电子商务网站
自动推荐算法研究和实现

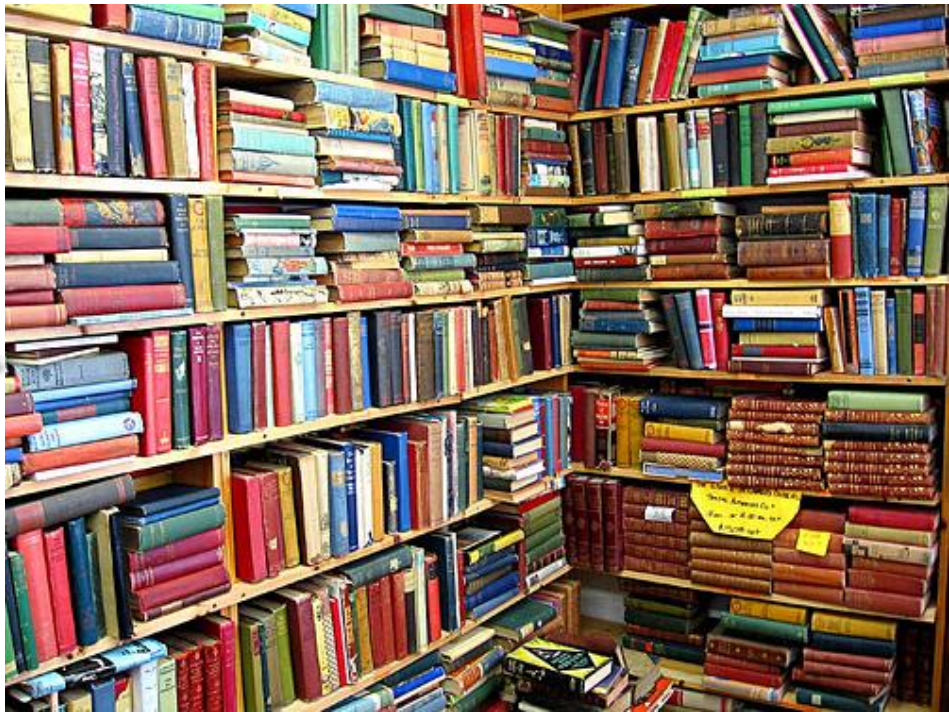
学　　生：康清波
指导老师：于中华

信息过载

推荐系统-目标

- 1. 帮助用户找到他们感兴趣的项目;
- 2. 帮助项目提供者向正确的用户提交他们的项目;
- 3. 帮助网站提高用户的参与。

推荐系统



推荐系统-卓越亚马逊


今日推荐
1 2 3 4
这里显示的是今日推荐的部分商品，点击此处查看所有推荐商品
加速世界6:净火神子
图书
古剑奇谭原声音乐集(4CD)
游戏/娱乐
变形金刚3(DVD9 珍藏版)
音像
伊苏7 豪华版(亚马逊独家'菲娜'特典版)
游戏/娱乐
赛车总动员2(DVD)
音像
捕捉图像(C) <CTRL><SHIFT><C>

推荐系统-卓越亚马逊

经常一起购买的商品

顾客购买此书时也通常购买图像处理、分析与机器视觉(第3版) - 桑卡(Milan Sonka) 平装 ¥ 48.70

价格合计: ¥ 74.90

立即购买组合

查看发货和库存信息

购买此商品的顾客也同时购买

图像处理、分析与机器视觉(第3版) - 桑卡(Milan Sonka)

(13)

¥ 48.70

数字图像处理(MATLAB版) - 冈萨雷斯

(102)

¥ 37.50

数字图像处理(第3版) - 阮秋琦

(23)

¥ 53.90

图像局部不变性特征与描述 - 王永明

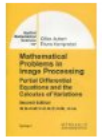
(8)

¥ 24.90

数字图像处理与机器视觉:Visual C++与Matlab实现(附CD-ROM光盘1张) - 张铮

(23)

¥ 51.80



图像处理中的数学问题(第2版)(英文版) - 奥伯特(Gjlles Aubert)

(6)

¥ 33.80

推荐系统-豆瓣猜

豆瓣 douban　首页　友邻广播　我的豆瓣　我的小组　我的小站

[redacted]，你可能喜欢

勿忘蛛

★★★☆☆ 7.9

かつて上代に手練れの陰陽師が大蜘蛛を封印したと言われる一冊の本。それを入手した古書店の視とビルのオーナーの孫

情敌大战

导演: 约瑟夫·麦克金提·尼彻

★★★☆☆ 7.0

塔克（汤姆·哈迪 Tom Hardy 饰）和FDR（克里斯·派恩 Chris Pine 饰）是美国中情局的顶尖探员，两人各自身怀绝技，

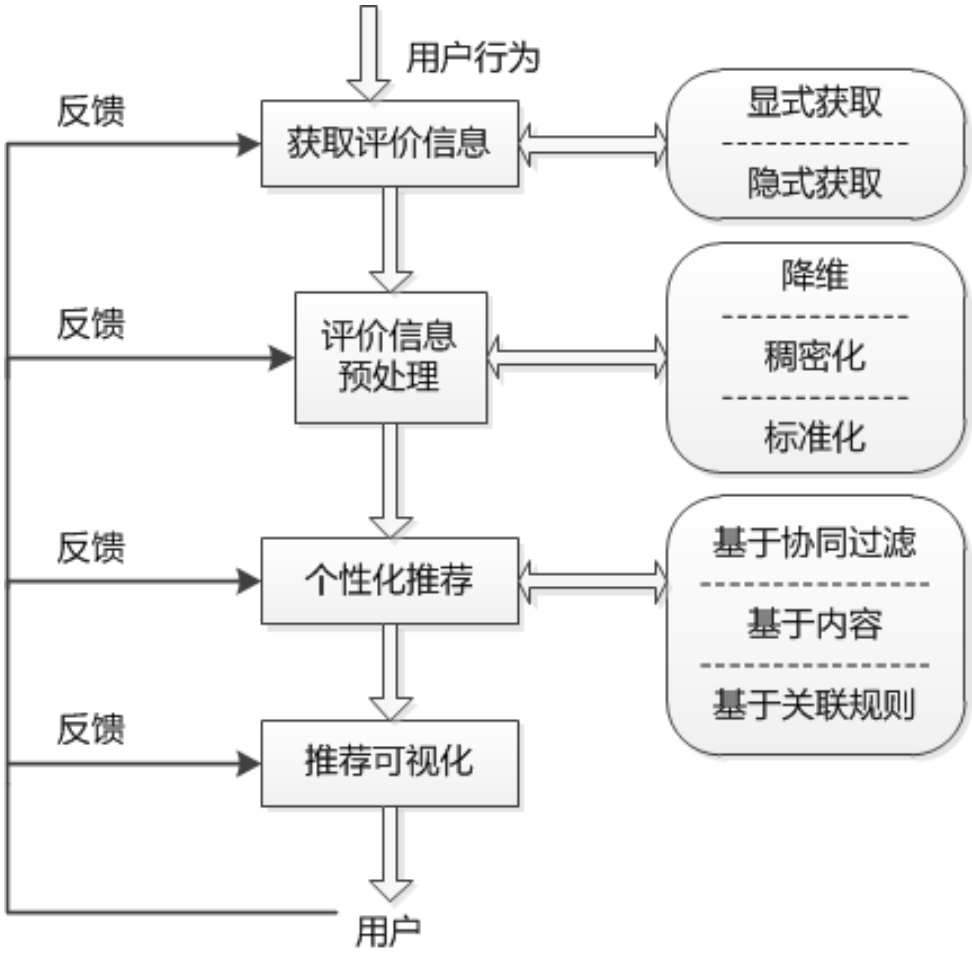
黑衣人

导演: 巴里·索南菲尔德

★★★☆☆ 7.8

地球并不只是人类的天下，其实有1500名外星人生活中我们当中，而星际移民局则处理和外星人相关的事情。

推荐系统-工作流程


用户行为
反馈
获取评价信息
显式获取
隐式获取
反馈
评价信息
预处理
降维
稠密化
标准化
反馈
个性化推荐
基于协同过滤
基于内容
基于关联规则
反馈
推荐可视化
用户

