# 用户画像

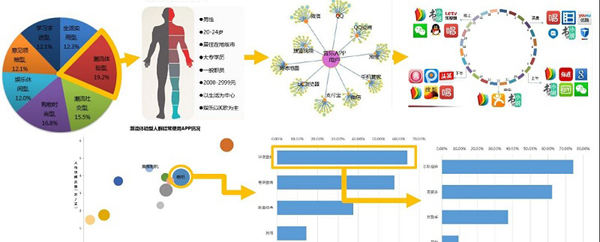
# 用户画像

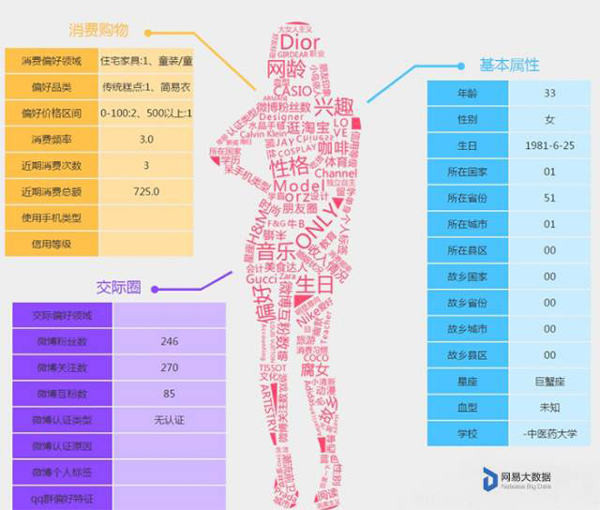
## 什么是用户画像

用户画像的含义用户画像(personas)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。

用户画像可以简单理解成是海量数据的标签，根据用户的目标、行为和观点的差异，将他们区分为不同的类型，然后每种类型中抽取出典型特征，赋予名字、照片、一些人口统计学要素、场景等描述，形成了一个人物原型 (personas)。

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：

* **精准营销** ：精准直邮、短信、App消息推送、个性化广告等。
* **用户研究** ：指导产品优化，甚至做到产品功能的私人定制等。
* **个性服务** ：个性化推荐、个性化搜索等。
* **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

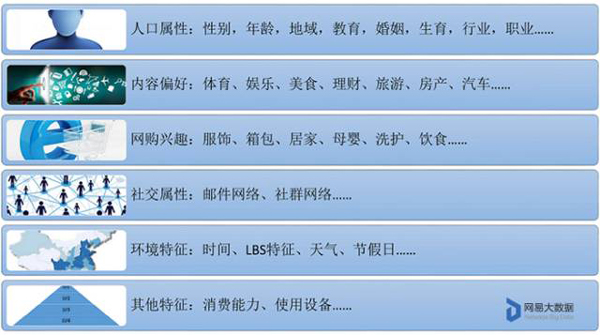
用户画像包含的内容并不完全固定，根据行业和产品的不同所关注的特征也有不同。 对于大部分互联网公司，用户画像都会包含**人口属性**和**行为特征**。 人口属性主要指用户的**年龄**、**性别**、所在的**省份**和**城市**、**教育程度**、**婚姻情况**、**生育情况**、**工作所在的行业**和**职业**等。行为特征主要包含**活跃度**、**忠诚度**等指标。

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。

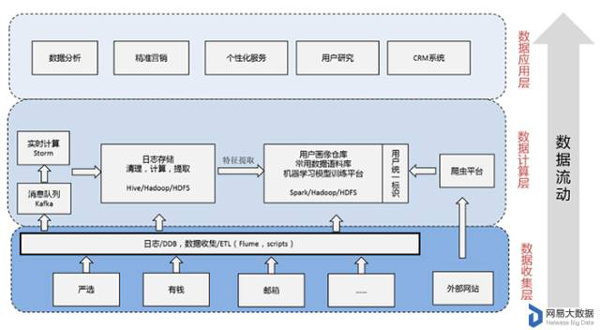
当然，对于特定的网站或App，肯定又有特殊关注的用户纬度，就需要把这些维度做到更加细化，从而能给用户提供更精准的个性化服务和内容。



