## 关于FACEBOOK用户数据爬取的想法

### 用户画像的职能：

用户画像需要产生实际的作用，为其他部门提供辅助、支持，从游戏角度来说，用户画像的主要职能有：

1，助力运营团队，实现精准化营销。运营团队的职责是让游戏尽可能长的健康运营，为公司带来稳定、大量的利润。通常，运营通过各类活动来实现这些目标，在上级要求、游戏本身具有的付费点、玩家需求三者平衡之间来开展。因此，用户画像主要提供玩家的需求描述，提供给运营团队的玩家标签具有动态特征，这些动态标签描绘的是玩家在游戏当中的需求，具有很强的时效性。

主要的动态标签以如下形式展现如附件：用户画像标签计算.xlsx

2，广告投放，降低获取新玩家的成本。游戏的拉新需要渠道将广告曝光给用户，经过注意、理解、下载、安装过程，达到获取新玩家的目的。对于海外的广告投放，FACEBOOK在2015年占全球社交网络广告市场的75%份额，占全球广告销售额的51%。FACEBOOK提供了为了广告主提供了方便的管理工具。通过广告目标设定、创建广告账户、广告的受众选定、预算和排期、广告的投放形式来完成在FACEBOOK平台的广告投放。广告投放中除了投放目标、收费形式以外，最重要的是完成受众的设定，如图1.1~1.2所示。用户画像为广告投放部门提供的接口主要为静态标签，帮助运营人员进行有效的受众定向设定。



图1.1：FACEBOOK受众选定-1



图1.2：FACEBOOK受众选定-2

### FACEBOOK平台的广告投放策略

FACEBOOK的广告投放方式，如表2.1所示。FACEBOOK综合广告主竞价、广告内容和用户相关度来决定是否投放，等实际投放成功后，收取竞价第二高广告主的费用，来规避恶意竞价。广告的优先度=广告竞价+相关性分数，广告竞价和预算都会影响到广告的优先度，这部分主要由运营广告投放组负责，去除这两项影响因子，提高广告内容和用户相关度是非常重要的，对于CPM的投放方式来说，粗制投放会很大程度浪费广告费用；对于CPC和CPA的投放方式来说，会产生费用花不出去的现象，尤其对于精确的受众定向，将会面临很多公司的竞争。

表2.1 FACEBOOK的广告收费模式

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 广告形式 | 简称 | 适用场景 | 广告主 | 广告提供商 |
| 按展示量收费 | CPM(cost per impression) | 品牌广告 | 风险较大，效果难以评估 | 要求较低 |
| 按安装量收费 | CPA(cost per action) | 效果广告 | 风险较低，效果可以量化评估 | 要求较高 |
| 按点击量收费 | CPC(cost per click) |
| 优化的按照展示量收费 | oCPM(optimized cost per impression) |  | FACEBOOK根据投放结果，帮助广告商来优化投放策略 | 要求较高 |

### FACEBOOK平台的用户信息

谷歌、百度是信息检索的公司，掌握着每个人历史搜索记录，但是不能够精确描绘用户，而FACEBOOK是依托用户精准信息的社交平台，利用玩家详细的个人信息进行广告推广。从FACEBOOK平台上可以获取丰富的玩家信息，包括对方的好友列表、时间线上的发帖内容、体育运动、电影、电视、书籍、音乐、应用和游戏、点赞过的内容等。如图3.1~3.3是FACEBOOK用户JOSHUA JAMES WILLIAMS的偏好信息展示。

