

大数据编程

---

7-1

# 个性化推荐

中央财经大学 商学院  
姚凯  
2016

# PageRank思想

- ❖ PageRank是Google专有的算法，用于衡量特定网页相对于搜索引擎索引中的其他网页而言的重要程度。它由Larry Page 和 Sergey Brin在20世纪90年代后期发明。PageRank实现了将链接价值概念作为排名因素。
- ❖ PageRank让链接来“投票”

# 基本假设

- ❖ 数量假设：如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要
- ❖ 质量假设：指向页面A的入链质量不同，质量高的页面会通过链接向其他页面传递更多的权重。所以越是质量高的页面指向页面A，则页面A越重要。

# PageRank算法原理

- ❖ 网页的重要性排序是由网页间的链接关系所决定的，算法是依靠网页间的链接结构来评价每个页面的等级和重要性，一个网页的PR值不仅考虑指向它的链接网页数，还有指向'指向它的网页的其他网页本身的重要性。

$$PR(p_i) = \frac{1-d}{n} + d \sum_{p_j \in M(i)} \frac{PR(p_j)}{L(j)}$$

- PR(pi): pi页面的PageRank值
- n: 所有页面的数量
- pi: 不同的网页p1,p2,p3
- M(i): pi链入网页的集合
- L(j): pj链出网页的数量
- d: 阻尼系数, 任意时刻, 用户到达某页面后并继续向后浏览的概率。  
(1-d=0.15) : 表示用户停止点击, 随机跳到新URL的概率  
取值范围:  $0 < d \leq 1$ , Google设为0.85

# 例子

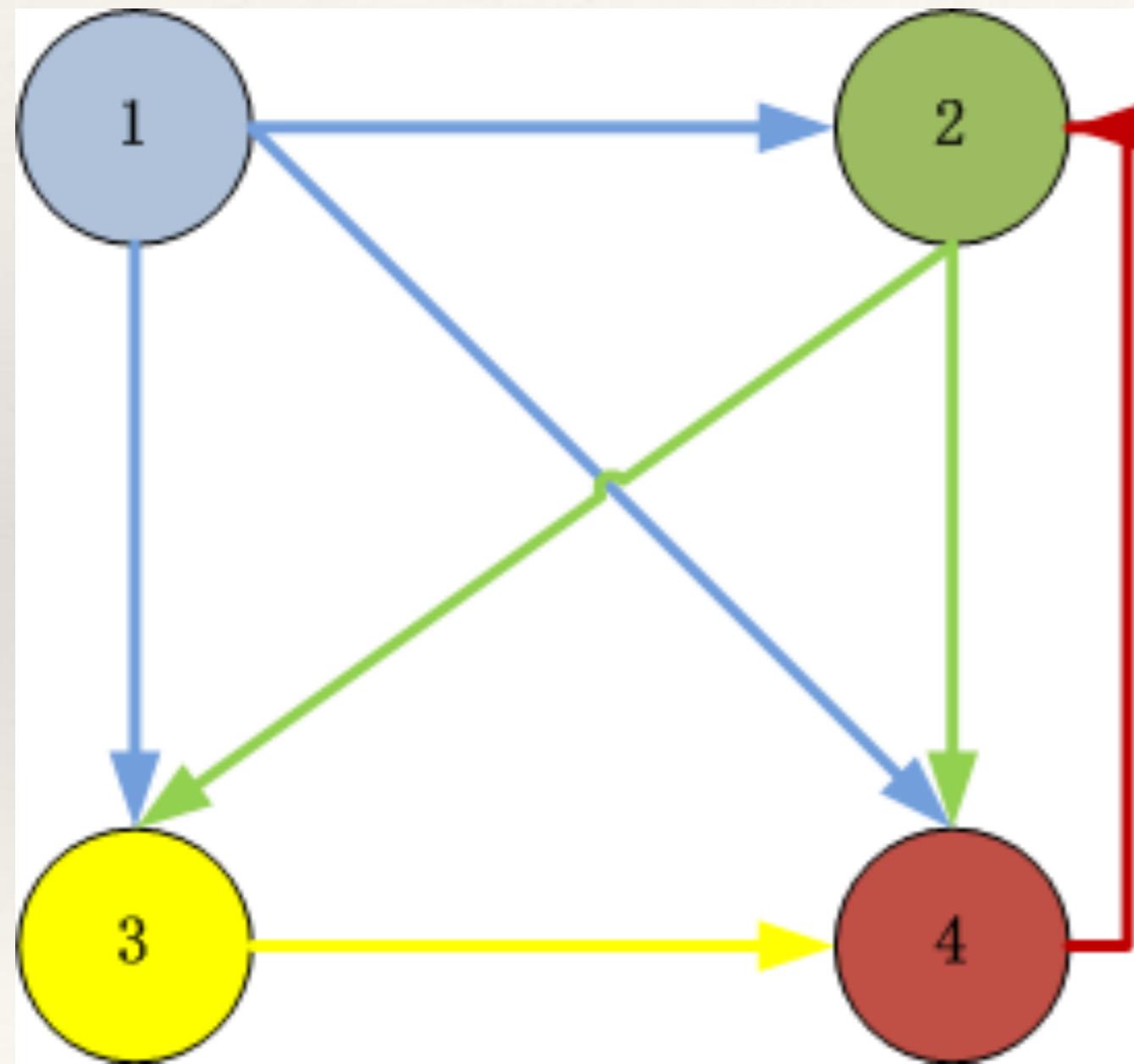
图片说明：

ID=1的页面链向2,3,4页面，所以一个用户从ID=1的页面跳转到2,3,4的概率各为 $1/3$

ID=2的页面链向3,4页面，所以一个用户从ID=2的页面跳转到3,4的概率各为 $1/2$

ID=3的页面链向4页面，所以一个用户从ID=3的页面跳转到4的概率各为1

ID=4的页面链向2页面，所以一个用户从ID=4的页面跳转到2的概率各为1



# 实现步骤

构造邻接矩阵(方阵):

列: 源页面

行: 目标页面

|      | [,1] | [,2] | [,3] | [,4] |
|------|------|------|------|------|
| [1,] | 0    | 0    | 0    | 0    |
| [2,] | 1    | 0    | 0    | 1    |
| [3,] | 1    | 1    | 0    | 0    |
| [4,] | 1    | 1    | 1    | 0    |

转换为概率矩阵(转移矩阵)

|      | [,1] | [,2] | [,3] | [,4] |
|------|------|------|------|------|
| [1,] | 0    | 0    | 0    | 0    |
| [2,] | 1/3  | 0    | 0    | 1    |
| [3,] | 1/3  | 1/2  | 0    | 0    |
| [4,] | 1/3  | 1/2  | 1    | 0    |

# 主要内容

- ❖ 个性化概述
- ❖ 个性化推荐算法
- ❖ 协同过滤推荐算法实现

# 个性化

个性化与数据市场是大数据精细化和融聚力的两个发展方向

- ❖ 用户信息饥饿感**与日俱增**
- ❖ 用户对非关联信息的容忍度**与日俱减**
- ❖ 用户兴趣数据**与日俱增**
- ❖ 用户甄别信息能力占比**与日俱减**

# 什么是个性化？

从大众营销到客户化定制到个性化营销

市场细分理论的终极目标

消费者个体层面的海量数据收集与挖掘

# 从大规模生产到客户化定制

## 戴尔的定制

Dell™ Contact Us Products Services Support Purchase Help

Cart | Sign In

Keyword Search

Dell recommends Windows Vista™ Home Premium.