实验用数据集

- MovieLens数据集（来自Minnesota大学GroupLens Research项目组）

- 数据集描述：
- 1. 包含943个用户对1682部电影共10万条的评分记录；
- 2. 每个用户至少评价了20部电影；
- 3. 提供少量关于项目（电影）的信息。

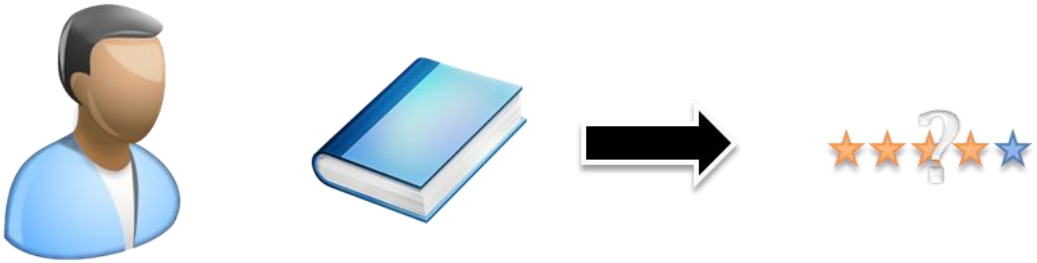
- 将实验数据随机划分为训练集和测试集两个部分，其中训练集占全部数据的80%，测试集占20%。

实验任务（推荐系统主要任务）

- 1. 预测评分
- - Input

user	item	rating
A	a	★★★★½
B	a	★★★★★
B	b	★½☆☆☆
…	…	…

- - Output



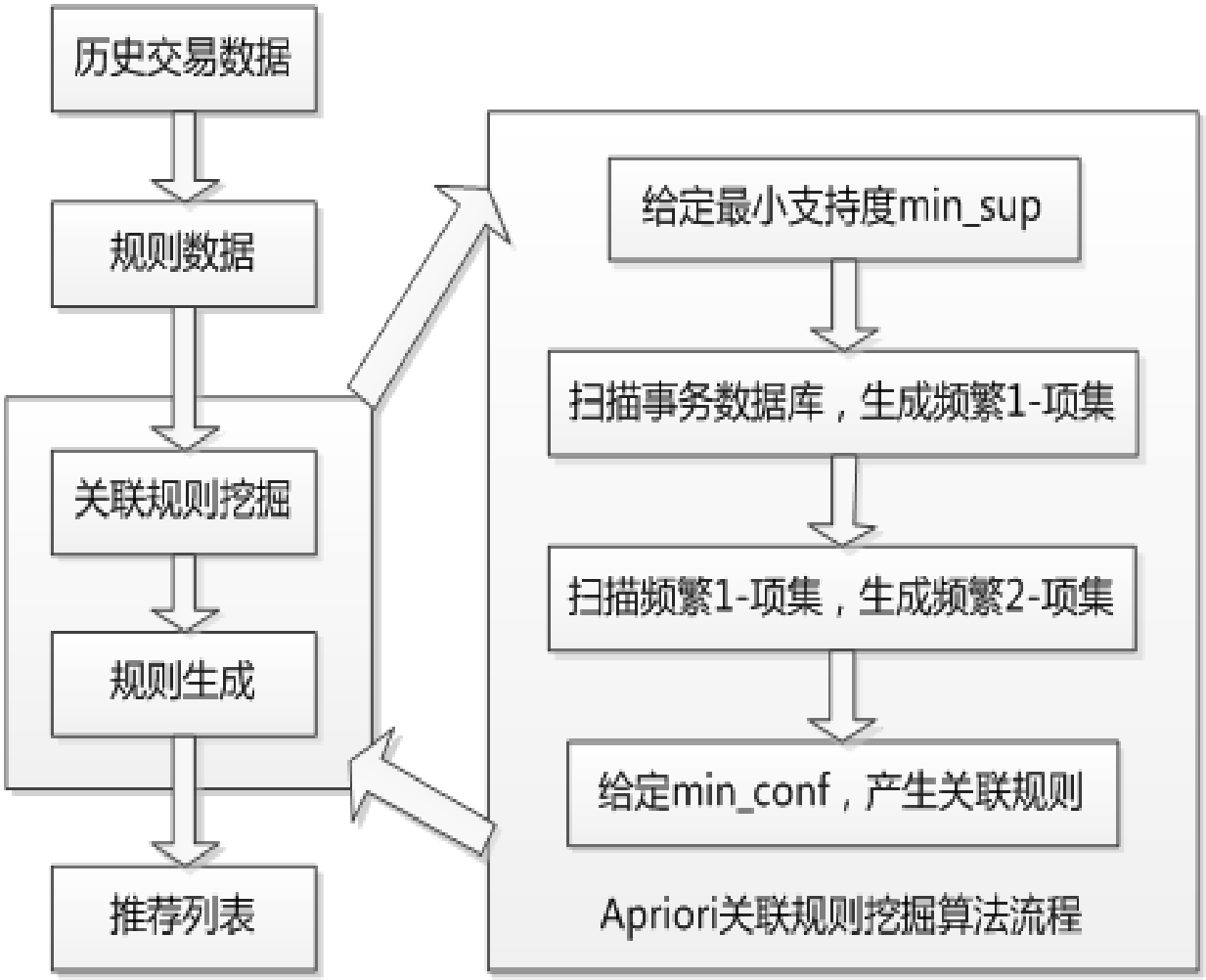
实验任务（推荐系统主要任务）

- 2. Top-N 推荐
- - Input

user	item
A	a
B	a
B	b
...	...

