百度推荐系统实践

RecSys Engineering

姚旭

推荐与个性化部

2013

背景介绍

Background

以搜索

满足用户的主动表达的需求

以推荐

挖掘并满足用户的潜在需求

Baidu 百度

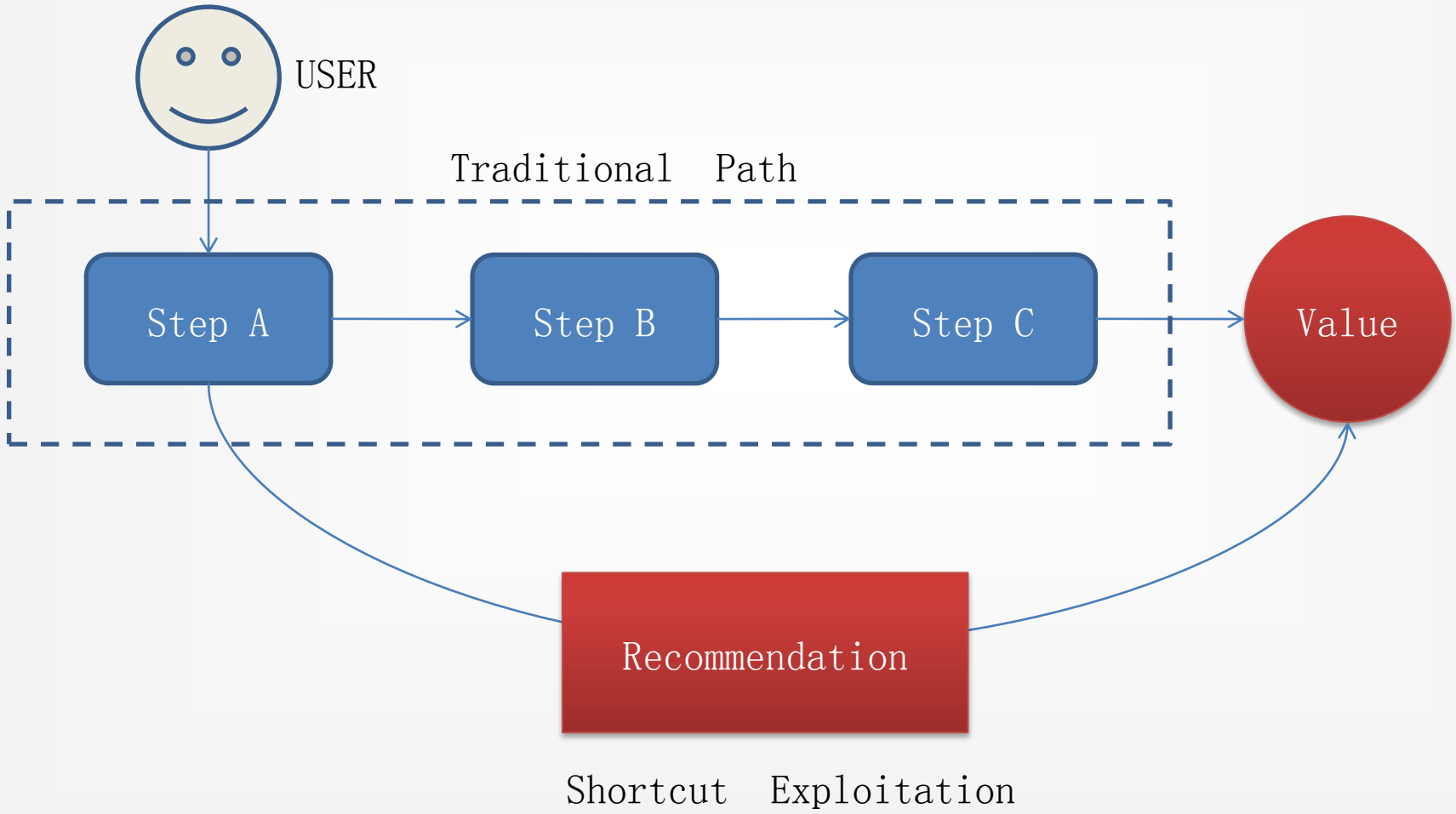
背景介绍

Background

精确需求	泛需求	潜在需求
搜索	搜索 + 浏览	浏览

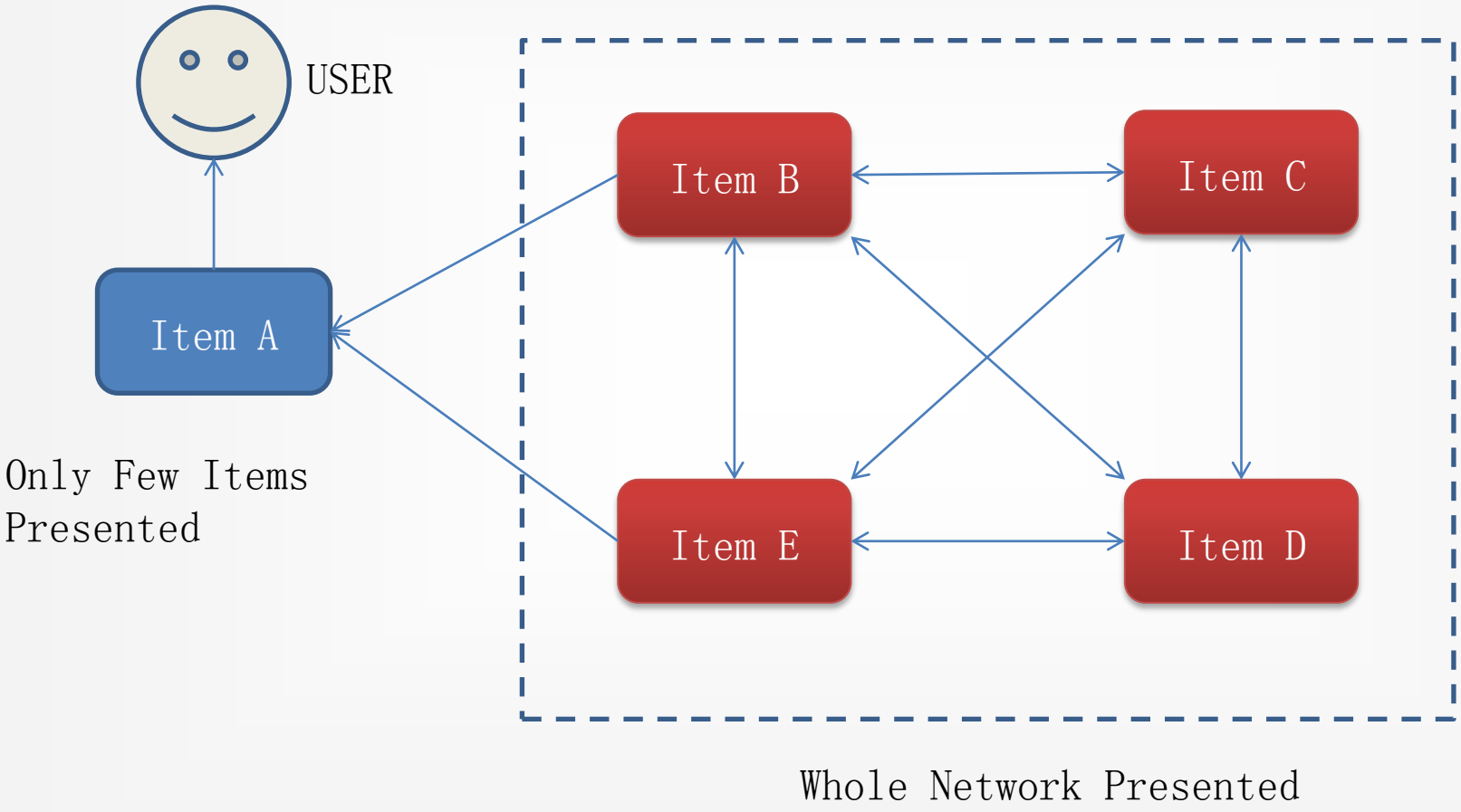
核心功能之一：过程优化

RecSys Engineering



核心功能之二：兴趣发现

