

大数据构建标签体系

黄良懿@OPPO

目录

○ 用户标签体系

- 从埋点说标签系统的意义
- 标签系统的应用

○ 标签的前世今生

- 标签的定义
- 常用算法
- 标签的量化方法
- 标签验证及稳定性验证

○ 标签系统

- 数据拉通之上帝视角
- 系统介绍
- 三大调用模式（批量、个体、检索）

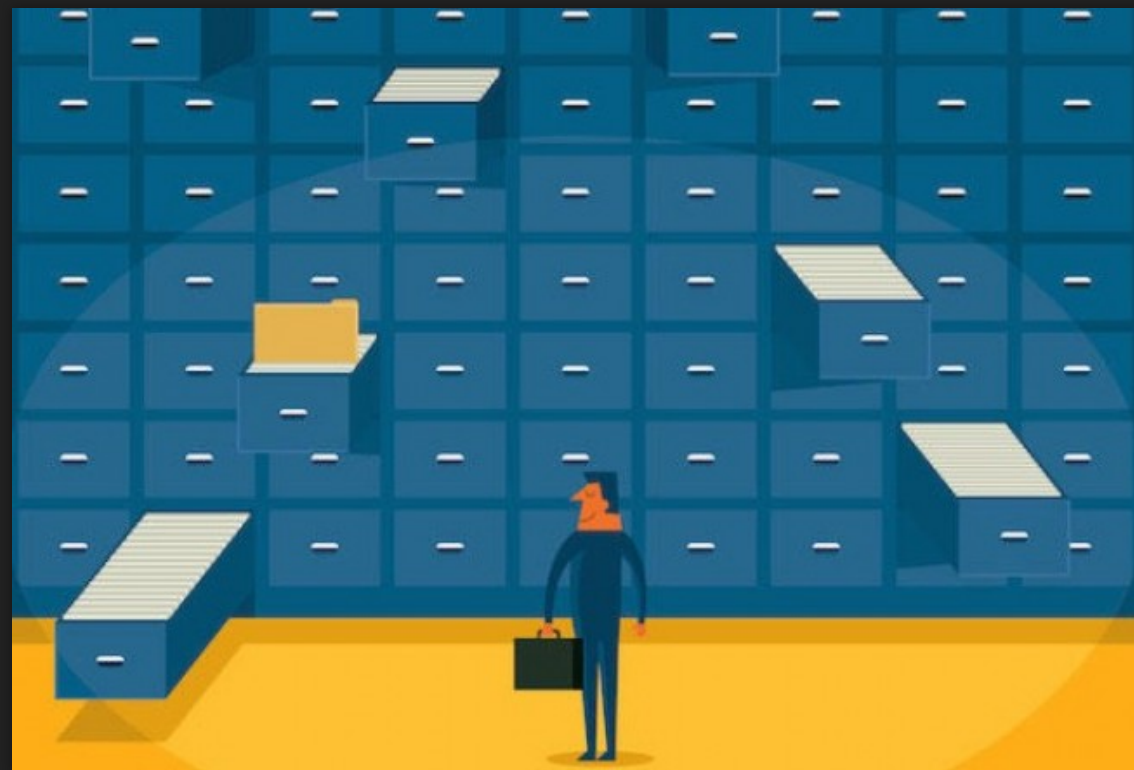
从埋点说起

不管有用没用，先收集再说

按需求埋点，还是全埋点？

数据存储压力
数据清洗压力
需求仍然难以满足
数据价值难体现

标签系统需要的大量属性信息是全埋点较难解决的



标签系统的意义

对于有产品系列的公司
容易出现数据各自为政的现象
数据孤岛
埋点规范不一
跨产品困难

良好的标签体系规划
可以引导各产品的埋点，形成产品矩阵
使数据价值在不同产品上分别体现



标签的应用场景

○ “更懂” 用户的三大类应用场景

- [在线] 提升用户体验，如利用关联规则，向不同人群推广不同游戏
- [近线] 提升业务转化率，如消息推送、个性推荐、广告投放等
- [离线] 利于用户研究，如分析R7男女老少用户分别都有哪些特征？

