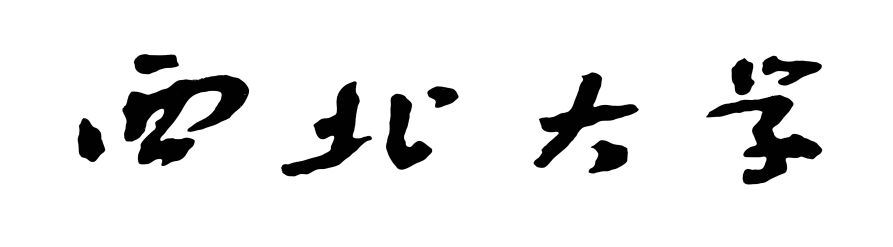
****

**《GIS设计与开发实习报告》**

**基于Python网络爬虫的陕西省气象数据实时发布系统**

**姓 名 何豪**

**学 号 2016116013**

**代课老师 黄昌**

**院 系 城市与环境学院**

**专 业 地理信息科学**

**年 级 2016级**

**教务处制**

目录

[第一部分、作品介绍 3](#_Toc13834607)

[一、需求分析 3](#_Toc13834608)

[二、系统设计 4](#_Toc13834609)

[1.总体设计 4](#_Toc13834610)

[2.数据采集模块设计 4](#_Toc13834611)

[2.数据存储模块设计 7](#_Toc13834612)

[3.GUI功能设计 8](#_Toc13834613)

[4.关键技术 13](#_Toc13834614)

[三、作品亮点与不足 14](#_Toc13834615)

[1.作品亮点 14](#_Toc13834616)

[2.作品不足与解决 14](#_Toc13834617)

[第二部分、操作手册 15](#_Toc13834618)

[一、系统安装 15](#_Toc13834619)

[二、系统使用 16](#_Toc13834620)

[三、系统卸载 17](#_Toc13834621)

# 第一部分、作品介绍

## 一、需求分析

气象数据气是开展天气预警预报、气候预测预估及各类气象服务、科学研究的基础，是推动气象科学发展的原动力。实时气象发布系统是现代气象业务体系的重要组成部分，是提升公共气象服务能力的重要基础。其作用主要体现在以下几个方面：

在防灾减灾方面。我国是世界上气象灾害最为严重的国家之一，灾害种类多，分布地域广，发生频率高，造成损失重。为了有效防御和减轻所造成的损失，提高应对突发公共事件应急的气象保障能力，迫切需要实现实时气象数据的发布，实现对实时气象数据的发布和预测。

在应对气候变化方面。气候和气候变化导致人类生存条件的变化，这不仅是科学问题，更是世界各国政府共同关注的政治问题、经济问题和外交问题。为了推进气候变化综合影响评估工作，减少或消除有关气候变化的不确定性，增强对陕西省气候变化和极端气候事件的监测能力，提高对气候变化的定量描述和预估水平，制定适应和减缓气候变化的对策及措施，需要开展代表性好、精度高、长期稳定运行的气象发布系统。

在提高气象预报预测准确率和精细化水平方面。提高预报预测准确率和精细化水平是气象预报预测业务的核心。为了建立满足公共气象服务需求的各类气象预报预测业务系统，提高预报预测准确率和精细化水平，提供均一性更好、时间序列更长的观测资料，需要建立陕西省气象数据发布系统，以分析陕西省气象服务区、天气气候敏感区的高精度、高准确度、及时、稳定、可靠的观测数据，提高综合气象系统能力。

因此，我设计了一个面向普通群众和各界研究人员的陕西省气象数据实时发布系统。用户可以通过该系统查询、分析陕西省的实时气象数据和历史气象数据。用例图如下：



图1 系统用例图

## 二、系统设计

**注：系统展示的主要功能设计放在GUI设计模块讲解**

### 1.总体设计

系统总体上可以分为三个部分——数据采集模块、数据存储模块和用户图形界面。其中数据采集部分主要负责数据的采集和部分预处理，由Python3代码实现；数据存储模块使用的是MongoDB数据库，它是连接数据采集部分和数据存储部分的桥梁；用户图形界面是系统的核心部分，实现了对数据的分析、管理、显示等功能。总体设计模块图如图2所示。



图2 总体设计模块图

### 2.数据采集模块设计

#### 1）数据采集模块总体设计

数据采集模块，顾名思义主要负责数据的采集。目前模块主要实现了对陕西省地级市气象数据的爬取、解析、入库和导出数据库，其主要功能如图3所示。该模块的工作原理主要是通过Python代码向数据源网页URL发送Get请求，取得数据后通过解析、编码取得需要的数据后存入对应的数据库集合或文档中。主要工作流程如图4所示。此外，代码也留有抓取区县气象数据的URL和方便后续开发的数据类。



图3 数据采集模块功能模块图



图4 数据采集模块主要工作流程图

数据采集模块文件组织如图4所示。其中venv文件夹为运行该模块所必须的虚拟环境；\_\_init\_\_.py为该模块的主程序，可以说它是负责的是数据采集模块的调度；configs.py是数据采集模块的配置文件，通过该文件配置数据库连接参数、爬虫代理（目前使用的两个网站均无任何反爬措施，可以不设置，但实际上程序留有代理，设置好后任然可以使用）、浏览器标识等信息；其他三个文件为主要的爬虫文件，它们由爬虫类组成。