RecSys Engineering



产品应用

RecSys Engineering

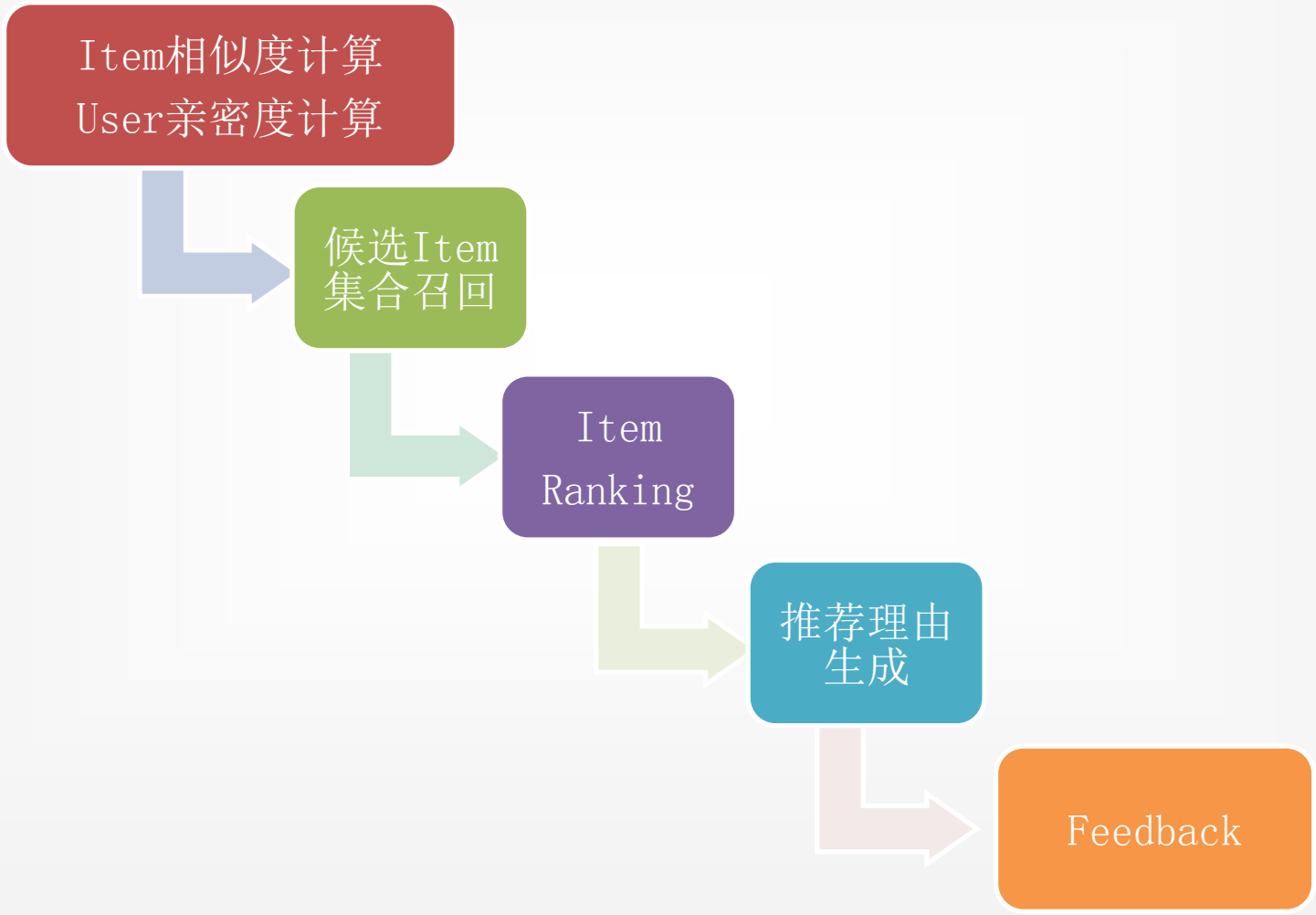
	Item数量级	稀疏性	多样性	时效性需求	反馈速度
电影	100,000	1%	低	中	中
音乐（单曲）	1,000,000	3%	低	低	快
网络小说	1,000,000	<0.5%	中	中	慢
APP	100,000	<1%	中	中	中
资讯	10,000,000	<0.1%	高	高	快
短视频	10,000,000	<0.1%	高	高	快
文档	10,000,000	<0.1%	高	低	快

