共334784条评分数据，用户15090个，电影5000部，训练集一共273762条数据，测试集一共61021条数据。

## 解决方案

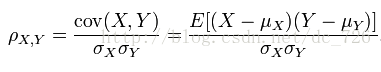
总共用三种进行用户评分预测。

#### 标准方法：

用训练集的平均评分作预测。

#### User-Based CF：

基于用户的协同过滤的主要思想就是找评分最相近的用户，常用的相似度测量方法有余弦相似度，调整余弦相似度，皮尔森相似度。这里我们使用皮尔森相似度进行用户相似度的测量，其公式如下：



#### SVD

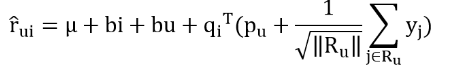
SVD的思想是分解用户-电影评分矩阵，得到用户的特征矩阵P和电影的特征矩阵Q。其优化目标如下：





#### SVD++

SVD++在SVD的基础上加上了用户历史评分的影响，其优化目标是：



其中y表示物品j所隐含的用户信息。

### 实验结果

所有的方法采用RMSE来判断好坏。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | Standard | User-based | SVD | SVD++ |
| RMSE | 1.384 | 1.126 | 1.026 | 1.033 |