**人人用户性格分析SRT第十一周进度报告**

**组员：汪忆辰、叶紫、周若凡**

1. **人人APP数据收集情况**

现在通过我们的性格测试APP收集到的用户样本为86个，虽然尚未达到我们预先制定的目标样本数量，但是可以直接进入下一个阶段的工作，对用户数据的特征进行抽取和量化分析。

每一个用户性格测试的结果单独存在一个文件中，格式如下：

220956324.txt——文件名（用户账号token）

{'A': '0.917', 'C': '0.833', 'E': '0.656', 'M': u'5', 'O': '0.65', 'N': '0.062'}——用户性格测试标准结果，取自权威测试网站<http://www.outofservice.com/bigfive>。

1. **机器学习算法**

我们进行对比测试的机器学习方法是针对量化数据（arff格式）进行估值的模型，包括线性回归模型、M5Rules和高斯过程回归。利用现有的机器学习工具Weka，这一过程并不困难。我们将对每个模型的参数进行调整，已达到该模型的最好性能。在此基础上对比各模型的性能选择最好的模型用到最终的APP中。

M5Rules是用于解决回归问题的机器学习算法，基于分割覆盖的思想根据给定的训练集生成一系列规则，每一次迭代会建立一棵M5的决策树，并把最“好”的叶子节点添加到规则中。

高斯过程回归类似于贝叶斯回归，只是用核函数替代了贝叶斯中的线性函数进行回归。

1. **数据特征抽取**
2. 状态

|  |  |
| --- | --- |
| **In** | status.dat: content, createTime, shareCount, commentCount, sharedStatusId |
| **Out** | 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 |
| **TODO** | 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

|  |  |
| --- | --- |
| **In** | blog.dat: title, content length, type(0=“TYPE\_DEFAULT”, 1=“TYPE\_WAP”), creatTime, shareCount, accessControl(0 = “PUBLIC”, 1=others), viewCount, commentCount |
| **Out** | 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 |
| **TODO** | 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

|  |  |
| --- | --- |
| **In** | share.dat: title, url, comment count, shareTime |
| **Out** | 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 |
| **TODO** | summary和标题中的tag，如果是图片/视频分析缩略图 |

1. 个人信息

|  |  |
| --- | --- |
| **In** | profile.dat: name, star, basicInfo(只记录填写比例), education(只填填写比例), work(个数), like(把所有填写的数量), app count, visitor count, page count, zhancount, music count, movie count, friend count, density（好友关系密度） |
| **Out** | 除了名字以外的上述特征 |
| **TODO** | 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

|  |  |
| --- | --- |
| **In** | status.dat share.dat blog.dat |
| **Out** | 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

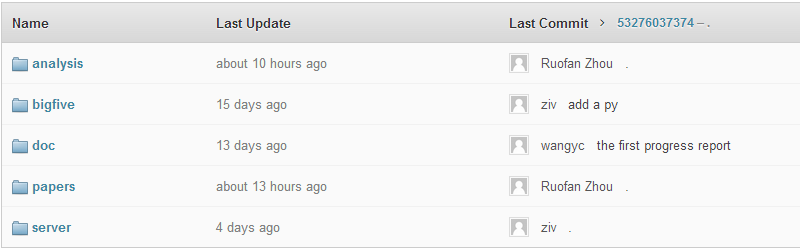
|  |  |
| --- | --- |
| **In** | photo.dat |
| **Tools** | CxImage的c++库，可以方便地取出每个像素上的颜色和灰度 |
| **Out** | 颜色直方图、明暗度，如果可能考虑人脸识别的特征 |
| **TODO** | 使用CxImage的API接口 |

1. 音频特征

|  |  |
| --- | --- |
| **In** | photo\_vocal.dat |
| **Tools** | libsndfile的c库，matlab工具 |
| **Out** | 语速、基频、平均功率等 |
| **TODO** | 针对wav格式的音频分析并不困难，可以自己实现音频特征提取的代码 |

1. **文件说明**

我们的代码和文档统一放在git9上：[git@git.net9.org:zrfan/srt.git](mailto:git@git.net9.org:zrfan/srt.git)



文件的具体说明如下：

|  |  |  |
| --- | --- | --- |
| **目录** | **文件** | **功能** |
| **papers** | 16篇项目相关论文，包括社交网络的用户性格分析、音频和图片的特征提取 | 对每一步实验起指导性作用 |
| **doc** | 选题报告和项目进度报告 | 总结每一阶段的完成情况 |
| **bigfive** | get.py  index.html  test.html | 完成标准bigfive性格测试网站的自动填表和结果获取功能 |
| **server** | 基于Django架构的服务器代码和获取的数据 | 为人人APP提供服务器  为模型训练提供数据 |
| **analysis** | data  blog.h  profile.h  share.h  status.h  test.h  features.rtf | 定义提取出的特征的接口，包括状态、日志、分享、个人主页等。后续会增加数据处理的代码。 |

1. **下一步工作分工**

|  |  |
| --- | --- |
| **叶紫** | 特征抽取工作，利用人人上现有的API提取分享、相册、日志、状态、照片，以及好友列表和个人主页的信息，进行量化后作为特征。 |
| **汪忆辰** | 比较不同的机器学习算法的性能，训练模型并进行评估。 |
| **周若凡** | 研究音频和图片的可抽取特征，完成后期人人APP的制作。 |