Hello 个性化推荐

马恒阳
2013-10-26

货架与购物车的矛盾

货架与购物车的矛盾

根本原因：信息过载

电商中的购物路径

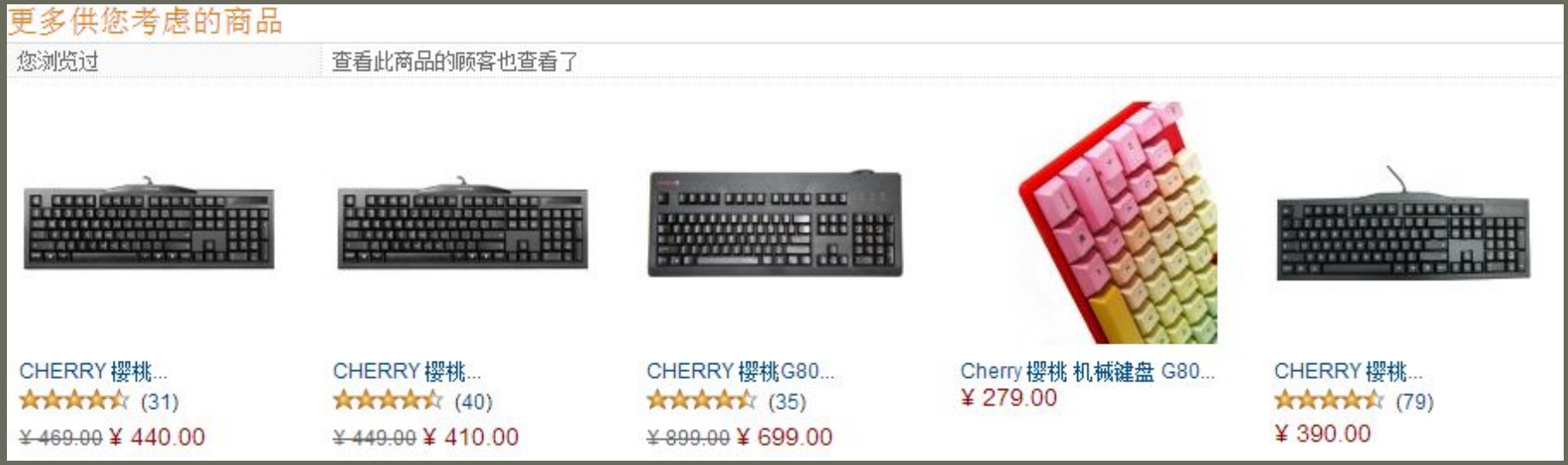
- 普通青年
- 每次购物，必逐层进入其所需物品的类目，疯狂的转动鼠标滚轮，直至目标出现，并乐此不疲

电商中的购物路径

- 文艺普青
- 搜索搞定一切，搜不出来或者排的太靠后就是不存在

电商中的购物路径

- 文艺青年
- 直接打开喜欢的商品，通过关联推荐选择比较



推荐所需技能

- 做个性化推荐系统需要什么技能？

推荐所需技能

- 做个性化推荐系统需要什么技能？

- 答：
- 数学、统计学、数据挖掘、网站分析、python、java、R、SPSS、hadoop、OpenStack、BigData、人工智能、机器学习、语义分析、决策树、聚类、贝叶斯置信网络、神经网络、&*#@%$……

推荐所需技能

- 做个性化推荐系统需要什么技能？
- 答：
- 数学
  pyth
  Ope
  学习
  置信

析、
op、
、机器
贝叶斯
……

Hello 推荐

- 现有用户订单如下，只考虑购物车中有多个商品的记录
- 简单的计算购物车中任意两个商品的组合在所有的购物车中出现的次数，可得商品关联度矩阵

用户	购物车中商品
101	a b c
102	a c
103	d c
104	a c
105	a c d

Hello 推荐

- 现有用户订单如下，只考虑购物车中有多个商品的记录
- 简单的计算购物车中任意两个商品的组合在所有的购物车中出现的次数，可得商品关联度矩阵

用户	购物车中商品
101	a b c
102	a c
103	d c
104	a c
105	a c d

组合	出现的次数
a b | b a	1
a c | c a	4
a d | d a	1
b c | c b	1
d c | c d	2

Hello 推荐

- 根据商品关联度矩阵，给用户推荐商品
- 现有一用户正在浏览商品a，因为a与c同时出现的次数最多，所以应该把商品c推荐给他
- 当然，如果推荐位不只一个，那么应该按照关联度从高到低选取与a最相关的商品做为推荐

