校名-小

# 大学生创新训练项目 立项申请书

项目来源：导师科研类☑ 自主探索类□

校企合作类□ 滚动支持类□

科研院所合作类□ 校际合作类□

项目名称： 基于大数据技术的企业无组织排放

行为算法的研究与实施

项目名称（英文）：Research and Implementation of Enterprise’s Fugitive Emission Behavior Algorithm Based on Big Data Technology

项目依托学院： 计算机学院（国家示范性软件学院）

项目负责人： 梁国燊

联系电话： 19801192780

E-mail： 2938978311@bupt.edu.cn

指导教师： 孙艺

E-mail： sunyisse@bupt.edu.cn

起止年月： 2021年3月1日—2022年3月1日

填报时间： 2021年 6月 17日

**填写说明**

1. 本申请书所列各项内容均须实事求是，认真填写，表达明确严谨，简明扼要。
2. 申请书为大16开本（A4），在网上下载后，根据填报项目类别可自行删减和加页，但须保持格式和内容与原件一致。
3. 第五部分“推荐、评价及审批意见”不用填写。

**一、基本情况**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **项目名称** | 基于大数据技术的企业无组织排放行为算法的研究与实施 | | | | | | | |
| **项目负责人** | 梁国燊 | | **学号** | 2019212920 | **所在学院** | 理学院 | **手机号** | 19801192780 |
| **专业** | 应用物理 | **班级** | 2019214104 | **邮箱** | 2938978311@qq.com |
| **指导教师** | 孙艺 | | **职称** | 高级工程师 | **所在学院** | 计算机学院 | **手机号** | 13910988583 |
| **邮箱** | sunyisse@bupt.edu.cn |
| **项目类别**  **（类别说明见立项指南）** | **□**智能硬件**□**社交媒体**□**数字娱乐**□**通信网络**□**医疗健康**□**公共服务**□**电子商务**□**教育文化**□**房产家居**□**理论研究**□**机器人**□**无人机**☑**智能制造**□**智能交通  **□**创意设计**□**其他**\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | | | | |
| **检索关键词** | **大气污染监控；无组织排放；空气扩散模型；源强核算；大数据分析；分布式计算；Hadoop；Web前后端开发；Django；数据可视化** | | | | | | | |
| **项目成员**  **基本信息** | **姓名** | **学院** | | **专业** | **班级** | **学号** | **电话** | **邮箱** |
| 梁国燊 | 理学院 | | 应用物理 | 2019214104 | 2019212920 | 19801192780 | 2938978311@qq.com |
| 唐良爽 | 计算机 | | 软件工程 | 2019211501 | 2019210010 | 18190326484 | 2809867884@qq.com |
| 胡逸同 | 国际学院 | | 电子商务及法律 | 2020215111 | 2020213350 | 13661192968 | huyt@bupt.edu.cn |
| 谢滨竹 | 国际学院 | | 电子商务及法律 | 2020215111 | 2020213356 | 15656533660 | Xbz\_nicous@bupt.edu.cn |
| **团队主要成**  **员介绍** | 梁国燊 应用物理专业，对项目中空气污染传播物理场景的理解有专业基础。大一、大二期间分别参与雏燕计划、大创项目都顺利结项。熟悉python,java等编程语言，有机器学习、深度学习开发经验，进行过网页前后端的部署，有数据结构、计算机网络等计算机课程经历。  唐良爽 高中时期获奖情况：第35届全国物理竞赛省级二等奖 。就读于软件工程专业，熟悉Java、Python以及Linux环境下的编程和部署，对WEB前端也有一定了解，有算法与数据结构、操作系统、数据库等课程经历。  胡逸同 高中时期获有：全国青少年科创大赛二等奖，全国物理竞赛二等奖。重视数理基础，求知欲强，有创新思维，动手实践能力强。熟悉计算机领域的基础概念(计算机基础90+)，掌握C, Python, Java基础，有机器学习经历(GANs)，对计算机网络和云计算有较多的探索。（在阿里云服务器上实现了家、校、云之间P2P异地组网）  谢滨竹 就读于电子商务及法律专业。高中时期获得：全国信息学竞赛二等奖，SPSC二等奖。大一获得校程序设计竞赛银奖。熟悉C语言，掌握Python、Java基础。 | | | | | | | |
| **指导教师承担科研课题情况** | 承担河北省科学技术厅重点项目，要求确定无组织排放企业。  1、分析企业的大量数据，利用相关算法区分企业排放是否符合标准  2、领用气体扩散模型溯源，判断企业的污染源，并展示经纬度等位置信息  3、构建大数据算法库，实现数据的分析，导入以及各种直观的视图 | | | | | | | |
| **指导教师对本项目支持情况** | * + - 1. 每周1-2次的项目指导       2. 每周的项目例会       3. 阶段检查       4. 相关知识或课程讲解       5. 一定的经费支持 | | | | | | | |

