.

**硕 士 学 位 论 文**

MASTER’S DISSERTATION

论文题目 关于京津冀地区空气污染物关键传播

路径和脆弱点挖掘方法研究

作者姓名吴瑶

学科专业计算机科学与技术

指导教师黄国言教授

**2018年6月**

中图分类号：xxx 学校代码：10216

UDC：xxx 密级：公开

**工学硕士学位论文**

关于京津冀地区气污染物关键传播路径和

脆弱点挖掘方法研究

|  |  |  |
| --- | --- | --- |
| 硕士研究生 | ： | 吴瑶 |
| 导师 | ： | 黄国言教授 |
| 申请学位 | ： | 工学硕士 |
| 学科专业 | ： | 计算机科学与技术 |
| 所属学院 | ： | 信息科学与工程学院 |
| 答 辩 日 期 | ： | □□□□年□月 |
| 授予学位单位 | ： | 燕山大学 |

A Dissertation in Mechanical Manufacturing and Automation

（Times New Roman 字体4号）

Bio-syncretic Rehabilitation Mechanism theory and Application

（Times New Roman 2号字加粗，题目太长时可用小2号字）

by □□□

（Times New Roman字体4号）

Supervisor: Professor□□□

（Times New Roman字体4号）

**Yanshan University**

（Times New Roman 4号字 加粗）

June,2009

（Times New Roman字体4号）

燕山大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《 》，是本人在导师指导下，在燕山大学攻读硕士学位期间独立进行研究工作所取得的成果。论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签字： 日期： 年 月 日

燕山大学硕士学位论文使用授权书

《 》系本人在燕山大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归燕山大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解燕山大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权燕山大学，可以采用影印、缩印或其它复制手段保存论文，可以公布论文的全部或部分内容。

保密□，在 年解密后适用本授权书。

本学位论文属于

不保密□。

(请在以上相应方框内打“√”)

作者签名： 日期： 年 月 日

导师签名： 日期： 年 月 日

摘 要

随着我国经济的的快速发展，中国大部分地区的空气污染状况日趋严重。严重的空气污染会引发一系列的疾病，因此城市空气质量对保护人民群众身体健康，提高人民生活质量，起着关键的作用。但是，由于我国地域辽阔，无法全方位的覆盖空气质量监测站，因此导致了不能及时并全面的反映当前的空气质量状况，以及不能有效的对空气污染做出有效的防治。于是，本文提出一种基于基于污染物传播机理的污染物传播模型，在此基础上挖掘出实时污染物传播路径和一个周期内污染物的关键传播路径，从而以此为依据来决定监测站的选址以及有效的防治大气污染物的传播。

首先，由于无法获取污染源的详细信息，因此无法采用传统的基于污染源的扩散模型来挖掘传播路径。于是，本文充分的分析了影响污染物扩散的因素和机理，提出了一种基于污染物传播代价的传播模型，进而基于此传播模型挖掘出实时的污染物传播路径。

其次，通过每一时刻的污染物传播路径构建成污染物传播网络图，经过分析发现该网络图具有复杂网络的特性，因此本文采用复杂网络的度量方法来分析污染物传播网络的特性。

再次，本文基于污染物传播网络提出一种图矩阵的方法，来存储每条边在每个子图中出现的情况，并通过生成污染物传播网络子图、构建有向图矩阵、构建简化图矩阵来挖掘污染物关键传播路径。

最后，本文提出一种大气污染物传播网络脆弱点判断的算法—NodeRank算法，该算法不仅考虑了站点之间的传播方向以及污染物的传播代价，同时对大气污染物传播网络中节点的转移概率进行了改进。应用NodeRank算法对大气污染物传播网络中的节点进行评分排序，从而挖掘出大气污染物传播网络中的脆弱点。

关键词：大气污染物传播；复杂网络；关键路径；图矩阵；PageRank；脆弱点

Abstract

With the rapid development of China's economy, the air pollution in most parts of China is becoming more and more serious. Severe air pollution will cause a series of diseases. Therefore, urban air quality plays a key role in protecting people's health and improving people's quality of life. However, due to the vast territory of China, it cannot cover the air quality monitoring station in an all-round way, so it can not reflect the current air quality in time and comprehensively, and can not effectively prevent and control air pollution effectively. So, this paper proposes a propagation model based on pollutant pollutant propagation mechanism based on the basis of digging out the key pollutants pollutants real-time propagation path propagation path and a period of time, and as a basis to determine the location of monitoring stations and the dissemination of prevention and treatment of air pollutants effectively.

First, because of the inability to obtain the detailed information of the source of pollution, it is impossible to use the traditional diffusion model based on the source of pollution to mine the propagation path. Therefore, this paper analyzes the factors and mechanisms that affect the diffusion of pollutants, and proposes a propagation model based on the cost of pollutant propagation, and then extracts the real-time pollutant transmission path based on this propagation model.

Secondly, a pollutant transmission network map is constructed through the pollutant propagation path at each time point. After analysis, it is found that the network graph has the characteristics of complex network. Therefore, this paper adopts the measurement method of complex network to analyze the characteristics of pollutant propagation network.

Again, this paper presents a new method of communication network based on graph matrix of pollutants to store, each edge in each sub graph of the situation, and through the communication network graph generated pollutants and construct a directed graph matrix, constructing simplified graph matrix to excavate the propagation path of key pollutants.

Finally, this paper proposes a network vulnerability of air pollutants spread judgment algorithm - NodeRank algorithm, this algorithm not only considers the propagation direction between the site and the spread of pollutants cost, at the same time the transfer probability of nodes in the network transmission of air pollutants has been improved. The NodeRank algorithm is used to rank the nodes in the air pollutant transmission network, so as to excavate the fragile point in the air pollutant transmission network.

**Keywords:** air pollutant transmission; complex network; critical path; graph matrix; PageRank; fragile point

目 录

摘 要 I

Abstract II

目 录 IV

第1章 绪 论 1

1.1 课题背景及研究的目的和意义 1

1.2 国内外研究现状 2

1.3 论文研究内容 5

1.4 论文的组织结构 6

第2章 相关理论和基础知识 8

2.1 基本概念 8

2.1.1 空气质量指数 8

2.1.2 污染物传播机理 10

2.2 复杂网络特性 12

2.3 PageRank算法介绍 13

2.4 本章小节 14

第3章 大气污染物传播建模 15

3.1 数据预处理 15

3.1.1 监测站之间的距离 15

3.1.2 监测站方向与风向的夹角 15

3.1.3 监测站之间污染物的相关性 16

3.1.4 数据标准化 16

3.2 污染物传播模型构建 16

3.3 实时关键传播路径挖掘 17

3.3.1 传播约束条件分析 17

3.3.2 实时关键传播路径挖掘算法 19

3.3.3 算法实例 21

3.4 实验结果与分析 23

3.4.1 实验环境配置 23

3.4.2 实验数据 23

3.4.3 实验结果分析 25

3.4.4 实验结果验证 28

3.5 总结 30

第4章 基于图矩阵的关键传播路径挖掘 32

4.1 引言 32

4.2 构建大气污染物传播网络 32

4.2.1 基本概念与定义 33

4.3 基于图矩阵的污染物关键传播路径挖掘 34

4.3.1 生成污染物传播网络子图 35

4.3.2 构建有向图矩阵 38

4.3.3 构建简化图矩阵 39

4.3.4 挖掘污染物关键传播路径 40

4.3.5 实例说明 42

4.4 实验结果分析 46

4.4.1 实验环境配置 46

4.4.2 实验数据 46

4.4.3 实验结果分析 46

4.5本章小结 52

第5章 基于NodeRank的污染物传播网络脆弱点挖掘 53

5.1 引言 53

5.2 基于的污染物传播网络脆弱点挖掘 54

5.2.1 PageRank算法理论 54

5.2.2 PageRank算法改进 54

5.2.3 NodeRank算法实现 56

5.3 实验结果与分析 57

5.4 本章小结 60

结 论 61

参考文献 63

攻读博士学位期间承担的科研任务与主要成果 64

致 谢 65

第1章 绪 论

## 课题背景及研究的目的和意义

### 1.1.1 课题研究背景

近年来，我国快速的经济发展、日益增长的城市化使我国大部分地区出现了严重的空气污染问题。大气污染对人类健康有着负面影响，而且严重影响了经济和社会的可持续发展，因此防治空气污染成为中国社会的一个重要而亟待解决的问题。

空气污染就像个“隐形杀手”。世界卫生组织估计每年有700万人死于与空气污染有关的疾病，仅在美国，由于暴露在发电厂排放的微小颗粒环境中，每年就有7500至5.2万人死亡。空气污染也是肺癌的潜在病因，有大量的证据表明，它也可能增加了人类患上心脏病和中风的概率。最新发布的2018年环境绩效指数(EPI)显示，空气污染已经成为公众健康最大的环境威胁。

近些年来，京津冀地区空气质量问题日趋严重，曾多次出现污染物浓度严重超标的情况，严重的影响了人们的身体将康和严重扰乱了人们的生活规律。已有研究结果表明，北京地区与周边地区的大气污染物有着频繁的互相交换和输送关系，因此，周边地区的大气污染物排放对北京地区的空气质量也会造成一定程度上的影响。由此可见我国大气污染呈现出明显的区域性复合污染的特征，区域性复合污染的一个现象就是地理位置相近的城市颗粒污染物具有非常明显的一致性和同步性。区域性污染物特征另一个表征现象就是单纯控制单个城市的污染物浓度对污染物排放对大气中污染物浓度控制效果欠佳，必须在区域的尺度上协同控制才能解决区域性问题。

因此，一个地区的空气污染不仅仅与该地区自身产生的污染物有关，还与其周边地区的的大气污染物输送有着紧密的联系。所以，要想有效的治理大气污染，就需要从全方位来进行研究，不可仅限于对本地区污染状况的研究。

为了及时的监测空气质量的变化并迅速的对重污染天气采取出有效的应对措施，在每个地区的重污染频发的地区建立了空气质量监测站。但是，由于我国国土辽阔，无法在每个地区都建立完善的空前质量监测体系，因此，就会出现无法有效的采集污染物相关信息，从而导致无法正确对空气污染防治做出正确指导的情况。为此，合理的建立监测站对采集污染数据以及有效的防治大气污染起着至关重要的作用。

### 1.1.2 课题目的和意义

通过大量研究发现，一个地区的大气污染状况不仅是受当地污染物源的影响，同时很大一部分是来自周边地区的大气传输以及交换，因此大气污染具有跨区域输送的特点。因此，本文在空气质量监测站历史数据和气象监测站历史数据的基础上，根据污染物传播机理，构建污染物传播模型，搭建污染物传播网络，最终挖掘出污染物关键传播路径和脆弱点。

京津冀地区具有北临燕山山脉，西临太行山，东临渤海以及高度密集城市群的地理特征。独特的地理特征、气象条件以及各个地区不同的污染情况对污染物的传输都造成了显著的影响。目前，京津冀地区一共部署了大约200个空气质量监测站，其中北京地区大约有30个，天津地区大约有20个，从而在京津冀地区内形成了一个比较完善的污染物检测体系，可及时有效的获取各个地区的空气质量情况，并及时的加以防治。

通过研究发现，污染物的传输具有显著的区域性，往往一个地区受到污染后，它相邻的区域很快也会受到不同程度的污染，因此研究污染的传输规律，挖掘出污染物关键传输路径对污染物的防治起到至关重要的作用。因此，空气质量监测站监测的数据就具有重要的意义，但是，由于京津冀地区十分辽阔、昂贵的监测站建设成本以及维护费用，导致了不能在整个地区全方位的覆盖空气质量监测站。因此为了使用尽可能少的空气质量监测站来监测范围尽可能大的区域，从而获取经可能完善的数据，研究大气污染物区域内关键传播路径以及挖掘区域内脆弱点是十分有意义的。

## 国内外研究现状

十九世纪后期在日本足尾铜矿发生了一起震惊世界的大气污染事件。随后，世界各地工业国家也相继发生数起大气污染事件。第二次世界大战以后，许多工业发达国家，由于现代工业迅速地发展，带来了范围更大、情况更加严重的环境污染问题。美国洛杉矶市由于汽车尾气的排放，在本世纪40年代以后，经常在夏季出现光化学烟雾。欧洲由于燃煤造成大气污染，使北欧许多国家降酸雨，多次引起国际争端。在20世纪60年代后期，由于环境污染日益严重，各国人民的公害斗争此起彼伏，兴起环境保护运动，要求政府采取措施解决环境污染问题。进入20世纪80年代以后，由于人民对环境保护的要求日益提高，迫使各国政府开始重视环境保护，着手治理环境污染。

Chu等人介绍了兰州每个季节SO2浓度与气象条件的关系；Xu等人连续观察了一个地区连续七个月的空气质量，揭示了该地区的空气质量与污染物特征、区域交通状况、气象条件相关。Liu等人通过对北京2008年奥运会期间大气污染物的大量检测，发现了控制污染物的排放对污染趋势有着显著的效果；吕炜等人采用CMAQ模型系统获得了珠江三角洲主要污染物浓度时空分布和大气传输季节变化特征,并分析了长距离传输对珠三角区域空气质量的影响。薛文博等人基于CAMx空气质量模型定量模拟了全国PM2.5及其化学组分的跨区域输送规律。王扬锋等人通过新一代空气质量模式系统Models-3对冬季采暖期间沈阳市大气污染物的输送与化学转化进行了数值模拟研究，发现模拟值与观测值变化趋势有较好的一致性。张稳定等人利用嵌套网格空气质量模式系统(NAQPMS)较为合理的模拟出污染情况和输送过程。

