大数据编程

8-1

个性化推荐算法

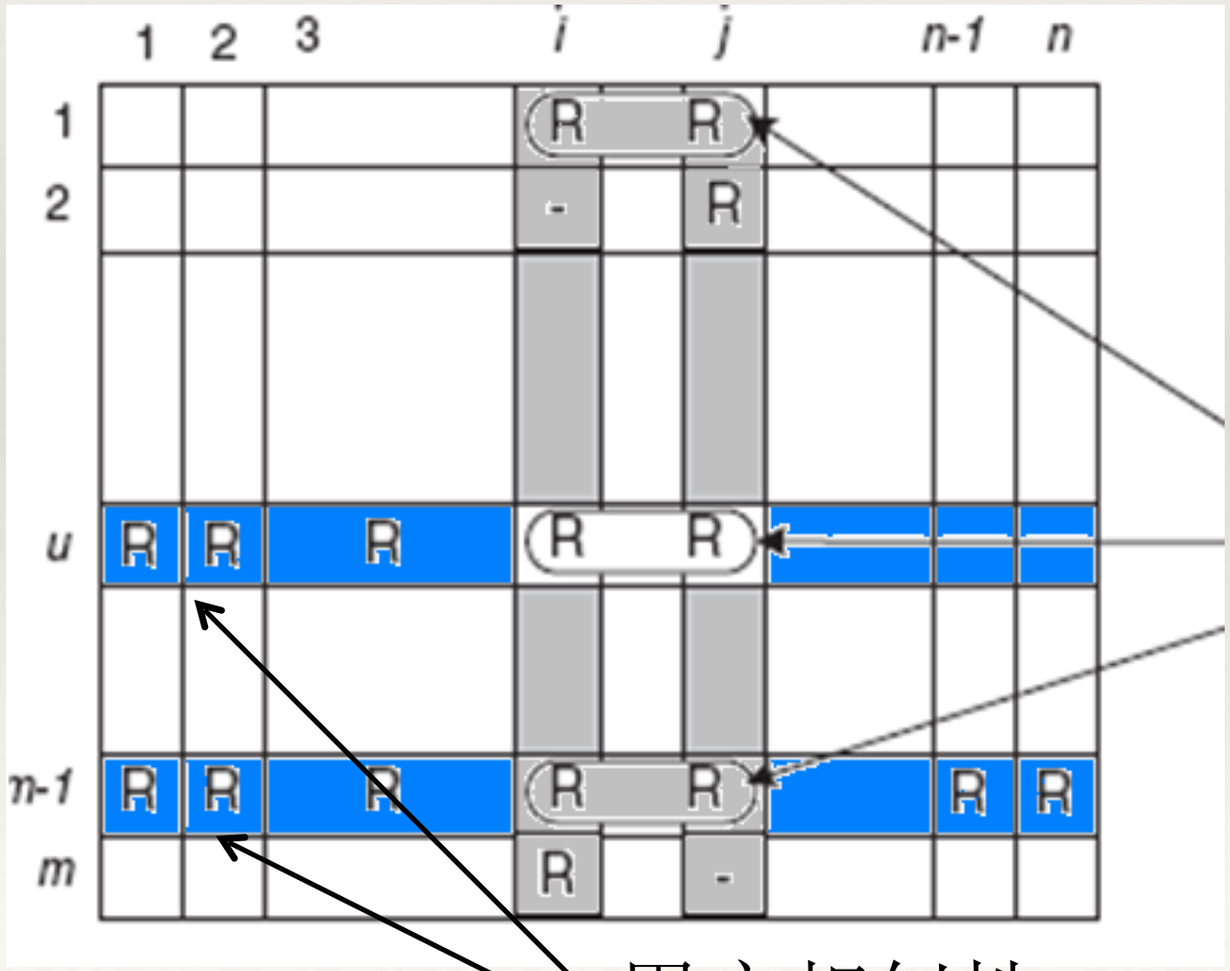
中央财经大学 商学院
姚凯
2016

主要内容

