用户画像系统实践

王富平

1号店精准化部架构师

“我想强调的是，同一个人有多样的自画像。与其追求照相般的相似性，不如深入地 发掘相似处”—梵高

用户画像的定义

使用标签来量化用户特征属性，达到描述用户的目的

用户画像难点

1、数据源
2、业务结合
3、动态更新

假设现有用户画像有性别、地域两个属性，你将如何使用？

1、分析不同性别的群体特征，做特定营销

2、分析广州、北京用户的群体特征，做特定营销

3、分析90后、80后的群体特征，做特定营销

分类—聚类
迈出个性化的第一步，用户画像的应用开始

1号店建立用户画像的初衷，来自于《千人千面》项目，简言之：分析不同群体特征，针对群体进行推荐调整。

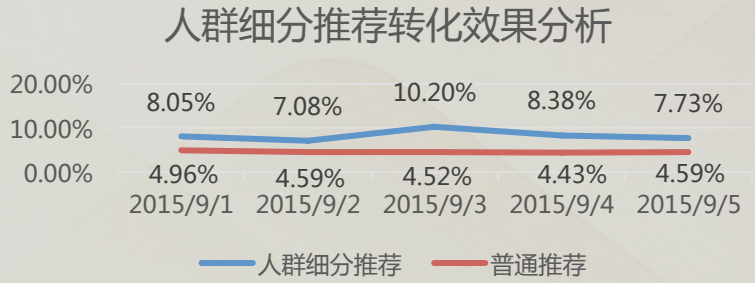
典型的群体有：小区、学校、公司等

补充用户命名实体识别的标签

公司、小区、校园标签：

用户群体	数量
公司	覆盖3558家公司，591个行业
小区	覆盖293个城市的4.26万个小区
校园	覆盖全国1334所高校

校园、小区千人千面引擎优化上线



完整的地址处理系统包含三部分：

- **地址结构化**
- **命名实体识别**

公司名识别模型的F1值（提高到80.6%）

- **地址匹配**

1号店从零开始打造了自己的用户画像系统,包含了用户标签画像、用户偏好画像。经历了全量版画像、Storm版实时画像、电商用户标签画像等慢慢演进和完善，在两年的时间里，遇到了性能瓶颈、数据质量评估、用户标签的膨胀、画像在精准化营销等应用场景的摸索，一步步成长，在推荐系统发挥了巨大作用。

用户标签画像

基本特征	社会身份	顾客用户生命周期	类目偏好	购物属性	风险控制
• 性别 • 母婴年龄预测 • 顾客消费层级 • 顾客年龄 • 地域气候	• **家庭用户** • **学生** • **公司白领** • 中老人 • 顾客职业的行业	• 注册用户转新客 • PC转移动 • 类目半新客转化 • 流失得分	• 果粉 • 吃货 • 高品质生活 • 家庭日用品 • 手机数码达人 • 礼物礼券	• 跨区域购买用户 • 日用品周期购买 • 顾客价值得分 • 促销敏感 • 辣妈、丽人	• 黄牛小号判别得分 • 注册异常用户判别得分 • 积分获取异常用户得分

类目标签（主题推荐）

女装	饼干/糕点	茶叶	流行首饰	身体护理	公共
• 甜美文艺 • 职业通勤 • 个性街头 • 妩媚性感 • 气质名媛	• 三高人群 • 瘦身减脂 • 独爱花香 • 香甜 • 鲜咸	• 清热解暑 • 补血益气 • 清肝明目 • 呵护女性 • 健胃消食	• 恋恋深情 • 卡通图案 • 平安 • 乔迁 • 金饰	• 抗敏感 • 滋润型 • 中草药 • 清香型 • 防晒隔离	• 儿时回忆 • 懒人必备 • 便携旅游 • 送礼必备 • 宴会待客

挑战

a）亿级画像系统实践和应用

b）记录和存储亿级用户的画像，支持和扩展不断增加的维度和偏好，毫秒级的更新，支撑个公司性化推荐、广告投放和精细化营销等产品

怎么做到的

1. 用户画像算法模型不断优化
2. 引入Storm等实时技术
3. 主题推荐标签、用户命名实体等新增标签补充进画像
4. HBase的离线和在线分离、Hbase的KV读和Solr的批量读分离、region热点监控和切分
5. 数据流不断优化
6. 数据存储改进

