百度推荐引擎实践系列（一）：策略篇

赵岷
百度 – 推荐与个性化部
2012. 10.20

Baidu 百度

目录

搜索 VS. 推荐：互为补充

信息爆炸
信息过载

\+

知识匮乏
时间有限

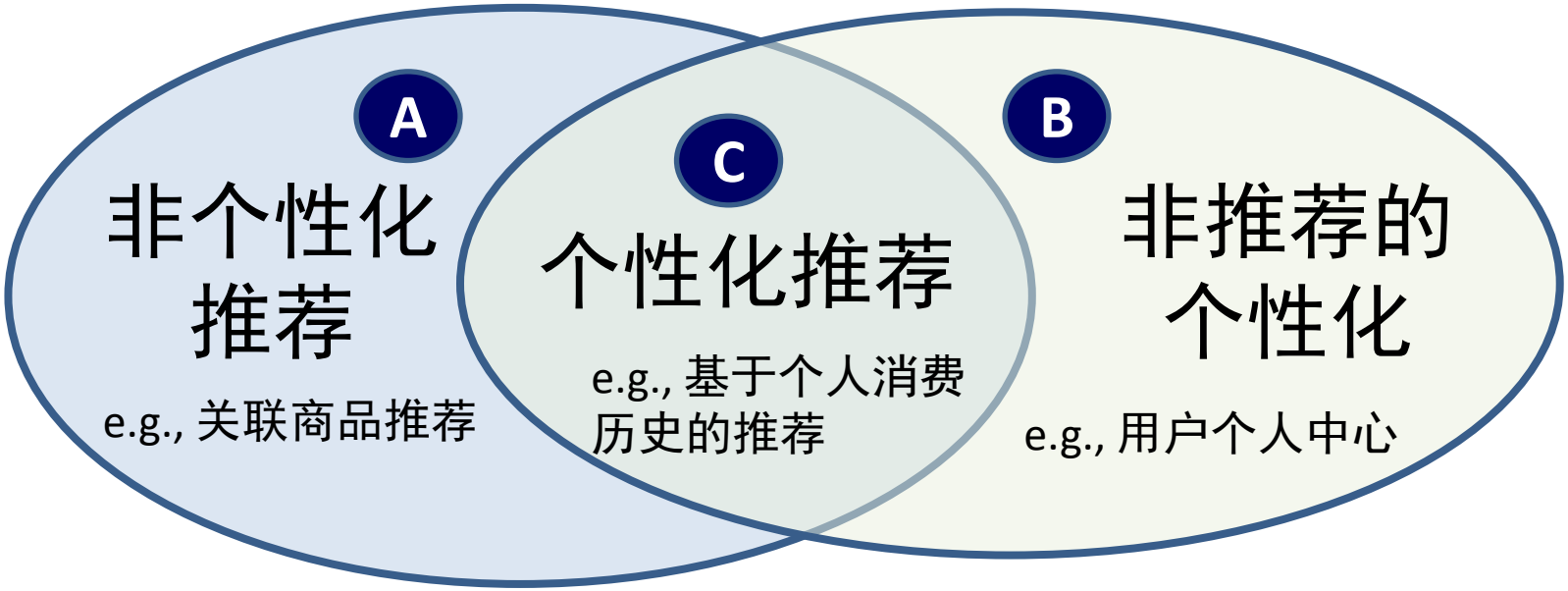
方法 1
用户主动搜索

方法 2
系统主动推荐

用户知道自己想要什么
&& 知道如何描述自己的需求

用户有需求 &&
不知道怎样描述自己的需求
or 不知道去哪寻找 / 懒得找

- ✓ 目标：协助用户高效便捷地寻找/发现信息
  - ✓ 管理&组织、搜索&引导、浏览&发现



- ✓ 产品可兼顾三者，搜索与推荐功能有机结合

为什么个性化推荐越来越热？

- ✓ 提升用户体验和满意度，增强用户粘性
  - 消费需求的变化：
    - 单一/从众 → 多样/个性/品位
  - 信息的极大丰富
    - 需要有效的信息过滤工具
- ✓ 用户数据的积累已经可以支撑个性化应用
  - 能够在线获得大量用户行为、偏好数据
  - SNS的流行，用户逐渐养成分享和接受推送的习惯

个性化推荐：“钱”景

- ✓个性化营销
  - → 电子商务公司
- ✓个性化广告
  - → 以面向个人用户的广告为主要盈利模式的互联网公司

但是……

- ✓ — 除了广告/商品推荐之外，推荐还能带来什么显著收益？
- ✓ — 推荐是锦上添花，还是雪中送炭？独立推荐产品能成功吗？
- ✓ — 会出现像搜索引擎一样成功的推荐引擎吗？或者，推荐引擎将与搜索引擎合二为一？
- ✓ 以上，期待大家的实践 ☺