Hello 推荐

- 根据商品关联度矩阵，给用户推荐商品
- 现有一用户正在浏览商品a，因为a与c同时出现的次数最多，所以应该把商品c推荐给他
- 当然，如果推荐位不只一个，那么应该按照关联度从高到低选取与a最相关的商品做为推荐

恭喜你！第一版推荐系统已经完成

Hello 推荐

- 根据产品关联度矩阵，给用户推荐产品
- 现有一
同时出现的次数最
- 当然
按照关联度从高
推荐

恭喜
完成

来看看我们的成果吧

这是对苏宁易购的用户浏览数据应用上述算法得出的结果示例，结果还不算坏吧

浏览了"语义网数据管理技术及应用 '746921'的人

还浏览了	相关度
计算机网络实验指导书(第2版)	1
结网@改变世界的互联网产品经理	1
NS网络模拟和协议仿真	1
精彩网址收藏夹	1
反黑风暴 网络渗透技术攻防高手修炼	1
现场总线系统设计与应用丛书 LONWORKS总线系统设计与应用	1
电子政务系统建设与管理(1CD)	1
中国互联网协会全国大学生网络商务创新应用大赛优秀案例选辑3	1
物联网技术导论	1
电子政务（第2版）	1
物联网:影响未来 国务院发展研究中心研究丛书	1
智慧的物联网--感知中国和世界的技术	1
计算机网络工程与实训教程	1
TD-SCDMA网络规划与工程	1
网上支付-网商成功之道	1
电子商务概论(第2版)	1
电脑硬道理-网管实战(第12版)(附盘)	1
虚拟计算环境中的覆盖网构建技术	1
淘宝网店拍摄/拍摄/装修/推广完全攻略(附盘)	1
电子政务系统的体系结构	1
数据通信与计算机网络(第3版)/高等学校规划教材	1
IP多播网络的设计与部署.第1卷	1
新手学网络攻防	1

面临的问题

- 数据稀疏
- 二两醋与一斤螃蟹
- 自推
- 促销的干扰
- 推荐结果趋于单一化
- 过时

面临的问题

- 数据稀疏 大部分购买记录都不可用
- 二两醋与一斤螃蟹 活性碳推冰箱
- 白推 手机推内存卡，实际上内存卡是送的
- 促销的干扰 凑单，推出的结果风马牛不相及
- 推荐结果趋于单一化 推荐结果大部分是强相关的
- 过时 推荐出来的商品大都是过时的热销商品