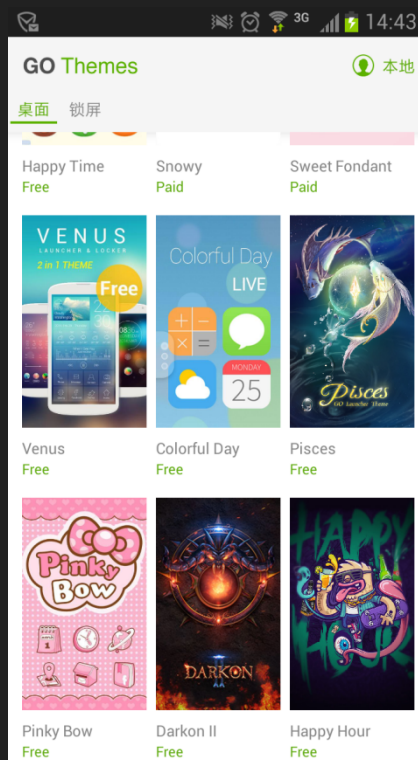
大数据下标签系统的变化



标签系统的线上应用场景



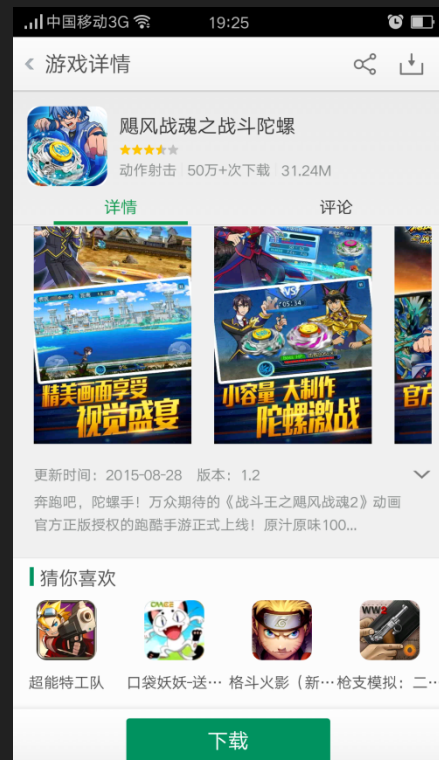
消息推送



付费意愿

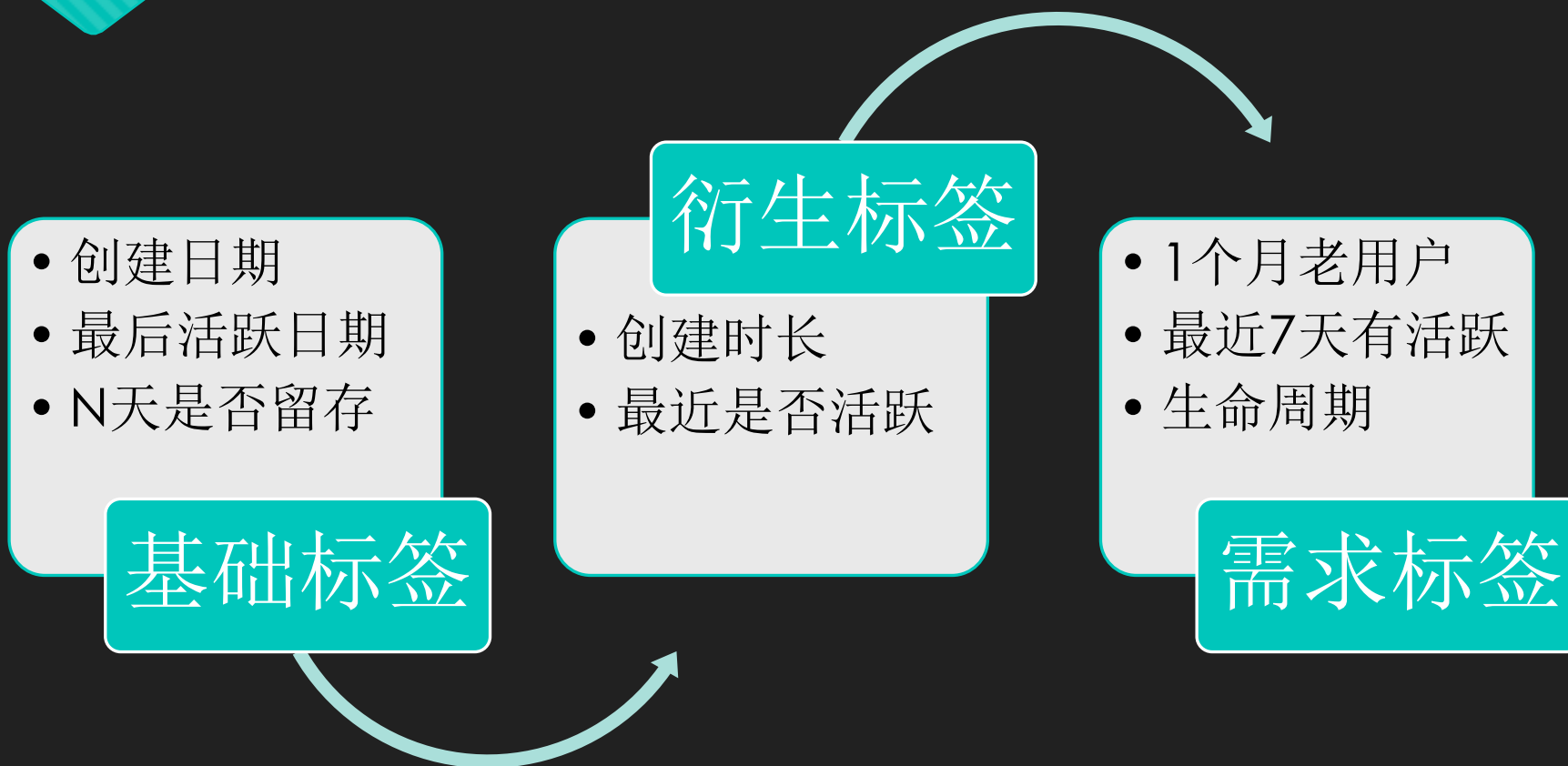


大家还喜欢



猜你喜欢

用户标签的繁衍



目录

○ 用户标签体系

- 从埋点说标签系统的意义
- 标签系统的应用

○ 标签的前世今生

- 标签的定义
- 常用算法
- 标签的量化方法
- 标签验证及稳定性验证

○ 标签系统

- 数据拉通之上帝视角
- 系统介绍
- 三大调用模式（批量、个体、检索）

用户标签的定义

标签是用户某一特征的符号化

标签体系

化整为零

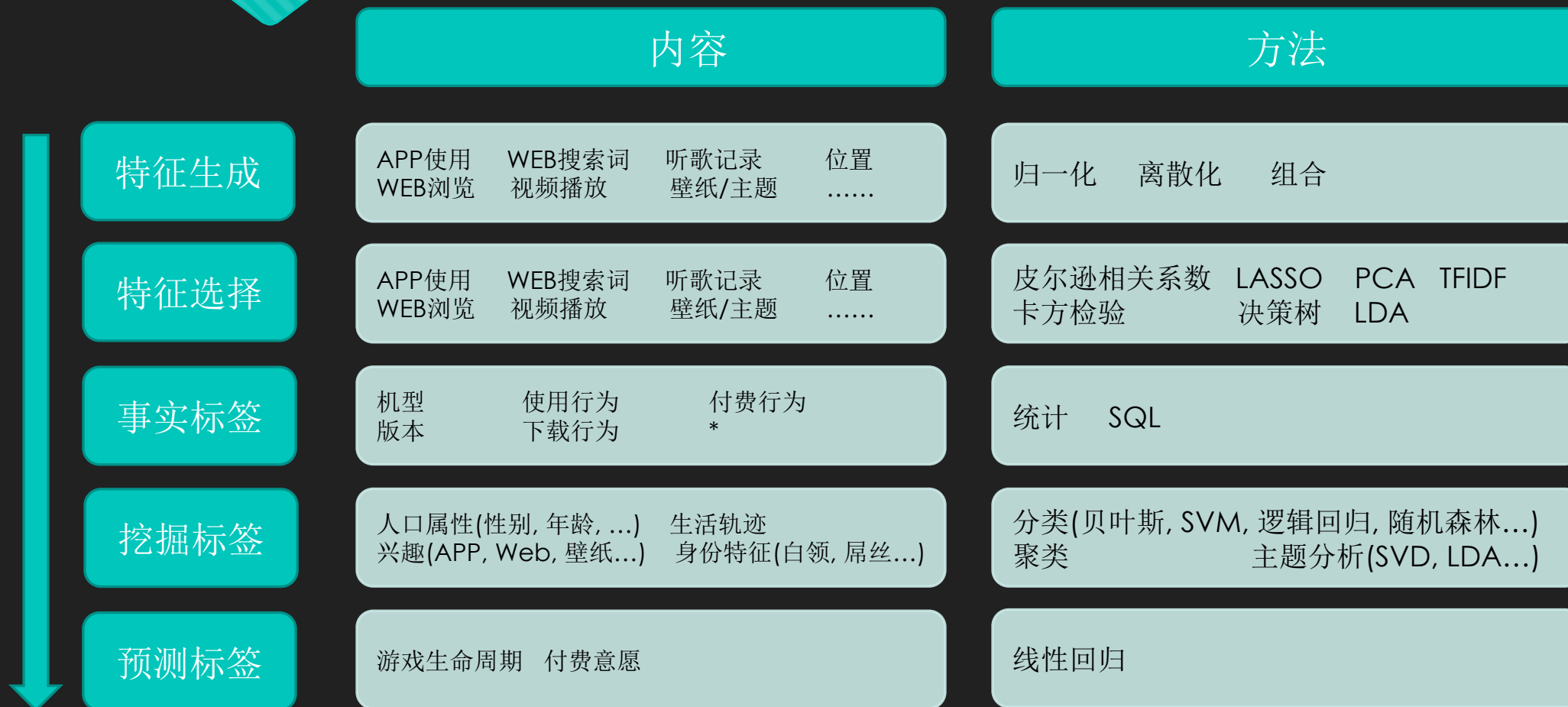
每个标签都是我们观察、
描述用户的一个角度

化零为整

标签之间并不孤立，标签集合
