**信号处理原理SRT选题报告**

计14 2011011279 汪忆辰

计15 2011011300 周若凡

计15 2011011308 叶 紫

1. **选题内容**

通过对社交媒体（如微博、人人）中的文本评论、音频数据、图片数据、视频信息（URL、标签、评论等）为基础，分析用户的性格特点。

1. **选题背景**

现在使用社交网络的人越来越多，而一些学术研究表明人们在社交网络上的一些行为和自身的性格有一定的联系。我们希望能通过对社交媒体上各种信息的分析给予用户性格上的一个预测，进而可以获得一些统计意义上的结论。

1. **文献综述**

Golbeck J, Robles C, Turner K在Predicting personality with social media[C]中介绍了他们关于利用Facebook的用户数据预测用户性格的工作。他们使用的是近年来公认为比较有参考价值的Big-Five人格特征作为性格模型，包括Openness to experience（开放性）、Conscientiousness（严谨性），Extraversion（外向性），Agreeableness（宜人性），Neuroticism（神经质）。一般Big-Five性格测试是问卷测试的形式，得到五个方面的分数比重，类似下图。



他们通过Facebook上的应用向279位用户收集了性格调查结果和个人数据，总结出六大类特征，包括语言特征、个人信息、活动和偏好、Facebook内部数据、线性回归结果、朋友圈结构（见下图）。然后分析了各种特征和性格之间的相关度，筛选出74项特征项，并用这些数据作为训练集训练出来两种机器学习的模型：M5’ Rules和Gaussian Processes，作为预测用户性格的模型。对比10折检验的效果，他们认为效果最好的是M5’ Rules模型。文章最后还讨论了用户性格预测的潜在市场价值。



另一篇文章，Quercia D, Kosinski M, Stillwell D的Our Twitter profiles, our selves: Predicting personality with Twitter，是基于Twitter用户数据的用户性格预测分析，其中也使用了前一篇中提到的M5’ Rules模型，但是只考虑了3个特征项（following, follower and listed counts）。

他们的分析结果表明，Twitter上不同的用户类型，包括listeners, popular, highly-read, and influential users，都在情感上表现的较为稳定，大部分属于外向型性格，而非神经质性格。

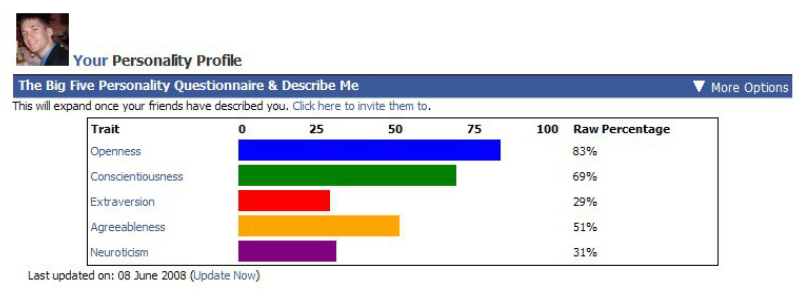
1. **研究方案**

1、我们计划调查的用户社交平台为人人网，初步拟在人人上发布以大五类人格特性为模型的性格测试应用，收集200份以上的用户数据和性格测试结果，数据采集的前提是经过用户的同意。

2、在之前的基础上，我们希望增加考虑用户相册、视频、表情符号等多媒体资料中可抽取的特征（标签、类型、评论、分享次数等）。通过对这些特征和用户性格的相关度的分析，加上前人已经总结出的一些可用特征，构成我们自己的特征集。

3、用不同算法（M5’ Rules和Gaussian Processes等）训练出性格预测的模型，通过N折交叉验证对不同模型的预测性能作出评价。

4、用我们训练出的模型，构建一个人人App，能够自动分析用户信息，对用户的大五类人格特性各项指标做出预测，预测结果类似前文中的五边形分布图表或下图，并附加一定的性格文字描述。如果项目时间允许，我们会收集用户体验并反馈结果。



注：社交网络我们暂限制于人人网（www.renren.com）。

数据收集范围也暂定在本系以及周围的好友同学。

1. **进度安排**

前期准备在第8周周末完成（包括收集标准集数据、算法的寻找）

模型部分完成：基本代码在第10周周中完成；之后1周利用数据调整参数和模型。

最后的时间完成在社交网络（人人）上的相关应用制作。

1. **预期成果**

能够通过对人人网上信息的抓取给出一个较为科学的用户的性格分析（五型性格各自的分数值），并制作成一个应用给用户娱乐。