目录

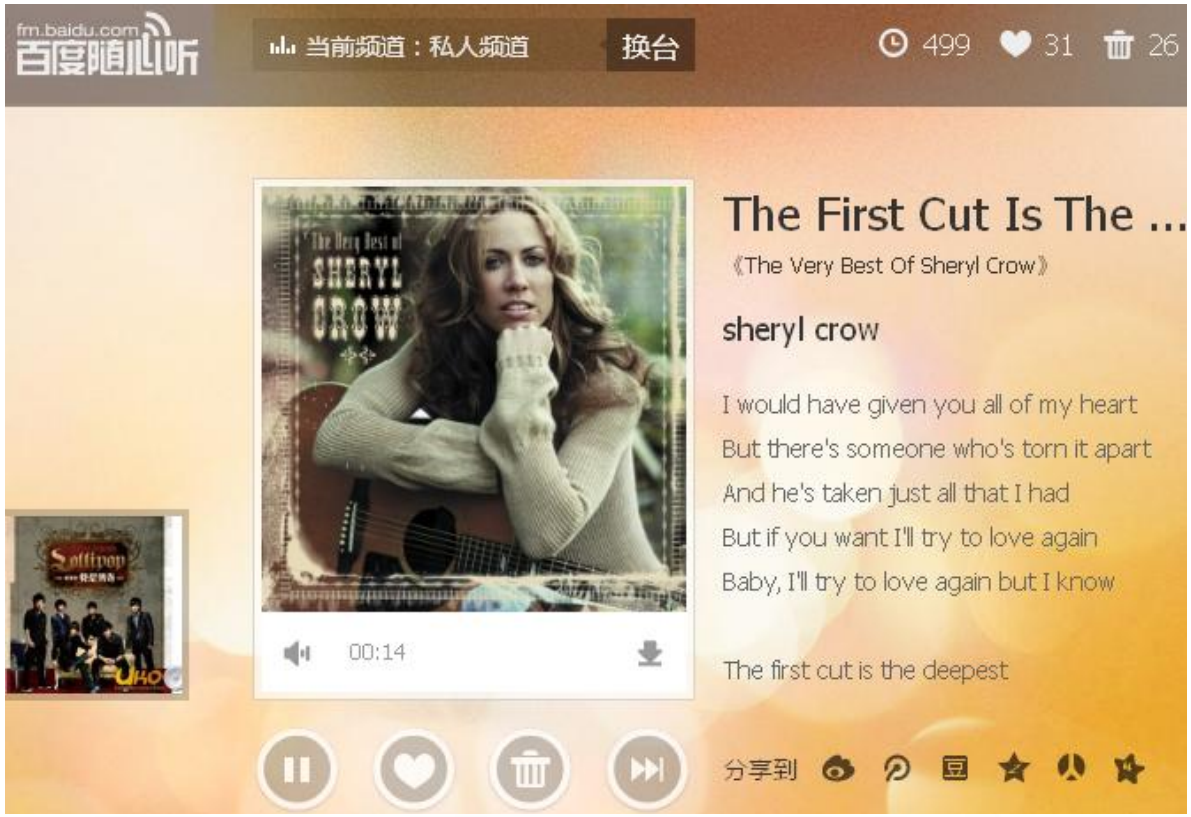
例：个性化上网入口 -- 百度新首页

例：知道问题推荐

标题（共247条）	回答数	提问时间
安徽有BEC高级考点吗？有的话在哪报名以及报名... [英语考试]	0	2012-5-20
10 BEC口语考试时间不合适，可以申请改时间吗，本... [英语考试]	1	2012-5-20
高级考试心得？ [英语考试]	2	2012-5-20
急求BEC中级考试的最终复习资料啊……... [英语考试]	1	2012-5-20
20 谁有英语BEC证件的清晰版扫描件发给我一份 要... [英语考试]	0	2012-5-13
10 新东方BEC主讲谢姣岳的中级口语教材 谁能帮忙... [英语考试]	0	2012-5-17
黄山英语BEC在哪里报名 [英语考试]	1	2012-5-17
推荐一些考bec中级的好方法，我打算下半年考，... [英语考试]	2	2012-5-16
bec中级写报告与商务建议的格式是什么 [英语考试]	1	2012-5-15
求助BEC中级的复习资料 [英语考试]	0	2012-5-19
求教BEC高级 四级630 六级590 托福9... [英语考试]	0	2012-5-21
BEC考试还有15天，做了半个月的题，错的比较... [英语考试]	1	2012-5-18
有无2012中级BEC全套练习资料。 [英语考试]	0	2012-5-14
我想大三上个学期考BEC（就是下个学期），老师... [英语考试]	2	2012-5-17
考BEC的资料谁有啊？ [英语考试]	2	2012-5-12

