**用户画像**

目录

[一、 概念 3](#_Toc476217493)

[二、 标签库设计： 3](#_Toc476217494)

[三、 用户性别、年龄识别过程： 3](#_Toc476217495)

[四、 用户画像本体表数据入库hbase 5](#_Toc476217496)

[五、 用户画像程序执行 6](#_Toc476217497)

# 概念

**用户画像:**

用户画像也叫用户信息标签化、客户信息。 根据用户的信息和行为动作，用一些标签把用户描绘出来，描绘的标签就是用户画像。

**appID:**

用于标记APP，APP的唯一编号，判断终端安装的APP。

**年龄段说明：**

|  |  |
| --- | --- |
| **名称** | **说明** |
| 年龄段1 | 24岁以下 |
| 年龄段2 | 25-30岁 |
| 年龄段3 | 31-35岁 |
| 年龄段4 | 36-40岁 |
| 年龄段5 | 40岁以上 |

# 标签库设计：

标签库的设计，主要是针对应用市场上主流的App，对App进行分类，App分类信息包括：App名称，男女权重信息，各个年龄段的权重信息。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AppID** | **App名称** | **男性权重** | **女性权重** | **年龄段1** | **年龄段2** | **年龄段3** | **年龄段4** | **年龄段5** |

见appTab表:

10001|QQ|0.001|0.001|0|0.2|0.3|0.2|0.3

10002|飞信|0.001|0.001|0|0.2|0.3|0.2|0.3

10003|MSN|0.001|0.001|0|0.2|0.3|0.2|0.3

10004|阿里旺旺|0.001|0.001|0|0.2|0.3|0.2|0.3

10005|微信|0.001|0.001|0|0.2|0.3|0.2|0.3

10006|陌陌|0.001|0.001|0|0.2|0.3|0.2|0.3

10007|米聊|0.001|0.001|0|0.2|0.3|0.2|0.3

10008|啪啪|0.001|0.001|0|0.2|0.3|0.2|0.3

10009|飞聊|0.001|0.001|0|0.2|0.3|0.2|0.3

10010|来往|0.001|0.001|0|0.2|0.3|0.2|0.3

10011|连我|0.001|0.001|0|0.2|0.3|0.2|0.3

10012|有你|0.001|0.001|0|0.2|0.3|0.2|0.3

# 用户性别、年龄识别过程：

**必备条件：**性别比例、年龄比例必须在识别出AppID后进行，如果AppID编号为空，则不做处理。

性别和年龄的判断，都依赖于标签库的设计，也就是用户手机安装的App，根据手机App计算用户的性别和年龄权重信息，最后判断用户的年龄和性别。

源代码：

1. 对男女性别的计算:

**public** **void** protraitSex(**double** male2, **double** female2, **long** times) {

**double** sum = (**this**.male + **this**.female + (male2 + female2) \* times);

**if**(sum != 0){

**this**.male = (**this**.male + male2 \* times) / sum;

**this**.female = (**this**.female + female2 \* times) / sum;

}

}

2、对年龄段的计算:

**public** **void** protraitAge(**double** pAge1, **double** pAge2, **double** pAge3, **double** pAge4, **double** pAge5, **long** times) {

**double** sum = (age1 + age2 + age3 + age4 + age5 ) // 之前的APP的

+ (pAge1 + pAge2 + pAge3 + pAge4 + pAge5 ) \* times;// 当前的APP的

**if**(sum != 0){

**this**.age1 = (pAge1 \* times + age1) / sum;

**this**.age2 = (pAge2 \* times + age2) / sum;

**this**.age3 = (pAge3 \* times + age3) / sum;

**this**.age4 = (pAge4 \* times + age4) / sum;

**this**.age5 = (pAge5 \* times + age5) / sum;

}

}

**表结构**：**用户画像本体表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **字段名** | **类型** | **长度** | **默认值** | **说明** | **备注** |
|  |  |  |  | **日期** | **记录历史时间** |
| MSISDN | byte | 16 | 全F | 手机号 |  |
|  |  |  |  | 男性权重 |  |
|  |  |  |  | 女性权重 |  |
|  |  |  |  | 年龄段1权重 |  |
|  |  |  |  | 年龄段2权重 |  |
|  |  |  |  | 年龄段3权重 |  |
|  |  |  |  | 年龄段4权重 |  |
|  |  |  |  | 年龄段5权重 |  |