图4 数据采集模块文件组织图

#### 2）数据爬虫设计

爬虫是数据采集模块的核心，该系统数据采集模块的爬虫由3个爬虫文件3个爬虫类组成，具体介绍如下:

①airSpider.py：

该文件下主要含有五个个类CoderSpider，Spiders，AQI，AirAQI，AirCon五个类，其中AirAQI以评价空气质量的各种污染物AQI指数为属性，AirCon以以评价空气质量的各种污染物浓度为属性，AQI类继承了这两个类的全部属性并重写了time和city\_code，由于这三个类暂时未在系统中使用，不做过多描述。

CodeSpider类：该类主要用于爬取城市编号以及对城市编号的Base64加密（URL中要用到）



图5 CodeSpider类对象模型图

Spiders类：该类主要用于抓取各种污染物的AQI、数据入库以及数据库数据的导出和编码。



图6 Spiders类对象模型图

②weatherSpider.py：该类主要用于抓取天气数据并实现入库。



图7 WeatherSpider类对象模型图

③AQForecastSoider.py：由于目前并不完善，该文件内目前只有AQF()函数用于爬取空气质量预报信息，将来还将实现爬取网页源中空气质量预报的各种浓度、AQI信息和一些统计信息。总之，目前的想法是将该文件设计为一个静态HTML的爬虫类，而不是目前的单一函数。（暂未完全完成）

#### 3）主函数设计

在\_\_init\_\_.py的主要工作是将其他的Python代码“打包“，在这里定义了3个函数，它们调用了3个爬虫文件的类，分别用来爬取所需的空气质量数据、天气数据和空气质量预报数据，最后通过主函数创建三个线程来运行，爬取速度更快。

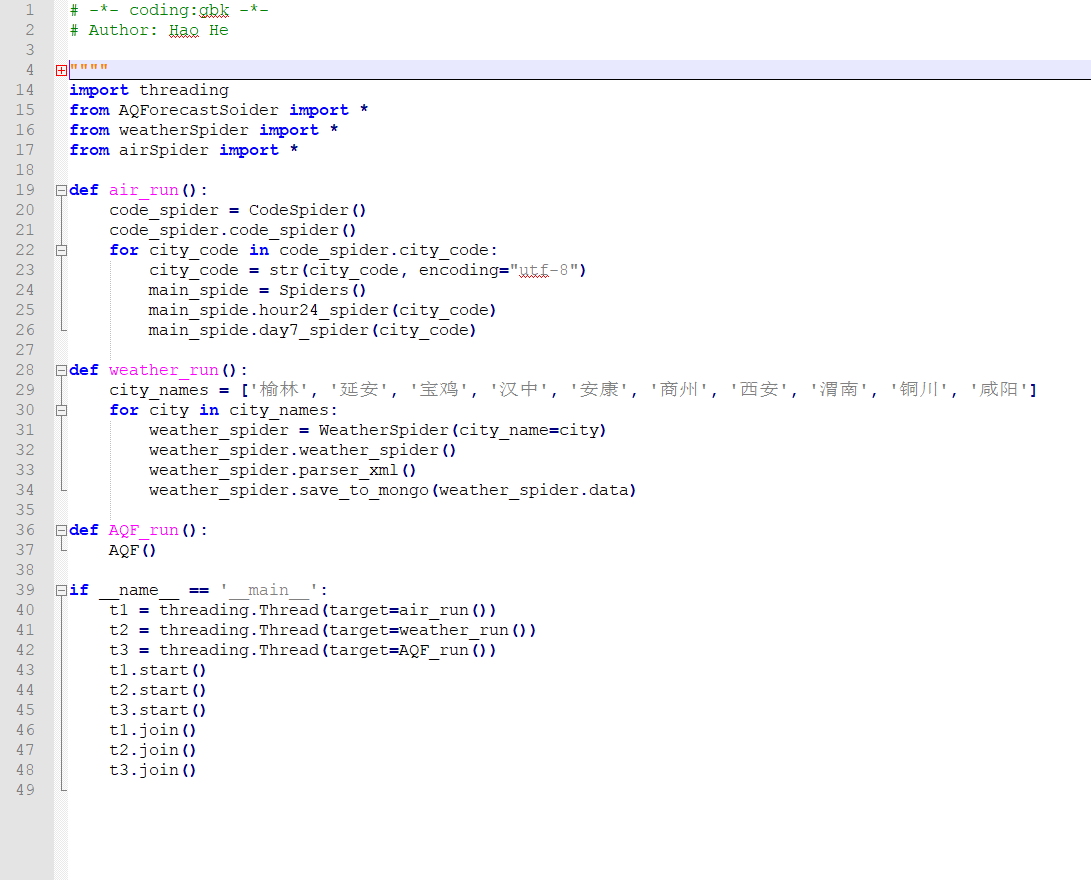


图8 数据采集模块主函数

### 2.数据存储模块设计

该系统使用的数据库为MongoDB数据库，数据库含有一个数据集合，集合中有三个表，分别是day7、hour24、weather，由于MongoDB数据库的数据表是一种类似于JSON的数据结构。因此，所有数据都是以键值对的形式存在，可以把所有数据都看成是字符串类型。由于数据结构单一，也为了简化操作，数据表结构设计得非常简单。

如图9分别是weather、hour24、day7的表结构。



图9 数据库设计

但可以说，目前数据库结构并不完善。因为在之前为了更具快速完成GUI中统计制表功能而省去了day7的时间字段，因此在存储是也使用的是更新而非插入，因此数据库仍有待改善，仍需稍作修改。（尚不完善）

