【数据分析技术系列】

之用户画像数据建模方法

目录

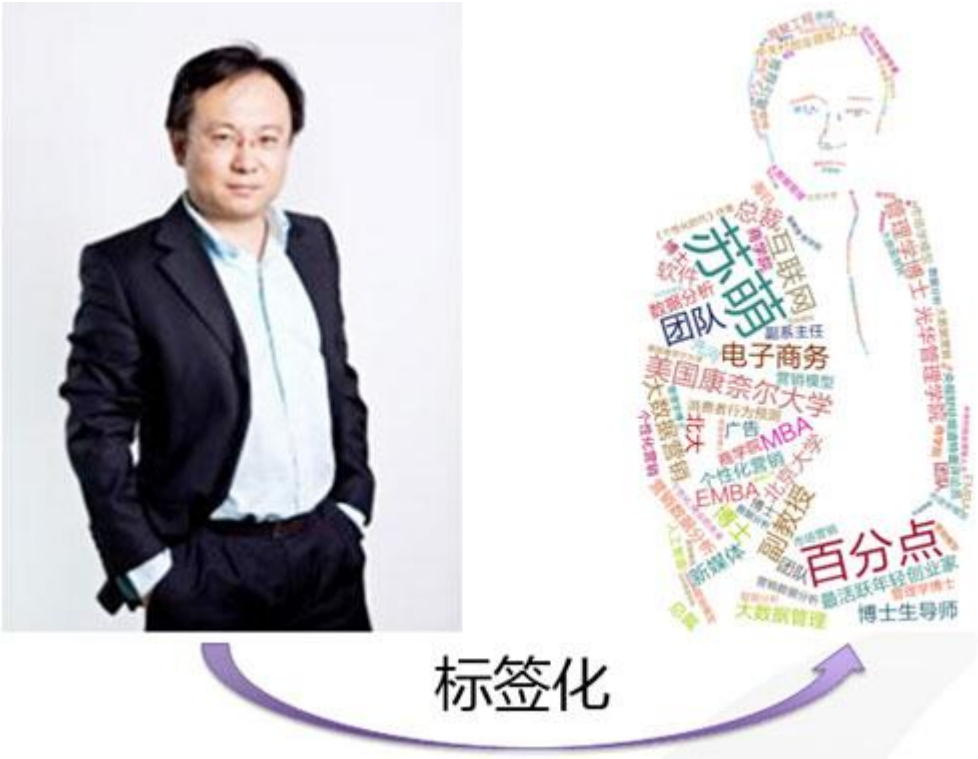
从 1991 年 Tim Berners-Lee 发明了万维网（World Wide Web）开始到 2011 年，互联网真正走向了一个新的里程碑，进入了“大数据时代”。经历了 12、13 两年热炒之后，人们逐渐冷静下来，更加聚焦于如何利用大数据挖掘潜在的商业价值，如何在企业中实实在在的应用大数据技术。伴随着大数据应用的讨论、创新，个性化技术成为了一个重要落地点。相比传统的线下会员管理、问卷调查、购物篮分析，大数据第一次使得企业能够通过互联网便利地获取用户更为广泛的反馈信息，为进一步精准、快速地分析用户行为习惯、消费习惯等重要商业信息，提供了足够的数据基础。伴随着对人的了解逐步深入，一个概念悄然而生：用户画像（UserProfile），完美地抽象出一个用户的信息全貌，可以看作企业应用大数据的根基。

一、什么是用户画像？

男，31 岁，已婚，收入 1 万以上，爱美食，团购达人，喜欢红酒配香烟。

这样一串描述即为用户画像的典型案例。如果用一句话来描述，即：用户信息标签化。

如果用一幅图来展现，即：



二、为什么需要用户画像

用户画像的核心工作是为用户打标签，打标签的重要目的之一是为了让人能够理解并且方便计算机处理，如，可以做分类统计：喜欢红酒的用户有多少？喜

欢红酒的人群中，男、女比例是多少？也可以做数据挖掘工作：利用关联规则计算，喜欢红酒的人通常喜欢什么运动品牌？利用聚类算法分析，喜欢红酒的人年龄段分布情况？

大数据处理，离不开计算机的运算，标签提供了一种便捷的方式，使得计算机能够程序化处理与人相关的信息，甚至通过算法、模型能够“理解” 人。当计算机具备这样的能力后，无论是搜索引擎、推荐引擎、广告投放等各种应用领域，都将能进一步提升精准度，提高信息获取的效率。

