# 年龄信息抽取概要文档

版本：V1.0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **修订记录** | | | | | |
| 日期 | 版本 | 修订内容摘要 | 修订 | 审核 | 批准 |
| 2020-04-23 | V1.0 | 完成初稿 | 曹博聪 | 詹军、元方 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

上海创蓝传播文化有限公司

**目录**

[年龄信息抽取概要文档 1](#_Toc7723)

[1. 背景和当前存在的问题](#_Toc13949) 2

[2. 解决方案 3](#_Toc19861)

## 背景和当前存在的问题

从短信文本中抽取年龄信息分为直接抽取和间接定义两种方式，直接抽取目前通过规则提取身份证号码，今年xx，xx岁三种方式来定义号码的具体年龄，而间接抽取需要通过号码接收的短信内容和签名所属行业来打标签，间接抽取年龄信息存在难定义，难分类，难验证的问题。难定义在于年龄区间难以定义；难分类在于一个号码会接收到的短信内容具有多样性及重复性；难验证在于通过特征和逻辑来间接定义的年龄区间难以验证。本文档目的希望解决间接抽取年龄属性的难定义、难分类、难验证的问题。

### 1.2 解决方案

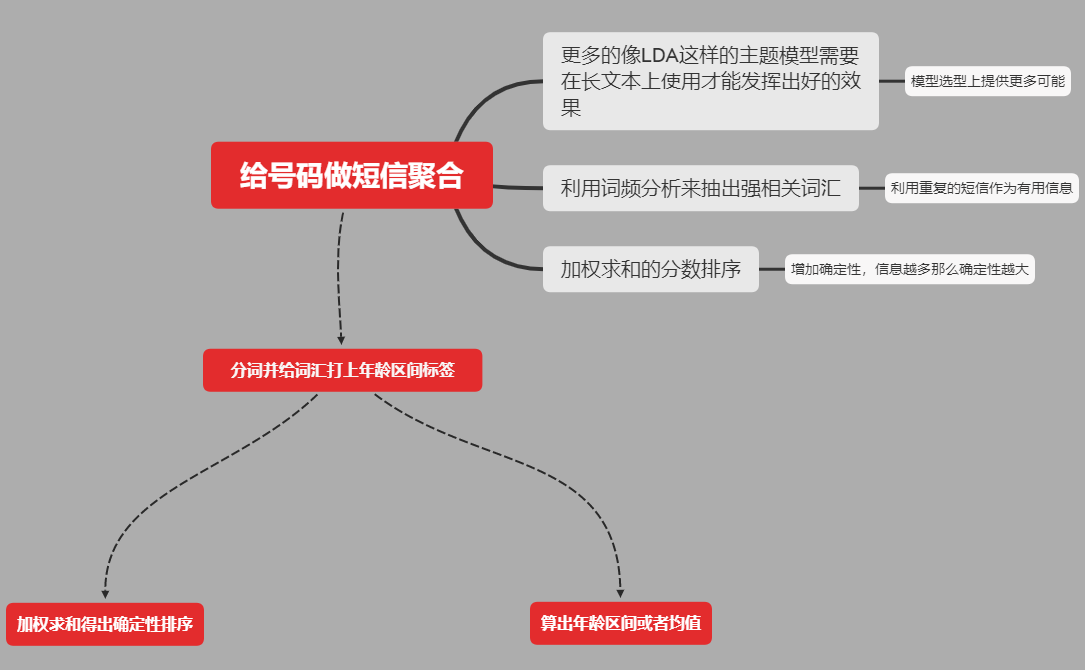


图1 年龄抽取算法设计

1. 年龄定义：

由于目前没有指定年龄区间的具体定义，所以我们需要要从平台已有的短信内容给出一个合理的年龄区间定义，为此我们需要抽出每个号码接收到的高频词汇、关键词、产品名称来定义年龄区间。

