

架构师  
ARCHITECT

特刊

# 用户画像实践

SPECIAL ISSUE  
March, 2017

架构师特刊



# ■ 卷首语

在 2016 年，王兴说，互联网已经进入了“下半场”，互联网人口红利的时间已经过去了，需要对用户的深耕细作获得更多的收入和利润。过去的一年里，各家将大数据从嘴上落到实际的运营体系当中，“用户画像”就是其中必不可少的一环。

无论是“增长黑客”还是“精益数据分析”，所有公司精细化运营者面对成千上万的用户，都会问那三个哲学上的终极问题：“你是谁？”（用户画像与特征），“你从哪里来？”（用户来源渠道与效果），“你到哪里去？”（用户流失与召回），其中用户画像系统会在业务和技术领域中不可或缺的组件。

由于产生用户画像会用到大量的数据挖掘算法，很多的 CTO/CDO 都认为将用户画像系统想当然的放置到挖掘团队来执行，而笔者认为，用户画像系统，是与大数据存储平台、大数据调度平台、元数据管理平台等平行的大数据基础业务组件，它执行力度层次应该以 CTO/CDO 执行领导的项目

体系。一个优秀的用户画像系统存在以下几个挑战，需要 CTO/CDO 亲自重视。

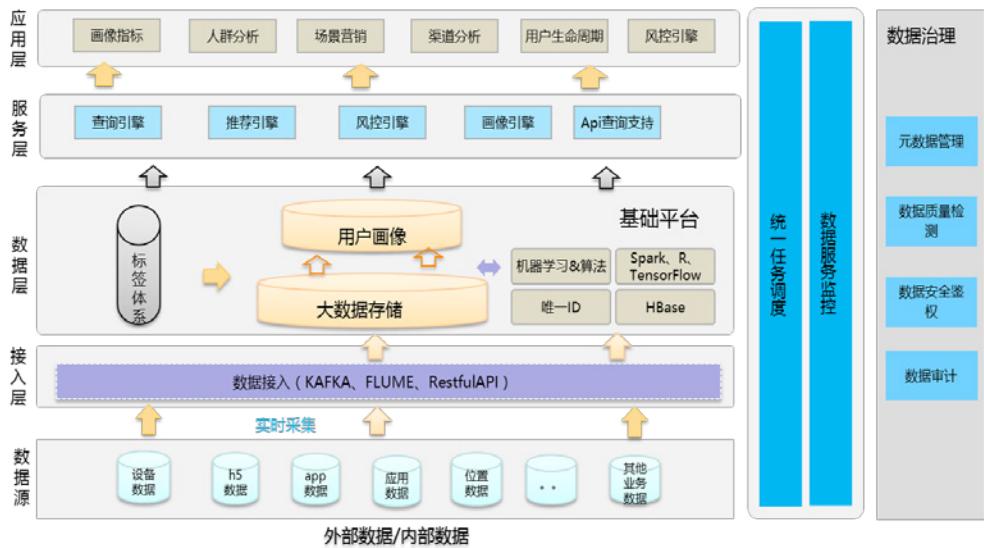
**用户画像系统的基础是用户统一 ID 系统：**用户统一 ID 系统，在传统公司里叫做 ECIF，它横跨了数据治理、数据整合、业务打通等几个难关；在互联网公司中叫用户跨屏唯一 ID，对于跨屏 ID 整合算法，APP 硬件设备指纹 / 防刷量等技术门槛有很高的要求，而做好这几点的业内公司少之又少。

**用户画像标签体系是业务技术共同合作的结晶：**在大数据融合的背景下，很多不同公司之间进行数据补全的工作，经常会遇到标签打通的难点。其实，一个公司好的标签体系与其业务是强绑定的，通用性较强的只有用户基本属性一层，越良好的标签体系越是和公司业务与运营密切相关，例如万达的线下品牌偏好度标签与易观线上 APP TGI 标签就是典型不同维度的指标体系分支。

**用户画像系统与各系统打通：**一个完备的用户画像系统，不仅仅为搜索推荐引擎服务，也会为数据分析 BI 展示、风控系统、数据挖掘引擎、数据元数据管理平台等提供有效的用户全生命周期的标签以及计算指标。技术和业务整合难度非常大，需要跨多个技术和业务部门进行协同，是一个技术“一把手”工程。

**用户画像的时时并发挑战：**一个优秀的画像系统经常会被各种系统时时访问，很多动态标签也需要实时更新，今日头条和一点资讯的时时推荐系统就是基于一个庞大的时时用户兴趣标签集群计算而得；而大量数据 Ad-hoc 查询经常体现在这里，最常见的案例就是要求秒级的用户标签与用户行为的交叉查询（十亿级别用户 v.s. 千亿级别的用户行为），InfoQ 中我和各位专家有很多类似文章，跟兴趣的同学可以去观看。

简而化之，用户画像系统的大致关系位置如下图：



综上，用户画像系统是一个涉及到各种知识体系的综合系统，本电子书中几个作者介绍一个公司如何从无到有的搭建用户画像系统，以及其中的技术难点与实际操作中的注意事项，实为用户画像的实操精华之选，推荐各位收藏阅读，也希望各位大数据从业人士在各自领域里有所斩获，算法精进，数据大成！

易观 CTO 郭炜

# 目录



**06 美团外卖 O2O 的用户画像实践**

**15 去哪儿的用户画像构建策略及应用实践**

**26 40 亿移动设备的用户画像和标签架构实践**

**34 携程是如何做用户画像的**

**42 百分点苏海波博士：为什么你做的用户画像模型不精准？**

**53 易观用户画像实践**

**60 让机器读懂用户：大数据中的用户画像**



# 美团外卖 O2O 的用户画像实践

李滔

美团外卖经过 3 年的飞速发展，品类已经从单一的外卖扩展到了美食、夜宵、鲜花、商超等多个品类。用户群体也从早期的学生为主扩展到学生、白领、社区以及商旅，甚至包括在 KTV 等娱乐场所消费的人群。随着供给和消费人群的多样化，如何在供给和用户之间做一个对接，就是用户画像的一个基础工作。所谓千人千面，画像需要刻画不同人群的消费习惯和消费偏好。

外卖 O2O 和传统的电商存在一些差异。可以简单总结为如下几点：

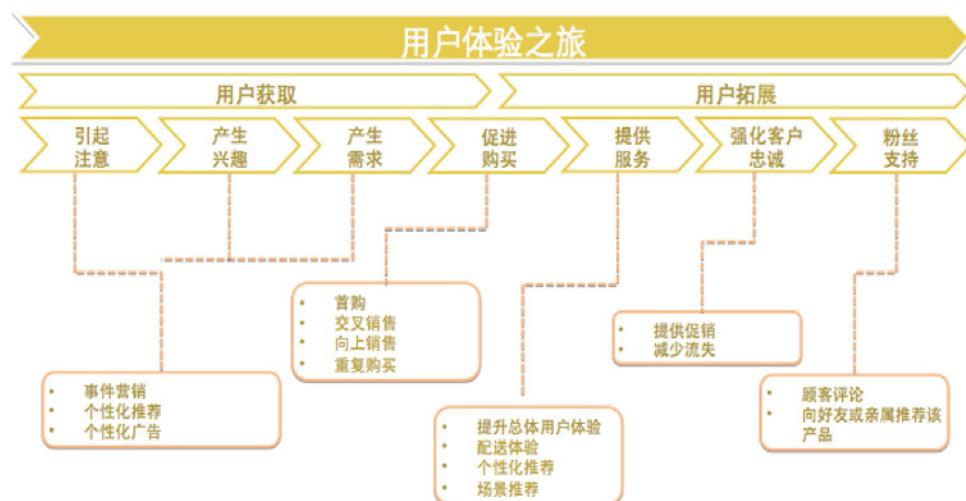
1. 新事物，快速发展：这意味着很多用户对外卖的认知较少，对平台上的新品类缺乏了解，对自身的需求也没有充分意识。平台需要去发现用户的消费意愿，以便对用户的消费进行引导。
2. 高频：外卖是个典型的高频 O2O 应用。一方面消费频次高，用户生

命周期相对好判定；另一方面消费单价较低，用户决策时间短、随意性大。

3. 场景驱动：场景是特定的时间、地点和人物的组合下的特定的消费意图。不同的时间、地点，不同类型的用户的消费意图会有差异。例如白领在写字楼中午的订单一般是工作餐，通常在营养、品质上有一定的要求，且单价不能太高；而到了周末晚上的订单大多是夜宵，追求口味且价格弹性较大。场景辨识越细致，越能了解用户的消费意图，运营效果就越好。
4. 用户消费的地理位置相对固定，结合地理位置判断用户的消费意图是外卖的一个特点。

## 外卖产品运营对画像技术的要求

如下图所示，我们大致可以把一个产品的运营分为用户获取和用户拓展两个阶段。在用户获取阶段，用户因为自然原因或一些营销事件（例如广告、社交媒体传播）产生对外卖的注意，进而产生了兴趣，并在合适的时机下完成首购，从而成为外卖新客。在这一阶段，运营的重点是提高效



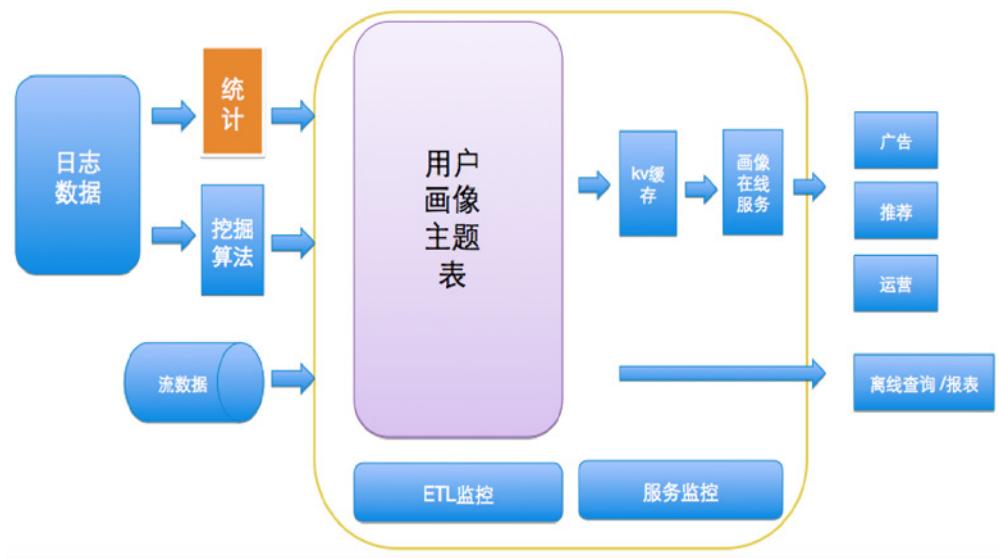
率，通过一些个性化的营销和广告手段，吸引到真正有潜在需求的用户，并刺激其转化。在用户完成转化后，接下来的运营重点是拓展用户价值。这里有两个问题：第一是提升用户价值，具体而言就是提升用户的单均价和消费频次，从而提升用户的 LTV (life-time value)。基本手段包括交叉销售（新品类的推荐）、向上销售（优质高价供给的推荐）以及重复购买（优惠、红包刺激重复下单以及优质供给的推荐带来下单频次的提升）；第二个问题是用户的留存，通过提升用户总体体验以及在用户有流失倾向时通过促销和优惠将用户留在外卖平台。

所以用户所处的体验阶段不同，运营的侧重点也需要有所不同。而用户画像作为运营的支撑技术，需要提供相应的用户刻画以满足运营需求。根据上图的营销链条，从支撑运营的角度，除去提供常规的用户基础属性（例如年龄、性别、职业、婚育状况等）以及用户偏好之外，还需要考虑这么几个问题：1) 什么样的用户会成为外卖平台的顾客（新客识别）；2) 用户所处生命周期的判断，用户是否可能从平台流失（流失预警）；3) 用户处于什么样的消费场景（场景识别）。后面“外卖 O2O 的用户画像实践”一节中，我们会介绍针对这三个问题的一些实践。

## 外卖画像系统架构

下图是我们画像服务的架构：数据源包括基础日志、商家数据和订单数据。数据完成处理后存放在一系列主题表中，再导入 kv 存储，给下游业务端提供在线服务。同时我们会对整个业务流程实施监控。主要分为两部分，第一部分是对数据处理流程的监控，利用用内部自研的数据治理平台，监控每天各主题表产生的时间、数据量以及数据分布是否有异常。第二部分是对服务的监控。目前画像系统支持的下游服务包括：广告、排序、

运营等系统。



## 外卖 O2O 的用户画像实践

### 新客运营

新客运营主要需要回答下列三个问题：

1. 新客在哪里？
2. 新客的偏好如何？
3. 新客的消费力如何？

回答这三个问题是比较困难的，因为相对于老客而言，新客的行为记录非常少或者几乎没有。这就需要我们通过一些技术手段作出推断。例如：新客的潜在转化概率，受到新客的人口属性（职业、年龄等）、所处地域（需求的因素）、周围人群（同样反映需求）以及是否有充足供给等因素的影响；而对于新客的偏好和消费力，从新客在到店场景下的消费行为可以做出推测。另外用户的工作和居住地点也能反映他的消费能力。

对新客的预测大量依赖他在到店场景下的行为，而用户的到店行为对

于外卖是比较稀疏的，大多数的用户是在少数几个类别上有过一些消费行为。这就意味着我们需要考虑选择什么样的统计量描述：是消费单价，总消费价格，消费品类等等。然后通过大量的试验来验证特征的显著性。另外由于数据比较稀疏，需要考虑合适的平滑处理。

我们在做高潜新客挖掘时，融入了多方特征，通过特征的组合最终作出一个效果比较好的预测模型。我们能够找到一些高转化率的用户，其转化率比普通用户高若干倍。通过对高潜用户有针对性的营销，可以极大提高营销效率。