构成用户画像

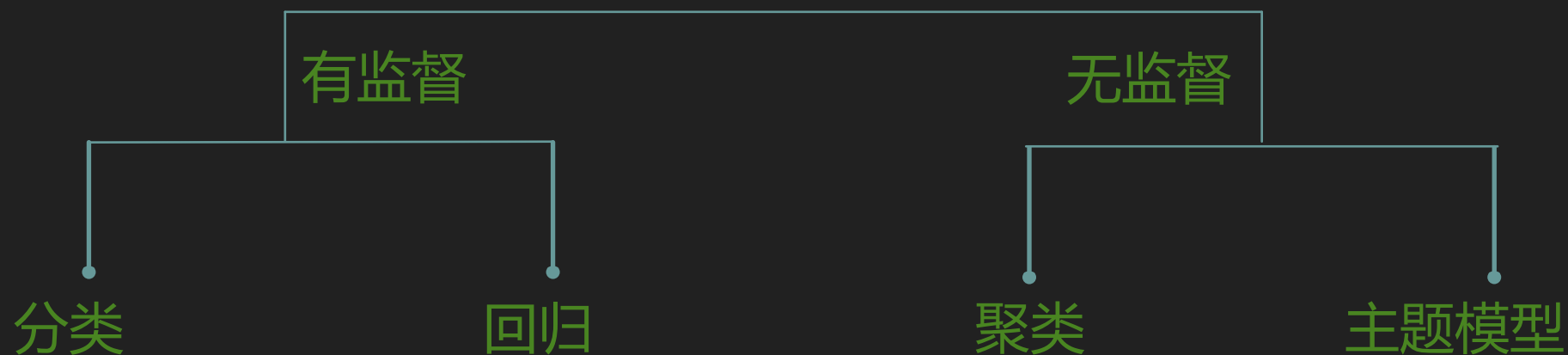
标签集合构成用户的整体画像

用户标签开发 - 总览



用户标签开发 - 算法

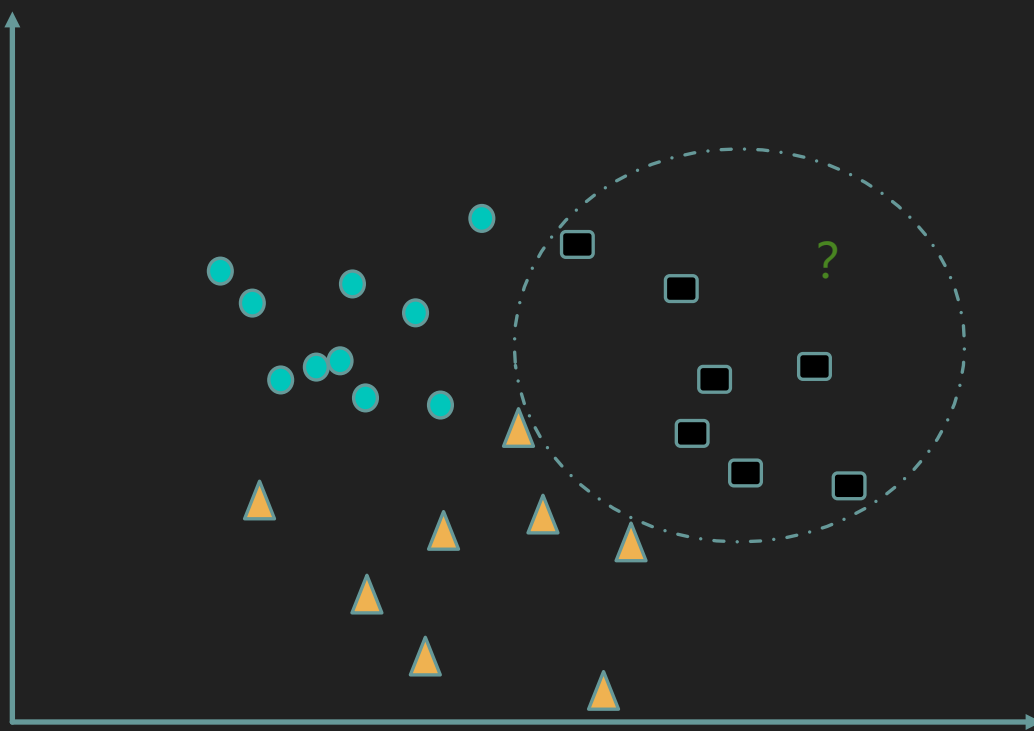
标签开发常见算法



主要问题:

1. 适用场景
2. 效果评价
3. 计算性能

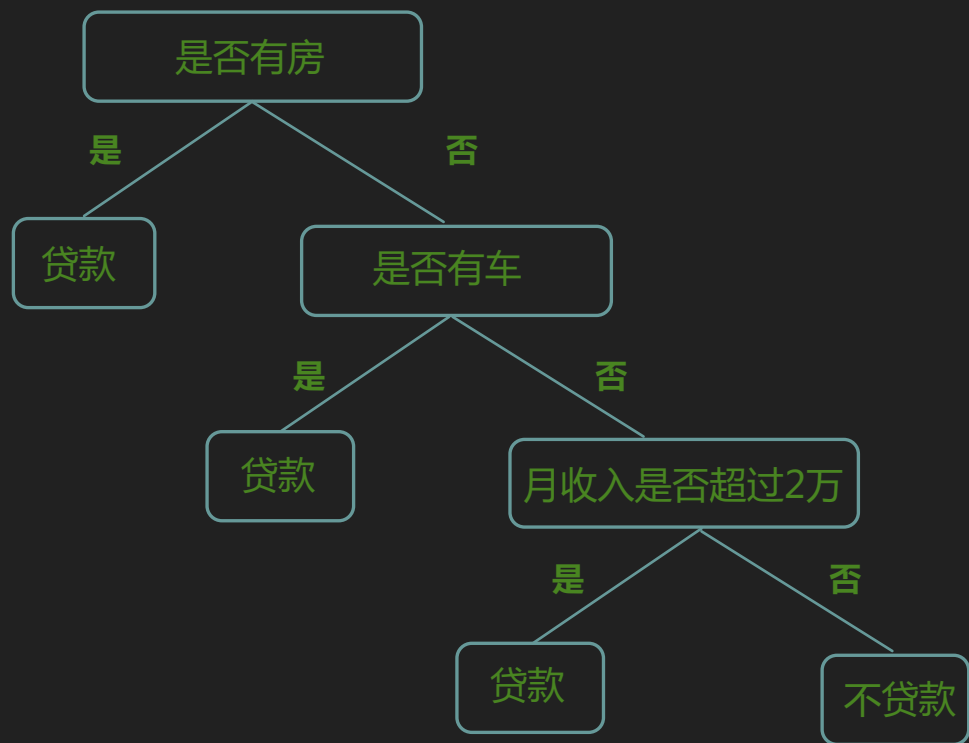
用户标签开发 - 算法 - 分类



将未知分类的样本划分到已知的类别里

用户标签开发 - 算法 - 分类

决策树



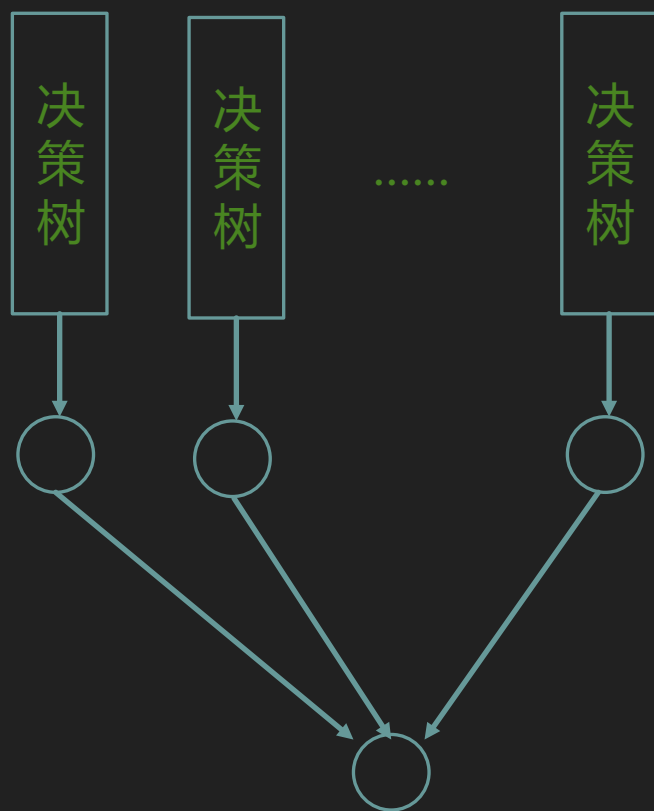
关键点：

- 每个决策或事件（即自然状态）都可能引出两个或多个事件，导致不同的结果，把这种决策分支画成图形很像一棵树的枝干
- 找到最好的特征来进行划分
- 如果尚未划分完全，则在剩下的特征里找到最好的特征来进行划分，以此类推
- 特征的“好坏”量化一般采用香农熵