You are here: USA > Home & Home Office

1 Build My Dell 2 Add My Accessories 3 Choose My Software 4 Protect My Investment 5 Review & Continue

► SWITCH TO LIST VIEW



Sample image only

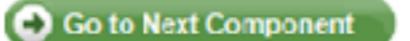
**SELECT MY PROCESSOR**

 Video Learn More

Intel® Core™2 Extreme X6800 (4MB,3.73GHz Factory overclocked)  
add \$0

Intel® Core™2 Extreme QX6700 (8MB,3.46GHz Factory overclocked)  
[Included in Price]

Intel® Core™2 Extreme QX6800 (8MB L2 Cache,3.73GHz Factory overcl  
[add \$100 or \$3/month<sup>1</sup>]  
**Dell Recommended**



 Processor

 Operating System

 Memory

 Hard Drive

 Additional Hard Drive

 Optical Drive

 Monitor

 Large Displays

**XPS 720 H2C**

**\$5,999** As low as \$180/month<sup>1</sup>

 Apply Now | Learn More

 Preliminary Ship Date: 6/21/2007<sup>3</sup>

 Print Summary

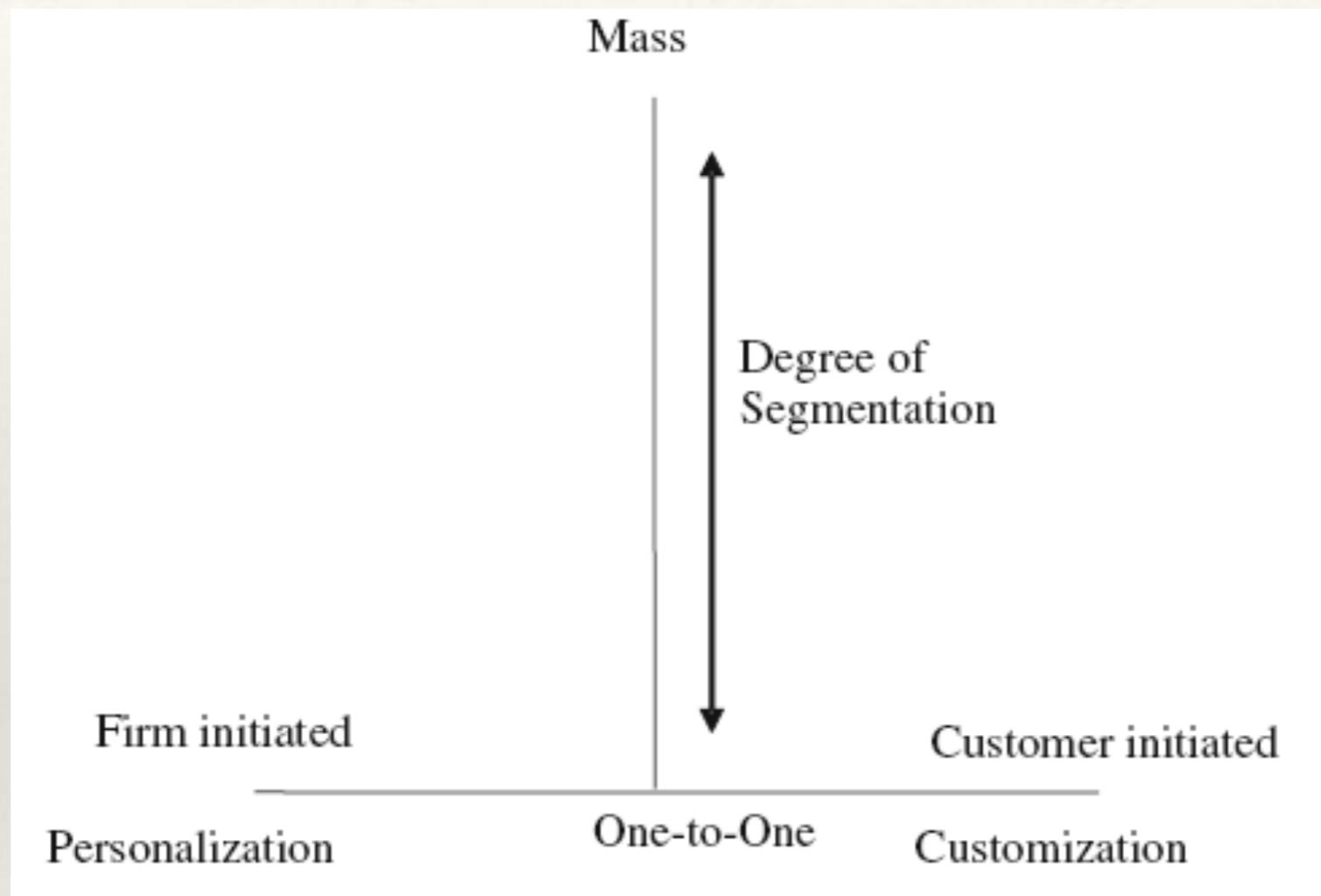
**My Components**

- Intel® Core™2 Extreme QX6700 (8MB,3.46GHz Factory overclocked)
- Genuine Windows® XP Media Center 2005 Edition with re-Installation CD
- 2GB Corsair Dominator DDR2 SDRAM 900MHz OC'd to 1066MHz 2 DIMMs
- 320GB Performance RAID 0 (2 x 160GB WD Raptor SATA 1.56Ms 10,000 RPM HDDs)
- Single Drive: Blu-ray Disc Drive (BD/DVD/CD burner w/double layer BD write)
- 24 inch UltraSharp™ 2407WFP Widescreen Digital Flat Panel
- Dual 768MB Nvidia GeForce 8800 GTX
- Sound Blaster® X-Fi™ XtremeMusic(D)
- Sound Card

**My Accessories**

# 个性化和客户化定制的区别？

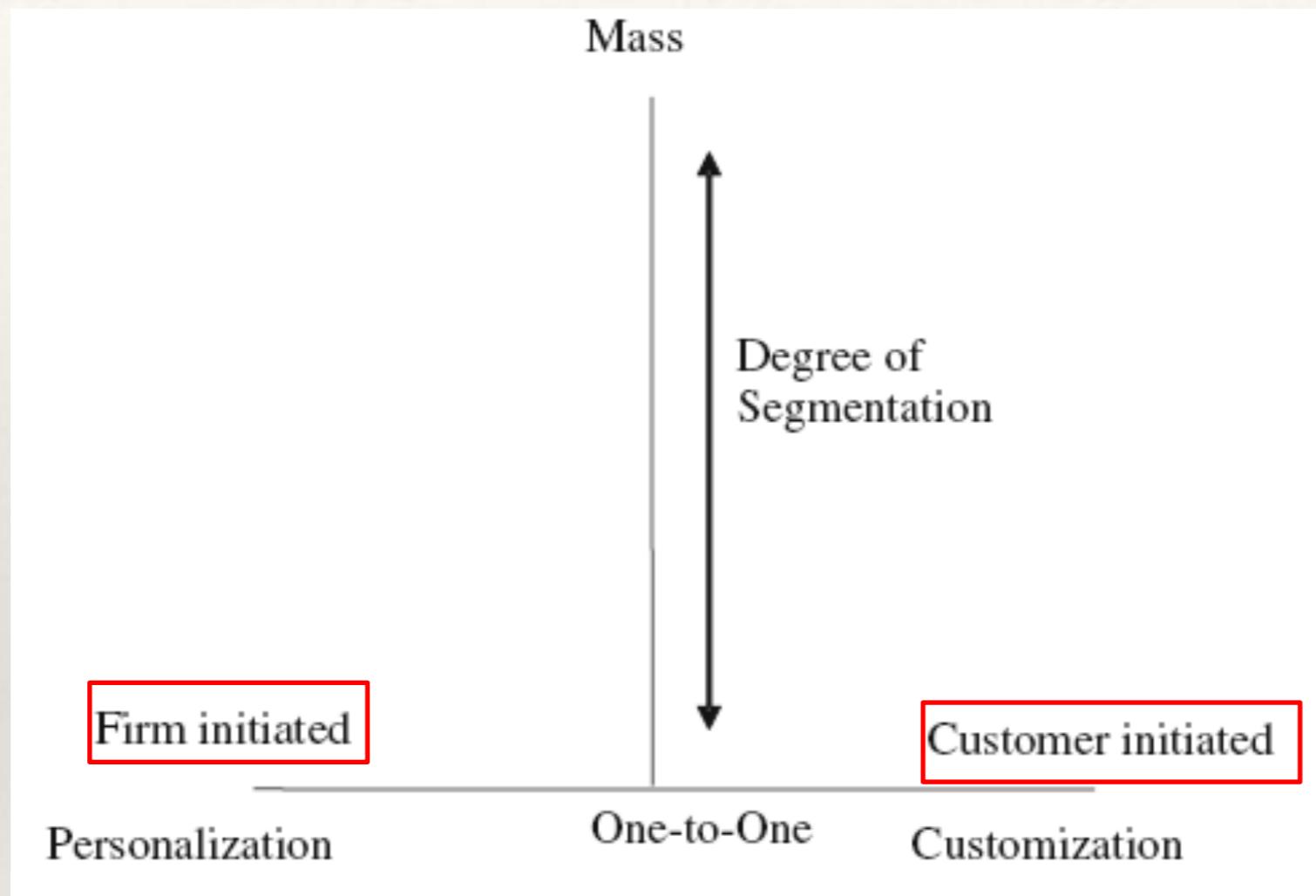
# 个性化和客户化定制的区别?