例：贴吧帖子推荐图片、视频

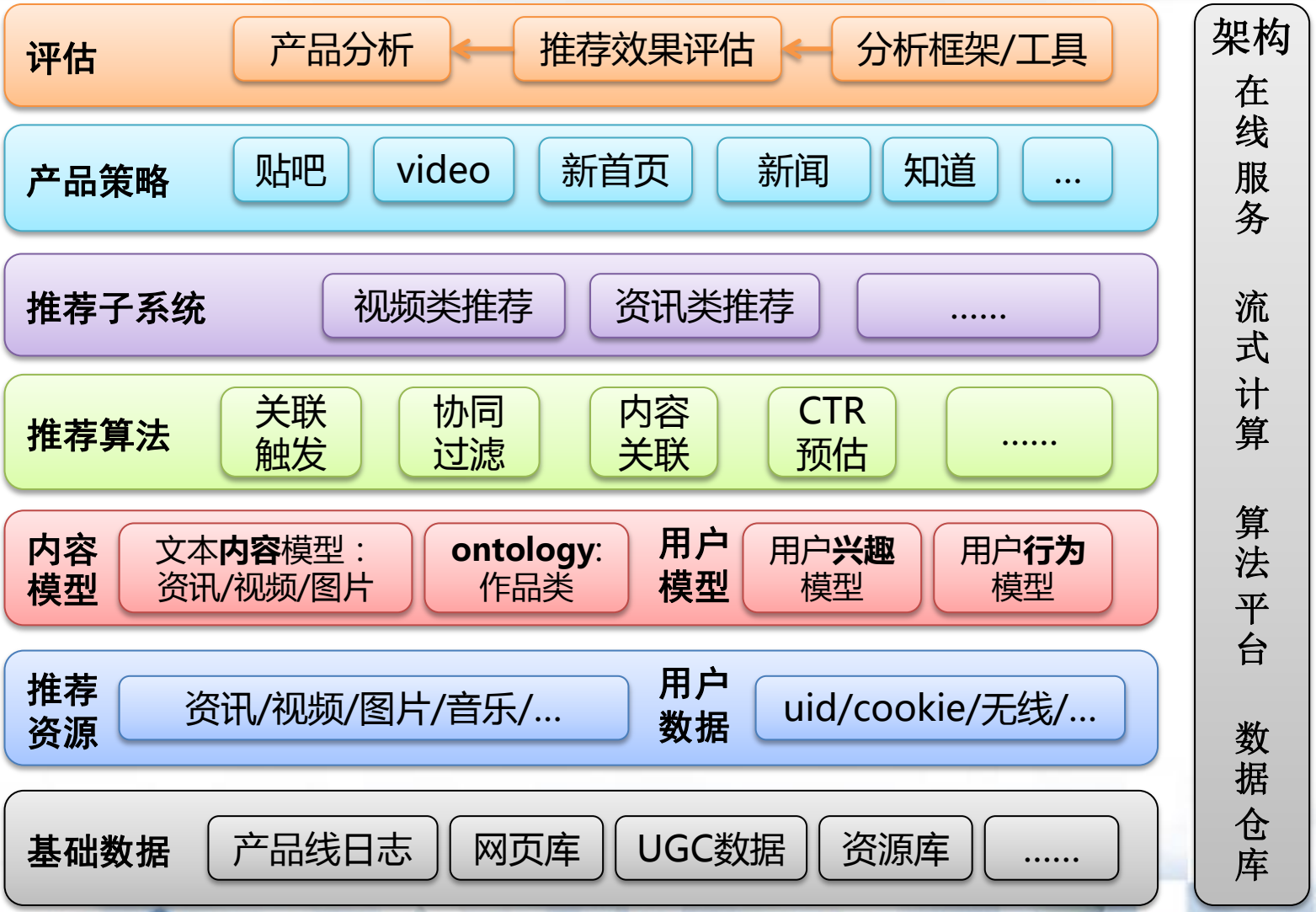
例：音乐推荐


fm.baidu.com
百度随心听
当前频道：私人频道
换台
499
31
26
The First Cut Is The ...
《The Very Best Of Sheryl Crow》
sheryl crow
I would have given you all of my heart
But there's someone who's torn it apart
And he's taken just all that I had
But if you want I'll try to love again
Baby, I'll try to love again but I know
The first cut is the deepest
00:14
分享到

百度推荐实践小结

- ✓ 全类型
  - 资讯  贴吧  新闻  知道  百科  文库
  - 多媒体  图片  视频  百度音乐
  - LBS、APP  地图  百度身边  百度应用  百度游戏
- ✓ 全方位
  - 个人上网入口、各垂直领域、PC+无线

推荐引擎技术Overview



目录

推荐的核心任务

- ✓ Task 1: 通过人的行为/偏好/兴趣、事物的特性等建立事物间和人之间的关联
  - 行为：浏览购买、地理位置、Social Network、……
  - 口味：吃喝玩乐、衣食住行、……
- ✓ Task 2: 把关联的人或物推荐给人
  - 书籍、电影、音乐、文章、网站……
  - 商品、广告……
  - 人、团体、活动……

设计要素

1. ✓ 需求分析和用户调研
2. ✓ 功能设计
3. ✓ 界面设计
4. ✓ 架构设计
5. ✓ 算法设计
6. ✓ 系统评测

~~ 不同推荐系统各部分重要程度不同 ~~

需求分析和用户调研 - 为谁？推荐什么？何时？何地？

- ✓ 为谁（例）：
  - 新用户：兴趣未知，着重多样性、新热
  - 老用户 ：兴趣已知，着重个性化
- ✓ 推荐什么（例）：
  - 价格一致，用户经常购买的类别
    - 书、电影、音乐、文章 → 以用户对内容主题的兴趣为主
  - 价格不一致，用户经常购买的类别
    - 服饰、日用百货 → 视觉、品牌、价格、内容
- ✓ 何时（例）：
  - Email VS. 手机短信 VS. APP推送
  - 短期、长期、周期（节假日、季节、……）
- ✓ 何地（例）：
  - 商家、优惠券推送

功能设计 – 具体实现哪些功能？ 2

- ✓ 产品分类（例）
  - 文本：新闻、博客、小说、论文、……
  - 图片：风景、商品、旅游、……
  - 音频：歌曲、歌手、专辑、……
  - 视频：电影、电视剧、综艺节目、短视频、……
  - 其他：app、位置服务、……
  - SNS：人、群组、……
  - 混合类别~~
- ✓ 数据（例）：文本或其他内容 + metadata + 用户行为 + SNS
- ✓ 功能（例）
  - 1) item → item list：e.g., 关联商品、关联视频、关联app、关联网站
  - 2) item set → item set list：e.g., 关联列表、关联专辑
  - 3) user → item list、item set list：e.g., 您可能喜欢的XXX
  - 4) user → user list、 user set list：e.g., 您可能感兴趣的XXX（人、群组）

