推荐系统@淘宝

魏虎(空望) 2012.7
http://weibo.com/skyhope
kongwang@taobao.com

主要内容

- 推荐系统概念
- 淘宝的数据
- 淘宝推荐系统应用场景
- 淘宝推荐系统核心算法
- 淘宝推荐系统的设计

- 推荐系统概念
- 淘宝的数据
- 淘宝推荐系统应用场景
- 淘宝推荐系统核心算法
- 淘宝推荐系统的设计

推荐系统定义

- 维基百科：form or work from a specific type of information filtering system technique that attempts to recommend information items (item, music, books, news, images etc.) or social elements (e.g. people, events or groups) that are likely to be of interest to the user.

推荐系统作用

- 从用户角度：
  提高用户忠诚度
  帮助用户快速找到宝贝
- 从网站角度：
  提高网站交叉销售能力
  提高成交转化率
- 好的推荐系统更像一个有经验的网站导购员

分类

- 个性化推荐
- 非个性化推荐

推荐与搜索

- 相同点：

  帮助用户找到商品

- 不同点：

  搜索是通过用户主动输入的关键字进行查询。

  推荐则是用户在浏览网站的过程中，不一定需要用户输入，根据当前网页的上下文进行个性化的信息输出。

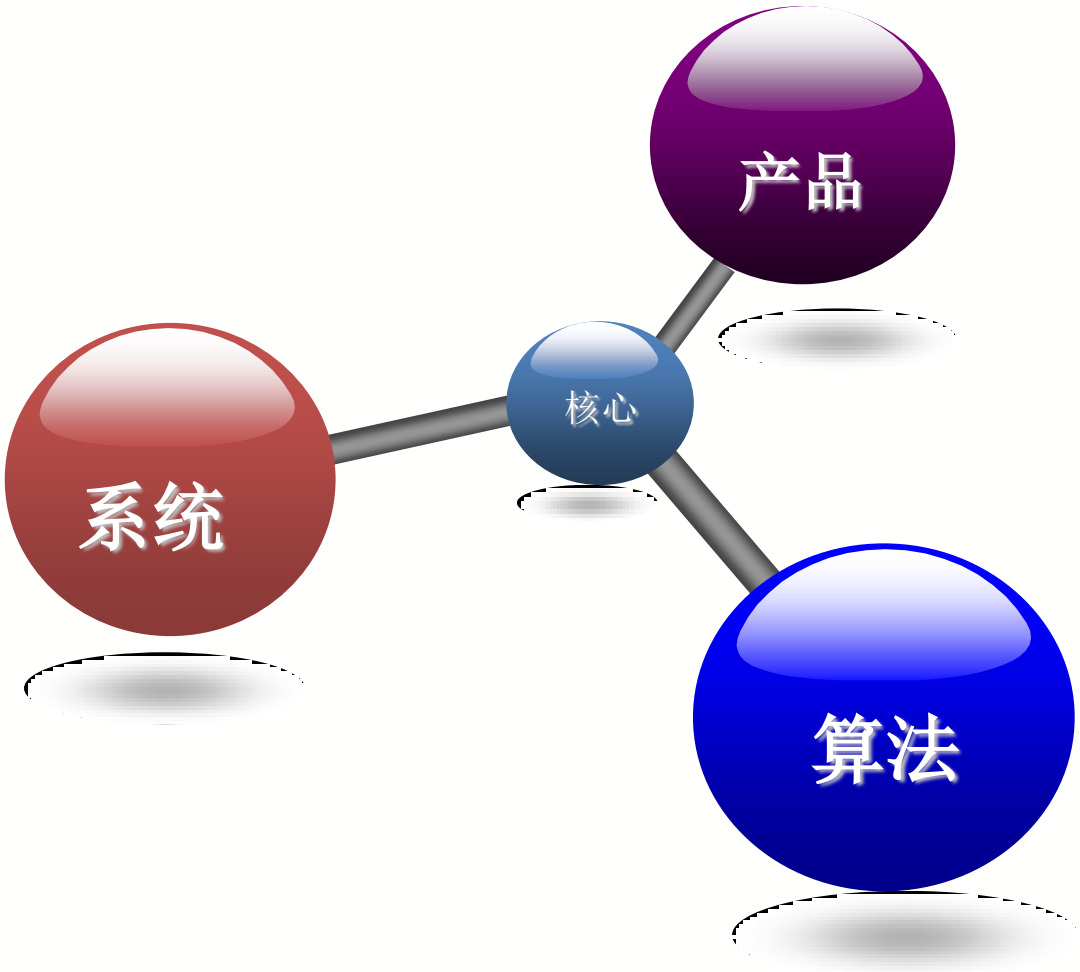
推荐与广告

- 相同点：
  - 基于用户行为
- 不同点：
  - 广告目的是帮助商家推广商品等
  - 推荐系统帮助用户找到想要的商品等

推荐与SNS

- 相同点：它们都有基于人群的共同点产生推荐
- 不同点：一个是机器，一个是人工

推荐系统的核心



推荐系统产品

- 同类或者相关商品、店铺推荐
- 买了还买、看来还看等
- 猜你喜欢
- 群体信息披露
- 热门排行榜
- etc

大型网站推荐系统组成

- 数据
- 算法（online & offline）
- Messaging system
- Search engine
- NoSQL
- 分布式计算
- 效果评测

数据

- explicit（显式）：能准确的反应用户对物品的真实喜好，但需要用户付出额外的代价
  用户收藏
