《软件工程课程设计》课程报告

课题名称： 微博爬虫系统

课题负责人名（学号）： 秦浩均

同组成员名单（角色）：

何长鸿 2016141482154

宁永豪

指导教师： 陈虎

评阅成绩：

评阅意见：

提交报告时间：2018 年 12 月 24 日

目录

微博爬虫系统设计与开发

计算机科学与技术

学生　何长鸿　　**指导老师**　陈虎

[摘要]　微博作为当下最为流行的社交软件之一，拥有广泛的用户群体。本系统通过多线程爬虫采集微博用户数据存档，包括用户微博、关注、粉丝人际关系。采集速度可调节，生成 excel 文件，通过邮件方式将数据发送到客户邮箱，并利用自然语言处理、人工智能等手段对数据进行情感分析、主题抽取等工作。

关键词：微博、多线程爬虫、数据分析

一、引言

并且已成为社会舆论在中国传播的最主要途径之一。新闻媒体、明星、政府机构等都有官方微博及时发布相关信息。[1]为了快速地获取到微博中人际关系信息,根据网址的特点,文章提出了一种基于新浪微博的爬虫程序设计方法。本方法通过模拟登录新浪微博,实现抓取微博中由指定用户出发的关注对象的名称等信息;该程序利用解析关键路径,广度遍历等技术,匹配符合规定条件的人物名称,并抓取相关内容;最后对该程序又进一步地优化与改进。实验结果表明:本程序具有针对性强,数据采集速度合理,易推广开发,稳定性强等

优点,为寻求人际关系的研究者提供了寻求微博用户关注者的方法,有利于对微博的后续数据挖掘研究。

除了基本的数据采集功能，本项目初步完成了部分数据分析功能，例如针对用户微博生成词云，未来还将继续加入用户为博情感组成成分分析，用户微博主题分析，图片分析等相关功能。

二、主要承担角色

本团队成员共三人，分别负责项目管理、服务器开发、爬虫系统开发与数据分析处理。

本人负责的部分为爬虫设计与开发，一下文档仅描述爬虫相关内容。

三、实验目的

1. 学习使用软件工程的需求分析、建模相关技术进行软件系统开发
2. 开发稳定高效的微博爬虫系统，为后续数据挖掘工作奠定基础。

四、系统环境

(一) 开发环境

1. Python 3.6.5
2. Windows 10
3. 30 Mbps 带宽网络
4. MongoDB

(二)　运行环境

1. Python>=3.0.0
2. 网络带宽>= 10 Mbps
3. MongoDB stable

五、需求分析

(三)　软件需求功能说明

1. 通过微博 id 获取微博用户基本信息、发布的文章、关注用户、粉丝用户
2. 将数据通过 email 发送到用户邮箱
3. 对数据进行分析，如生成词云

(四)　对功能的一般性规定

软件采用可视化图形界面，界面格式统一，界面功能键排版能使用户能较快找到所需功能按钮，设定默认值以防每次操作要求输入太多，统一的错误提示风格。

(五)　对性能的一般性规定

1. 精度

浮点型数据保存 2 位以上小数，百分比数据小数点后 2 位小数。

2. 输入输出要求

信息录入时，数字、字符、时间日期的格式和长度应遵照提示要求，否则不能提交或提示出错。系统输出时，与金额相关的

数值都保留 2 位小数；百分比数值小数点后保留 2 位小数。

3. 数据管理能力要求

需要管理的问卷

4. 其他专门要求

1) 用户的信息加密、信息认证（登陆访问）
2) 使用方便
3) 可维护性、可补充性、易读性、可靠性

六、接口说明

(六)　用户接口：

输入微博 UID，输出：微博用户博文、关注用户 uid 列表，粉丝用户 uid 列表，微博用户基本信息。

(七)　系统接口：

主要为爬虫系统与服务器系统之间的接口，以下使用代码及注释方式说明，函数参数为服务器需要提供给爬虫系统的接口，返回值为爬虫系统传送给服务器的数据。

1. 根据 uid 列表生成用户基本 profile 信息的 excel 文件
   uids = [6029786152, 5102089477, 2988799167,3554683503]
   path = xls_gener.profile_to_xls(uids)
2. 生成一个用户所有微博到 excel,返回 excel 文件路径
   path = xls_gener.blogs_to_xls(3554683503)
3. 邮件发送附件
   MAIL('Weibo Spider','微博内容数据','709531006@qq.com', path)
4. 生成词云，返回词云图片路径，返回 excel 文件路径
   path = xls_gener.word_cloud(3554683503)
5. 根据 uid 生成粉丝基本信息的 excel，返回 excel 文件路径

path = xls_gener.fans_profile_to_xls(3554683503)

6. 根据 uid 生成关注用户基本信息的 excel，返回 excel 文件路径
7. path = xls_gener.folows_profile_to_xls(3554683503)
8. 微博情感组成分析，生成统计图
9. 生成用户报告