$$H = -\sum_{i=1}^n p(x_i) \log_2 p(x_i)$$

用户标签开发 - 算法 - 分类

随机森林

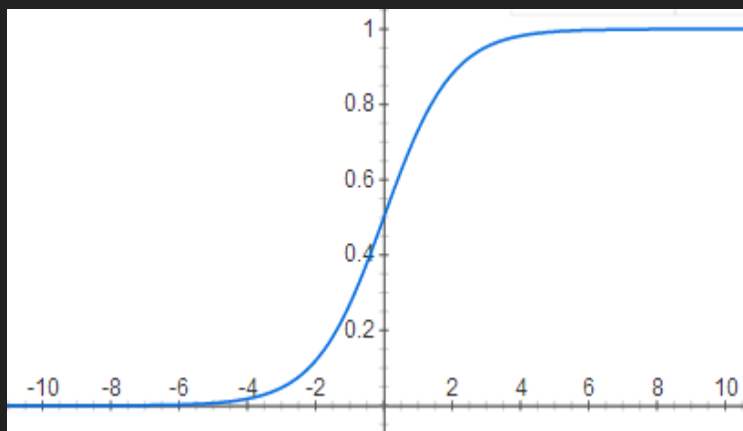


关键点：

- 随机的方式建立一个森林，森林里面有很多的决策树组成
- 根据每棵树的结果进行投票决定属于哪种分类
- 多个弱的分类器组成一个强的分类器

用户标签开发 - 算法 - 分类

Logistic回归



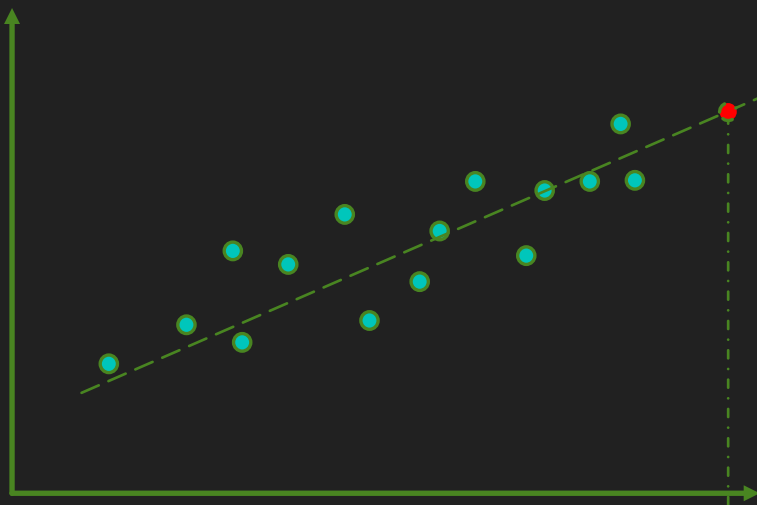
关键点：

- Logistic回归是一种广义线性回归，其输出范围在[0,1]之间，表示分类结果为1的可能性

$$y = \frac{1}{1+e^{-z}}, \text{ 其中 } z = a_0x_0 + a_1x_1 + \cdots + a_nx_n$$

用户标签开发 - 算法 - 回归

线性回归

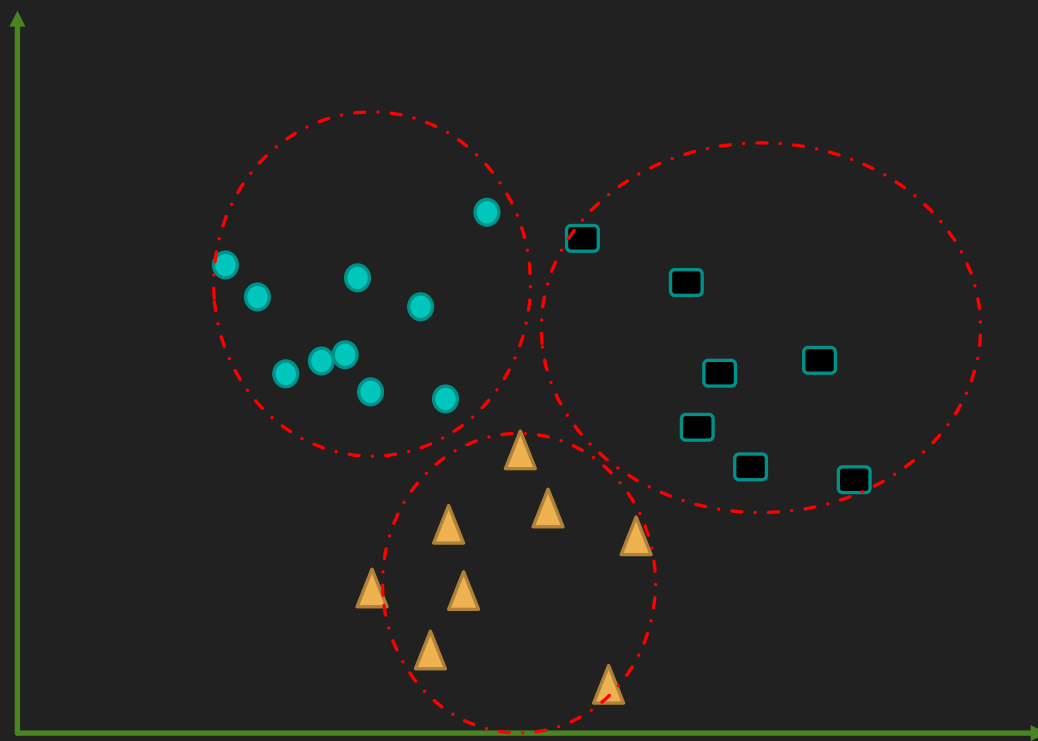


关键点：

- 通过最小二乘法找到 $Y=AX$ 的参数
- 可通过变换将原来非线性的转化为线性的

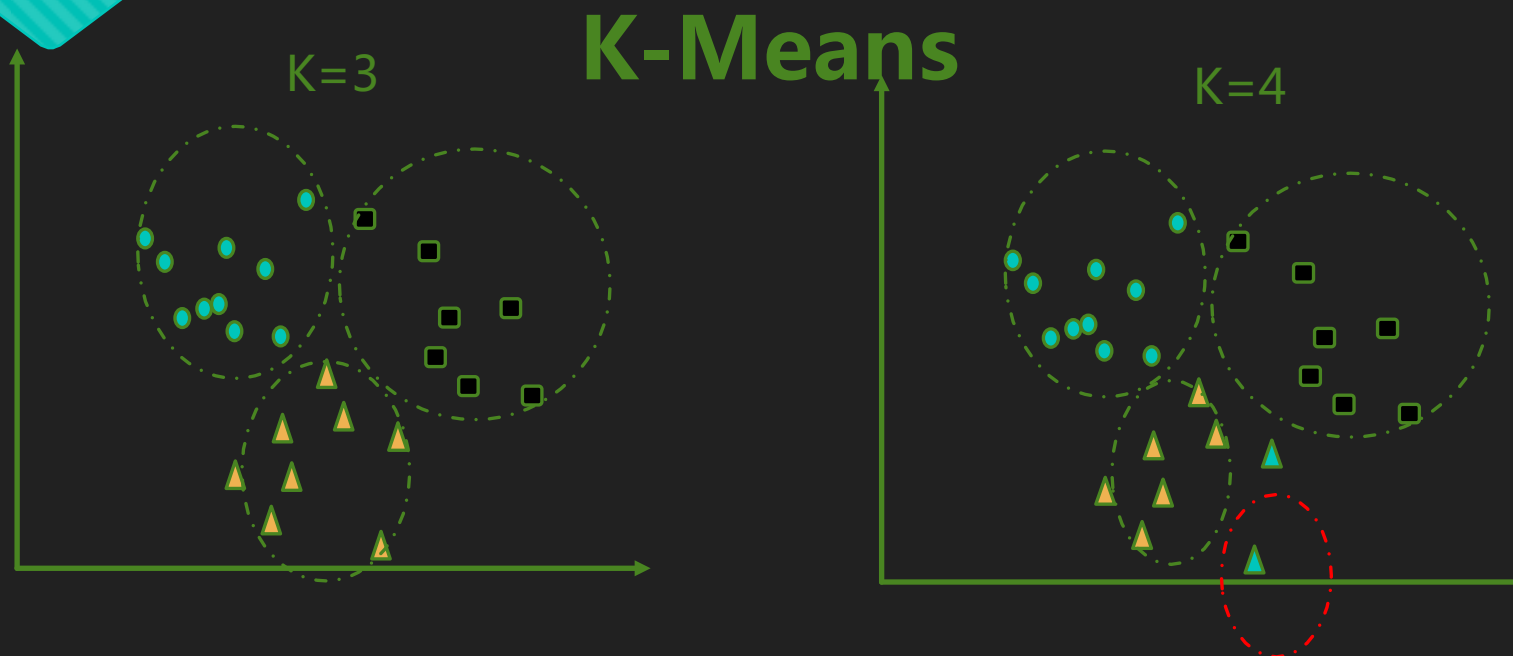
用户标签开发 - 算法 - 聚类

聚类



将样本空间进行分割或者划分

用户标签开发 - 算法 - 聚类



方法：

- 把n个点划分到k个聚类中，使得每个点都属于离他最近的聚类中心对应的类别
- 点之间的距离有多种定义，常用的距离计算方法有欧式距离，余弦距离等。选用那种方法需要具体问题具体分析