### 3.GUI功能设计

GUI是系统的核心，是用户操作该系统的必要组成部分，图10该系统功能的功能模块图（其中带“\*”号的代表尚未实现，功能模块图只显示到二级功能），图11为系统的主界面。系统的主要功能时对陕西省气象数据的实时可视化显示、分析统计以及气象预警。



图10 GUI功能模块图

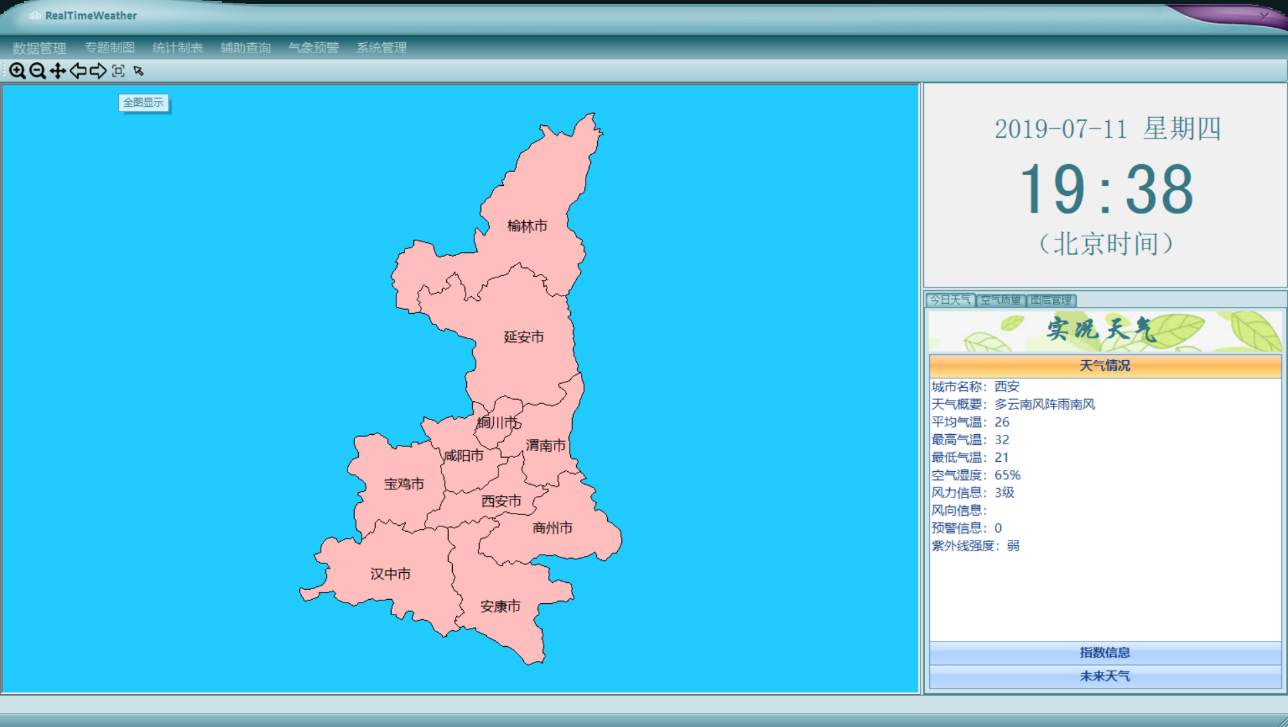


图11 系统主界面

①基本工具条：实现了常用的GIS功能，如放大、缩小、全图显示、上一视图、下一视图、漫游等。

②陕西省地图显示：实现了点击地图矢量面块实现右侧气象数据显示地区的切换，同时，长按可以显示当前地区的空气质量状况。

③数据管理：目前已经实现了数据的导出和地图的导出，数据库的清理尚未完成。

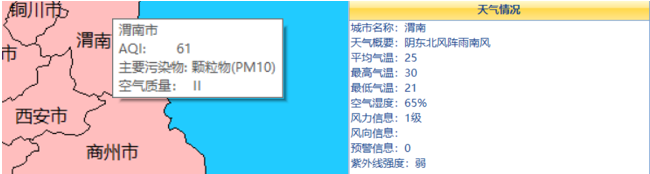


图12 点击查看和切换右侧显示数据

④专题制图：专题制图已经全部实现，代码参考的是GIS\_Demo，但参考代码存在许多BUG和对本系统无用的代码块，这里我们不需要加入全部字段，因此对参考代码进行了较多的修改。