## 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。

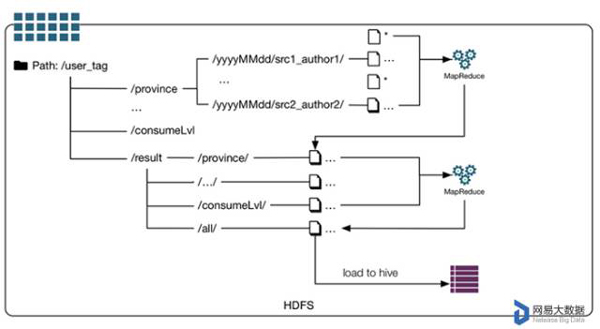


下面以用户性别为例，具体介绍特征提取的过程 ：

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+罒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wij代表商品Ii和Ij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=[ i1, i2, ……, in ]。于是U\*W可以看成用户对每个商品的感兴趣程度V=[ v1, v2, ……, vn ]，这里v1即是用户对商品I1的感兴趣程度，v1= i1\*w11 + i2\*w12 + in\*w1n。如果把相关系数w11, w12, ……, w1n 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。



# 构建用户画像

用户画像可以简单理解成是海量数据的标签，根据用户的目标、行为和观点的差异，将他们区分为不同的类型，然后每种类型中抽取出典型特征，赋予名字、照片、一些人口统计学要素、场景等描述，形成了一个人物原型 (personas)。



## 企业在进行营销时会遇到的问题

如何进行产品的定位、如何去优化用户的体验、如何进行广告的精准投放，以及衡量我们的用户价值的大小等等，这类问题的解决需要结合受众群体的需求特性以及行为组合，去优化自身产品的调整策略。

**然而用户画像才是解决问题的基础，用户画像能帮助广告主找到对的人。**

了解广告营销的人都知道，确定目标受众是营销的第一步，如果说传统线下广告还可以根据不同场合大概确定一个人的社会背景，然而，在互联网这个虚拟世界中，隐藏在幕后的用户拥有太多可能性，因此在互联网营销逐渐占据主流的时代，互联网产品或服务的人群画像愈加重要。