**汇总周期：**每天汇总一次，建议当天凌晨统计前一天数据

画像数据：

20160812|+2SJcwT8h0TSsxVd2pBMjA==|0.508|0.492|0.1|0.2|0.2|0.2|0.3|

20160812|+2YYprC2jdMfQntMAn8lmA==|0.4|0.6|0.0|0.1|0.3|0.3|0.3|

20160812|+2k9y/1uapW4IW6LUq0LGQ==|0.7|0.3|0.0|0.2|0.3|0.2|0.3|

20160812|+3M/cSZerfq+XDC9kMWWcA==|0.5|0.5|0.1|0.2|0.2|0.2|0.3|

20160812|+3NZW71R8R8nL78M3JB8tw==|0.5|0.5|0.1|0.2|0.2|0.2|0.3|

20160812|+40FbnopHa5sY/CzDDf8cg==|0.5|0.5|0.0|0.2|0.3|0.2|0.3|

20160812|+5k9calLrKDGvHyzvO0pvg==|0.5|0.5|0.1|0.2|0.2|0.2|0.3|

20160812|+5wU5NlU1iJUZ5sMN91DUw==|0.5|0.5|0.2|0.2|0.2|0.2|0.2|

20160812|+69Ng7CRkKTr2hbYtgeiAw==|0.8|0.2|0.0|0.2|0.3|0.2|0.3|

20160812|+6YeD7QUw8Hy6qN7Dxk7vQ==|0.8|0.2|0.0|0.2|0.3|0.2|0.3|

20160812|+A4H4S++fiGw3pKi8xpwrw==|0.698|0.302|0.0|0.199|0.3|0.201|0.3|

20160812|+AJEIaelS8LGbHqdDaorzg==|0.7|0.3|0.0|0.6|0.3|0.05|0.05|

20160812|+AbbMhECgaMivqZBr/U+Aw==|0.5|0.5|0.1|0.2|0.2|0.2|0.3|

# 用户画像本体表数据入库hbase

创建hbase表：

create 'user\_draw',{NAME=>'draw',VERSIONS=>1,BLOCKCACHE=>true,BLOOMFILTER=>'ROW',COMPRESSION=>'SNAPPY'},{SPLITS => ['/','+','0','1','2','3','4','5','6','7','8','9','a','b','c','d','e','f','g','h','i','j','k','l','m','n','o','p','q','r','s','t','u','v','w','x','y','z','A','B','C','D','E','F','G','H','I','J','K','L','M','N','O','P','Q','R','S','T','U','V','W','X','Y','Z']}

入库hbase代码：

**for** (Text v2 : val) {

String[] splited = v2.toString().split("\\|");

//rowkey

**if**(k2.toString().length()!=0){

Put put = **new** Put(Bytes.*toBytes*(k2.toString()));

//跳过写入Hlog，提高写入速度

put.setDurability(Durability.***SKIP\_WAL***);

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("mdn"), Bytes.*toBytes*(splited[1]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("male"), Bytes.*toBytes*(splited[2]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("female"), Bytes.*toBytes*(splited[3]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("age1"), Bytes.*toBytes*(splited[4]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("age2"), Bytes.*toBytes*(splited[5]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("age3"), Bytes.*toBytes*(splited[6]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("age4"), Bytes.*toBytes*(splited[7]));

put.~~add~~(Bytes.*toBytes*("draw"), Bytes.*toBytes*("age5"), Bytes.*toBytes*(splited[8]));

context.write(NullWritable.*get*(), put);

}

}

# 用户画像程序执行

程序运行环境：hadoop集群、zookeeper、hbase。

appTab.txt规则库表hdfs路径：/home/appTab.txt

程序执行jar包：userdraw.jar

程序执行命令：hadoop jar userdraw.jar 数据输入路径 job1输出路径 job2输出路径

# 文件编码问题需要注意！！！