三、如何构建用户画像

一个标签通常是人为规定的高度精炼的特征标识，如年龄段标签：25~35 岁，地域标签：北京，标签呈现出两个重要特征：语义化，人能很方便地理解每个标签含义。这也使得用户画像模型具备实际意义。能够较好的满足业务需求。如，判断用户偏好。短文本，每个标签通常只表示一种含义，标签本身无需再做过多文本分析等预处理工作，这为利用机器提取标准化信息提供了便利。

人制定标签规则，并能够通过标签快速读出其中的信息，机器方便做标签提取、聚合分析。所以，用户画像，即：用户标签，向我们展示了一种朴素、简洁的方法用于描述用户信息。

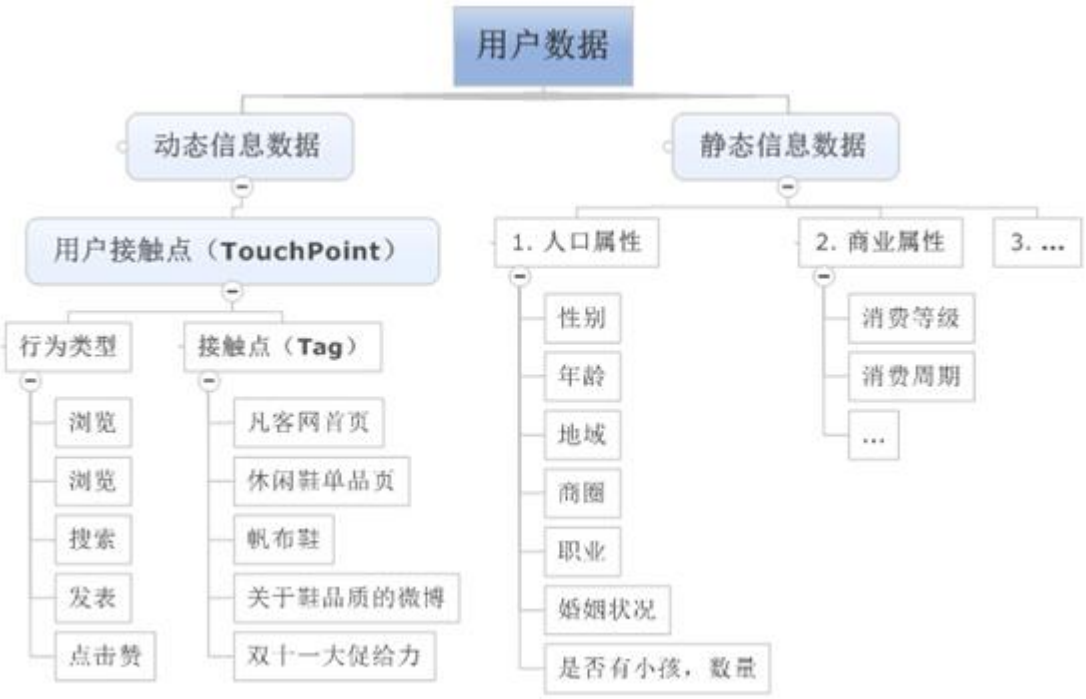
3.1 数据源分析

构建用户画像是为了还原用户信息，因此数据来源于：所有用户相关的数据。

对于用户相关数据的分类，引入一种重要的分类思想：封闭性的分类方式。如，世界上分为两种人，一种是学英语的人，一种是不学英语的人；客户分三类，高价值客户，中价值客户，低价值客户；产品生命周期分为，投入期、成长期、成熟期、衰退期…所有的子分类将构成了类目空间的全部集合。

