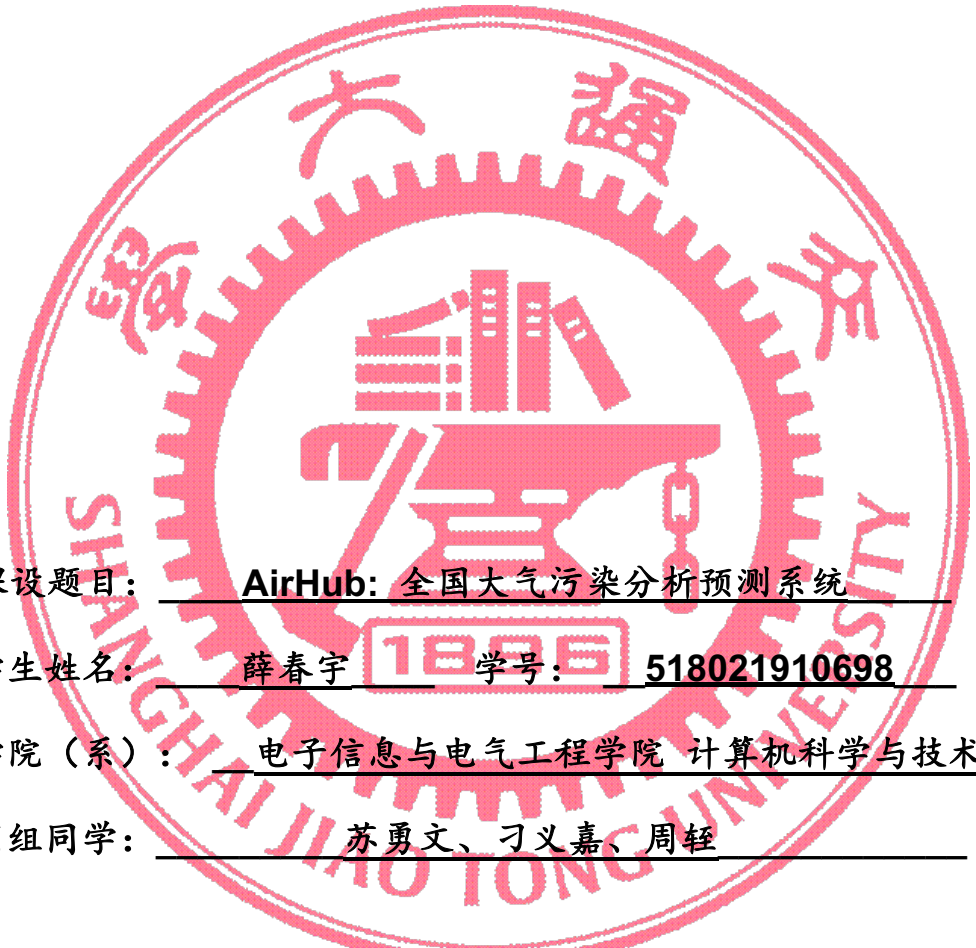
上海交通大學

# 《数据可视化与可视分析》课程设计论文



课设题目：AirHub: 全国大气污染分析预测系统

学生姓名：薛春宇　学号：518021910698

学院（系）：电子信息与电气工程学院 计算机科学与技术

同组同学：苏勇文、刁义嘉、周轻

2021 年 6 月

# AirHub: 全国大气污染分析预测系统

薛春宇 518021910698
电子信息与电气工程学院 F1803304

## 1. 问题描述及设计思想概述

### 1.1 问题描述

空气质量，是衡量空气污染程度的重要指标，其可以根据空气中污染物的浓度高低进行判断，清洁的空气对于改善我国人民的身体健康，创造宜居的环境至关重要。得益于我国日益完善的空气质量检测网络，空气质量的分析人员可以收集到大量具有高维、时序特点的空气质量数据。而如何通过大数据分析技术和可视化方法，有效利用这些数据以理解大气污染时空态势，分析大气污染物来源与扩散方式。数据可视化与可视分析将数据智能处理、视觉表征和交互分析有效结合，为大气污染防治工作的分析、指挥和决策提供有效手段和决策依据，对于防治大气污染具有重要意义。

### 1.2 设计思想概述

本团队借助 2013-2018 年中国高分辨率大气污染再分析开放数据集，构建了一个全国大气污染分析预测系统。该系统通过提取污染物的局部与全国范围内的分布特征，与气象指标结合，对污染物的时空分布趋势进行了可视化与一定时间范围内的预测。该系统的目标用户为空间质量分析监控人员及污染防控机构，可被用于指挥大厅或宣传科普大屏幕进行展示。