用户标签开发 - 量化方法

- 布尔值/分类值: 二分类/N分类问题
 - 性别=(男 or 女)、年龄=(18-24 or 24-30 or others)
- 离散值: 带权重信息
 - 标准分 $\in [100, 900]$
- 哈希值: 稀疏特征
 - 购物偏好 = {"化妆品": 90, "包包": 75, "手机": 45}
- 标签的模型多样化

用户标签开发 - 算法 - 效果评价

○ 准确率/召回率/F1

- 准确率: 模型判为正的样本中有多少是真正的正样本

$$\text{Precision}(A,B) = |A \text{ and } B| / |A|$$

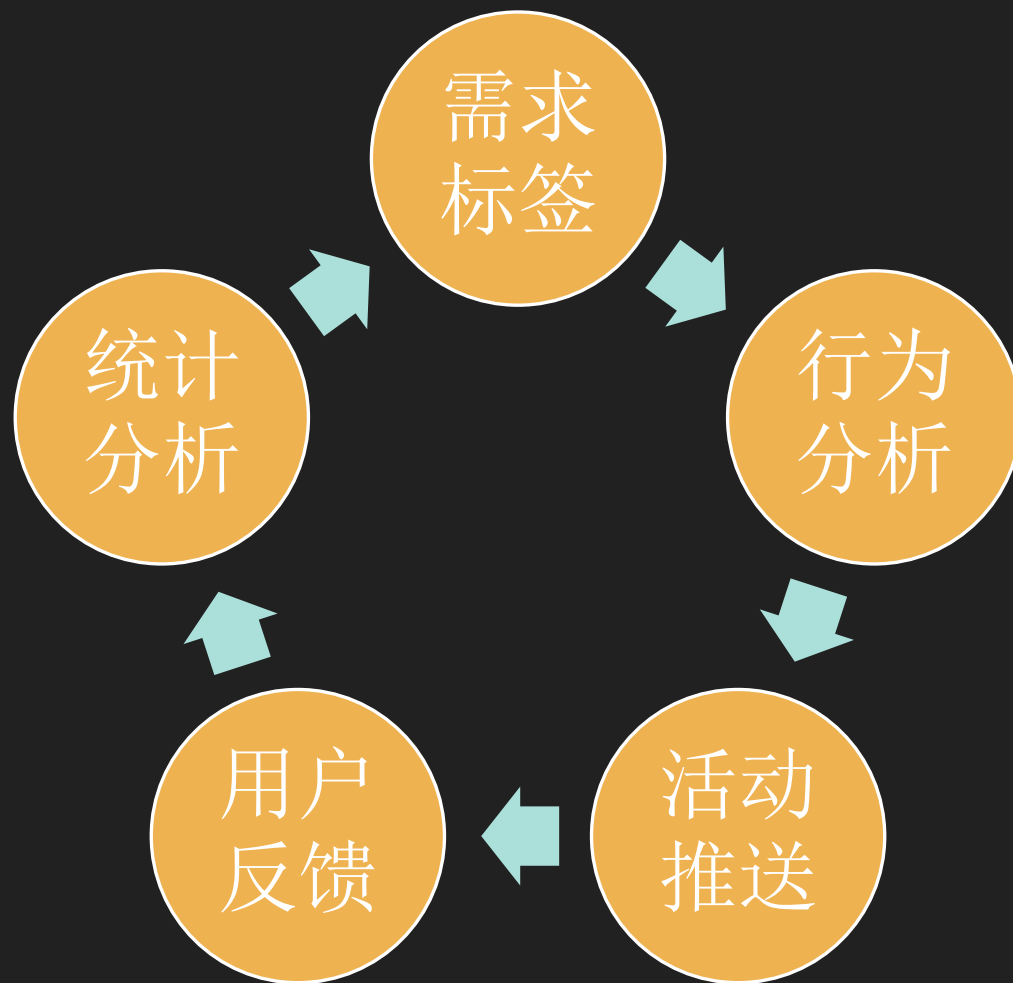
- 召回率: 所有真正的正样本中被模型判为正样本的比例

$$\text{Recall}(A,B) = |A \text{ and } B| / |B|$$

- F1-score: 准确率和召回率的综合指标

$$F1 = 2 \times P \times R / (P + R)$$

用户标签线上验证流程



用户标签开发 - 算法 - 效果监控

○ 效果如何保证持续

- 数据异常
- 集群异常
- 资源抢占失败

○ 效果监控和报警

- 召准率
- 经验规则
- 稳定性

标签数据稳定性规则规范

由 黄良懿(Louis)创建, 最终由 高洪涛修改于大约1分钟以前

标签	规则名	规则说明	无效阈值范围	生效时间
所有	执行失败	任务没有正常执行完成	False	20160715
性别	diff比例	diff比例高于指定阈值	>8%, 挑战>%6	20160830
	经验分布	男女比例正常范围	<40% >60%	20160801
	召准率	召回率和准确率	召回<90% 准确 <80%	20160730
年龄段	diff比例	diff比例高于指定阈值	>8%	20160830
	经验分布	各个年龄段的比例	待定	N/A
	召准率	召回率和准确率	召回<90% 准确 <65%	20160830