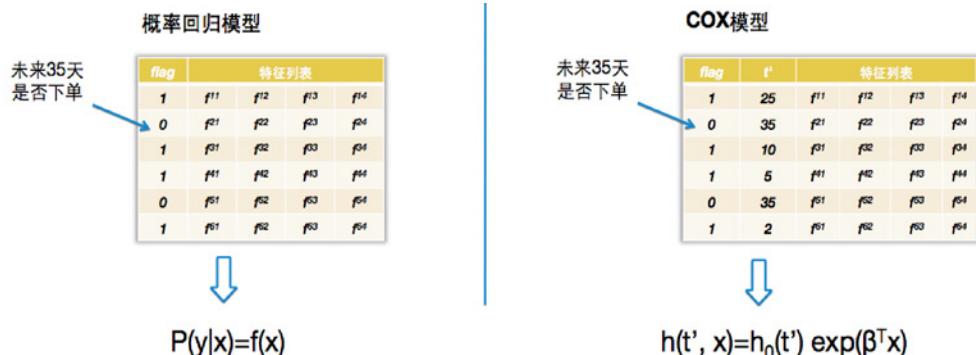
## 流失预测

新客来了之后，接下来需要把他留在这个平台上，尽量延长生命周期。营销领域关于用户留存的两个基本观点是（引自菲利普·科特勒《营销管理》）。

- 获取一个新顾客的成本是维系现有顾客成本的5倍！
- 如果将顾客流失率降低5%，公司利润将增加25%~85%。

用户流失的原因通常包括：竞对的吸引；体验问题；需求变化。我们借助机器学习的方法，构建用户的描述特征，并借助这些特征来预测用户未来流失的概率。这里有两种做法：第一种是预测用户未来若干天是否会下单这一事件发生的概率。这是典型的概率回归问题，可以选择逻辑回归、决策树等算法拟合给定观测下事件发生的概率；第二种是借助于生存模型，例如 COX-PH 模型，做流失的风险预测。下图左边是概率回归的模型，用户未来 T 天内是否有下单做为类别标记  $y$ ，然后估计在观察到特征  $X$  的情况下  $y$  的后验概率  $P(y|X)$ 。右边是用 COX 模型的例子，我们会根据用户在未来 T 天是否下单给样本一个类别，即观测时长记为 T。假设用户

的下单的距今时长  $t < T$ , 将  $t$  作为生存时长  $t'$ ; 否则将生存时长  $t'$  记为  $T$ 。这样一个样本由三部分构成: 样本的类别 ( $flag$ ), 生存时长 ( $t'$ ) 以及特征列表。通过生存模型虽然无法显式得到  $P(t' | X)$  的概率, 但其协变量部分实际反映了用户流失的风险大小。



生存模型中,  $\beta^T x$  反映了用户流失的风险, 同时也和用户下次订单的时间间隔成正相关。下面的箱线图中, 横轴为  $\beta^T x$ , 纵轴为用户下单时间的间隔。

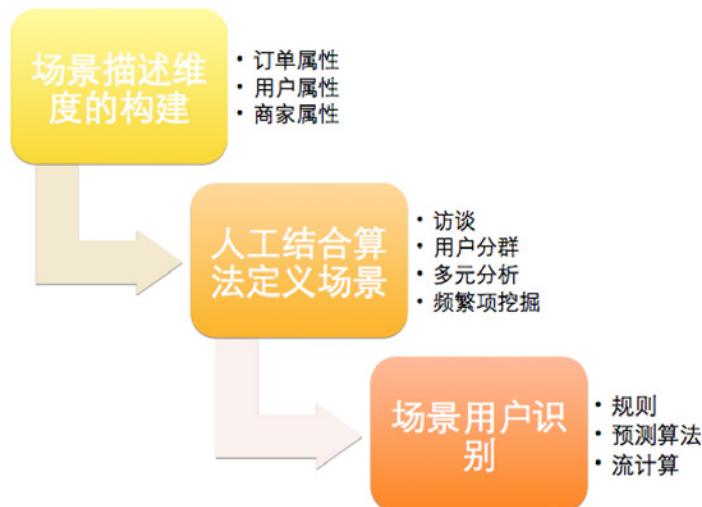
A box plot showing the distribution of user purchase time intervals (Y-axis, 0 to 40) versus  $\beta^T x$  (X-axis, ranging from -39.4 to 9.8). The plot indicates a positive correlation, where higher values of  $\beta^T x$  correspond to shorter purchase intervals.

我们做了 COX 模型和概率回归模型的对比。在预测用户 XX 天内是否会下单上面，两者有相近的性能。

美团外卖通过使用了用户流失预警模型，显著降低了用户留存的运营成本。

## 场景运营

拓展用户的体验，最重要的一点是要理解用户下单的场景。了解用户的订餐场景有助于基于场景的用户运营。对于场景运营而言，通常需要经过如下三个步骤。

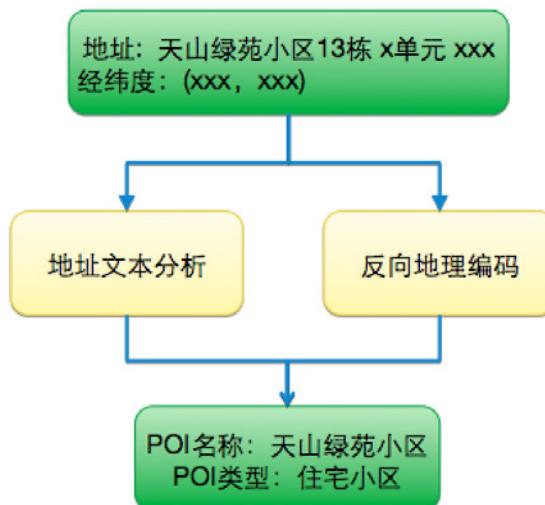


场景可以从时间、地点、订单三个维度描述。比如说工作日的下午茶，周末的家庭聚餐，夜里在家点夜宵等等。其中重要的一点是用户订单地址的分析。通过区分用户的订单地址是写字楼、学校或是社区，再结合订单时间、订单内容，可以对用户的下单场景做到大致的了解。

上图是我们订单地址分析的流程。根据订单系统中的用户订单地址文本，基于自然语言处理技术对地址文本分析，可以得到地址的主干名称（指去掉了楼宇、门牌号的地址主干部分）和地址的类型（写字楼、住宅小区

等）。在此基础上通过一些地图数据辅助从而判断出最终的地址类型。

另外我们还做了合并订单的识别，即识别一个订单是一个人下单还是拼单。把拼单信息、地址分析以及时间结合在一起，我们可以预测用户的消费场景，进而基于场景做交叉销售和向上销售。



## 总结

外卖的营销特征，跟其他行业的主要区别在于：

外卖是一个高频的业务。由于用户的消费频次高，用户生命周期的特征体现较显著。运营可以基于用户所处生命周期的阶段制定营销目标，例如用户完成首购后的频次提升、成熟用户的价值提升、衰退用户的挽留以及流失用户的召回等。因此用户的生命周期是一个基础画像，配合用户基本属性、偏好、消费能力、流失预测等其他画像，通过精准的产品推荐或者价格策略实现运营目标。

用户的消费受到时间、地点等场景因素驱动。因此需要对用户在不同的时间、地点下消费行为的差异做深入了解，归纳不同场景下用户需求的差异，针对场景制定相应的营销策略，提升用户活跃度。

另外由于外卖是一个新鲜的事物，在用户对一些新品类和新产品缺乏认知的情况下，需要通过技术手段识别用户的潜在需求，进行精准营销。例如哪些用户可能会对小龙虾、鲜花、蛋糕这样的相对低频、高价值的产品产生购买。可以采用的技术手段包括用户分群、对已产生消费的用户做 look-alike 扩展、迁移学习等。

同时我们在制作外卖的用户画像时还面临如下挑战。

- 数据多样性，存在大量非结构化数据例如用户地址、菜品名称等。需要用到自然语言处理技术，同时结合其他数据进行分析。
- 相对于综合电商而言，外卖是个相对单一的品类，用户在外卖上的行为不足以全方位地描述用户的基本属性。因此需要和用户在其他场合的消费行为做融合。
- 外卖单价相对较低，用户消费的决策时间短、随意性强。不像传统电商用户在决策前有大量的浏览行为可以用于捕捉用户单次的需求。因此更需要结合用户画像分析用户的历史兴趣、以及用户的消费场景，在消费前对用户做适当的引导、推荐。

面临这些挑战，需要用户画像团队更细致的数据处理、融合多方数据源，同时发展出新的方法论，才能更好地支持外卖业务发展的需要。而外卖的上述挑战，又分别和一些垂直领域电商类似，经验上存在可以相互借鉴之处。因此，外卖的用户画像的实践和经验累积，必将对整个电商领域的大数据应用作出新的贡献！

## 作者简介

**李滔**，美团外卖数据组。多年算法经验，在图像、文本、推荐、广告等领域都有涉猎。目前在美团外卖负责数据挖掘和精准营销业务。目前对强化学习及其在推荐、对话中的应用感兴趣，希望能和志趣相投的朋友相互交流。



# 去哪儿的用户画像构建策略及应用实践

李国芳

## 1 用户画像的构建原则

我们做用户画像的目的有两个。

1. 必须从业务场景出发，解决实际的业务问题，之所以进行用户画像要么是获取新用户，或者是提升用户体验，或者是挽回流失用户等有明确的业务目标。
2. 根据用户画像的信息做产品设计，必须要清楚知道用户长什么样子，有什么行为特征和属性，这样才能为用户设计产品或开展营销活动。

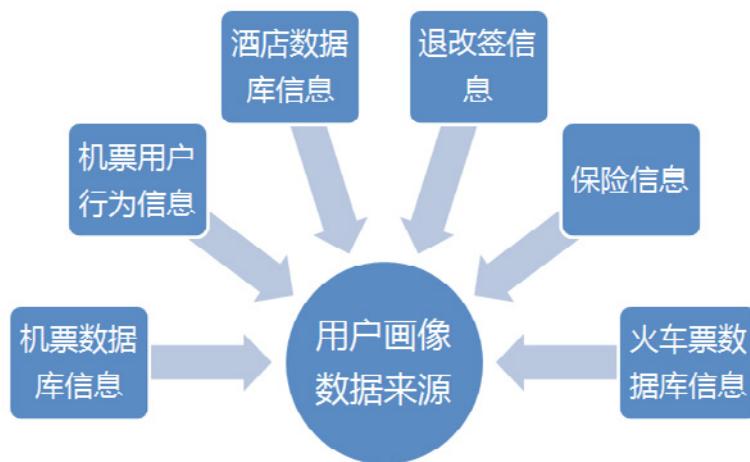
一般常见的错误想法是画像维度的数据越多越好，画像数据越丰富越好，费了很大的力气进行画像后，却发现只剩下了用户画像，和业务相差甚远，没有办法直接支持业务运营，投入精力巨大但是回报微小，可以说

得不偿失。鉴于此，我们的画像的维度和设计原则都是紧紧跟着业务需求去推动。

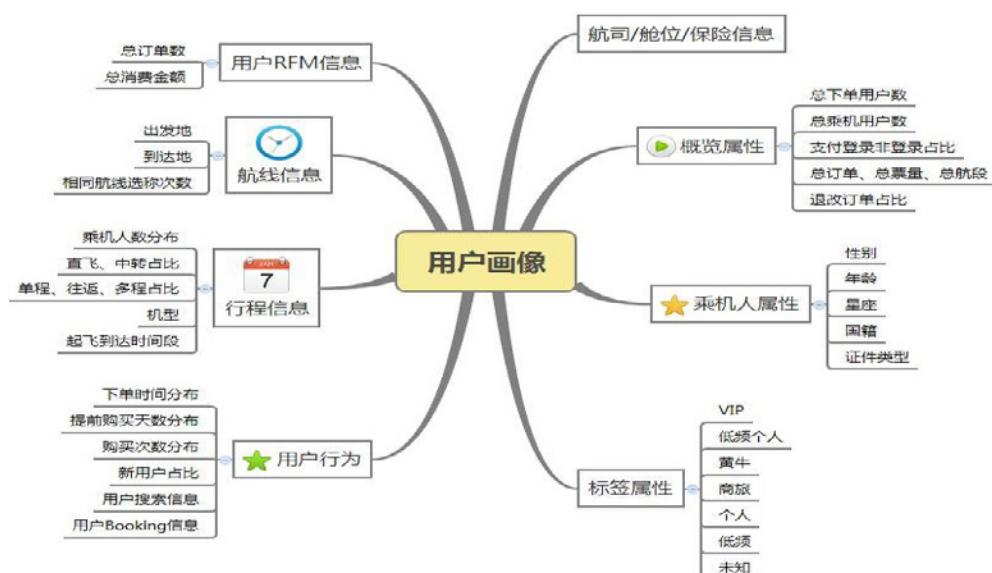
## 2 用户画像数据仓库构建

### 2.1 数据源的集成

目前哪儿网用户画像数据仓库中的数据源来自业务数据库的数据和用户行为日志数据，目前数据仓库中基本涵盖了机票、酒店、火车票以及保险等业务系统的数据，可以从全方位的了解去哪儿的一个用户的画像。

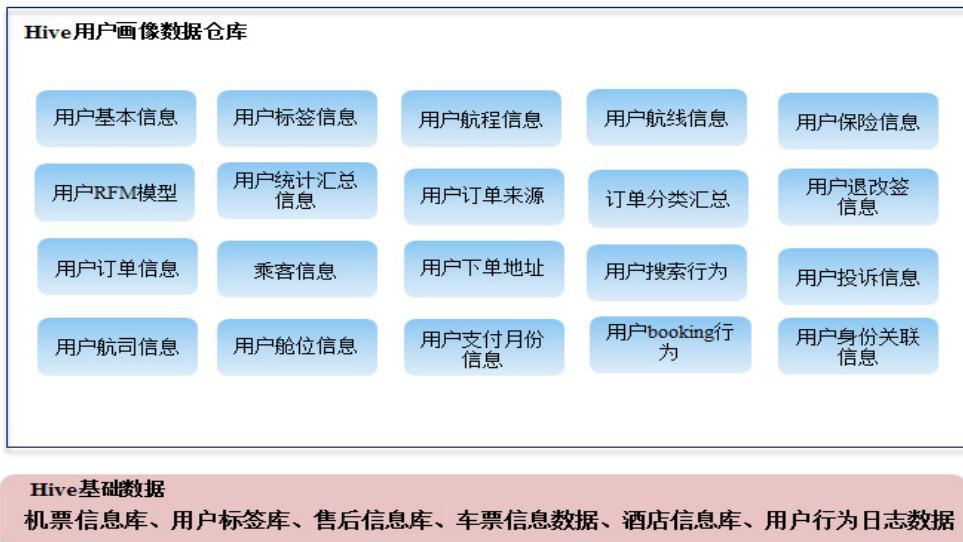


### 2.2 我们有哪些数据？ - 数据维度



## 2.3 我们有哪些数据？数据仓库

目前我们画像数据仓库的构建都是基于哪儿网基础数据仓库进行构建，并按照维度进行划分。



目前数据仓库中包括的信息如下：

- 画像数据仓库表20个
- 画像数据仓库
- 国内、国际 2年+数据
- 标签数据
- 每日增量
  - 基本数据
  - 业务数据
  - 搜索 - Booking