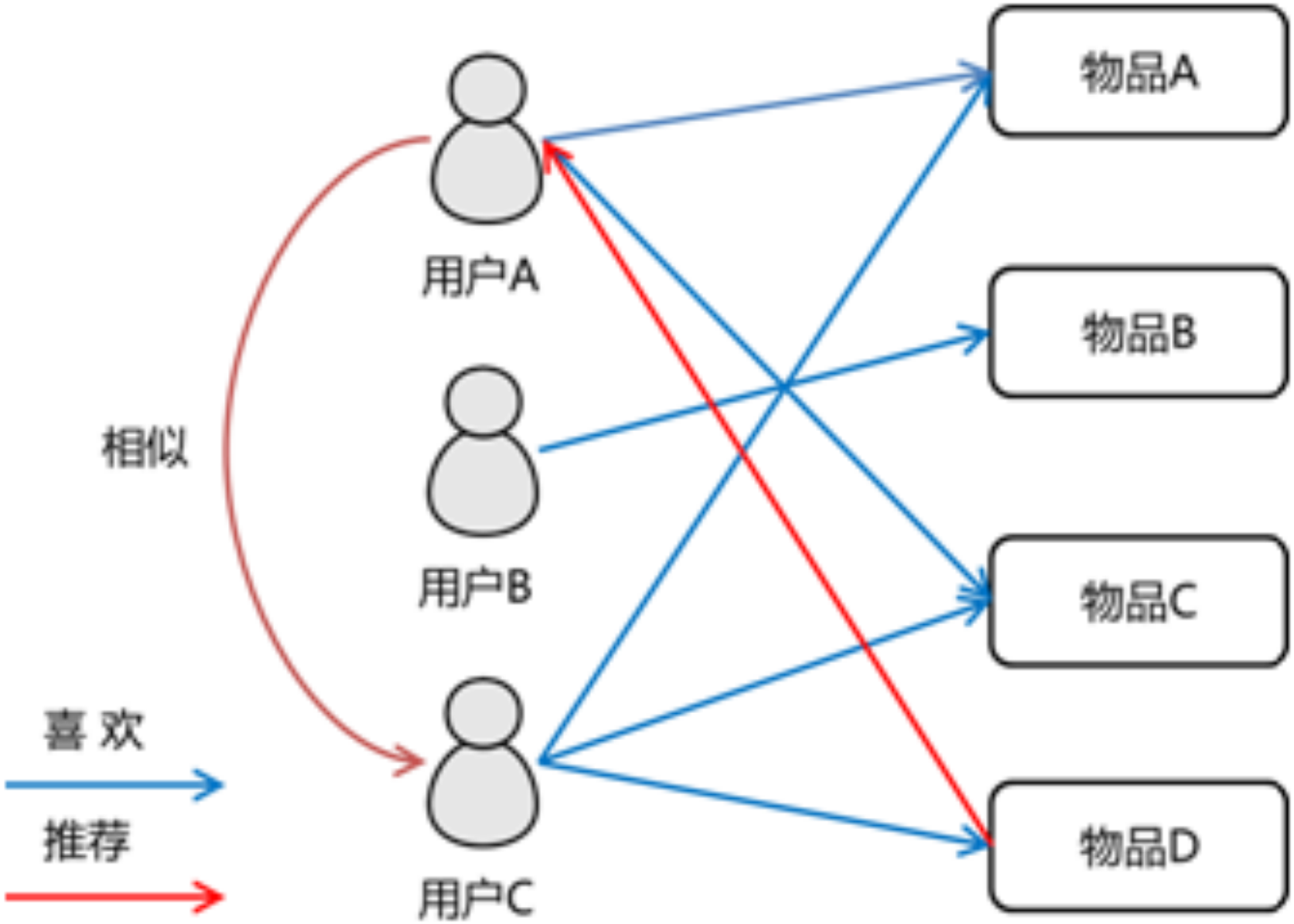
- 基于用户的协同过滤推荐算法实现
- 基于商品的协同过滤推荐算法实现
- 两种推荐算法对比
- 评价推荐算法标准