考虑因素：

-- 用户是否需要？

-- 系统收益？

-- 数据是否支持？

-- ……

- ✓ 如何将推荐结果呈现给用户？
- ✓ 如何收集用户信息和反馈数据？

- ✓ 目的：
  - 提高用户满意度，达到推荐目的
  - 更多更好地收集高质量的用户反馈
  - 准确评测推荐算法效果

- ✓ 大规模存储
- ✓ 分布式计算
- ✓ 用户量、访问频次、峰值
- ✓ 实时响应的要求：
  - 毫秒级、秒级、小时级？
- ✓ 硬件资源的最大利用

- ✓ 优化准则：
  - 准确性、多样性、新颖性、覆盖率、时效性、……
- ✓ 数据预处理
- ✓ 离线算法
- ✓ 在线算法
- ✓ 功能实现策略
- ✓ 推荐解释
  - 对消费代价大的（时间、金钱）item尤其重要

- ✓ 上线前：基于人工标注评测集
- ✓ 上线后：
  - 基于用户点击数据
    - 将用户显示/隐式反馈转化为评测集
  - 基于A/B测试
    - 点击率、后续步长、转化率、……
    - 整体收益 VS. 各模块内部收益
      - 产品指标
      - 用户指标：高收益用户、低收益用户
    - 每个产品特性导致不同的评估指标
    - 如何评估用户需求满足度？

Baidu 百度

目录

推荐系统设计之策略篇

功能分析、数据分析、算法设计

用户群体、推荐功能

- ✓ 用户数：万 → 十万 → 百万 → 千万→ 亿
- ✓ 用户群体：低端/高端、大众/小众、职业、年龄……
- ✓ 推荐功能：
  - 推荐内容：资讯、视频、图片、……
  - 个性化？非个性化？
  - Session？Cookie？用户？
  - Top-N？列表浏览？
  - 实时反馈的更新：点击、收藏、喜欢、删除、换一批
  - 用户模型的更新：实时、小时级、天级、周级？

例：功能分析

- ✓ 例1：知道问题推荐
  - 用户：知道产品相对资深用户，各领域都有
  - 推荐功能：
    - 推荐内容：知道待回答问题
    - 是否个性化：针对特定用户的个性化推荐，和用户历史行为偏好相关
    - 展现形态：个人中心列表浏览 & 特定场景推送
    - 实时反馈：点击查看、回答
    - 时效性需求：固定周期更新 or 根据用户行为实时调整

- ✓ 例2：贴吧帖子推图片、视频
  - 用户：浏览该帖子的用户，可能是贴吧忠实用户或搜索带来的非贴吧用户
  - 推荐功能：
    - 推荐内容：帖子相关的图片或视频
    - 是否个性化：非个性化的关联推荐，每个用户看到的都一样
    - 展现形态：关联列表（文字标题+多媒体内容）
    - 实时反馈：点击查看
    - 时效性需求：固定周期更新（旧帖）or 实时关联计算（新帖）

推荐系统设计之策略篇

功能分析、数据分析、算法设计

数据类型

- ✓ Item
  - 内容：文本、图片、音频、视频
  - Ontology、tag
- ✓ 用户
  - profile
- ✓ 用户-item行为数据
  - 点击、收藏、删除、观看、评分历史

- ✓ 关键：各类数据是否充足？可用性如何？

例：item基础数据的重要性

界面1：

普罗米修斯
主演: 劳米·拉佩斯

机器人总动员
主演: 本·贝尔特


哈利·波特7：...
主演: 丹尼尔·雷...

变形金刚3：黑...
主演: 希亚·拉博夫

源代码
主演: 杰克·吉伦...

界面2：

普罗米修斯
主演: 劳米·拉佩斯

机器人总动员
主演: 本·贝尔特

哈利·波特7：...
主演: 丹尼尔·雷...

变形金刚3：黑...
主演: 希亚·拉博夫

源代码
主演: 杰克·吉伦...

界面3：

普罗米修斯　机器人总动员　哈利·波特7：...　变形金刚3：黑...　源代码

例：用户-item行为数据

- ✓ Explicit feedback
  - 评分、收藏、推荐/分享、购买、评论
- ✓ Implicit feedback
  - 点击浏览、下载、停留观看时间