- - Output

推荐算法-关联规则

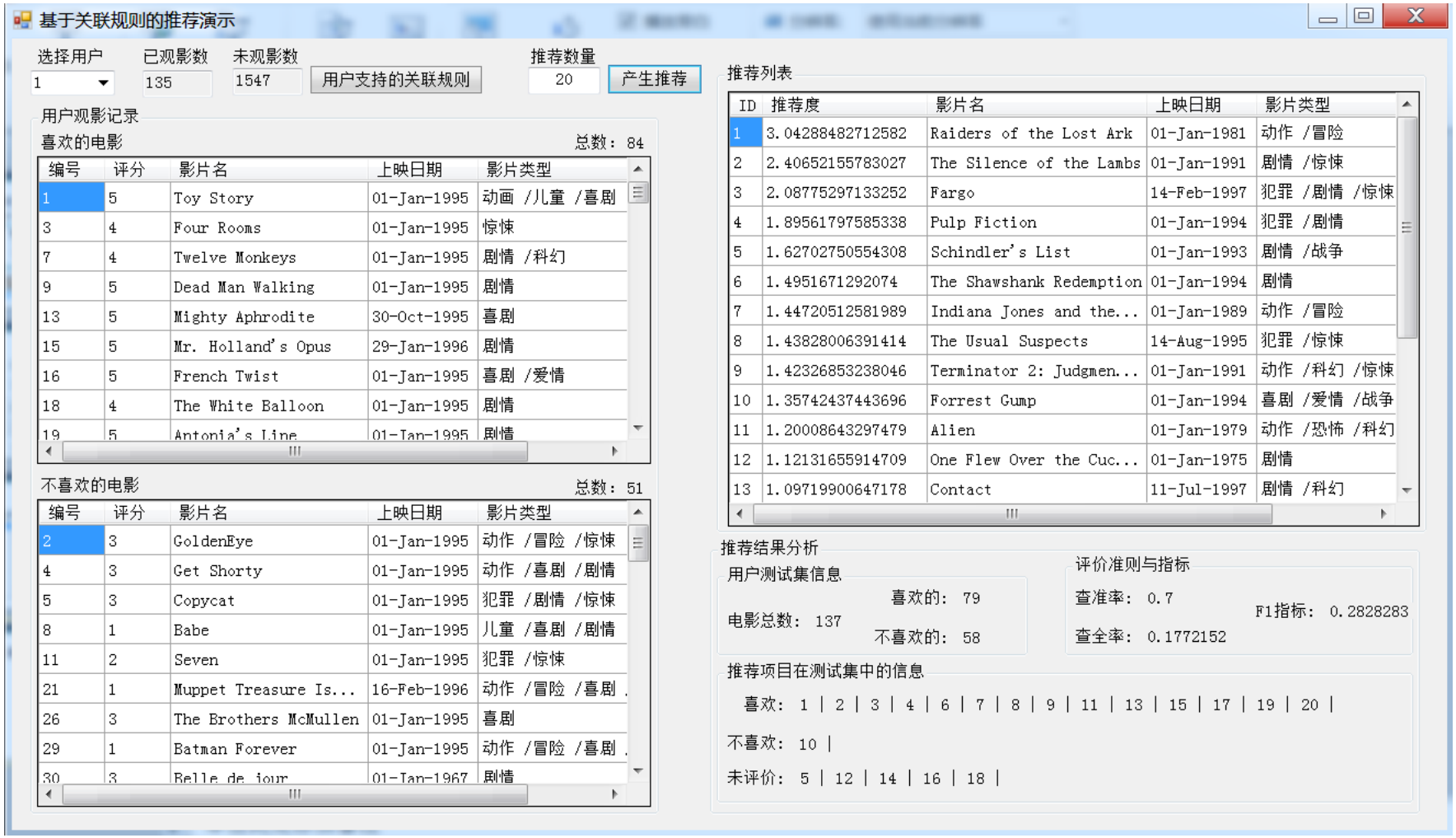

历史交易数据
规则数据
关联规则挖掘
规则生成
推荐列表
给定最小支持度min_sup
扫描事务数据库，生成频繁1-项集
扫描频繁1-项集，生成频繁2-项集
给定min_conf，产生关联规则
Apriori关联规则挖掘算法流程

推荐算法-关联规则

关联规则库

规则编号(ID)	规则先导(Antecedent)	规则后继(Consequent)	支持度(Support)	置信度(Confidence)
1	Star Wars	Return of the Jedi	0.2627119	0.6138614
2	Return of the Jedi	Star Wars	0.2627119	0.8184819
3	Star Wars	Raiders of the Lost Ark	0.217161	0.5074257
4	Raiders of the Lost Ark	Star Wars	0.217161	0.7269503
5	Star Wars	The Empire Strikes Back	0.2065678	0.4826733
6	The Empire Strikes Back	Star Wars	0.2065678	0.8333333
7	Star Wars	The Godfather	0.1980932	0.4628713
8	The Godfather	Star Wars	0.1980932	0.6726618
9	Star Wars	Fargo	0.1853814	0.4331683
10	Fargo	Star Wars	0.1853814	0.5663431
11	Toy Story	Star Wars	0.1800848	0.6439394
12	Star Wars	Toy Story	0.1800848	0.4207921
13	The Empire Strikes Back	Raiders of the Lost Ark	0.1769068	0.7136752
14	Raiders of the Lost Ark	The Empire Strikes Back	0.1769068	0.5921986
15	Star Wars	The Silence of the Lambs	0.1737288	0.4059406

基于关联规则的推荐演示



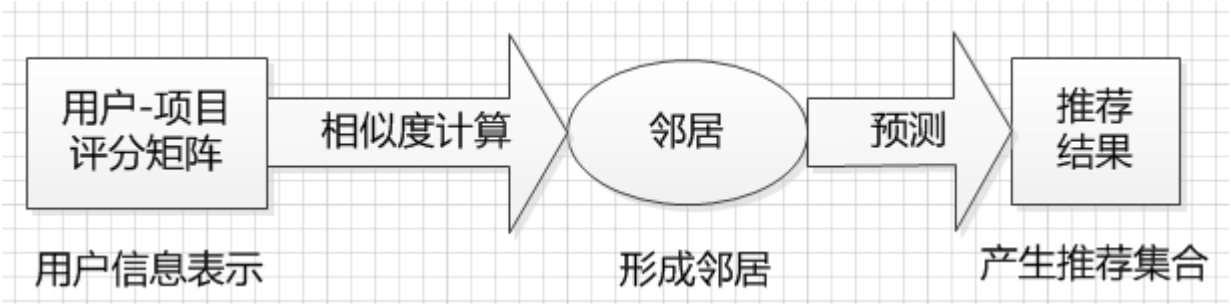
推荐算法-协同过滤算法

- 协同过滤核心思想：用户的兴趣偏好是可以通过对具有类似行为或偏好的用户群进行分析和预测得出的，强调人与人之间的协作。

- 其基于这样的一个假设：如果一组用户对一些项目的评分比较相似，则他们对其它项目的评分也比较相似；如果大部分用户对一些项的评分比较相似，则当前用户对这些项的评分也比较相似。

