开题报告

学生姓名：鲍颖

所在班级：通信工程1310班

学号：03131313

选题名称：基于Python网络爬虫的天气信息搜索与预报系统

选题目的（为什么选该课题）：

互联网是一个非常庞大的数据库，各种各样的数据与资源以各种形式存储在互联网上，在网页展示形式也非常多：文字、图片、Flash动画或者视频等等。当用户试图在网页上获取一些信息并点击链接的时候，往往会收到大量外带的数据，例如广告推送、网页浮动窗口等，不但容易干扰视线，使用户较难筛选出有效信息，而且某些广告弹窗可能会暗藏一些恶意链接，使用户造成不必要的损失。本课题的主要目的是设计一个可以主动定向收集天气信息的网络爬虫程序，使用户试图去在网络上获取信息的时候，避开这些可能存在的“坑”，高效、快速的完成信息的获取，此程序还可支持用户进行指定城市的天气信息的订阅，被订阅的城市近期的天气数据会主动推送到用户端进行显示，并可根据历史数据，得到每日的降水率，进行一定程度的天气预测。在满足功能的同时还需要考虑到对程序性能上的要求。要考虑到网络爬虫的各种需求，以及网络爬虫主体网站的特性。网络爬虫使用 scrapy 实现多线程，让爬虫具备更强大的抓取能力和灵活性。

前期基础（已学课程、掌握的工具，资料积累、软硬件条件等）：

到目前我已经学习了数据结构、计算机操作系统、网络编程等课程，掌握了mysql数据库、VS2015、Qt等软件工具的使用方法，可以熟练使用的计算机语言有Python/Shell、C++，具备了开发linux服务器端程序的能力，已经查阅的资料包括《Python核心编程》（美）丘恩（Chun，W.J.）人民邮电出版社、《Unix网络编程》（美）史蒂文斯 人民邮电出版社等。软件条件包括windows/Linux操作系统、VS2015、Mysql、PyQt等，硬件条件包括个人开发机器一台(内存8G,CPU:i5-4200H)，阿里云服务器一台(内存1G,单核CPU,带宽100M)。

要解决的问题（做什么）：

1.开发环境的搭建，包括：Python环境的搭建以及PyQt、VS2017、Mysql等软件的安装。

2.熟练掌握Python编程以及Linux操作系统的常用命令。

3.了解网络爬虫技术研究现状，学习网络爬虫的技术原理。

4.学习已经有的爬虫开源框架，了解其优点与缺点，结合自己的使用场景进行优化。

5.设计数据的存储格式，使用什么数据结构。

6.设计与选择客户端与服务器端之间的数据传输协议与格式。

7.对所爬取网站的主体内容与结构的分析。

8.若网站有防爬策略，比如IP限频、IP黑名单等，该如何解决。

9.对网络爬虫工具性能进行测试，对程序功能进行完善。

10.对所爬取到的数据进行分析与归类，让推送到客户端的数据更准确。

11.根据历史数据进行一定程度上的天气分析与预测。

工作思路和方案（怎么做）：

我的工作思路：软件开发中，搭建一个开发环境是最先需要完成的任务，开发环境包括windows上的VS2015、PyQt，Linux上的Mysql、Python3以及一些软件的环境变量的配置等。之后需要进行一些相关资料与知识的搜索与积累，例如windows下使用VS2015和PyQt进行GUI开发的技术，linux下的网络编程的知识，和网络爬虫程序相关的开源框架scrapy的原理的学习与研究，以及如何使用python第三方库request+beautifulsoup进行html网页内容的获取与解析等，预计学习时间为2017年2月27日至2017年3月15日。之后需要开始进行程序的开发，设计与实现网络爬虫程序，并将其作为一个长期运行的服务器程序进行开发，其中编写爬虫程序需要使用scrapy框架，网页内容的请求与获取需要使用第三方库，对所获取网页的内容的解析与信息的提取需要用到正则表达式等技术 。服务器程序要能够在指定时间完成数据的更新功能，这需要使用到到一些Linux下命令行操作与Shell脚本编程的技术，且由于是先进行服务器程序的开发，此时尚未有客户端程序，所以需要在代码中加入调试信息，观察每一个步骤获取到的数据以对服务器程序进行调试，编写测试程序模拟客户端对服务器发起请求，以测试服务器的性能，服务器程序开发时间预计2017年3月16日到2017年4月25日。之后进行客户端程序的开发，此环节较服务器程序开发较为简单，使用PyQt进行GUI界面的开发，并向服务器发起请求，取回已经被服务器分析、处理过的数据，预计开发时间2017年4月26日到2017年5月10日。撰写论文，完成论文初稿，2017年5月15日至2017年5月30日。完善并修改毕业论文，2017年6月1日至2017年6月7日。 最后准备答辩，2017年6月8日至2017年6月17日。