数据稀疏

- 数据稀疏是推荐系统面临的头等问题，考虑前面的推荐系统，如何缓解因可用的购买记录相比总记录数少很多带来的关联矩阵稀疏的问题？

数据稀疏

- 数据稀疏是推荐系统面临的头等问题，考虑前面的推荐系统，如何缓解因可用的购买记录相比总记录数少很多带来的关联矩阵稀疏的问题？

所谓购物车就是一起购买的商品的集合，如果我们把用户一周之内购买的商品都算进一个购物车，可用记录数是不是就多了很多？

过时

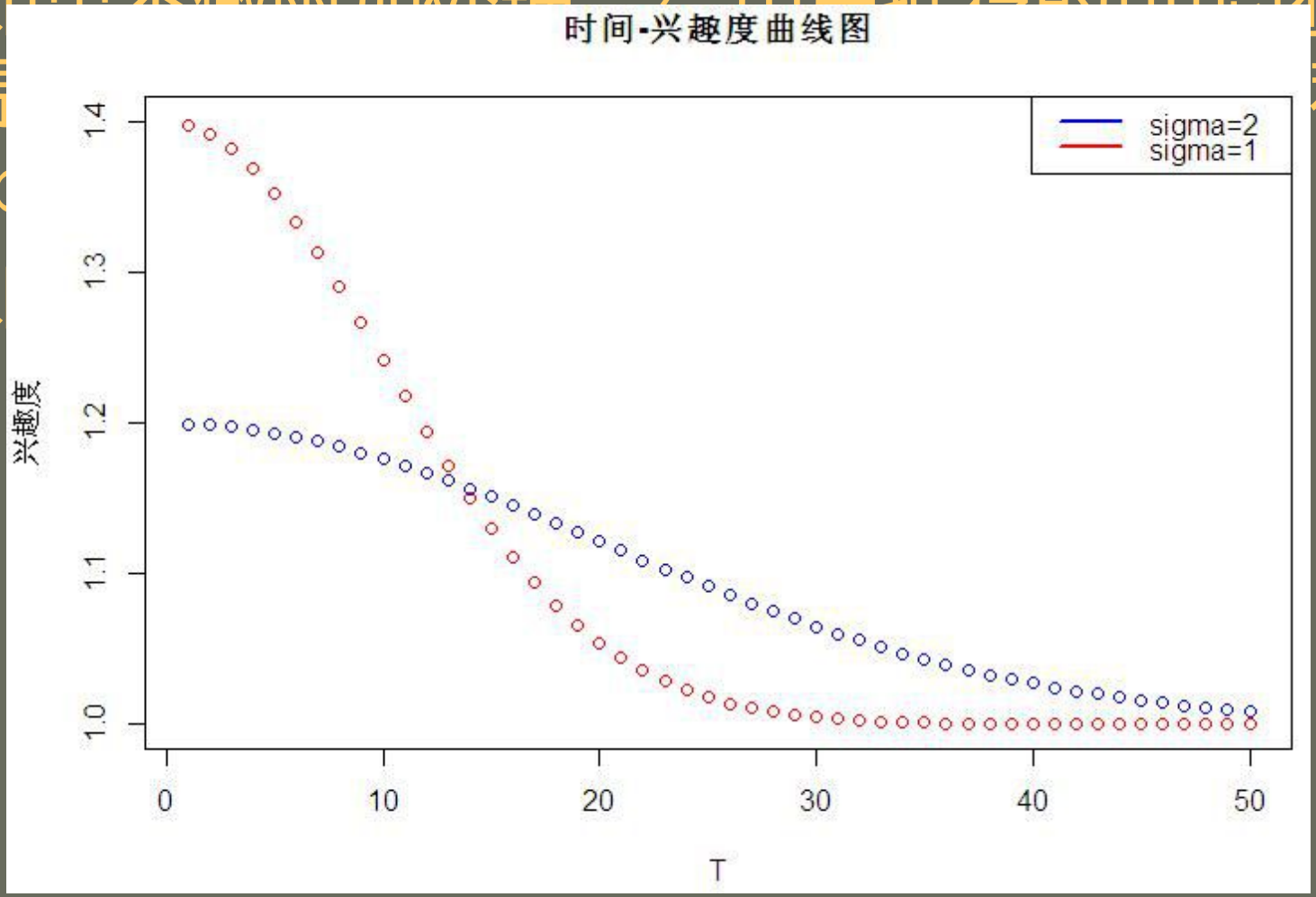
- 如果不做附加处理，给用户推荐的可能还是1年前的热销商品，比如：移动电源推荐iphone4。
- 如何让推荐系统跟上时代的潮流？

$$f(u,i) = (\frac{1}{\sigma * \sqrt{2\pi}} * e^{-\frac{T^2}{2\sigma^2}}) * log_{10}(N+10)$$

- δ：衰减速度　T：相隔天数　N：购买数量

过时

- 如果不做附加处理，给用户推荐的可能还是
- 如



事后诸葛亮

- 根据前面的用户兴趣模型，用户对当前购买的商品的兴趣度应该是最高的，就会得到下面的推荐

离线评测

- 有评测的推荐才是完整的推荐
- “三个代表”：命中率、准确率、覆盖率
- 按时间把订单分为两部分，比如2013-10-1之前的做为训练集，之后的做为测试集
- 在训练集上应用推荐算法，得到关联矩阵，用测试集检验关联矩阵的质量
- 特色指标：热门度、热门比例、覆盖率、新颖度、补全比例