目录

○ 用户标签体系

- 从埋点说标签系统的意义
- 标签系统的应用

○ 标签的前世今生

- 标签的定义
- 常用算法
- 标签的量化方法
- 标签验证及稳定性验证

○ 标签系统

- 数据拉通之上帝视角
- 系统介绍
- 三大调用模式（批量、个体、检索）

标签建设需要上帝视角

○ 打通公司各部门数据，统一用户视图

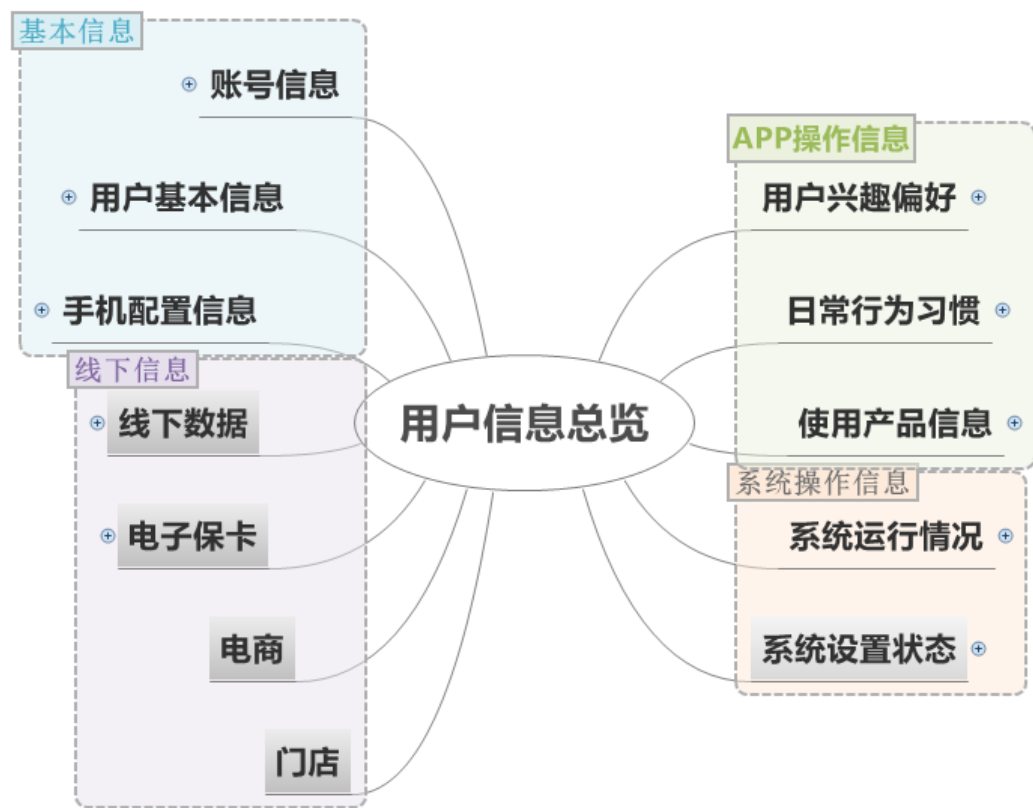
- 线下销售渠道(体验店/门店/销售点)
- 用户服务部门(在线客服/呼叫中心/售后服务)
- 社交媒体(微薄/微信/OPPO社区)

○ 标签分为属性、行为和需求3个一级类，12个子类

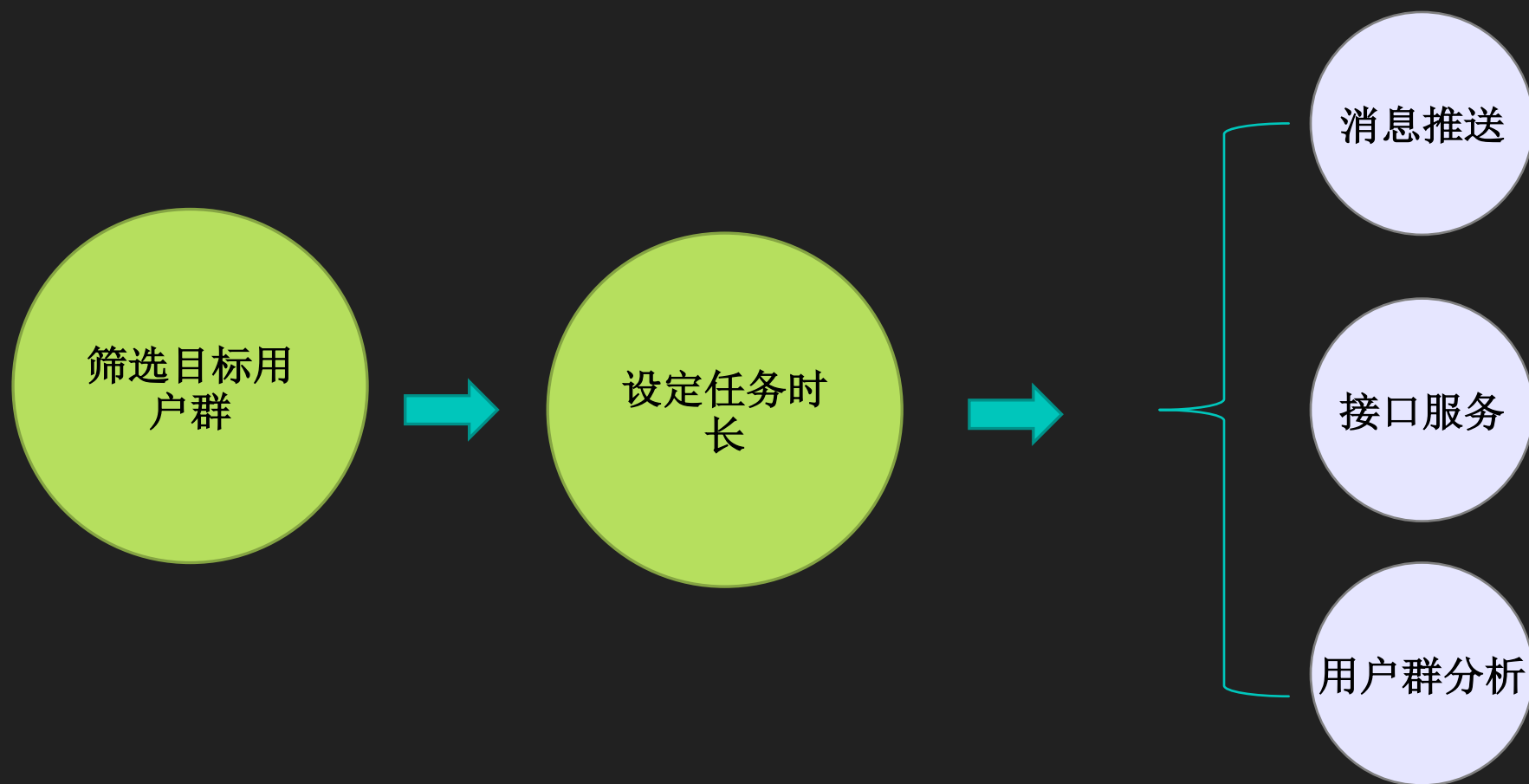
- **属性**：用户本身的固有特征
- **行为**：用户已发生的事件，用于挖掘用户的偏好和倾向
- **需求**：未来用户可能发生的行为，由属性和行为关联分析产生，用于指导具体的**用户运营**工作。