Barmpadimos等人研究了气象因素对瑞士1991～2008年间PM10趋势和变化的影响，指出风和温度是两个最重要的因素；Lalas等人研究了复杂的海陆相互作用和特定区域不同地貌对大气污染物的影响，结果表明，海风环流影响Athens地表O3的日变化；Maraziotis等人发现PM10和PM2.5与水平风速呈负相关关系；SS Abdalmogith等[23]采用轨迹聚类模拟了大气运动对空气污染物长距离传输的影响；Garcia Menendez F [17]以三维欧拉输送模式研究了大气流动及预测了相应的空气质量变化；S Freitag等[24] 结合HYSPLIT4 模拟了西班牙气溶胶背向轨迹并对其进行聚类分析发现主要传输路径。

## 1.3 论文研究内容

本文利用京津冀区域内空气质量监测站的历史数据和气象监测站的历史数据，采用污染物传播机理、复杂网络、和PageRank改进算法来挖掘大气污染物在传播过程中的关键路径和脆弱点，为大气污染物的区域联防提供理论支持和帮助。主要内容如下：

（1）由于无法获取详尽的污染源清单等信息，因此无法使用传统的扩散模型来表征污染物的传播规律。本文通过采用分析污染物传播机理的方法，筛选出影响大气污染物扩散的气象因素和地理因素，从而建立污染物传播代价模型表征污染物传播关系。

（2）基于污染物传播代价模型，通过分析污染物传播约束条件来确定站点之间的路径。将监测站点视为节点，将站点之间的路径视为边，污染物传播代价视为边上的权重，从而将污染物传播过程抽象成一个有向加权污染物传播网络，并对该网络进行了度分布的特性分析，发现污染物传播网符合复杂网络特征。

（3）将污染物的每一条传播路径映射成一个序列，用有向图矩阵将每个序列进行表示，根据频繁阈值和边出现的重复次数判断向图矩阵中的边是否为有效边，最后将有效边拼接起来即为污染物传播关键路径。

（4）污染物传播网络中站点所处地区越是经常出现重污染天气，说明该节点脆弱性越强，其重要程度越高。为了找到污染物传播网络的这些十分脆弱，易受污染的站点，提出了一个基于PageRank算法的改进算法—NodeRank算法，该算法对传统算法中的节点数目以及转移概率进行了改进，充分考虑了边的权重以及相邻节点的重要。对所有监测站进行评分排序，评分越高，节点越脆弱，越易受到污染。

## 1.4 论文的组织结构

论文共分5章，本章主要介绍了课题的研究背景、目的和意义。从第2章起，论文结构如下：

第2章首先介绍了空气污染指数和污染物传播机理等基本概念；然后，简要阐述了复杂网络的基本概念；最后简要介绍了网页排序算法PageRank的基本原理。

第3章首先对京津冀地区的气象数据和污染物数据进行数据预处理；然后，根据污染物传播机理构建大气污染物传播模型；最后，设定污染物传播约束条件，并依据构建的传播模型挖掘实时的污染物关键传播路径，并对该大气污染物传播模进行了实验分析。

第4章首先在第3章实时污染物关键传播路径的基础上，构建一个周期内的污染物传播网络，并验证其复杂网络的特性。最后，在建立的京津冀大气污染物传播网络基础上，通过图矩阵的方法对污染物关键传播路径进行挖掘，并对关键传播路径挖掘方法进行了实验分析。

第5章首先对京津冀地区目前空气质量的现状进行了分析并阐述了网络节点重要性的评估方法；然后，在论述了传统PageRank算法的缺点以及应用于空气污染物传播有向加权网络的不适应性，并在此基础上提出NodeRank算法；最后，详细分析了在京津冀空气污染物传播网络基础上挖掘关键节点的实验过程与结果。

最后总结了本文研究成果，并对本文研究工作做出了总结和展望。

第2章 相关理论和基础知识

2.1 基本概念

### 2.1.1 空气质量指数

AQI，即空气质量指数（Air Quality Index），是定量描述空气质量状况的无量纲指数。由于AQI评价的6种污染物浓度限值各有不同，在评价时各污染物都会根据不同的目标浓度限值折算成空气质量分指数AQI。AQI范围从0到500，大于100的污染物为超标污染物。AQI分优（绿色）、良（黄色）、轻度污染（橙色）、中度污染（红色）、重度污染（紫色）、严重污染（褐红色）6种评价类别和表征颜色。当AQI类别为优或良，一般人群都可以正常活动；为轻度污染以上，各类人群就需要关注建议采取的措施。

空气质量分指数计算公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | (2-1) |

式中 ——污染物的空气质量分指数；

——污染物的质量浓度值；

——与相近的污染物浓度限值的高位值；

——与相近的污染物浓度限值的低位值；

——与对应的空气质量分指数；

——与对应的空气质量分指数。

空气质量指数计算方法如式(2-2)：

|  |  |  |
| --- | --- | --- |
|  |  | (2-2) |

式中 ——空气质量分指数；

——污染物。

当时，最大的污染物为首要污染物。若最大的污染物为两项或两项以上时，并列为首要污染物。

表2-1 空气质量分指数及对应的污染物项目浓度限值



### 2.1.2 污染物传播机理

空气污染物传播是一个十分复杂的过程，在这个传播过程中会受到多方面因素的影响。其中主要的三个方面是地形因素、气象因素以及污染物因素。地形因素包括：两个地区之间的水平距离 、两个地区之间是否有山脉、两个地区的海拔。气象因素包括：风速、风向、温度、湿度、降雨、雾、沙尘等。污染物因素包括：两地之间的浓度差。

水平距离：由于污染物传播具有远距离传输的特性，所以在一定范围的的地区，常常会与受污染的区域有着同步增长和减少的趋势，因此这些地区的相关性就会很高，这样就会导致当一个地区受到严重的污染时，与其相邻的其他地区的污染物浓度也会同步增加。

山脉：山脉对污染物的传播既有抑制作用，也有促进作用。当山脉位于两地之间时，这时山脉就会对污染物的传播起到一个阻挡和削弱的作用。当两个地区位于山脉的同侧时，污染物会沿着山脉的走势进行传播，此时山脉对污染物的传播就会起到促进的作用。

海拔高度：当两个地区的海拔高度相差不大，并且之间不存在山脉时，此时污染物的传播主要受气象因素的影响。如果两地之间的海拔高度差十分大时，低海拔地区对高海拔地区的影响很小，低海拔地区地势较低，污染物就会产生汇聚现象，导致该地区的污染物浓度很高。

风向、风速：风是影响一个地区大气污染物运动和分布的重要气象因子。风向决定了污染物的传播方向，风向决定了污染物的传播速度、传播的范围的大小以及局地地区污染物浓度的大小。当风速较大时，会使局地地区的空气变得稀薄，污染物的沉降效果降低，因此局地地区的空气质量会很好，但是这样污染物影响的范围会扩大。当风速很低时，不仅使污染物很难得到扩散，而且对污染物的聚集起到了促进作用，很容易会造成局部严重污染。

温度：逆温是指地面温度比较冷、上层温度比较热的“下冷上热”，形成比较稳定的形势，同时高空的暖温区与地面辐射降温相配合，形成逆温层，不利于污染物的对流和垂直扩散。

湿度：如果一个城市长时间无降雨，造成空气质量不佳，之后虽降了一场雨，但其效果只是“湿润空气”。同时由于空气湿度加大，让累积悬浮的污染物如同穿上一层水衣，给扩散加重负担。如果没有风，“重装”污染物基本上停滞于空气中，就会形成烟雨蒙蒙的效果，使空气质量不佳。在湿度较高的情况下，空气中的污染物会“吸湿增长”，导致其质量增加，污染物浓度也会相应提高，这样一来，空气质量反而会变差。只有在雨量较大，一般是中到大雨的情况下，才容易通过物理沉降的方式降低空气中污染物的浓度，起到净化空气的作用。

降水：雨水可以使大气污染物溶解于其中，对大气污染物的有着净化的作用，但降雨的净化作用与降雨量和降雨持续时间相关，雨水量大，持续时间长，净化作用就越明显，反之，则没有净化作用，甚至会加重污染。

沙尘：沙尘天气会使空气污染加剧，强沙尘暴会在极短的时间内使空气中的污染物急剧增加，造成严重的颗粒物污染。空气中高浓度的悬浮颗粒物，在静风、逆温等特殊的气象条件相互作用下，易导致雾霾天气的发生，具体表现为区域性或大范围内的空气质量恶化。

污染物浓度差：一个地区的污染物浓度十分高时，它对周围地区污染物浓度有着很大的影响，根据气体扩散原理，污染物浓度高的地区的污染物会向污染物浓度低的地方进行扩散，到这整个区域受到污染，另外受到风等气象因素的影响，会导致整个区域污染程度的加剧，因此，两个地点之间的污染物浓度差也是影响空气污染物扩散的一个重要因素。

2.2 复杂网络特性

如果存在一个网络，它具有自组织、自相似、吸引子、小世界、无标度中部分或者全部特性，则称该网络为复杂网络(Complex Network)。复杂网络具有高度的复杂性，其特点的具体表现如下：

（1）小世界(Scale-free)特性又称为六度分割理论，通俗来讲就是网络中任何一个成员和另一个成员之间间隔都不会超过六个人。复杂网络中小世界特性对网络中信息传播有着重要的作用。在这样的网络中，信息传播十分迅速，如果在网络中去除或者修改几个边时，就会导致整个网络性能的巨大变化。

（2）无标度(Scale-free)特征主要表现在网络中少数节点存在大量连接，而大量节点存在少数连接。这些节点的度数分布呈现幂律分布的特性。复杂网络的无标度性与网络的鲁棒性有着紧密的联系。无标度网络中的幂律分布特性很大程度上提升了高度数节点存在的可能性，正是因为这一点，也极大的降低了复杂网络的鲁棒性，也就是说，如果恶意攻击者攻击网络中度数很高的一部分节点时，很可能会造成网络迅速的瘫痪。

很多研究成果发现现实世界中很多网络都可以进行复杂网络抽象化，应用复杂网络理论研究其网络性质，这也为更复杂的网络关系研究提供了新思路。因此，将现实网络抽象为具有某些特殊拓扑特性的复杂网络，使人们对此网络进行理论性研究，应用复杂网络的性质对其进行定性和定量分析。

## 2.3 PageRank算法介绍

PageRank（网页排序）又称网页级别、Google左侧排名或佩奇排名[37]。由Stanford University 的SergeyBrin和Lawrence Page工程师发明。其理论是把节点间链接权重作为网页排名的因素。每个网页都会有一个PageRank值来表示网页的等级，其级别从1到10，网页的搜索结果会按照等级由高到低显示，即PageRank值越高表示网页越受欢迎。在网络中，如下图每个球代表一个站点，球的大小反应了站点的PageRank值的大小。指向站点B和站点E的链接很多，所以B和E的PageRank值较高，另外，虽然很少有站点指向C，但是最重要的站点B指向了C，所以C的PageRank值比E还要大。通过计算各个节点的PageRank值，PageRank值越高的节点越重要，最后对节点按照PageRank值由大到小排序，按需取排序靠前的几个节点为重要节点。



图2-1 PageRank示例图

因此，网页的PageRank值由三个因素决定：一是网页链入当前网页的数量；二是链入网页本身的重要度，表示了链接源是否为质量高的网页；三是链入网页本身的链出数量，即该链入网页给其它网页的投票数目。

设表示一个页面，为页面指向的页面集合，为指向页面的页面的集合，为页面指向的页面数，为页面总数，为阻尼系数。PageRank值用公式表示为：

|  |  |  |
| --- | --- | --- |
|  |  | (2-10) |

为向后搜索其它页面的概率，为用户随机浏览到新URL的概率。根据公式可以看出每个页面的PageRank值即为多个连接它的页面的投票值。PageRank算法步骤如下：

1. 最初阶段：

首先建立站点节点间网络关系图，其次初始化每个站点一样的PageRank值，通过在每次迭代过程中，不断更新各个站点的PageRank值，当前后两次迭代的PagePank差值达到一定阈值，迭代终止，得到每个站点的最终PageRank值。

(2)迭代计算站点PageRank得分方法：

每轮迭代中站点的PageRank值会赋值到本站点的出边上，使每条边都有相同的值。站点对链入自己的边权值求和，即获得新的PageRank得分。一次次迭代直到计算完成。

2.4 本章小节

本章主要介绍了本文用到的相关理论以及基础知识，首先介绍了空气质量指数的基本概念以及计算公式、影响污染物传播的相关因素。其次，介绍了复杂网络的相关特性。最后，对PageRank算法做了简单的介绍。

第3章 构建大气污染物传播模型

3.1 数据预处理

数据去噪：由于数据中存在一些值为空值，这些空值将会对后边的计算产生一定的误差，考虑到样本数据十分的多，因此本文采用了直接去除空值的手法。即如果某条数据中的一项为空值，则舍弃该条数据。