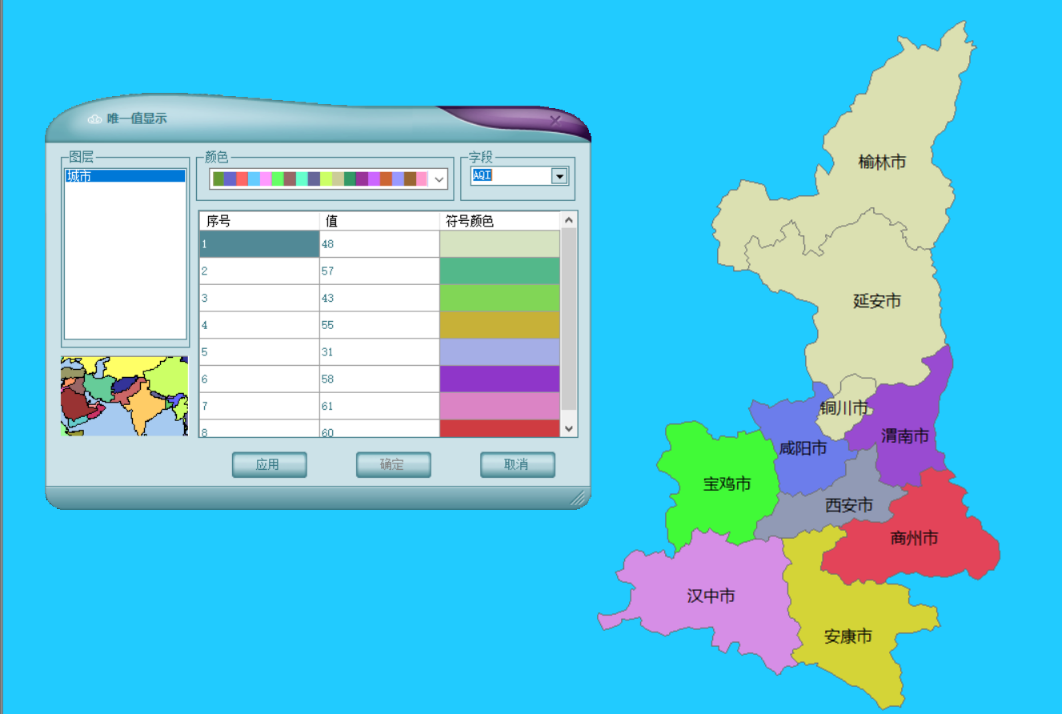


图13唯一值显示

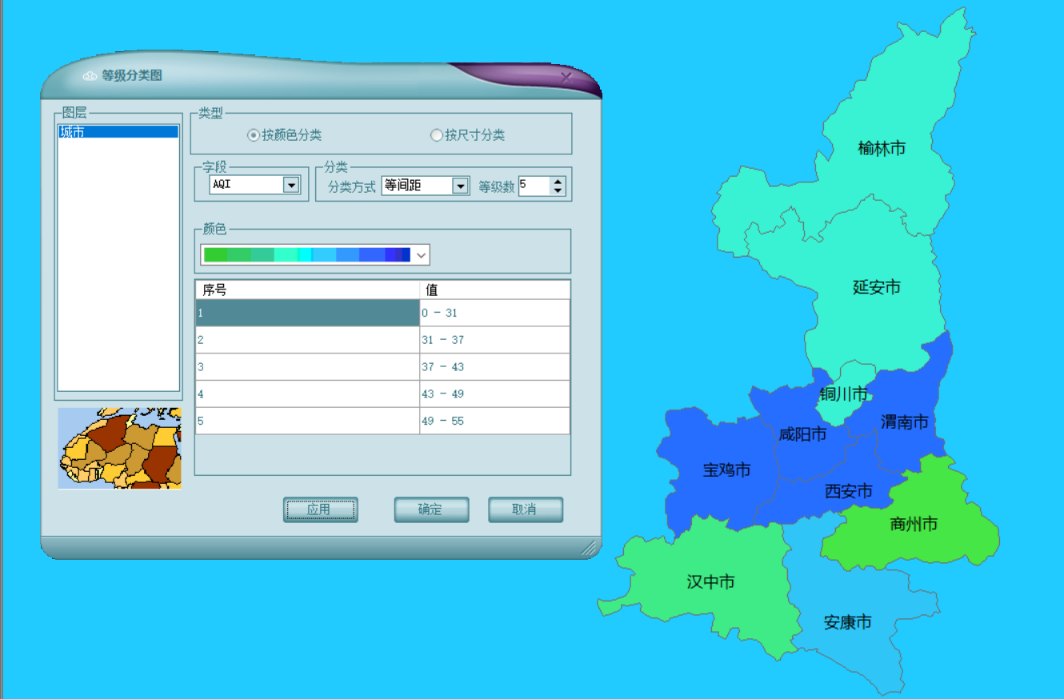


图14按等级分类-1

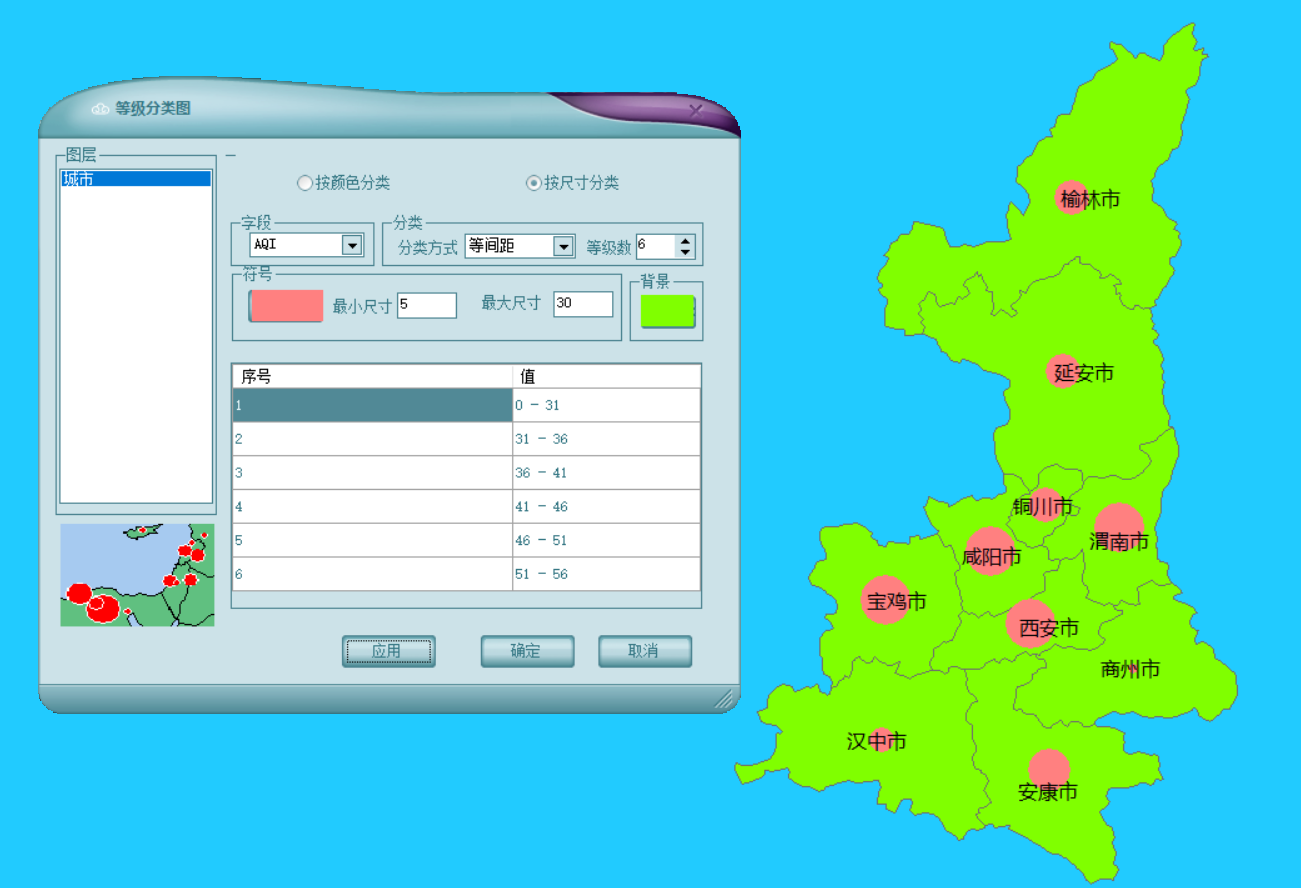


图15按等级分类-2



图16柱状图

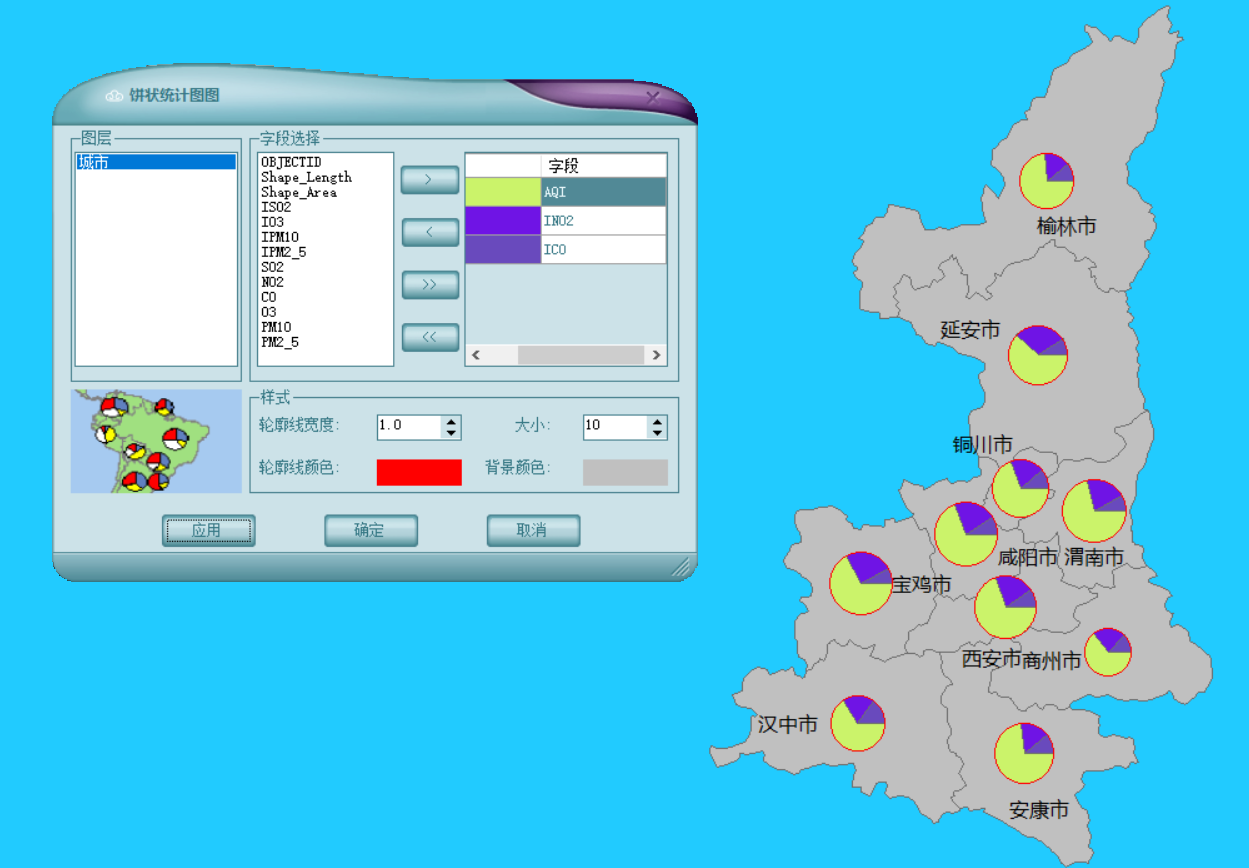


图17 饼状图

⑤统计制表功能尚未实现。预期是实现一些简单的统计表，类似于如图18、图19的形式。

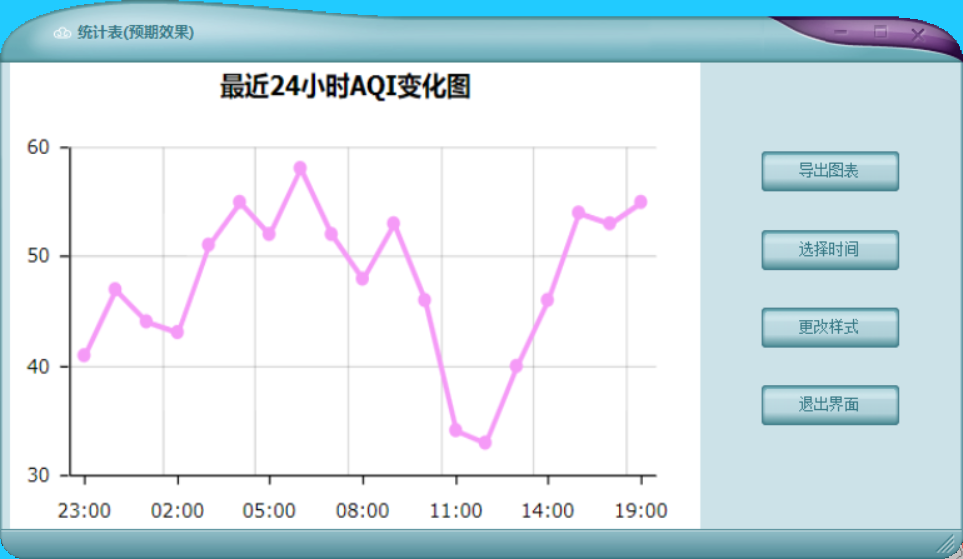