## 2.4 用户唯一标识设计

用户唯一标识是整个用户画像的核心，它把从用户开始使用 App 到下

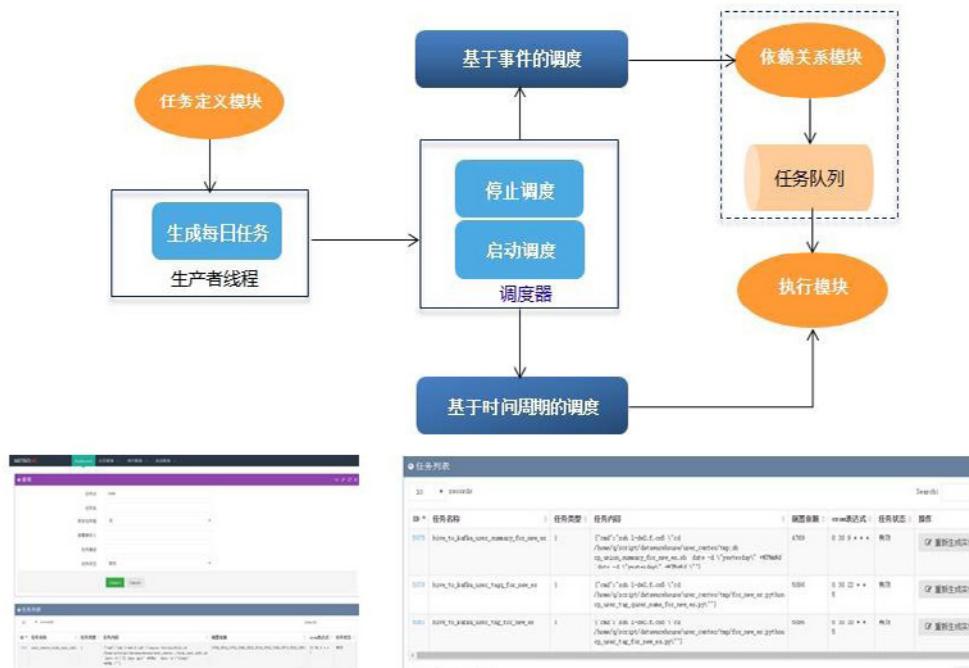
单到售后整个所有的用户行为轨迹进行关联，可以更好的去跟踪和描绘一个用户的特征。

为什么要设计唯一标识？



## 2.5 ETL 过程设计 - 调度系统

- 依赖数据平台调度系统
- 定时触发和Job依赖触发两种模式



## 2.6 ETL 过程设计 - 任务执行

- ETL的过程主要是将数据源的清洗到数据仓库表的过程（每天更新增量）
- Summary表的处理逻辑（每天更新全量）
- 标签库的处理（每周更新，2年全量）



## 2.7 用户主题分析及数据挖掘

有了丰富的画像数据后，产品和运营人员可以根据用户主题进行数据分析和数据挖掘相关的工作。用户主题 Cube 的定义如下：

- Measure:
  - 订单数量
  - 订单金额
  - 搜索次数
  - Booking 次数
- Dimension:
  - 下单时间

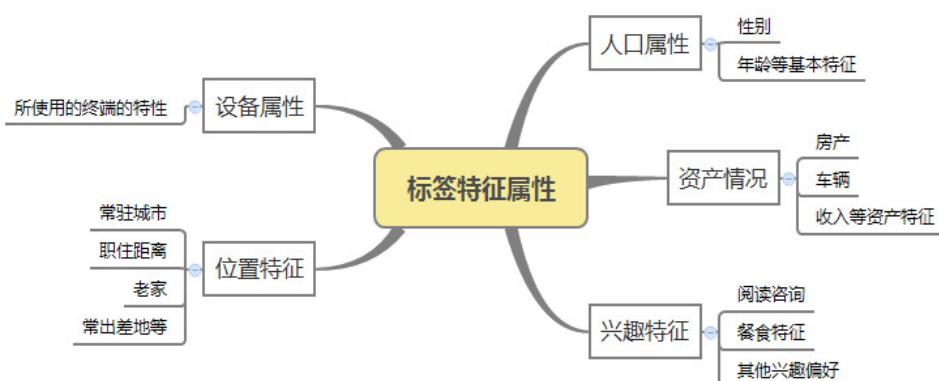
- 出发时间
- 航司信息
- 舱位信息
- 航班（出发地、目的地）
- 基本信息（年龄、性别等自然属性）



### 3 用户画像标签构建策略

#### 3.1 用户标签特征属性

用户的特征属性可以是事实的，也可以是抽象的；可以是自然属性，比如性别，年龄，星座等，可以是社会属性，比如职业，社交，出生地等；还可以是财富状况，比如是否高收入人群，是否有豪车豪宅等固定资产，对于机票用户来讲位置特征也是比较重要的属性，比常驻地，常出差地，老家等。这些属性都可以清楚的描绘一个用户的画像特征。



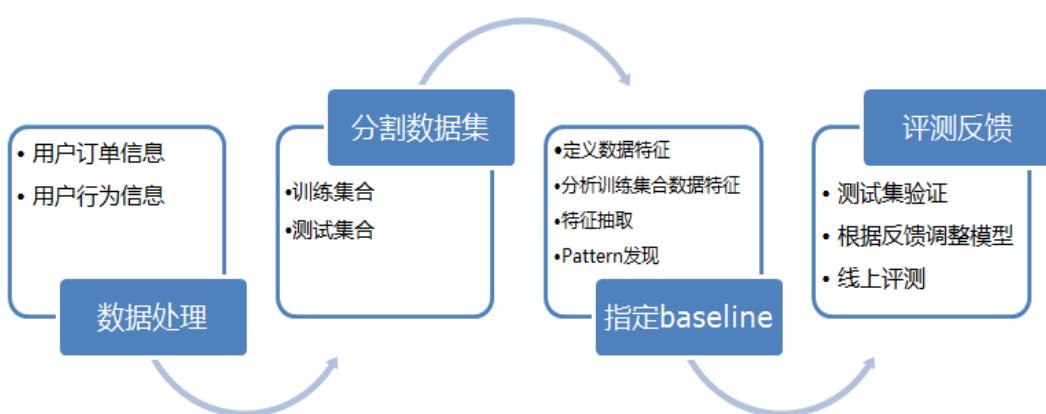
- 画像标签一般根据公司的业务体系来设计，存储有HDFS，HBASE，ES
- 标签的更新频率：每日更新，每周、每月更新
- 标签的生命周期：有的数据随时间衰减迭代

## 3.2 用户标签分类及特征项



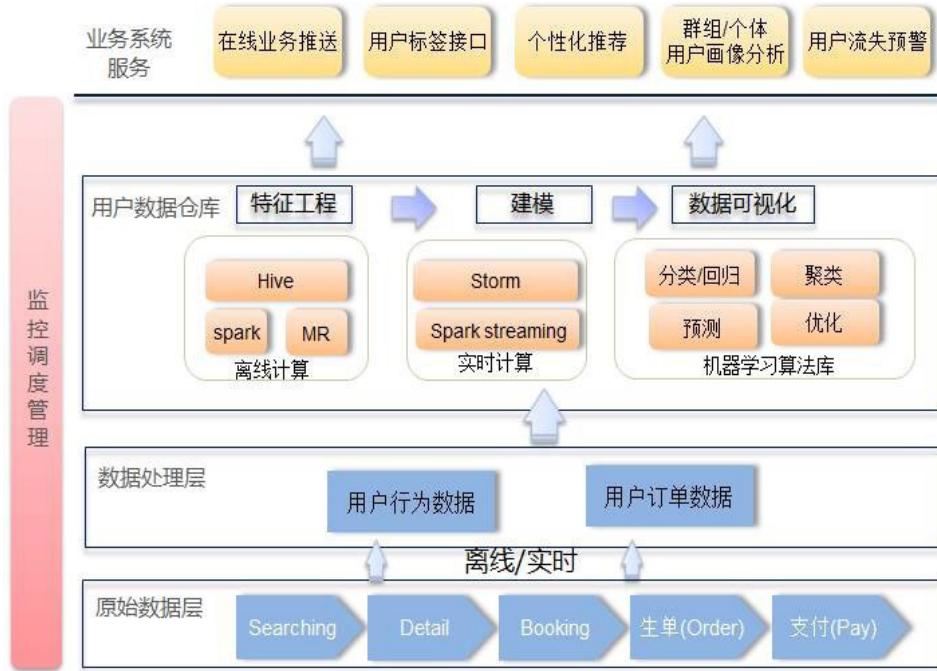
提到用户画像就不得不提到一个词“标签”。标签是表达人的基本属性、行为倾向、兴趣偏好等某一个维度的数据标识，它是一种相关性很强的关键字，可以简洁的描述和分类人群。标签的定义来源于业务目标，基于不同的行业，不同的应用场景，同样的标签名称可能代表了不同的含义，也决定了不同的模型设计和数据处理方式。我们给机票用户画像打标签分类为两大类，基础类标签和个性化标签，这些标签可以有重复，但是都是通过不同的角度去定义和刻画一个用户，来满足不同的业务营销需求。

## 3.3 用户标签库构建流程

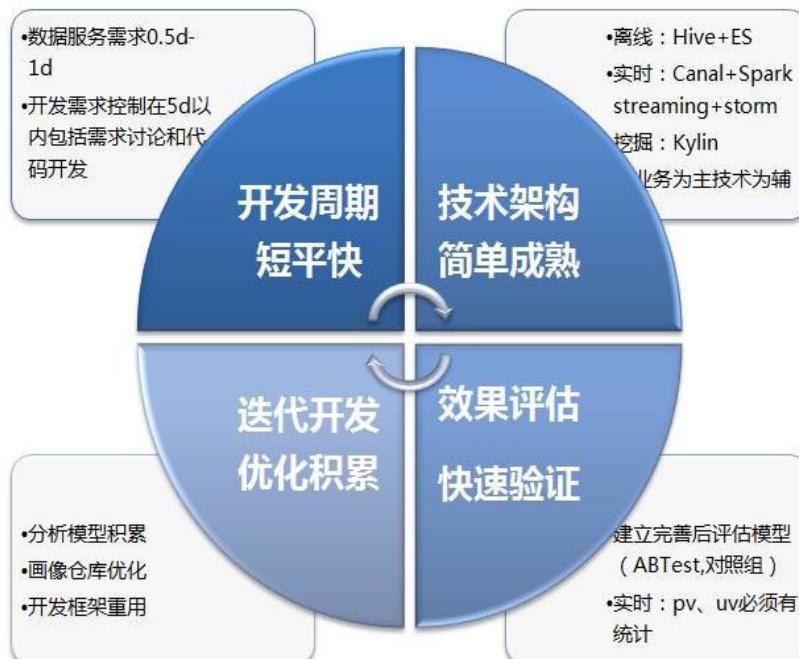


## 4 用户画像技术架构

### 4.1 技术架构



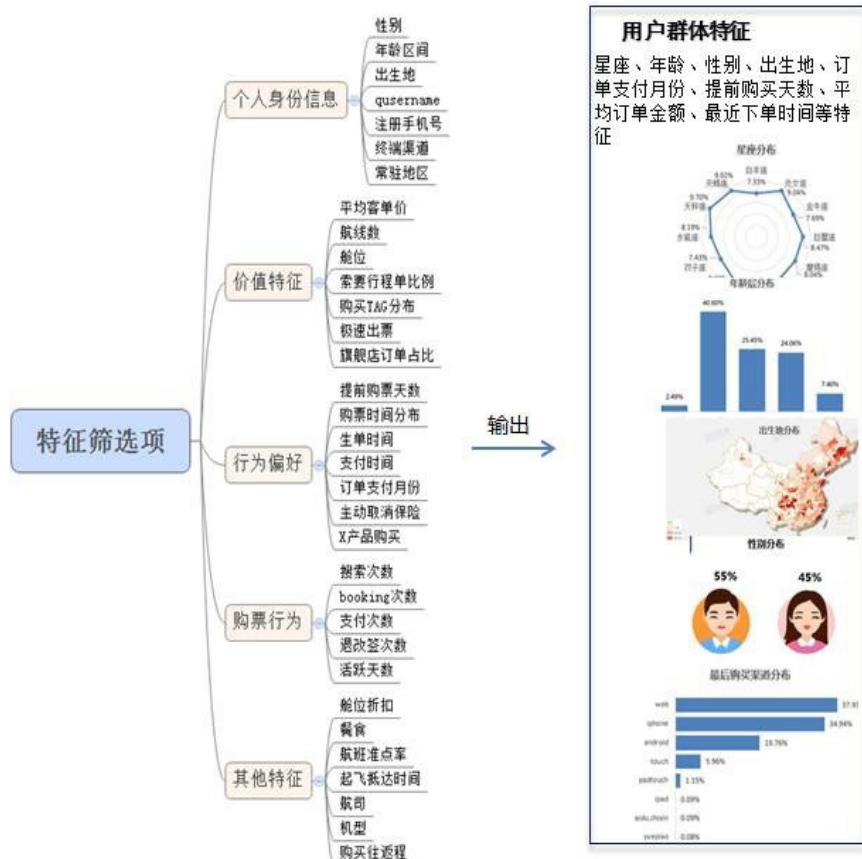
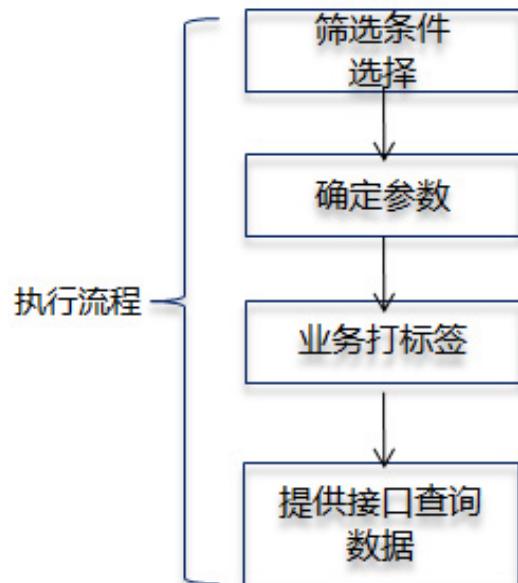
### 4.2 实施方法论