  用户评价

- Implicit（隐式）：通过一些分析和处理，才能反映用户的喜好，只是数据不是很精确，有些行为的分析存在较大的噪音
  用户浏览
  用户页面停留时间、访问次数

算法

- 算法计算方式

  离线： 用户类目偏好、用户购买力分析、关联性分析、相似矩阵计算等等

  在线：排序、过滤、增量计算

- 算法配合大量业务规则
- 没有最好，只有更好！

Messaging system

- 大型系统不可或缺的重要组成部分
- 与其他系统解耦，消息转发

Search engine

- 文本分析 抽取关键词
- 作为推荐系统的一个信息检索技术 内容相关性匹配

NoSQL

- 简单
- 高性能
- 方便定制

分布式计算

- 大规模数据统计和运算
- 大数据集合的ETL

MapReduce , Hive、Hadoop

主要内容

- 推荐系统概念
- **淘宝的数据**
- 淘宝推荐系统应用场景
- 淘宝推荐系统核心算法
- 淘宝推荐系统的设计

淘宝数据特点

- 数据量巨大

数百万店铺
数亿激活用户
数亿的在线商品
数十亿的收藏信息
……

- 商品问题

同一类商品多个卖家
标类 非标类
类目属性正确性
恶意收藏、刷信誉

主要内容

- 推荐系统概念
- 淘宝的数据
- **淘宝推荐系统应用场景**
- 淘宝推荐系统核心算法
- 淘宝推荐系统的设计

应用场景

目前覆盖大小场景60多个，主要包括

- Detail 浏览了还浏览
- 收藏夹弹出层推荐
- 购物车弹出层推荐
- 已买到宝贝 你可能感兴趣
- 淘宝无线应用
- EDM（重复购买提醒）
- 各个垂直频道
- 个性化list排序
- 开放平台api
- 。。。

淘宝推荐产品

- 淘宝业务产品丰富，推荐功能穿插其中
- 推荐功能涵盖的范围更广
- 很多场景推荐算法与业务规则相关

主要内容

- 推荐系统概念
- 淘宝的数据
- 淘宝推荐系统应用场景
- **淘宝推荐系统核心算法**
- 淘宝推荐系统的设计

淘宝推荐系统算法

- 基础算法

  聚类算法，预测算法，分类算法等，主要用于产生基础知识库

- 推荐算法

  content-based，collaborative-based，Association Rules等等

基础算法

- 预测算法
  logistic 回归，通过以点击率为目标，以商品，卖家等因素作为指标，建立预测模型构建淘宝优质宝贝库
- 分类算法
  朴素贝叶斯
  商品性别判断（男性，女性，中性）
  用户性别判断