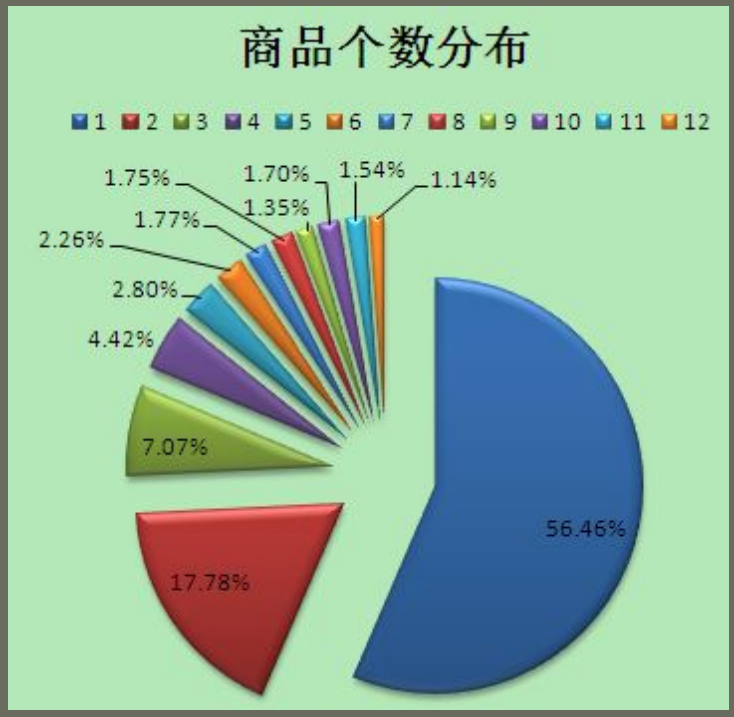
离线评测

购物车大小：99128
推荐结果集大小：70852
处理后的购物车大小：32227
推荐结果集大小：32227

命中率：94099 / 3095981 = 0.0303
准确率：10258 / 188586 = 0.0543
覆盖率：6119 / 72671 = 0.0842

推荐结果分布

- 右图是推荐结果中的商品跨目录数分布
- 可以看出56%的推荐结果都属于同一个目录，推荐结果内聚性非常强



基于物品的协同过滤（一）

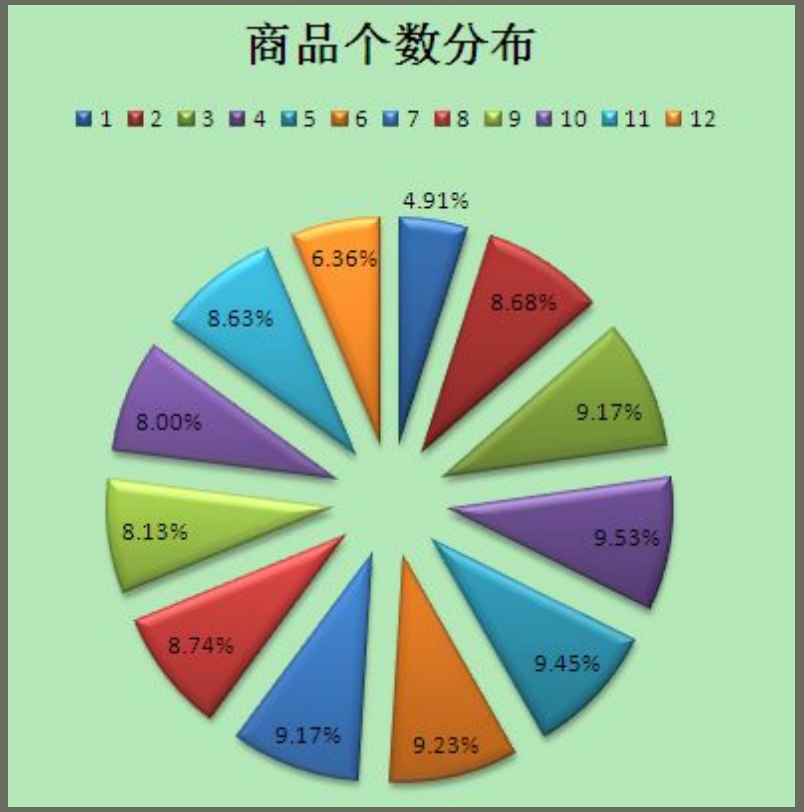
- 基本思想是购买两个商品的人群重合度越高，则两个商品的关联性越强。

$$w(a,b) = \frac{N_a \bigcap N_b}{\sqrt{\|N_a\| * \|N_b\|}}$$

- Na：购买商品a的用户集合
- Nb：购买商品b的用户集合

推荐结果分布

- 计算结果较发散性，除了空调与挂架之类的强相关结果，还会得出手机与笔记本这样的关联关系看起来不是很明显的结果，且推荐结果较多



基于物品的协同过滤（二）

- 理论依据：不活跃的用户对商品相似度的贡献应该大于活跃用户

商品a、b相似度计算公式

$$w(a,b)=\frac{\sum_{i}^{k}\frac{1}{\ln(1+h(i))}}{\sqrt{size(N(a))*size(N(b))}}*\ln(1+k)$$

