**SVD应用于推荐系统**

数据集中行代表用户user，列代表物品item，其中的值代表用户对物品的打分。基于SVD的优势在于：用户的评分数据是稀疏矩阵，可以用SVD将原始数据映射到低维空间中，然后计算物品item之间的相似度，可以节省计算资源。

整体思路：先找到用户没有评分的物品，然后再经过SVD“压缩”后的低维空间中，计算未评分物品与其他物品的相似性，得到一个预测打分，再对这些物品的评分从高到低进行排序，返回前N个物品推荐给用户。

具体代码如下，主要分为5部分：

**第1部分：加载测试数据集；**

**第2部分：定义三种计算相似度的方法；**

**第3部分：通过计算奇异值平方和的百分比来确定将数据降到多少维才合适，返回需要降到的维度；**

**第4部分：在已经降维的数据中，基于SVD对用户未打分的物品进行评分预测，返回未打分物品的预测评分值；**

**第5部分：产生前N个评分值高的物品，返回物品编号以及预测评分值。**

2.slope one 算法思想

Slope One 算法是由 Daniel Lemire 教授在 2005 年提出的一个Item-Based 的协同过滤推荐算法。和其它类似算法相比, 它的最大优点在于****算法很简单, 易于实现****, 执行效率高, 同时推荐的准确性相对较高。

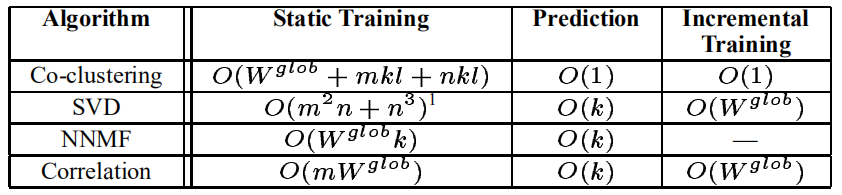
Slope One算法是基于不同物品之间的评分差的线性算法，预测用户对物品评分的个性化算法。主要两步：  
Step1:计算物品之间的评分差的均值，记为物品间的评分偏差(两物品同时被评分)；  
Step2:根据物品间的评分偏差和用户的历史评分，预测用户对未评分的物品的评分。  
Step3:将预测评分排序，取topN对应的物品推荐给用户。

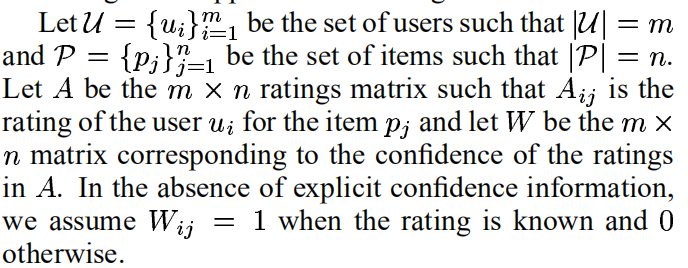
4.slopeOne使用场景

该算法适用于物品更新不频繁，数量相对较稳定并且物品数目明显小于用户数的场景。依赖用户的用户行为日志和物品偏好的相关内容。  
优点：  
1.算法简单，易于实现，执行效率高；  
2.可以发现用户潜在的兴趣爱好；  
缺点：  
依赖用户行为，存在冷启动问题和稀疏性问题。

验证系统：

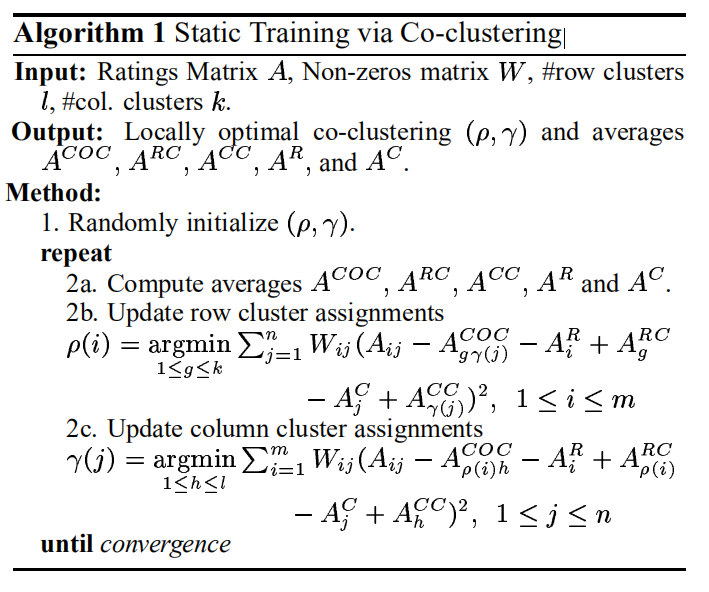
1. 用户体验
2. 加标签

张浩@大数据分析与处理 11:39:18  
各位同学，大家好。本周一无需上传任何视频。今天下午将以组为单位对两周工作做桌面检查。具体如下： 1. 分组进行   2. 组内成果展示（文字、GITHUB、SPRINT成果、接下来的工作、其它与项目相关的工作）3、组员逐个介绍工作进展  4、本次检查为考试的一部分。我们将在今后的课程中采用这个这种方式积分。即使在同一组，分值也可能因此有较大差距。  
  
张浩@大数据分析与处理 11:44:02  
另外，本课程鼓励用英文写文档（BONUS)，开发的软件必须至少是英文界面且界面友好，本课程的软件是以 1. 完整的文档资料（按时上传阶段性文档）  2. 每次的SCRUM MEETING记录（视频上传、现场讨论）   3. 课堂分组考察  4. 实现软件所规定的功能且可稳定运行 5. 积极的课堂活动（出勤、发言、讨论）为完结标志的。大家只要认真做，应该能取得很好的成绩的。但如果没有按时去完成，也是很容易不及格的。  
  
  
张浩@大数据分析与处理 11:46:36  
下午各组带好相关文档和演示工具，将分组讨论，我将根据大家表现给予课程评分。



k:In case of SVD and NNMF, k is the rank of the approximation matrix and for correlation based methods, k=#neighbors

Wglob: is the number of non-zeros in A



5折验证

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| k | 1 | 2 | 3 | 4 | 5 | 方差 | 均值 |
| SVD | 1.6385 | 1.6384 | 1.6350 | 1.6366 | 1.6294 | 1.12x10-5 | 1.63558 |
| BaseLine | 1.6381 | 1.6385 | 1.6353 | 1.6368 | 1.6316 | 6.23x10-6 | 1.63606 |
| NMF | 2.4765 | 2.4777 | 2.4688 | 2.4815 | 2.4656 | 3.48x10-5 | 2.47402 |
| CoClustering | 1.8438 | 1.8494 | 1.8500 | 1.8516 | 1.8424 | 1.33x10-5 | 1.84744 |

时间对比：三次运行取平均

|  |  |
| --- | --- |
| 算法 | 时间(秒) |
| SVD | 2.59 |
| BaseLine | 4.16 |
| NMF | 2.75 |
| CoClustering | 2.87 |