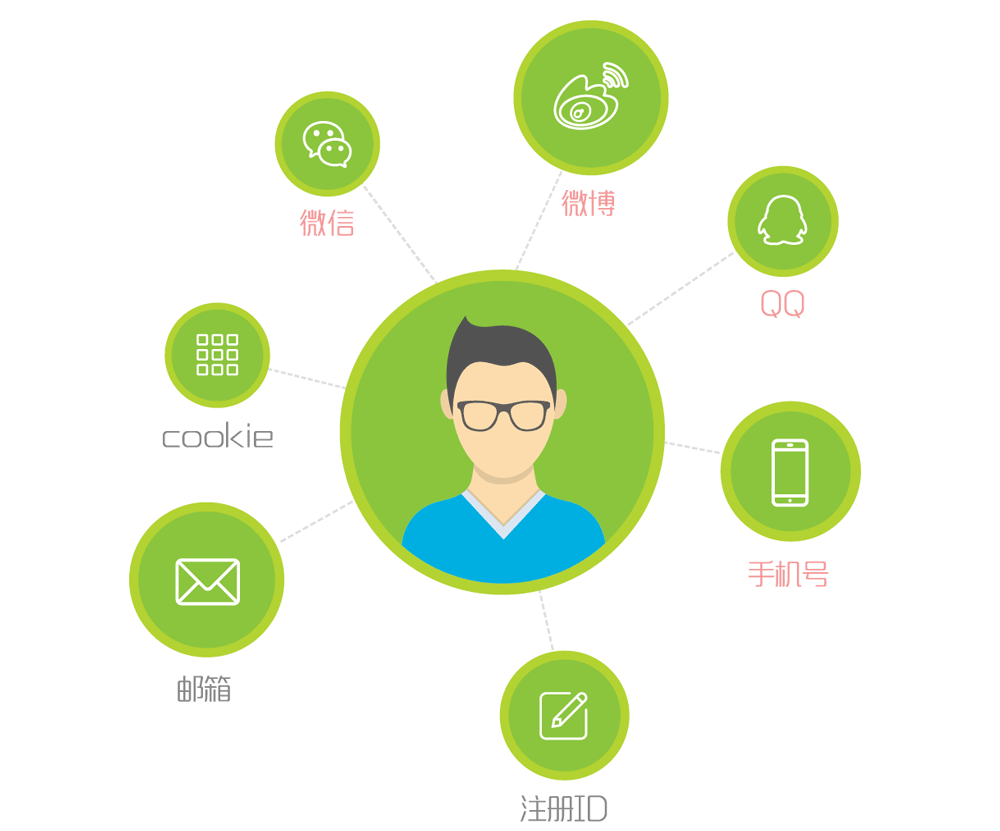


## 构建用户画像

### 准确识别用户

**微博/微信/QQ等第三方登录成企业识别用户的折中选择**

用户识别的目的是为了区分用户、单点定位。用户识别的方式有很多种，如cookie、注册ID、邮箱、微信/微博/QQ等第三方登录、手机号等，这些都是互联网用户相对于传统线下渠道所特有的身份标识，其中手机号是目前移动端最为准确的用户标识，但随着用户的注册意愿越来越低，微博/微信/QQ等第三方登录成为越来越多企业的折中选择--单点登录。

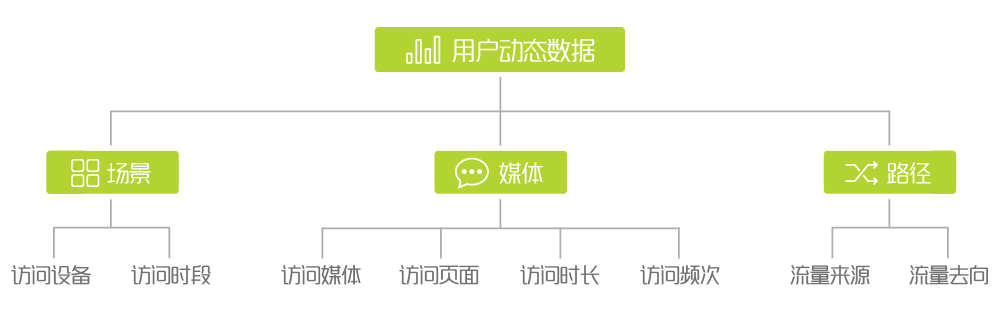


### 动态跟踪用户行为轨迹

**三大维度：场景+媒体+路径**

动态行为数据可以确认用户不同场景下的不同访问轨迹，助力广告主跨端控频营销。

用户网络行为动态跟踪主要包括三个维度：场景+媒体+路径，应用到互联网中，场景主要包括访问设备、访问时段，媒体指某一时段下用户具体访问的媒体，如资讯类、视频类、游戏类、社交类等，路径指用户进入和离开某媒体的路径，可以简单理解为用户的站内与站外行为，如是通过搜索导航进入还是直接打开该APP，离开时是站内跳转到其他网页还是直接关闭，一方面有助于媒体自身优化流量运营，另一方面帮助广告主有效控制不同页面的投放频次，避免产生用户倦怠。



### 结合静态数据评估用户价值

**五大维度：人口属性+商业属性+消费特征+生活形态+CRM**

静态数据获取后，需要对人群进行因子和聚类分析，不同的目的分类依据不同：如对于产品设计来说，按照使用动机或使用行为划分是最为常见的方式，而对于营销类媒体来说，依据消费形态来区分人群是最为直接的分类方式。