$N(a)$：购买a的用户集合

N(b)：购买b的用户集合

k = $size(N(a)\cap N(b))$：交集大小

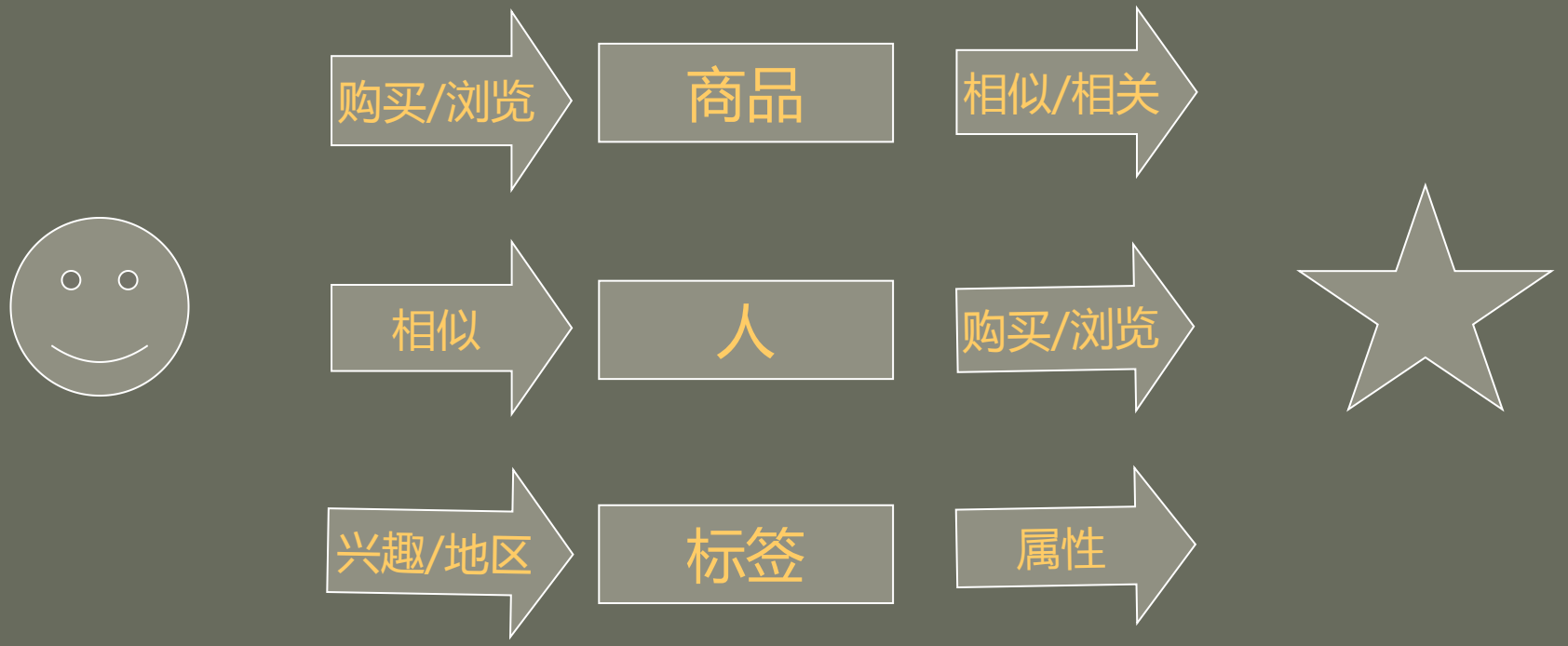
i $\in N(a)\cap N(b)$：i是N(a)与N(b)的交集中的用户