# 5 用户画像数据应用实践

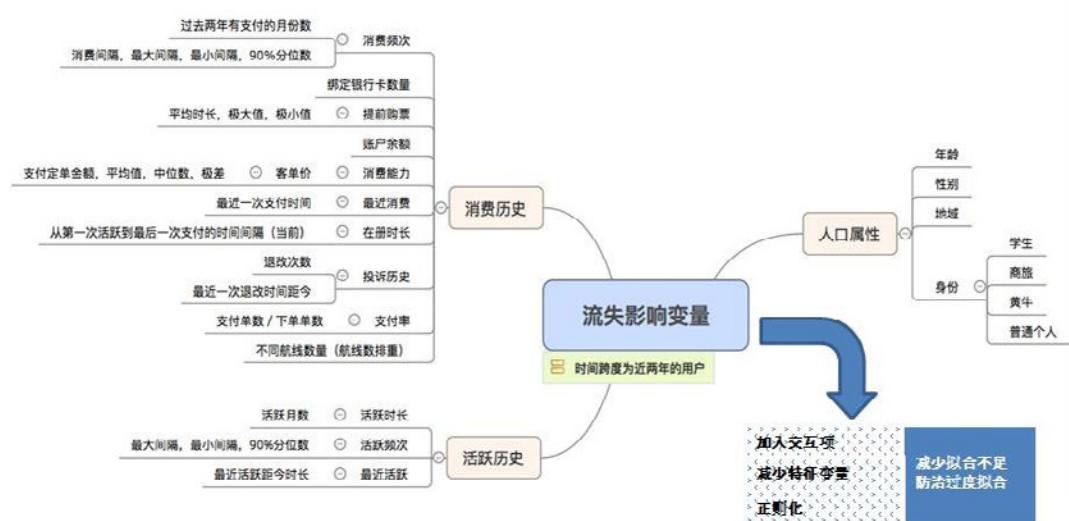
## 5.1 用户群体特征分析

- 设计目标
  - 根据条件可选项，输出筛选用户群体
  - 图形展示用户群体属性特征
- 应用场景
  - 如果筛选的用户群组满足业务的要求，将筛选条件形成参数
  - 根据参数提供接口查询



## 5.2 客户行为预测

客户行为预测建立步骤：



- 建模数据准备
- 客户流失节点判断
- 模型应用变量确定
- 模型构建
- 模型应用
- 模型验证

可以对用户流失做及时预测指导建议用户维系运营。

## 5.3 数据和业务在一起

用户画像与业务产品互相依赖，相辅相成：

- 用户画像标签库丰富优化
- 快速提供数据服务
- 数据分析+机器学习+模型训练



## 6 总结

用户画像作为大数据的根基，它完美的描述了一个用户的信息全貌，为进一步精准、快速的分析用户行为、消费等重要信息，用户画像仓库同时也提供了足够的数据基础，让我们去哪儿网更好的为用户提供高价值的服务，满足用户智慧出行的需要。

## 作者简介

**李国芳**，去哪儿网机票大数据组，精于架构，追求极致。曾先后搭建起机票实时数据处理系统，并主导用户画像项目，指导各业务线精准定位用户。在数据仓库方面，搭建 saiku + kylin + hive 平台，为运营及业务分析人员提供更快速精确的 OLAP 工具。技术涉猎广泛，乐于分享和激励新人。



# 40亿移动设备的用户画像和标签架构实践

王鹏

说起大数据的应用可能很多朋友们脑子里边第一映像就是画像，我想从以下几个方面跟大家聊聊画像相关的事情：1、什么是画像；2、画像的用处；3、如何进行用户画像；4 画像应用中的难点。

**什么是画像呢？**可能大家看到过一些外文资料或者演讲中出现过 profile 一词，其实和画像是一个概念，都是从不同的维度来表达一个人，这些维度可以是事实的，可以是抽象的；可以是自然属性，比如性别、年龄；可以是社会属性，比如职业、社交特征；可以是财富状况，比如是否高收入人群，是否有固定资产；可以是家庭情况，比如是否已经结婚，是否有孩子；可以是购物习惯，比如喜欢网购还是喜欢逛商场；可以是位置特征，比如在哪个城市生活；可以是其他行为习惯。总之，所有大家能想到的描述一个人的特征的都可以算作是画像的范畴，画像其实就是想方设

法用数据来描述人的特征。

**画像有什么用处呢？**大家之所以要进行用户画像，就是为了解决业务问题，或者为了拓展一个新用户，或者为了获得一个新订单。想要获得新用户，首先必须知道自己产品定位的用户画像（也就是用户长什么样子，有什么行为特征），而很多产品设计的时候可能由于定位不清晰，对用户的了解不够，导致最后产品上线后效果与预期大相径庭。

这里举一个例子，A 银行的电子支付团队计划与 Uber 公司合作，在春节后以短信推送优惠券的方式进行营销，选择了多类人群进行投放，其中有“有打车需求且有车”和“有打车需求且无车”两类人群，本以为“有需求且无车”人群的广告触达的营销效果会更好，结果却完全相反，“有需求且有车”人群的广告触达的比例反而最高。这可能映射出无论是开车还是打车，习惯了车反而离不开车。用数据来画像正是帮助企业了解用户和定位产品的最直接的方法。

综上我们可以看到要向更好的解决业务问题，首先必须明确业务目标，而用户画像是帮助企业明确目标客群的重要手段之一。当企业了解了自己的用户都长什么样子以后，接下来的任务就是如何将有类似画像特征人群的潜在用户变成自己的用户，也就是在营销上获新客的过程。所以，从大的框架来看，用户画像承载了两个业务目标：一是如何准确的了解现有用户；二是如何在茫茫人海中通过广告营销获取类似画像特征的新用户。

如果仔细琢磨这两个目标，其实在根源上逻辑是有些相悖的。了解现有用户的画像，需要的是少量、画像特征覆盖度全面的无倾斜的精准样本，这样能更精确的定位产品的用户。而通过画像结果做广告营销获取新用户，在一定程度上需要的是大量的相似样本。量的大小和精准度的不同决定了后续画像模型在应用设计中的不同。

提到用户画像就不得不提到一个词“标签”。标签是表达人的基本属性、行为倾向、兴趣偏好等某一个维度的数据标识，它是一种相关性很强的关键字，可以简洁的描述和分类人群。标签的定义来源于业务目标，基于不同的行业，不同的应用场景，同样的标签名称可能代表了不同的含义，也决定了不同的模型设计和数据处理方式。

举个例子，如果一款卖男装的 app 想在近期做营销，只筛选“男性”和“网络购物”这两个标签进行投放，可能效果并不一定理想。因为“性别（男 / 女）”可能有多种维度，真实性别男女是一种维度，网络购物特征男女是一种维度，性取向男女可能又是另外一种维度。因为网络的发展，你甚至都不知道网络的另一端是不是一个人，更何况是男女呢。想要正确的设计标签模型和计算处理数据，必须了解画像标签应用的场景和目标。

**如何进行用户画像呢？**这完全取决于业务目标（需要什么样的画像标签）和有什么样的原材料（有什么类型的数据源），基于这两样才能确定使用什么样的模型设计和数据计算处理方式。就像做菜一样，要做一顿美味的晚餐，必须知道客户是想吃中餐还是西餐，配菜都有哪些鱼蛋肉和蔬菜，然后才能确定牛肉是红烧还是煎炸。

仍然以性别（男 / 女）为例，尝试演绎一下刚才的三个场景。

如果业务是征信场景，想知道的是这个人的真实性别（男 / 女），在没有全量真实数据的前提下可以采取如下的方法来处理，可以选取少量真实样本，使用这些真实样本追加一些特征因子，使用 lookalike 算法进行样本扩展，将该少数样本特征扩展到大量或者全量数据。当然，这些数据的准确度取决于样本的均衡程度和算法的质量。

如果业务是网络购物的电商场景，我们先不尝试判断真实购买男装的是否是男性（很多已婚人士是妻子负责网购丈夫的装备），仅仅考虑将来

该网络账户实体是否会购买男装的角度考虑，需要的是“男装购买倾向”的标签，可以直接基于所有账户实体以往购买记录来计算处理该标签。

如果是业务场景是 blued（一款同志交友 app）定义的男性又是另外一个特殊群体，基于客户想拓展新客，这里定义的特殊男性群体或许可以定义为“男性同志”标签，而实现该标签可以考虑通过安装了类似同志交友的 app 人群或者以同志人群经常出现的聚集地进行计算处理。

所以说针对不同的行业，不同的应用场景，需要使用不同的数据源进行不同的标签设计和计算。

说起标签，可能每个行业有每个行业的标签体系，各个公司基于自己的数据源和特征不同也设立了不同的标签体系。我认为这些标签都可以归纳为以下几个方面。

- 人口属性：包含性别、年龄等人的基本特征
- 资产情况：车辆、房产、收入等资产特征
- 兴趣特征：阅读资讯、运动健康等兴趣偏好
- 消费特征：网上/线下消费类别品牌等特征
- 位置特征：常驻城市、居住距离等
- 设备属性：所使用终端的特性等

要支持以上这些标签的设计和计算，需要多种维度的数据源，从产生维度来看：可以包含 PC 端的数据、移动终端的数据、线下的数据；从数据拥有者来看：可以包含一方客户自己的数据、外部官方渠道的数据、市场采集的数据；从数据类型来看：有社交数据、交易数据、位置数据、运营商数据等。

使用这些不同源的数据，我们如何计算处理业务需要的标签呢？一般都会经过如下几步：

- 数据抽取：从不同数据源抽取要计算标签的数据原材料。
- 数据标准化：针对抽取的数据将其清洗为标准格式，将其中的错误数据和无效数据剔除。
- 数据打通：不同来源的数据有不同的主键和属性，如何将这些数据关联起来是数据打通的关键，比如有设备的wifi信息，又有设备的poi信息，就可以通过wifi将设备终端和POI建立起关联。
- 模型设计：针对不同的数据内容和业务目标设计不同的规则和算法进行模型的构建，并使用小样本数据来验证模型的可靠性。
- 标签计算：在模型可靠性验证的基础上，部署生产运营环境来进行标签计算。

一般标签计算无外乎以上过程，以“大学生”标签为例，假如我们需要针对移动终端人群设计一个大学生标签，而我们并没有每个大学生的入学信息和证件信息，我们该如何操作呢。首先进行业务分析，发现大学生的行为特征，一般大学生都会在大学校园内活动比较多，我们可以将全国2000多所高校的位置找到，根据移动终端设备的位置信息来筛选“大学生”人群；另外大学生可能还会使用一些特殊的App比如考研类、四六级、超级课程表等这些特殊App，我们可以通过App进行“大学生”人群的筛选。

如果不用算法，就只用规则，我们想找精确的“大学生”人群，可以将位置和App行为两个特征叠加使用；如果我们想要扩展样本进行大规模广告投放，可以考虑含有位置、app行为任意一个特征的人群，同时还可通过算法进行lookalike的扩展样本学习。

注：以上表达的都是数据和标签处理的逻辑过程，实际业务中的数据处理要视具体情况而定。

最后说一说用户画像和标签设计 / 计算中的一些难点。

## 1. 如何定义画像主体？也可以理解为如何唯一标识一个实体？

可以理解真实世界每个人都只是一个实体，但是虚拟世界他可能就变身为多个，比如人可能有一个身份 ID，但是可能有多个手机，就对应了多个手机号，多个设备终端 ID，那就对应多个移动终端的使用行为；这多个终端 ID 分别代表了这个实体的不同特征，只有将这个实体拼接起来才能代表完整的画像。一个人可能有多个 qq 号，如果从 qq 行为的角度分析，同样的逻辑。这是终端实体多对一的体现。

反过来也会有一对多的情况，比如就一个家庭用的 ipad，孩子用 ipad 来玩游戏，父亲用 ipad 来查收邮件，母亲用 ipad 来购物，这一个 ipad 代表了多个实体的行为特征，并且无法分拆。所以要想唯一完整的定义一个实体其实很难。所以在业务领域中追求标签的完整性有时候是一个很难达到的目标，反过来应该更多的关注标签的代表性，无论是一对多还是多对一，只要能通过标签筛选出来想寻找的受众群体就可以，即便是家庭公用的 ipad，有游戏标签也表明了家庭中有成员有该方面的兴趣偏好。

## 2. 如何打通不同源的数据？

pc 端的行为信息、移动终端的行为信息和 TV 端的行为信息，如何将这些信息关联起来？核心问题在于如何将这些终端的唯一标识 ID 打通。TalkingData 的数据体系已经建立了以 TDID 为核心 ID 的关联图谱，TalkingData 的 IDmapping 能力已经实现了跨设备 ID 的关联映射。所以要解决不同源 ID 的打通只要接入一家类似 TalkingData 的数据即可。

## 一些问题

**Q1：**画像的时候常用的算法有哪些，比如什么类型的标签适合什么

类型的算法？怎么评估画像画的好坏？

**王鹏：**我们除了用常规的算法以外，还有自研的开源算法系统fregata，基于spark，支持10亿样本1亿维度的超大规模运算，无需调参，超高速度。

评估画像的好坏：小样本的真实验证；在实际的case中迭代验证。



**Q2：** 画像的标签体系一般怎么设计，还有你们的标签体系如何存储？更便于不同画像标签之间的追溯？

**王鹏：** 画像的体系一般参考你们公司的业务体系来设计。标签的存储也取决于你的服务应用场景：我们的存储有多种：hdfs、vertica、hbase。标签的追溯属于另外一个问题，取决于你的标签的生命周期，有的标签就是最新的，有的标签就是每周每月加工的，有的标签是有时间衰减迭代的。

**Q3：** 标签的确定一般是怎么样的？是人工打标签，想业务场景，还是通过自动的算法跑出来？TalkingData目前又是如何实现的呢？

**王鹏：** 标签的确定，一般是先人工筛选小样本规则，进行验证，规则

合理后，在通过算法扩展。

人工和自动是结合的，取决于该标签的具体场景、字典数据的使用、主数据的量等各种因素。

TD 的标签也是结合这二者一起完成的。

**Q4:** 在做用户画像时如何解决数据准确性不足的问题，毕竟非 BAT 公司太多，很多数据都不完善，或者说没有准确数据。

**王鹏：**说到这个问题，谈谈我对大数据的看法：我理解大数据本身不存在所谓的正确性，大数据是用来验证人的先验知识 / 经验的一种工具，这个里边应该考虑的不止是准确性的问题，而是如何能更好的提高你认为的准确率的问题，大数据由于体量大，需要的是数量、时间等多维的迭代，维度的扩展。

## 作者简介

**王鹏**，TalkingData 数据产品总监 & 数据负责人，负责公司数据架构设计及质量管理，自有数据及第三方数据的收集、处理、加工全过程管理，数据标签和数据市场等产品的管理工作。曾任四维图新数据中心品保部经理、阿里 - 高德地图数据产品事业部数据产品总监等职位。在数据领域有深厚的积淀，对数据应用和数据管理有深刻的见解。



# 携程是如何做用户画像的

周源

用户画像作为“大数据”的核心组成部分，在众多互联网公司中一直有其独特的地位。作为国内旅游 OTA 的领头羊，携程也有着完善的用户画像平台体系。目前用户画像广泛用于个性化推荐，猜你喜欢等；针对旅游市场，携程更将其应用于“房型排序”“机票排序”等诸多特色领域。

本文将从目的、架构、组成等几方面，带你了解携程在该领域的实践。

## 1 携程为什么做用户画像

首先，先分享一下携程用户画像的初衷。一般来说，推荐算法基于两个原理“根据人的喜好推荐对应的产品”“推荐和目标客人特征相似客人喜好的产品”。而这两条都离不开用户画像。