个性化：  
公司主导

客户化定制：  
客户主导

# 个性化和客户化定制的区别?



个性化：  
公司主导

客户化定制：  
客户主导

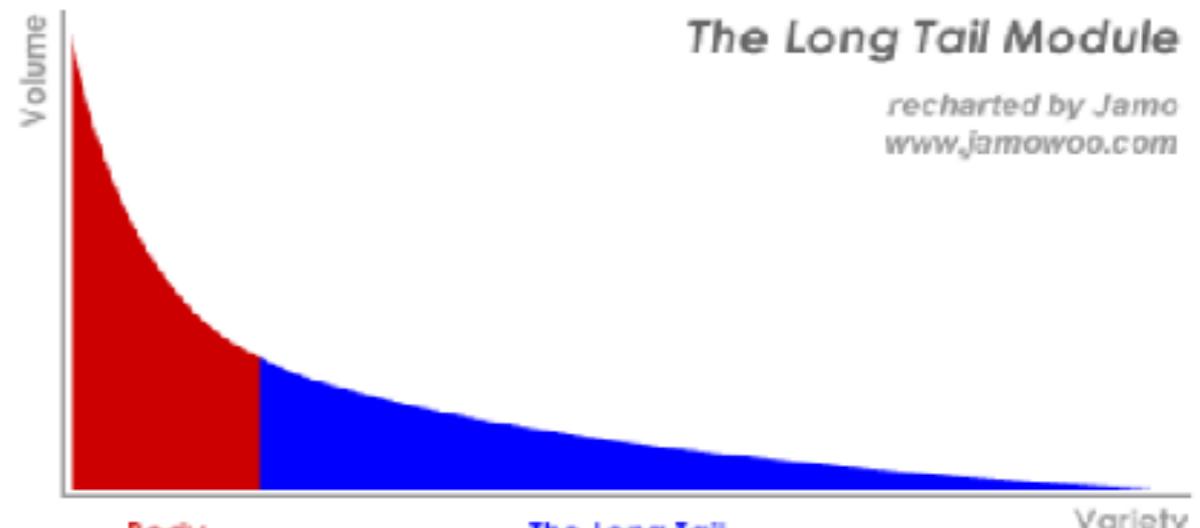
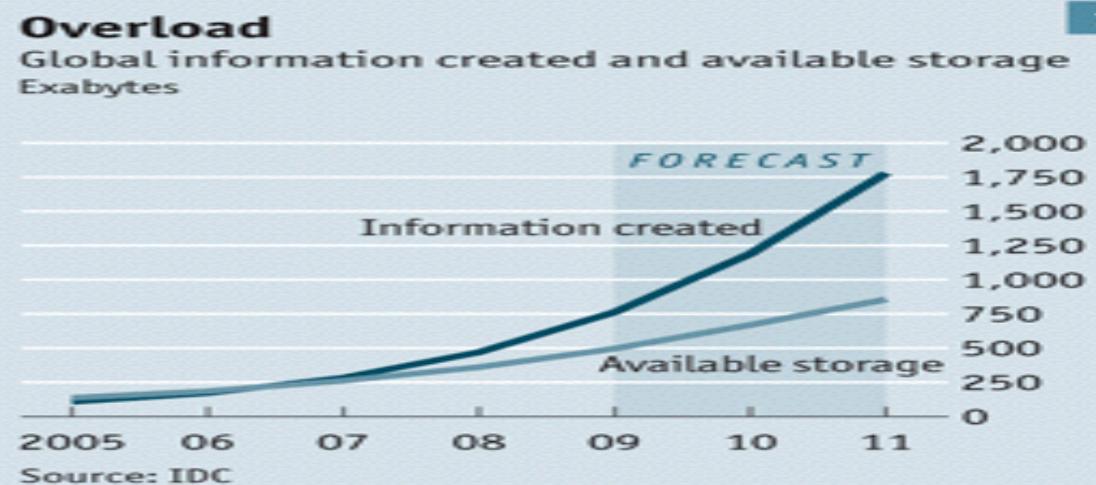
# 为什么要个性化?

## 信息过载

- 海量数据与海量信息

## 消费者异质性

- 消费者需求呈现长尾化趋势



# 个性化营销：数据来源

## 1. 消费者的偏好数据

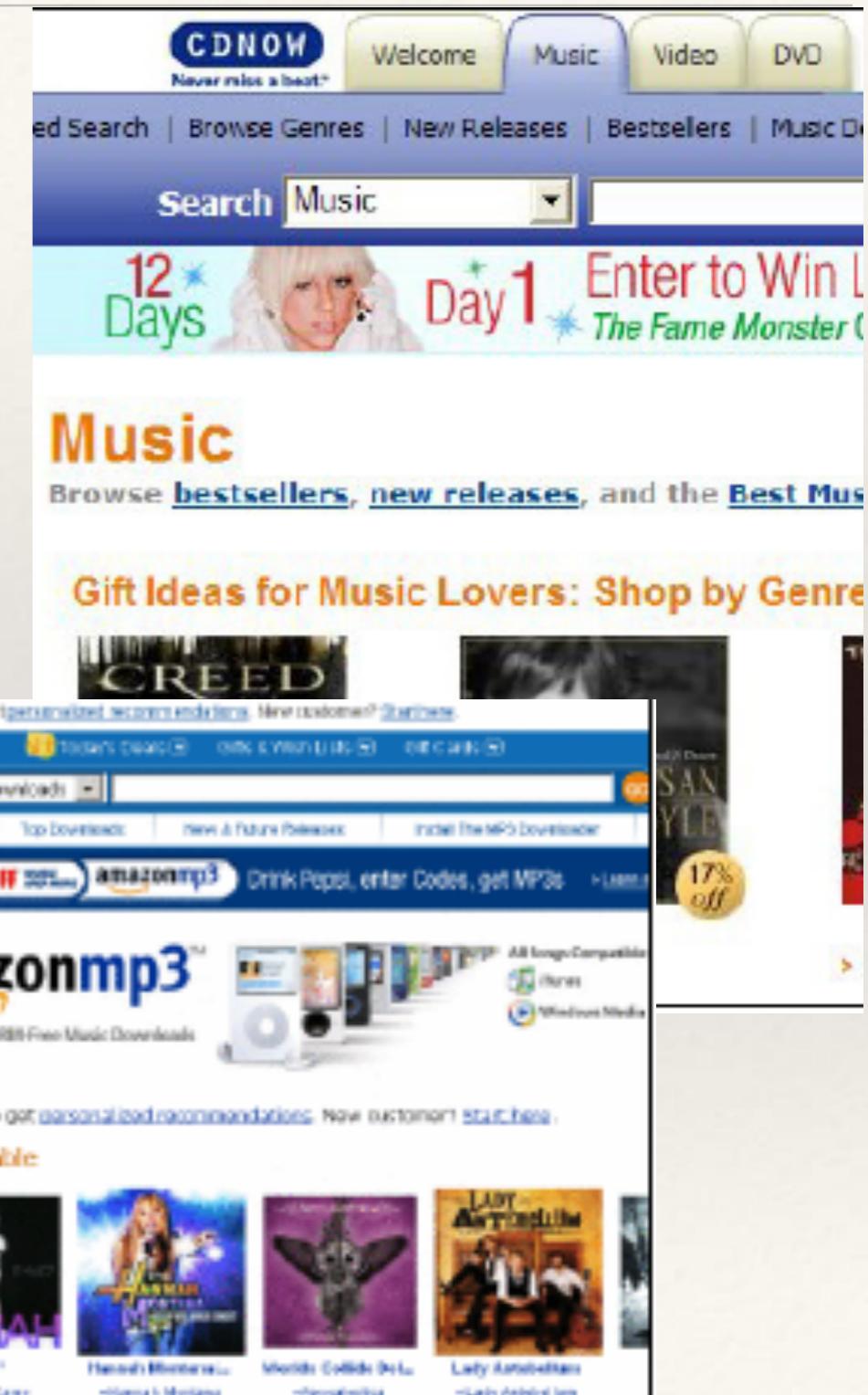
- ❖ 外显数据: 对商品的评分
- ❖ 内隐数据: 购买, 点击

## 2. 商品信息

- ❖ 例如商品属性, 新闻内容

## 3. 消费者特点

- ❖ 例如人口统计变量



# 个性化技术

计算机学

统计学

营销学

机器学习

协同过滤

人工智能

语义分析

云计算

样本推测整体

数据挖掘

商业智能

用户行为分析

偏好预测

广告效果评估

再营销

互联网正从搜索时代进入推荐时代，从用户主动搜索演变为网站主动向用户呈现他们感兴趣的内容。

“如果我有2百万个网络顾客，我就应该有2百万个网络商店”

--- Jeff Bezos, 亚马逊CEO

“所有的媒体都将在未来的3-5年内实现个性化”

--- Sheryl Sandberg, Facebook COO

“推荐系统将成为未来十年里最重要的变革，社会化网站将由推荐系统所驱动”

--- John Riedl, 明尼苏达大学教授

# 个性化商业应用



## 购物

- 电子商务
- 团购
- 定价与促销
- 线下超市

## 内容

- 新闻
- 广告
- 搜索引擎

## 渠道

- 移动互联网
- 社交网络
- 微博

## 生活

- 电影音乐
- 求职招聘
- 约会交友

PERSONALIZED  
DEALS



# 个性化时代的网络营销实践

- 搜索引擎营销  
SEM
- Email营销
- 门户广告
- 网址导航
- 热销/热览榜
- 站内搜索
- 组合销售
- .....