### 3.1.1 监测站之间的距离

地球是一个近乎标准的椭球体，平均半径为6356.755km。假设监测站A的经纬度为(lonA,latA)，监测站B的经纬度为(lonB,latB)。然后根据三角推导，可以得到计算两监测站距离的公式如下：

|  |  |
| --- | --- |
|  | （3-1） |
|  | （3-2） |

其中，R为平均半径6356.755km，pi为3.1415。

### 3.1.2 监测站方向与风向的夹角

根据向量乘积的原理，假设存在向量和，因此存在如下等式：

|  |  |
| --- | --- |
|  | （3-3） |

其中和代表向量的横纵坐标，和代表向量的横纵坐标，和表示向量和的模，为两个向量之间的夹角。

假设风向的单位向量为，从监测站A到监测站B方向的向量为，如图3-1所示。

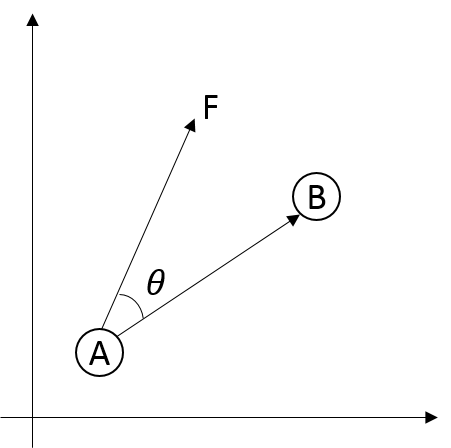


图3-1风向夹角示意图

因此可将上述向量乘积的原理应用到计算两个监测站方向与风向之间夹角的计算中，公式如下：

|  |  |
| --- | --- |
|  | （3-4） |

于是可以求出夹角为：

|  |  |
| --- | --- |
|  | （3-5） |

其中为风向的弧度，为向量的模。

### 3.1.3 监测站之间污染物的相关性

假设监测站A在t时刻的污染物浓度为，监测站B在t时刻的污染物浓度为，监测站A上污染物浓度均值为，监测站B上污染物浓度均值为。因此，A、B监测站之间污染物的相关性为可根据公式(6)求得：

|  |  |
| --- | --- |
|  | （3-6） |

### 3.1.4 数据标准化

由于本文中涉及到的各个因素的单位是不同的，如果对他们直接进行计算，那计算得出的结果不具有科学性。因此需要对每个因素做去除单位的操作—数据标准化。数据标准化就是将不同单位的数据进行无量纲转换，这样不同单位的数据就相互进行计算而不受单位限制。

本文采用标准化方法，对原始数据进行线性转换。公式如下：

|  |  |
| --- | --- |
|  | （3-7） |

其中为样本数据的最大值，为样本数据的最小值。

3.2 污染物传播模型构建

空气污染物传播起点的气象因素，如：温度(T)、湿度(H)、风向(WD)、风速(WS)等都会影响空气污染物的向外扩散，并且空气污染物的传播难易程度不仅受本地气象因素的影响，也会受到传播目的地的影响。其中，两地之间的距离(D)，两地之间的地形(hill)，两地之间的海拔差)，监测站污染物的相关性()，两地之间的污染物浓度差()，以及两地之间的气象因素的差异都会影响到空气污染物的传播。根据第2.12节提出的大气污染物传播机理，t时刻空气污染物从站点A输送到站点B的传播代价为：

|  |  |
| --- | --- |
|  | （3-8） |

如式中表示起始点A在t时刻的湿度，表示起始点B在t时刻的温度，表示站点A与站点B之间的水平距离，表示站点A与站点B之间时刻的差值，表示站点A与站点B之间的相关性，表示站点A与站点B之间t时刻的风力系数差值，表示传播代价的波动值，表示校正系数，，取值范围均为。

通过污染物在站点A与站点B之间的传播代价可以衡量出污染物传播的难易程度，污染物传播代价越大，代表污染物越是不易传播，反之，越容易传播。该公式不仅考虑了污染物起始地的气象因素，也考虑到了污染物目的地的气象因素，以及两地之间的地形因素。而不是像其它文献中的模型只考虑风向风速等气象因素。

3.3 实时关键传播路径挖掘

由于地形、地貌等一些自然条件的限制，和气象因素的影响，导致在一些地区出现一些污染物的传播通道，例如：以京津冀为例，其西临太行山、北临燕山、东临渤海，导致在偏西南风气流的影响下，污染物会沿着太行山系中的洋河河谷和燕山山脉向京、津、冀地区传播，导致大面积、区域性的污染。因此研究污染物的关键传播路径是十分重要的，不仅可以有效的对污染进行区域联防，而且可以及时阻断污染物的传播。并且由于京津冀地区十分广阔，不能实现监测站的全方位的覆盖，因此寻找污染物的关键路径，并合理的在关键路径上设置空气质量监测站，不仅减少了监测站的建设成本，也对整个地区起到了有效的防控作用。

### 3.3.1 传播约束条件分析

水平距离(D)：根据数据集station.csv中的经纬度以及公式(1-2)计算出站点之间的距离。如果监测站和监测站的实际水平距离小于或者大于50km，则断开监测站和监测站之间的路径。因为当距离小于10km时，基本上两个站点所有状态以及数据都是相同以及同步的，这样就会导致在一个局地地区存在大量的路径，使网络中存在大量的环，这样对研究污染物的区域性传播十分不利。从数据可知，大多数县级区域有且至少有一个监测站点，并且距离都在50km以内，又因为空气污染物的传播具有远距离传输特性，并且传输距离远大于50km，因此如果两监测站距离如果大于50km，就断开连接。

山脉(hill)：当两个监测站之间存在山脉时，则认为污染物是无法在这两个地区之间进行传播的。

海拔差()：当海拔差超过一定数值时，则认为污染物无法从低海拔的地区传播到高海拔的地区。

相关性()：当两个在适当范围内的站点上的污染物浓度出现同增或者同减的现象时，往往是因为一个站点上的污染物浓度对另一个站点产生了影响。因此，根据公式(6)对网络中的任意两个站点进行污染物浓度的相关性计算。当时表示站点之间有紧密依赖关系，也就是说一个站点的污染物浓度很容易就会影响到另一个站点，也就说明了这两个站点之间存在一条污染物传播通道。反之不存在传播通道。

风力系数差()：设站点A到站点B的方向为，站点A和站点B在t时刻时的风速风向分别为、，则分别与站点A和站点B风向的夹角为，根据三角形勾股定理可知，风速、在方向上的分量分别为，如图3-2所示。

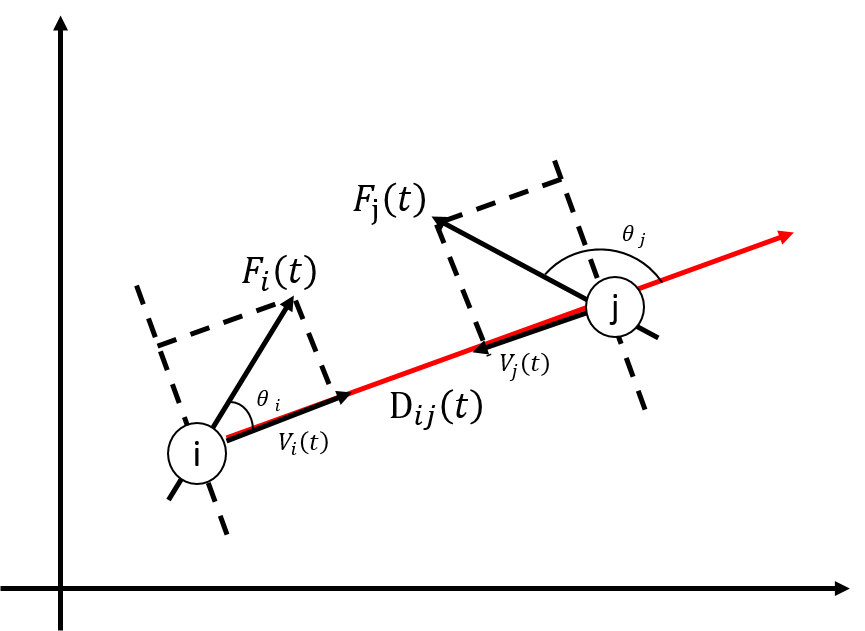


图3-2 监测站上风力系数示意图

因此可以求出站点A与站点B之间t时刻的风力系数差值为：

|  |  |
| --- | --- |
|  | （3-9） |
|  | （3-10） |
|  | （3-11） |
|  | （3-12） |

其中可根据公式(3-5)计算得出。

如上式，当站点A处的风向与A、B站点方向的夹角超过时，则认为此时站点A处的空气污染物不会对站点B处的空气质量造成影响，或者当小于等于0时，表示站点A在方向的风力系数小于等于站点B在方向上的风力系数，即站点A处的空气污染物是无法对站点B处造成影响的。

### 3.3.2 实时关键传播路径挖掘算法

当污染物从一个地区向相邻地区进行扩散传播时，往往不只是一条路径进行传播，而是存在多条不同的传播路径。由于地形因素以及气象素的影响，会导致其中一些路径对污染物的传播有极小的抑制作用，这样就使得污染物在这条路径上很容易被传播，而有些路径则对污染物的传播有极大的抑制作用，导致污染物很难在这条路径传播，其中这种难易程度本文称之为传播代价(Cost)。因此，从传播原理来看，传播代价越小的路径越是容易形成污染物的传播通道，所以，将传播代价最小的路径定义为污染物关键传播路径。

根据污染物传播的约束条件以及关键传播路径的定义，可总结出站点A与站点B存在边的条件，如下式：

(3-13)

其中，表示监测点A和监测点B之间的距离，表示监测点A和监测点B在t时刻的风力系数的差值，表示监测点A和B之间污染物的相关系数，代表监测点A和B在t时刻的污染物传播代价。当为0时，表示站点A、B之间在t时刻不存在一条污染物关键传播路径，为1时表示存在一条污染物关键传播路径。

t时刻污染物从站点A传播到站点B的传播路径)：

其中，表示时刻t时路径)内包含边的集合。

t时刻污染物从站点A传播到站点B的传播路径对应传播代价：

其中，表示时刻t时路径)内所包含边的传播代价的集合。

根据污染物传播模型以及污染物传播代价即可计算出t时刻的从任意一点出发的污染物关键传播路径。如算法3.1所示：

算法3.1 Search Critical Path(SCP)

输入： ，

，

，

输出：

执行过程：

BEGIN

1. Initialize:
2. zero List nodeList,
3. ;
5. ;







14. end for

END

如算法3.1所示，首先任意选择一个污染物传播的起始点，方法根据公式（3-2）计算起始点与每个监测站的距离。并选择出距离在10km到50km范围内的站点。将这些站点的集合记作，然后接下来的计算将基于集合中的站点进行。由于数据集中风向不是用角度进行表示的，因此需要通过方法进行角度转换。通过方法计算时刻t时中每个站点与的风向夹角，并筛选出夹角在0度到90度的站点，并将这些站点的集合记作AngleList，然后在AngleList的基础上通过方法计算每个站点与的在t时刻时的传播代价，并将符合传播代价小于0.01的站点的集合记作CostList，最后在CostList中找到传播代价最小的站点，因此路径即为在t时刻以为起点的一条最容易传播的路径，这条路径因此称之为污染物关键传播路径。

### 3.3.3 算法实例

下面将通过一个实例展示出算法1的过程。该实例选取井陉县气象局2014年4月26日时的气象数据以及污染物浓度数据来寻找以井陉县气象局为起点的污染物关键传播路径。井陉县气象局的经纬度为。根据方法计算出满足与井陉县气象局距离在10km到50km范围内的监测站，如表3-1所示。

表3-1 距离井陉县气象局的监测站

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 监测站 | 监测站 | 监测站 | 监测站 | 监测站 |
| 市区世纪公园 | 市区西南高教 | 市区化工学校 | 市区职工医院 | 市区人民会堂 |
| 市区西北水源 | 市区高新区 | 市区封龙山 | 鹿泉一中 | 正定联通公司 |
| 栾城通讯公司 | 元氏住建局 | 赞皇县政府 | 平山冶河 | 灵寿供水 |

通过方法计算出表3-1中监测站与井陉县气象局风向的夹角，如表3-2所示。

表3-2 各监测站与井陉县气象局风向夹角

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 角度 | 角度 | 角度 | 角度 | 角度 |
| 63 | 135 | 81 | 150 | 165 |
| 45 | 143 | 35 | 132 | 145 |
| 87 | 56 | 156 | 180 | 90 |

由于当夹角为钝角时，视为污染物不能传播到下一个污染物监测站，因此去除夹角为钝角的点，则剩下的监测站为污染物可以影响到的最大范围，其中符合要求的监测站如表3-3所示。

表3-3 符合夹角为锐角的监测站

|  |  |
| --- | --- |
| 监测站 | 夹角 |
| 市区世纪公园 | 63 |
| 市区西北水源 | 45 |
| 栾城通讯公司 | 87 |
| 元氏住建局 | 56 |
| 市区化工学校 | 81 |
| 市区封龙山 | 35 |

最后通过方法计算出监测站井陉县气象局到表3各点的传播代价，计算结果如表3-4所示。

表3-4 井陉县气象局到各个监测站的传播代价

|  |  |  |
| --- | --- | --- |
| 监测站 | 夹角 | 传播代价 |
| 市区世纪公园 | 63 | 0.002 |
| 市区西北水源 | 45 | 0.0015 |
| 栾城通讯公司 | 87 | 0.04 |
| 元氏住建局 | 56 | 0.025 |
| 市区化工学校 | 81 | 0.035 |
| 市区封龙山 | 35 | 0.015 |

根据表3-4的传播代价就可找出污染物的关键传播路径，即传播代价最小的站点即为下一个传播起点。其中由井陉县气象局为起点的关键传播路径为：

由表4中的风向夹角和传播代价可以看出，并不是所有的风向夹角的越小的监测站上的传播代价越小。这是因为，由于两个监测站之间污染物的传输，不仅仅受风向影响，风向只是决定了污染物的主要的传播风向，而无法反映出污染物对下一个站点的影响程度，其中地形因素、气象因素，都很大程度上决定了污染物是否能影响到其他站点。因此就需要通过污染物传播代价来衡量污染物的传播难易程度，并以此来找出关键传播路径。

3.4 实验结果与分析

本文的最终目的是找到污染物实时的传播规律，并依据此传播规律来寻找污染物的实时关键传播路径。第二章通过对污染物传播机理的详细分析，找出了影响污染物传播的关键因素，并基于污染物传播机理建立了污染物传播模型。第三章提出了污染物传播的约束条件和寻找污染物传播关键路径的详细步骤以及方法，本章针对上述寻找出的污染物关键路径进行试验，并对实验结果进行分析。

