开放型实验设计文档

一、 背景及原理介绍

在新浪微博中, 每天都有许多热门话题引发微博用户的广泛讨论。分析这些热门话题的特征（如粉丝在地理上的分布，粉丝间组成的社交网络等），对于发现新浪微博中有相似兴趣的用户群体，理解信息在社交网络中传播的行为有重要意义。

使用网页爬虫技术对新浪微博进行爬虫，得到热门话题的相关信息。利用这些信息分析该热门话题的特征。

二、方案设计

所需环境：一台能联网的电脑；

具体实现步骤：

1. 对新浪微博进行爬虫，难点包括模拟登陆和之后的分析网页结构以便进行有效爬虫。此处注意爬取微博移动版，因为其结构简单。
2. 数据结构：爬虫获得的数据包括关注者/讨论者（非大V）的所在位置，年龄，性别，关注/讨论话题的时间，关注的人，每个占一行；大V的信息同上，另存，加入一条其所属类别（第一项吧）。
3. 统计关注话题的粉丝和讨论用户的地点分布，年龄分布和性别分布。
4. 分析粉丝和讨论用户各自组成的社交网络（每个用户为一个节点，用户间的关注关系为边），统计节点的出入度的分布，社交网络中的孤岛个数（内部连通且与外部无连接），平均每个孤岛的节点数等数据特征。
5. 分析大V在热门话题传播中的作用，观察大V参与话题前后关注量增长曲线的斜率，分析前后关注者中为大V粉丝的比例变化。
6. 观察大V所属类别，和这个话题类别的关系。

三、分析及验证方案

爬虫获得了相关数据，通过使用手机观察对应热门话题，看是否相同，如果相同则说明爬虫成功。数据分析部分程序能运行且结果符合逻辑即说明达到预期目标（就是嘛，这里怎么会出错）。

四、实验分工及时间安排

分工：罗惠文负责爬虫，单萌和李智涛负责数据处理和写实验报告。

时间安排：周日晚之前爬虫结束，周二晚之前完成数据处理，周三晚之前写完实验报告。