# 个性化时代的网络营销实践

- 搜索引擎营销  
SEM
- Email营销
- 门户广告
- 网址导航
- 热销/热览榜
- 站内搜索
- 组合销售
- .....

+ 个性化 = ?

# 搜索引擎营销SEM + 个性化

SEM + 个性化着陆页  
= 二跳率大幅提升



- 不改变SEM策略
  - 优化着陆页，智能匹配用户偏好和商品——“千人千面”
  - 打通用户全网消费行为

# 邮件营销EDM + 个性化

个性化  
标题

个性化  
品类排序

个性化  
单品推荐



# 站内导购 + 个性化

潜在的消费者为何只逛不买？

用鼠标投票：

Baynote：如果消费者在三次点击之内不能找到需要的商品，95%的人会离开。

•个性化帮助B2C购物网站提升用户体验、保留更多的顾客

Top 10: 个性化热销榜



看过还看过

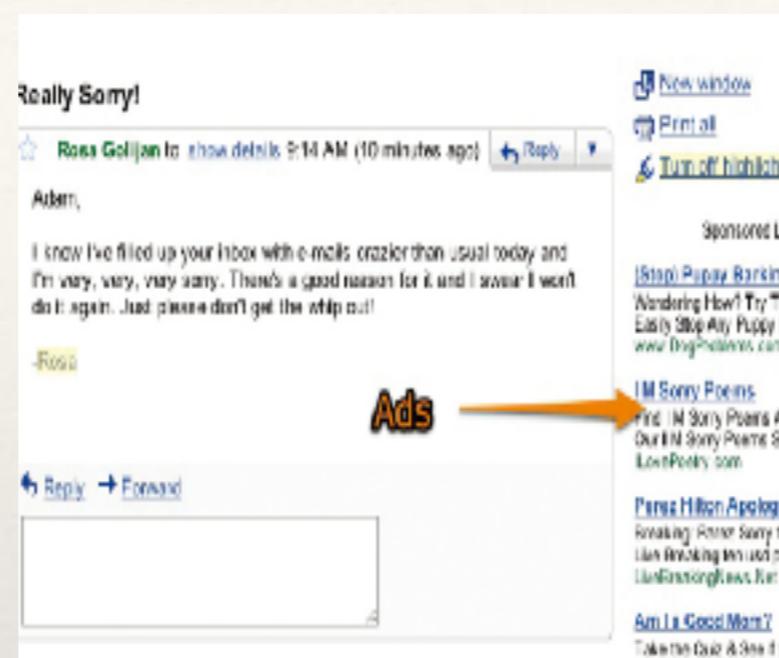


买过还买过



# 更多个性化应用.....

- 个性化广告
- 个性化搜索
- 个性化定价
- 个性化促销
- .....



# 个性化定价和促销



如何通过“完美”的价格，挖掘隐藏利润？

# 个性化定价和促销

- ❖ 如何制定商品的最优价格?
  - ❖ 统一定价减少了利润
  - ❖ 不同消费者的价格弹性不同
  - ❖ 例如航空公司：对每位消费者收取的价格不同



# 个性化定价和促销

## 个性化定价和促销



# 个性化促销的核心思想

由于不同消费者对某商品的支付意愿不同，为了让更多消费者产生购买，在一定机制下给**不同消费者**发放**不同面值的优惠券**，以产生最大的销售。

# 个性化定价的问题

- ❖ 亚马逊个性化定价
- ❖ 不同的消费者购买CD的价格不同

# 个性化定价的问题

- ❖ 亚马逊个性化定价
  - ❖ 不同的消费者购买CD的价格不同
    - ❖ 存在什么问题?

# 个性化定价的问题

- ❖ 亚马逊个性化定价
  - ❖ 不同的消费者购买CD的价格不同
    - ❖ 存在什么问题?
    - ❖ 如何解决?

# 个性化定价和促销

- 效仿亚马逊，却不会造成公共危机的可能性？

# 个性化定价和促销

- 效仿亚马逊，却不会造成公共危机的可能性？
- ❖ 维多利亚的秘密：人口统计信息



# 个性化定价和促销

- ❖ 消费者什么时候不会产生不公平感？
  - ❖ 个性化折扣券比个性化定价更好
  - ❖ 设置烦扰成本，让消费者进行自我选择



- 执行中应注意：
  - 识别价格敏感性
  - 设置合理的烦扰成本
  - 开始先解释价格差异的原因
  - 雇佣优秀的价格分析师

# 个性化定价和促销

- ❖ 案例1：山姆俱乐部的个性化折扣券
  - ❖ 个性化的电子省钱计划：E-value
  - ❖ 找出价格敏感性的消费者及其喜欢的商品
  - ❖ 送出个性化的折扣券



# 个性化定价和促销

- ❖ 案例2: 埃森哲的个性化定价工具
  - ❖ 对历史成交数据进行数据挖掘，识别消费者的购买习惯
  - ❖ 结合商店的信息（如库存等），通过计算决定最优的促销计划

Search

**accenture**

About Accenture | Newsroom | Careers

High performance. Delivered. | Home Consulting Technology Outsourcing | Industries Served Research & Insights |

Home > Services > Accenture Technology Labs > Personalized Pricing Tool

**Personalized Pricing Tool**



Watch the Video: Windows Media | Real Player | Quicktime

Print Article

E-mail to a Colleague

**Contact Us**

To discuss how we can help your organization, call us toll-free at 1 (877) 889-9009. Outside the United States and Canada please dial 1 (312) 842-5012.

► Send Us an E-mail  
► More Contact Information

**How may we help you?**

► Your Content  
► Request for Services  
► Alerts & Newsletters  
► Send Site Feedback

# 个性化促销：两个核心技术问题

- ❖ **估计消费者的支付意愿**：可以根据用户浏览和购买行为与其它用户属性来估计。
- ❖ **机制的设计**：需要概率模型，需要 randomize。
  - ❖ 比如即使一个消费者满足了某些规则，可以发放某个优惠券，也可以用一个随机概率来决定是否发放。这样可以在一定程度上防止消费者学习这些规则。

# 个性化促销：优势

- ❖ 交叉销售
- ❖ 向上销售 / 升级销售
- ❖ 固货管理 + 新品促销
- ❖ 吸引新客户
- ❖ 保留老客户
- ❖ 品牌 / 商家转换（从竞争对手转化）
- ❖ 提升销售额
- ❖ 提升总利润
- ❖ 精准定位目标客户
- ❖ 一度价格歧视

# 需要解决的问题

- ❖ 哪个商品需要发coupon
- ❖ 什么情况下发
- ❖ 发给谁
- ❖ 发多少面值的coupon

# 个性化优惠券的发放方法

- ❖ 确定商品
- ❖ 确定coupon面值范围
- ❖ 确定有效期+用户限定+商品限定等
- ❖ 确定coupon的触发条件
- ❖ 发放（当前访问页面跳出javascript窗口，或建立用户的个人coupon 袋，EDM）
- ❖ 用户接受，放入e-coupon wallet

## 规则引擎 (Rules engine) :

### 单规则+多规则:

- 某用户过去某段时间内访问某商品次数 > 阈值；
- 某用户过去某段时间内浏览某商品总时长 > 阈值；
- 某用户完成某商品的购买后；
- 某用户距离上次购买（访问）时间 > 阈值；
- 某商品销售速度 < 阈值；
- 某商品存货数量 > 阈值；

# 个性化新闻

- ❖ 如何吸引消费者持续地到你的网站阅读新闻？
- ❖ 个性化的新闻内容
  - ❖ 新闻主页, 例如我的Washington Post
  - ❖ 新闻整合: 谷歌新闻, [百度新闻](#)
  - ❖ RSS, 如Google Reader
- ❖ 然而, 定制并不能自动更新消费者的偏好。消费者需要不断地自行更新订阅。