第一版画像现状

偏好系统包括类目偏好和导购属性偏好两个部分，第一版的偏好系统接口调用数每天达千万次，主要服务于推荐栏位和EDM。但改版的偏好系统存在性能低下，偏好得分分布不合理之类的问题。详情如下：

- 运行一次全量的数据更新太慢
- 用户的偏好得分数据分布不合理，得分呈多波峰分布，且在[6.0，8.0]区间的得分数目几乎为0
- 用户强偏好和弱偏好的阈值界限未有明显规定
- 用户未产生新的行为，兴趣偏好分值将不会发生变化（未按时间进行衰减）

新版画像系统流程

画像模型优化1

偏好画像的得分应满足三个条件：

- 用户在此类目或导购属性上的操作越多，得分越高
- 用户对类目或导购属性的喜好程度不同，可以通过偏好得分区间体现
- 用户的历史行为应有衰减

对于类目偏好，需先将用户对类目偏好离散化提高某些场景性能，最简单的行为可划分为两档【喜欢|一般】。

参数调整原则：

- 衰减系数的设置满足两个月衰减一半

（结合用户在不同类目下的购买周期，见下页）

- 各类行为权重之间的比例设置等同于用户各种行为数目的比例
- 偏好得分分布应与用户对类目的权重分布一致

画像模型优化2

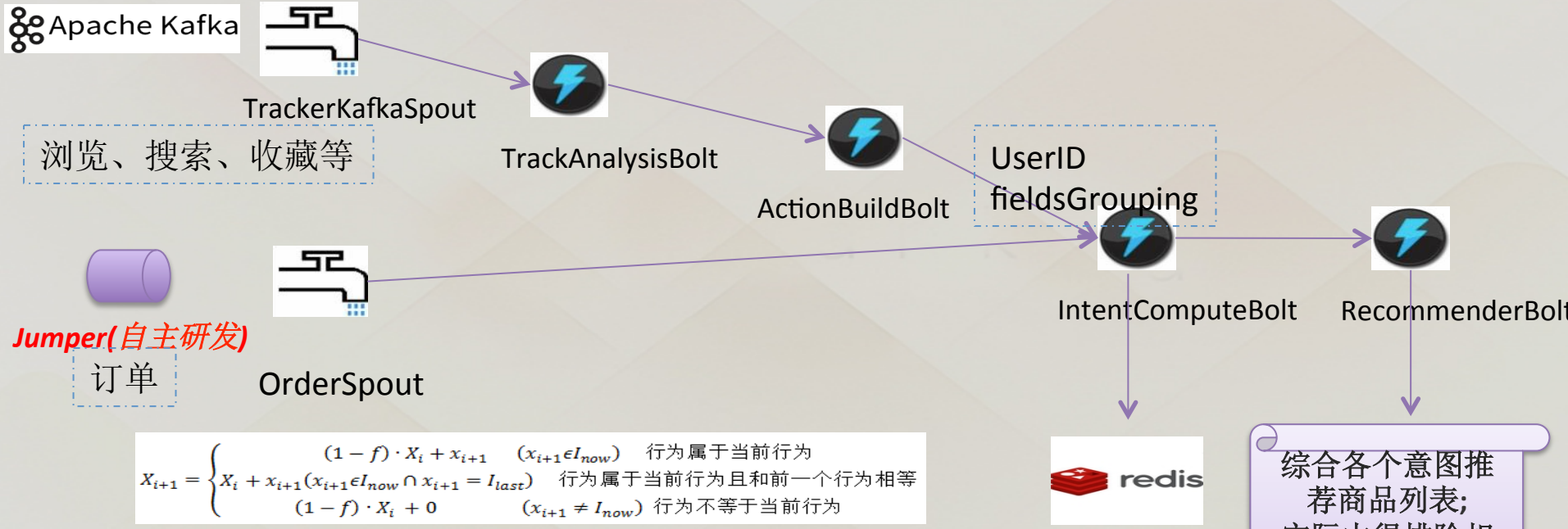
用户不同类1目的购买周期

商品类目

- 母婴
  - 奶粉系列
    - 牛奶粉
      - 1段牛奶粉
      - 2段牛奶粉
      - 3段牛奶粉
      - 4段牛奶粉
  - 尿裤/湿巾
    - 拉拉裤