基础架构

RecSys Engineering

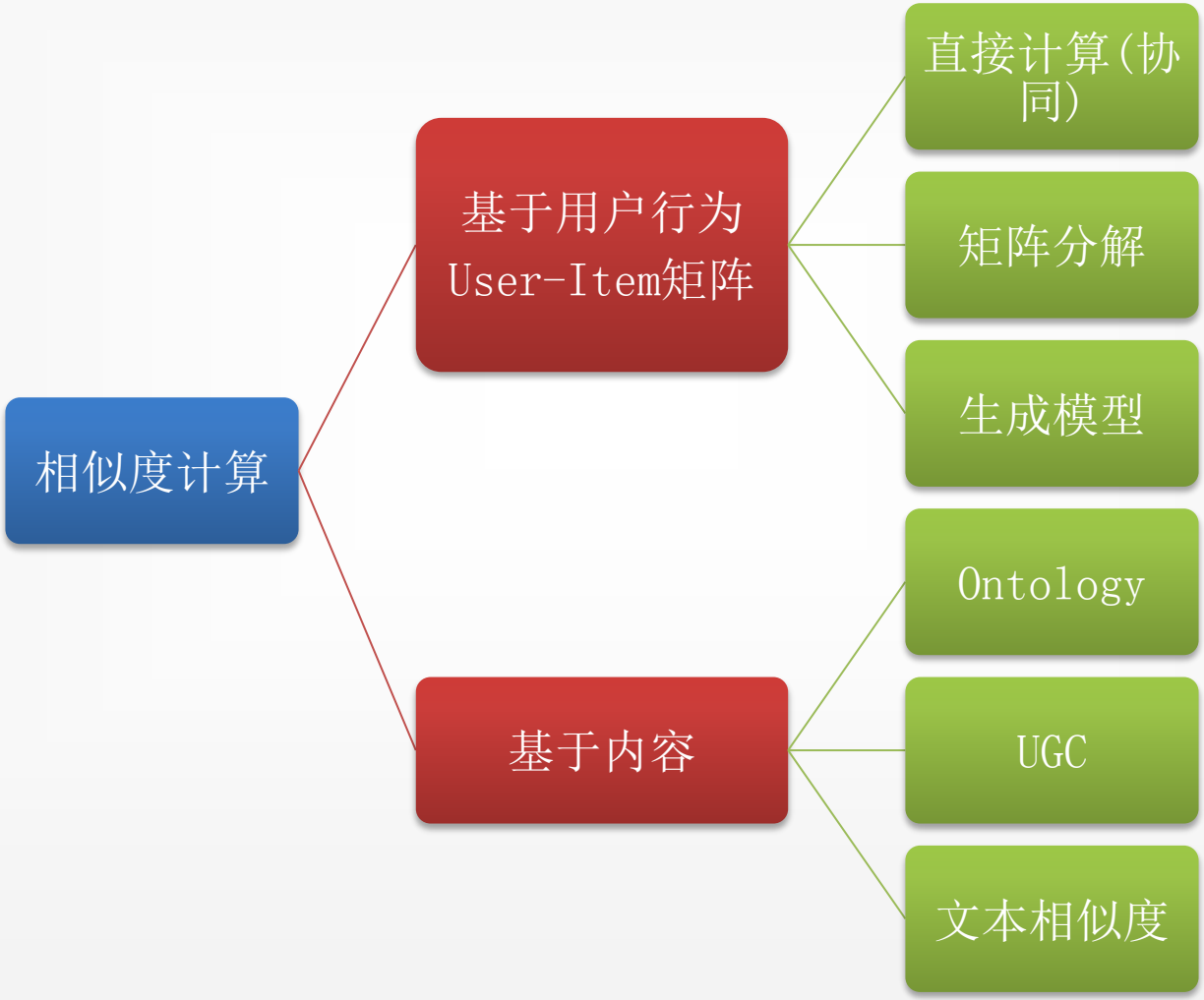
基础架构

RecSys Engineering



相似度计算

RecSys Engineering



基于内容推荐

RecSys Engineering

优点

- 无需依赖用户数据，回避产品初期用户不足和数据稀疏性问题
- 覆盖率高

缺点

- 数据建设成本大，不具有推广性
- 人对于内容理解的多样性，多层次

基于邻域推荐

RecSys Engineering

优点

- 利用群体智慧，无需依赖背景知识
- 通用性高

缺点

- 强依赖于用户行为数据
- 数据稀疏性问题

相似度计算

RecSys Engineering

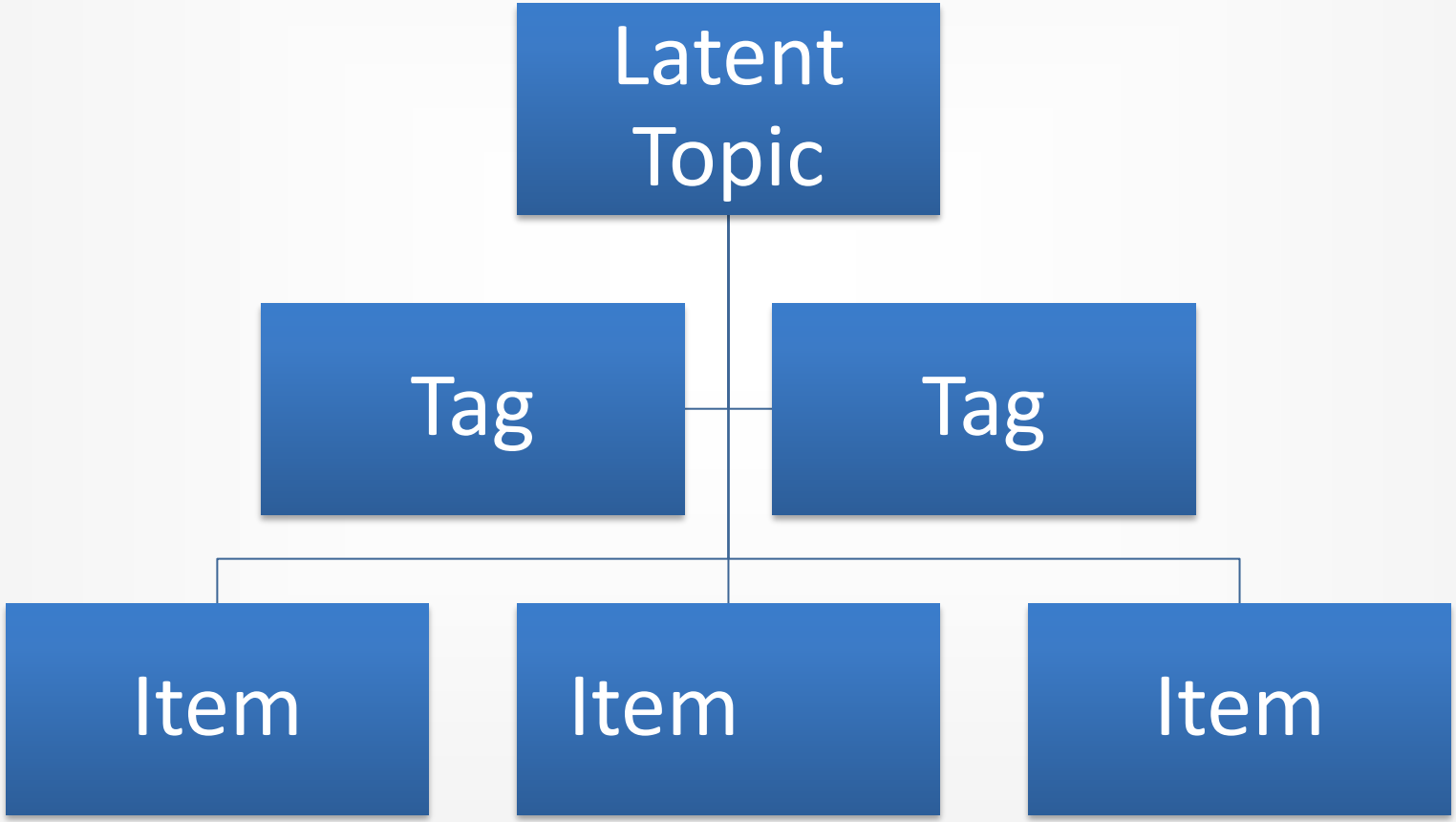