系统在宏观上展现了全国各地主要城市的空气污染情况，并在各个子系统中可交互地展示了地方空气质量相关新闻，污染程度排名，污染物和气象指标的时空变化趋势，以及污染物预测结果。为用户了解全国与各地空气质量提供了多维度的分析渠道，可以帮助当地政府深入了解污染物来源和变化趋势，从而制定更加合理高效的空气质量管理政策。

## 2. 个人工作

出于文体规范，下文仍以“我们”作为第一人称进行描述。

### 2.1 后端的搭建

本次项目的后端搭建由本人独立负责，主要可以分为**数据预处理**、**Flask 后端框架**、**LSTM 污染物预测模型**以及 **MySQL 数据库**，后端总代码量在 1500 行左右。

#### 2.1.1 数据预处理

基于 ChinaVis2021 提供的大气污染数据集，本项目借助百度地图 API 和 AQI 国际转换标准等工具，对原始数据集进行一系列的预处理及特征提取操作，进而构建基于城市的污染物数据库。注意，数据预处理全部使用自行编写的 Python 脚本进行实现。该数据库中包含如下数据类别：

(1) **城市与位置的映射关系**。由于原始数据集中仅存在经纬度信息，本项目通过调用百度地图 API 的逆地址编码功能，将经纬度信息逆向转化为行政地区信息，再经过筛选和清洗，得到 90 个主要城市从 2013 年到 2018 年的污染物及气象指标信息。该城市集合覆盖了每个省热度前三的城市，以及北京、上海等直辖市，能够较好地表征全国大气污染的情况。

(2) **城市中心位置**。为了在前端的主地图模块上进行城市位置的表示，城市中心位置的确定十分必要。基于(1)中提取完成的映射关系，我们遍历该映射哈希表，将属于某城市的经纬度全部线性相加并取平均，作为该城市的唯一位置。

(3) **城市特征向量**。我们通过将属于某城市的各项污染物及气象指标同类相加取平均，得到该城市长度为 11 的唯一特征向量，包括 6 个污染物指标和 5 个气象指标。在数据集覆盖的 2192 天中，每个城市在每天均对应一个用以描述该城市当天污染及气象情况的特征向量。

(4) **城市 AQI 指数**。AQI 指数的计算参见 2012 年出台的《环境空气质量指数(AQI)技术规定（试行）》中的相关方法，首先根据区间关系计算各污染物的 IAQI 指数，再取 6 种污染物中最大的 IAQI 作为该城市当天的 AQI 指标。每个城市每天对应唯一特定的 AQI 值。

(5) **城市 IAQI 指数**。将(4)中各污染物的 IAQI 指数保存为 6 维的向量，存储为该城市当天的污染物 IAQI 指标向量，以供前端进行更加细粒度的可视化分析。

### 2.1.2 Flask 后端框架

本项目使用 Python + Flask 框架的形式进行后端服务器的搭建，并使用 JSON 作为前后端交互的数据格式。

**前后端接口的设置**是本次项目里最为耗时的工作之一，在将前端可视化模块移植到本地的前后端框架中时，必须首先通过该模块开发者提供的所需数据的信息，确定后端应该提供给该模块的数据接口，再在前后端分别实现特定的函数来满足该接口的数据传输及处理要求，一个前后端接口的实现往往需要前后端上百行的代码。

同时，后端包含支持**数据库操作的相关接口**，以更加快速地实现前端对后台数据的查询，该部分的实现将在 2.1.3 节中详细给出。

### 2.1.3 MySQL 数据库

本项目使用关系性数据库管理系统 MySQL 来搭建后端数据库，该数据库支持数据的快速筛选、分类、排序和查询，能够很好地支持前端在可视化分析过程中对不同数据的快速获取。MySQL 本身仅支持命令行操作，借助 pymysql 工具包，可以直接在 Flask 后端框架下进行**基于 Python 脚本的 MySQL 数据库操作**，包括数据的创建、插入、删除和查询等。

注意，项目进行了**存入数据类别的选择**，将经过 2.1.1 节数据预处理与清洗后得到的几类数据条目存入数据库，而非直接将原始数据集不加筛选的悉数存入。上述操作能够显著提高数据访存的效率，以及前端模块的响应速度。