- ✓ 理想：大量准确的Explicit 反馈
- ✓ 折中：用Implicit 反馈补充
- ✓ 问题：Explicit与Implicit数据的整合

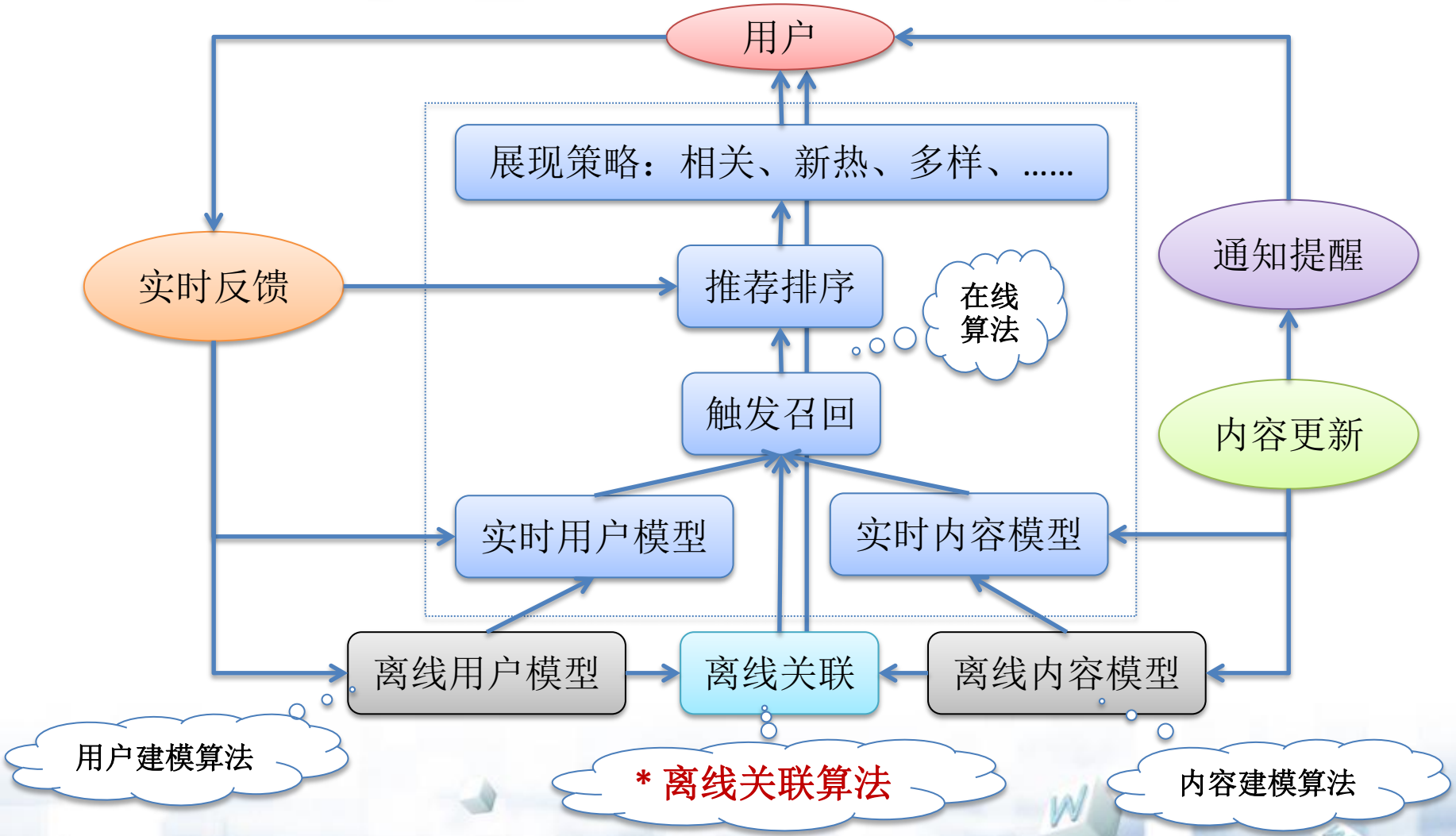
用户-item行为数据的价值

- ✓ 推荐算法设计与评估的基础
  - 数据充足，简单算法性能可以很好
  - 数据缺失，任何算法也不可能有好的性能
- ✓ 要求：不仅要吸引用户提供反馈，而且要吸引用户提供准确反馈
  - 给用户充足便利的反馈机会
  - 促使用户提供准确反馈/反馈鉴别机制
    - 购买行为：主动搜索购买 VS. 促销购买
    - 浏览行为：排行榜的强引导作用