- 聚类算法
  人群，用户细分
  用于降维

推荐算法

- 基于内容推荐

  通过给用户和商品标注Tag，通过内容匹配算法，推荐商品给用户

优点：简单，搜索引擎支持，解决部分冷启动问题

缺点：难以区分商品信息的品质,而且不能为用户发现新的感兴趣的商品,只能发现和用户已有兴趣相似的商品

推荐算法

- 协同思想

  优点：新奇特，个性化程度高

  缺点：冷启动，稀疏性

推荐算法

- 关联规则
  类目的相关性、商品相关性、人的相关性

效果评测

- 推荐系统的效果需要数据来评测

Offline: 给定输入输出，验证系统的输出

Online : ABTest

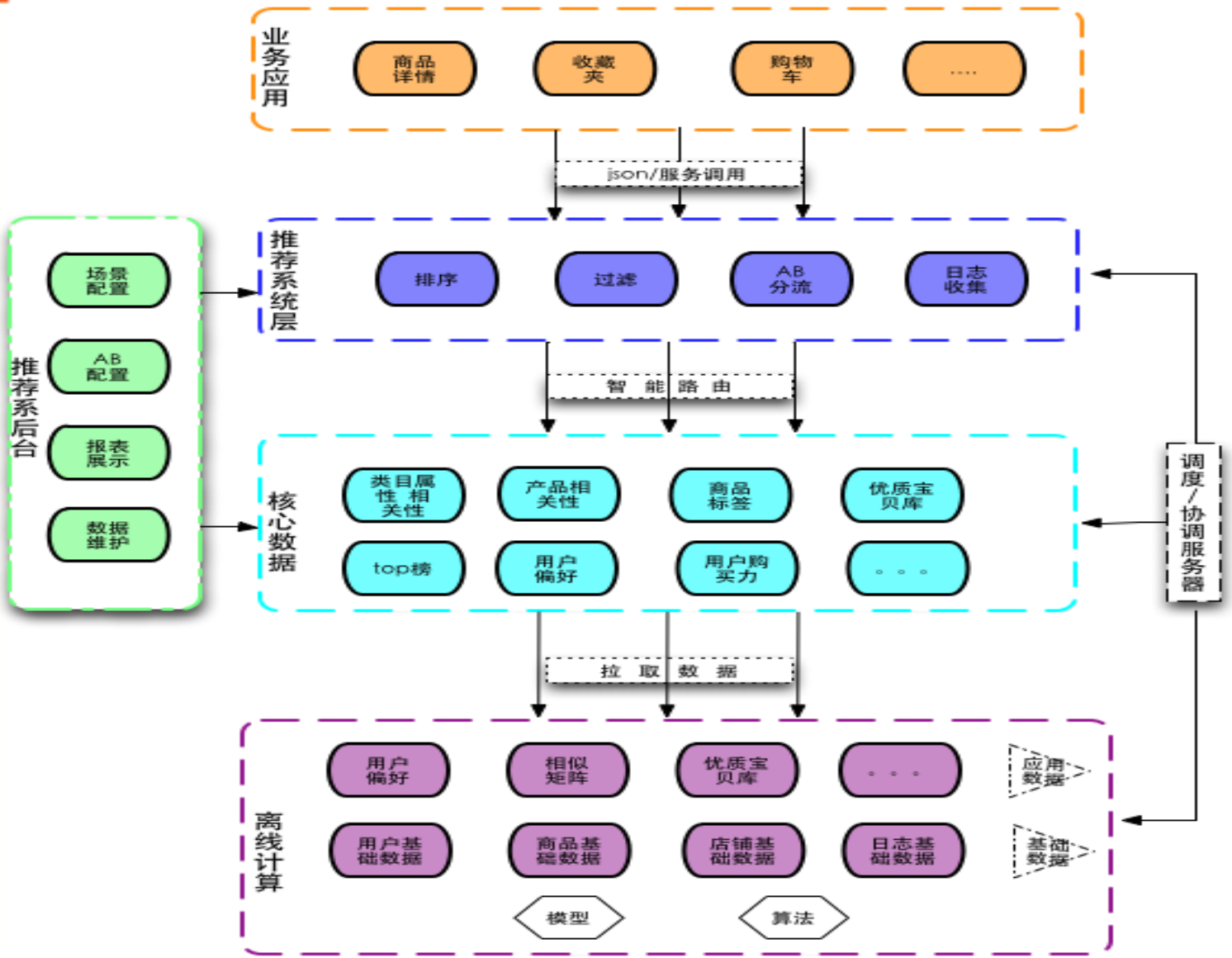
衡量指标 CTR GMV 转换率

主要内容

- 推荐系统概念
- 淘宝的数据
- 淘宝推荐系统应用场景
- 淘宝推荐系统核心算法
- **淘宝推荐系统的设计**

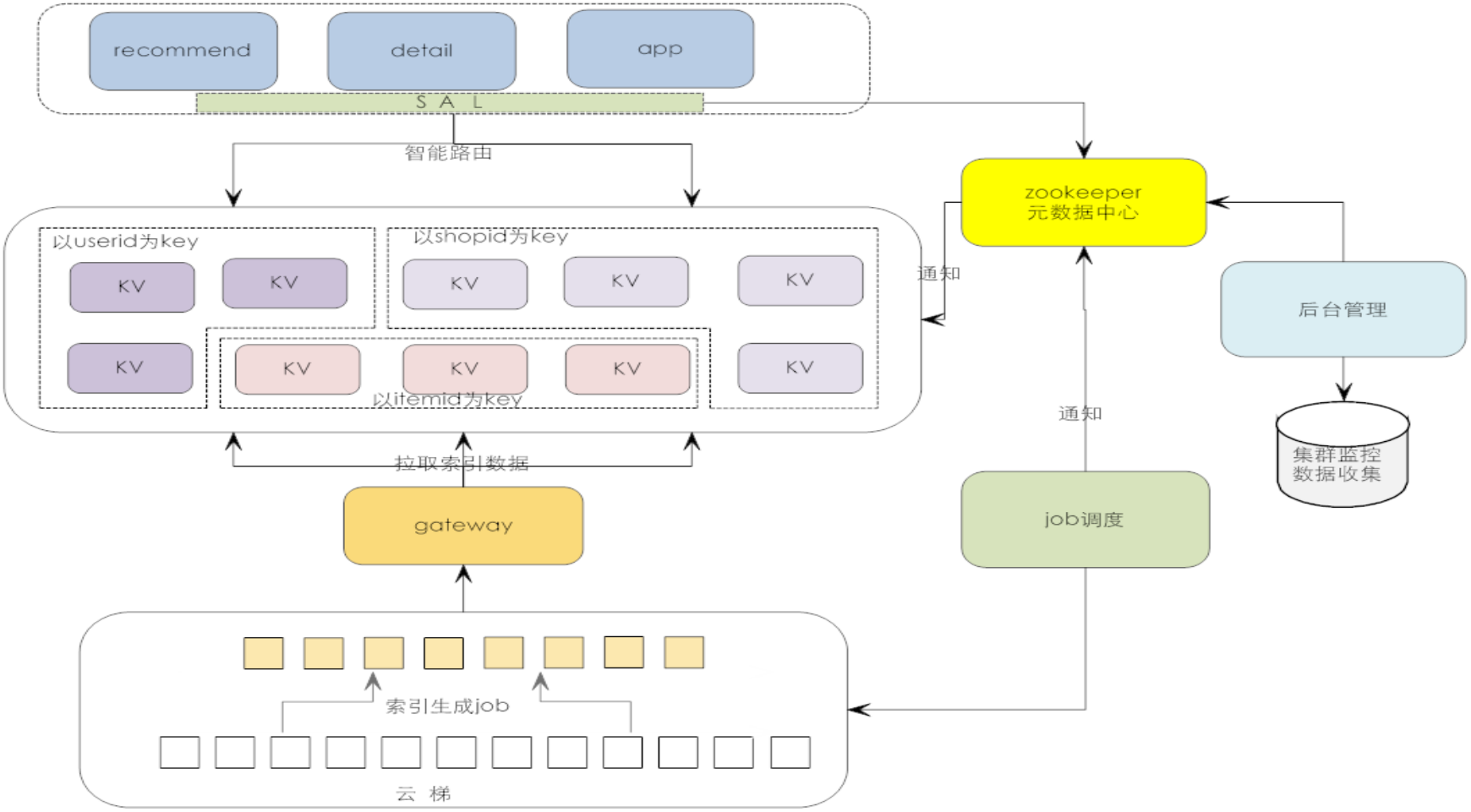
淘宝推荐系统设计

- 提供统一的平台管理各个推荐模块
- 提供高性能分布式存储
- 提供算法的AbTest和效果统计
- 提供灵活算法配置


业务应用
商品详情
收藏夹
购物车
....
json/服务调用
推荐系统层
排序
过滤
AB分流
日志收集
智 能 路 由
推荐系后台
场景配置
AB配置
报表展示
数据维护
核心数据
类目属性 相关性
产品相关性
商品标签
优质宝贝库
top榜
用户偏好
用户购买力
。。。
调度/协调服务器
拉 取 数 据
离线计算
用户偏好
相似矩阵
优质宝贝库
。。。
应用数据
用户基础数据
商品基础数据
店铺基础数据
日志基础数据
基础数据
模型
算法

分布式存储

treasure2系统结构



Treasure存储的数据

- 存储云梯（hadoop）上对用户、商品等原始数据分析的结果
- 云梯周期性同步，无实时更新
- 为推荐系统提供ABTest存储支持
- 可直接存储部分推荐算法的结果供推荐使用
- 动态部署

调度系统

- 负责周期性云梯（hadoop）任务调度
- 分布式
- 生产者 消费者

协调系统

- Zookeeper集群
- 智能路由
- 线上与线下联动通知
- Job依赖通知

- 推荐系统是需要不断创新
- 推荐系统与场景和行业相关

谢　谢