$h(i)$：用户购买的sku数（用户活跃度）

$size()$：集合大小

推荐的本质

- 把人和物通过某种方式联系起来



个性化

- 数字化一个人
- 行为：浏览、购买、收藏……
- 属性：城市、年龄、性别……
- 以计算用户对商品的兴趣度为终极目标
- 最终的推荐结果按兴趣度从高到低排序

行为加权


流量来源
全部来源
进入访次：235,087
访次235,087
网站着陆页
流失访次 129,857
跳出率 55.24%
访次 105,230
商品详情页到达率 44.76%
站内浏览
流失访次 84,613
流失率 80.41%
访次 20,617
购物车到达率 8.77%
加入购物车
流失访次 13,959
购物车流失率 67.71%
下定订单访次 6,658
转化率 2.83%
下定订单

行为加权

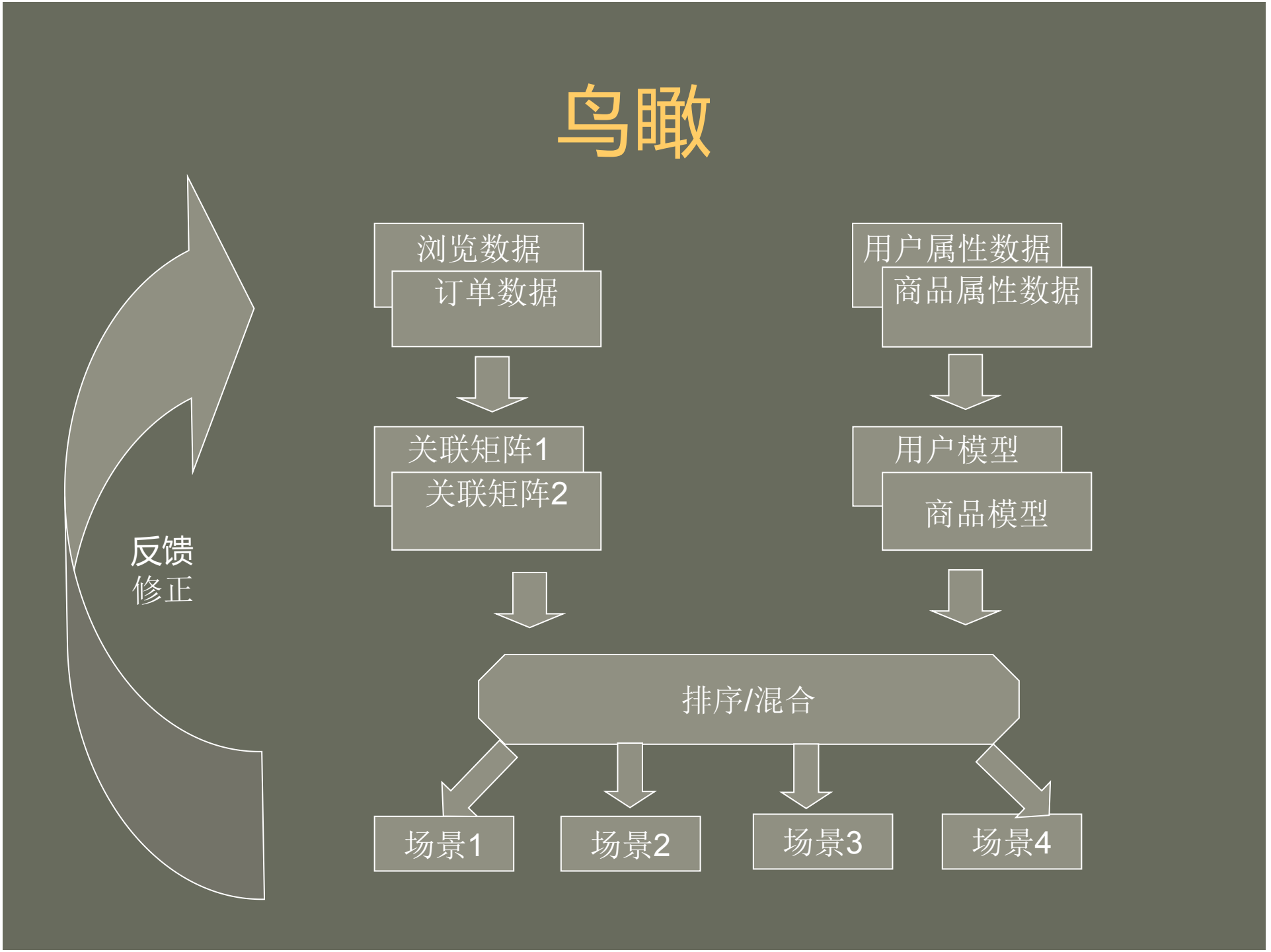
- 用户为每种行为付出的代价是不一样的
- 浏览多少次付出的代价相当于成交付出的代价？
- 设用户购买权重为1