### 3.4.1 实验环境配置

实验的系统环境配置为macOS 10.13.2操作系统，硬件环境配置为Intel Core i5，主频为3.10GHz，内存为16 GB 2133 MHz LPDDR3。算法采用Java，Matlab编写。数据展示采用Echarts.js以及Excel，运行环境为IntelliJ IDEA，JDK版本为1.7，采用的数据库为MySQL数据库。

### 3.4.2 实验数据

本文所需的数据集来源于微软亚洲研究院郑宇博士所提供京津冀地区2014年5月1日到2015年4月30日的数据集，数据集集中包含airquality.csv、city.csv、district.csv、meteorology.csv、station.csv五个文件。其中数据集详细解析如下：

表3-5 station.csv

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 站点编号 | 名称 | | 纬度 | | 经度 | 区域编号 |
| 1001 | | 海淀北部新区 | 40.090679 | 116.173553 | | 101 |
| 6010 | | 天津前进道 | 39.092699 | 117.201676 | | 607 |
| 14004 | | 建设大厦 | 39.942 | 119.537 | | 1401 |

station.csv文件中展示了京津冀范围内所有空气质量监测站的详细信息：站点编号、站点名称、站点所在位置的经纬度，站点所属区域编号。

表3-6 city.csv

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 城市编号 | 名称 | | 纬度 | | 经度 | 类别 |
| 001 | | 北京 | 39.904210 | 116.407394 | | 1 |
| 006 | | 天津 | 39.084158 | 117.200982 | | 1 |
| 011 | | 石家庄 | 38.042307 | 114.514860 | | 1 |

city.csv文件中展示了京津冀范围内所有的市级城市的详细信息：城市编号、城市名称、城市所在位置的经纬度、城市类别（1：北方；2：南方）。

表3-7 district.csv

|  |  |  |  |
| --- | --- | --- | --- |
| 区域编号 | 名称 | | 城市编号 |
| 00101 | | 海淀区 | 001 |
| 00102 | | 石景山区 | 001 |
| 00103 | | 丰台区 | 001 |

district.csv文件中展示了京津冀区域内所划分的区域（区、县）的详细信息：区域编号、区域名称、区域所属城市。

表3-7 airquality.csv

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 站点编号 | 日期 | PM2.5 | PM10 | NO2 | CO | O3 | SO2 |
| 1001 | 2014/5/1 00 | 138 | 159.4 | 56.3 | 0.9 | 50.8 | 17.2 |
| 1002 | 2015/4/20 10 | 45 | 41.6 | 56.7 | 0.7 | 39.8 | 2.9 |
| 1003 | 2015/3/10 02 | 11 | 30 | 34 | 0.6 | 37 | 5 |

airquality.csv文件展示了京津冀区域内每个空气质量监测站每个时刻（小时）采集到的污染物浓度信息，其中需要采集的污染物包括：PM25、PM10、NO2、CO、O3、SO2。它们的单位均除了污染物CO为以外，其余均为。由于本文中需要用到的是污染物的日均AQI值进行计算，因此需要将时均浓度值转化为日均值，并按照AQI的计算公式将浓度转化为AQI。

表3-8 meteorology.csv

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 区域编号 | 日期 | 温度 | 压强 | 湿度 | 风速 | 风向 |
| 616 | 2015/4/18 03 | 15 | 1009 | 65 | 1.8 | 1 |
| 1108 | 2015/4/6 10 | 7 | 1027 | 29 | 6.12 | 13 |
| 1405 | 2015/3/30 05 | 5 | 1010 | 85 | 3 | 24 |

meteorology.csv展示了每个区域在每一个时刻（小时）的气象状况。其中包含温度、压强、湿度、风速、风向五个数据。由于本文中风向需要用角度表示，因此需要根据十六风向图将该数据集中的风向转换为角度，其中风向标识与风向角度的对应关系如下表所示：

表 3-9 风向标识与风向角度对应关系

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 风向标识 | 描述 | 角度 | 风向标识 | 描述 | 角度 |
| 0 | 无风 | Null | 9 | 不稳定 | Null |
| 1 | 东风 |  | 13 | 东南风 |  |
| 2 | 西风 |  | 14 | 东北风 |  |
| 3 | 南风 |  | 23 | 西南风 |  |
| 4 | 北风 |  | 24 | 西北风 |  |

### 3.4.3 实验结果分析

本实验采用2015年2月6日凌晨3点京津冀各个监测站的气象数据、污染物浓度数据，并将算法SCP中的传播代价上限设置为0.01。通过算法SCP的计算，可计算出两个监测站之间是否会存在传播路径，结果如图3-3所示。

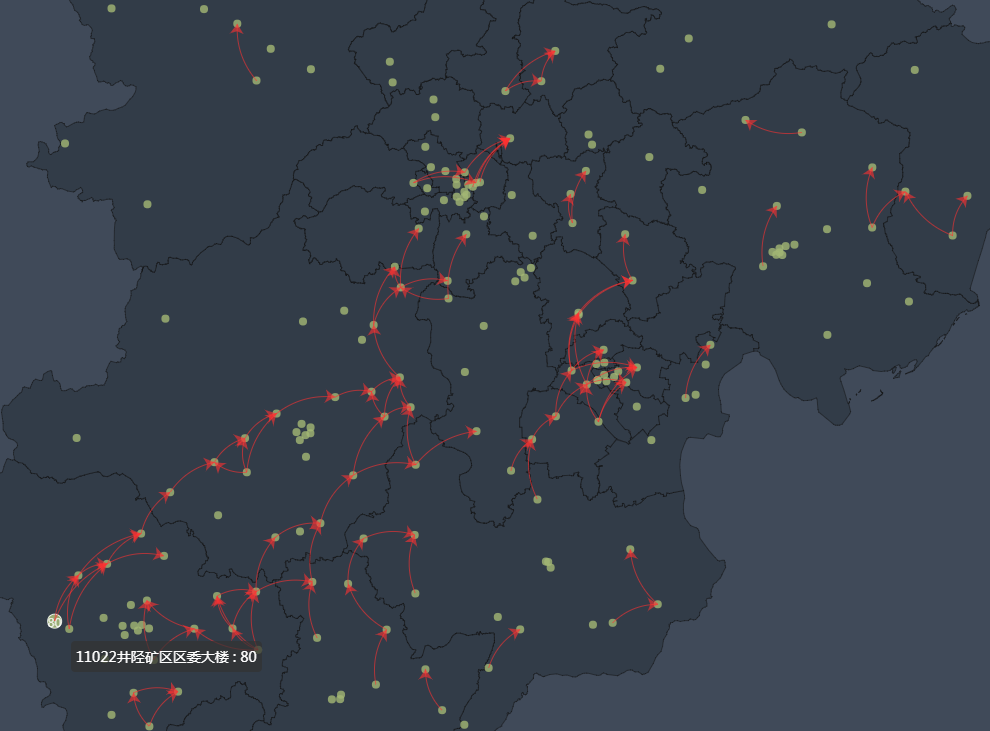


图3-3 京津冀地区2015年2月6日凌晨3点污染物传播路径

根据第3.3节所描述的污染物关键路径的搜索方法以及算法实例，本章的实验以2015年2月6日凌晨3点井陉矿区区委大楼监测站点为起点。由于所研究站点的数据量十分大，因此需要截取其中一部分进行实验分析，截取的路径部分如图3-4所示。

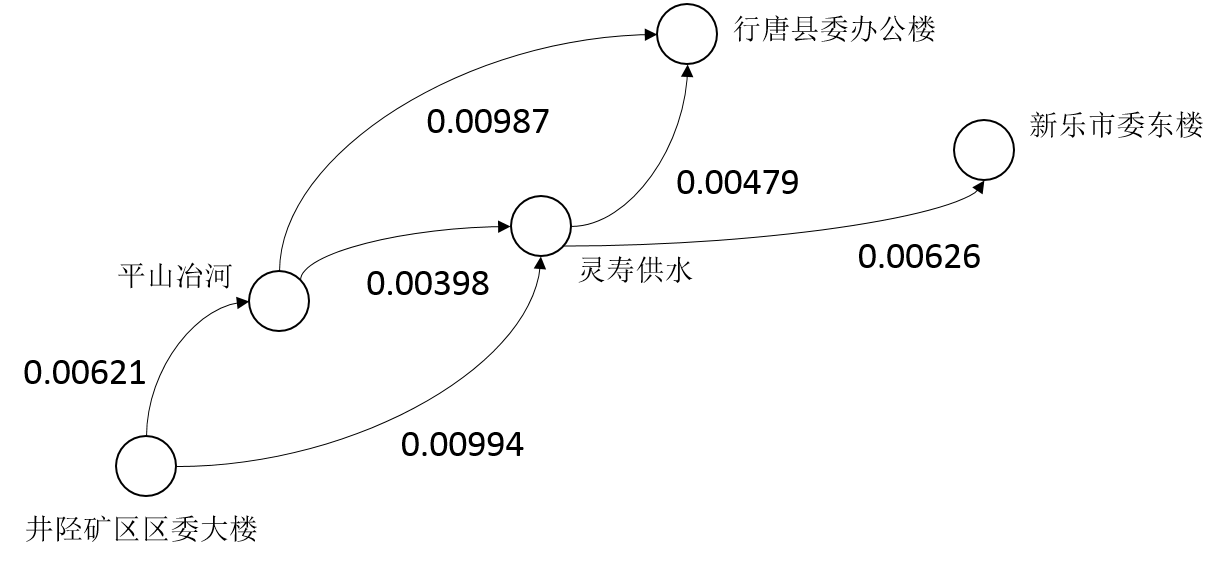


图3-4 2015年2月6日凌晨3点部分污染物传播路径

由图3-4可知，以井陉矿区区委大楼为起点的污染物传播路径存在两条，即

其边上的数值为此条路径上污染物的传播代价，根据污染物关键传播路径的定义，选取传播代价最小的一条边为关键传播路径，因此路径为关键路径，根据此方法，接下来以平山冶河监测站为起点，选取传播代价最小的路径为关键路径。然后，以灵寿供水监测站为起点，选取传播代价最小的路径为关键路径。最终，以井陉矿区区委大楼为起点的污染物关键传播路径为：

表3-10展示出了2015年2月6日凌晨3点以井陉矿区区委大楼为起点的完整的污染物关键传播路径：

表3-10 2015年2月6日凌晨3点污染物关键传播路径

|  |  |  |  |
| --- | --- | --- | --- |
| 监测站名称 | 监测站编号 | 监测站名称 | 监测站编号 |
| 井陉矿区区委大楼 | 11022 | 徐水环保局 | 17009 |
| 平山冶河 | 11023 | 容城县环境保护局 | 17015 |
| 灵寿供水 | 11024 | 白沟新城行政中心楼 | 17022 |
| 行唐县委办公楼 | 11025 | 高碑店环保 | 17012 |
| 唐县政府楼 | 17017 | 涿州监测站 | 17011 |
| 满城税务局 | 17007 | 京南榆垡 | 1035 |

其中这条关键路径的方向是由上至下，即从井陉矿区区委大楼监测站到徐水环保局监测站的方向。由于这条关键传播路径是根据传播机理以及传播代价公式计算所得，因此它表征了污染物的在这个时刻的一个最易扩散的趋势，即当井陉矿区区委大楼的污染物浓度上升时，这条路径所经过的区域将有很大的概率也受到污染。并通过Echarts将关键路径在地图上展现出来，如图3-5所示。

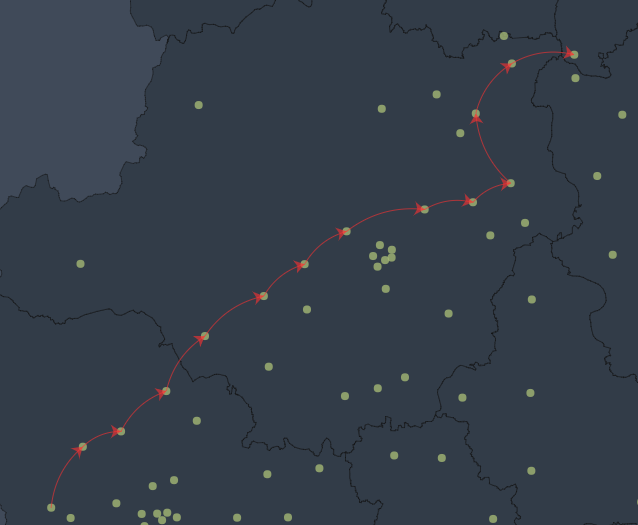


图3-5 关键传播路径

由图3-5可以直观的看出污染物传播路径的一个大致的方向，其中，在该路径的左侧是太行山山脉，由于受山势的影响，这里常年盛行西南风，图3所示的路径也恰恰反映出了这一特点，该路径正是沿着太行山山脉的走势，在西南风的影响下，途径保定，将污染物输送至北京，从而对北京的空气污染造成了一定的影响。

### 3.4.4 实验结果验证

接下来以路径为例，来分析验证污染物关键传播路径的科学性以及可行性。为了与这条关键路径形成对比，于是，选择了一个不在这条关键路径但与起始站点井陉矿区区委大楼距离非常接近的站点，即井陉县气象局站点。并根据各个监测站每个时刻的污染物浓度绘制出变化曲线，如图3-6、表3-11所示。