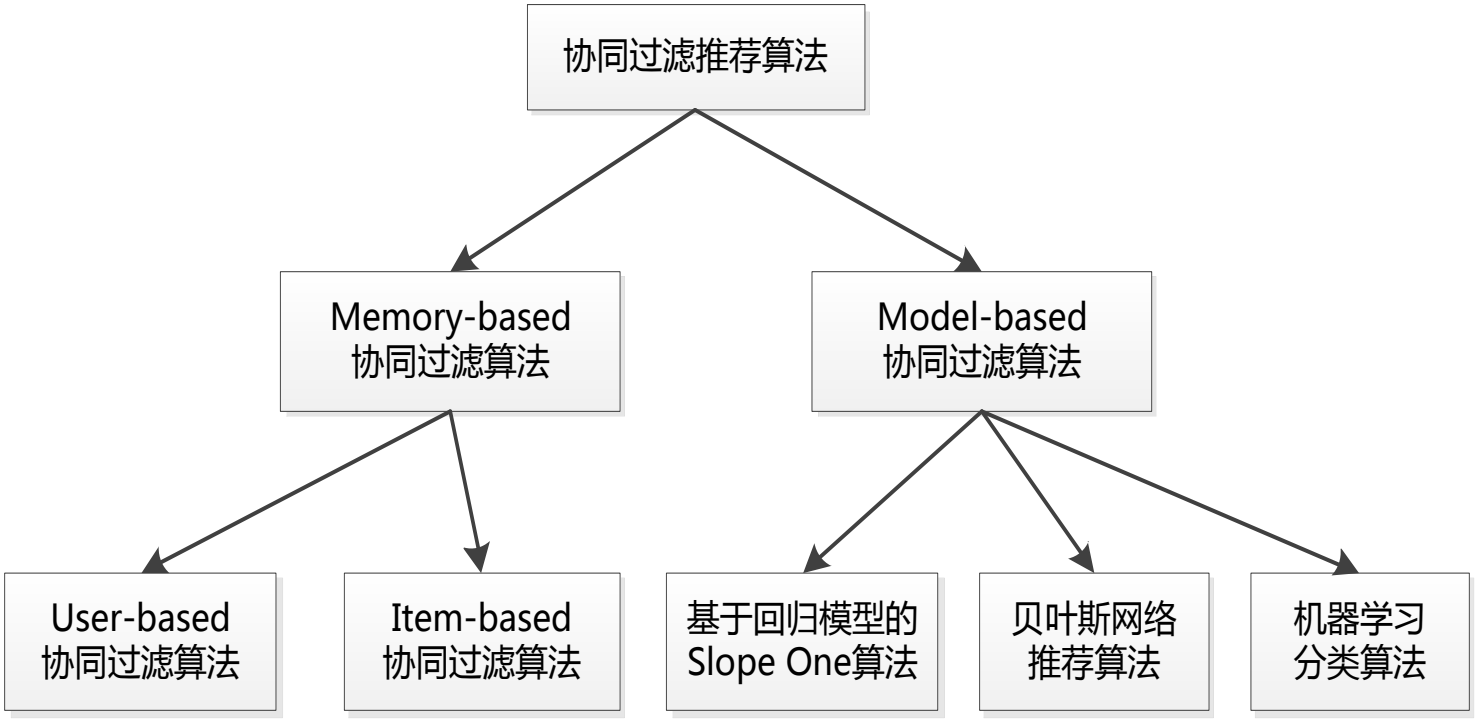
推荐算法-协同过滤算法

- 协同过滤算法的实现步骤：



推荐算法-协同过滤算法

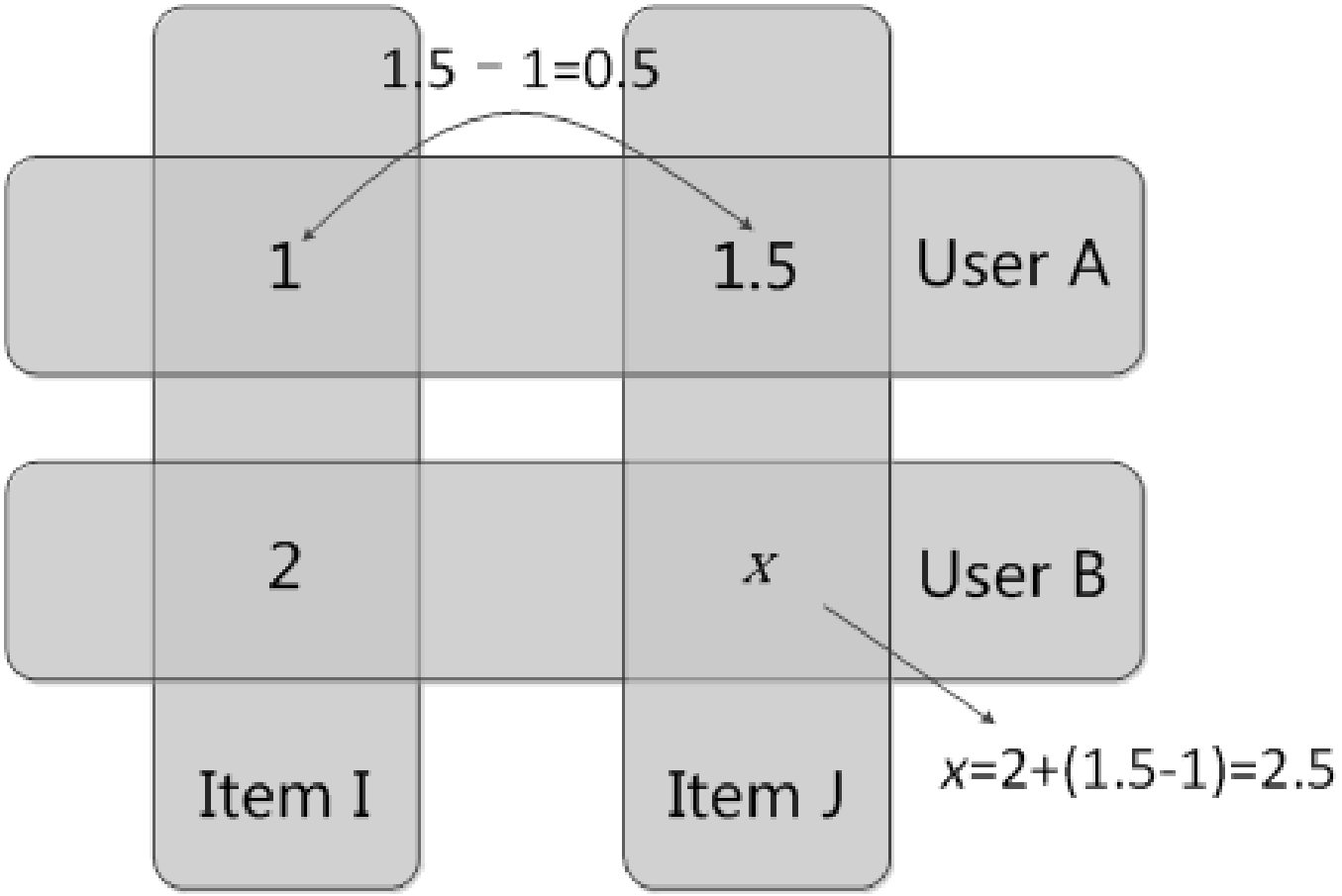
- 协同过滤算法的分类：



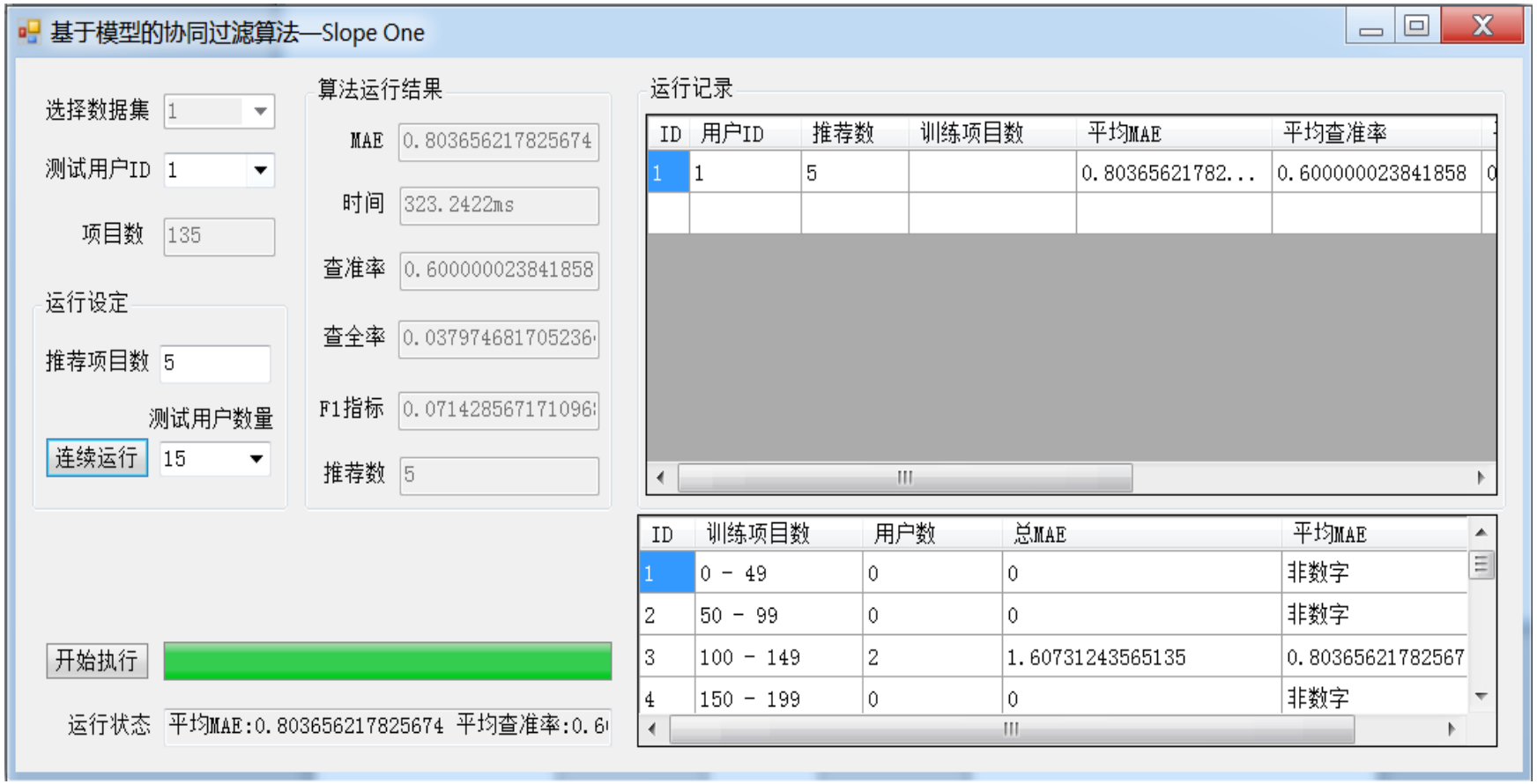
基于回归模型的Slope One算法

- 基本思想：设用户u对某两个项目的评分为 x, y，该算法假设所有用户对这两个项目的评分x与y之间符合线性关系 $y = x + b$。通过已经对这两个项目评过分的大量用户评分数据拟合该线性函数获得参数 b 的估计值 $\hat{b}$，最后将目标用户已有的项目评分 x 代入拟合公式 $y = x + \hat{b}$，从而获得 u 对项目的评分估计值$\hat{y}$。