    - 纸尿裤
      - undefined
    - 婴儿湿巾
  - 洗护清洁
    - 清洁液/衣物洗护/柔
      - 宝宝洗衣液
- 美容护理
  - 女性护理
    - 卫生巾
    - 护垫
  - 口腔护理
    - 牙膏1
  - 美发护发/假发

用户周期

引入实时Storm



$$X_{i+1} = \begin{cases} (1-f)\cdot X_i + x_{i+1} & (x_{i+1} \epsilon I_{now}) \quad \text{行为属于当前行为} \\ X_i + x_{i+1} (x_{i+1} \epsilon I_{now} \cap x_{i+1} = I_{last}) & \text{行为属于当前行为且和前一个行为相等} \\ (1-f)\cdot X_i + 0 & (x_{i+1} \neq I_{now}) \quad \text{行为不等于当前行为} \end{cases}$$

主题推荐标签

主题和标签的映射关系：

主题	类目	标签	关键词
数码极客	手机通讯	高端	高端 有档次 上档次 大气 高贵 贵族 金属质感 金属机身 做工精细 奢华 拉风
数码极客	手机通讯	性价比高	低调 价廉物美 物有所值 价格合理 价格公道 强烈推荐

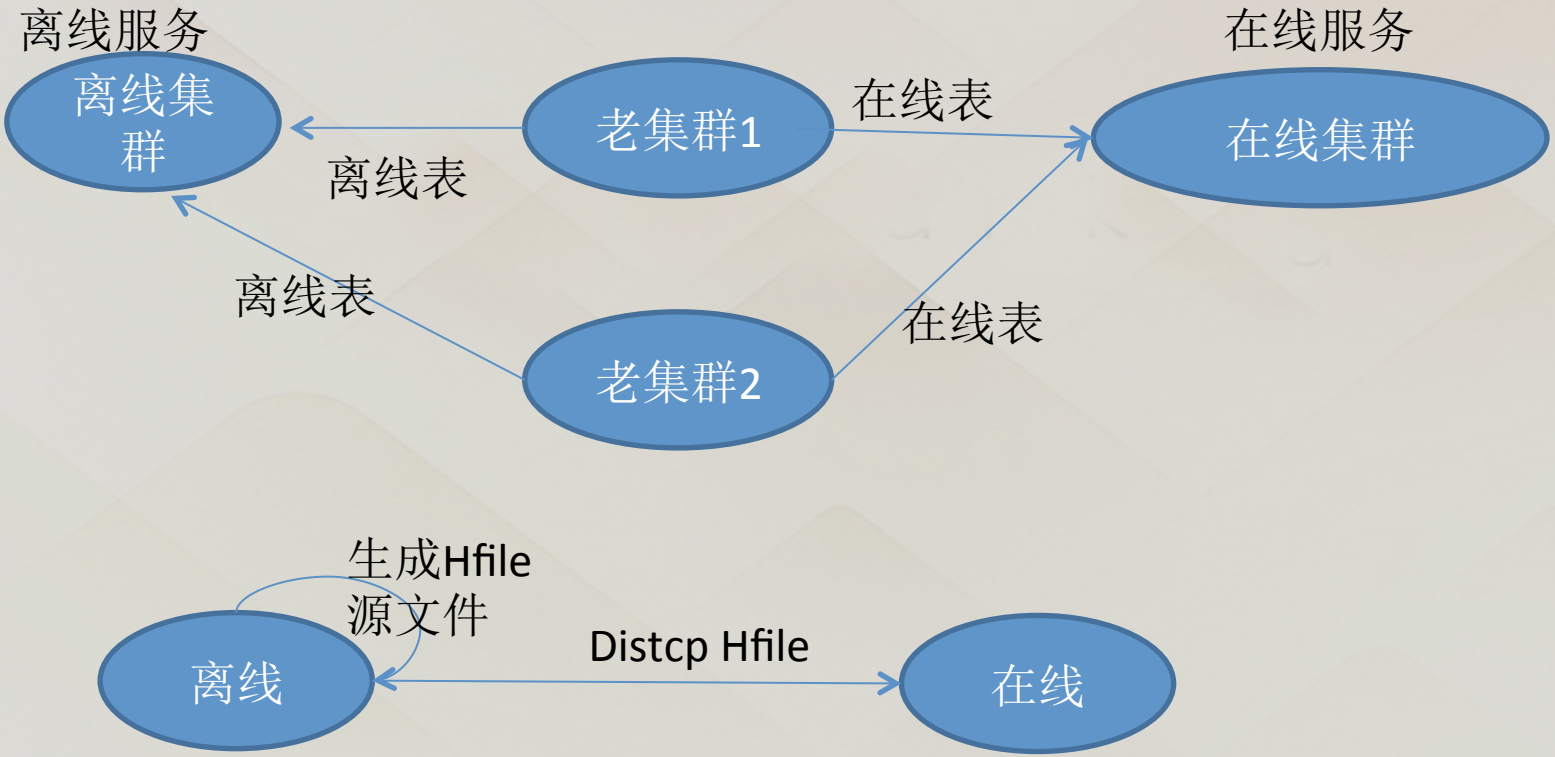
使用标签表中的关键词列表，结合商品的评论、标题数据给商品打标签。
商品打标签公式为：