图18 统计制表预期效果图-1

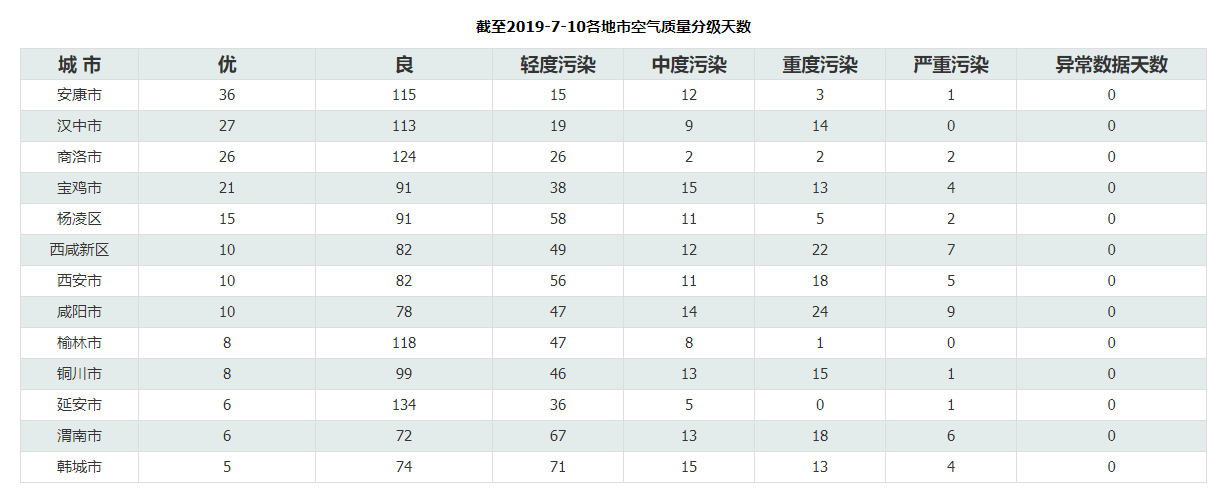


图19 统计制表预期效果图-2

⑥辅助查询功能主要分为天气查询功能、空气质量查询功能、按地区查询功能和辅助查询功能，目前主要实现了属性查询功能，效果如图20所示。

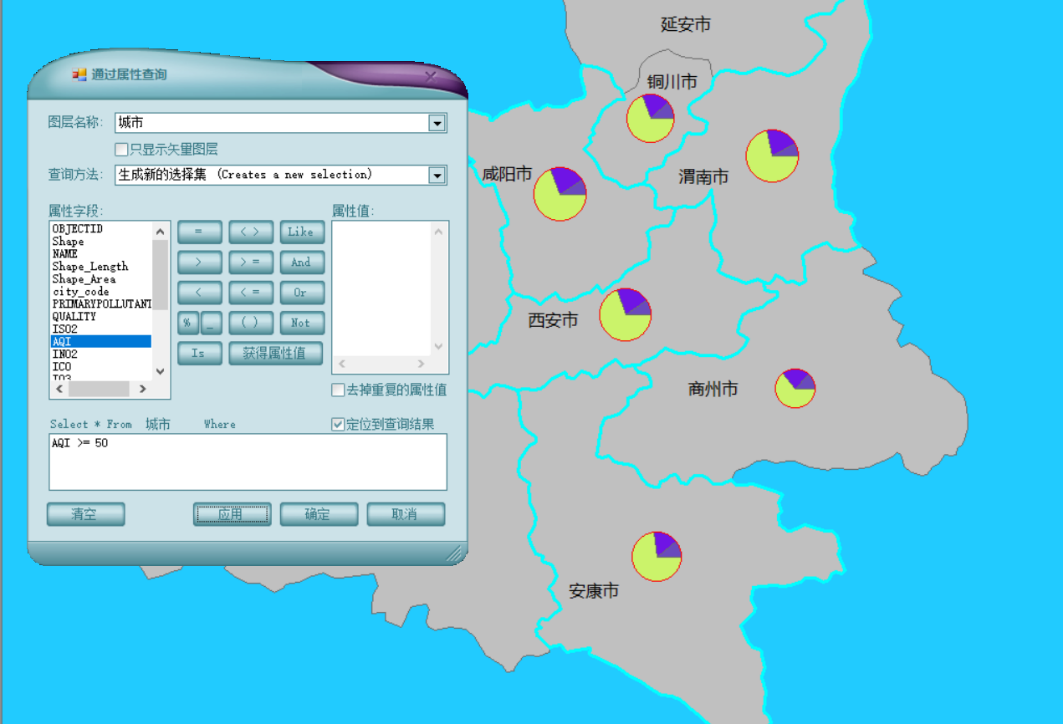


图20属性查询

此外，系统预期实现一些如AQI日报查询和空气质量预报表查询的功能，预期效果如下



图21 AQI历史信息查询预期效果图

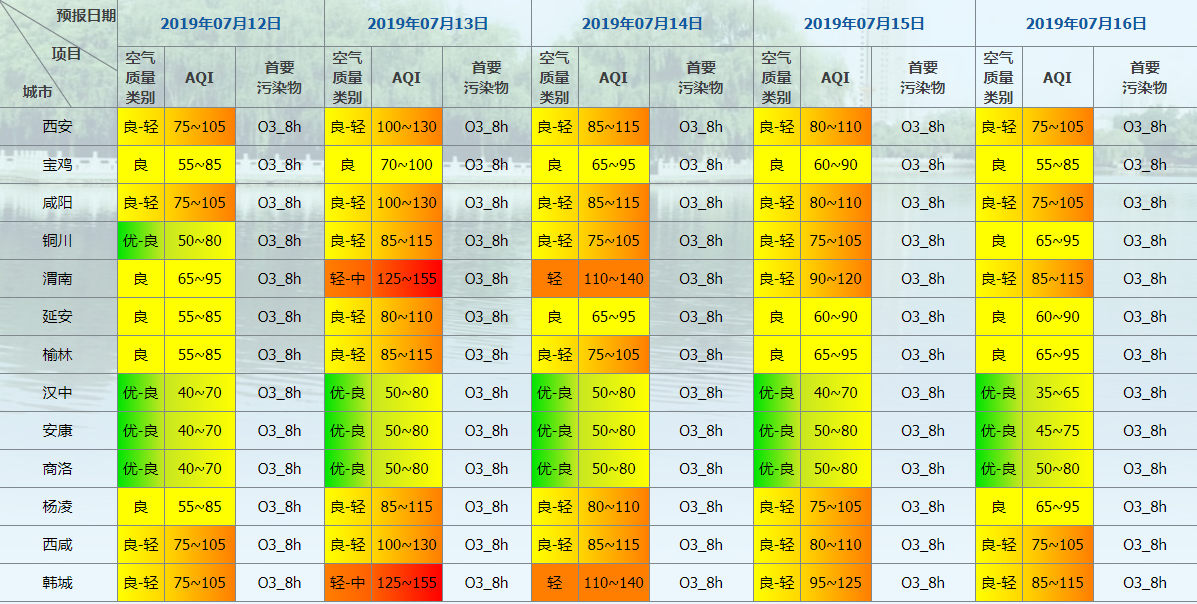