静态数据主要包括用户的**人口属性、商业属性、消费特征、生活形态、CRM五大维度**，其获取方式存在多种，数据挖掘是最为常见也是较为精准的一种方式，如果数据有限，则需要定性与定量结合补充，定性方法如小组座谈会、用户深访、日志法、Laddering 阶梯法、透射法等，主要是通过开放性的问题潜入用户真实的心理需求，具象用户特征，定量更多是通过定量问卷调研的方式进行，关键在于后期定量数据的建模与分析，目的是通过封闭性问题一方面对定性假设进行验证，另一方面获取市场的用户分布规律。



### 用户标签定义与权重

**从繁杂的数据中抽取共同的特征值**

根据特征值对群体进行定义，有助于广告主一目了然掌握该群体的特性，如“时尚小咖”，可以快速的联想到针对这类人，时尚感至关重要，即产品的设计感、外观等，并且小咖两字表明该类人并不盲目追求潮流，他们有自己的审美观，并且能够影响身边的人。

同时，一个群体会有多个标签，不同的群体之间也会有标签的重合，此时标签的权重反映了不同群体的核心特征。如“时尚小咖”和“科技先锋”两类人群中都有女性标签，此时需要比较女性在不同人群中的标签权重，以决定将该标签解读给哪类群体。通常，一个好的用户画像，不同人群之间的标签重合度较小，只有在那些权重较小的标签上会有些许重合。