# 个性化新闻

- ❖ 案例1: Digg和协同过滤方法
  - ❖ 用户分享阅读过的新闻链接
  - ❖ Digg通过阅读历史识别用户间的相似性
  - ❖ 将有类似偏好的消费者读过的文章推荐给某个用户

The screenshot shows a Digg user profile for a user named 'Leo'. The profile includes a blue placeholder profile picture, the name 'Leo', a location 'Las Vegas, NV', and a 'Following' button. Below the profile, there are tabs for 'All', 'Diggs', 'Submissions', 'Comments', 'Followers', and 'Following'. The main content area displays three news items:

- Lighting EVER - High Quality LED Lighting Store**  
sqdoo.com — Lighting EVER offers high quality LED lighting. We try our best to help our customer save energy bills, enjoy better sense and green living. Lighti... 4 days ago  
via HappyL 0 Comments 0 Save 0 Dury
- Outdoor LED Flood Lights, Outdoor LED Lighting**  
lightingeve.com — High quality outdoor LED flood lights here. It saves over 50% energy by our LED flood lights. Aug 30, 2010  
Ma rockII 0 Comments 0 Save 0 Dury
- Press Release - Lighting EVER Partners with Energy Star to Promote LED Lighting**  
24-7pressrel.com — Lighting EVER partners with Energy Star. You can buy Energy Star qualified products from Light EVER Retail.com

On the left sidebar, there are summary statistics: 61 Followers, 1 Following, 8 Diggs, and 0 Daily Diggs.

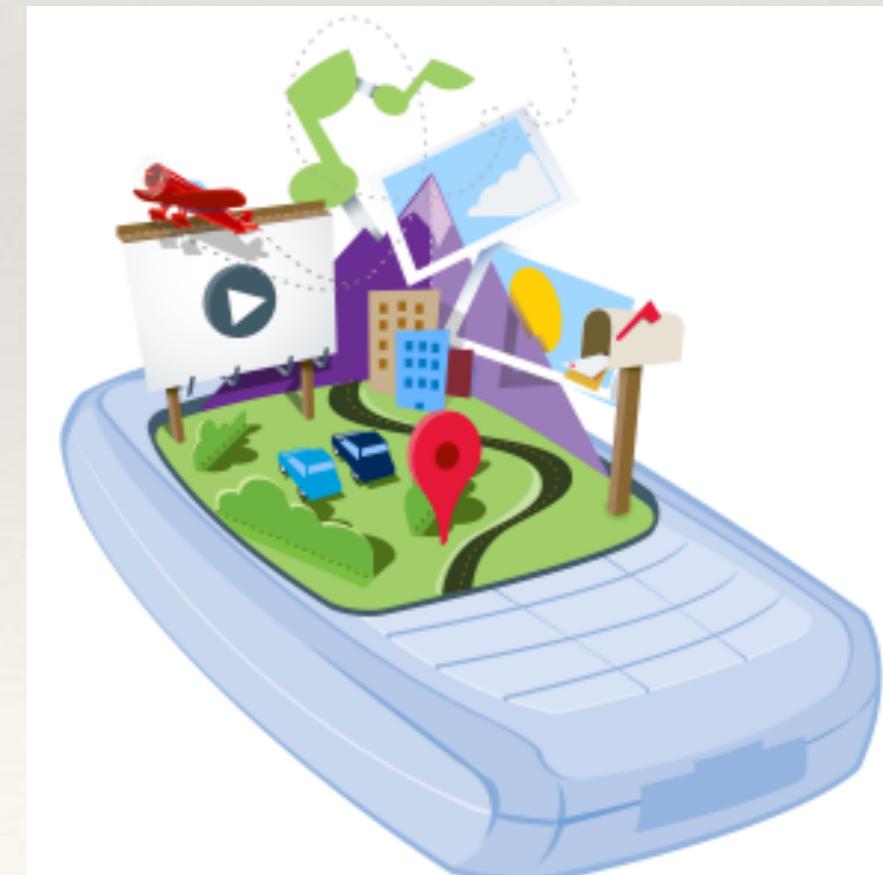


# 个性化新闻

- ❖ 案例2: Reddit和用户决定内容
  - ❖ 用户分享阅读过的新闻链接和自己原创内容
  - ❖ 其他的用户可对发布的链接或内容投支持票或反对投票，得分突出的链接会被放到首页，低的不再显示。
- ❖ 案例3: Meehive和社交网站
  - ❖ 将Meehive账户和Twitter连接起来
  - ❖ 记录用户或者其朋友在Twitter上发表或分享的关键词
  - ❖ 推荐包含这些关键词的新闻

# 个性化手机内容

- ❖ 手机上的个性化
- ❖ 消费者的需要：
  - ❖ 展示内容：用户信息过载，小屏幕限制了手机可以展示的内容
  - ❖ 数百万的应用程序
  - ❖ 时间压力
- ❖ 移动数据来源的不同
  - ❖ 不通过cookie，但能获得精准的位置信息



# 个性化手机内容

- ❖ 案例1: Sidebar和个性化的手机内容
  - ❖ 跨产品类别进行推荐，例如新闻、音乐、铃声、应用和视频等
  - ❖ 收集消费者是否喜欢某个商品或应用
  - ❖ 推荐类似的手机应用

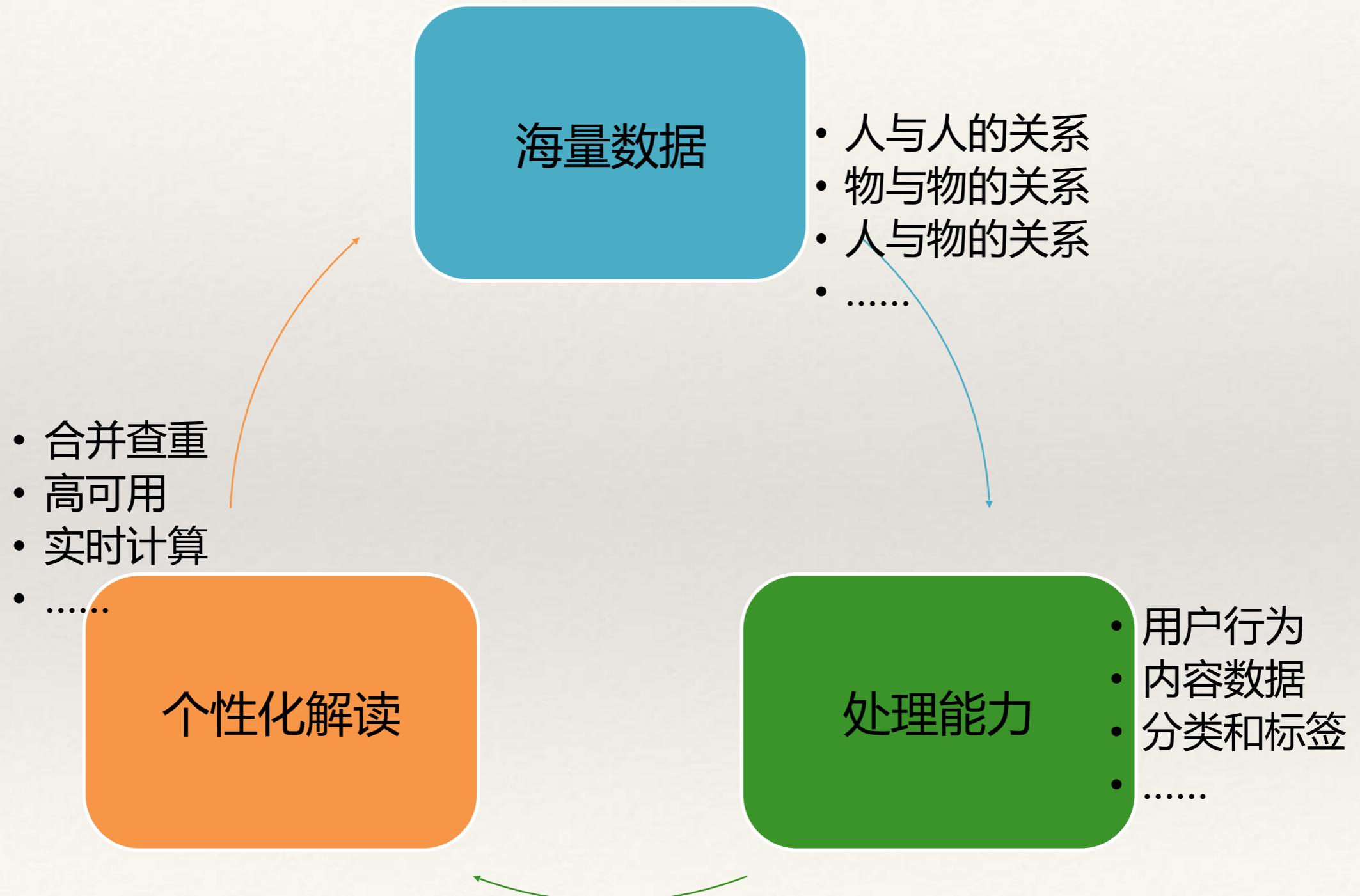


# 个性化手机内容

- ❖ 案例2: Foursquare和基于位置的个性化
  - ❖ 消费者可以用手机确认其朋友是否在附近
  - ❖ 根据消费者的位置信息提供个性化的广告，例如最近的咖啡店