图22 AQI预测信息查询预期效果图

⑦气象预警功能：该功能的主要作用是从实时返回的气象数据中取得预警相关参数，如天气状况、空气质量状况等，如何综合这些数据向用户发送预警邮件。该功能目前并不完善，打算后期改用用户数量更多的微信进行通讯。界面和发送的邮件如图23，由于西安目前并无预警，因此采用相同接口下北京市某日的气象数据样例进行了测试。



图23 邮件预警功能

⑧系统管理功能：该功能主要是退出系统，同时我在该功能条下加入了获取源码的功能，通过该功能可以链接到GitHub获取源码，方便用户对系统进行进一步的拓展和维护。

⑨右侧展示功能：该功能主要实现了双缓存的时钟、空气质量预报展示、天气展示和图层管理功能。其中，天气展示主要展示的是实时的天气和各种与人类活动有关的指数。当然，也有最近几天的天气预报。

### 4.关键技术

①网络爬虫：任何系统的操作都必须基于数据，否则毫无价值，因此网络爬虫对于该系统而言十分重要。数据通过网络爬虫实时从数据源网页获取数据后对数据进行整理汇总，综合多方面数据完善气象系统。不过，这也是该系统的缺点，因为采用该方法获得的数据往往并不是真正的实时数据，而是已经发布的数据。并且，即使采用三线程数据抓取速度任然较慢，导致加载数据时间过长，尝试将数据爬取和数据采集分离，但试想为了保证实时性并不能这么做，故改成在数据加载时无法操作系统。