协同过滤法: 集体智慧

- 使用目标消费者和其他消费者的历史数据
- 基于用户 vs. 基于产品
- 计算用户之间或产品之间的相似性



基于用户的协同过滤算法



实现步骤

1.构造用户商品矩阵

2.计算用户间相似度

3.根据用户间相似度推荐商品

构造用户商品矩阵

Consider the data sample:

CRITIC	TITANIC	BATMAN	INCEPTION	SUPERMAN RETURNS	SPIDERMAN	MATRIX
MICHEL	2.5	3.5	3	3.5	2.5	3
SATYA	3	3.5	1.5	5	3	3.5
PARANAV	2.5	3	N/A	3.5	N/A	4
SURESH	N/A	3.5	3	4	2.5	4.5
TOM	3	4	2	3	2	3
LEO	3	4	N/A	5	3.5	3
CHAN	N/A	4.5	N/A	4	1	N/A

计算用户间相似度

- 将用户看过的商品列表用向量表示
- 计算两个向量之间的相似度(欧式内积，皮尔森相关系数)

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

计算用户间相似度

- 实际应用中利用矩阵运算得到所有用户间相似度

```
user_sim = cosine(as.matrix(t(x)))
```

根据用户间相似度推荐商品

1.计算权重矩阵

User_sim for CHAN		TITANIC	INCEPTION	MATRIX		TITANIC	INCEPTION	MATRIX
0.7125006	*	2.5	3	3	==>	1.7812515	2.1375	2.1375
0.760215		3	1.5	3.5		2.280645	1.1403	2.66075
0.6831639		2.5	N/A	4		1.7079098	N/A	2.73266
0.7028414		N/A	3	4.5		N/A	2.1085	3.16279
0.7341787		3	2	3		2.2025361	1.4684	2.20254
0.80555		3	N/A	3		2.41665	N/A	2.41665
1		N/A	N/A	N/A		N/A	N/A	N/A

