# 数字家庭业务用户画像需求文档

## 需求分析

App客户端中可以采集到用户的大量数据。本项目通过收集用户在指定App应用的行为，进行数据分析得到用户对于该应用的用户画像。得到的用户画像可用于行为预测，分类，精准广告，个性化推荐等后续操作。

在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

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，自行设置调整。最终关联后得到用对于该业务的用户标签，则用户画像。对得到的用户画像的后续分析由使用者进行，本项目的目的是得到与业务相关的用户画像。

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。故而这方面的用户画像是不能够处理的。

## 过程大纲

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。用户数据主要从客户端App中获取，包括用户点击按键数据、用户手势收据（滑动速度等）、用户访问页面数据等。后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，进行后续历史记录等功能的使用。
3. 用户行为数据与具体业务数据的关联。在应用中收集到的数据仅仅为按钮id、页面id等不能够进行分析的数据，需要将特定按钮、页面与特定的业务属性对应起来。
4. 用户标签和建模。根据数字家庭业务的实际情况，我们把用户基本属性如下分类：
5. 静态属性，则从用户个人注册信息的数据。包括：用户ID、城市、性别、生日、注册时间
6. 预测属性：用户活跃时间段、爱好因子（如喜剧0.8等）、潜在汽车用户概率、职业预测、家庭成员、消费能力。
7. 用户画像的验证。用户建模拟用如今主流的一些数据挖掘算法：决策树、随机森林、神经网络、SVM等进行处理。通过对比不同算法以及同一算法的不同参数的结果，权衡运行时间、以及精确率和召回率，找到合适的算法。