盗梦空间：诺兰 莱昂纳多 科幻 动作 悬疑 剧情

基于内容：
X战警3
蝙蝠侠前传2：黑暗骑士
蝙蝠侠前传1：侠影之谜
我是传奇
X战警：第一战
神秘代码
关键第四号
勇士
源代码
美国队长

基于用户行为（无修正）：
十二生肖
人再囧途之泰囧
101次求婚
古惑仔
喜爱夜蒲2
一代宗师
血滴子
禁闭岛
肖申克的救赎
V字仇杀队

相似度计算

RecSys Engineering



相似度计算

RecSys Engineering

相同点

- 将Item映射到一个表示空间上，使相似度计算结果最优
- 在不丧失区分度的情况下，空间上尽量稠密
- 经验目标：稀疏度>1%
- 横向结合+纵向结合

不同点

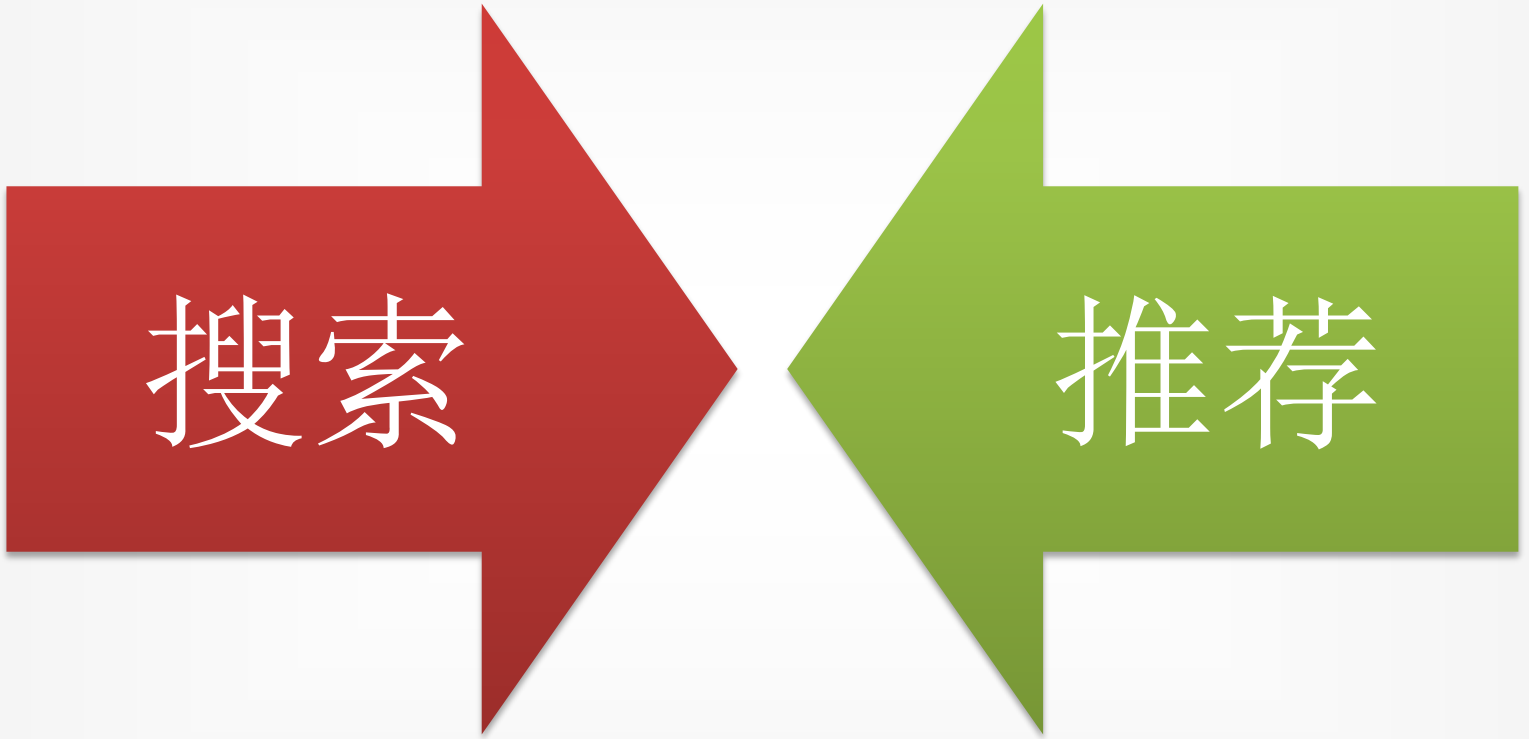
- 基于统计 VS 基于知识
- 黑盒策略 VS 白盒策略
- 推荐理由的可理解性