根据用户间相似度推荐商品

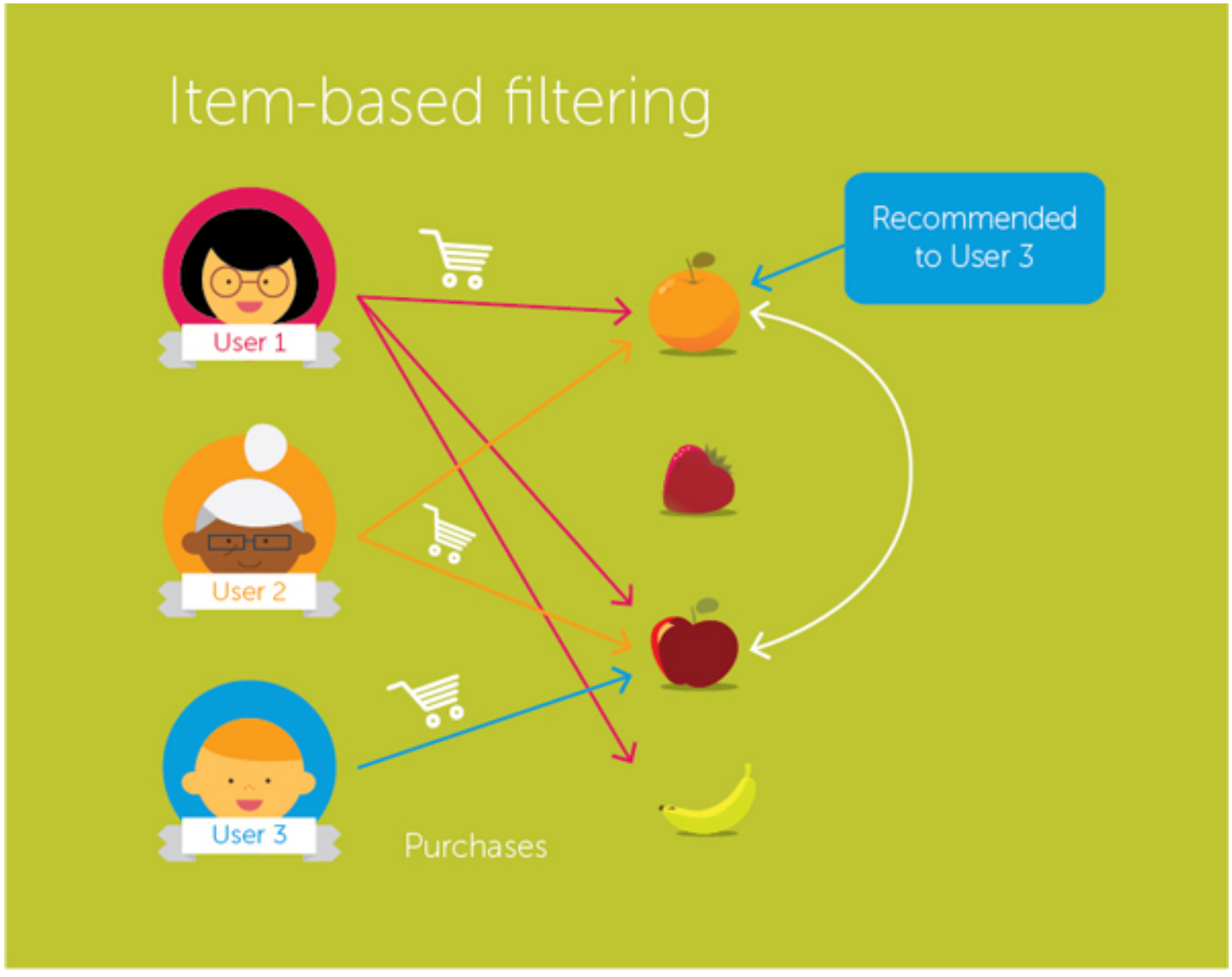
- 用每列权重和除以相似度和

		TITANIC	INCEPTION	MATRIX
sum of columns		10.38899235	6.854706	15.312882
Sum of Sim Users who have rated		3.6956082	2.909736	4.3984496
Divide		2.811172556	2.355783	3.4814273

User_sim for CHAN		TITANIC	INCEPTION	MATRIX		TITANIC	INCEPTION	MATRIX
0.7125006	*	2.5	3	3	==>	1.7812515	2.1375	2.1375
0.760215		3	1.5	3.5		2.280645	1.1403	2.66075
0.6831639		2.5	N/A	4		1.7079098	N/A	2.73266
0.7028414		N/A	3	4.5		N/A	2.1085	3.16279
0.7341787		3	2	3		2.2025361	1.4684	2.20254
0.80555		3	N/A	3		2.41665	N/A	2.41665
1		N/A	N/A	N/A		N/A	N/A	N/A

		TITANIC	INCEPTION	MATRIX
sum of columns		10.38899235	6.854706	15.312882
Sum of Sim Users who have rated		3.6956082	2.909736	4.3984496
Divide		2.811172556	2.355783	3.4814273