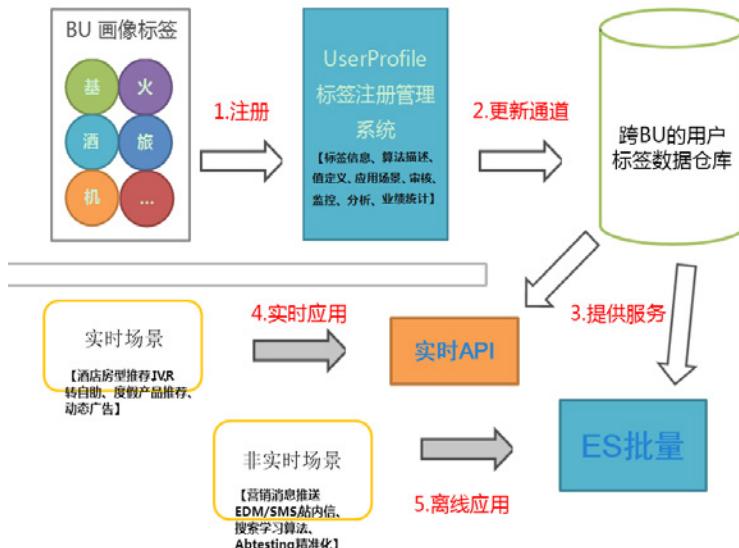
根据用户信息、订单、行为等等推测出其喜好，再针对性的给出产品可以极大提升用户感受，能避免用户被无故打扰的不适感。同时针对不同

画像的用户提供个性化的服务也是携程用户画像的出发点之一。

## 2 携程用户画像的架构

### 2.1 携程用户画像的产品架构

如下图所示，携程用户画像的产品架构大体可以总结为：



1. 注册
2. 采集
3. 计算
4. 存储/查询
5. 监控

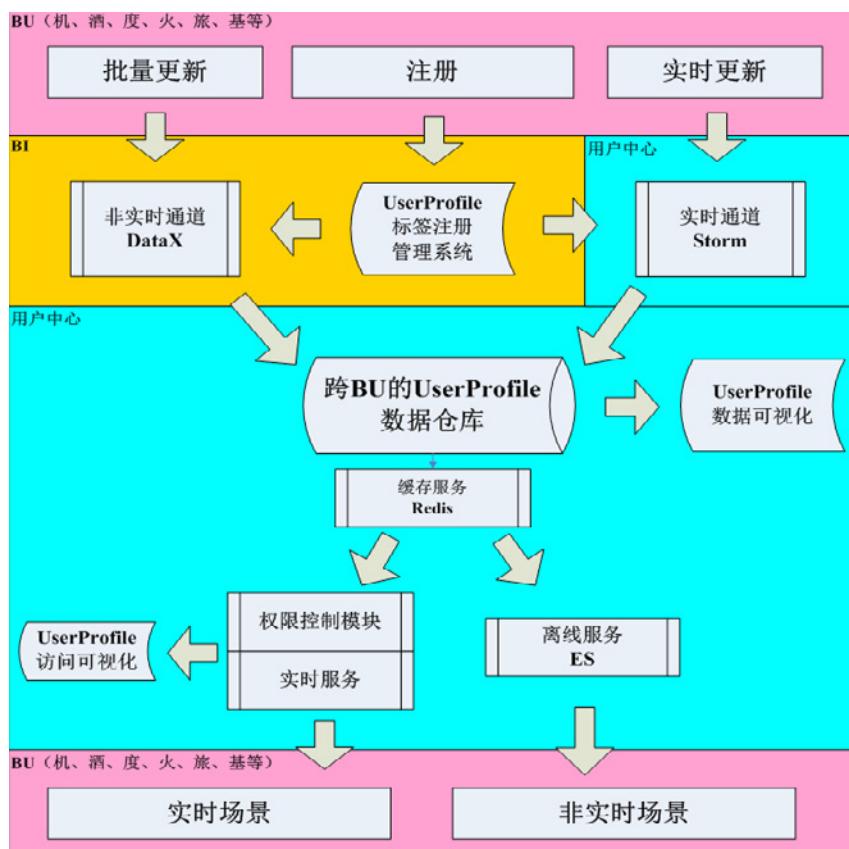
所有的用户画像都会在”UserProfile 平台”中进行注册，由专人审核，审核通过的画像才可以在“数据仓库”中流转；之后会通过用户信息、订单、行为等等进行信息采集，采集的目标是明确的、海量的、无序的。

信息收集的下一步是画像的计算，携程有专人制定计算公式、算法、模型，而计算分为批量（非实时）和流式（实时）两种，经过严密的计算，画像进入“画像仓库”中；而根据不同的使用场景，我们又会提供实时和

批量两种查询 API 供各调用方使用，实时的服务侧重高可用，批量服务侧重高吞吐；最后所有的画像都在监控平台中得到有效的监控和评估，保证画像的准确性。

## 2.2. 携程用户画像的技术架构

携程发展到今天规模，更强调松耦合、高内聚，实行 BU 化的管理模式。而用户画像是一种跨 BU 的模型，故从技术架构层面，携程用户画像体系如下图所示。



各 BU 都可以贡献有价值的画像，而基础部门也会根据 BU 的需要不断制作新的画像。画像经过开源且经我们二次开发的 DataX 和 Storm 进入携程跨 BU 的 UserProfile 数据仓库。在仓库之上，我们会有 Redis 缓存

层以保证数据的高可用，同时有实时和借助 elasticsearch 两种方式的 API，供调用方使用。

该架构有如下关键点：

1. 有异步和实时两种通道满足不同场景、不同画像的需要，事实类画像一般采用实时计算方式，而复合类画像一般采用异步方式。
2. 携程强调专人专用，每个人做自己最适合的事。故整个 UserProfile 是多个团队合作完成的，其中包括但不限于各 BU 的开发、BI，基础的开发、BI 等。
3. 所有 API 都是可降级、可熔断的，可以根据需要切数据流量。
4. 由于用户画像极为敏感，出于数据安全的考虑，我们查询服务有严格的权限控制方案，所有信息必须经过授权才可以访问。
5. 出于对用户画像准确性负责的目的，我们有专门的 UserProfile 数据可视化平台监控数据的一致性、可用性、正确性。

上述是用户画像的总体描述，下面我将详细分享各个细节。

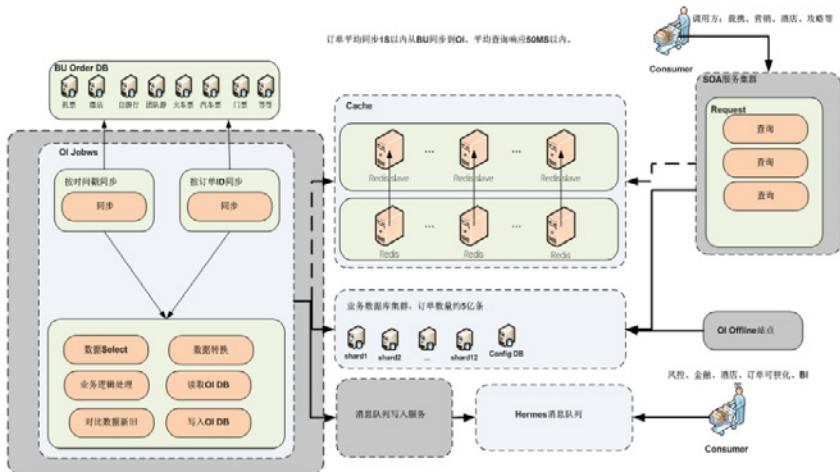
### 3. 携程用户画像的组成

#### 3.1. 信息采集

基础信息的采集是数据流转的开始，我们会收集 UserInfo（比如用户个人信息、用户出行人信息、用户积分信息）、UBT（用户在 APP、网站、合作站点的行为信息）、用户订单信息、爬虫信息、手机 APP 信息等。而上述每个基础信息的采集又是一个专门领域。比如下图展示了用户订单信息采集流程。

#### 3.2. 画像计算

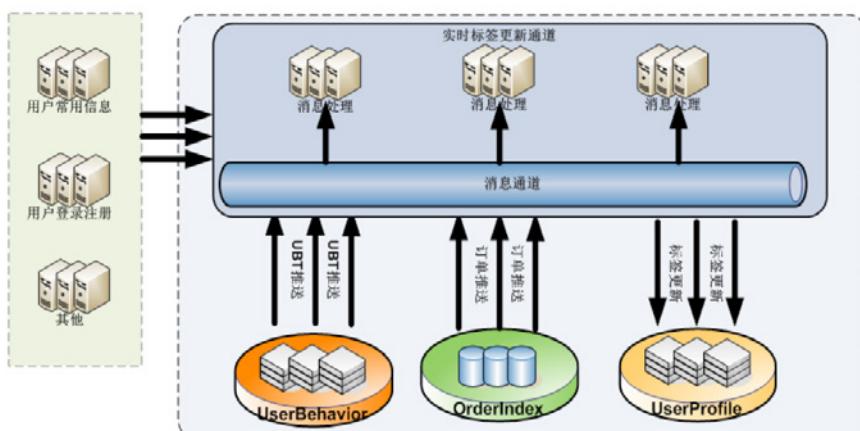
基础信息是海量的、无序的，不经加工没有太大的价值。故用户画像



的计算是数据流转的关键所在。我们的 BI 团队会制定严密的公式和模型，根据场景的需要，制定规则和参数，对采集信息做异步计算。这样的计算由于耗时较长，一般我们会采用 T+N 的方式异步更新，根据画像的不同，数据新鲜度的要求亦不同。动态和组合标签大多采用异步方式计算更新。Hive、DataX 等开源工具被使用在这个步骤中。

而有些画像是事实或对新鲜度要求比较高的，故我们会采用 Kafka+Storm 的流式方案去实时更新计算。比如下图，UBT（用户行为数据）使用消息通道 Hermes 对接 Kafka+Storm 为 UserProfile 的实时计算提供了有力的支持。

### 3.3. 信息存储



用户画像的数据是海量的，被称作最典型的”大数据”，故 Sharding 分布式存储、分片技术、缓存技术被必然的引入进来。

携程的用户画像仓库一共有 160 个数据分片，分布在 4 个物理数据集群中，同时采用跨 IDC 热备、一主多备、SSD 等主流软硬件技术，保证数据的高可用、高安全。

由于用户画像的使用场景非常多、调用量也异常庞大，这就要求用户画像的查询服务一定要做到高可用。故我们采用自降级、可熔断、可切流量等方案，在仓库前端增加缓存。如下图所示，数据仓库和缓存的存储目的不同，故是异构的。



### 3.4. 高可用查询

响应时间和 TPS 是衡量服务可用性的关键指标，携程要求所有 API 响应时间低于 250ms (包括网络和框架埋点消耗)，而我们用户画像实时服务采用自降级、可熔断、自短路等技术，服务平均响应时间控制在 8ms (包括网络和框架埋点消耗)，99% 响应时间控制在 11ms。

大部分场景都是通过单个用户获取用户画像，但部分营销场景则需

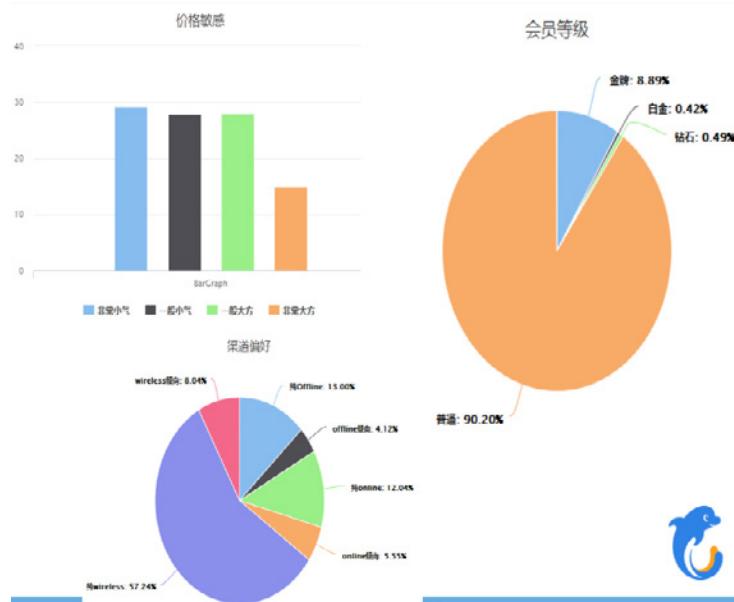
要满足特定画像的用户群体，比如获取年龄大于 30 岁、消费能力强、有亲子偏好的女性。这种情况下会返回大量用户，此时就需要借助批量查询工具。经过多次技术选型，我们决定采用 elasticsearch 作为批查询的平台，封装成 API 后很好的支持上述场景。

### 3.5. 监控和跟踪

Fail 率	成功率	应用服务器		数据库服务器 (参考 dbmonitor)		平均请求 响应时间 (ms)	响应时 间-90% Line
		应用服务 器 CPU	内存 (max-min)	CPU (增量)	IOPS (增量)		
0.00%	100.00%	3.00%	450M			8	11

在数据流转的最后，数据的准确性是衡量用户画像价值的关键指标。基于高质量信息优于大数量信息的基调，我们设置了多层监控平台。从多个维度衡量数据的准确性。同时我们还要监控数据的环比和同比表现，出现较大标准差、方差波动的数据，我们会重新评估算法。

上述所有环节组成了携程跨 BU 用户画像平台。当然技术日新月异，



我们也在不断更新和局部创新，或许明年又会有很多新的技术被引入到我们用户画像中，希望我的分享对你有所帮助。

## 作者简介

**周源**，携程技术中心基础业务研发部高级研发经理，从事软件开发10余年。2012年加入携程，先后参与支付、营销、客服、用户中心的设计和研发。

投稿或寻求报道可联系

tina.du@infoq.com





# 百分点苏海波博士： 为什么你做的用户画像模型不精准？

苏海波

对企业而言，得用户者得天下，能够有一套科学的精准营销、个性化推荐模型，无疑会促进业务的增长；对开发者而言，用户画像也是频繁被提及的技术，这样可以根据目标用户的动机和行为上进行产品设计，远远优于为脑中虚构的东西做设计。

用户画像的应用场景甚多，但即使是从事这方面研发的人，对其内部逻辑也是似是而非。大家都希望自己的用户画像模型更加精准，如何做到？这就要深入解剖，理解用户画像与标签的关系、根据何种理论建模更加有效？大数据时代，需要上帝的视角，有了科学的大数据思维方法和理论指导，才能在结合实际业务建模中游刃有余。