**二、立项依据**

**（一）项目创意来历及项目意义**

1. **项目构思来历、意义及目的：**

化工企业无组织排放严重威胁着人民群众的身体健康，且不满足我国社会经济的可持续发展。

《重点区域大气污染防治“十二五”规划》中指出，采用区域大气污染联防联控的工作机制，通过污染在线监测设备及网络的配套建设，环保部门已实现对上万个污染源的实时监控。然而，中国化工企业具有集聚性强，排污节点多，无组织排放量大，废气组分复杂、间歇排放等特点，对污染源的追溯和定位构成较大挑战，且传统方法只能检测企业排放量。

所以有必要开发一款利用大数据、人工智能技术预测预警、分析趋势、分析企业的污染贡献以及检测企业排放量是否合标、精准定位企业污染源并直观展示企业排放详情信息的系统，以供环保部门使用，实现对无组织排放的细颗粒度检测、管控，进而保障人民群众的身体健康，促进社会经济的可持续发展。

1. **国内外研究现状和发展动态：**

目前，国内化工企业排放检测的研究主要集中在有组织超低排放治理技术，而无组织排放检测是共性难题，大多数对化工企业无组织排放检测的研究是基于厂界检测方法，缺乏对污染排放的智能化研究。而国外，如美国，德国和日本等国家，有一套成熟的排放标准，如BAT体系等，并配有智能化的监控、检测、预测预警系统，对利用大数据和物联网技术进行无组织颗粒物排放的研究也领先国内。

1. **同类项目分析：**

同类项目识别了单一工厂无组织废气的主要来源，监测挥发性污染物浓度、排放特点及其中主要化合物等无组织废气的源项特征，评价了各产生源的主要贡献率，识别了不同无组织废气排放源中的主要挥发性大气污染物。但并未实现多工厂环境下对无组织排放工厂的识别。

1. **需求分析：**

开发一款利用大数据、人工智能技术预测预警、分析趋势、分析企业的污染贡献以及检测企业排放量是否合标、精准定位企业污染源并直观展示企业排放详情信息的系统。

1. **项目研究主要内容**
2. **项目主要内容简介**

目前国内外企业和研究者尚未对利用现代计算机技术对企业无组织排放的溯源和预测分析进行研究，而日益增多的无组织排放严重危害着环境，加上国内化工企业集聚性强，排污节点多，无组织排放量大，废气组分复杂、间歇排放等特点，使得传统的实地数据采集后再进行监测的方法变得繁琐，而风向、温度等自然因素对精准定位污染源带来了及大的干扰，用传统方法精准定位污染源有着很大难度，这些因素让传统方法已经满足不了对众多企业无组织排放监管的需求。所以利用现代计算机技术对无组织排放的监管成为了急需解决的问题。本项目立足于此，目的是利用大数据和人工智能技术对企业无组织排放进行精准溯源，对企业排放进行智能预测和分析，并对企业排放行为进行智能识别，从而极大地减轻了环保部门对企业无组织排放监管的压力，减少了监管的人力和财力开销。最后再并结合WEB前后端技术对监控信息进行集成展示，并提供对企业以供政府环保部门和人民群众参考，使得信息更加透明化。

1. **概述项目研究方法，要解决的问题**

采用的调查法主要是通过对多家化工企业实地调查，并记录每家化工企业排放气体中的主要污染成分。另外在化工企业厂区之外选择合适的采样点，在样本点采样收集污染气体的成分、浓度以及风向，并记录下与该采样点距离最近的几个化工厂，从而得到本项目成果系统的数据来源，系统对这些数据进行分析后得出相应结果，并作为第一批源数据最终展示到本系统的展示界面中。

通过文献研究法，收集并整理各个化工行业的具体排放标准，定义无组织排放行为判定的阈值；搜集并学习环境科学领域对于排放源识别的专业方法，形成对待解决问题的总体认识，进而选取适合的数学方法建模，提高模型的适应性；