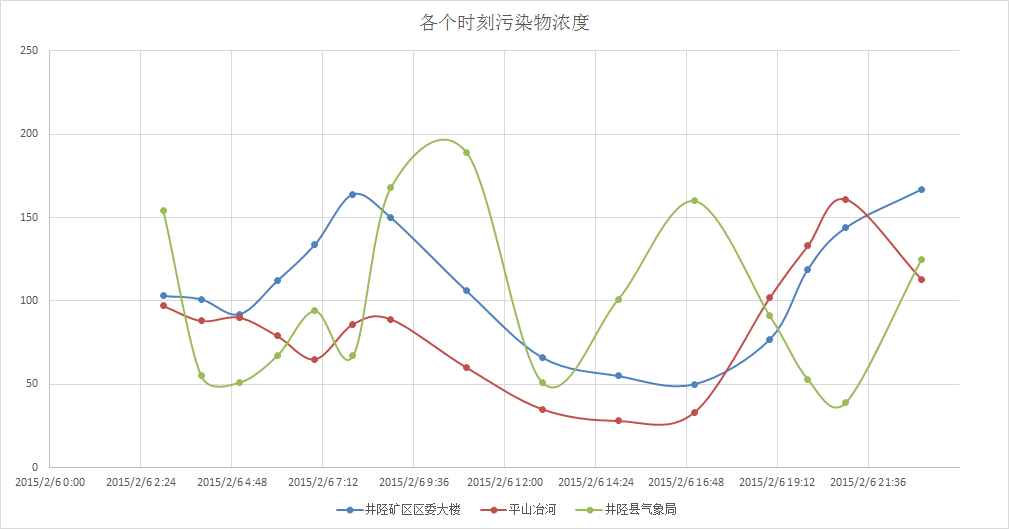


图3-6 各个时刻污染物浓度

表3-11 三个监测站各个时刻污染物浓度

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 时刻 | 井陉矿区区委大楼 | | 平山冶河 | 井陉县气象局 |
| 2015-02-06 03:00:00 | 103 | 97 | | 154 |
| 2015-02-06 04:00:00 | 101 | 88 | | 55 |
| 2015-02-06 05:00:00 | 92 | 90 | | 51 |

由于这是一条出现在6日凌晨3点的以井陉矿区区委大楼监测站为起点的关键路径，因此将着重分析凌晨3点及以后的各个时刻的污染物浓度的变化情况。

通过查询数据库中6日凌晨3点井陉矿区区委大楼的气象信息可知当时的风速为6.12m/s，并根据距离公式可计算出井陉矿区区委大楼与平山冶河的距离为24.7km。由于两地距离很近，因此不考虑风速变化的问题，这样就可计算出污染物扩散到平山冶河的大致耗时为1.12小时，即平山冶河监测站的污染物浓度将会在6日凌晨4点07分以后开始出现上升趋势。并且由图3-6可知，平山冶河监测站的污染物浓度确实在6日凌晨4时之后开始上升，其中就是由于受到了井陉矿区区委大楼污染物扩散的影响。

然后再看并不在这条路径上的井陉县气象局污染物的浓度，它的上升趋势是出现在了6日凌晨4点之后。根据距离公式可计算出井陉县气象局与井陉矿区区委大楼的距离为8.15km。如果路径为6日凌晨3点的污染物关键传播路径，那井陉县气象局污染物的浓度的上升点应该会在凌晨4点之前而不是之后。由此可以证明路径并不是6日凌晨3点以井陉矿区区委大楼监测站为起点的污染物关键传播路径。然后通过表8展现出其余各个子路径的实际情况，如表3-12所示。

表3-12 关键路径中子路径实际情况

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 起始点 | 风速(m/s) | 历时(h) | 下一站点 | 下一站点上升点 | 是否符合 |
| 平山冶河 | 7.92 | 0.54 | 灵寿供水 | 2015-02-06 03:32:00 | 是 |
| 灵寿供水 | 8.64 | 0.71 | 行唐县委办公楼 | 2015-02-06 03:43:00 | 否 |
| 行唐县委办公楼 | 6.84 | 2.0 | 唐县政府楼 | 2015-02-06 05:00:00 | 否 |
| 唐县政府楼 | 2.52 | 4.22 | 满城税务局 | 2015-02-06 07:13:00 | 是 |
| 满城税务局 | 6.12 | 1.35 | 徐水环保局 | 2015-02-06 04:21:00 | 是 |
| 徐水环保局 | 3.96 | 1.26 | 容城县环境保护局 | 2015-02-06 04:15:00 | 是 |
| 容城县环境保护局 | 8.28 | 0.52 | 白沟新城行政中心楼 | 2015-02-06 03:31:00 | 是 |
| 白沟新城行政中心楼 | 4.68 | 1.66 | 高碑店环保局 | 2015-02-06 04:39:00 | 是 |
| 高碑店环保局 | 4.68 | 1.32 | 涿州监测站 | 2015-02-06 04:19:00 | 是 |
| 涿州监测站 | 3.6 | 1.77 | 京南榆垡 | 2015-02-06 04:46:00 | 是 |

由表7可以计算出此条关键路径的准确度为：

因此可以说明该路径与实际情况基本吻合，能够准确的反映出在该时刻污染物的一个潜在的传播通道，可以为污染发生提供预警以及及时的防控作用。所以，由此可以证明该传播模型的科学性、准确性和实际应用中的可行性。

## 3.5 本章小结

为了更加全面的研究大气污染物的传播规律，本文对大气污染物的传播机制进行了深入的研究，分析出了影响大气污染物传播的几个必要条件。一个地区的大气污染不仅仅与本地的污染物排放有关系，而且与该地区的相邻地区的大气污染物的输送有着紧密的联系。影响大气污染物传输的因素有许多，其中风速、风向、降雨降雪、湿度、压强、地面粗糙度、地形等都对大气污染物的传输起着至关重要的作用。因此本文依据大气污染物的传播机制，推导出大气污染物的传播代价，以此来衡量大气污染物的传播的难易程度。并以此为依据，筛选出下一个将要传播至的点。并根据实际的大气污染物浓度变化的趋势对该路径做验证，证明了该路径的准确性，从而也证明了本文提出的污染物传播代价公式的科学性。同时，也为防治大气污染物传播、及时阻断大气污染物传播提供了理论依据。

第4章 基于图矩阵的关键传播路径挖掘

## 4.1 引言

由于大气污染物具有区域性传输的特点，往往会因为一个地区发生污染而导致相邻区域也遭受不同程度的污染，当两个地区之间存在一条在某种气象条件下对污染物传输十分有利的路径时，这条路径在这个时刻就会成为这两个地区的污染物关键传播路径。因此，为了全面的了解一个地区的污染状况，并以此制定出有效的防治策略，就需要挖掘在一个周期内频繁出现的路径，并将频繁出现的路径就视为污染物关键传播路径，并针对这些路径经过的地区制定相应的污染物防治的策略，才能在整体上对污染物的起到防治的作用。

## 4.2 构建大气污染物传播网络

根据第三章提出的大气污染物传播代价公式以及大气污染物传播模型，可以计算出每个时刻（小时）的大气污染物传播路径，由于本文是以年为周期进行的统计分析，因此在一个周期中会产生大量的污染物传播路径，然后将每个时刻的路径存储到一个矩阵当中，并将该位置的值置为1，若在另一个时刻该条路径再次出现时，则在之前的基础上进行加1，以此类推，最终会得到大气污染物在一个周期内传播的邻接矩阵，其中的矩阵中的值表示了该条路径在一个周期内出现的次数。其中，一个周期内大气污染物的传播网络图如下图所示。

在大气污染物传播网络中，节点代表空气污染物监测站点，边代表污染物从一个区域传播到另一个区域的路径，当在一个周期（年）中，大气污染物多次从一个地区传播到另一个地区时，就表明了这两个地区之间很可能存在一条有利于污染物传播的通道，因此，考虑节点之间的依赖关系主要是该路径在一个周期中出现的次数，即边上的权重代表可以用路径在一个周期中出现的次数来表示。

因此，这个大气污染物传播网络图可以表示为三元组，其中，表示网络中点的集合，表示网络中边的集合，表示网络中边上的权重的集合，由于本文中用到的大气污染物传播网络图是一个有向图，因此存在。其中计算权重的算法具体过程如下所示：

算法4.1 统计周期内大气污染物传播网络边上的权值

输入：站点的集合stationList和周期内传播路径集合pathList

输出：边出现次数集合martixNum

BEGIN

1. //用于存储每条边的出现次数


5. //存储终点，并初始化为0
6. r
7. //存储起点和终点
8. r

11. r
12. r

END

如算法4.1所示，首先对京津冀内所有的监测站集合进行遍历，并根据这些监测站初始化矩阵martixNum，矩阵中的行表示大气污染物传播的起点，列表示大气污染物传播的终点，矩阵的初始值均设置为0。然后对周期内大气污染物传播路径集合进行遍历，并将每条大气污染物传播路径中的路径片段存储到martixNum矩阵中，并在之前的数值基础上进行累加。最终会得到一个周期内的大气污染物传播网络的邻接矩阵。

### 4.2.1 基本概念与定义

通过上一节获得了大气污染物传播网络有向图，为了获得大气污染物的关键传播路径，本文提出了基于图序列的挖掘方法，此方法在挖掘的过程中需要对访问过的边和节点进行记录。因此，为了对网络中节点或者边的访问情况进行记录，接下来定义了大气污染物传播网络有向加权标识图。

定义4.1 染物传播网络有向加权标识图：图是一个有向加权标识图，其中：是一个有向加权图，其符合大气污染物传播网络有向加权图的定义，是图 ，其作用是表示监测站被访问情况。的值只有true和false两个。是一个映射函数，其作用计算有向加权图的节点和边被访问的情况，并赋予相应的标识，。

定义4.2 子图：存在大气污染物传播网络有向加权标识图和，若图和的节点、边以及映射函数满足条件且且,则将图称作图的子图，记作。

定义4.3 边序列：如果大气污染物传播网络有向加权标识图G中站点之间存在边，则将置为1，否则将置为0。最终将一条大气污染物传播路径映射成一个0/1的序列，并将该序列称之为边序列。

定义4.4 有向图矩阵APTM：如果已经知道图 G中的边与节点的出现情况，则创建一个矩阵，矩阵中每个位置的值由0/1字符串构成，表示该对站点对应的边的出现情况，将该图序列矩阵称之为大气污染物传播网络有向图矩阵 APTM。

定义4.5 关键路径序列：在大气污染物传播网络有向图 G中，若存在一个频繁连通子图可以通过序列表示，并且该频繁联通子图中所有的边都满足了预先设置的频繁阈值和重复次数的限制，则称该序列为大气污染物传播的关键路径序列。

## 4.3 基于图矩阵的污染物关键传播路径挖掘

本文通过图矩阵的数据结构存储大气污染物在两个地区之间的传播关系，并在此基础上运用图序列的挖掘方法挖掘出大气污染物的关键传播路径。

首先，构建大气污染物传播网络，根据第三章提出的大气污染物传播机理以及传播代价公式，将站点之间的调用关系存储至一个三列矩阵中。

其次，根据一个周期内的空气质量监测站的污染物浓度的数据信息，统计出京津冀地区各个城市污染物的超标率，并将这些超标的城市分别作为大气污染物传播的起点，获取若干个污染物传播的路径，可将这些污染物传播路径看成是大气污染物传播网络有向图的一个子图。并根据定义三，将路径中边的出现情况按照子图的顺序依次用0/1表示，构建有向图矩阵APTM。

再次，设置频繁阈值，筛选出满足频繁阈值要求的序列，并将这些满足要求的序列转换成十进制数，这样便于比较。

最后，统计出矩阵中十进制数相同的边，并将这些边拼接成序列，即大气污染物传播网络的关键传播路径。

### 4.3.1 获取污染物传播网络子图

为了生成大气污染物传播网络子图，则需要获取大气污染物传播的起始点，本文采用一个周期（年）内污染物超标率比较严重的地区作为起始点，具体算法过程如算法4.2所示。然后分别以这些地区为起点对污染物传播网络进行深度优先遍历，从获得若干条大气污染物传播路径序列，这些序列即为大气污染物传播网络子图，具体算法过程如算法4.2所示。

算法4.2 统计一个周期内京津冀各城市的污染物超标率

输入：

输出：

执行过程：

1. //对每行数据按逗号进行分割,获取每个时刻的污染物浓度信息








11. 将中的值都设置为初始值0

END

算法4.2中首先初始化用于存储一个地区的污染物浓度总和、污染天数、污染物超标天数，然后对一个周期（年）内的京津冀大气污染物浓度的信息进行遍历，由于每一条信息是按逗号分隔的，因此将其分割，并取得对应城市的编号以及该地区的污染物浓度值（5-7行），若中不存在对应的地区的值，则初始化该地区的污染物浓度总和、污染天数、污染物超标天数均为0，否则将该地区每一时刻（小时）的污染物浓度进行累加，污染天数进行累加，如果该时刻的污染物浓度超出了阈值，则将污染物超标天数进行累加（9-12行）。最后，按对进行遍历，计算出每个城市一个周期内平均污染物浓度以及污染物超标率。算法结束。

算法4.3 根据传播网络计算传播路径

输入：传播网络邻接矩阵M、城市编号cityId

输出：传播路径集合pathList

BEGIN

1. Initialize pathList
2. Initialize list
3. for(each key in cityId)
4. int[] len = {0}
5. list.add(DFS(key, map, new Array<String>(), new Array<String>(), len))
6. end for
7. function DFS(String key, Set map, List marked, List sList, int[] len)
8. if(marked.contains(key))
9. if(sList.size > len[0])
10. maxList = new ArrayList(sList)
11. len[0] = sList.size()
12. end if
13. else
14. marked.add(key)
15. sList.add(key)
16. Set values = map.get(key)
17. for(each value : values)
18. DFS(value, map, marked, sList, len)
19. end for
20. slist.remove(key)
21. end else
22. return maxList
23. end DFS