基于商品的协同过滤算法



基于商品的协同过滤算法实现

1.构造用户商品矩阵

2.计算商品间相似度

3.根据商品间相似度推荐与消费者已买商品相似的其他商品

实现步骤

1.构造用户商品矩阵

2.计算用户间相似度

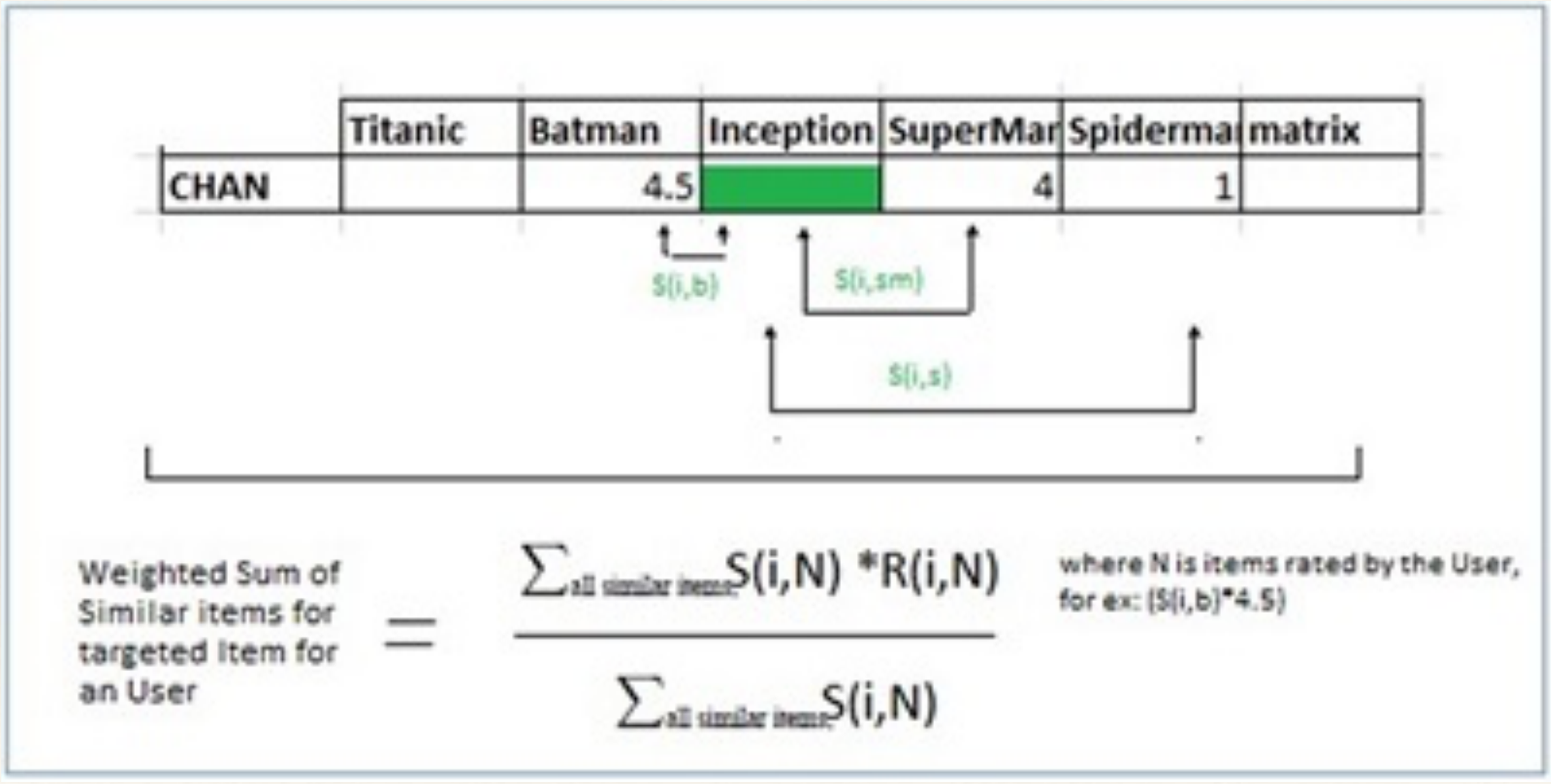
3.根据用户间相似度推荐商品

构造用户商品矩阵

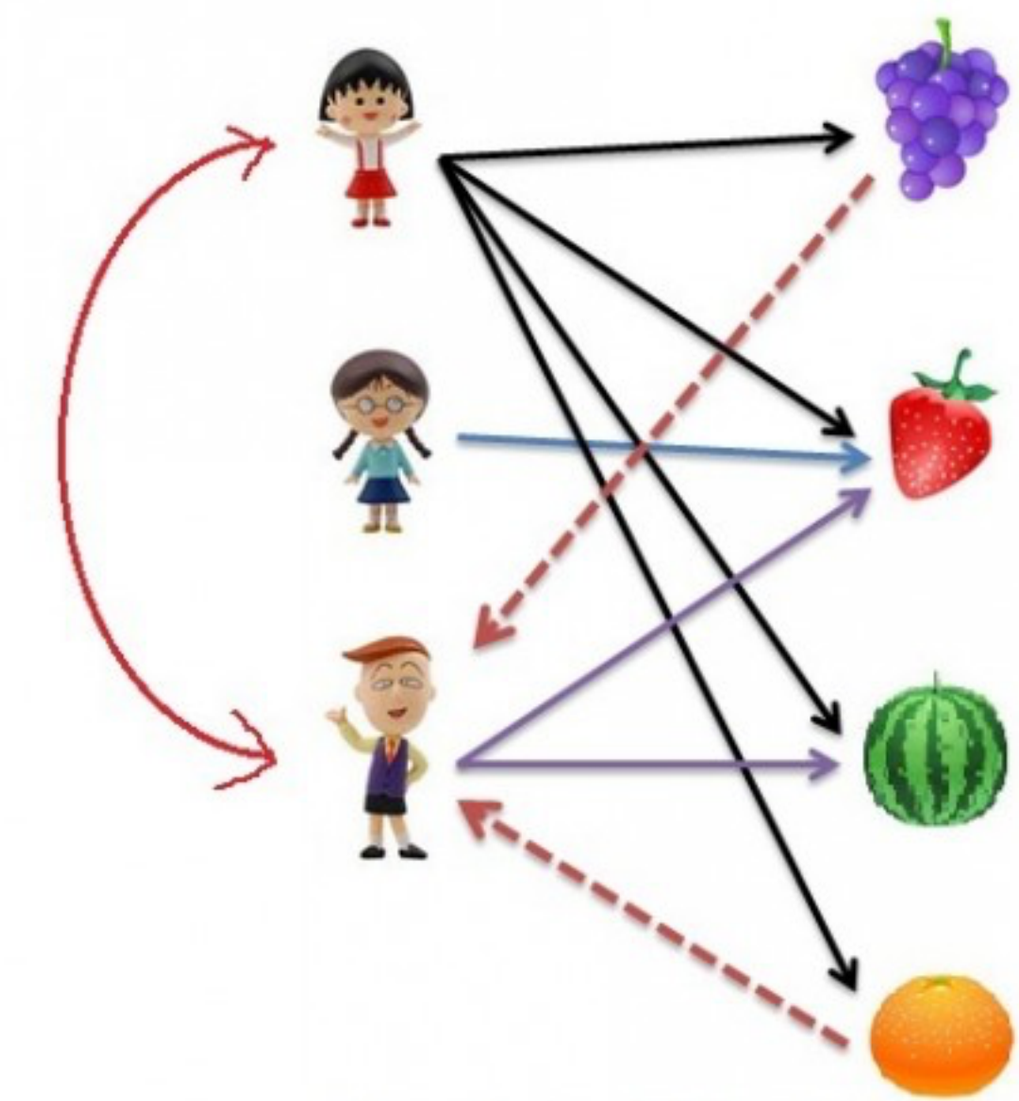
Consider the data sample:

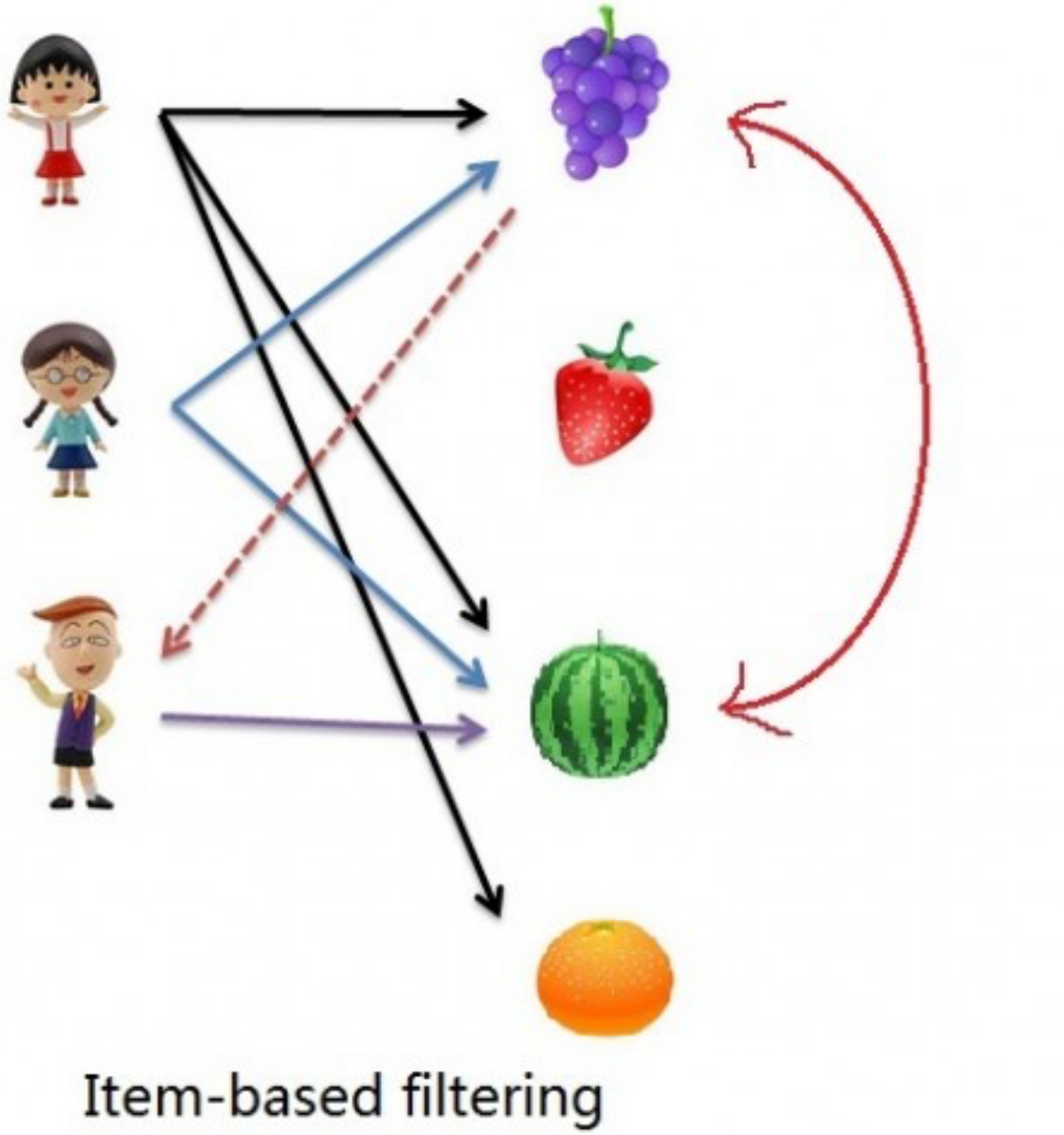
CRITIC	TITANIC	BATMAN	INCEPTION	SUPERMAN RETURNS	SPIDERMAN	MATRIX
MICHEL	2.5	3.5	3	3.5	2.5	3
SATYA	3	3.5	1.5	5	3	3.5
PARANAV	2.5	3	N/A	3.5	N/A	4
SURESH	N/A	3.5	3	4	2.5	4.5
TOM	3	4	2	3	2	3
LEO	3	4	N/A	5	3.5	3
CHAN	N/A	4.5	N/A	4	1	N/A