基于回归模型的Slope One算法



基于回归模型的Slope One算法



推荐算法 - 评价指标

1. 预测评分评价指标 – MAE（平均绝对误差，Mean Absolute Error）

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

其中，p_i 是预测值，q_i 是实际评分，共有n个预测项目。

推荐算法 - 评价指标

2. Top-N 推荐评价指标

Top-N 推荐结果：

	推荐	未推荐
访问	推荐-访问（tp）	未推荐-访问（fn）
未访问	推荐-未访问（fp）	未推荐-未访问（tn）

推荐算法 - 评价指标

（1）查准率（准确率，Precision）

$$Precision = \frac{tp}{tp+fp}$$

表示推荐命中物品数量与推荐物品总数之比。

（2）查全率（召回率，覆盖率，Recall）

$$Recall = \frac{tp}{tp+fn}$$

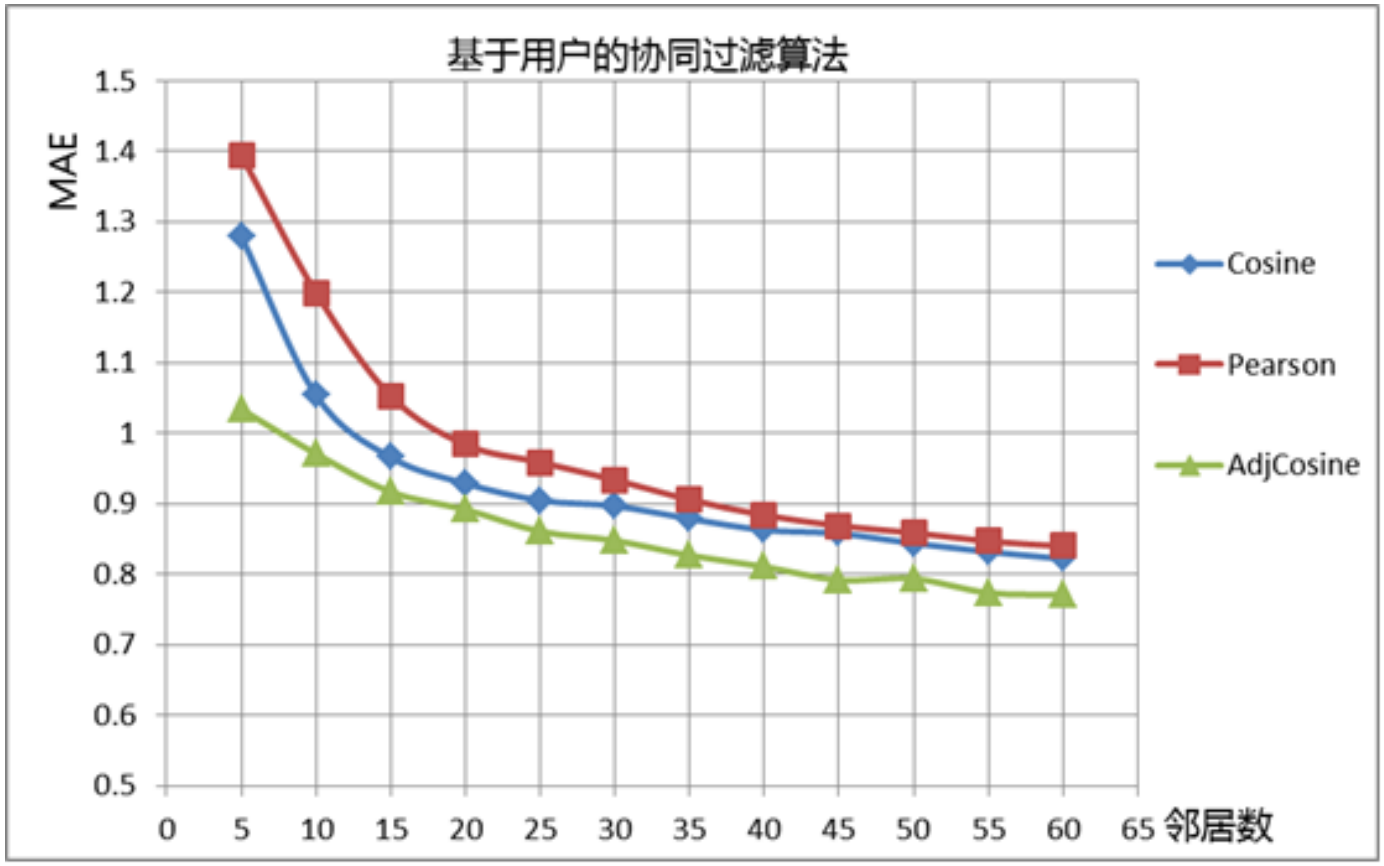
表示推荐命中物品数量与测试集中用户所访问物品总数之比。

（3）F1值

$$\mathrm{F1} = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

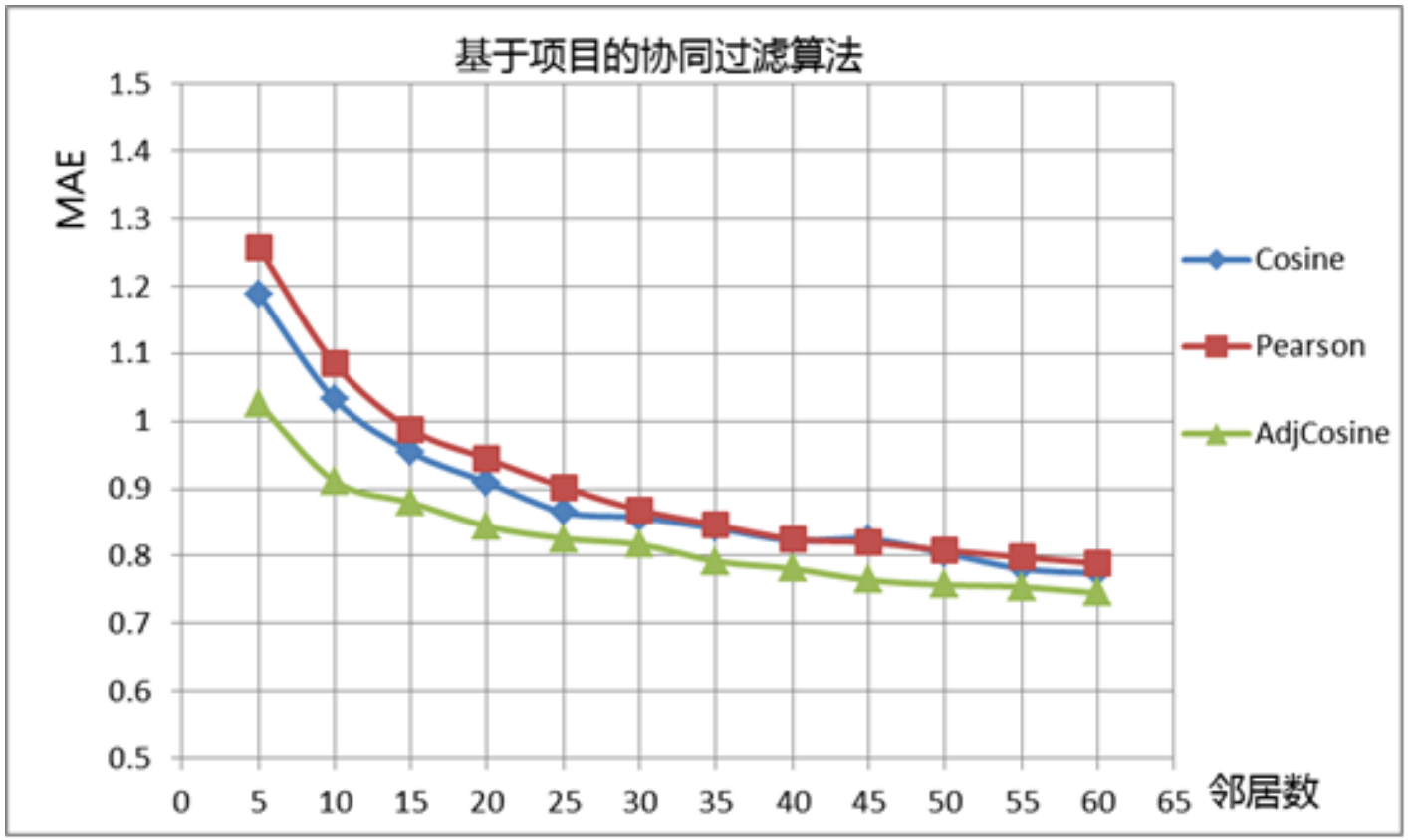
推荐算法 – 实验结果

基于用户的协同过滤（UBCF）



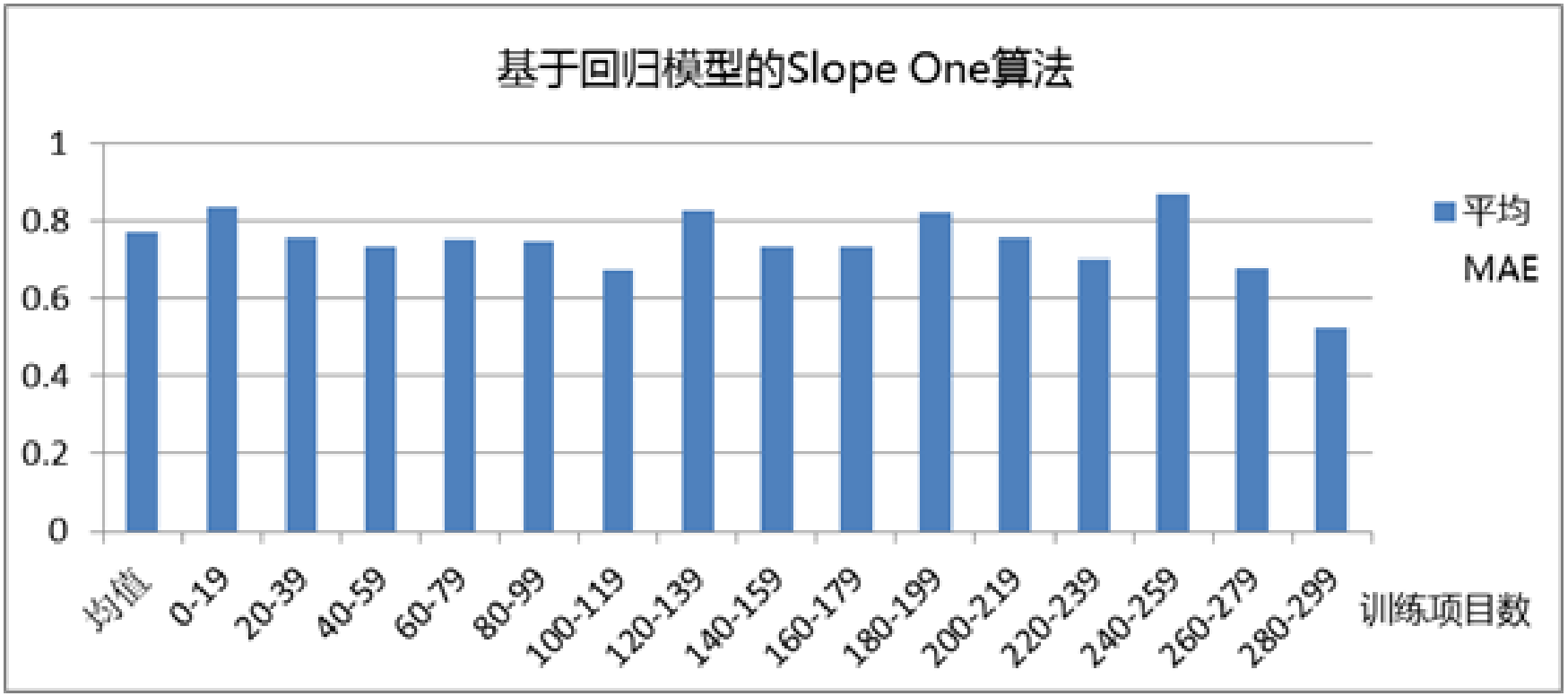
推荐算法 – 实验结果

基于项目的协同过滤（IBCF）



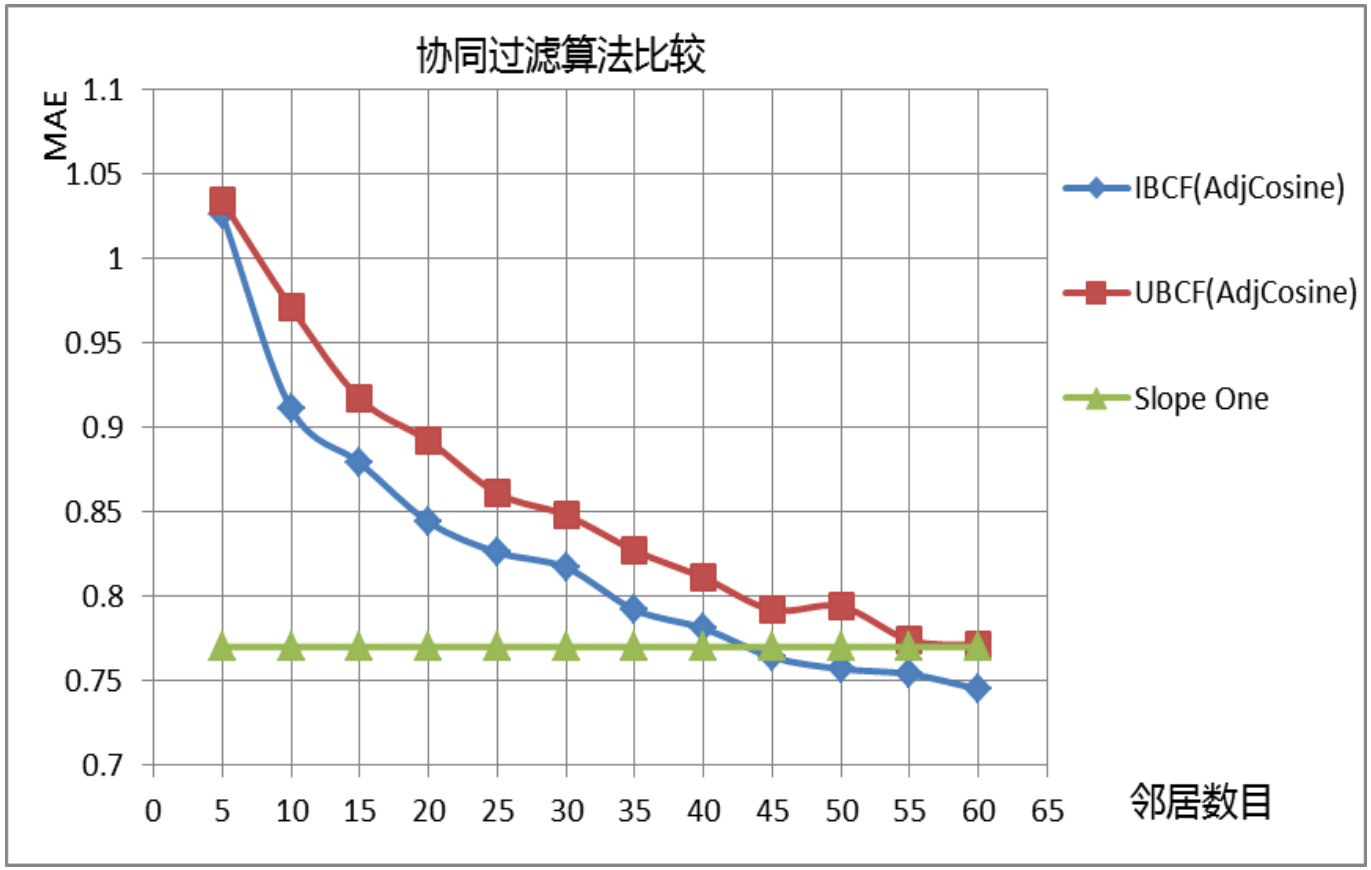
推荐算法 – 实验结果