## DT 时代要从比特流中理解人类行为

水有源木有本，之所以需要用户画像，是因为 DT 时代相较传统 IT 时

代发生很大变化：DT 时代的数据是现实世界的虚拟化表现，数据本身构成了一个虚拟世界，这使得 IT 系统构建在虚拟系统上，也变得更加智能。

尤其表现在信息化建设、可穿戴设备、信息网络的发展，使全社会的信息化程度越来越高，越来越多的业务需要计算机应用，将设备和人连接在一起，用户与这些应用、设备交互中产生大量数据。

在这种社会科技发展趋势下，人与人沟通的方式发生了根本变革，这就导致“要学会从比特流中解读他人”，因此要构建用户画像；但数据这么大，人工显然无法应对，所以“还要教会机器从比特流中理解人类”，再在画像的基础上构建一些应用，比如个性化推荐、精准广告、金融征信等，进行机器与人的交互。

## 你真的理解用户画像是什么意思吗？

用户画像、标签、360 度用户视图等这些词经常被提起，但实际上连从事研发工作的人，对这些概念也不甚了解。要想搞清楚，还需要从理论层面解读。

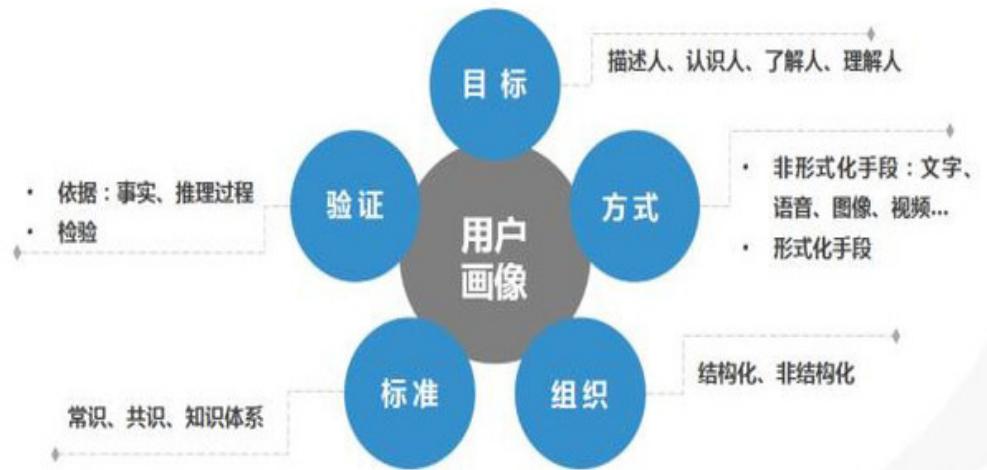
用户画像从某种程度上说来源于对事物的描述，但每个人描述事物的方式和角度不一样，梳理共性，可将用户画像分为五个层次：

第一个是目标，目标都是为了描述人、认识人、了解人、理解人。这是用户画像最大的目标。

第二是描述的方式，分为非形式化（语音、文本、视频、图像……）和形式化（读卡器读取信息的形式）两种手段。

第三是组织方式，就是结构化和非结构化的组织方式，我们前面看到的球员数据它就是结构化的。

第四个就是用户画像标准，包括常识、共识、体系。这个很重要（比



比如说某个人特别二次元，这个词对方就可能听不懂，是因为双方对二次元这个词没有达成共识，所以必须有一套达成共识的知识体系，不然用户画像这件事是没办法达到的。）

最后一个就是验证，依据：事实、推理过程、检验。为什么一定要验证？举个例子，比如说某个人“特别不靠谱”，相当于打上标签，但会被反问为什么不靠谱、依据是什么？所以要提前验证，否则会丧失可信力。

据此，可以得出用户画像的定义：用户画像是对现实世界中用户的数学建模。

一方面，用户画像是描述用户的数据，是符合特定用户需求的形式化描述。从业务中抽象出来，可以形容为“来源于现实，高于现实”。另一方面，用户画像是一种模型，是通过分析挖掘用户尽可能多的数据信息得到的。对数据做抽象，可以形容为“来源于数据，高于数据”。反过来，根据这个模型，可以挖掘出更多用户画像。

## 如何构建用户画像？

在 90 年代流行一种”本体论”方法，但非常复杂。所以重点来了，



## 大数据 + 洞察

用户画像构建需要根据一套原则，在这里分享一套相对朴素的方法：

朴素的知识表现方法：符号 – 概念法。符号与概念是相对应的，比如，狗这个词是一个符号，但人们脑子中的概念是”四条腿、看家的、一个能汪汪叫的动物”。

朴素的用户特征表现方法：标签 – 模型法。标签的定义是用户特征的符号表现，模型定义是经验总结的用户特征。什么是标签？举例来说，比如”收入高、坐办公室” 这个群体可以打上白领这个标签；同时标签是跟业务场景绑定在一起的，脱离业务场景的符号没有明确的含义。比如在阿里内部，关于男女，这样最简单的标签，也有 12 个男和女，它与业务密切相关，不仅仅是指生理上的男和女，还包括在互联网喜欢买男性的商品或者女性的商品定义的男女等等。

那么，用户画像和标签有什么关系？其实二者是整体和局部的关系，用户画像是整体，标签是局部，而整体和局部的关系可以通过“标签体系”体现。

根据这个逻辑，可以得出，用户画像可以用标签的集合来表现，即“标签体系”方法，用户画像（整体）和标签（局部）还包含两方面的关系：



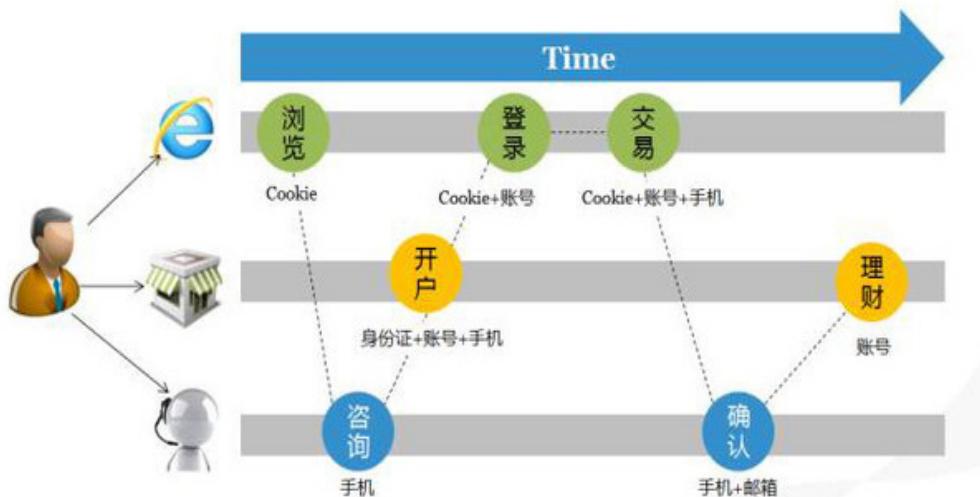
化整为零，整体如何反映在局部；化零为整，局部如何组成整体。

举例来说：“人都有一双眼睛一个鼻子”，化整为零来看：应该观察到每个人都有一双眼睛和一个鼻子；化零为整：只有位置合适的一双眼睛和一个鼻子才被认为是一个人。

至于标签体系，因为标签是和业务密切相关的，对应的标签体系也要搜集所有业务方的需求，制定出标签体系后，给每一个标签标准进行定义，最后进行标签开发。

另外，在用户画像建模方面，可以将标签建模分为四层：第一层是事实类标签，譬如用户购物了什么品类；第二层是机器学习模型的预测标签，





譬如当下需求、潜在需求等；第三层是营销模型类标签，譬如用户价值、活跃度和忠诚度等；第四层是业务类的标签，譬如高奢人群、有房一族等，它是由底层的标签组合生成的，通常由业务人员定义。

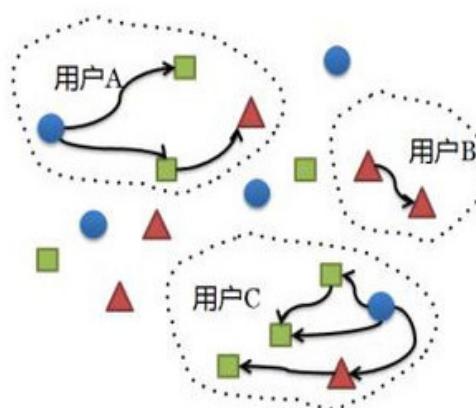
最后是验证，对模型的验证可以分成两个方面，一个是准确率的验证，标签打得准不准；第二个是标签打得全不全。但这两个方面没有办法同时满足的。现实业务中无法追求 100% 完备的标签体系。不过，目前谈得最多的是准确率。其分为两种，一种是有事实标准的，譬如生理性别；另外一种是无事实标准的，譬如用户的忠诚度，只能验证过程，具体效果需要通过线上业务 A/B Test 进行验证。

## 构建用户画像的关键难题：需要上帝视角

要想精准构建用户画像还面临着许多技术难题，比如用户多渠道信息打通、多渠道的产品打通、实时采集用户数据，以及用户数据挖掘建模等方面。重点解读下用户多渠道信息打通和多渠道的产品打通两个关键问题。

首先是用户多渠道信息打通，大数据时代我们需要上帝视角。

因为用户与企业的触点非常多，譬如手机、邮箱、Cookie 等，要将同一个用户的多个触点进行打通。方法就是把用户 ID 视为图中的顶点，



## 图中

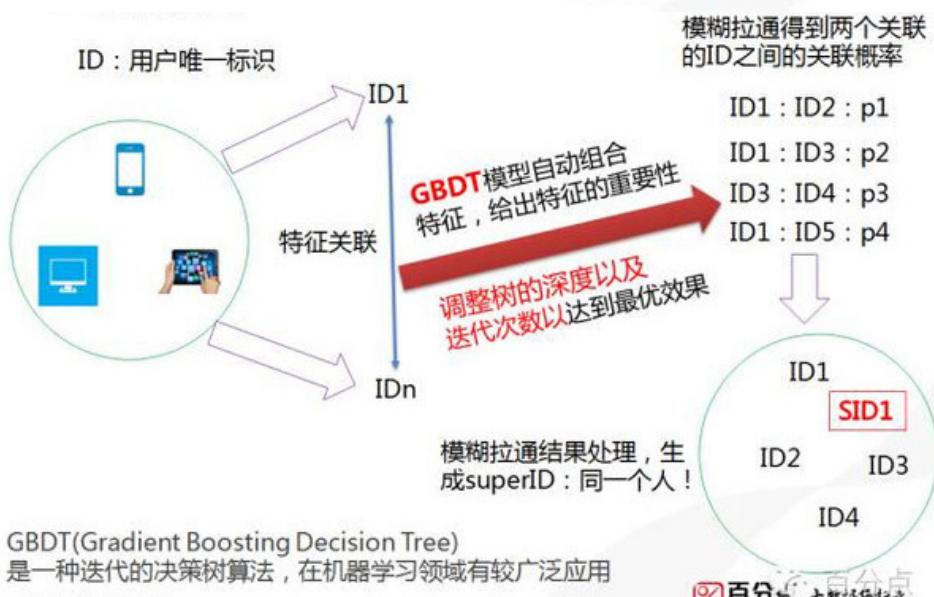
- 有三类ID
- ID间有相互联系
- 相互联系的ID反映出它们很有可能代表同一个用户

## 低密级业务

- 可以仅使用多重ID中的任意一个
- 最大程度打通，跨平台一致体验

## 高密级业务

- 使用特定ID，或者多种ID的组合
- 保证数据的准确和安全



如果用户的两个触点在同一个场景出现（比如用邮箱登陆），那么就可以把在用户的邮箱和 Cookie 用一条边进行连接，从而构建一张图。

用户打通可以基于图例的方法进行强拉通，也可以采用机器学习方法进行模糊拉通，预测出拉通的概率。

除了用户打通，不同渠道的产品也需要拉通，可采用标签体系拉通方法：建立一套标准的分类标签体系，比如一颗分类树，任何商品都能划分到这个分类树的叶子节点。根据百分点的实践经验，手工映射的方法成本

高、难以大规模开展，实际工作中会采用机器学习模型 + 少量的人工规则来实现。

但要实现自动分类，其中难点不在于模型，而在于获得训练数据、feature engineering，以及分类树层级节点之间的依赖问题。

## 用户画像应用，是业务和技术的最佳结合点

可以说，“用户画像”在行业应用中算是曝光率最高的技术之一，有很多用武之地，总结来说，包括：售前的精准营销、售中的个性化推荐，以及售后的增值服务等；用户画像的标签维度包括人口属性、上网特征、购物偏好等。

需要强调的是，标签和应用是相互相承的关系，一方面可以根据现有的标签维度开发应用，另一方面也可以根据应用的需求扩展标签的维度，两者互相促进。



首先，根据用户画像进行精准营销。不同于门户广告等 DSP 公司投放的程序化广告，百分点着眼点在于帮助企业整合、拉通自己的第一方数据，

建立企业用户画像、实现全渠道营销。

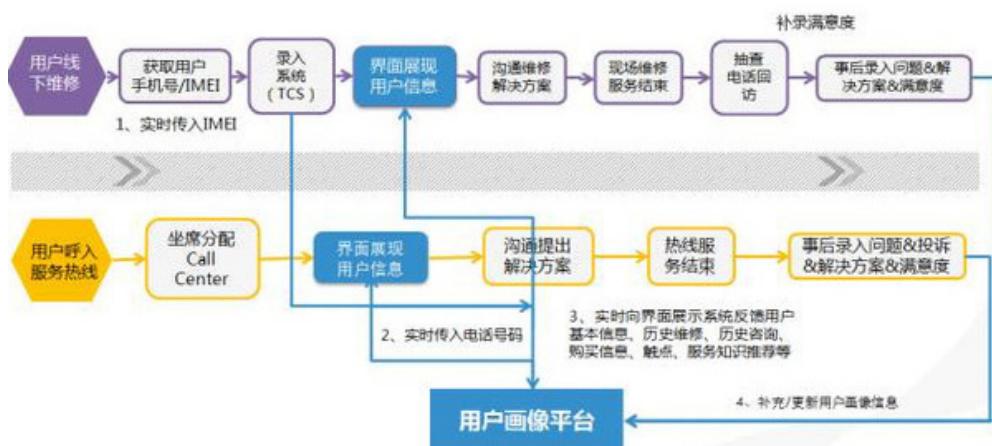
而且结合百分点的营销管家产品，可以实现触发式的营销。

比如，用户在某网站下单购买一款手机，便可以立马给他推送该品牌手机对应的手机配件广告。最终效果是，通过用户拉通用户画像，对 59 万个潜在消费者形成 4 个精准人群，并进行投放，是盲投点击率的 10 倍。