### 2.1.4 LSTM 污染预测模型

具有记忆选择性的 LSTM 长短期记忆神经网络则能够更好地学习到大气时序污染物数据中的有效特征，同时降低某些无意义信号在全局上的影响。在大气污染物时序数据中，某一天的污染指标不仅和之前几天的状态相关，还和未来几天的各项指标相联系，因此，本项目使用**双向长短期记忆神经网络 BLSTM**[1]来构建污染预测模型，并在重构造后的数据集上进行模型的训练和测试。网络有两层 BLSTM，三层 Dropout 和两层 Dense 组成，网络结构图见附录 A。

**数据集重构造**的具体方法为：在基于城市的污染物数据集的基础上，以城市为单位，将该城市对应的 2192 个特征向量按照 8:2 的比例划分为训练集和测试集，并顺序遍历这些特征向量，取每七天的污染物及气象指标作为一个训练样本的 features vector，第八天的污染物及气象指标作为 label。在基于上述数据集的充分调参和训练后，该模型能够在已知最近七天污染物及气象指标的前提下，预测未来第八天各项污染物指标的具体数值。

通过将预测结果加入到已知数据的方式，我们能够对未来几天进行**污染物指标的迭代预测**。然而，由于迭代会导致预测误差的逐渐积累，预测结果的准确性与预测天数成反比关系。因此在本项目中，为保证污染预测模块的可靠性，仅提供未来七天，即 2019 年 1 月 1 日到 7 日的预测结果。

## 2.2 前端的部分搭建

本次项目中，本人在后端之外同时负责了部分前端的搭建工作，包括**前端整体框架的实现**，**前端各模块移植后的适配**，**前端各模块联动的逻辑实现**，以及**总览模块及污染预测模块的实现**。

### 2.2.1 前端整体框架的实现

为构建美观、合理、高效的可视化分析系统，必须首先**对前端框架进行设计和实现**。基于上述思想，本项目参考大量已有的可视化系统，设计并实现了适配于“大气污染分析预测”主题的前端可视化框架，效果参见附录 B。特别地，将本作品的名称正式命名为“AirHub: 全国大气污染分析预测系统”，并为之设计了主界面 LOGO。

### 2.2.2 前端各模块移植后的适配

由于大部分前端模块交由其他组员完成，在将其成果移植到本地前后端框架中时，大概率会出现各式各样的错误，因此需要**对这些错误进行逐一解决**，以保证可视化系统能够正常运行。

例如，在将新闻板块的代码移植到系统中时，后端一直出现”undefined”的报错，但在我们将新闻模块代码中报出该错误的代码注释掉后，污染分析与预测板块却出现了新的 bug：当我们点击污染分析/污染预测按钮，再在主地图中进行地图缩放时，主地图视图中便会出现污染分析中的图标，见附录 C。出现上述的原因是新闻板块代码中的 option 为未被设置为局部变量，会影响到整个前端系统内的其他模块。再比如，当我们频繁在主地图中点击切换城市时，前端会报错“Error 500”，查阅资料后发现是与后端数据库的连接断开。经过充分的代码分析后，我们在后端建立数据库连接的操作中将每次运行后端只建立一次改成每次前端执行对数据库的查询操作均新建一次连接，上述问题得以解决，见附录 D。

在可视化前端的开发过程中，类似上述错误的过程不断发生，通过不断发现问题，分析问题，解决问题，我们得以开发出一套美观且可执行的系统。

### 2.2.3 前端各模块联动的逻辑实现

在完成前端各模块的设计后，我们的工作的重点是**如何建立各个模块在逻辑上的联动**。例如，当点击主地图中的城市时，其他模块的目标城市也要被修改成该城市，以形成可视化系统的逻辑完整性。对此，我们创造性地设置了包括当前城市、当前日期、当前污染物三个全局变量，并通过可视化办法直接展示在前端的总览模块中。若某个模块需要对全局变量进行设置，则该操作会触发其余模块对该全局变量进行读取，进而修改自己部分所显示的内容，实现图表之间的联动。