# 如何“个性化”才能成功?

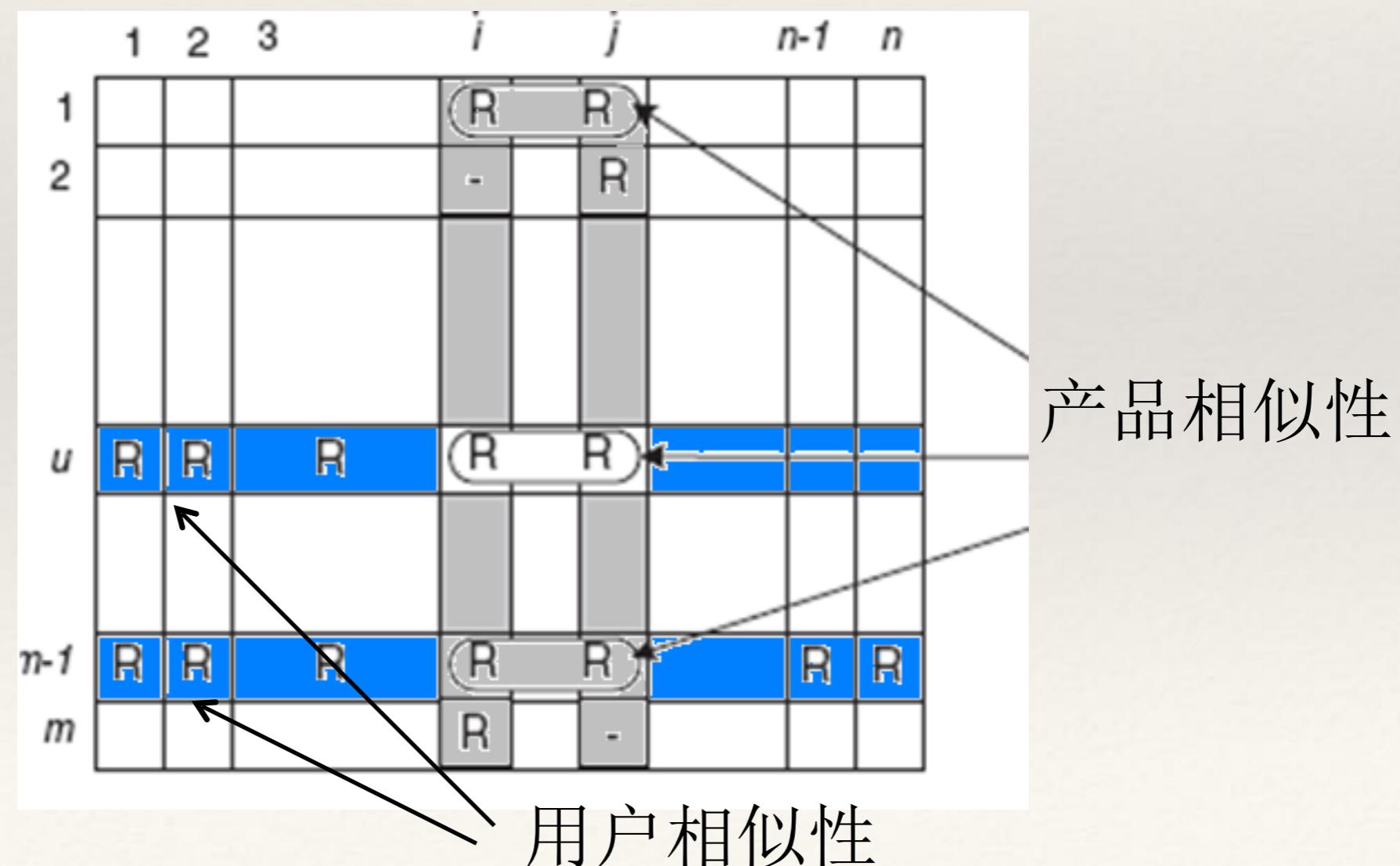


# 主要内容

- ❖ 个性化概述
- ❖ 个性化推荐算法
- ❖ 协同过滤推荐算法实现

# 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



# 协同过滤推荐方法

- ❖ 如何计算用户之间或产品之间的相似性?
  - ❖ 基于记忆: 采用没有随机项的模型来预测用户偏好
    - ❖ 例如, 皮尔森相关系数, 用户或产品向量的Cosine值
  - ❖ 基于模型: 用统计模型预测用户偏好
    - ❖ 例如, 贝叶斯聚类法和贝叶斯网络

# 示例：基于产品的协同过滤推荐

Amazon.com

Hello, Meng Su. We have recommendations for you. (Not Meng?)

Meng's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Get FREE 2-Day Shipping for Father's Day Sponsored by Black & Decker

Your Account | Help

Shop All Departments | Search | All Departments | GO | Cart | Wish List

Your Amazon.com | Your Browsing History | Recommended For You | Rate These Items | Improve Your Recommendations | Your Profile | Your Communities | Learn More

Meng, Welcome to Your Amazon.com (If you're not Meng Su, click here.)

Today's Recommendations For You

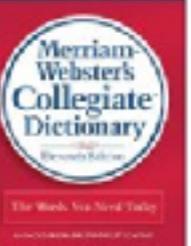
Here's a daily sample of items recommended for you. Click here to see all recommendations.

Page 1 of 37

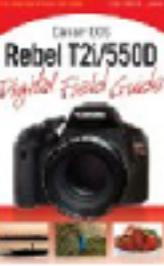
  
[Opteka Battery Pack Grip / Vertical Shutter...](#)  
★★★★★ (20) \$79.95  
Fix this recommendation

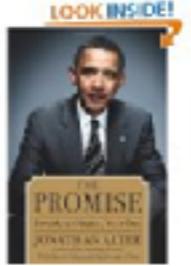
  
[Cokin Dedicated \(Bayonet\) Lens Hood for...](#)  
★★★★★ (1) \$3.99  
Fix this recommendation

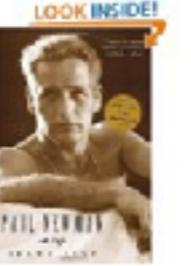
  
[Primary Colors: A Novel of Pol... \(Paperback\) by Anonymous](#)  
★★★★★ (1) \$11.92  
Fix this recommendation

  
[Merriam-Webster's Collegiate... \(Hardcover\) by Merriam W...](#)  
★★★★★ (1) \$14.90  
Fix this recommendation

New For You

  
[Canon EOS Rebel T2i / 550D...](#)

  
[The Promise: President Obama's Life Story by Jon Meacham](#)

  
[Paul Newman: A Life \(Paperback\)](#)

Improve Your Recommendations

The American Heritage Dictionary of the English Language, Fourth Edition: Print and CD-ROM Edition ()

Rate this item:  
★☆☆☆☆

This was a gift  
 Don't use for recommendations

# 示例：基于产品的协同过滤推荐

Amazon.com

Hello, Meng Su. We have recommendations for you. (Not Meng?)

Meng's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Get FREE 2-Day Shipping for Father's Day Sponsored by Black & Decker

Your Account | Help

Shop All Departments | Search | All Departments | GO | Cart | Wish List

Your Amazon.com | Your Browsing History | Recommended For You | Rate These Items | Improve Your Recommendations | Your Profile | Your Communities | Learn More

Meng, Welcome to Your Amazon.com (If you're not Meng Su, click here.)

Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to see all recommendations.