### 不同人群优先级排列

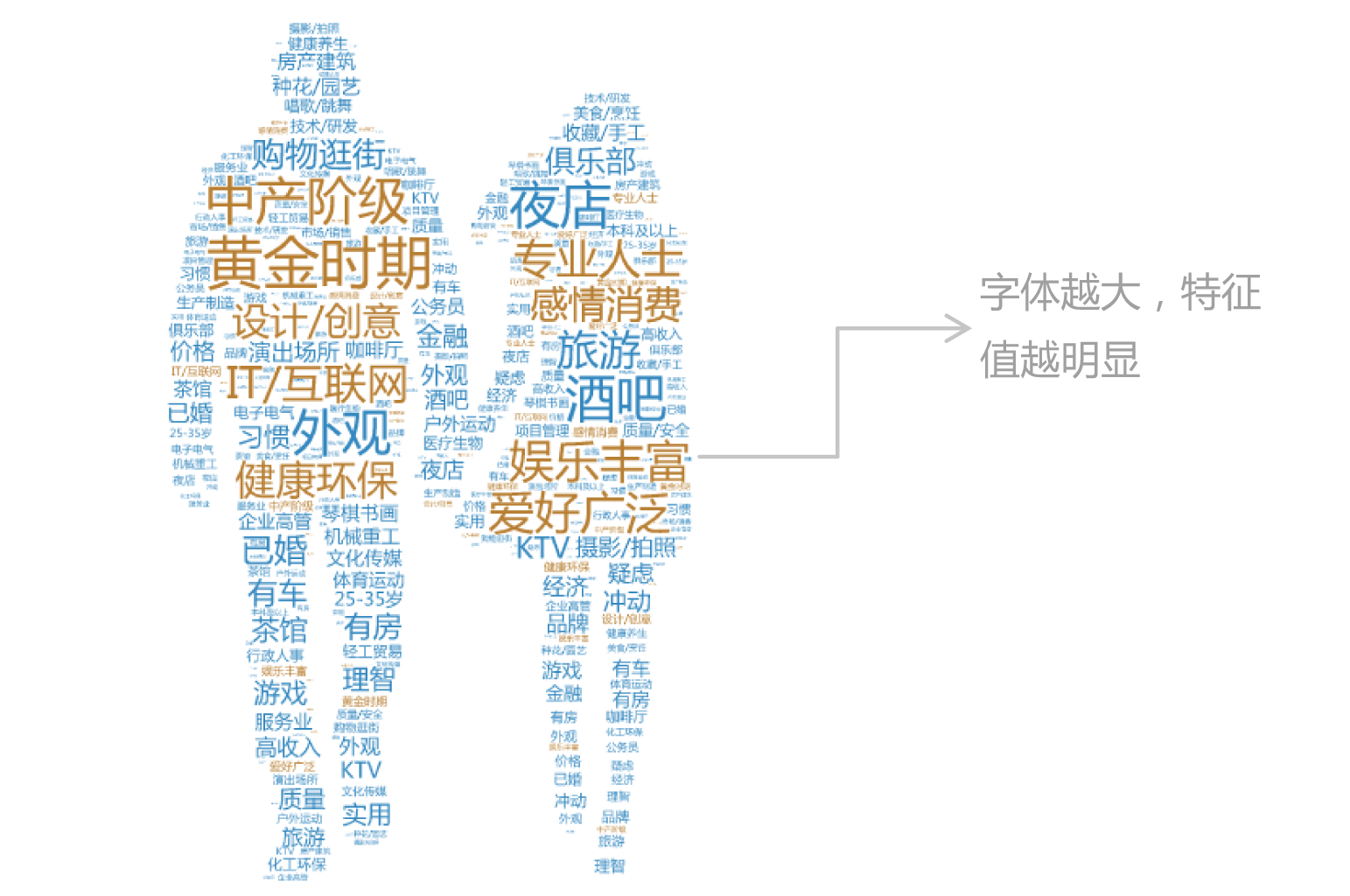
**根据企业自身情况排列不同组合**

目前，大部分画像只完成上述4步就结束了，然而最后一步决定了最终效果的落地，对于广告主来说可以理解为媒介的组合策略。组合策略可以按照频率的高低、市场的大小、收益的潜力、竞争优势等，根据企业自身情况排列不同组合。

比如：品牌刚刚建立，需要快速提升知名度，可以按照不同媒体目标人群覆盖率的高低进行预算分配；当品牌具备一定知名度，企业核心领域营收处于快速增长期时，可以按照不同媒体目标人群贡献的市场大小进行分配；当企业想开拓新市场时，可以按照不同媒体目标人群的收益潜力进行分配，另外如企业品牌需增强差异化的竞争优势时，可按照不同媒体目标人群的竞争优势进行投放。