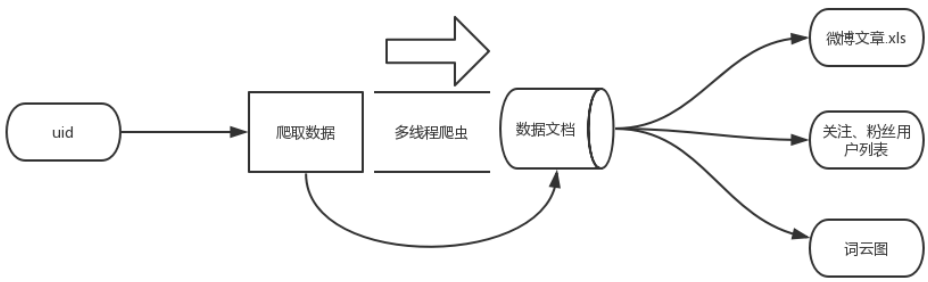
七、设计

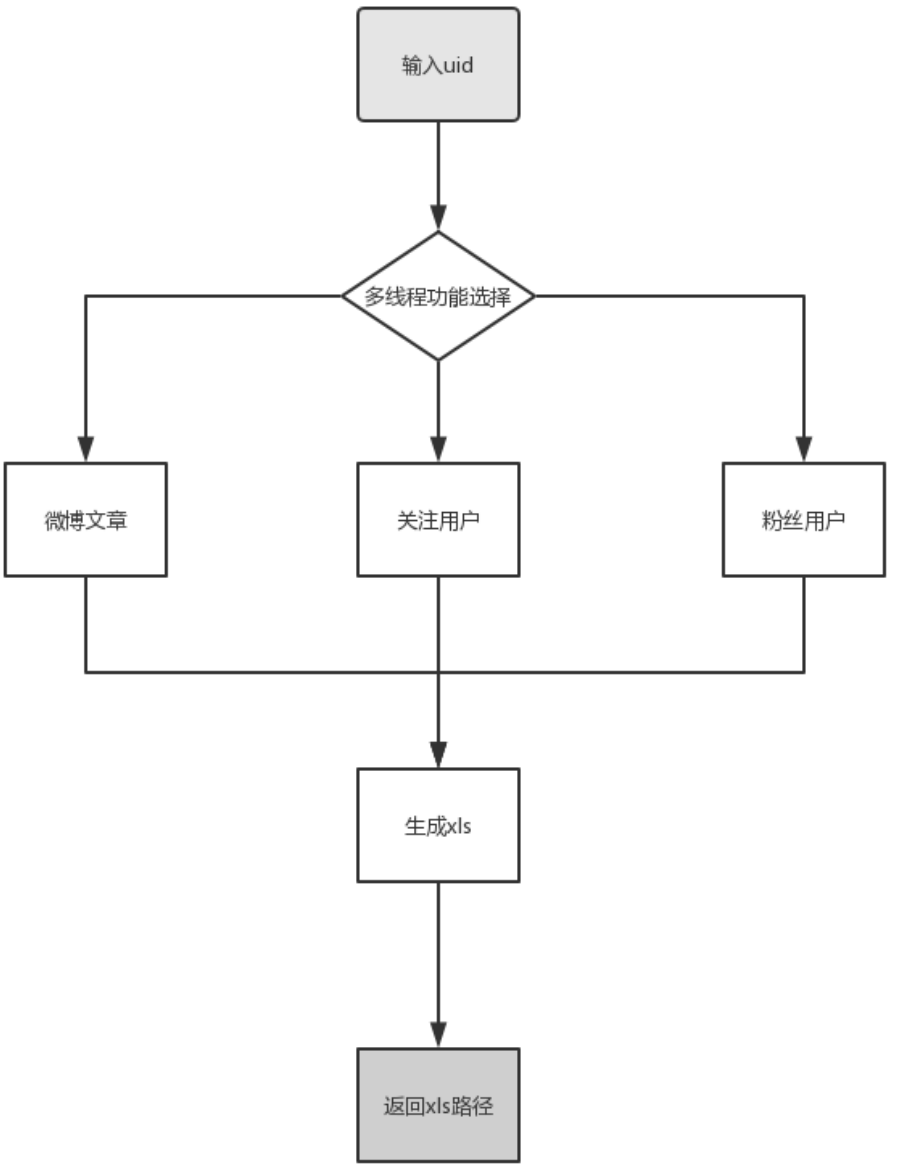
1. 数据库设计：

本系统数据库使用 MongoDB, 数据组织采用文档方式储存，文档结构如下

```
_id: 6029786152
screen_name: "Sillyemon"
statuses_count: 4151
verified: false
description: "有缘也不想再见"
gender: "m"
urank: 14
mbtype: 12
avatar_hd: "https://wx3.sinaimg.cn/orj480/006A4mPmly8fy8xy9fwgtj30ig0igjra.jpg"
profile_api: "containerid=2302836029786152"
weibo_api: "containerid=1076036029786152"
album_api: "containerid=1078036029786152"
▸ weibo: Array
▸ fans: Array
▸ follows: Array
```


2. 系统设计





八、算法说明

1. 微博数据抓取

主线程先获取用户微博数据总条数，计算需要请求的数据次数。然

后使用多线程并行抓取，以克服网络请求造成的 CPU 等硬件资源的浪费。并行速率主要通过线程并发延时来调节，经过测试，当延时为 0.05 秒时，可比较稳定的运行。在完成所有线程的创建后进行线程同步，确保请求执行完毕。

2. 词云生成

在词云生成过程中，需要先对中文自然语言进行分词，本程序使用 thulac 分词软件完成。通过使用蒙版生成不同图像的图片。

3. 情感分析

本系统使用人工智能进行情感分析。情感分析算法为基于 BiRNN-LSTM 的深度神经网络，在看看 AI 提供的数据集上进行模型训练。该模型在测试集三分类（积极、消极、中性）的精度为 72%。

九、参考文献

[1] 胡海潮，基于新浪微博的爬虫设计实现[J]，无线互联网科技，2018 年 09 期

[2] 袁婷婷、杨文忠、仲丽君、张志豪、向进勇，一种基于性格的微博情感分析模型 PLSTM[J], 计算机应用研究