Page 1 of 37

**Opteka Battery Pack Grip / Vertical Shutter...**  
★ ★ ★ ★ (1) \$79.95  
Fix this recommendation

**Coolpix Dedicated (Bayonet) Lens Hood to...**  
★ ★ ★ ★ (1) \$3.99  
Fix this recommendation

**Primary Colors: A Novel of Pol... (Paperback) by Anonymous**  
★ ★ ★ ★ (1) \$11.92  
Fix this recommendation

**Merriam-Webster's Collegiate Dictionary (Hardcover) by Merriam W...**  
★ ★ ★ ★ (1) \$14.90  
Fix this recommendation

New For You^

**Canon EOS Rebel T2i / 550D...**

**The Promise: President Obama...**

**Paul Newman: A Life (Paperback)**

Improve Your Recommendations

The American Heritage Dictionary of the English Language, Fourth Edition: Print and CD-ROM Edition ()

Rate this item:  
★ ★ ★ ★ ★

This was a gift  
 Don't use for recommendations

LOOK INSIDE!

LOOK INSIDE!

LOOK INSIDE!

# 协同过滤方法优缺点

- ❖ 优点
  - ❖ 简单易用
  - ❖ 实际应用效果好
- ❖ 缺点
  - ❖ 存在冷启动问题

# 基于内容的方法

- ❖ 基于内容的方法
  - ❖ 采用目标用户的历史数据、产品信息，例如关键词和产品属性
  - ❖ 对于可以用关键词概括的文本产品(如新闻)比较适用
- 例子: **Pandora**
  - 对音乐识别出400多个“基因”



# 示例：基于内容的方法



(点击观看高清组图)

新浪体育讯 北京时间10月11日晚，2018世界杯预选赛亚洲区12强小组赛第四轮赛事展开，中国国家男足客场0-2负于乌兹别克斯坦，国足四轮赛事1平3负只拿到1个积分，乌兹别克斯坦3胜1负积分提升到9分。比赛中乌兹别克斯坦不仅打入2球，还3次射中中国队球门门框。[\[小炮今日抛开感情因素再战巴甲\]](#)[\[巴甲不假！近期命中率达83%\]](#)[\[1元可知赛果！\]](#)

10月6日的第三轮赛事中国足主场0-1负于叙利亚，该场比赛“零射正”的低迷表现以及三轮仅拿到1个积分的情况令国足承受了极大压力。此战高洪波大幅调整阵容，首发名单较之对叙利亚一战进行了7个位置的调整。乌兹别克斯坦在前两轮拿到连胜，但上轮主场负于伊朗。海因里希与效力长春亚泰的伊斯梅洛夫累计黄牌停赛，此前有伤病麻烦的核心骨干杰帕罗夫、艾哈迈多夫以及效力于北京国安的克里梅茨、谢尔盖耶夫均进入首发阵容。

乌兹别克斯坦在开场快速拼抢的短兵相接中迅速体现了传递运转能力、阵型大幅前压；中国队在防守中全线回撤己方半场。第3分钟，图赫塔胡贾耶夫回传失误送给中国队角球机会，蒿俊闵快速传球制造了一定威胁，迫使门将单拳将球击出。上一场对叙利亚一战中国队也是开场就获得角球，但贯穿比赛的传中球高度与落点均不理想。

- 1 [火箭主帅:哈登能胜任控卫?](#)
- 2 [火箭前主帅:哈登能胜任控卫?](#)
- 3 [21+14+4!热火巨灵神碾碎篮网](#)
- 4 [丁霞颜妮率辽宁女排京城热身](#)
- 5 [英格兰=欧洲中国队？国足配？](#)
- 6 [火箭腾飞只因放弃一人](#)
- 7 [杨智闪光让队友进球荒继续](#)
- 8 [调查-如何看高洪波突然辞职？](#)
- 9 [球迷吐槽皇马新伯纳乌:电饭煲](#)
- 10 [球迷对国足态度已180度转弯](#)

咪咕善跑  
中国马拉松队  
的选择

中国马拉松队官方合作伙伴  
中国马拉松队官方使用APP

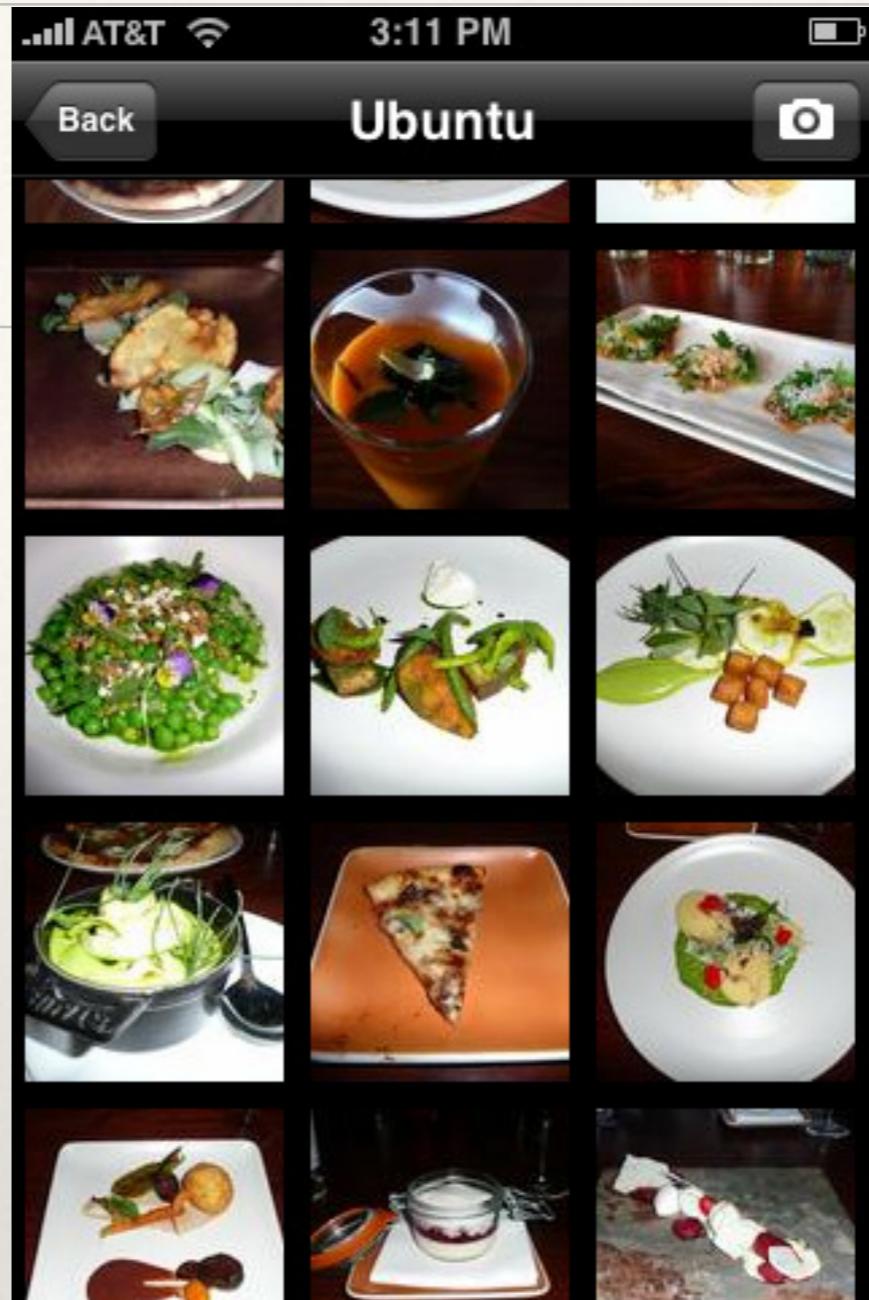
广告

# 基于内容的方法的优缺点

- ❖ 优点
  - ❖ 可避免冷启动问题
  - ❖ 可应用于任何网站
- ❖ 缺点
  - ❖ 复杂度高
  - ❖ 特征提取困难

# 混合方法

- ❖ 混合方法
  - ❖ 结合前两种推荐的结果
  - ❖ 以协同过滤法为基础，融合基于内容的方法
  - ❖ 以基于内容的方法为基础，融合基于协同过滤方法
- ❖ 将来还可以融合基于属性的模型和协同过滤方法