## 用户画像关键示例

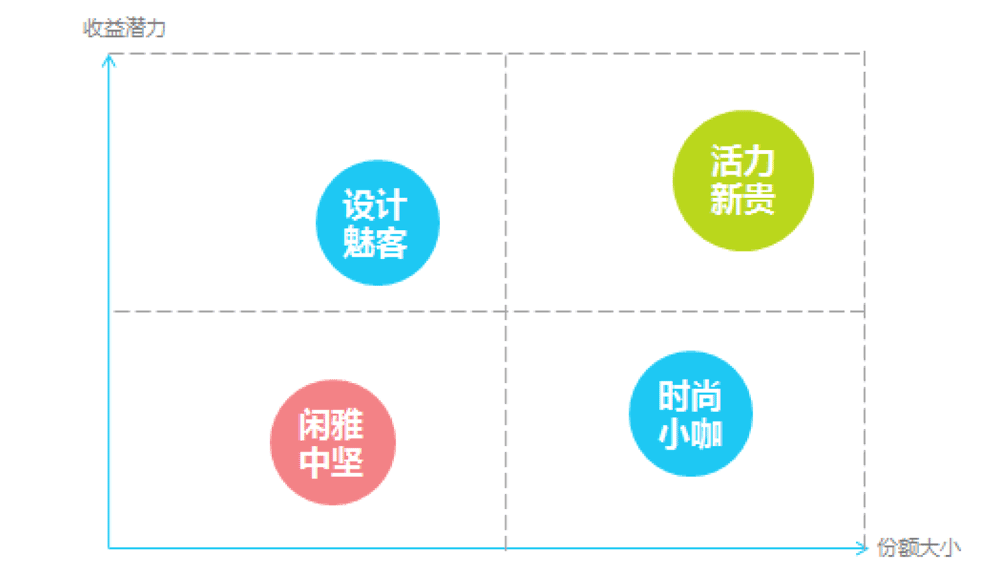
A：某图片分享类媒体整体用户画像



B：某图片分享类媒体细分用户特征



C：某图片分享类媒体细分用户优先级排列

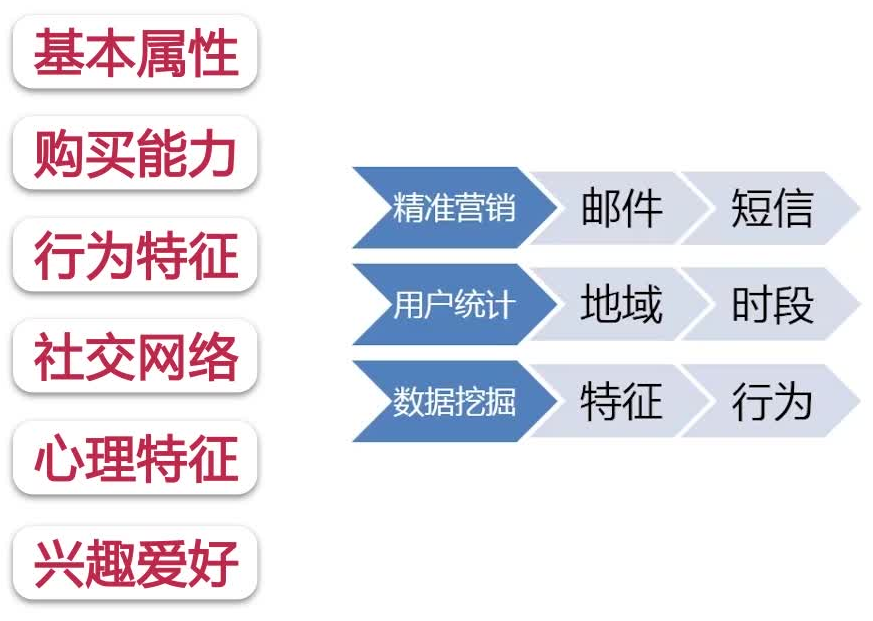


# 电商大数据之用户画像

## 概念

用户画像也叫用户信息标签化、客户标签。