推荐系统之策略设计

功能分析、数据分析、算法设计

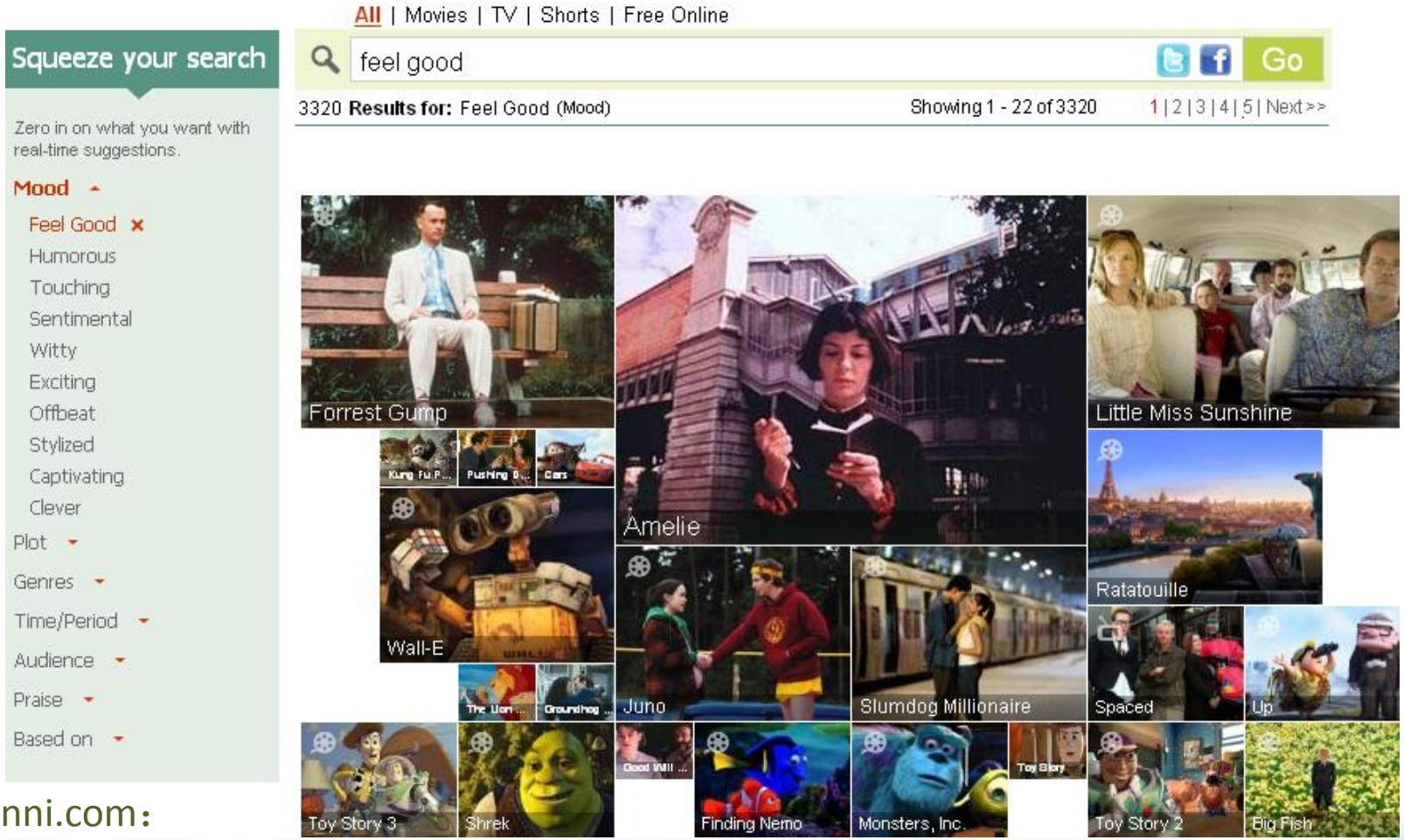
例：简单的个性化推荐流程图



算法选择：基于数据和功能

- ✓ 数据
  - 内容：文本、图片、音频、视频、……
  - Metadata：Ontology/类别信息、tag、……
  - 用户行为日志：点击、评分、……
  - SNS：好友关系、群组关系、……
- ✓ 同一个算法可实现不同功能；同一个功能可用不同算法实现
- ✓ 用户建模、内容建模：将用户、内容用特征向量描述
  - 属性、term、topic、……
- ✓ 离线关联算法：计算<用户-用户>/<用户-item>/<item-item>关联并排序
  - **关联/相似度计算**
    - 基于内容的：专家标注、ontology、tag、文本/音频/图像/视频、……
    - 基于用户行为的：统计方法、关联规则、相似度经验公式
    - 混合算法
  - **机器学习**
    - 协同过滤 ：knn、基于模型的、……
    - 各种经典算法：分类、回归、聚类、图算法、……

例：关联计算 – 基于内容的（专家标注）



jinni.com：
Movie Genome

基于内容的关联计算：解决冷启动的好办法

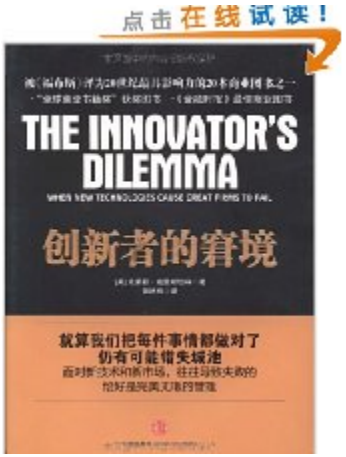
- ✓ 和其他领域紧密结合
  - 新闻、博客、… – 自然语言处理
  - 音乐 – 音频处理；图像 – 图像处理；视频 – 视频处理

  精度取决于相关领域的研究进展

- ✓ 专家标注：限于item数量少且有相对客观标准的领域
  - 电影 VS. 书籍
  - 自动专家发现？

- ✓ 可与Metadata结合：
  - ontology(量少准确)：商品分类
    - 分类排行榜：很土很有效的推荐列表
  - tag(量大不准确)

例：关联计算 – 基于用户行为的（统计）



创新者的窘境 [平装]

~ 克莱顿•克里斯坦森 (Clayton M. Christensen) (作者), 胡建桥 (译者)

(10 个用户评论)

市场价: ~~¥ 38.00~~

卓越价: ¥ 22.40 此商品可以享受**免费送货** 详情

为您节省: ¥ 15.60 (5.9折)

现在有货。

由**卓越亚马逊**直接销售和发货。

关键在于用户是否需要此功能
不在于算法简单或复杂

浏览此商品的顾客最终购买

81% 的顾客看完这一页后购买了此商品
创新者的窘境 - 克莱顿•克里斯坦森 (Clayton M. Christensen) 平装 (10)
¥ 22.40

7% 的顾客看完这一页后购买了
创新的艺术 - 汤姆·凯利(Tom Kelley) 平装 (11)

5% 的顾客看完这一页后购买了
重来:更为简单有效的商业思维 - 贾森•弗里德(Jason Fried) 平装 (47)
¥ 22.80

4% 的顾客看完这一页后购买了
设计心理学 - 唐纳德·A·诺曼(Donald Arthur Norman) 平装 (46)
¥ 19.50

关联计算 – 基于用户行为的（关联规则）

- ✓ 基本假设
  - 过去经常被一起频繁消费的商品，今后也会被一起消费
- ✓ 算法：
  - 根据事先确定的支持度、置信度、提升度等，计算关联商品
- ✓ 成熟的商业应用
  - 电信套餐定制、超市捆绑销售
- ✓ 特点：
  - 适合Session/Transaction数据
  - 难以对长尾商品作有效预测
  - 用户的消费差异性被忽略，不是很适合个性化推荐