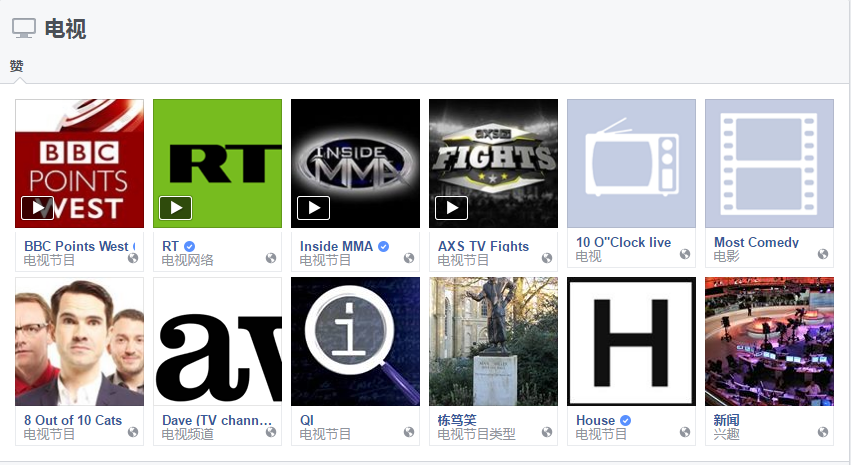


图3.1 用户JOSHUA JAMES WILLIAMS喜爱的电视节目

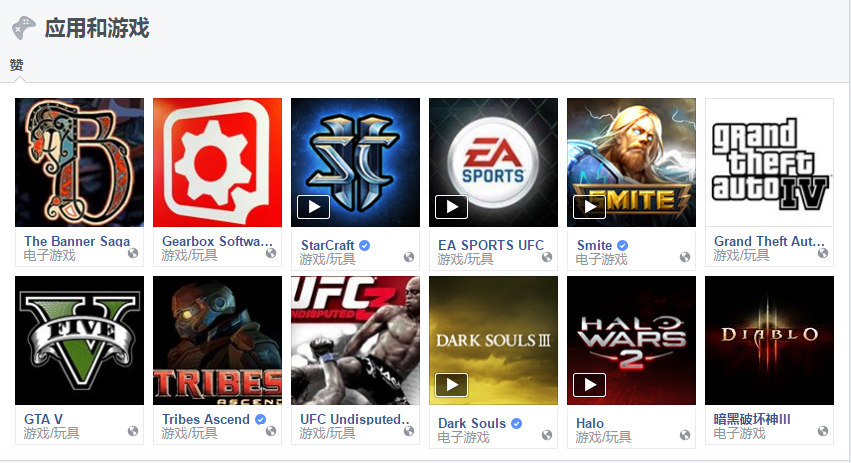


图3.2 用户JOSHUA JAMES WILLIAMS关注的应用和游戏



图3.3 用户JOSHUA JAMES WILLIAMS喜欢的书籍

### 常用的FACEBOOK广告投放优化策略

FACEBOOK广告投放优化策略：

1. 对于公司在FACEBOOK已有的粉丝，可以提取他们的信息，对这些真实粉丝的营销效果将远远优于陌生的用户。
2. 上传手里的客户邮箱列表到FACEBOOK，FACEBOOK会自动匹配邮箱列表中的FACEBOOK用户，所有匹配的用户会被挑选出来。这些用户的转化率会非常高。
3. 在大规模导量之前，进行小规模的广告投放，根据投放的结果，总结受众的共性特征，来降低获取新用户的成本。

利用爬虫技术积累大量用户，可以为运营广告投放组提供有价值的参考方案，节省广告费用和测试的时间成本。

### 爬虫技术

5.1 Selenium

Selenium 是一个可以让浏览器自动化地执行任务的工具，常用于自动化测试。与bs4等结合使用，也适合爬取动态网页数据。可以用于网页截图，而且由于可编程性，用法更具想象空间。目前，Selenium 支持 Java、C#、Ruby 以及 Python 四种客户端语言。

5.2 代理IP

爬虫的访问量必然比正常用户的量要大很多，因而，使用代理IP，可以实现大量不同IP的访问。所谓代理就是介于用户与网站之间的第三者：用户先将请求发到代理，然后代理再发到服务器，这样看起来就像是代理在访问那个网站了。这时，服务器会将这次访问算到代理头上。同时用多个代理的话，单个IP的访问量就降下去了，于是就有可能逃过平台的检测。

5.3 爬取内容包括不限于以下内容：

* 用户出生日期
* 用户学校
* 感情状况
* 喜爱的电影
* 用户好友
* 喜爱的音乐
* 喜爱的书籍
* 喜欢的运动员
* 喜爱的餐馆
* 接触过的应用与游戏
* 居住的地区
* 性别
* 家庭成员
* 受教育程度