从电商角度看，根据客户在电商网站上所填的信息和客户行为，可以用一些标签把客户描绘出来，描述客户的标签就是用户画像。

## 用户画像的内容



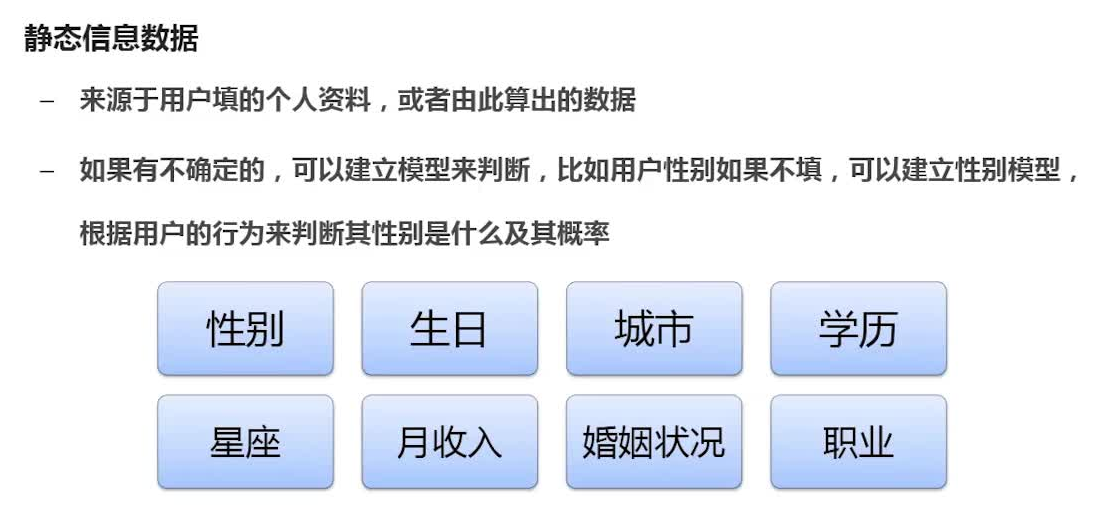
## 用户画像的构建技术



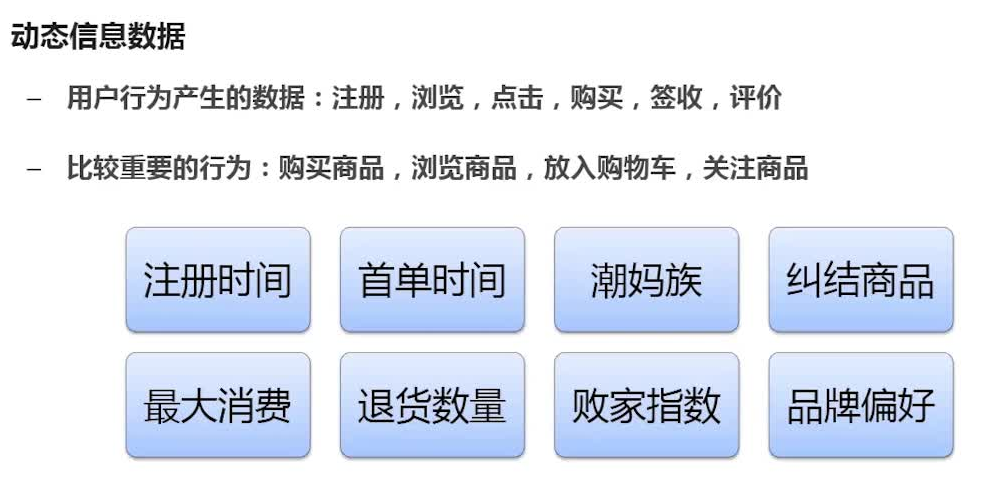
## 用户画像的源数据



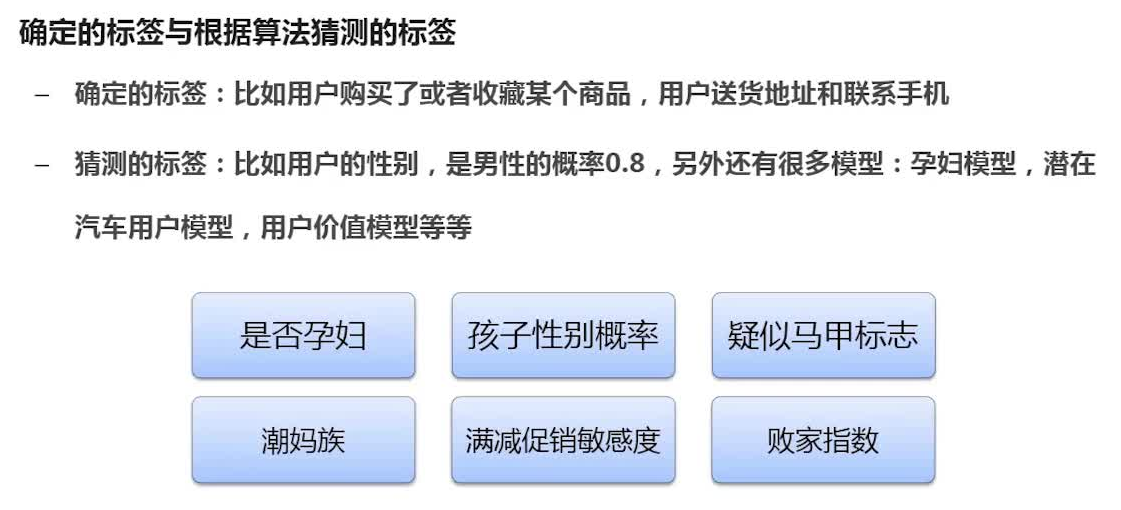