$$score(producti,tagj)=\sum k=0\uparrow n\blacksquare titleW*isFind+commentW*count(tagj)/countAll+attributeW*isFind$$

用户打标签公式为：

$$score(user\downarrow i,tag\downarrow j)=weight(tag\downarrow j)*\sum prd\in prdset(user\downarrow i)\cap prdset(tag\downarrow j)\uparrow\blacksquare score(prd,tag\downarrow j)*weight(action)*decay\uparrow day$$

HBase的离线和在线分离

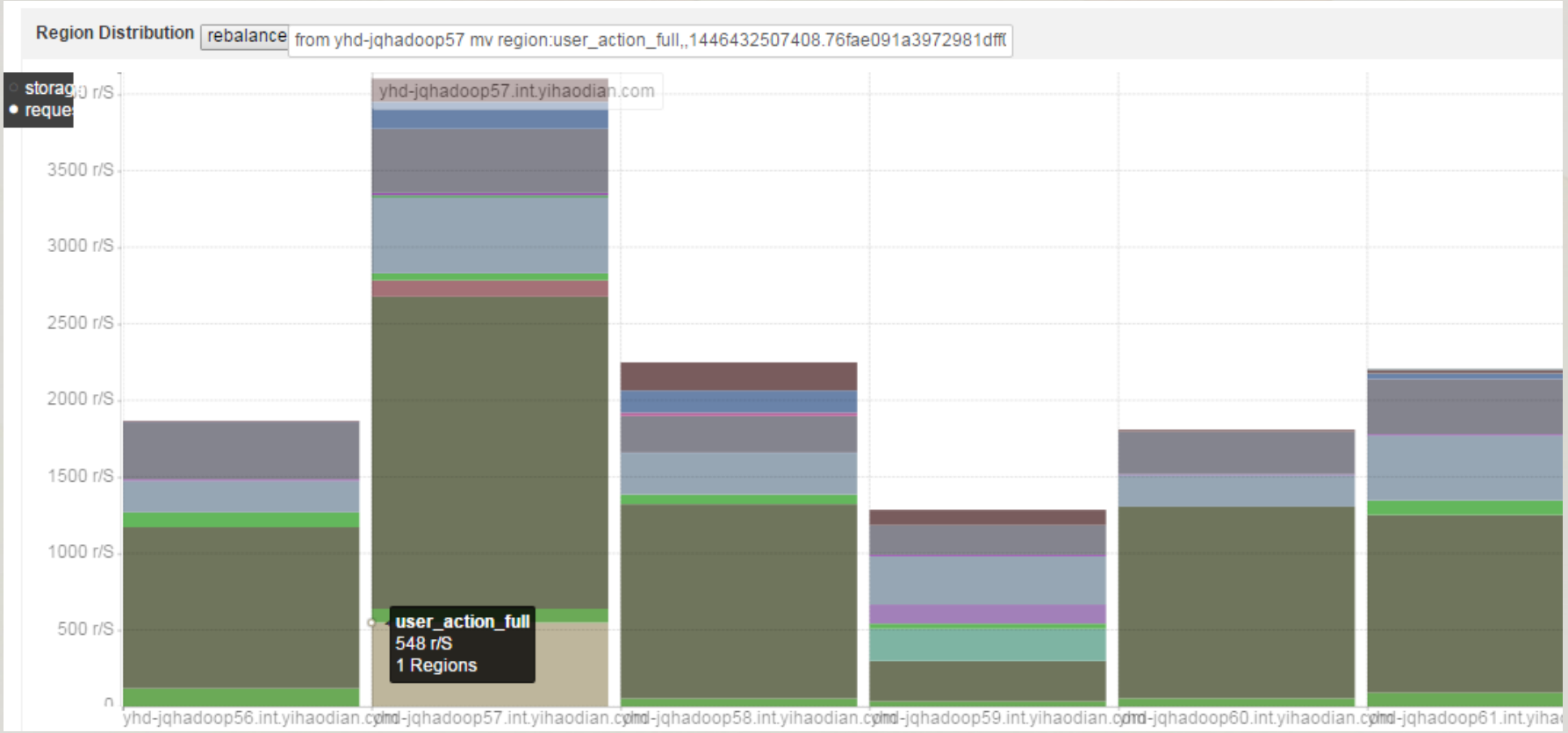


Solr解决批处理选人

Hive，Solr和HBase满足不同的使用场景

调优相关表，提高读写性能

根据画像表每一台机器的热点，迁移或者切分



数据流优化