本项目采用的信息研究法主要体现在利用大数据技术和深度学习相结合的方法，使系统从采集到的的大量企业无组织排放数据中自动学习到有效的特征表示，而后经过多层的特征转换，使得原始数据变成更高层、更抽象的表示，并不断用已有的样本作为训练集和测试集，使系统的预测模型更加准确。从而打造一款能够精准地识别、分析和预测企业无组织排放行为的信息系统。

1. **项目创新点与项目特色**
2. **行为识别**

以两个监控点为系统为例，当检测到相关数据之后，不可通过检测时刻的数据进行判断。需要通过风向、相对位置、行为习惯等各项数据通过概率论相关原理将多种可能组合的概率算出来，并取可能性最高的一种可能组合做为结果。

1. **展示平台**

本项目的所有检测数据标准，均采用国家白皮书最新标准，并放在WEB相应板块进行展示。本项目的行为识别算法也将进行公示，确保每个企业对其知晓并认可，对结果以及处罚无异议。

1. **分析报告**

本项目不仅提供完善的算法，可靠的结果报告，更会为每个企业和检测点生成完善的分析报告，这些报告不仅会为各企业的今后排放提供蓝图，更会做为之后的行为识别算法的行为习惯，不断自我迭代。

1. **系统方案和技术路线**
2. **设计思路**
   1. **标准确定**

我们把中国人民共和国生态环境部目前实施的标准：GB 16297-1993《大气污染物综合排放标准》、GB 3095-2012《环境空气质量标准》以及一系列行业污染排放作为污染判断标准录入到系统中，作为阈值判断是否存在无组织排放行为。

* 1. **污染识别**

我们把检测站的数据录入系统，进行数据预处理（数据清理、数据集成、数据归约与数据转换），然后对数据进行遍历，与录入的标准进行逐一对比，判断数据是否超标，对超标数据进行标识，并录入到指定的关系表中，用于进一步的溯源判断。

* 1. **污染溯源**

对污染场景进行数学建模，引入气体扩散模型，构建一个能够精准识别复杂场景污染源的模型。

* 1. **平台搭建**

搭建一个集成展示数据，多角色用户共享的web平台，以高度可视化的形式，向用户直观展示分析成果。

1. **技术关键**
   1. **场景模型建立**

项目初期，我们做出以下假设以简化我们的模型构建：

（1）工厂位于监测点。

（2）检测时间段风力，风向，温度，湿度恒定。

（3）释放的气体的密度与空气相差不多(不考虑重力或浮力的作用)，且气体扩散过程中没有发生化学反应。

（4）气体的传播服从扩散定律，即单位时间通过单位法向面积的流量与它的浓度梯度成正比。

首先考虑只有两家工厂的情况，如果检测点存在污染现象，那么就有以下三种情形：

（1）污染是其排放的。

（2）污染是另一家工厂排放的。

（3）污染是两家工厂共同排放的。

通过引入检测点坐标，构建网络图，通过图论的相关模型与算法，推广到多工厂下的复杂图网络情形。

* 1. **数据处理相关算法**

由于工厂基数大，检测点采样频率高，污染气体种类多，导致系统需要处理海量数据，通过引入Hadoop分布式大数据处理系统，结合Spark，和MapReduce，利用机器学习中的极大似然估计和最大后验概率估计，以及朴素贝叶斯分类算法和隐马尔可夫模型，实现离线数据分析和计算，并不断训练、优化上述的场景模型，从而得出可疑工厂对污染数据的贡献度，从而实现精准溯源。

* 1. **平台搭建技术**

本项目具有高度可视化的特点，采用前后端分离的方式，构建一款web系统，web前端使用流行的vue框架，后端使用Django框架+NoSQL数据库的模式，通过在后端制定与前端的通信接口，实现数据的传输、展示和存储。

1. **系统模块图与功能概述（子模块功能）**

电脑萤幕

中度可信度描述已自动生成

本项目的最终成果是一个Web平台，在前端界面部分按照用户群体分成了三个模块，分别是政府模块、企业模块以及网站管理员模块。政府部门的权限包括查询各个企业的排放数据、查看平台生成的企业排放数据的分析报告，以及更新各项污染气体的排放标准；企业的权限包括查看企业的信息，这些信息主要分为企业的基本信息，如地理坐标、主要排污气体等，也包括该企业的历史排放记录，如某年某月某日该企业排放气体中某项或多项指标超标等信息。企业也有申诉权限，可以在网站的指定申诉入口进行申诉。网站管理员除了运维网站外，还可以导入和更新标准，且具有管理政府用户、企业用户的权限；平台的大数据分析与处理部分在后端进行，重要的功能模块包括数据分析、离线计算、文件存储以及数据库管理。本平台通过使用大数据环境下的机器学习框架Mahout以及Spark MLlib，构建出能够精准溯源、智能分析预测企业排放行为的数据模型，从而实现本平台的核心功能；通过数据挖掘技术，高度自动化地分析企业排放数据，并做出归纳性的推理，从中挖掘出潜在的模式，使得构建的数据模型的决策能力更加强大；离线计算部分采用的是传统的Hadoop + MapReduce + Spark的技术，从而能够长时间保存巨大的数据量，并在大量数据上进行高效的复杂批量计算，同时也能很方便地查询批量计算的结果；文件存储主要采用HDFS分布式文件系统，辅以Tachyon和KFS；数据库方面采用MongoDB结合NoSQL的方式进行数据存储。