END

由算法4.3可知，以每个城市为起点对污染物传播网络进行深度优先遍历，其中在深度优先遍历时，用表示站点是否被访问过，如果被未访问过，则将该节点存储至序列中，并将该节点标示为已被访问状态（13-14行），并继续进行深度优先遍历（18行）。否则，将序列输出。最终，将可获取以每个城市为起点的所有大气污染物传播路径。

### 4.3.2 构建有向图矩阵

由算法4.4获取的大气污染物传播路径即为大气污染物传播网络有向图的子图，然后根据子图中的信息来构建定义4.4中的有向图矩阵APTM。首先，获取子图中所有的节点。其次，顺序遍历所有子图，将站点编号看作矩阵的行列号，如果子图中有对应边的存在，则在对应子图的位置上置为1，否则置为0。再次，在所有的子图中，可能某条边在多个子图中出现过，所以将所有子图对应边位置上的0/1值拼接起来形成新的字符串。最后，将刚才得到的字符串序列存储到矩阵的对应位置中，即形成了有向图矩阵APTM。其具体的算法过程如下所示。

算法4.4 构建有向图矩阵

输入：

输出：APTM

BEGIN



END

在算法4.4中，是所有大气污染物传播网络的子图的集合，是根据中的三列构造的邻接矩阵。首先，初始化矩阵，用来存储所有子图中涉及到的节点（2-8行），根据中的节点来初始化矩阵，节点编号作为行列号。其次，将每个大气污染物传播网络子图转换成邻接矩阵（12-16行），并将每个矩阵中的元素转换成字符，把每个每个矩阵中对应行列位置的字符拼接起来，形成字符串序列（17-18行）。最后，将所有的序列按照对应的行列存储到矩阵中（19行）。算法结束。

### 4.3.3 构建简化图矩阵

对算法4.4构造出的大气污染物传播网络有向图矩阵按照规则依次对每一项进行判断筛选，从而获得简化图矩阵。首先，统计出矩阵中每一项中1出现的次数，通过预先设置的频繁阈值对每条边进行判断，如果超过了频发阈值，则保留，否则，将该边对应项中的每个为都置为0。最后，将矩阵中的所有二进制数转换为十进制，为后续挖掘污染物关键传播路径提升效率做准备。

算法4.5 构建简化图矩阵

输入：矩阵和频繁阈值

输出：

BEGIN

1. //表示计算中1的个数
2. 置0

5. 置0

END

在算法4.5中，是对序列中1的个数的统计方法。首先，对有向图矩阵进行遍历搜索判断，对每个序列中1的个数进行统计判断（2行），若的值小于预先设置的频繁阈值,则认为该边不是频繁出现的，则将置为0，并将该序列从向图矩阵中移除。若大于，则判断是否至少t个连续路径中出现1，如果没有，则认为其不是频繁出现的，将置为0（3行）。否则，将序列转换成十进制数并根据对应的行列号存入简化图矩阵中（4行），最后得到简化图矩阵。算法结束。

### 4.3.4 挖掘污染物关键传播路径

大气污染物的关键传播路径是由若干符合频繁阈值要求并有周期规律的路径片段组成。通过算法4.5获得污染物传播网络中的频繁出现的边，并根据这些边生成了简化图矩阵。然后采用深度优先遍历的方法对矩阵进行遍历，搜索数值相同的路径，从而得到该路径片段周期出现的次数。为了更加体现得到的路径片段的重要性以及关键性，本文中预先设定了路径片段周期出现次数的阈值,即由超出阈值的路径片段组成的路径为大气污染物关键传播路径。具体的算法实现如下所示。

算法4.6 污染物关键传播路径挖掘算法

输入：，

输出；

Begin

10. 将中的序列按顺序拼接成字符串

END

算法4.6 是对矩阵中每一个路径片段进行判断是否呈现出周期性性来获得污染物关键传播路径。首先，对矩阵进行遍历，判断矩阵中的路径片段是否为0，如果为0，则判断下一条路径片段。否则将该路径片段存储到数组a中（2-4行），并使用函数去除矩阵a中的重复路径片段。其次，如果数组a不为空，则对矩阵进行深度优先遍历，判断是否存在有与数组a的数值相同的路径片段。如果存在，则将该路径片段对应的行列号存储到字符串q中，加1，否则将当前值从数组a中删除。然后继续依据上述方法，继续遍历数组a其他的值（7-12行）。最后，计算每个的值，如果小于1，则删除对应的（13-15行）。最后将满足条件的拼接成字符串即为污染物关键传播路径。算法结束。

### 4.3.5 实例说明

接下来对以衡水地区为起点的污染物传播网络进行关键传播路径的挖掘。首先对一个周期内的污染物传播网络图进行子图挖掘，根据算法4挖掘出以衡水地区为起点的所有传播路径，共挖掘出10条传播路径，其中共有30个节点，这些路径组成了污染物传播子图集合。

由于数据量较大，因此本文只展示了其中的4条传播路径，如下表所示。

表4-1 传播路径节点序列

|  |  |
| --- | --- |
| 序号 | 路径 |
| 1 | 23012，23011，23006，23009，21011，21009，21006，22007，22009，22005，17011，1005，1007，1024，1027，1028，1032 |
| 2 | 23007，23010，23005，23013，23006，23009，21011，21009，21006，,22007，22009，22005，17011，1005，1007，1024，1027，1028，1032 |
| 3 | 23010，23005，23013，23006，23009，21011，21009，21006，22007，22009，22005，17011，1005，1007，1024，1027，1028，1032 |
| 4 | 23004，17024，17027，17007，17009，17015，17022，22009，22005，17011，1005，1007， 1024，1027，1028，1032 |

如表4-1所示，由于每条传播路径的长度各不相同，其中包含的节点也不同，因此需要将每条传播路径转换邻接矩阵，矩阵的大小应该由节点的个数决定。因此，该邻接矩阵是一个大小的矩阵，然后根据每条路径片段在每个子图中的出现情况，进一步生成以石家庄地区为起点的污染物传播网络有向矩阵APTM，矩阵中的每一项表示该路径片段在每个子图中出现的情况。由于矩阵过大，因此只展示部分矩阵中的数据，如表4-2所示。

表4-2 大气污染物传播网络有向图矩阵APTM

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 23011 | 23006 | 23009 | 21011 | 21009 | 21006 | 22007 | 22009 | 22005 | 17011 |
| 23011 | 00000  00000 | 10000  00100 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 |
| 23006 | 00000  00000 | 00000  00000 | 11101  11100 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 |
| 23009 | 00000  00000 | 00000  00000 | 00000  00000 | 11101  11101 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 |
| 21011 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 11101  11101 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 |
| 21009 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 11101  11101 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 |
| 21006 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 11101  11101 | 00000  00000 | 00000  00000 | 00000  00000 |
| 22007 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 11101  11101 | 00000  00000 | 00000  00000 |
| 22009 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 11111  11111 | 00000  00000 |
| 22005 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 11111  11111 |
| 17011 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 | 00000  00000 |

通过算法4-6来统计有向图矩APTM中每一项里1的个数，并设置频繁阈值为2，通过统计可知，边<23012景县县委，23011阜城县法院局>、<23007故城县县政府，23010枣强县环保局>、<23004安平县环保局，17024安国市政府>、<23008深州市政府，11016深泽供电局>中只出现了1次1，因此将其所在位置只为0，然后将剩余的边序列由二进制转换为十进制数，所生成的矩阵即为简化图矩阵。具体结果如下表所示。

表4-3 简化图矩阵

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 23011 | 23006 | 23009 | 21011 | 21009 | 21006 | 22007 | 22009 | 22005 | 17001 |
| 23011 | 0 | 516 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23006 | 0 | 0 | 956 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23009 | 0 | 0 | 0 | 957 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21011 | 0 | 0 | 0 | 0 | 957 | 0 | 0 | 0 | 0 | 0 |
| 21009 | 0 | 0 | 0 | 0 | 0 | 957 | 0 | 0 | 0 | 0 |
| 21006 | 0 | 0 | 0 | 0 | 0 | 0 | 957 | 0 | 0 | 0 |
| 22007 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 957 | 0 | 0 |
| 22009 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1023 | 0 |
| 22005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1023 |
| 17001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

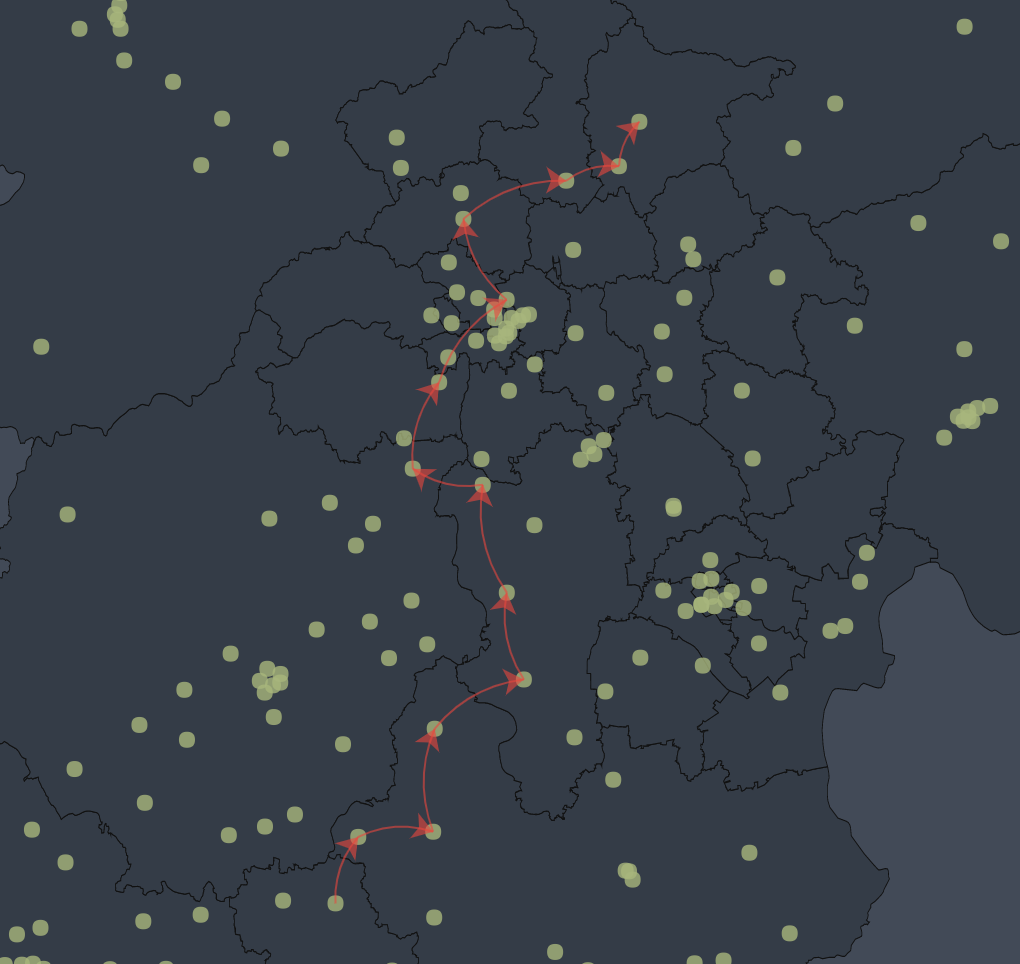
在表4-3中，经过频繁阈值的筛选和十进制数转换之后，数据的分布情况即可清晰可见。在简化图矩阵中，经过算法6的筛选之后，存在着相同二进制数的边，其值分别为957、1023、66、516、956、432、384、400。其中各个值对应的边的情况如下表所示。

表4-4 各值对应路径片段

|  |  |
| --- | --- |
| 值 | 路径片段 |
| 957 | 23009，21011、21011，21009、21009，21006、21006，22007、22007，22009 |
| 1023 | 22009，22005、22005，17011、17011，1005、1005，1007、  1007，1024、1024，1027、1027，1028、1028，1032 |
| 66 | 17024，17027、17027，17007、17007，17009、17009，17015、  17015，17022、17022，22009、 |
| 516 | 23011，23006、 |
| 956 | 23006，23009 |
| 432 | 23013，23006 |
| 384 | 23010，23005 |
| 400 | 23005，23013 |

由表4-4可知数值为516、956、432、384、400的边只有一条，设,则去除数值为516、956、432、384、400的边。最后以衡水地区的监测站为起点，将路径片段组合起来，结果可得污染物关键传播路径为：

23009饶阳县环保局，21011肃宁县政府，21009河间市交通局，21006任丘华油八处，22007文安县环保局，22009霸州顺达燃气，22005固安党校，17011涿州监测站，1005房山良乡，1007朝阳奥体中心，1024昌平镇，1027怀柔镇，1028密云镇，1032京东北密云水库，该大气污染物的关键传播路径图如图所示。



由图\*可以看出，该污染物关键传播路径由衡水地区出发，途径沧州、廊坊、保定最终到达北京地区。

## 4.4 实验结果分析

### 4.4.1 实验环境配置

实验的系统环境配置为macOS 10.13.2操作系统，硬件环境配置为Intel Core i5，主频为3.10GHz，内存为16 GB 2133 MHz LPDDR3。算法采用Java，Matlab编写。数据展示采用Echarts.js以及Excel，运行环境为IntelliJ IDEA，JDK版本为1.7，采用的数据库为MySQL数据库。

### 4.4.2 实验数据

本文采用的实验数据为4.2节中产生的一个周期内的大气污染物传播网络的邻接矩阵以及京津冀地区2014年5月1日到2015年4月30日的空气质量数据集airquality.csv。