这样的分类方式，有助于后续不断枚举并迭代补充遗漏的信息维度。不必担心架构上对每一层分类没有考虑完整，造成维度遗漏留下扩展性隐患。另外，不同的分类方式根据应用场景，业务需求的不同，也许各有道理，按需划分即可。

本文将用户数据划分为静态信息数据、动态信息数据两大类。



静态信息数据

用户相对稳定的信息，如图所示，主要包括人口属性、商业属性等方面数据。这类信息，自成标签，如果企业有真实信息则无需过多建模预测，更多的是数据清洗工作，因此这方面信息的数据建模不是本篇文章重点。

动态信息数据

用户不断变化的行为信息，如果存在上帝，每一个人的行为都在时刻被上帝那双无形的眼睛监控着，广义上讲，一个用户打开网页，买了一个杯子；与该用户傍晚溜了趟狗，白天取了一次钱，打了一个哈欠等等一样都是上帝眼中的用户行为。当行为集中到互联网，乃至电商，用户行为就会聚焦很多，如上图所示：浏览凡客首页、浏览休闲鞋单品页、搜索帆布鞋、发表关于鞋品质的微博、赞“双十一大促给力”的微博消息。等等均可看作互联网用户行为。

本篇文章以互联网电商用户，为主要分析对象，暂不考虑线下用户行为数据（分析方法雷同，只是数据获取途径，用户识别方式有些差异）。

在互联网上，用户行为，可以看作用户动态信息的唯一数据来源。如何对用户行为数据构建数据模型，分析出用户标签，将是本文着重介绍的内容。

3.2 目标分析

用户画像的目标是通过分析用户行为，最终为每个用户打上标签，以及该标签的权重。如，红酒 0.8、李宁 0.6。

标签，表征了内容，用户对该内容有兴趣、偏好、需求等等。

权重，表征了指数，用户的兴趣、偏好指数，也可能表征用户的需求度，可

以简单的理解为可信度，概率。

3.3 数据建模方法

下面内容将详细介绍，如何根据用户行为，构建模型产出标签、权重。一个事件模型包括：时间、地点、人物三个要素。每一次用户行为本质上是一次随机事件，可以详细描述为：什么用户，在什么时间，什么地点，做了什么事。

什么用户：关键在于对用户的标识，用户标识的目的是为了区分用户、单点定位。

用户标识方式	效果	备注（局限性）
Cookie	互联网使用最为广泛的方式，能够标识匿名、未注册用户。	通常有一定的有效期，不易跨浏览器、设备。
注册ID	各家网站的用户标识，最常见的互联网会员管理方式。	用户注册意愿越来越低，需要投入大量推广运营成本。
Email	互联网早期较为常用的用户标识方式。目前依然有一定的占有率。	一人有多个email很常见。因此标识会损失些准确性。
微博、微信、QQ	当下业内共识的第三方登录ID，提供OAuth授权机制	标识准确性，持久性上是个较好的折中方案。
手机号	移动端最精准的标识	较难获取到，视产品激励用户填写意愿。
身份证	最官方的标识	难获取到，视产品激励用户填写意愿。

以上列举了互联网主要的用户标识方法，获取方式由易到难。视企业的用户粘性，可以获取的标识信息有所差异。

什么时间：时间包括两个重要信息，时间戳+时间长度。时间戳，为了标识用户行为的时间点，如，1395121950（精度到秒），1395121950.083612（精度到微秒），通常采用精度到秒的时间戳即可。因为微秒的时间戳精度并不可靠。浏览器时间精度，准确度最多也只能到毫秒。时间长度，为了标识用户在某一页面的停留时间。

什么地点：用户接触点，Touch Point。对于每个用户接触点。潜在包含了两层信息：网址 + 内容。网址：每一个 url 链接（页面/屏幕），即定位了一个互联网页面地址，或者某个产品的特定页面。可以是 PC 上某电商网站的页面 url，也可以是手机上的微博，微信等应用某个功能页面，某款产品应用的特定画面。如，长城红酒单品页，微信订阅号页面，某游戏的过关页。