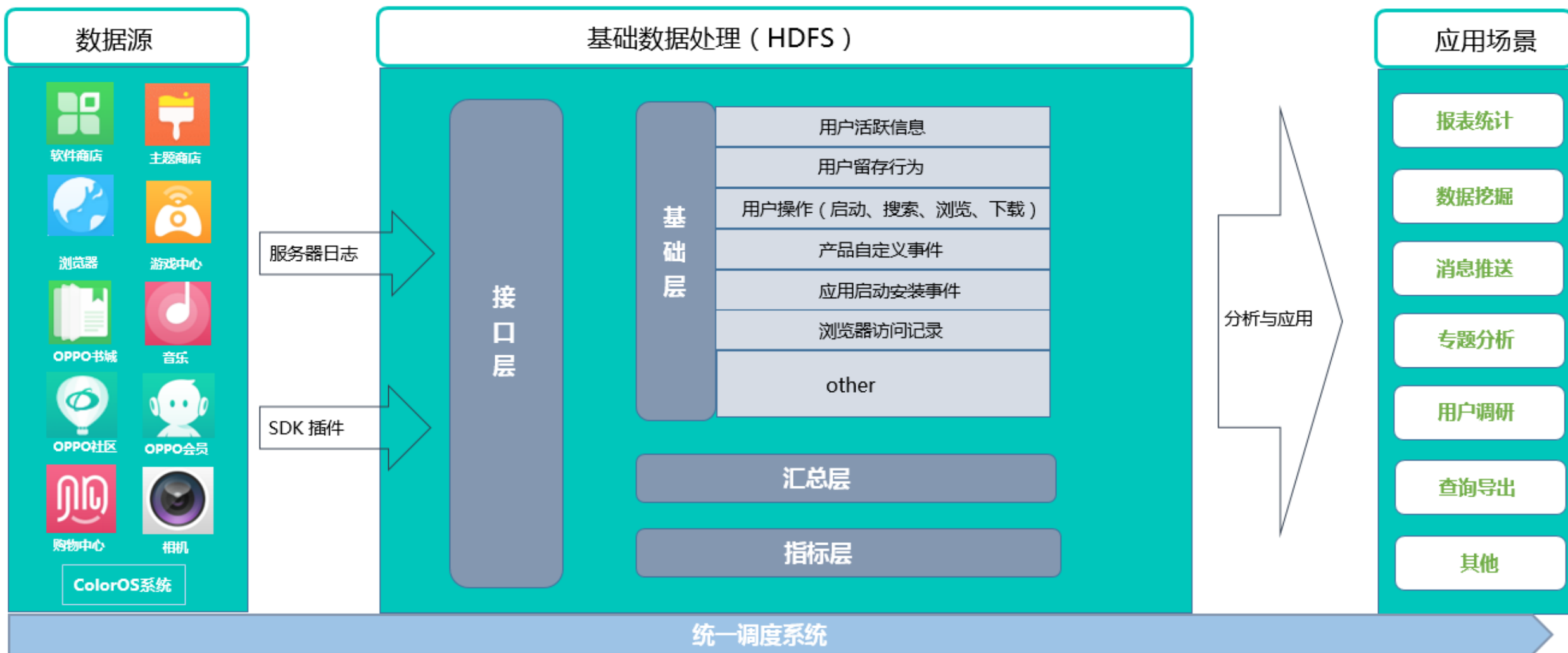
用户信息总览-例子



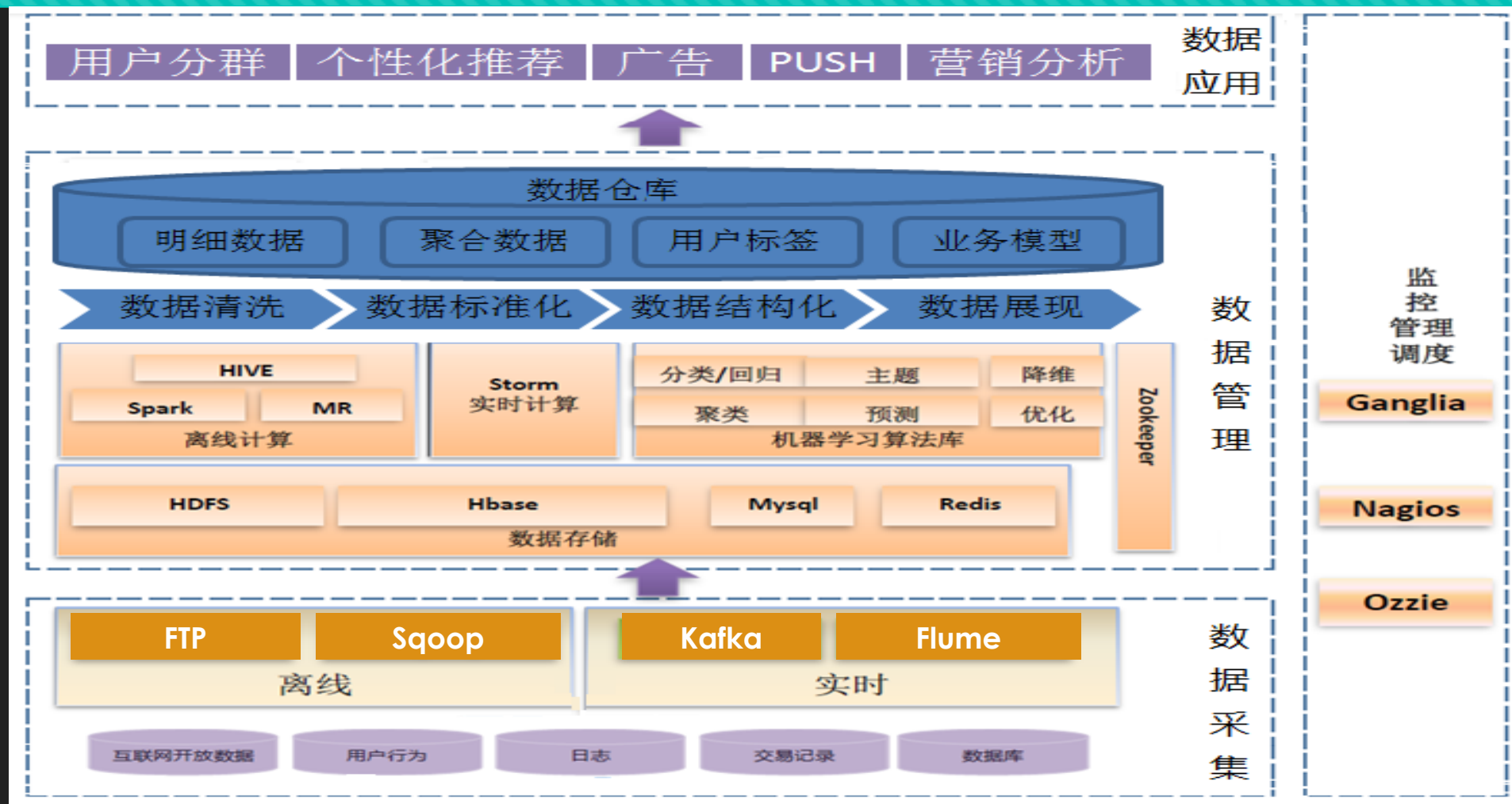
用户标签系统使用



我们的工作-数据仓库



OPPO标签技术生态

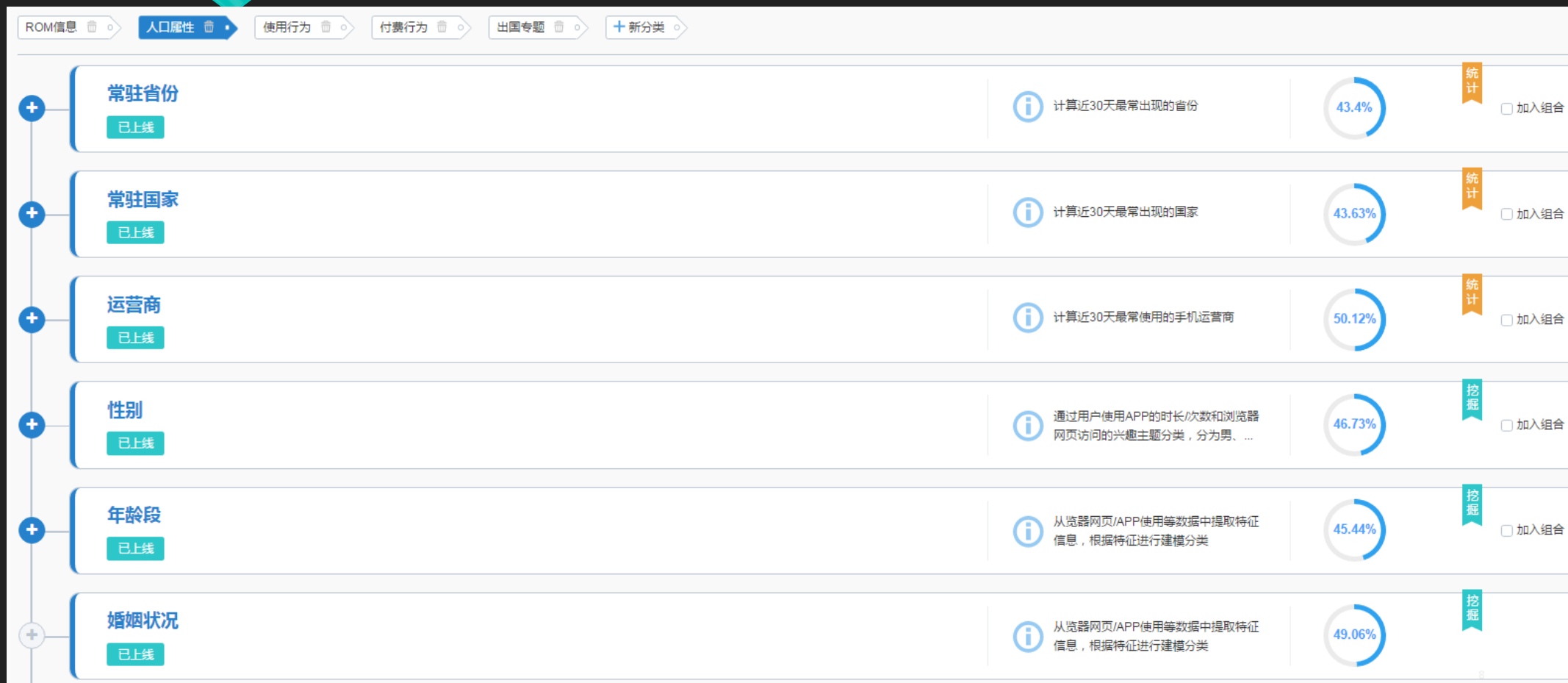


用户画像列表

输入用户ID
显示完整的用户标签树



用户标签浏览



用户群分析-例子

活跃行为

☐ 全选 ☐ 首次活跃时间 ☐ 最后活跃时间 ☒ 近7天活跃天数 ☐ 近30天活跃天数 ☐ 产品使用天数 ☐ 未活跃天数

添加

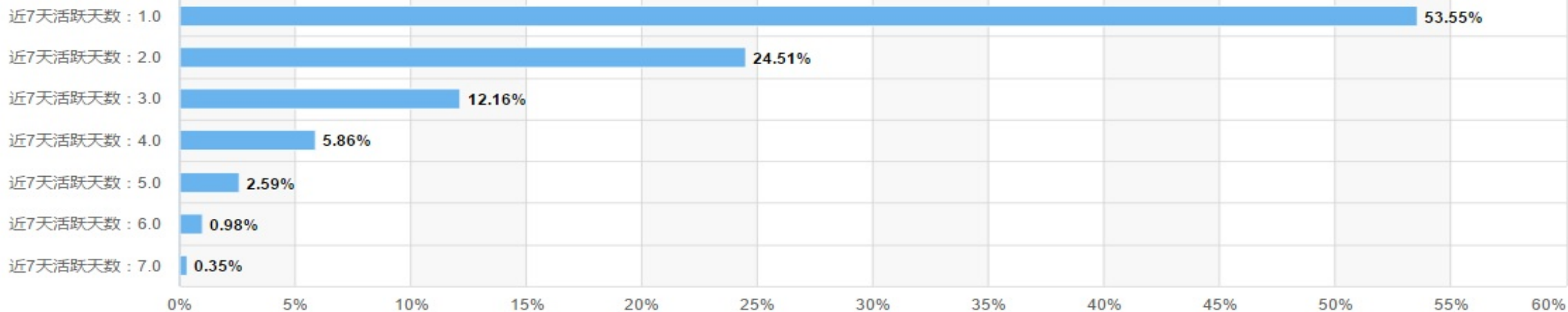
特征发现报告

清空

近7天活跃天数

显示条数: 10

查看



筛选目标用户群 – 组合标签的管理

The screenshot displays a user selection interface with a sidebar on the left and a main content area. The sidebar contains a list of categories: 组合标签, 平台, ROM信息, 人口属性, 使用行为, and 付费行为. Under 人口属性, 常驻省份 and 性别 are highlighted with red boxes. The main content area is titled 常驻省份 and includes a search bar with the text '请输入标签名称' and a '搜索' button. Below the search bar is a list of provinces with checkboxes: 全选, 广东, 山东, 河南, 浙江, 江苏, 河北, 湖南 (checked), 四川, 安徽, 云南, and a '更多>>' link. A red arrow points from the 常驻省份 category in the sidebar to the '添加' button. Below the province list is a '条件' section with buttons for '(', ')', 'AND', 'OR', and 'NOT'. The '已选条件' section