其次是售中的个性化推荐。这是百分点最开始创立时做的事情，目前已经服务超过 1500 家的电商和媒体客户，是国内最大的第三方推荐服务提供商。

值得一提的是百分点推荐引擎的设计架构，核心为四大组件：场景引擎、规则引擎、算法引擎和展示引擎，尤其是规则引擎非常强大，可以根据客户的业务需求可视化配置推荐逻辑，譬如推新品、清库存等等，而不仅仅是点击率最优。

比如百分点的某个团购网站客户，采用这个推荐引擎解决下单率的问题，通过分析发现了该网站用户的一系列特征，譬如忠诚度低、区域性购买等。



最后是如何结合用户画像提供“售后”增值服务。上图是百分点客户

的应用系统方案，通过数据接口实时反馈用户相关信息，包括历史维修、历史咨询并进行知识推荐等内容，支撑服务效率、提升客户满意度；同时收集用户的服务满意度数据，进一步补充、完善用户画像信息。

## 小结

在大数据时代，机器要学会从比特流中解读用户，构建用户画像变得尤其重要，是上层各种应用的基础。

用户画像不是数学游戏，而是严肃的业务问题。构建用户画像的核心是进行标签建模，标签不仅仅是个符号，更要和业务紧密关联，是业务和技术的最佳结合点，是现实与数据化的最佳实践。不断从更深的逻辑角度思考建模理论，并有效匹配业务应用，用户画像在实际业务中的重要价值将会越来越大。

## 一些问题

**Q1：** 用户画像有哪些常用算法，是否有成熟的开源算法？

**苏海波：** 用户画像里常见的分类算法和聚类算法都会用到，譬如 svm、lr 分类，k-means 聚类等，这些模型都有开源版本。

**Q2：** 用户画像的商业价值体现在哪些方面？

**苏海波：** 用户画像的商业价值和它支撑的应用密切相关，譬如应用于营销，可以提升广告效果和流量变现的效率。

**Q3：** 用户画像标签属性字段一般都去哪些系统里取呢 涉及多少系统比如在金融业数据的特点移动行业数据的特点？

**苏海波：** 用户画像的数据源可以从业务和 crm 等系统中提取，不同企业可能有区别。根据需要加工的标签去判断需要什么样的数据源。

**Q4：** 定义一个用户一般都用什么样的标签属性，有没有规律或者规

则？

**苏海波：**这个与业务密切相关，根据用途来制定。

**Q5：**请问用户画像如何用在实时系统中，如何快速查询画像数据？

**苏海波：**用户画像放在 Hbase 里，就可以快速查询微观画像，放在 Redis 里可以应用于实时业务。

## 作者简介

**苏海波**，百分点集团研发总监，清华大学电子工程系博士。擅长文本分析、机器学习，精于个性化推荐以及计算广告学；多篇论文发表于 GLOBECOM、ICC、IEICE Transactions 等国外顶尖学术会议和期刊；曾负责当当网百货搜索以及 AdSmart 广告系统的算法效果优化；曾负责新浪微博信息流广告产品整体算法策略的设计及研发。



# 易观用户画像实践

代立冬

随着当今互联网进入到存量发展阶段，企业进入到精细化运营阶段，如何识别自己的客户和潜在客户就显得尤为重要。这也是今天所要谈论的主题——用户画像。所谓用户画像，其实就是给客户打上不同的数据标签，形成个人画像，以便了解客户的行为特征和偏好，然后根据业务需要，挑选出目标客户群。

## 一 什么是用户画像

用户画像是根据用户的一系列行为和意识过程建立起来的多维度标签。

通常的标签分类如下：

1. 人口学属性：性别、年龄、学历、收入等
2. 兴趣偏好：爱玩篮球、德扑、打电竞等

3. 消费偏好：线上/线下的一些消费行为等
4. 位置信息：WiFi定位、常住城市、商圈等
5. 设备属性：品牌、机型、操作系统等

## 二 用户画像的应用分类

### 1. 风险控制

风险控制包括个人及企业级信用评分、欺诈识别，芝麻信用就是一个对个人信用评分的典型案例。

### 2. 个性化推荐

根据每个人的不同喜好推荐与之相关的内容，今日头条、天天快报是个性化推荐的典型。

### 3. 精细化运营

精细化运营包括产品优化、市场和渠道分析、漏斗分析等。提升用户体验还有广告投放、数据交易、行为预测等。

## 三 做好用户画像的前提

### 1. 优质数据源

有数据是能做好用户画像的基础条件，而本身拥有的数据源质量则在一定程度上决定了画像的质量。举个稍浅显的例子，运营商的数据源质量和靠爬虫抓取的数据源质量明显不同。

### 2. 统一设备 ID

当前每个用户往往同时使用电脑、Pad、手机等终端设备，识别多个设备上的同一用户，做好统一 ID 识别是第一步，易观目前已经建立以易观 ID 为核心的一套 ID mapping 结构，用于识别用户跨设备 ID 的打通。

### 3. 技术积累

用户画像本身需要对用户身份预测，比如性别、年龄、兴趣、商业偏好等，这里面涉及分类、聚类，现在还会使用到神经网络等算法，以及算法模型的构建和优化，需要相当技术积累和时间验证。

### 4. 数据源补充

每个企业拥有的数据是很有限的，这就需要跟各数据伙伴合作，进行数据补充，解决数据孤岛问题。

### 5. 标签体系

业务需求决定用户画像的成败，根据业务需求建立的用户标签体系是否合理极其重要。

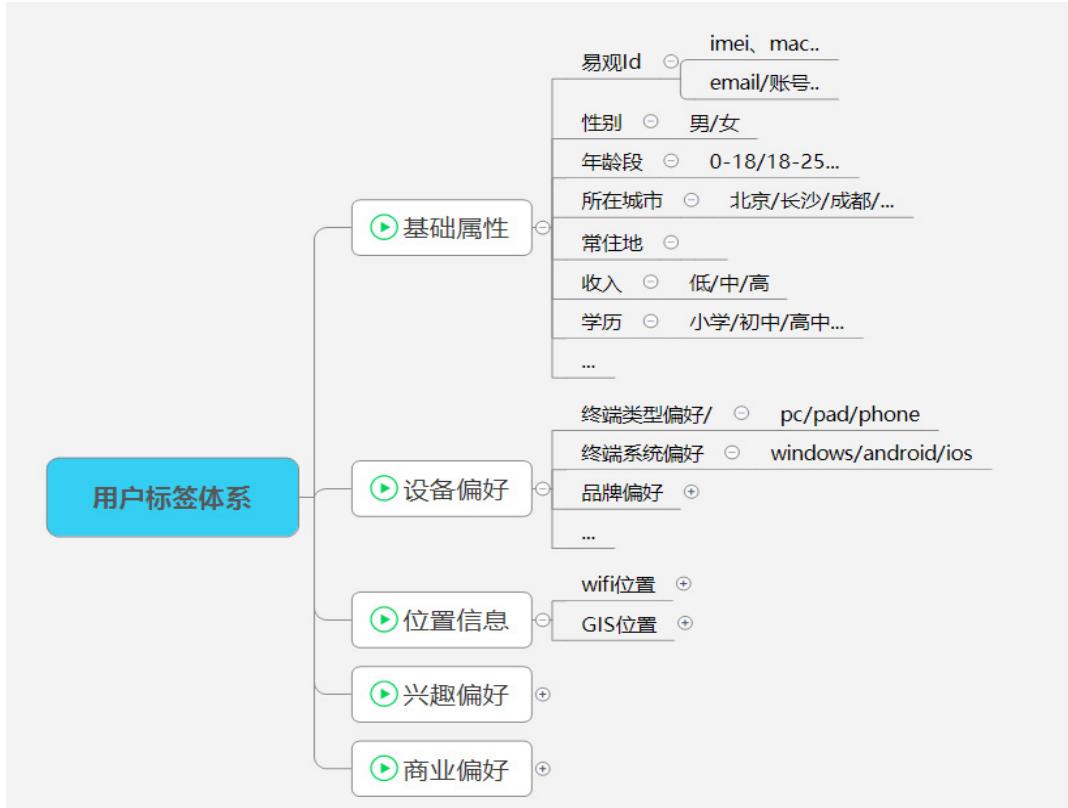
### 6. 计算能力

数据挖掘需要大量的矩阵和迭代计算，周知 GPU 通常是 CPU 计算能力的 20 倍以上，尤其是在深度学习逐渐成为主流的今天，没有计算能力也不太容易做好挖掘。

## 四 用户画像的标签体系构建

考虑到易观现有的数据，易观有以下维度的标签。

人口学属性	年龄、性别、职业、教育、收入...
设备属性	设备品牌/机型/价格/制式...
位置信息	省/国家行政区划/GIS/商圈...
兴趣偏好	金融资讯/健康...
商业偏好	游戏/母婴/汽车/旅游...



利用以上维度的标签，我们可以建立用户标签体系，大体如上图。

拿兴趣偏好来说，目前又会细分成3级，示例如下图。

## 五 用户画像的技术架构

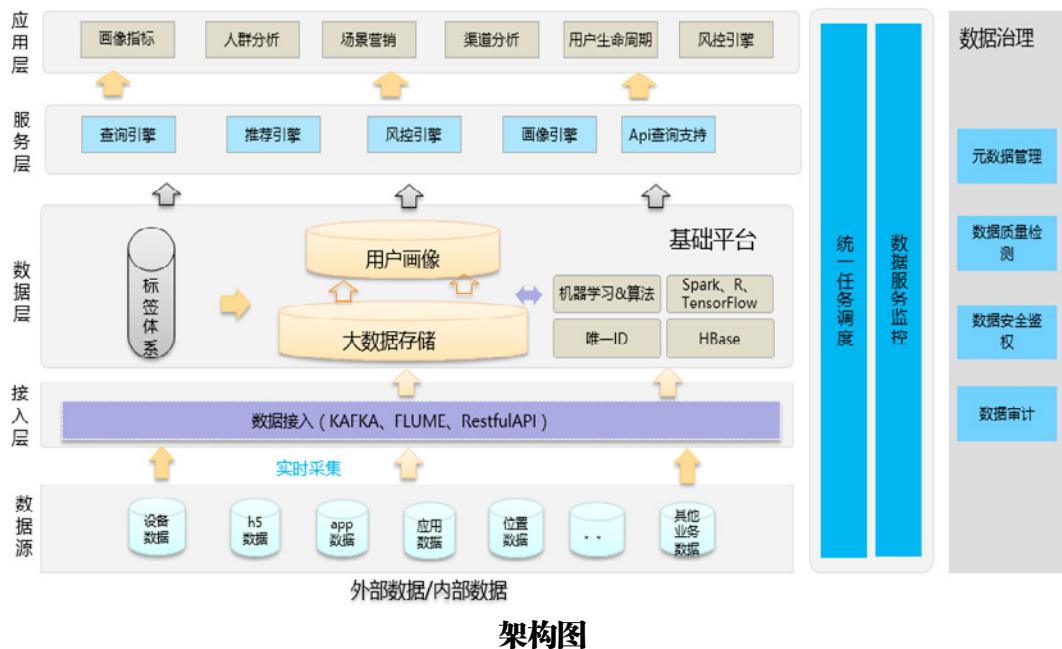
易观用户画像架构图见下页。

## 六 用户画像的实施流程

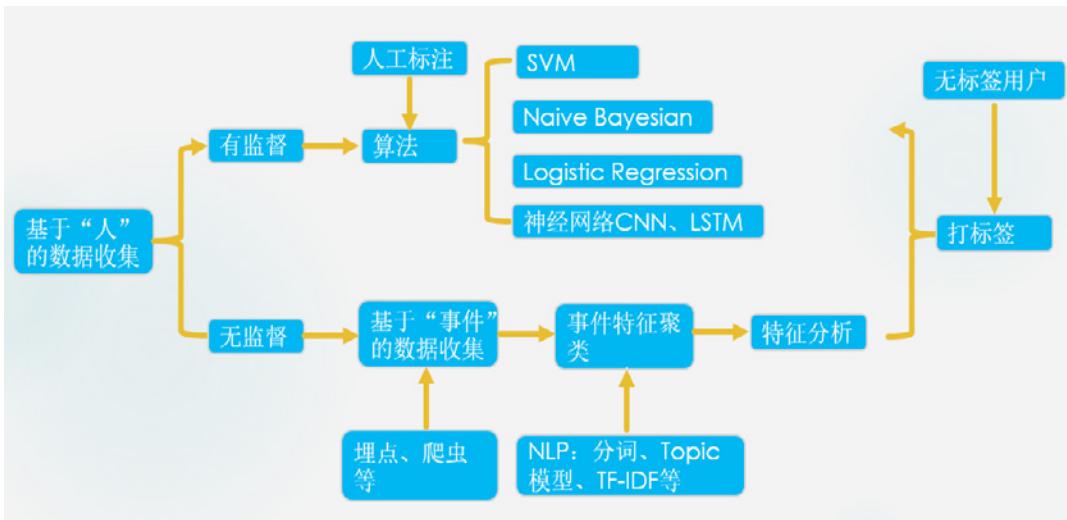
用户画像实施是根据用户的历史行为去反推用户的基本属性及偏好的过程，实施流程如下页所示。

以易观用户基本属性预测为例，需要以用户的APP安装列表、APP埋点事件以及爬虫数据为基础，推测用户的性别、年龄、学历等人口学属性。除数据获取流程外，画像过程通常还包括特征工程和数据建模等。





架构图



实施流程图

## 1. 特征工程

所谓特征工程，就是提取 APP 特征、事件特征、浏览内容特征等。对非结构化数据来说，通常要经历“分词”、“过滤”和“特征提取”三个步骤。

目前易观使用比较流行的 Jieba 分词器，并结合易观内部不断沉淀得

到的自定义分词库、标签词库等，得到了比较好的分词效果。

对于无用词的过滤，除了利用词性进行简单过滤外，易观同样沉淀了大量的通用停用词，并结合业务，建立行业专有停用词库，同样得到了比较好的过滤效果。

对于文本类内容的特征提取，易观采用业内流行的 LDA 算法，并结合业务，对 LDA 算法进行改进。将无监督的 LDA 算法，改进为半监督算法，在分类准确性上提高 20% 左右。除此之外，我们还尝试利用 TF-IDF、Word2Vec 等对分类结果进行校验、优化。

## 2. 数据建模

目前易观尝试使用的算法模型有很多，常见模型比如朴素贝叶斯，逻辑回归，SVM，神经网络等。易观内部通常会根据画像目的、数据量大小等情况，分别选择不同的模型，同时不断总结各种模型的适用情况，尝试将多种模型混合使用，以达到一个更好的效果。