工业界 VS 学术界

RecSys Engineering

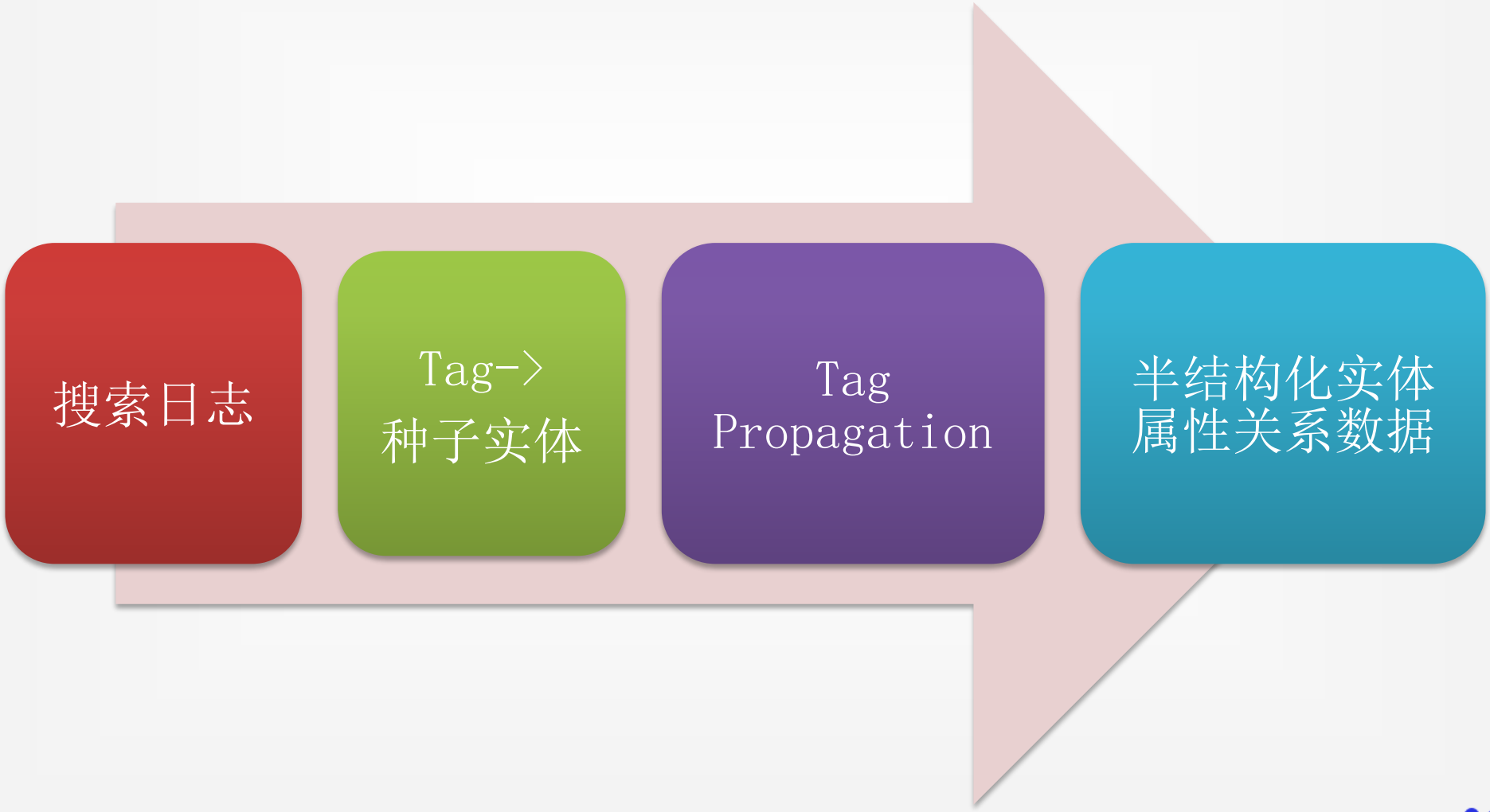
特征

RecSys Engineering



基于搜索日志的实体属性挖掘

RecSys Engineering



基于搜索日志的实体属性挖掘

RecSys Engineering

男主腹黑的小说
凤囚凰
盗情
绮梦璇玑
芊泽花
且试天下
兔子爱吃窝边草
昏前婚后
老公是腹黑大人

关于分手的电影
失恋33天
前度
那些年，我们一起追的女孩
和莎莫的500天
阿郎的故事
立春
李米的猜想
一声叹息

美食漫画
美食的俘虏
深夜食堂
将太的寿司
妙手小厨师
美食侦探王
中华小当家
日式面包王
料理仙姬

基于搜索日志的实体属性挖掘

RecSys Engineering

优点

- 主流垂直领域全覆盖
- 通用的挖掘方法
- 表述习惯和推荐目标用户保持一致
- Item的关键属性和推荐目标用户保持一致

缺点

- 需要做一定的数据清理，可直接作为中间数据，直接应用于产品使用时还需要再加工
- 单层扁平结构，缺乏层级关系，缺乏语义关联

基于搜索日志的实体属性挖掘

RecSys Engineering

盗梦空间： 高智商 心理 哲学(?)