②MongoDB数据库技术：MongoDB数据库是连接数据采集模块和GUI的桥梁，因此是必不可少的一部分。

③基于C#的ArcGIS Engine技术：哪怕数据再这么完美如果不能通过可视化形式向人们传递信息那么都是空想。系统中ArcGIS Engine + C# 主要用于数据的可视化展示，通过此项技术将数据库内的数据可视化为契合人体视觉美感的、具有对比性的、高价值的信息。

④电子邮件通讯技术：该项技术是实现气象预警功能的核心，它使得该系统更具有实用价值。这也是我在参考C#编程从入门到精通是突发气象

## 三、作品亮点与不足

### 1.作品亮点

#### 1）实时数据获取

该系统通过网络爬虫实现，不仅解决了找数据难的问题，也在一定程度上保证了数据的时效性。并且该网络爬虫采用了多线程技术，加快了数据获取的速度。

#### 2）动态预警功能

动态预警设置的实现使得该系统具有更高的使用价值，对普通用户的意义更大。

#### 3）统计图表功能

统计制表功能使得数据更具有参考价值，可以直观的从统计图表中发现规律，数据的表现形式更加直观。

#### 4）专题制图功能

将数据以色差或图表的形式在地图上展示，可以结合生活常识和地理知识从数据中挖掘出更多的有价值的信息。

### 2.作品不足与解决

#### 1）作品功能尚未完善

目前系统只实现了少部分功能，仍然有大部分功能只是在预想状态，如果要完全实现，肯定还有比较多的坑踩，前面的代码也肯定还有许多的坑要填。等到时间充足的时候，可能会继续对系统进行更新。

#### 2）数据加载时间太长

在运行程序时先要执行一遍爬虫程序，但这个过程可能比较长（测试一般10~25秒），其次，使用现有的皮肤控件加载窗体时要全部重绘一次窗体速度较慢。其次，在数据加载方面，我们可以考虑把数据采集模块传到云服务器运行，这样省去了启动系统必须先运行爬虫的恶心步骤。

#### 3）界面任然不完善

目前右侧的界面在放大缩小时任然会产生尴尬，尤其是右上角的时钟，由于刚刚制作程序时使用Label控件来展现时间，但后面经过实践效果非常不令人满意。目前已经想到了解决方案，使用Button控件加容器，并在代码中设置字体高可以解决这一问题。其次，本系统主要是做系统的展示，并不具有太多的数据处理，因此使用HTML5+CSS+JS+ArcGIs API for JavaScript来实现系统会更加的美观、实用、大众化。

#### 4）必须安装相应数据库

使用文件分离形式的数据库可能更好，或者将MongoDB安置在云服务器可能会更好，这样可以省去用户安装配置数据库的麻烦。

# 第二部分、操作手册

**注：该操作手册只含已经实现的操作**

## 一、系统安装

下载系统安装文件Setup.exe按提示步骤进行即可完成安装。此外，系统已经打包好了虚拟Python环境，但仍然需要安装MongoDB才能运行。其次，要是有导出文件的EXCEL文件的操作必须将MongoDB中默认导出数据库的EXE文件加入到系统的环境变量才能运行，成功导出后中文会出现编码错误，这是因为导出的数据默认为UTF-8编码，而EXCEL默认打开CSV文件为GB2312编码，可以找到系统已经预置好的Python代码进行转码。当然，如果觉得麻烦，可以使用记事本或者Notepad++等软件打开也能看到你想要的数据。

下面是系统安装相关的一些图片。

1）安装文件的图标：



2）安装好后系统后安装路径文件夹：



3）安装好后的图标：



致歉：由于系统为默认窗体全屏，所以如果未安装相应数据库而使用该系统可能会导致您的电脑处于无法操作状态，但您可以按下win键调出菜单栏，然后通过菜单栏关闭系统。如果给您带来不便深感歉意。

## 二、系统使用

**1）打开系统：**

双击如下快捷方式图标，或找到运行的EXE文件双击运行系统。



**2）查看某地区气象数据：**

鼠标移动到想查看的地区并点击，此时地图上会出现气象数据的显示，同时右侧的气象数据显示被切换到你选择的地区。单机右侧文字标签查看你想要查看的信息

**3）基本操作：**

系统右上角工具栏具有地图放大、缩小、漫游、上一视图、下一视图、全图显示和默认鼠标键几个操作，点击后即可使用。

**4）导出数据：**

点击数据管理菜单栏→导出EXCEL文件，选择路径后单击确定即可。

**5）专题制图：**

点击专题制图菜单栏→选择需要的专题图类型，设置好相关参数单击确定即可。（注：单击应用只是相当于一种预览的效果，不会保存在地图图层上）

**6）按属性查询：**

单击辅助查询菜单栏，更具字段设置好查询语句，单击确定即可进行查询。

**7）自动预警：**

单击自动预警测试向作者发送预警邮件，在预警设置中输入邮件地址向邮件地址发送邮件。

**8）获取系统源码：**

单击系统管理菜单栏→选择获取源码即可转到作者GitHub地址下载源码。

**9）退出系统：**

单击单击系统管理菜单→选择退出系统即可退出系统，也可以单击右上角的关闭按钮

## 三、系统卸载

双击以下文件即可实现卸载操作：