关联计算 – 基于用户行为的（相似度公式）

- 将用户用item向量表示，或将item用用户向量表示
  - 向量上的取值可以是用户对item的评分或其他行为取值
- 常用的相似度计算公式（也可用于内容关联计算）
  - cosine

$$sim(i,j) = cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|i\| * \|j\|} = \frac{\sum_{u\in U} R_{u,i}R_{u,j}}{\sqrt{\sum_{u\in U} R_{u,i}^2}\sqrt{\sum_{u\in U} R_{u,j}^2}}$$

  - adjusted cosine

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_u)(R_{u,j}-\bar{R}_u)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_u)^2}}$$

  - pearson correlation

$$sim(i,j) = \frac{Cov(i,j)}{\sigma_i\sigma_j} = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_i)(R_{u,j}-\bar{R}_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_j)^2}}$$

关联计算 – 混合算法

- ✓ 关联融合
  - 数据融合 → 关联算法
  - 不同关联算法 → 结果融合
- ✓ 关联结果应用
  - 直接用于相关推荐
  - 个性化推荐：用户对特定item的偏好 → 关联扩展

✓ 优点：不依赖domain

	协同过滤	
基本假设	过去行为偏好相似的用户，今后行为偏好也相似	
基本思路	**基于近邻的** 为每个用户/商品计算相似用户/商品，再利用相似用户/商品的历史进行预测： 基于部分user-item关系 相似度计算 → k近邻 → 偏好预测	**基于模型的** 用隐变量刻画用户和商品间的关系： 部分user-item关系 → 用隐变量刻画user-item关系 →偏好预测
商业应用	在线购物网站，如Amazon的商品关联推荐	学术研究热点，实际应用不太普及
特点	**适于个性化推荐**	

例：基于模型的协同过滤 – MF & PLSA

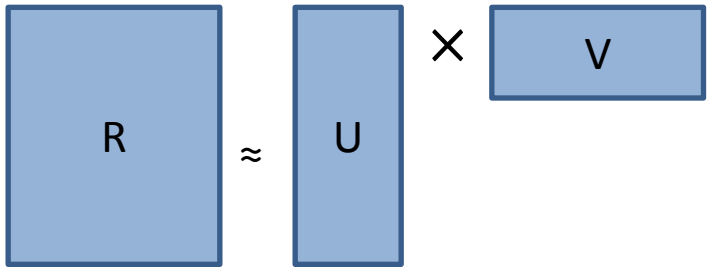
✓ MF（矩阵分解）

$$\mathbf{R} \approx \mathbf{U}^\top \mathbf{V}$$

$$\widehat{r}_{ui} = \sum_{k=1}^{K} u_{ku} v_{ki} = \mathbf{u}_u^\top \mathbf{v}_i$$

$$\min_{\mathbf{U} \in \mathbb{R}^{k \times m}, \mathbf{V} \in \mathbb{R}^{k \times n}} \sum_{(u,i) \in \mathcal{R}} (r_{ui} - \mathbf{u}_u^\top \mathbf{v}_i)^2 + \lambda (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)$$

SVD, set A=U Σ, B=V^T

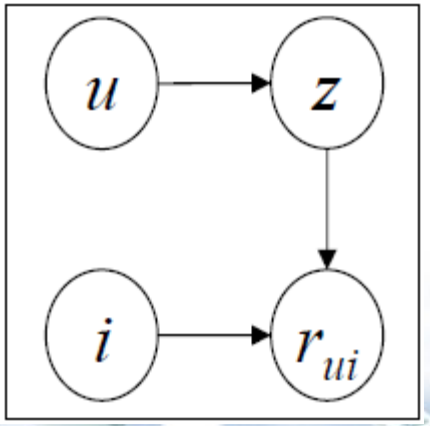


✓ PLSA（概率隐语义分析）

$$P(r_{ui}|u,i) = \sum_{z=1}^{k} P(z|u) P(r_{ui}; \mu_{zi}, \sigma_{zi})$$

$$P(r_{ui}; \mu_{zi}, \sigma_{zi}) = \frac{1}{\sqrt{2\pi}\sigma_{zi}} \exp\left[-\frac{(r_{ui} - \mu_{zi})^2}{\sigma_{zi}^2} \right]$$

$$\widehat{r}_{ui} = \sum_{z=1}^{k} P(z|u) \int r P(r; \mu_{zi}, \sigma_{zi}) dr = \sum_{z} P(z|u) \mu_{zi}$$



机器学习算法 -- 分类

✓ 将用户-item行为数据：<user, item, 点击>，转换为

	用户属性1	用户属性2	…	item属性1	item属性2	…	<u,i>属性1	<u,i>属性2	…	是否点击
<u1,i1>	a	II	…	A	t	…	22	0.4	…	1
<u1,i2>	b	III	…	B	p	…	587	0.8	…	0
…	…	…	…			…			…	…
<un,im>	a	I	…	B	m	…	31	0.01	…	1

- 可使用各种机器学习分类算法（把难点转换成属性构造和选择问题）
- 参考：个性化广告的CTR（点击率）预估模型
  - 但是，很多推荐不是只以提升CTR为目标
  - CTR提升，用户满意度不一定提升（吸引眼球的推荐不一定是好的推荐）