### 4.4.3 实验结果分析

由于本文中将小时作为一个时刻，将一年作为一个周期。因此，在一个周期中就会有8760个时刻，并且每个时刻中会出现若干条污染物传播路径，因此数据量是十分巨大的，所以本文只展示部分实验数据。通过对京津冀地区2014年5月1日到2015年4月30日期间每个站点之间的污染物传播代价进行计算，并最终计算出了每条边的平均传播代价，如表4-5所示。

表4-5 周期内站点间平均传播代价

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 监测站名称 | 海淀北  部新区 | 海淀北京植物园 | | 石景山古城 | 丰台  云岗 | 房山  良乡 | 海淀  万柳 | 朝阳奥  体中心 |
| 海淀北部新区 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 |
| 海淀北京植物园 | 0.6456 | 1 | 1 | | 1 | 1 | 1 | 1 |
| 石景山古城 | 0.7369 | 0.6476 | 1 | | 1 | 1 | 0.6736 | 0.6424 |
| 丰台云岗 | 0.7675 | 0.745 | 0.6779 | | 1 | 1 | 0.7486 | 0.8037 |
| 房山良乡 | 0.9308 | 0.8633 | 0.7853 | | 1 | 1 | 0.8557 | 0.687 |
| 海淀万柳 | 0.7263 | 1 | 1 | | 1 | 1 | 1 | 1 |
| 朝阳奥体中心 | 0.7409 | 0.7536 | 1 | | 1 | 1 | 1 | 1 |

通过表4-5可以看出，该表不是一个对称结构的表格，也就是说明只有在特定因素条件的制约下，相应的路径才能成为比较有利于污染物传播的一条路径，而在其他的条件下可能不利于污染物的传播。并通过图0展示出周期内京津冀地区污染物传播网络图：

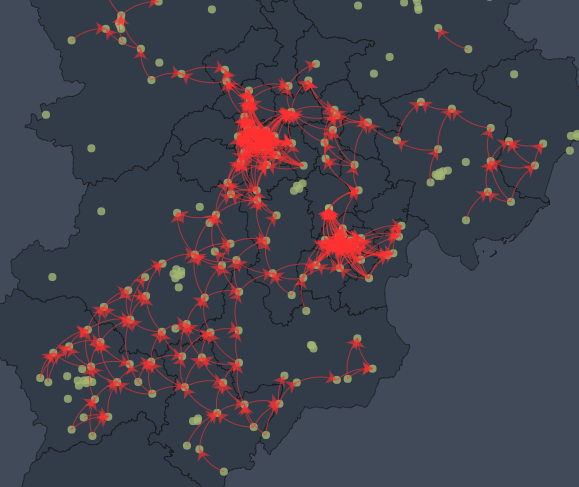


图0 周期内京津冀污染物传播网络图

由图0可知，北京和天津地区城市内部污染物传播路径十分的复杂以及繁多，而其他城市的污染物传播路径则稀疏许多。这是因为这两个城市的经济与其他城市相比比较发达，城市群落比较密集，高层建筑较多，因此地面粗糙度比较高，这使得气流沿地表流过时，同地面的建筑物发生摩擦作用，使风向风速发生变化，因此导致了这两个城市内部形成了复杂的污染物传播网络。

通过算法4.2统计出一个周期内京津冀各个城市的污染物超标率，如图1所示。由图可见廊坊、衡水、石家庄、唐山、沧州、保定六个城市的超标率超过了65%。

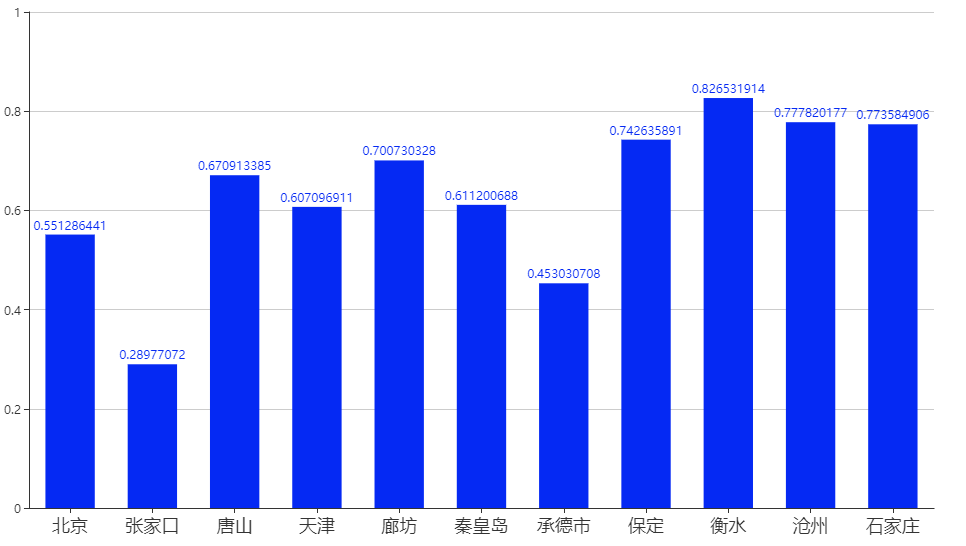


图1 京津冀各城市污染物超标率

由图1可知廊坊、衡水、石家庄、唐山、沧州、保定这几个地区的空气污染比较严重，而秦皇岛、北京、承德、天津空气污染情况相对较轻，只有张家口一个城市污染物超标率是低于30%的，是空气质量最好的城市。这些数据的统计结果与近几年京津冀各城市的空气质量状况十分吻合，由此证明了统计结果的准确性。因此，选取廊坊、衡水、石家庄、唐山、沧州、保定这几个地区中的监测站为起点，对污染物关键传播路径进行挖掘。

首先，根据算法4.2计算出以每个地区为起点的污染物传播子图个数以及节点个数。经计算，各地区的污染物传播子图个数以及节点个数情况如下表所示：

其次，根据算法4.4和算法4.5计算出每个地区污染物传播网络的有向图矩阵以及简化矩阵。但是，由于每个矩阵中的数据量十分巨大，因此本章不再对污染物传播子图、构建有向图矩阵以及构建简化矩阵的数据进行展示，具体的计算过程可见上一章的算法实例说明，本章只对符合频发阈值以及呈周期出现的边进行展示，如表4-6~4-11所示。

表4-6 以廊坊地区监测站为起点的路径片段表

|  |  |
| --- | --- |
| 值 | 路径片段 |
| 184 | 22010，1023、1023，1024、1024，1031、1031，19009 |
| 71 | 1024，1027、1007，1024、1027，1028、1028，1032 |
| 69 | 22005，17011、17011，1005、1005，1007 |

表4-7以衡水地区监测站为起点的路径片段表

|  |  |
| --- | --- |
| 值 | 路径片段 |
| 957 | 23009，21011、21011，21009、21009，21006、21006，22007、22007，22009 |
| 1023 | 22009，22005、22005，17011、17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032 |
| 66 | 17024，17027、17027，17007、17007，17009、17009，17015、17015，17022、17022，22009 |

表4-8 以石家庄地区监测站为起点的路径片段表

|  |  |
| --- | --- |
| 值 | 路径片段 |
| 400 | 11016，17024、17024，17027 |
| 65535 | 17027，17007、17007，17009、17009，17015、17015，17022、17022，22009、22009，22005、22005，17011、17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032 |

表4-9 以唐山地区监测站为起点的路径片段表

|  |  |
| --- | --- |
| 值 | 路径片段 |
| 168 | 13018，6015、6015，1033、6015，1033、6015，1033、6015，1033、6015，1033 |

表4-10 以沧州地区监测站为起点的路径片段表

|  |  |
| --- | --- |
| 值 | 路径片段 |
| 1729 | 21006，22007、22007，22009、22009，22005、22005，17011、17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032 |

表4-11 以保定地区监测站为起点的路径片段表

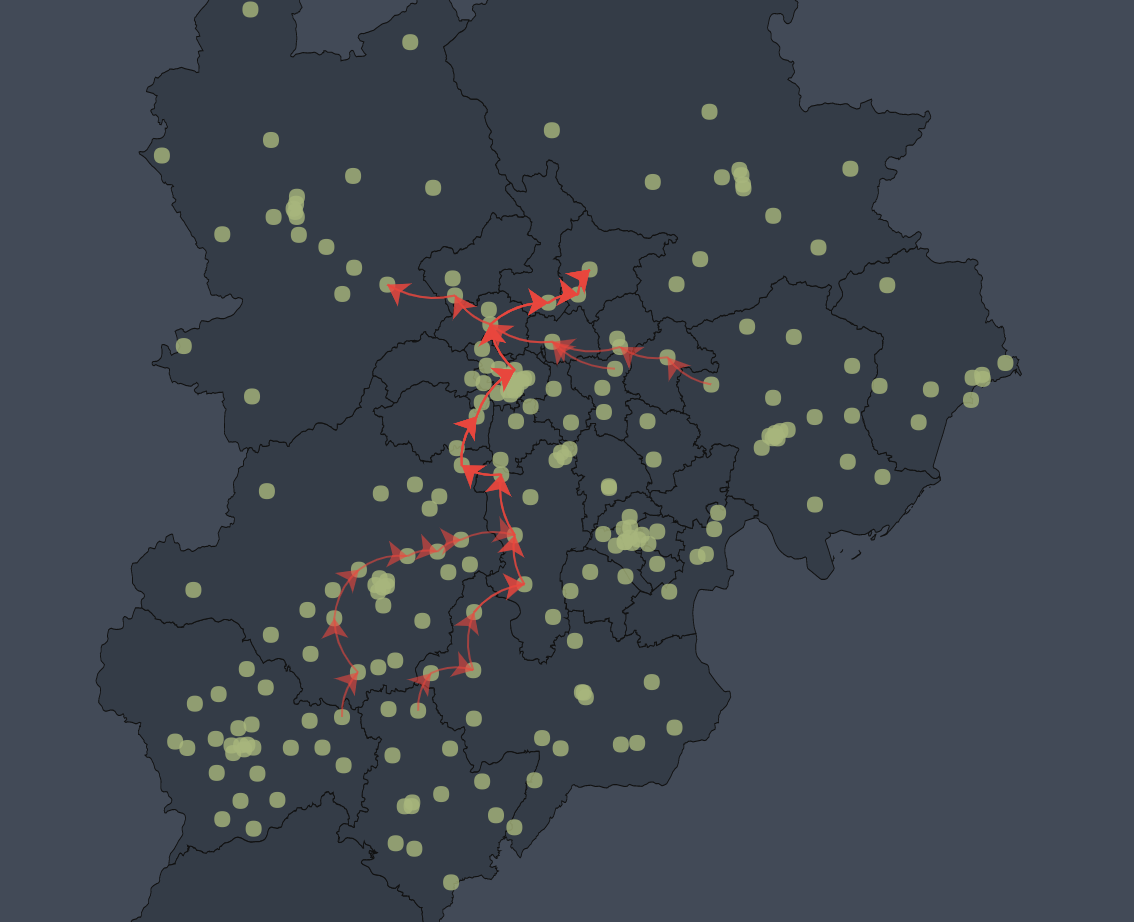
|  |  |
| --- | --- |
| 值 | 路径片段 |
| 30207 | 22009，22005、22005，17011 |
| 32767 | 17011，1005、1005，1007、1007，1024、1024，1027、1027，1028、1028，1032 |

根据表4-6~4-11中的数据，以每个地区中的监测站为出发点，通过算法4.7可计算污染物关键传播路径。如下表所示。

表4-12 个城市污染物关键传播路径

|  |  |
| --- | --- |
| 地区 | 关键传播路径 |
| 廊坊 | 22010，1023，1024，1031，19009 |
| 22005，17011，1005，1007，1024，1027，1028，1032 |
| 衡水 | 23009，21011，21009，21006，22007，22009，22005，17011，1005，1007，1024，1027，1028，1032 |
| 石家庄 | 11016，17024，17027，17007，17009，17015，17022，22009，22005，17011，1005，1007，1024，1027，1028，1032 |
| 唐山 | 13018，6015，1033，1023，1024，1031，19009 |
| 沧州 | 21006，22007，22009，22005，17011，1005， 1007，1024，1027，1028，1032 |
| 保定 | 17011，1005，1007，1024，1027，1028，1032 |

最后，通过Echarts将各地区的污染物关键传播路径绘制在地图上，如下图所示。



其中，红色路线为各城市的污染物关键传播路径，其中部分路径存在重叠部分。由上图可以看出，在京津冀地区的西南部，污染物沿着太行山脉的走势由石家庄地区将污染物输送至北京地区，途中与保定、廊坊地区的污染物传播路径想汇合；东南方向污染物由沧州和天津输送至北京地区，途中与廊坊地区的污染物传播路径相汇合。东北方向由唐山地区沿燕山山脉传播至北京地区，途中与天津地区的污染物传播路径相汇合。由此可见，此污染物的关键传播路径与实际中的地理特征以及气象特征基本吻合，即证明了此挖掘关键路径方法的科学性以及准确性。

## 4.5本章小结

本章研究内容是在上一章大气污染物传播机理以及传播代价公式的基础上进行的研究，主要对大气污染物传播网络的关键路径进行了挖掘。首先，构建大气污染物传播网络，并描述了传播网络中的基本概念以及关键传播路径的定义。其次。统计出一个周期内污染物超过超标率的城市，并以这些城市中的监测站为起始点，对传播网络进行深度优先遍历，获得若干个大气污染物传播网络子图。再次，详细介绍了挖掘大气污染物关键传播路径的详细的原理以及过程，并给出了相应的算法实例。最后，对挖掘关键路径的试验结果进行了展示以及分析。

