**人人用户性格分析SRT第十二周进度报告**

**组员：汪忆辰、叶紫、周若凡**

1. **数据收集和抓取情况**

现在通过我们的性格测试APP收集到的用户样本为89个。

这周我们已经开始对每个用户采集社交网络相关的信息，通过整理、分析、阅读paper和讨论，我们制定了若干个数据特征（详细列表可以见下面），并已经通过python的web编程抓取了用户有关状态、分享和博客的信息。之后还会通过人人的api接口抓取用户的个人信息、相册里的图片、和音频（语音相册）。

由于token使用有时间限制，现在抓取了已经填写性格测试问卷人数一半以上的用户的社交网络上的信息，还在进一步处理格式，以及尚未抓取的（个人信息、图片等）。

每个用户的社交网络相关数据特征抓取的结果存放在一个文件夹下（以该用户的人人id为名的文件夹），以方便往后的数据分析。

1. **数据特征抽取**

现在已经通过java、c和matlab编程分析抓取到的信息并量化了这些信息，为日后统计工作做了铺垫。具体的数据特征如下：

1. 状态（处理文件：status.h & StatusInfo.java）

|  |  |
| --- | --- |
| **抓取的信息** | 用户每条状态的内容，每条状态的创建时间，每条状态的分享数，每条状态的评论数量，状态是否为分享 |
| **量化的数据** | 状态数，平均被分享数，平均评论数，分享状态比例，状态长度，平均两条状态间隔 |
| **想到可能会做的** | 表情数统计，内容tag，把分享的状态和自己发的分开统计 |

1. 日志（处理文件：blog.h & BlogInfo.java）

|  |  |
| --- | --- |
| **抓取的信息** | 每篇日志的标题，内容长度，类型（通过什么发布），创建时间，分享数量，是否加密，浏览数量，回复数量 |
| **量化的数据** | 日志数，平均被分享数，平均评论数，平均浏览数，平均长度，手机发布占比例，公开比例，平均两篇日志的间隔 |
| **想到可能会做的** | 日志中插件数量（表情，图片，壁纸，音乐），评论中自己评论占的比重，标题和内容的tag |

1. 分享（share.h & ShareInfo.java）

|  |  |
| --- | --- |
| **抓取的信息** | 每个分享的标题、URL、回复数量、分享的时间 |
| **量化的数据** | 分享数，照片类型比例，日志类型比例，视频类型比例（只能识别url，现在识别的有youku tudou 56 youtube），其他类型比例，评论数量，平均两个分享的间 |
| **想到可能会做的** | summary和标题中的tag，如果是图片/视频分析缩略图 |

1. 个人信息（处理文件：profile.h & ProfileInfo.java）

|  |  |
| --- | --- |
| **抓取的数据** | 姓名，是否星级用户，basicinfo的填写比例，education的填写比例，work填写的数量，like的数量，app数量，来访数，关注小站数量，喜欢的音乐数量，填写看过的电影数，好友数，好友密度关系 |
| **量化的数据** | 除了名字以外的上述特征 |
| **想到可能会做的** | 名字非中文字多少，非中文和英文字数字多少 |

1. 其他时间相关特征（处理文件：DataProcessor.java）

|  |  |
| --- | --- |
| **抓取的数据** | status.dat share.dat blog.dat |
| **量化的数据** | 新鲜事发布的平均时间（24小时），过期的新鲜事，新鲜事平均时间间隔 |

1. 图片特征（处理文件：face\_detection.m & photo.m）

|  |  |
| --- | --- |
| **抓取的数据** | 尚未抓取（预抓取：所有相册里的照片） |
| **量化的数据** | 颜色直方图、明暗度，人脸识别（目前算法不是特别优，但有一定参考价值） |

1. 音频特征（尚未处理）

|  |  |
| --- | --- |
| **抓取的数据** | 尚未抓取（预抓取：所有语音相册里面的声音文件） |
| **工具** | 打算使用matlab来分析所有的音频 |
| **量化的数据** | 语速、基频、平均功率等 |

1. **统计方法**

收集到上述所有的数据之后，我们打算通过线性回归的方法来计算每个feature对于各个性格的贡献大小。我们会删除一些相关度很低的feature，并通过相关度来给每个feature制定不一样的权值，用机器学习的方法使最终模型逼近我们的数据集。

1. **下一步工作分工**

|  |  |
| --- | --- |
| **叶紫** | 进一步提取人人上的其他信息：包括照片、音频等。 |
| **汪忆辰** | 格式化数据，并尝试运用weka建立初步模型，调整参数。 |
| **周若凡** | 完成音频相关数据的量化，并开始写最终展示的app界面。 |