# 数字家庭业务用户画像需求分析

## 用户画像

用户画像的目标是描述人，理解人，把收集到的用户行为、社会属性、生活习惯等数据进行分析，抽象出一个用户的商业全貌。简而言之，则用户信息标签化，用标签的形式来表示用户。

举例子来说，通过一个用户经常会购买奶粉，那么通过分析，企业很有可能给这个用户贴上“有孩子”这个标签。

为用户打标签有什么用呢？其作用大致如下：

* 精准营销。分析出产品的潜在用户。例如在直播高尔夫球比赛会插播汽车广告，是因为观看直播的用户被贴上了“用户汽车模型”之类的标签。
* 个性化推荐。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。
* 产品开发定位。通过目标用户的标签化，可以理性地分析产品定位。用于业务经营的分析以及竞争分析，影响极其发展战略。
* 用户统计。
* 数据挖掘。
* 进行效果评估，完善产品运营。

用户画像为进一步精准、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，完美地抽象出了用户的信息全貌，可以看作企业应用大数据的根基。

## 需求分析

### 2.1. 项目目标

项目目标为建立一个处理数字家庭业务中的用户行为数据，生成用户画像的后台系统。生成的用户画像可用于企业自身的后续分析，并且后台系统提供数据可视化统计以方便使用。

### 数据来源

系统的数据来源于数字家庭业务中的客户端。数字家庭业务的客户端分为3种：手机端App、电视应用以及网站。这3种客户端的用户行为差异较大，不能够作为同一数据来处理，需要分别对这3种客户端进行数据收集和分析，然后再整合结果。

手机端应用中用户响应比较灵活，可以在里面收集到一些可以反应出用户习惯的数据，如开启App频率、滑动页面速度等。在手机端App应用中收集的用户数据包括：

1. 用户点击每个按钮的频率、次数
2. 用户在每个页面的访问频率
3. 用户在每个页面中停留的时间
4. 用户的滑动手势的速度以及频率
5. 用户使用应用时间段
6. 用户使用App的频率

电视端和手机的区别在于电视的响应较慢，一些用户行为特征不明显。在电视应用中收集的用户数据包括：

1. 用户点击界面中的按钮的频率、次数
2. 遥控器上各按键的点击频率。该数据可用于改进用户体验
3. 用户使用应用时间段、频率
4. 用户在每个页面的访问频率
5. 用户在每个页面中停留的时间

网页端的用户交互比较多，如果有一些动态的界面，还需要后台的数据关联。采集的客户端用户数据包括：

1. 用户访问网页的时间段
2. 用户访问网页的频率
3. 用户在各个页面的停留时间
4. 用户点击页面中各组件的频率、次数
5. 用户鼠标的移动速度

采集到用户数据之后，再和业务中特定的用户属性相关联。比如说按钮ID1对应喜剧电影等。这些用户属性根据业务的不同，自行设置映射表。关联后的数据用于机器学习分析，得到用户画像。

因为该项目进行的App应用数据采集方式有限，所以不能够对用户的文字、语音、图像这方面的输入进行采集。所以这方面的用户画像是不能够处理的。

### 用户模型

收集到的数据需要进行处理，这时候要用到机器学习算法，可以是决策树、随机森林、SVM、神经网络等，具体使用什么算法需要实验验证对比之后再做决定。除了机器学习算法，给用户贴标签还需要特定模型，来满足业务需求：

* 用户家庭成员模型

可以在家庭成员这个方面给用户贴上如下标签：单身男/女、二人世界、有小孩、家有老人。家庭成员的预测对用户画像的应用有很大帮助。

* 用户消费水平模型
* 用户喜好模型。用于表示用户家庭对各个类型的节目的爱好程度。
* 用户家庭成员职业预测模型。

## 功能设计

### 3.1. 用例图



**图1 系统用例图**

### 时序图



**图2 系统时序图**

### 后台项目流程

用户画像项目的流程拟分为以下几个步骤：

1. 用户数据收集。用户数据从客户端中的用户行为获取。客户端分为手机端App、电视端应用以及网站。后台分别为这三个平台提供数据接收的相应接口。
2. 用户数据持久化。数据库使用MySQL进行备份、保存。后台把接收到的数据通过特定策略（如定时保存）发送给数据库服务器，由后者保存到数据库，进行后续功能的使用。
3. 用户画像建模。用户画像建模采取人工辅助+机器学习的方法。人工辅助的方面在于：

* 用户行为数据与具体业务的关联。比如说把某个按键ID与某个行为（如购买商品）关联起来。
* 用户模型分类。

1. 用户画像结果的验证。
2. 用户画像数据的输出与可视化。

### 用户画像建模

用户画像建模过程分为以下几层：



**图3 用户画像建模过程**

第一层是客户端收集到的用户原始的、未经过处理数据，如用户点击的按钮Id、停留过一定时间的页面Id等。

第二层是通过管理员预先设置的映射表，将原始数据转化成业务内容。

第三层是通过机器学习方法，以不同维度的模型标准给用户“贴标签”，建立用户画像模型。

单单通过客户端的数据建立用户画像有一定的局限性，即使管理员能够建立起客户端组件与业务的映射。例如说一些动态的界面：电视节目等，管理员是不能够一一进行映射的。这时候需要后台方面的数据，直接告诉我们这个系统，用户观看了什么节目。处理从后台得到的数据比客户端得到的数据要简单，后台得到的数据就是图3中第二层的内容，可以直接进行数据挖掘。