## 静态信息数据



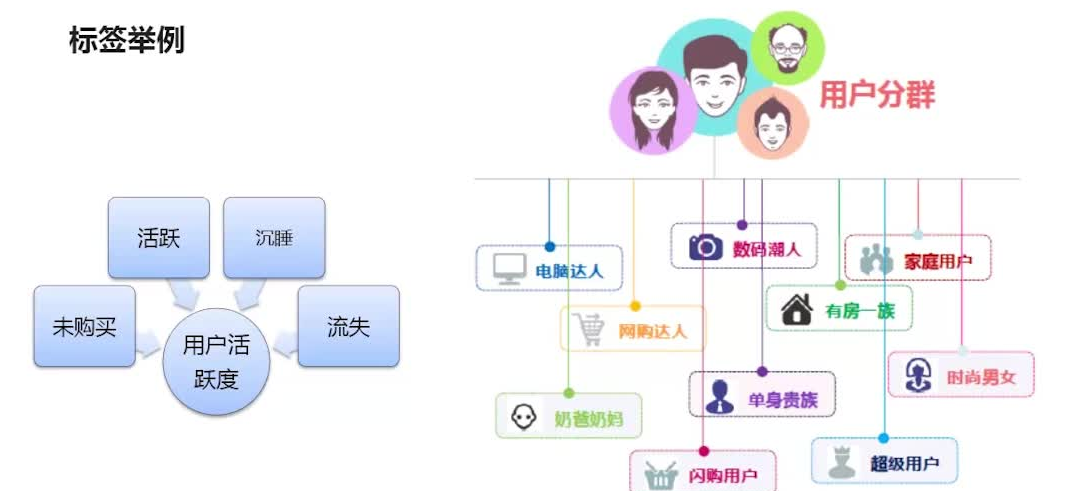
## 动态信息数据



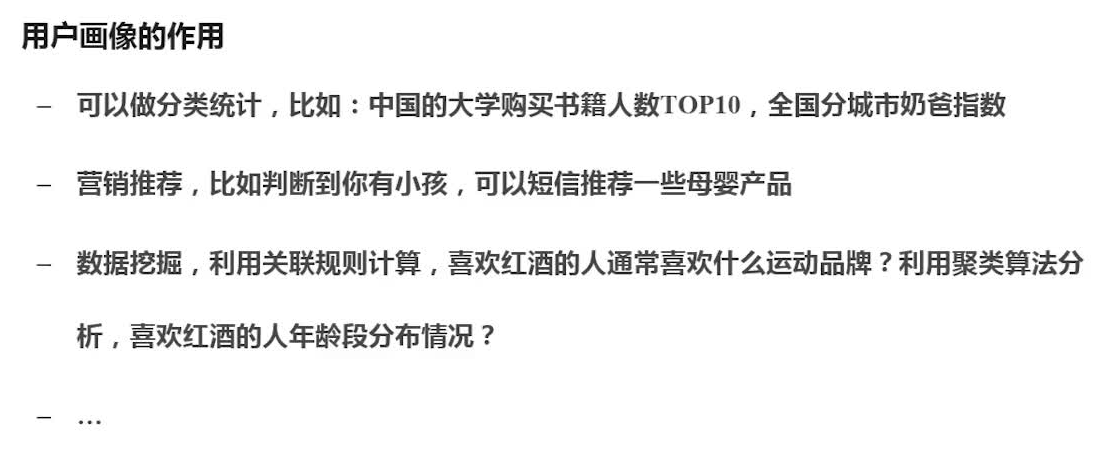
## 用户标签



## 标签示例



## 用户画像的应用



营销推荐

* 邮件营销
* 短信营销
* PUSH信息
* 数据挖掘

## 电商平台中用户数据

**客户消费订单信息**