高智商电影
盗梦空间
电锯惊魂
禁闭岛
恐怖游轮
猫鼠游戏
搏击俱乐部
致命魔术
记忆碎片
沉默的羔羊
穆赫兰道

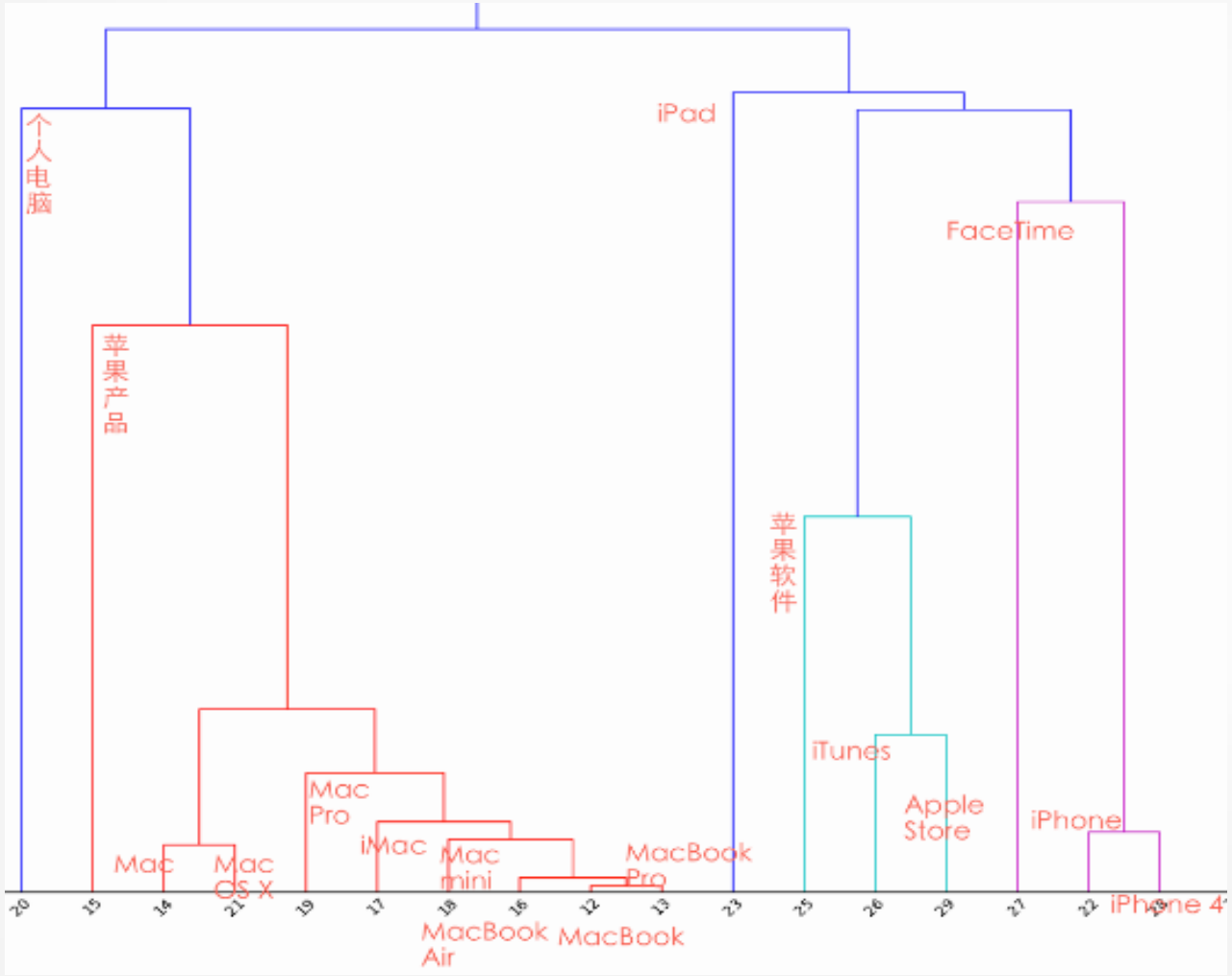
心理电影
少年派的奇幻漂流
盗梦空间
寒战
唐山大地震
花与蛇
电锯惊魂
入殓师
禁闭岛
海洋天堂
告白