内容：每个 url 网址（页面/屏幕）中的内容。可以是单品的相关信息：类别、品牌、描述、属性、网站信息等等。如，红酒，长城，干红，对于每个互联网接触点，其中网址决定了权重；内容决定了标签。

注：接触点可以是网址，也可以是某个产品的特定功能界面。如，同样一瓶

矿泉水，超市卖 1 元，火车上卖 3 元，景区卖 5 元。商品的售卖价值，不在于成本，更在于售卖地点。标签均是矿泉水，但接触点的不同体现出了权重差异。这里的权重可以理解为用户对于矿泉水的需求程度不同。即，愿意支付的价值不同。

标签 权重

矿泉水 1 // 超市

矿泉水 3 // 火车

矿泉水 5 // 景区

类似的，用户在京东商城浏览红酒信息，与在品尚红酒网浏览红酒信息，表现出对红酒喜好度也是有差异的。这里的关注点是不同的网址，存在权重差异，权重模型的构建，需要根据各自的业务需求构建。

所以，网址本身表征了用户的标签偏好权重。网址对应的内容体现了标签信息。

什么事：用户行为类型，对于电商有如下典型行为：浏览、添加购物车、搜索、评论、购买、点击赞、收藏 等等。

不同的行为类型，对于接触点的内容产生的标签信息，具有不同的权重。如，购买权重计为 5，浏览计为 1

红酒 1 // 浏览红酒

红酒 5 // 购买红酒

综合上述分析，**用户画像的数据模型，可以概括为下面的公式：用户标识 + 时间 + 行为类型 + 接触点（网址+内容），某用户因为在什么时间、地点、做了什么事。所以会打上**标签。**

用户标签的权重可能随时间的增加而衰减，因此定义时间为衰减因子 r，行为类型、网址决定了权重，内容决定了标签，进一步转换为公式：

标签权重=衰减因子×行为权重×网址子权重

如：用户 A，昨天在品尚红酒网浏览一瓶价值 238 元的长城干红葡萄酒信息。

标签：红酒，长城

时间：因为是昨天的行为，假设衰减因子为：r=0.95

行为类型：浏览行为记为权重 1

地点：品尚红酒单品页的网址子权重记为 0.9（相比京东红酒单品页的 0.7）

假设用户对红酒出于真的喜欢，才会去专业的红酒网选购，而不再综合商城选购。

则用户偏好标签是：红酒，权重是 0.95*0.7 * 1=0.665，即，用户 A：红酒 0.665、长城 0.665。

上述模型权重值的选取只是举例参考，具体的权重值需要根据业务需求二次建模，这里强调的是如何从整体思考，去构建用户画像模型，进而能够逐步细化模型。

四、总结：

本文并未涉及具体算法，更多的是阐述了一种分析思想，在计划构建用户画像时，能够给您提供一个系统性、框架性的思维指导。

核心在于对用户接触点的理解，接触点内容直接决定了标签信息。内容地址、行为类型、时间衰减，决定了权重模型是关键，权重值本身的二次建模则是水到渠成的进阶。模型举例偏重电商，但其实，可以根据产品的不同，重新定义接触点。

比如影视产品，我看了一部电影《英雄本色》，可能产生的标签是：周润发 0.6、枪战 0.5、港台 0.3。

最后，接触点本身并不一定有内容，也可以泛化理解为某种阈值，某个行为超过多少次，达到多长时间等。

比如游戏产品，典型接触点可能会是，关键任务，关键指数（分数）等等。如，积分超过 1 万分，则标记为钻石级用户。钻石用户 1.0。

百分点现已全面应用用户画像技术于推荐引擎中，在对某电商客户，针对活动页新访客的应用中，依靠用户画像产生的个性化效果，对比热销榜，推荐效果有显著提升：推荐栏点击率提升 27%， 订单转化率提升 34%。

（via：郭志金）