推荐算法

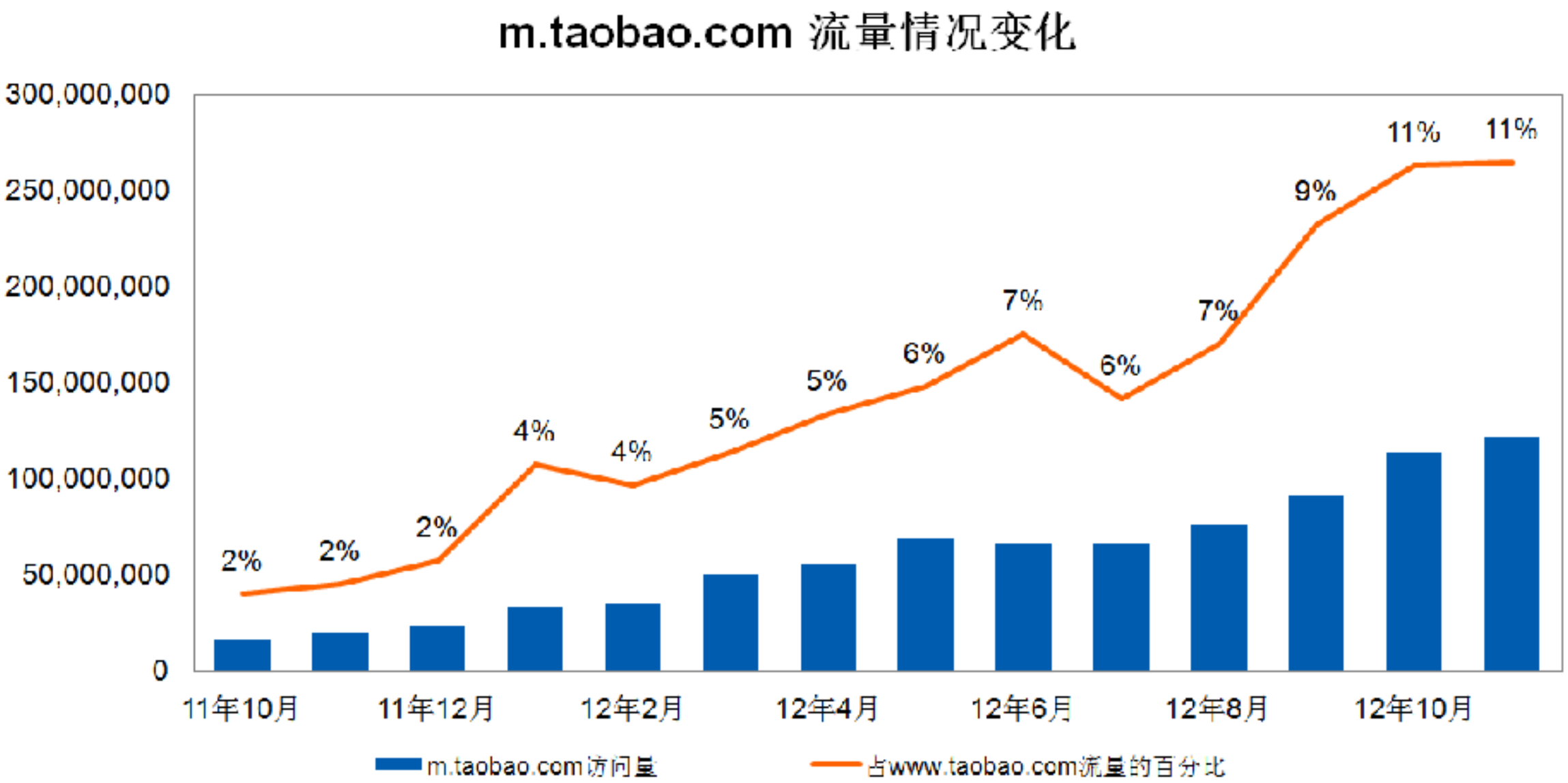

Titanic
Batman
Inception
SuperMar
Spiderma
matrix
CHAN
4.5
4
1
S(i,b)
S(i,sm)
S(i,s)
Weighted Sum of Similar items for targeted item for an User
=
$\sum_{\text{all similar items}} S(i,N) * R(i,N)$
$\sum_{\text{all similar items}} S(i,N)$
where N is items rated by the User, for ex: (S(i,b)*4.5)

协同过滤算法对比





手机/iPad购物成为淘宝增长点？
m.taobao.com访问量一年增6倍已超过淘宝流量的10%



注：m.taobao.com包含了来自于平板电脑(iPad等)浏览器和手机浏览器的http访问，不包含基于App客户端的访问方式

数据来源：Experian Hitwise 中国 2012年12月
http://www.experian.com.cn/hitwise

新浪微博
@Hitwise

改进协同过滤推荐系统的方法

1.根据用户场景进行过滤

2.优化协同过滤的用户商品矩阵

3.两种方法进行切换

4.采用混合推荐的方法

推荐算法指标

- 准确度
- 覆盖率
- 多样性
- 新颖性

如何检验推荐算法效果

- Offline：根据离线算法训练推荐系统，将训练好的算法应用在样本外数据上面看效果
- Online: 将不同的推荐算法部署在统一实际应用系统中，用户分组使用不同的推荐算法，实际测试推荐算法的好坏