机器学习算法的难点：优化目标的多样性

- ✓ 一般算法的优化目标：相对单一和明确
  - 分类：分类错误率
  - 信息检索：准确率、召回率、……
- ✓ 推荐系统
  - 不同功能的优化目标不同；不同发展阶段的优化目标不同
  - 不同优化目标可能需要不同的数据和算法
  - 同一算法在不同数据集上效果差异很大，数据在不断变化
    - Item/user比，新老用户比，稀疏度，时效性
  - 推荐系统本身影响收集的用户行为数据
    - 推荐列表 -> 用户点击 -> 根据点击数据优化推荐列表

算法设计小结

80% -> 90% √

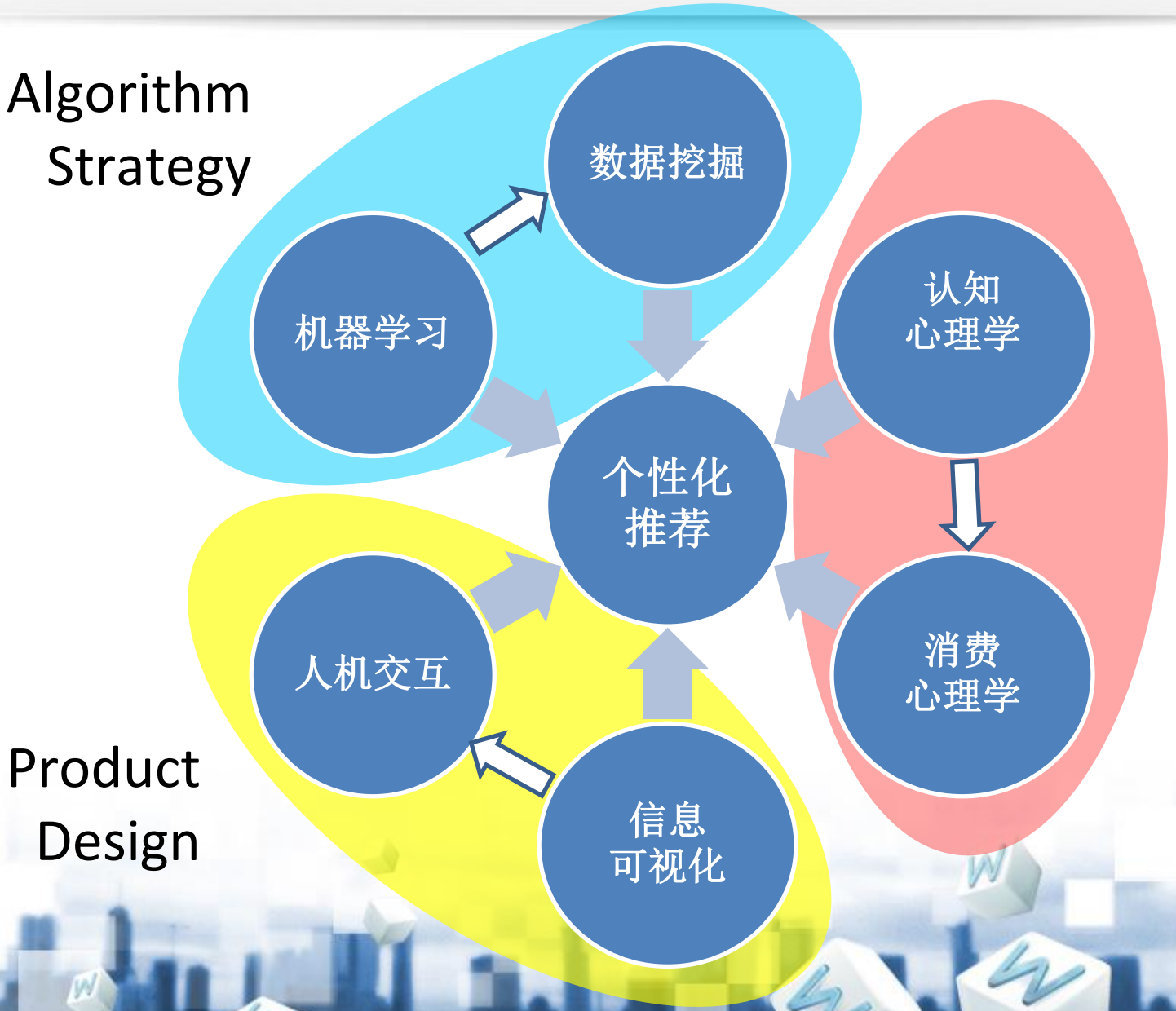
10% -> 20% × → **换一种方法吧~~**

不能因为手里有把锤子，就把所有问题都当作钉子~~

推荐系统之策略设计

一些感受

（1）木桶效应


Algorithm
Strategy
数据挖掘
机器学习
认知
心理学
个性化
推荐
User
Behavior
消费
心理学
人机交互
Product
Design
信息
可视化

（2）数据为王

- ✓ “某某说：XX 算法效果很好/不好”通常意味着“某某有/没有适合该算法的数据”
  - 例：“基于SNS的算法效果很好”、“内容关联不靠谱”
- ✓ “在现有数据上优化”VS. “寻找更多的数据”
  - 不同阶段的重点
- ✓ 数据清理&整合

（3）领域知识

- ✓ 推荐策略和产品业务紧密耦合 → 领域知识的大量使用
  - – 关联定义
  - – 数据处理
  - – 特征构建
  - – 推荐解释

- ✓ “通用推荐引擎” VS. “垂直推荐引擎”
  - – “通用系统平台+归一化数据+算法” + “垂直策略设计”

谢谢！