Slope One 算法



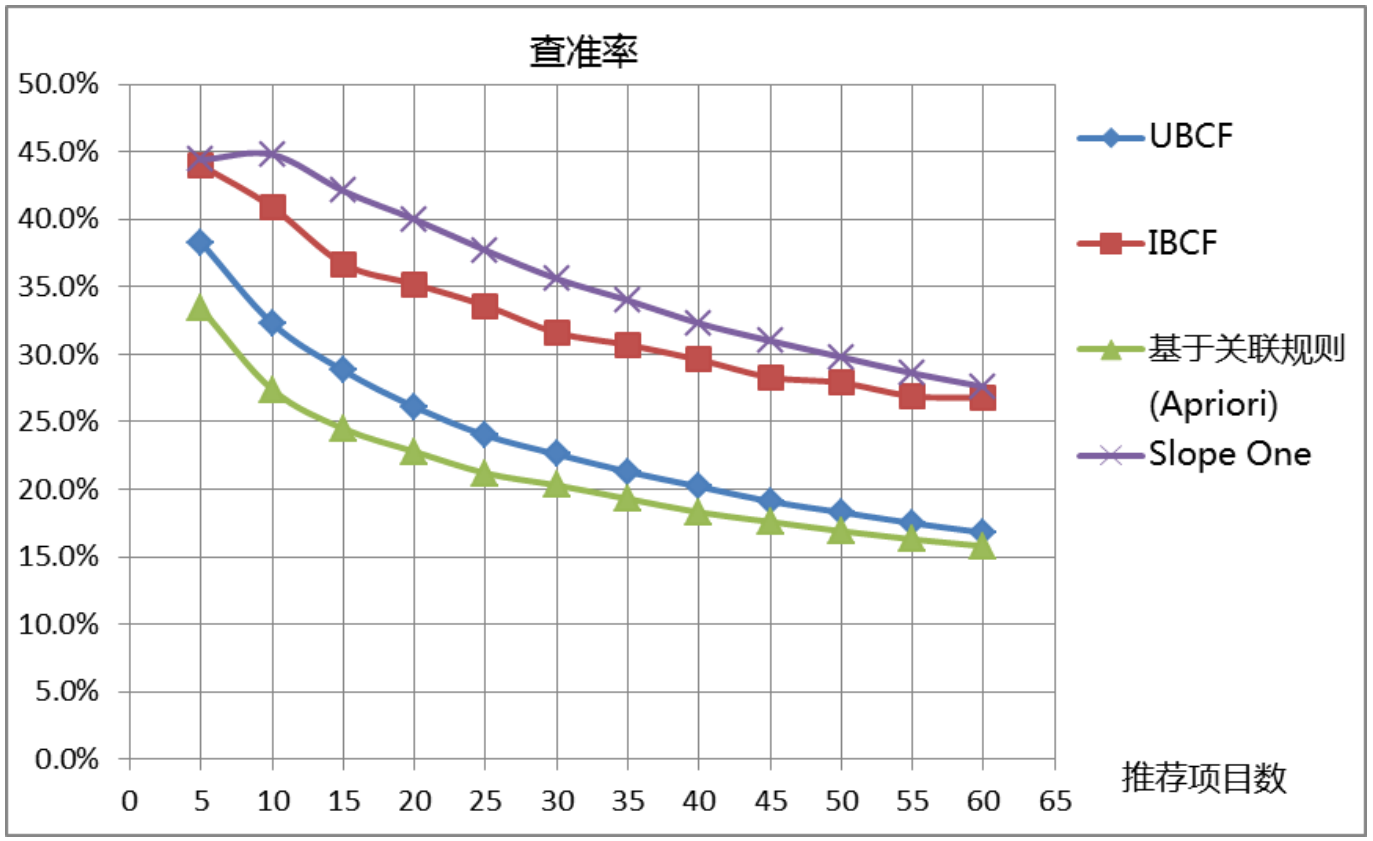
推荐算法 – 实验结果

协同过滤算法比较



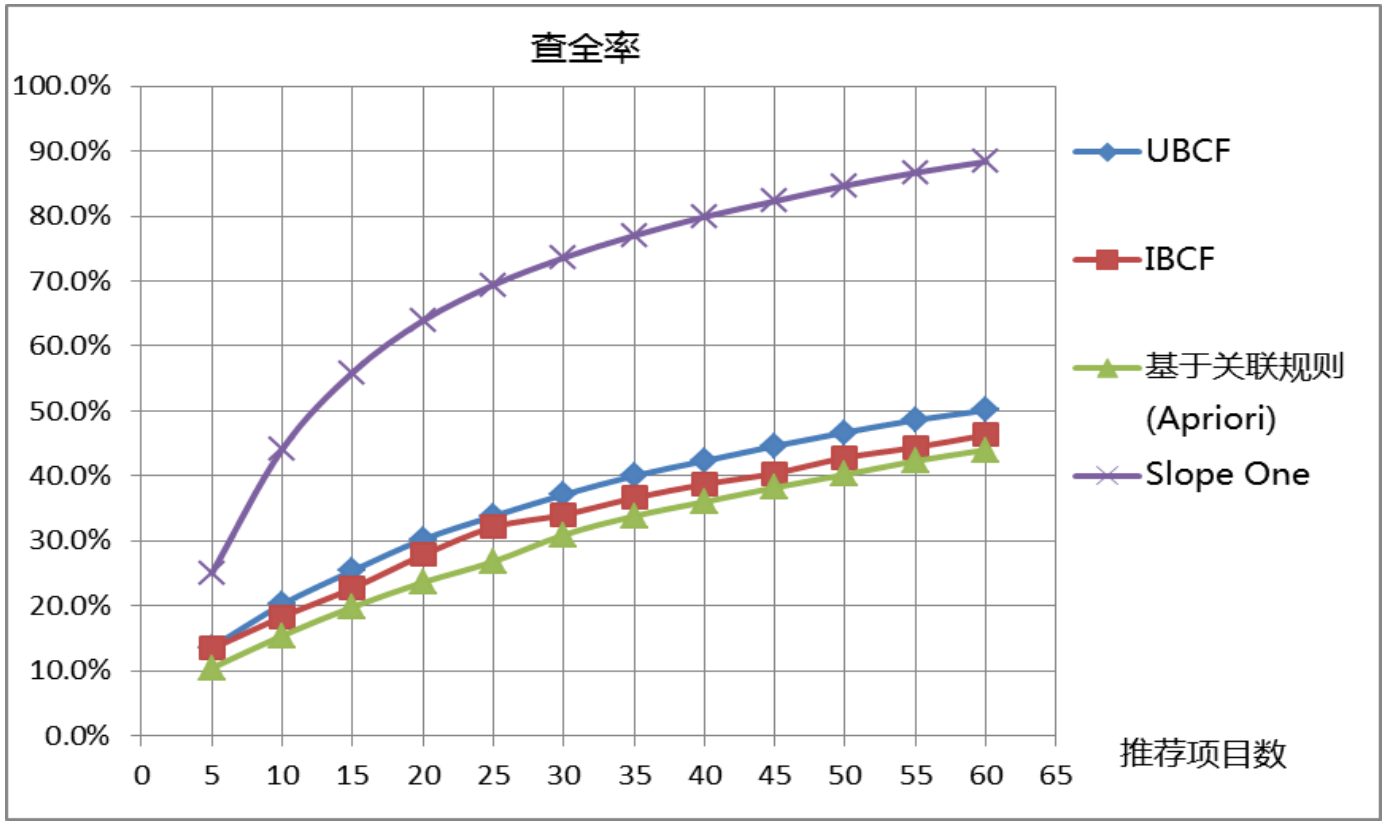
推荐算法 – 实验结果

各算法的Top-N推荐实验结果 – 查准率（ Precision ）



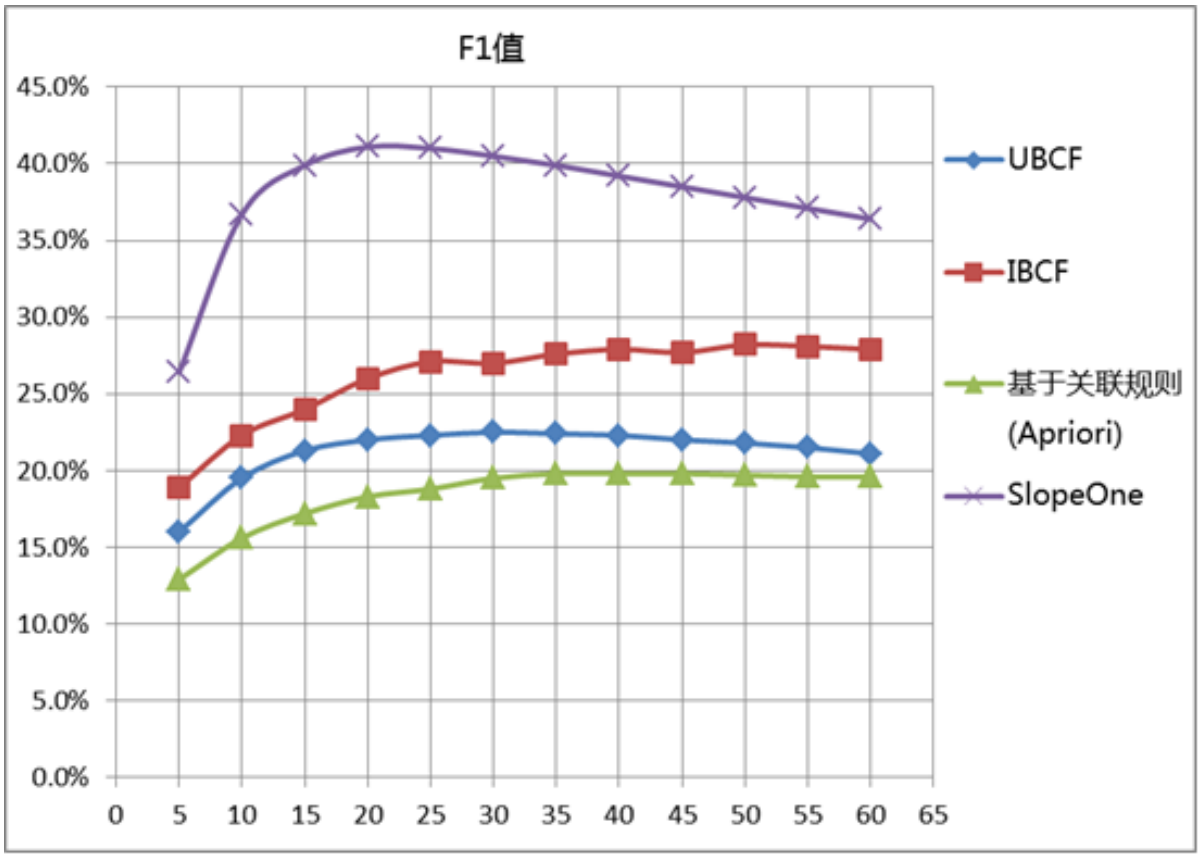
推荐算法 – 实验结果

各算法的Top-N推荐实验结果 – 查全率（ Recall ）



推荐算法 – 实验结果

各算法的Top-N推荐实验结果 – F1值



推荐算法 – 实验结果

从以上实验结果，我们可以得到：

1. 对于Top-N推荐而言，查全率和查准率是一个矛盾的指标，随着推荐项目数即N值的增大，查准率逐渐减小，而查全率上升。

2. 综合度量查全率和查准率的指标——F1值会随着推荐项目的增大而逐渐保持稳定。

3. 基于线性回归模型的Slope one算法综合来看，其Top-N推荐的性能最好。基于内存的两个协同过滤算法次之，最差的是基于关联规则（Apriori）的推荐算法。

4. Slope One算法对于历史评分项（即训练集合项目数）比较少的用户也能产生较好的预测评分精度，拥有不同历史评分项目数的用户之间的MAE值相差很小。

推荐算法 – 展望

1. 多种推荐算法的融合 —— 混合推荐；
2. 数据挖掘、人工智能等领域的技术、算法应用于推荐问题；
3. 海量数据的处理：如何在海量的数据中作出实时性的推荐；
4. 推荐系统应用领域的扩大化：只要是存在选择的地方就有推荐问题存在的意义，推荐系统完全可以应用于市场分析、客户关系、销售决策等领域；
5. 推荐系统的安全问题。

谢谢各位老师