模块联动的实现对开发者的逻辑思维以及调试能力都有较高的要求，也需要大量时间和精力的投入，不断在使用的过程中发现错误并进行相应的修改，我也在这个过程中收获颇丰。

### 2.2.4 总览模块及污染预测模块的实现

出于本人负责的 LSTM 污染预测模型及各模块之间实现逻辑联动的需求，我们需要自行实现总览模块及污染预测模块的可视化，见附录 E。

总览模块中，前三个变量分别表示当前系统选择的城市、日期和污染物，第四个变量表示当前城市在该日期内的空气质量等级，基于 AQI 标准进行计算和分级。

污染预测模块中，用户既可以通过该模块内置的下拉框选择需要预测的污染物，也可以直接在全局污染物的下拉框中进行选择，这样的设计能够有效提高用户的使用体验。该模块会根据当前的城市、日期和选择的污染物，从后端数据库中读取已经提前存储的预测结果，并在前端进行显示。

## 3. 结论

本项目基于 2013-2018 年中国高分辨率大气污染再分析开放数据集，借助 echarts、百度地图 API、Flask、Tensorflow 等工具，成功构建了全国大气污染分析预测可视化系统。该系统兼具设计美观性和功能使用性，能够很好地完成大气污染监测分析和预测的任务，为用户多方面了解全国各地空气质量提供帮助，同时也可帮助各地政府深入解读污染物来源和变化趋势，从而制定更加合理高效的空气质量管理政策。

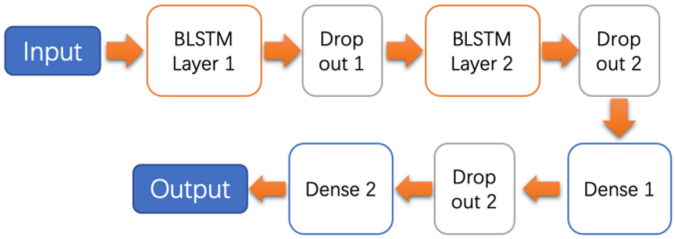
在未来大气污染分析预测的进一步研究过程中，本项目组会继续对当前系统存在的新闻模块显示错误，城市集合不够充分等问题进行改进和完善，并在系统界面和交互逻辑设计，和模型预测结构优化等潜在的研究方向进行深入挖掘，提高该可视化系统的设计美感和实用性。

## 4. 参考文献

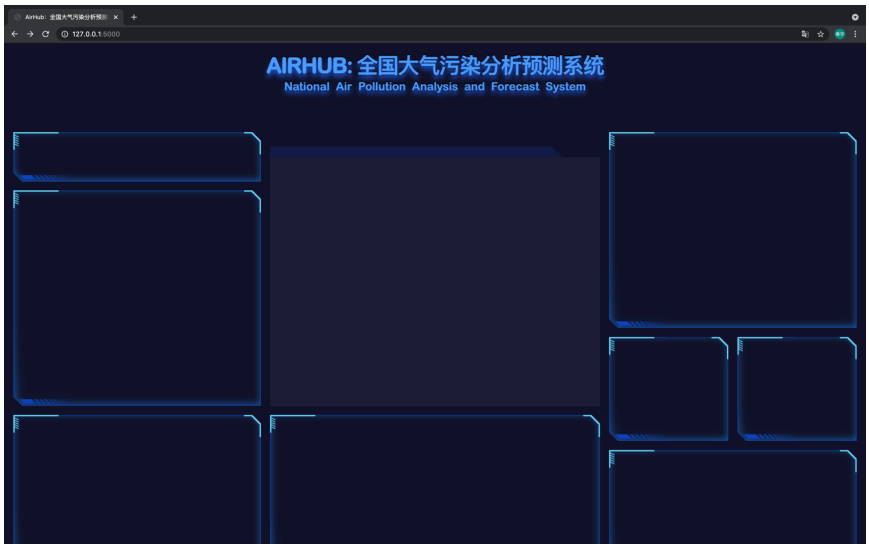
① S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

② M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

## 附录 A：LSTM 网络结构




## 附录 B：前端框架设计




## 附录 C：模块移植过程中的 bug

# 附录 D：后端数据库连接错误的日志记录

- 修复了数据库频繁访问导致的后端报错 `pymysql.err.InterfaceError: (0, '')` 以及前端报错 `Error 500` 数据库连接断开，原因是在本来的方案里，在 `app.py` 运行的过程中仅在最开始建立数据库连接，在 `app.py` 运行结束后断开连接，即主函数如下：

```python
if __name__ == '__main__':
    # Establish database connection
    db_connection = pymysql.connect(**config)
    # Create cursor
    cursor = db_connection.cursor()

    # Run server
    app.run()

    # Close database connection
    db_connection.close()
```

现在修改为在每一次访问数据库的时候都重新建立连接，在该次访问结束后关闭该连接。

# 附录 E：总览模块及污染预测模块的可视化实现