1. **技术架构**

我们的平台采用分层式开发，技术栈如图所示：

图形用户界面, 表格

中度可信度描述已自动生成

1. **技术亮点**
   1. **无组织排放行为的识别**

无组织排放行为难以监控，异常数据出现难以溯源。为此我们结合空气传播模型，复杂系统相关理论对污染场景进行了数学模型的构建，提出了污染源存在的多种情况，并结合大数据分析他们的可能性，提供了一种判断并溯源无组织排放行为的方案。

* 1. **网页平台的搭建**

使用了Django，基于Python的Web 应用框架，Python 语言可用的库和包非常多，开发应用程序非常迅速，这些特点使得Django可以快速进行Web 应用系统的开发。另外，微服务框架的应用也使系统具有低耦合度，大大提高了开发的灵活度。

* 1. **大数据处理：**

由于污染监控站点众多，采集频率较高，我们需要处理的数据量十分庞大，对算力的要求很高，我们便引入了Hadoop计算平台进行大数据的离线计算与分析。Hadoop计算平台采用分布式计算框架，可以在大量廉价的硬件设备组成的集群上运行应用程序，构建一个具有高可靠性和良好扩展性的并行分布式系统，很好的解决了大数据的分析的算力需求问题。其中的MapReduce技术采用了一种“分而治之”思想，把结构复杂，体量大的数据处理任务分发给若干个子节点，这些子节点由一个主节点进行管理，最后把子节点的处理结果进行汇总处理，得出最后的结果。节点的水平扩展可以很方便的增强数据处理能力，非常适合大数据

我们的用户层面向的是H5用户，页面使用H5/PUG、CSS/Stylus，使用前端框架: Vue.js , 其中包括Router, Axios等的使用。考虑到数据量大，用户群体广泛，使用LVM、Nginx技术实现负载均衡。网页后端使用Django框架进行搭建。由于数据量庞大，我们搭建Hadoop计算平台进行大数据的处理。其中考虑到Hadoop及Django的兼容性，选择使用MongoDB, NoSQL数据库进行数据的存储。文件存储使用Hadoop HDFS, Tachyon及KFS技术；离线计算方面我们使用Hadoop系统中的MapReduce, Spark技术，再通过数据挖掘、机器学习技术——Mahout，Spark MLlib，进行大数据的分析。考虑到开方的灵活性，维护更新的便捷性，我们使用开发架构使用微服务架构，微服务层包括了前后端的各种服务以及系统服务。

处理。其中机器学习我们使用了Spark MLIib, 该工具提供了Python的接口，我们可以在Tensorflow里进行嵌入。

整体而言，项目把现实中的无组织排放场景进行数学建模，再与大数据分析技术结合，最后通过web平台的开发把数据及分析结果进行展现，把各项技术的优点进行了很好的融合，是计算机技术在空气污染防治方面的应用与创新。

1. **使用的软件和开发测试环境**

本项目使用的软件包括CentOS8操作系统下的Hadoop3.2+Spark+MapReduce，以及集成开发环境Eclipse和Pycharm开发和测试环境都是在CentOS8操作系统下进行。

1. **开发难度**

本项目的开发难度较大，除了大数据环境下大量数据批量查询处理以及存储和备份等基础操作外，还要在大数据环境下使用机器学习算法构建精准溯源、分析预测的数据模型，并处理好数据集成、大规模资源描述框架处理和数据质量评定等高难度问题。同时为了搭建WEB平台，还需要使用Vue + Django的技术栈进行前后端的开发，并在后端框架中操作MongoDB以读取分析处理好的数据并传送给前端。由于数据量大，还需要考虑并发情况下的一系列问题。总体而言，本项目属于高难度开发项目。