在模型的优化过程中，调参优化是非常重要的一步，在调参优化过程中我们通常会遇到过拟合，样本不均等情况，我们也会单列一下在使用 CNN 方面的一些经验。

## 七 CNN 训练心得—调参经验

1. 样本要随机化，防止大数据淹没小数据。
2. 样本要做归一化。
3. 激活函数要视样本输入选择。（多层神经网络一般使用 relu）
4. Mini batch 很重要，几百是比较合适的。（很大数据量的情况下）
5. 学习速率 (learning rate) 很重要，比如一开始可以 lr 设置为 0.01，然后运行到 loss 不怎么降的时候，学习速率除以 10，接着训练。

6. 权重初始化，可用高斯分布乘上一个很小的数，这个可以看：权值初始化。

7. Adam 收敛速度的确要快一些，可结果往往没有 sgd + momentum 的解好。（如果模型比较复杂的话，sgd 是比较难训练的，这时候 adam 的威力就体现出来了）

8. Dropout 的放置位置以及大小非常重要。

9. Early Stop，发现 val\_loss 没更新，就尽早停止。

深度学习真是一门实验科学，很多地方解释不了为什么好，为什么不好。网络层数、卷积核大小、滑动步长，学习速率这些参数的设置大多是通过已有的架构来做一些调整。

## 八 用户画像总结

本文概括介绍了用户画像的定义、作用以及如何构建用户画像，在这个实践过程中，我们深刻体会到算法不是万能的，除了需要掌握那些挖掘算法的原理外，仍应以业务为中心做展开，一定要对自己的业务数据做分析。模型只是其中的一部分，即便在深度学习发展趋势迅猛的今天，我们也能看到很多传统的数据挖掘算法效果仍然优于深度学习。现在业界的整体模型也差不太多，能拉开差距的基本还是对数据的理解和数据处理上。

## 作者简介

**代立冬**，现任易观大数据架构师，曾担任多家公司数据平台架构师，从事数据领域开发与架构 9 年，对传统行业、互联网行业的数据分析及数据处理有丰富经验，对多个开源社区项目源码熟悉，偶尔研究下神经网络。



# 让机器读懂用户：大数据中的用户画像

杨杰

## 用户画像的含义

用户画像（persona）的概念最早由交互设计之父 Alan Cooper 提出：“Personas are a concrete representation of target users.” 是指真实用户的虚拟代表，是建立在一系列属性数据之上的目标用户模型。随着互联网的发展，现在我们说的用户画像又包含了新的内涵——通常用户画像是根据用户人口学特征、网络浏览内容、网络社交活动和消费行为等信息而抽象出的一个标签化的用户模型。构建用户画像的核心工作，主要是利用存储在服务器上的海量日志和数据库里的大量数据进行分析和挖掘，给用户贴“标签”，而“标签”是能表示用户某一维度特征的标识。具体的标签形式可以参考下图某网站给其中一个用户打的标签。

## 用户画像的作用

提取用户画像，需要处理海量的日志，花费大量时间和人力。尽管是



如此高成本的事情，大部分公司还是希望能给自己的用户做一份足够精准的用户画像。

那么用户画像有什么作用，能帮助我们达到哪些目标呢？

大体上可以总结为以下几个方面：

1. 精准营销：精准直邮、短信、App 消息推送、个性化广告等。
  2. 用户研究：指导产品优化，甚至做到产品功能的私人定制等。
  3. 个性服务：个性化推荐、个性化搜索等。
  4. 业务决策：排名统计、地域分析、行业趋势、竞品分析等。

## 用户画像的内容

用户画像包含的内容并不完全固定，根据行业和产品的不同所关注的特征也有不同。对于大部分互联网公司，用户画像都会包含人口属性和行为特征。人口属性主要指用户的年龄、性别、所在的省份和城市、教育程度、婚姻情况、生育情况、工作所在的行业和职业等。行为特征主要包含活跃度、忠诚度等指标。

除了以上较通用的特征，不同类型的网站提取的用户画像各有侧重点。

以内容为主的媒体或阅读类网站，还有搜索引擎或通用导航类网站，往往会提取用户对浏览内容的兴趣特征，比如体育类、娱乐类、美食类、理财类、旅游类、房产类、汽车类等等。

社交网站的用户画像，也会提取用户的社交网络，从中可以发现关系紧密的用户群和在社群中起到意见领袖作用的明星节点。

电商购物网站的用户画像，一般会提取用户的网购兴趣和消费能力等指标。网购兴趣主要指用户在网购时的类目偏好，比如服饰类、箱包类、居家类、母婴类、洗护类、饮食类等。

消费能力指用户的购买力，如果做得足够细致，可以把用户的实际消费水平和在每个类目的心理消费水平区分开，分别建立特征纬度。

另外还可以加上用户的环境属性，比如当前时间、访问地点 LBS 特征、当地天气、节假日情况等。

当然，对于特定的网站或 App，肯定又有特殊关注的用户纬度，就需要把这些维度做到更加细化，从而能给用户提供更精准的个性化服务和内容。

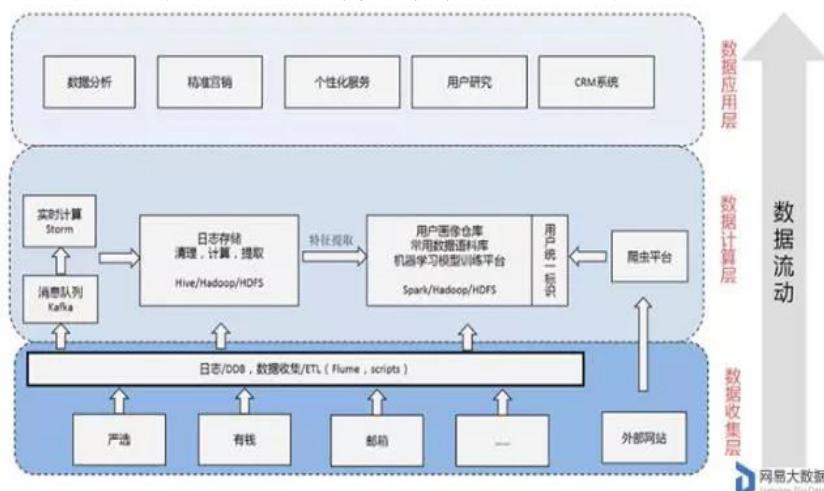
## 用户画像的生产



用户特征的提取即用户画像的生产过程，大致可以分为以下几步：

1. 用户建模，指确定提取的用户特征维度，和需要使用到的数据源。
2. 数据收集，通过数据收集工具，如 Flume 或自己写的脚本程序，把需要使用的数据统一存放到 Hadoop 集群。
3. 数据清理，数据清理的过程通常位于 Hadoop 集群，也有可能与数据收集同时进行，这一步的主要工作，是把收集到各种来源、杂乱无章的数据进行字段提取，得到关注的目标特征。
4. 模型训练，有些特征可能无法直接从数据清理得到，比如用户感兴趣的内容或用户的消费水平，那么可以通过收集到的已知特征进行学习和预测。
5. 属性预测，利用训练得到的模型和用户的已知特征，预测用户的未知特征。
6. 数据合并，把用户通过各种数据源提取的特征进行合并，并给出一定的可信度。
7. 数据分发，对于合并后的结果数据，分发到精准营销、个性化推荐、CRM 等各个平台，提供数据支持。

下面以用户性别为例，具体介绍特征提取的过程：



1. 提取用户自己填写的资料，比如注册时或者活动中填写的性别资料，这些数据准确率一般很高。

2. 提取用户的称谓，如文本中有提到的对方称呼，例如：xxx 先生 / 女士，这个数据也比较准。

3. 根据用户姓名预测用户性别，这是一个二分类问题，可以提取用户的名字部分（百家姓与性别没有相关性），然后用朴素贝叶斯分类器训练一个分类器。过程中遇到了生僻字问题，比如“甄嬛”的“嬛”，由于在名字中出现的少，因此分类器无法进行正确分类。考虑到汉字都是由偏旁部首组成，且偏旁部首也常常具有特殊含义（很多与性别具有相关性，比如草字头倾向女性，金字旁倾向男性），我们利用五笔输入法分解单字，再把名字本身和五笔打法的字母一起放到 LR 分类器进行训练。比如，“嬛”字的打法：『女 V+ 罂 L+ 一 G+ 衣 E = VLGE』，这里的女字旁就很有女性倾向。

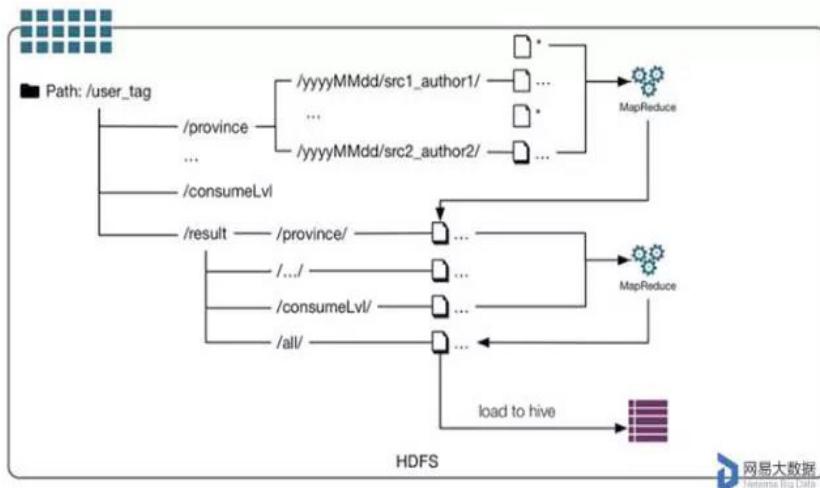
4. 另外还有一些特征可以利用，比如用户访问过的网站，经常访问一些美妆或女性服饰类网站，是女性的可能性就高；访问体育军事类网站，是男性的可能性就高。还有用户上网的时间段，经常深夜上网的用户男性的可能性就高。把这些特征加入到 LR 分类器进行训练，也能提高一定的数据覆盖率。

## 数据管理系统

用户画像涉及到大量的数据处理和特征提取工作，往往需要用到多数据来源，且多人并行处理数据和生成特征。因此，需要一个数据管理系统来对数据统一进行合并存储和分发。我们的系统以约定的目录结构来组织数据，基本目录层级为：/user\_tag/ 属性 / 日期 / 来源 \_ 作者 /。以性

别特征为例，开发者 dev1 从用户姓名提取的性别数据存放路径为 /user\_tag/gender/20170101/name\_dev1，开发者 dev2 从用户填写资料提取的性别数据存放路径为 /user\_tag/gender/20170102/raw\_dev2。

从每种来源提取的数据可信度是不同的，所以各来源提取的数据必须给出一定的权重，约定一般为 0-1 之间的一个概率值，这样系统在做数据的自动合并时，只需要做简单的加权求和，并归一化输出到集群，存储到事先定义好的 Hive 表。接下来就是数据增量更新到 HBase、ES、Spark 集群等更多应用服务集群。

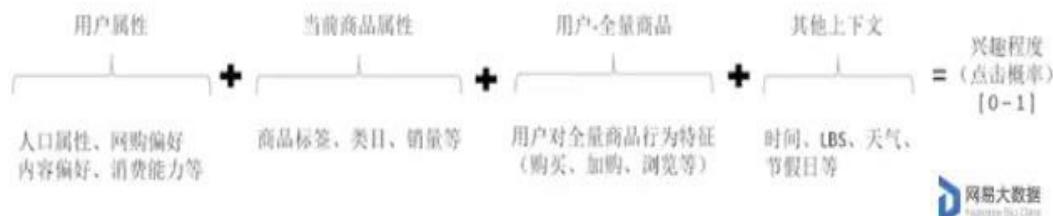


## 应用示例：个性化推荐

以电商网站的某种页面的个性化推荐为例，考虑到特征的可解释性、易扩展和模型的计算性能，很多线上推荐系统采用 LR（逻辑回归）模型训练，这里也以 LR 模型举例。很多推荐场景都会用到基于商品的协同过滤，而基于商品协同过滤的核心是一个商品相关性矩阵  $W$ ，假设有  $n$  个商品，那么  $W$  就是一个  $n * n$  的矩阵，矩阵的元素  $w_{ij}$  代表商品  $I_i$  和  $I_j$  之间的相关系数。而根据用户访问和购买商品的行为特征，可以把用户表示成

一个 n 维的特征向量  $U = [i_1, i_2, \dots, i_n]$ 。于是  $U * W$  可以看成用户对每个商品的兴趣程度  $V = [v_1, v_2, \dots, v_n]$ ，这里  $v_1$  即是用户对商品  $I_1$  的兴趣程度， $v_1 = i_1 * w_{11} + i_2 * w_{12} + \dots + i_n * w_{1n}$ 。如果把相关系数  $w_{11}, w_{12}, \dots, w_{1n}$  看成要求的变量，那么就可以用 LR 模型，代入训练集用户的行为向量  $U$ ，进行求解。这样一个初步的 LR 模型就训练出来了，效果和基于商品的协同过滤类似。

这时只用到了用户的行为特征部分，而人口属性、网购偏好、内容偏好、消费能力和环境特征等其他上下文还没有利用起来。把以上特征加入到 LR 模型，同时再加上目标商品自身的属性，如文本标签、所属类目、销量等数据，如下图所示，进一步优化训练原来的 LR 模型。从而最大程度利用已经提取的用户画像数据，做到更精准的个性化推荐。



## 点评

用户画像是当前大数据领域的一种典型应用，也普遍应用在多款网易互联网产品中。本文基于网易的实践，深入浅出地解析了用户画像的原理和生产流程。

精确有效的用户画像，依赖于从大量的数据中提取正确的特征，这需要一个强大的数据管理系统作为支撑。网易大数据产品体系中包含的一站式大数据开发与管理平台——网易猛犸，正是在网易内部实践中打磨形成的，能够为用户画像及后续的业务目标实现提供数据传输、计算和作业流调度等基础能力，有效降低大数据应用的技术门槛。

# 版权声明

**InfoQ** 中文站出品

## 架构师特刊：用户画像实践

©2017 极客邦控股（北京）有限公司

本书版权为极客邦控股（北京）有限公司所有，未经出版者预先的书面许可，不得以任何方式复制或者抄袭本书的任何部分，本书任何部分不得用于再印刷，存储于可重复使用的系统，或者以任何方式进行电子、机械、复印和录制等形式传播。

本书提到的公司产品或者使用到的商标为产品公司所有。

如果读者要了解具体的商标和注册信息，应该联系相应的公司。

出版：极客邦控股（北京）有限公司

北京市朝阳区洛娃大厦 C 座 1607

欢迎共同参与 InfoQ 中文站的内容建设工作，包括原创投稿和翻译，请联系 editors@cn.infoq.com.

网 址：[www.infoq.com.cn](http://www.infoq.com.cn)