- guid和userid的对应关系中，滤掉公用电脑和黄牛账户（全国有20万左右人从事刷单产业链）
- 为了进一步提高离线部分的计算速度，牺牲算法精确性，用户的行为权重计算亦可以增量计算

设Wh为用户对某个类目的历史行为权重，Wc为用户最新一天的行为权重，则总的行为权重

$Wt = \lambda Wh + Wc,\ 0 < \lambda < 1$

如果采用上述方法，则不必遍历用户的所有的行为数据，每次更新时，只需遍历一天的数据即可

优化数据存储

用户行为和行为统计表HBase替换为Hive，最后的画像表保留为HBase;
考虑到类目偏好使用比较频繁，而导购属性偏好数据量远大于类目偏好，解耦来将两者分开存储;

类目偏好离线数据结构-Hive

字段名	类型	备注
userid	string	用户id
category_id	bigint	类目id
category_level	int	类目层级
weight	double	得分

全量数据过滤

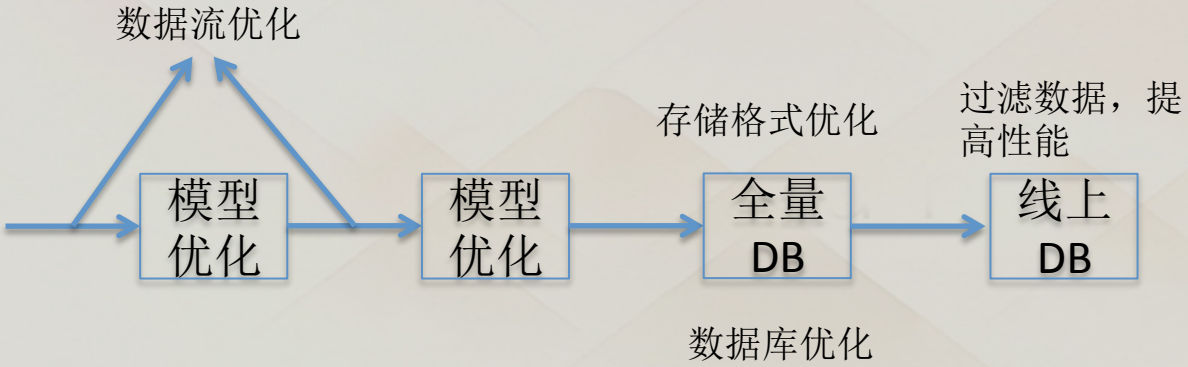
类目偏好离线的全量数据进行过滤之后，导入在线部分。过滤原则:

- 每个用户的偏好类目数量小于一个固定值
- 用户偏好得分大于下限，该下限可假设用户当天在某个类目只有一个加车行为，然后带入模型反推出来

导购属性偏好离线的全量数据进行过滤之后，导入在线部分。过滤原则:

- 属性偏好大于一个固定的下限
- 属性值的数量小于一个上限
- 属性值偏好大于一个固定下限

主要优化和改进点



- 长期兴趣和短期偏好解耦
- 类目和属性不同画像偏好解耦

And

曾经还尝试过什么但失败了/放弃了?

- 实时画像和离线画像融合，实时的权重融合进离线画像，最后权重算法过重，最终选择实时和离线画像分开。
- 中间过程全部采用HBase存储

未来想做：

- 使用HBase 镜像双集群
- Apache Ignite+ HBASE：提高在线服务集群的稳定性和速度
- 标签的分层治理

案例ROI分析

画像系统使得公司广告投放ROI提升3%；
画像（意图）对猜你喜欢栏位的贡献占比60%多
DMP和选人系统的核心部分
应用到首页猜你喜欢、团购、闪购、搜索、推荐、营销等栏位或者产品；
了解受众群体的变迁，适时推出适合的产品；
降低自营商品的采购数量，指导了厂商优化产品结构；

栏位覆盖率统计（ 11.02~11.08 ）：

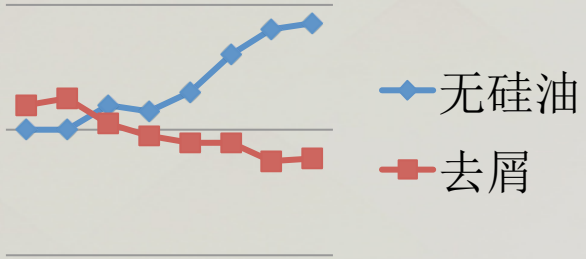
终端类型	页面	栏位	推荐算法B	推荐算法C	推荐算法D
APP	首页	1贵就赔	用户画像	热销补余	
		算法覆盖率	44.1%		
APP	首页	价比JD低	用户画像	指定CE类目选品	热销补余
		算法覆盖率	5.4%		
APP	首页	精选团购	用户意图相关分类	用户画像	热销补余
		算法覆盖率	2.1%	14.8%	
APP	首页	猜你喜欢	已购买分类的相关	用户画像	热销补余
		算法覆盖率	1.8%	47.2%	

用户画像在大数据营销中的应用

根据画像的校园和偏好标签做营销：
男生买女性用品销量=》暖男排行
零食销量=》吃货排行
化妆品销量=》颜值排行
单反等销量=》潮人排行
安全套销量=》性福排行；
等等。

捕捉到用户画像标签属性的变迁调整新品

用户偏好画像的标签是通过用户的搜索、浏览、购买等所有的站内行为计算而来，针对标签的监控，可以体现用户的喜好和关注度的迁移变化。



案例启示

提炼出该案例（或项目）的哲理、方法论。

- 算法准确度、数据规模、更新速度相互制衡，提高某些指标，必须牺牲其他指标
- 一个系统遇到性能瓶颈的时候，跳出系统本身，了解业务，根据业务解耦，以满足不同场景
- 数据流各个环节都可能出错，自动化检查各个节点的中间数据，考虑降级和延迟环境
- 系统的演进是个长期的过程，系统的分分合合和业务量有关，防止过度架构浪费资源
- 不同版本开发的时候，适度换些开发者，融入新的思路，避免思维定式
- 标签体系的管理规范比技术本身更重要，否则大部分标签会沉睡，后面基本用不到。
- 数据驱动，通过观察和研究数据，对数据有一定的敏感度，产生新的用户画像数据。

谢谢大家！