**（五）项目进度安排**

项目立项于2021年3月，预计项目周期为一年时间。

立项伊始，导师向我们介绍了项目背景、预期目标及实现路线。4月到5月两个月我们通过分析已有数据构建项目完成所需的方法并构建知识体系，熟悉大数据Hadoop系统环境，学习Python，Java等语言基础，学习JavaScript、CSS等前端语言基础。

中期安排为6月到11月，通过每周一次小组例会和导师指导会逐渐完善项目。一方面构建Web前端和后端内容，另一方面搜集论文并自助讨论形成完善的算法体系。

后期安排为11月到12月，借助已有数据并招募测试人员进行试运行，通过调查测试人员的体验度，数据结果的可靠度评估项目完成情况，并对内容进行改进。在此过程过整合并申请相关专利内容，同时发表《基于大数据技术的企业无组织排放行为算法的研究与实施》论文。

1. **已有基础**
2. 有真实数据
3. 有依据的实验模型
4. 有了一年省级项目的研究基础
5. 有省级项目的支持
6. 已经在知网、万方、国家标准化局等权威机构查阅相关文章：

《基于Hadoop和Django的大数据可视化分析Web系统》

《火电厂污染物排放异常数据检测》

《污染气体排放标准白皮书》等

1. 通过对相关领域的专家、教授咨询。
2. 已经具备的研究条件包括但不限于检测点数据、完善的硬件条件、云平台。
3. **预期成果形式，达到的技术指标**

本项目的最终成果是一个Web平台，同时发表《基于大数据技术的企业无组织排放行为算法的研究与实施》论文，并申请相关软件专利。

三、经费概算（单位：元）

|  |
| --- |
| （一）项目总经费： 94500 |
| （二）经费分项预算：**需附《北京邮电大学大学生创新训练项目经费预算表》** |

四、成员分工

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 项目前期调研学习情况 | 任务分工 | 预期投入精力 | 签字 |
| 梁国燊 | 搜集相关论文、了解开发工具 | 后端开发，负责大数据分析、机器学习部分的开发 | 教学周每周投入4到5小时，假期每天投入2到3小时 |  |
| 唐良爽 | 搜集相关论文、了解技术路线 | 后端代码编写、Linux运维 | 全程 |  |
| 胡逸同 | 搜集相关论文、了解阿里云产品 | 前端开发、部署开发环境 | 全程 |  |
| 谢滨竹 | 搜集相关论文、绘制图表 | 前端开发、文档撰写 | 全程 |  |
|  |  |  |  |  |

注：包括项目负责人

五、推荐、评价及审批意见

|  |
| --- |
| **指导教师意见与承诺：**  **导师签字：**  **年 月 日** |
| **学院推荐意见：**  **负责人签字：**  **学院公章 年 月 日** |
| **专家组评价意见：**  **组长签字：**  **年 月 日** |
| **学校审批意见：**  **负责人签字：**  **盖章 年 月 日** |

**附表：**

# 北京邮电大学 大学生创新训练项目经费预算表

项目所在学院：计算机学院（国家示范性软件学院） 日期：2021.6.17

|  |  |  |  |
| --- | --- | --- | --- |
| **支出项目类别** | **支出项目说明** | **支出金额（元）** | **测算依据** |
| 1、业务费 | 打印费、复印费、装订费、书费、资料费等费用 | 500 | 书籍、论文资料的购买等 |
| 2、仪器设备购置费 | 购置或试制专用仪器设备，对现有仪器设备进行升级改造等费用 | 20000 | 摄像头、水质/土壤/空气检测仪等 |
| 3、材料费 | 芯片、模块、元器件、电路板等低值易耗品费用 | 20000 | 芯片、内存卡、云服务器、云数据库、文件存储、域名、SSL证书报告、论文打印 |
| 4、外协费 | 支付给外单位的检验、测试、化验、维修、租赁和加工制作等费用 | 8000 | 产品的加工包装，硬件测试 |
| 5、差旅费 | 开展科学实验（试验）、科学考察、项目调研、学术交流等所发生的外埠差旅费 | 8000 | 调研、测试不同环境下系统的拓展应用 |
| 6、会议费 | 学术研讨、咨询、培训等费用 | 8000 | 学术研讨等 |
| 7、专项业务费 | 版面费、专利申请及其他知识产权事务等费用 | 30000 | 论文、版面费、专利申请等 |
| **合计**（元） | 94500 | | |

**图片包含 猫, 躺, 床, 房间

描述已自动生成**项目指导教师签字： 项目负责人签字：



项目组其他成员签字：