$$\text{浏览权重} = \frac{\text{订单数}}{\text{浏览数}} = \frac{6658}{105230} = 0.063$$

细化场景

- 在不同的页面展示不同的推荐结果，一般常用如下几种场景
- 浏览了还浏览了
- 买了还买了
- 浏览了最终购买了
- 猜你喜欢
- 促销邮件

鸟瞰

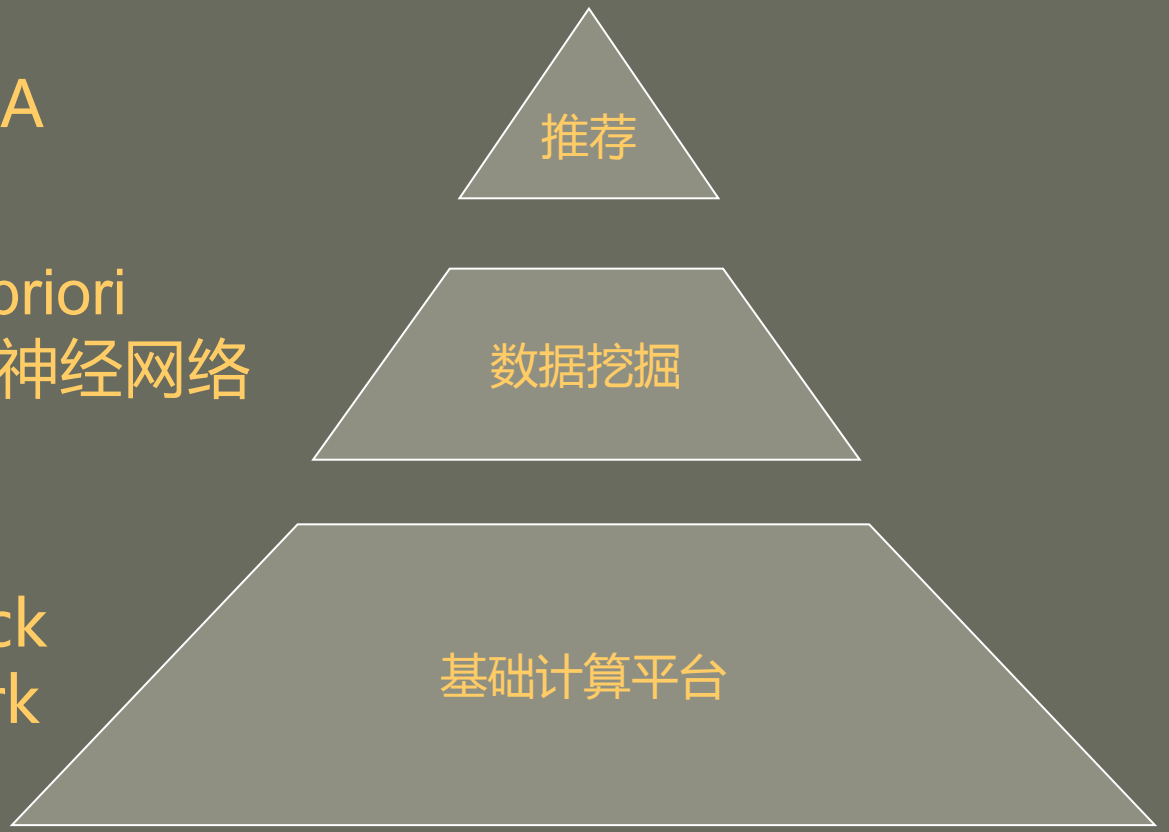

浏览数据
订单数据
用户属性数据
商品属性数据
关联矩阵1
关联矩阵2
用户模型
商品模型
反馈
修正
排序/混合
场景1
场景2
场景3
场景4

技术分布

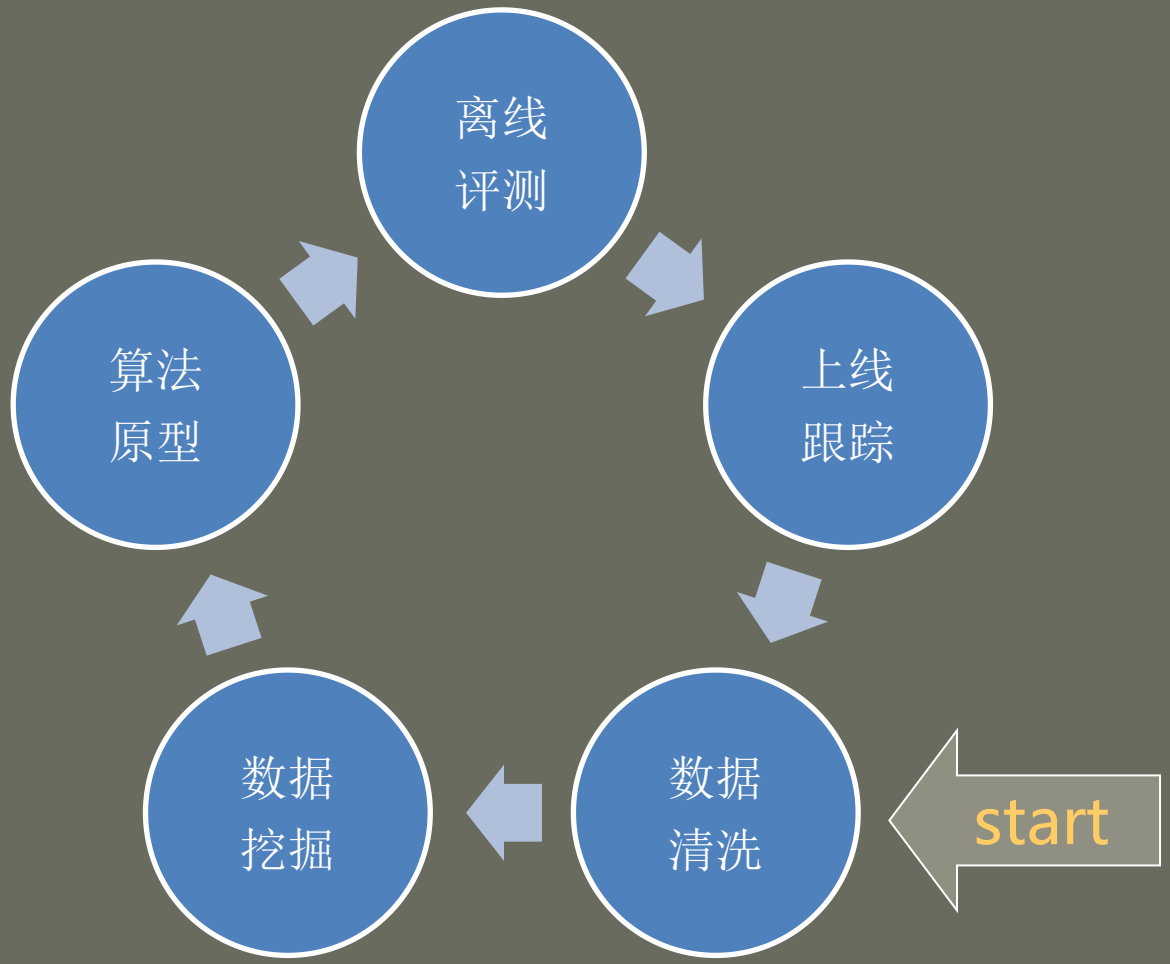
协同过滤、SVD、LSA

K-means、SVM、Apriori
回归分析、决策树、神经网络
数据可视化

hadoop、OpenStack
GPU、HANA、Spark



完整的个性化推荐



合格的个性化推荐攻城狮

- 数据分析与数据挖掘
- 网站分析
- 用户体验
- 消费者行为心理学
- 分布式计算
- 编程多面手

- 简而言之：数据科学家

切入点

- R语言、Mahout、SciPy
- 书籍推荐：
- 《数据之魅：基于开源工具的数据分析》
- 《数据挖掘与R语言》
- 《应用多元统计分析》（哈德勒、西马）
- 国内唯一一本专门讲推荐的《推荐系统实践》，可惜的是书中错误太多不建议初学者看

不懂数据分析的码农
不是一个好的产品经理

Blog: http://my.oschina.net/enyo/blog
E-mail: hengyangma@gmail.com