哲学电影
少年派的奇幻漂流
盗梦空间
三傻大脑宝莱坞
普罗米修斯
入殓师
大鱼
荒野生存
楚门的世界
本杰明巴顿奇事
闻香识女人

Baidu 百度

基于搜索日志的实体属性挖掘

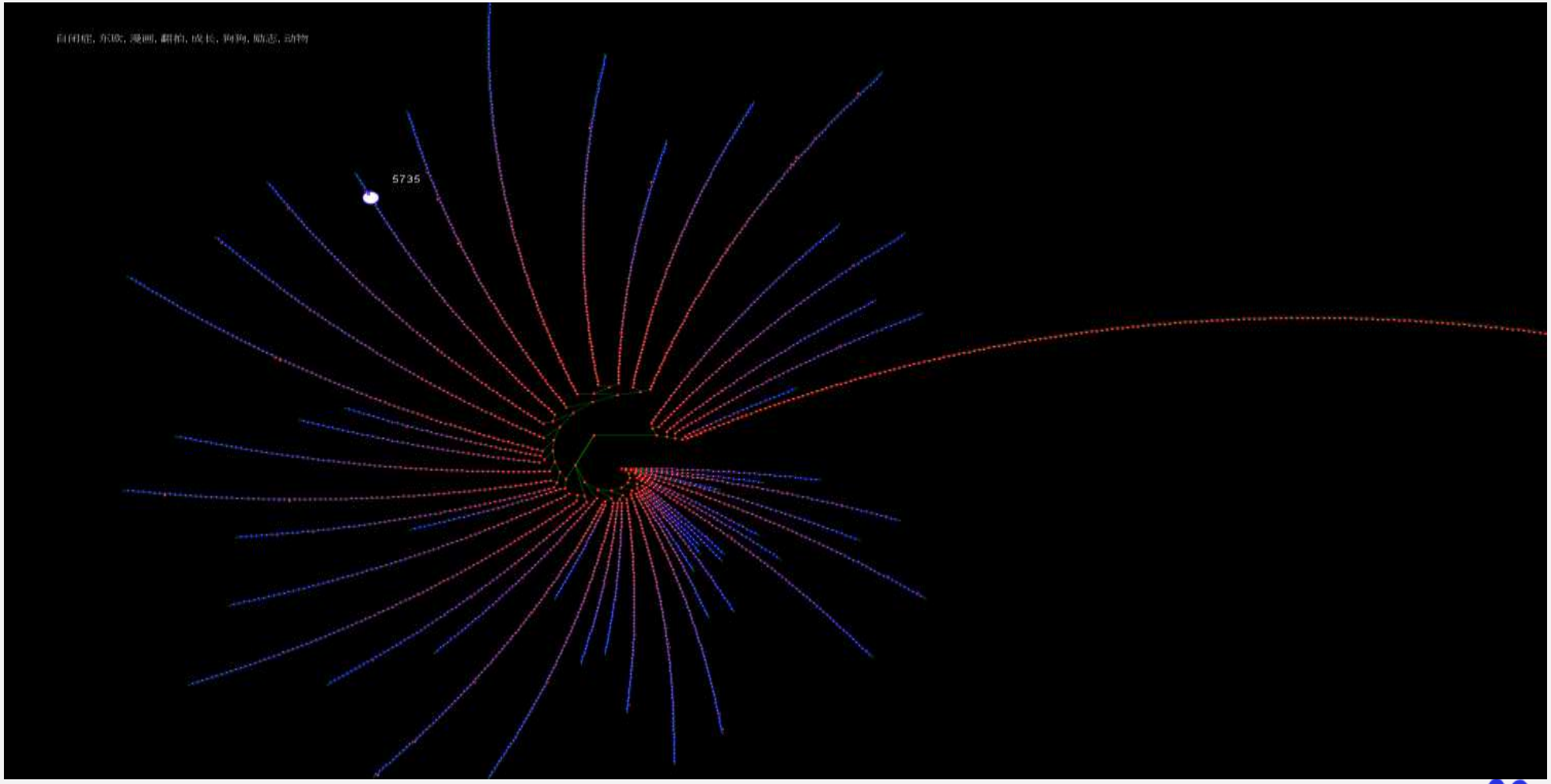
RecSys Engineering



Baidu 百度

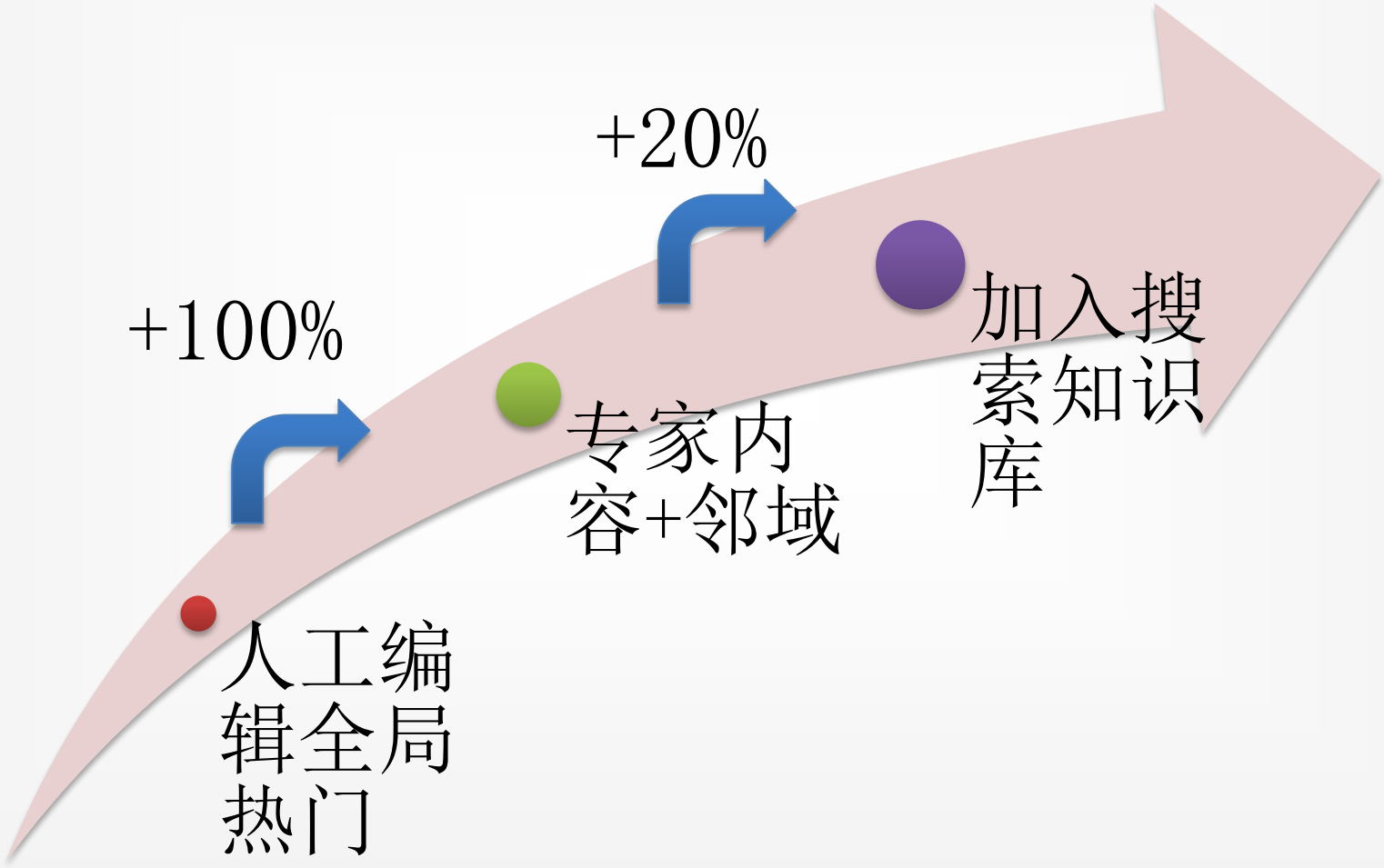
基于搜索日志的实体属性挖掘

RecSys Engineering



应用举例：线上效果

RecSys Engineering

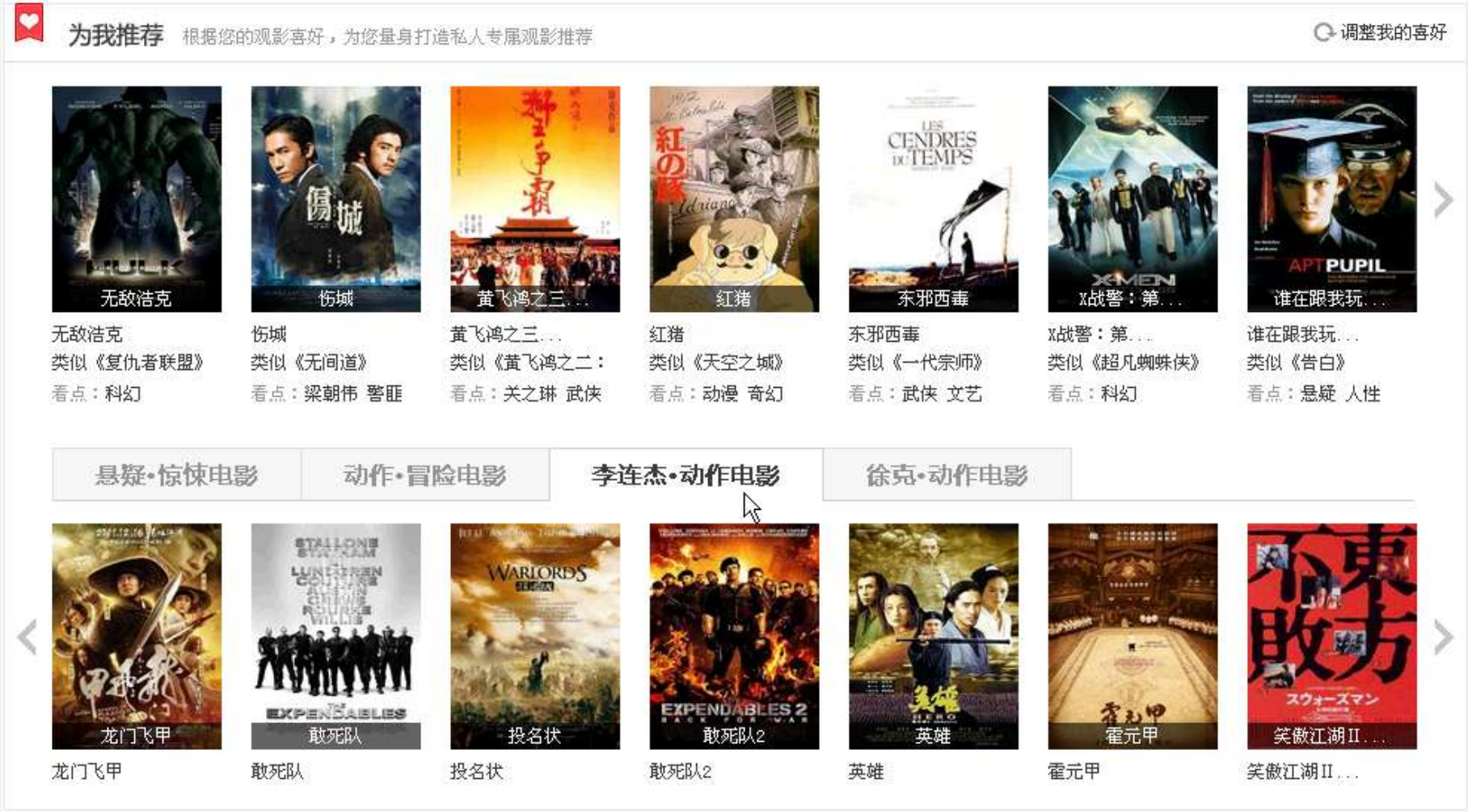


应用举例：关联推荐

RecSys Engineering

应用举例：个性化推荐

RecSys Engineering



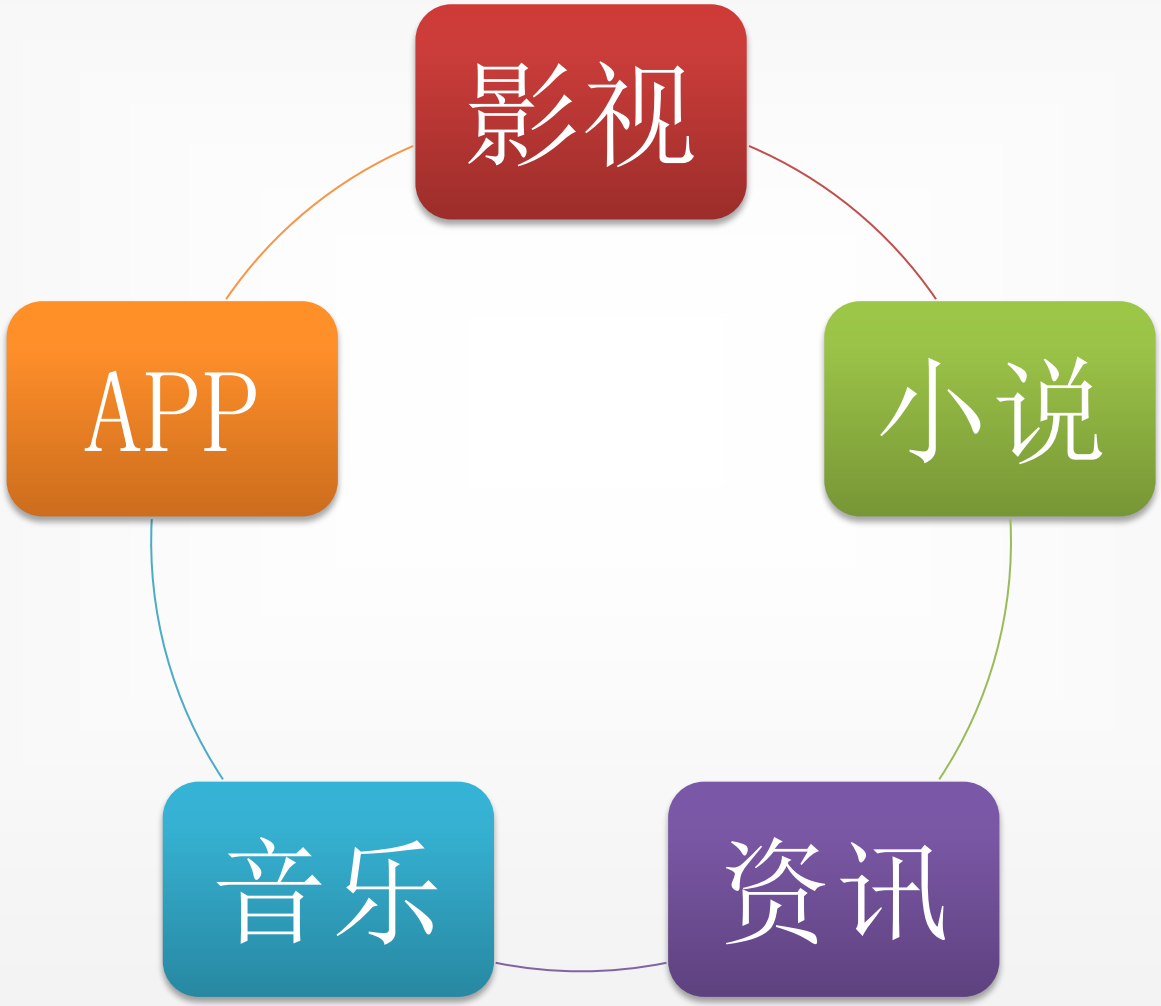
应用举例：泛需求搜索 搜索扩展

RecSys Engineering

Bai du 百度

应用举例：

RecSys Engineering



Tips

RecSys Engineering

- 基于统计 VS 基于知识库
- 优化特征 VS 优化模型
- 推荐理由生成 VS 推荐结果计算
- Item粒度 VS Cluster粒度
- 一步到位 VS 探索式发现
- 文本特征 VS 多媒体特征

百度推荐系统实践

RecSys Engineering

Q&A

微博：@姚旭_百度推荐

求贤：

Baidu 百度

百度技术沙龙

关注我们：t.baidu-tech.com

资料下载和详细介绍：infoq.com/cn/zones/baidu-salon

“畅想•交流•争鸣•聚会”是百度技术沙龙的宗旨。 百度技术沙龙是由百度与InfoQ中文站定期组织的线下技术交流活动。目的是让中高端技术人员有一个相对自由的思想交流和交友沟通的的平台。主要分讲师分享和OpenSpace两个关键环节，每期只关注一个焦点话题。

讲师分享和现场Q&A让大家了解百度和其他知名网站技术支持的先进实践经验，OpenSpace环节是百度技术沙龙主题的升华和展开，提供一个自由交流的平台。针对当期主题，参与者人人都可以发起话题，展开讨论。

InfoQ 策划·组织·实施

关注我们：weibo.com/infoqchina