shows two conditions: '平台-人口属性-性别: 女; 0.1 ~ 1' and '平台-人口属性-常驻省份: 湖南;', separated by an 'AND' button. A red arrow points from the 'AND' button in the '已选条件' section to the '查看覆盖情况' button. The bottom of the interface shows '总人数: --' and a '用户覆盖' bar. The '查看覆盖情况' button is circled in red.

组合标签

平台

ROM信息

人口属性

常驻省份

常驻国家

运营商

性别

年龄段

婚姻状况

职业

所在行业

受教育...

月收入...

城市类型

行政区域

商圈地区

身份类型

人生阶段

使用行为

付费行为

常驻省份

请输入标签名称 搜索

标签值

☐ 全选 ☐ 广东 ☐ 山东 ☐ 河南 ☐ 浙江 ☐ 江苏 ☐ 河北 ☒ 湖南 ☐ 四川 ☐ 安徽 ☐ 云南 更多>>

添加

条件

() AND OR NOT

已选条件

平台-人口属性-性别: 女; 0.1 ~ 1 AND 平台-人口属性-常驻省份: 湖南;

清空

总人数: --

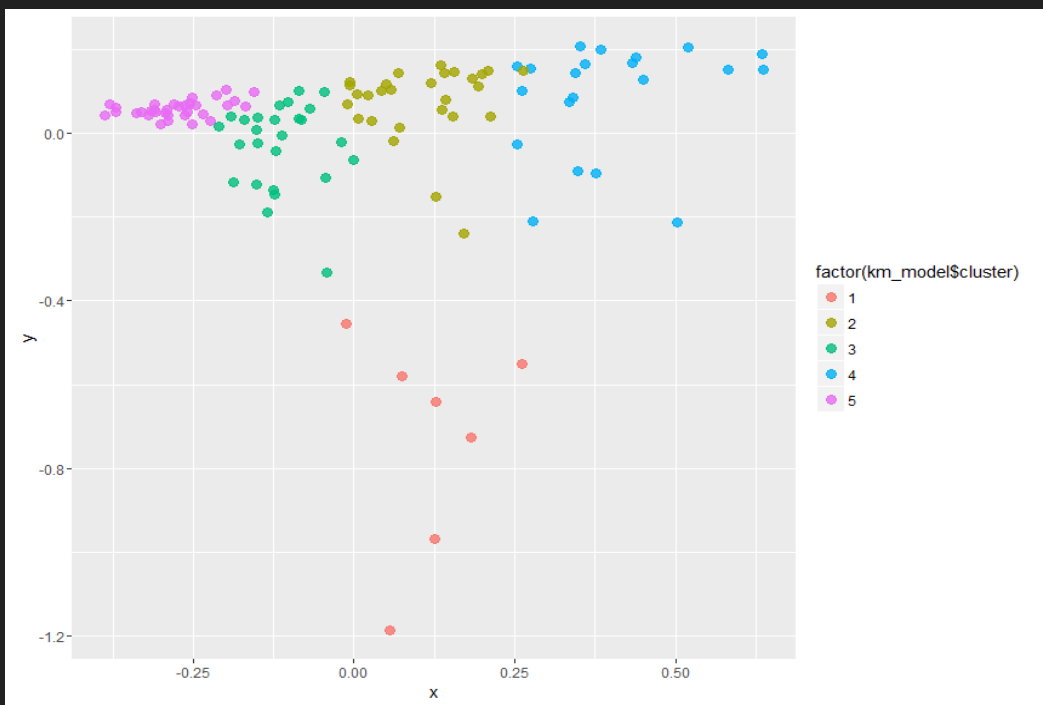
用户覆盖

查看覆盖情况

例子：湖南女生

标签的应用 – 妙龄少女APP使用分析

- 目的：针对年轻女性分析其APP兴趣，进行相应的营销和推荐活动
- 定义：介于18-24岁之间的女性



特征: 1. APP兴趣(110个APP兴趣标签); 2. 与整体人群的兴趣差值
模型: 对妙龄少女的特征空间进行聚类, 各类的APP分析如下:

1. 少女用的多, 整体人群用的也多, 包括社区, 视频, 聊天, 音乐等
2. 少女用的不多, 但相对整体人群用的多
3. 少女用的不多, 相对整体人群也不多
4. 少女用的多, 相对整体人群使用很突出: 包括相机, 美化, 美食, 乐器, 特效, 词典, 经期, 漫画, 减肥, K歌等.
5. 少女用的少, 相对整体人群用的更少: 包括天气, 斗地主, 股票, 育儿, 经营, 扫车, 听书, 等等

Thank you

○ Any Question?

- 诚聘数据挖掘工程师及高级数据分析师
- yeliang@oppo.com