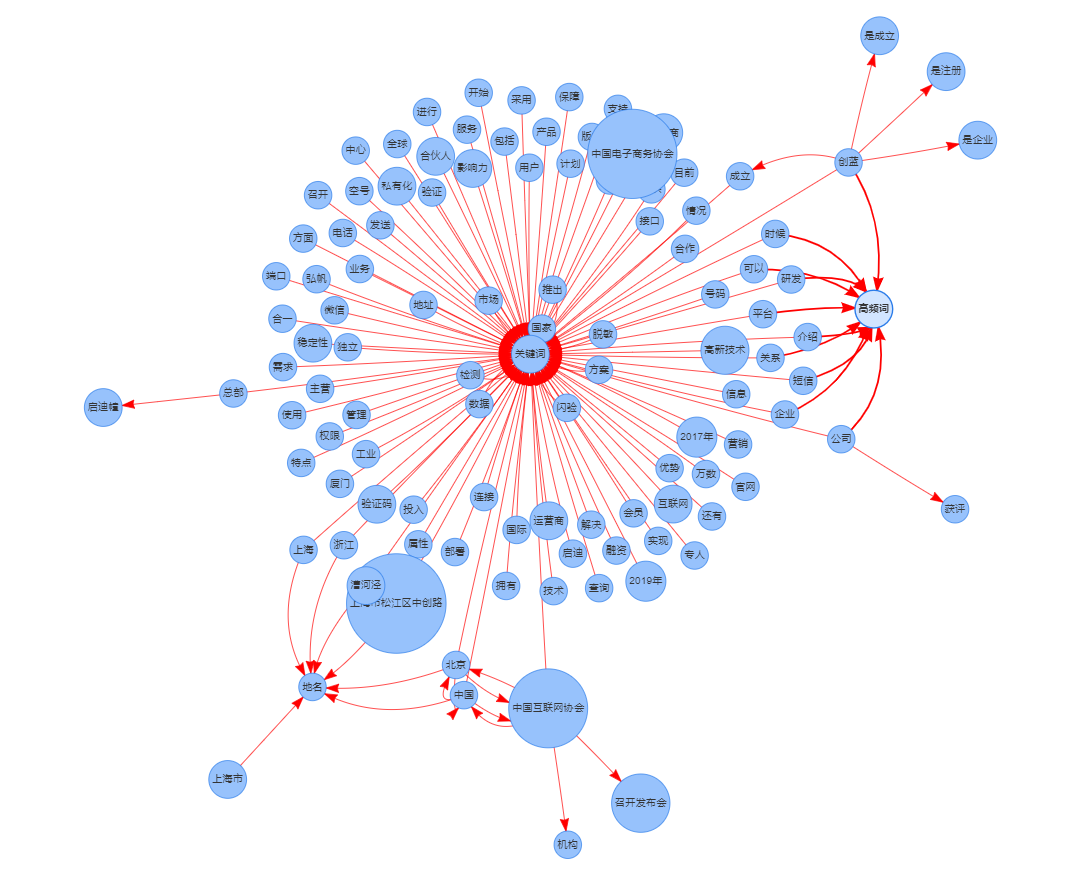


图2关键词、高频词、产品抽取结果的wordcloud呈现

方案明细：关键词算法比如LDA、textRank、加权TF-idf，统计词频此处省略，产品名称抽取目前依赖分词器，因为对细粒度要求较低。经过关键词，高词频词汇，产品名等抽取方式之后，将抽取结果人工打上年龄区间标签，举个栗子：

尿不湿：25-40岁

路虎：30-50岁

1.年龄分类：直接通过1的结果做分类。不同算法侧重点不一样：

取交集：30-40岁

取并集：25-50岁

取交集平均：（30+40）/2=35岁

取并集平均：（25+50）/2=38岁

取交集和并集平均：（35+38）/2=36岁

2.确定性排序：利用词频给分类结果做确定性排序

方案分析：为什么要用词频作为判断依据？

第一个原因是由于大部分号码接收到的短信条数有限，所以能抽取到的信息非常有限，也给抽取到的结果是否准确带来了很大的干扰。根据信息熵的原理，一个号码接收到短信越多，同时又抽取到了年龄信息，那么抽取的结果能反应的此号码的真实年龄结这个事件的确定性就越高；反之，如果一个号码只接受到一条短信，同时抽取到了年龄信息，那么这个抽取结果应该是很不确定的。

第二个原因是这样做抽取到的结果的抗干扰性会更强，如果用户直接收到很少量的短信，那么很可能是泛营销或错发等因素导致。

经分析，词频能很好的反映出结果的确定性。利用一个号码频繁接收到的词汇能很好的反应用户的关注度和用户本身在第三方企业或平台的重要性。举个栗子，如果一个手机号码累计接收到100条短信，在这100条短信组成的长文本中频繁出现车词汇，如路虎，那么很有可能那段时间关注过车类，如果这些短信是来自同一家平台发的短信，那么是因为此用户被这家平台判断为价值用户，从而反应用户的重要性；如果短信来自不同平台，那么可能他在不同平台留下过联系方式，体现了用户关注这款产品并进行了对比。所以统计用户接收到的短信里面的高频词汇能很好的反应用户的重要性和用户对相关产品的关注度，而且词汇出现的越多那么就越能说明用户被打上这个词汇所蕴含的年龄信息越确定。同样是出现了路虎车型词汇，在一个用户接收的所有短信里面抽取到多个路虎车的词汇，那么他被打上路虎（30-50岁）这个结果就会越确定；反之，如果只抽取到一个路虎车的词汇，那么他被路虎（30-50岁）这个结果的确定性值就会很低（实际情况中会根据多个词汇进行计算来判断一个号码的年龄，如下）

词频可以通过加权求和来做确定性排序。排序算法如下：尿不湿（3次\*2,）、路虎（5次\*2）、iphone（2次\*0）,胸罩（8次\*0） 这些抽出来的词汇，经过加权求和后得出：6+10+0+0=16 做归一化处理后是百分比

3.年龄验证：通过电话回访？