## 苹果手机应用：来自Goodrec的GoodFood

看看其他和你类似消费者都喜欢什么。使用这个应用，可以很方便的对一家餐馆做出评价，还能用文字简单描述餐馆的主要优点。

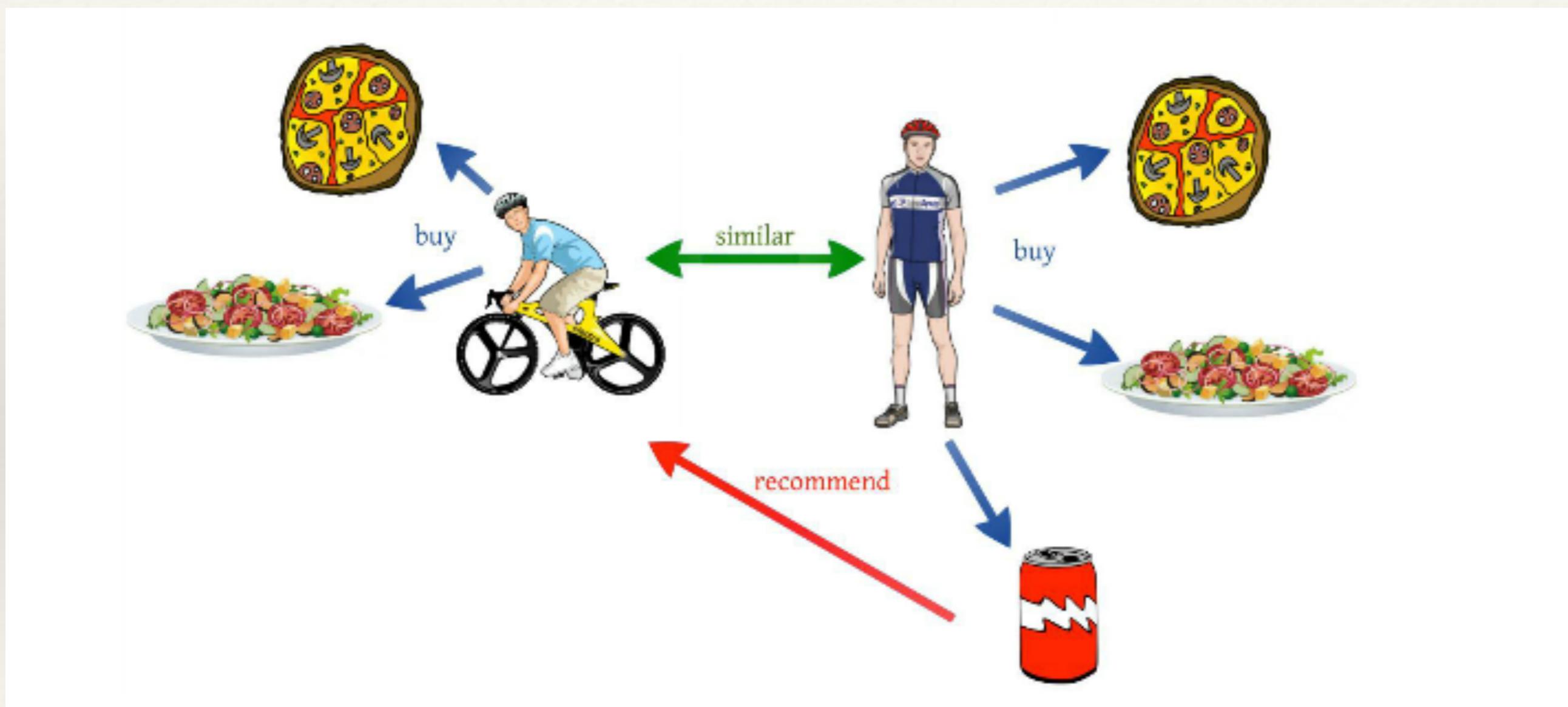
# 主要内容

- ❖ 个性化概述
- ❖ 个性化推荐算法
- ❖ 协同过滤推荐算法实现

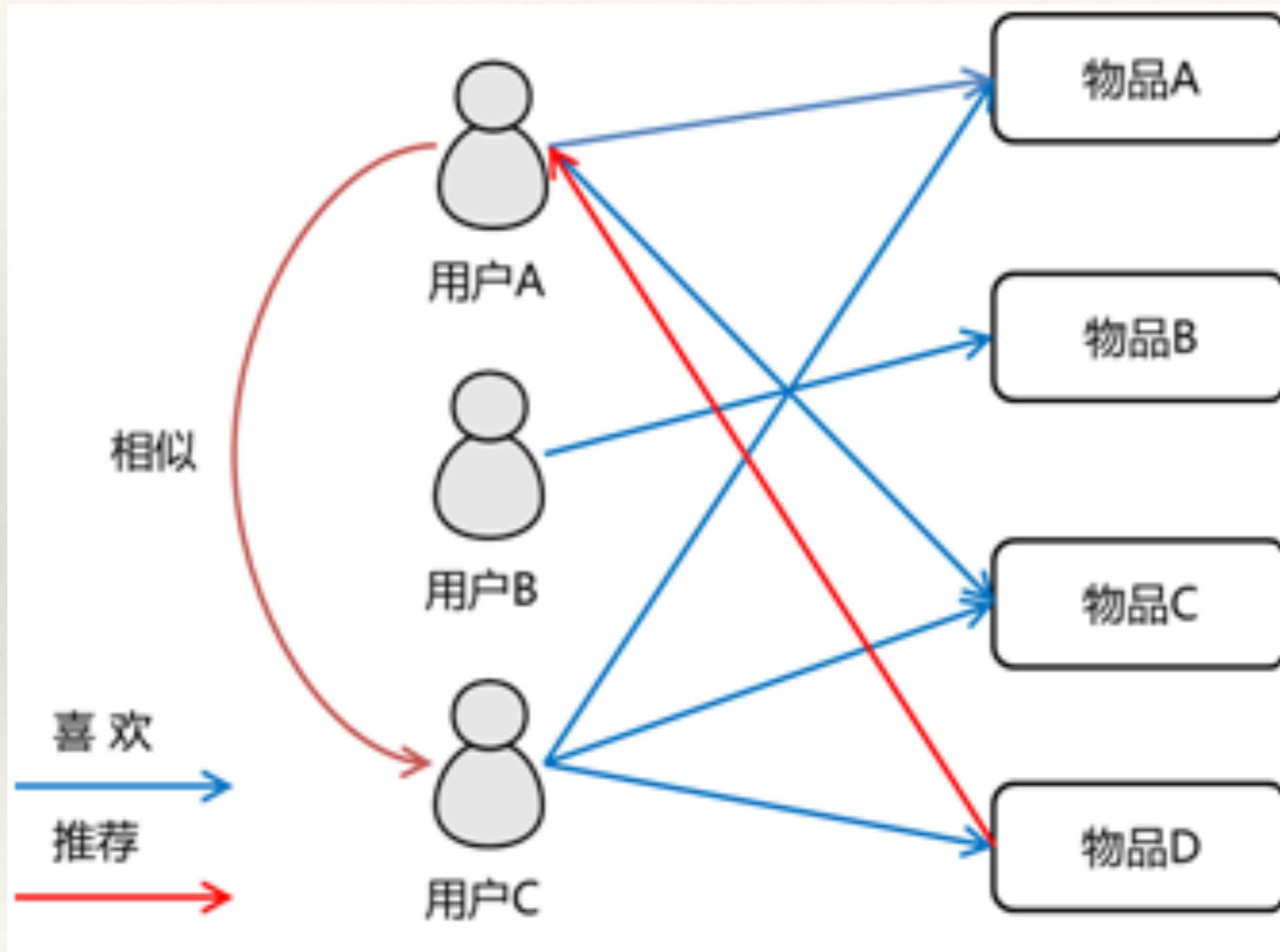
# 推荐算法实现

- ❖ 基于用户的协同过滤算法
- ❖ 基于商品的协同过滤算法

# 基于用户的协同过滤算法



# 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<br>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)

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

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

# 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

The diagram illustrates the calculation of a weighted matrix. It shows two matrices being multiplied (\*), followed by an arrow pointing to the result (=>).

**User\_sim for CHAN**

|           |         |           |        |
|-----------|---------|-----------|--------|
|           | TITANIC | INCEPTION | MATRIX |
| 0.7125006 | 2.5     | 3         | 3      |
| 0.760215  | 3       | 1.5       | 3.5    |
| 0.6831639 | 2.5     | N/A       | 4      |
| 0.7028414 | N/A     | 3         | 4.5    |
| 0.7341787 | 3       | 2         | 3      |
| 0.80555   | 3       | N/A       | 3      |
| 1         | N/A     | N/A       | N/A    |

**\* =>**

|  |           |           |         |
|--|-----------|-----------|---------|
|  | TITANIC   | INCEPTION | MATRIX  |
|  | 1.7812515 | 2.1375    | 2.1375  |
|  | 2.280645  | 1.1403    | 2.66075 |
|  | 1.7079098 | N/A       | 2.73266 |
|  | N/A       | 2.1085    | 3.16279 |
|  | 2.2025361 | 1.4684    | 2.20254 |
|  | 2.41665   | N/A       | 2.41665 |
|  | 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 |