第5章 基于NodeRank的污染物传播网络脆弱点挖掘

## 5.1 引言

通过第三章提出的大气污染物传播机理模型，构建出了一个周期内的京津冀地区的大气污染物关键传播路径网络，并证实了该网络符合复杂网络的基本特性。在大气污染物关键传播路径网络中，挖掘网络中脆弱点，即挖掘京津冀地区内最易受到大气污染物污染的地区。挖掘京津冀地区内最易受到大气污染物污染的地区的意义在于通过挖掘网络中的脆弱点，有助于在该地区设置更加完善的大气污染物监测体系，并有效的治理该地区的污染状况，实现有效的大气污染区域联防；也有利于在准确的位置设置空气质量监测站，从而大大减少了由于建设监测站带来的成本。

实际上，几乎所有的复杂系统，比如社会、生物、信息、交通运输系统都可以用复杂网络来表示。其中，复杂网络中的关键节点就是相比网络中其他节点而言，可以更大程度上影响网络的功能与结构的一些特殊节点。通常来说，关键节点的数量只是一小部分，但是这些节点却可以非常快的影响到网络中的大部分节点。如下列举了几种复杂网络基础上最简单直观的方法：

度中心性：认为网络中节点的邻居节点数目越多，其影响力就越大。即在网络图计算节点的度数（入度+出度）来衡量网络中节点的中心度，如果网络为加权网络，则通过计算节点的度强度来衡量节点的中心度。

特征向量中心法：认为节点的不仅仅取决于自身邻居节点的数目，也取决与每个邻居节点的重要性，即节点上的权重。因此可用节点邻居节点权重之和来衡量一个节点的重要性。

节点删除法：破坏性反映重要性。通过删除网络中的节点，来判断该节点的失去对网络的破坏程度，从而衡量该节点的重要性。

节点收缩法：将一个节点和它相连的节点进行收缩，形成一个新节点[93]，如果该节点十分重要，收缩后整个网络将会更好的聚集在一起，从而以此标准来衡量节点的重要性。

以上列举的网络关键节点的识别办法主要是应用于无向网络中，然而本文所要研究的大气污染物传播网络是一个有向加权网络，因此不适用于大气污染物传播网络网络。而PageRank算法应用于有向网络挖掘关键节点是十分有优势的，但是由于大气污染物传播网络是一个有向加权网络，而且每个站点的重要性受周围各站点的影响很大但不是等概率分配的，因此需要对原始的PageRank算法进行改进。最终，本文在原始PageRank算法的基础上提出了NodeRank算法，该算法结合了相邻边的权重以及邻居节点的重要性。

## 5.2 基于的污染物传播网络脆弱点挖掘

### 5.2.1 PageRank算法理论

PageRank简化计算模型[44]公式（5-1）如下所示：

|  |  |  |
| --- | --- | --- |
|  |  | (5-1) |

公式（5-1）中，表示页面的重要度，即值；表示每个链接到页面的页面的值；表示页面中所有页面链接的数量；表示所有链接到页面的页面集合。

根据PageRank理论[45,46]可知，PageRank的计算公式（5-2）如下：

|  |  |  |
| --- | --- | --- |
|  |  | (5-2) |

公式（5-2）中，表示页面的值；表示每个链接到页面的页面的值；表示页面中所有页面链接的数量；表示所有链接到页面的页面集合；表示阻尼系数，取值为；表示网页总数。

阻尼系数表示用户到达某个页面后继续访问其他页面的概率，页面中的值只有部分在其可以链接到的页面中进行分配，其余的部分分配在在整个网络的页面中，通常设置为0.85[47]。也就是一个页面的的值来自于能够链接到页面的某一个页面，剩下的则可能来自用户通过从浏览器地址栏输入URL访问页面。

### 5.2.2 PageRank算法改进

在现实世界中，社会关系网络、飞机航线网络、流行病毒网络等都为加权有向网络，PageRank算法在这些网络中都有这重要的应用。在本文构建的大气污染物传播网络中，空气质量监测站为网络中的节点，两个站点之间的污染物传播路径为网络中的边，边上的权重为一个周期内站点之间大气污染物传播代价的平均值。由此可见，大气污染物传播网络也是一个有向加权网络。但是，由于原始的PageRank算法无法满足于本文构建的大气污染物传播网络。因此，本文提出一种PageRank算法改进后的NodeRank算法应用于大气污染物传播网中。此算法不仅考虑了大气污染物传播的方向性、站点之间的传播代价，而且对PageRank算法中的转移概率以及N做出了适用于大气污染物传播网络的调整。

首先，由于大气污染物传播网络不同于一般的复杂网络，由于地理以及气象因素的限制，导致大气污染物无法进行远距离的输送，因此，在本文的3.3.1节中提出了约束大气污染物传输的5个限制条件。例如，污染物可能会因为两个站点之间距离十分遥远导致永远无法相互传输、可能会因为山脉的原因导致永远无法进行相互传输等，即在大气污染物传播网络中会存在若干个节点之间不会存在边的情况。因此，公式（5-2）中的表示网络中全部的节点数目就有些不恰当，因此本文将修改为站点可达范围内的所有的站点的数量。

其次，传统PageRank算中的从页面到页面转移概率是按照公式（5-3）进行的平均分配的。

|  |  |
| --- | --- |
|  | （5-3） |

公式（5-3）中，表示从页面到页面转移概率，表示链入页面的链接数目。因为大部分的页面的PageRank值是不同的，因此仅通过链入页面的链接数来计算转移概率是不合理的。PageRank值高的页面链接，它在页面链接中的排序也应该是靠前的，从而更容易被其他的页面访问到。因此提出新的转移概率公式，如公式（5-4）所示。

|  |  |
| --- | --- |
|  | （5-4） |

公式（5-4）中，表示页面的PageRank值，表示页面中所链接页面集合。但是从公式（5-4）中可以看出，与公式（5-3）的情况类似，仅仅依据PageRank值占总PageRank值的百分比来决定转移概率也是不合理的。因为有的页面虽然PageRank值很高，但是可能它会有很多的出链，因此可能会导致页面的转移概率比出链少的页面转移概率要低。

结合本文构建的大气污染物传播网络，可根据每对站点之间的传播代价来衡量站点之间污染物传播的难易程度，因此在此基础上结合公式（5-4），本文最终提出一种新的转移概率公式，如公式（5-6）所示。

|  |  |
| --- | --- |
|  | （5-5） |
|  | （5-6） |

公式（5-5）中，表示污染物从站点传播到站点之间的传播代价，越大表明越不易传播，反之容易传播。由于大气污染物的传播概率氛围两部分来确定：传播代价所占比、站点PageRank值所占比。考虑到传播代价所占比对污染物的传播概率起到的影响更大，因此将调整参数设置为0.7。该公式考虑到了站点之间的权值，同时也考虑到了站点的PageRank值，也保证了矩阵每行之和始终为1。因此，相比之前的转移概率公式，更加具有合理性。

因此最终得出了评价节点重要性的指标NodeRank值的计算公式为：

|  |  |
| --- | --- |
|  | （5-7） |

当将公式（5-7）应用于大气污染物传播网络时，表示监测站的NodeRank值；表示大气污染物输送路径i中监测站的NodeRank值；表示转移概率；表示大气污染物能够输送到站点的数目。

### 5.2.3 NodeRank算法实现

NodeRank算法以大气污染物传播网络为基础，对网络中的所有监测站进行打分排序。NodeRank算法的主要步骤如下：

算法 5.1 NodeRank算法

输入：大气污染物传播网络

输出：站点评分排序

BEGIN

（1）输入大气污染物传播网络

（2）根据公式（5-5）计算站点到站点的转移概率，并将每个值存储至矩阵

（3）根据公式（5-6）计算NodeRank矩阵

（4）使用幂法迭代[49]求解NodeRank矩阵中特征值为1对应的特征向量，即

平稳分布

（5）对结果进行节点评分排序

END

其中，第4步中求解NodeRankk矩阵平稳分布的幂法迭代的算法过程如下：

算法5.2 Pagerank-Interate

输入：NodeRank矩阵，转移概率矩阵

输出：NR

BEFIN

1. repeat

4. until
5. return

END

算法5.2 中初始每个站点的NodeRank依旧按照原始PageRank进行计算，若两次迭代后差值小于临界值，则停止继续迭代。

## 5.3 实验结果与分析

根据算法5.1的步骤，对京津冀地区空气污染物传播网络应用算法NodeRank，对其中的站点进行评分排序，并挖掘出网络中的脆弱点。表5-1为算法NodeRank评分前20的空气质量监测站，即脆弱点。

表5-1 排名前20空气质量监测站NodeRank值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 排名 | 站名 | NodeRank | 排名 | 站名 | NodeRank |
| 1 | 京西南琉璃河 | 0.03207 | 11 | 安平县环保局 | 0.01893 |
| 2 | 无极环保局 | 0.02705 | 12 | 大兴黄村镇 | 0.01783 |
| 3 | 满城税务局 | 0.02556 | 13 | 徐水环保局 | 0.01755 |
| 4 | 新乐市委东楼 | 0.02469 | 14 | 栾城通讯公司 | 0.01597 |
| 5 | 深泽供电局 | 0.02466 | 15 | 晋州博纳德 | 0.01563 |
| 6 | 高邑县政府 | 0.02315 | 16 | 赵县环保局 | 0.01503 |
| 7 | 正定联通公司 | 0.02084 | 17 | 房山良乡 | 0.01499 |
| 8 | 行唐县委办公楼 | 0.02060 | 18 | 任丘华油八处 | 0.01492 |
| 9 | 元氏住建局 | 0.01988 | 19 | 天津团泊洼 | 0.01466 |
| 10 | 亦庄开发区 | 0.01941 | 20 | 藁城实验学校 | 0.01461 |

根据表5-1中的排名前20空气质量监测站，统计出每个监测站一年的污染物超标率如下表所示

表5-2 脆弱点污染物超标率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 排名 | 站名 | 污染物超标率 | 排名 | 站名 | 污染物超标率 |
| 1 | 京西南琉璃河 | 44.08% | 11 | 安平县环保局 | 53.65% |
| 2 | 无极环保局 | 60.48% | 12 | 大兴黄村镇 | 36.34% |
| 3 | 满城税务局 | 57.11% | 13 | 徐水环保局 | 53.45% |
| 4 | 新乐市委东楼 | 56.46% | 14 | 栾城通讯公司 | 53.38% |
| 5 | 深泽供电局 | 55.69% | 15 | 晋州博纳德 | 52.12% |
| 6 | 高邑县政府 | 55.31% | 16 | 赵县环保局 | 51.58% |
| 7 | 正定联通公司 | 54.77% | 17 | 房山良乡 | 34.44% |
| 8 | 行唐县委办公楼 | 55.40% | 18 | 任丘华油八处 | 49.64% |
| 9 | 元氏住建局 | 54.28% | 19 | 天津团泊洼 | 34.01% |
| 10 | 亦庄开发区 | 36.98% | 20 | 藁城实验学校 | 52% |

由表5-2可以看出，虽然监测站点的污染物超标率部分不符合NodeRank值的排序，但是，总体上监测站点的污染物超标率的趋势是与NodeRank值的排序的趋势是相同的。

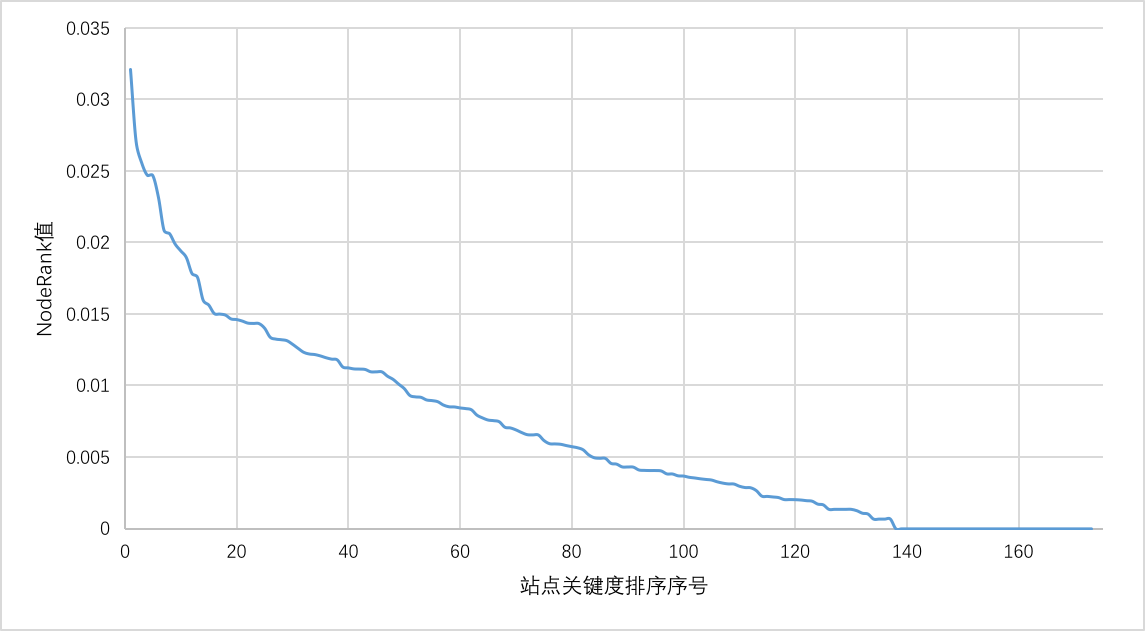


图5-1空气质量监测站NodeRank值排序

图5-1为通过NodeRank算法计算出的空气质量监测站NodeRank值排序的曲线图，横坐标为站点关键度排序编号，纵坐标为各站点的NodeRank值。

然后，根据NodeRank值排序后的站点编号，对每个站点的入度强度进行统计，

如图5-2所示，横坐标为站点关键度排序编号，纵坐标为各站点的入度强度。

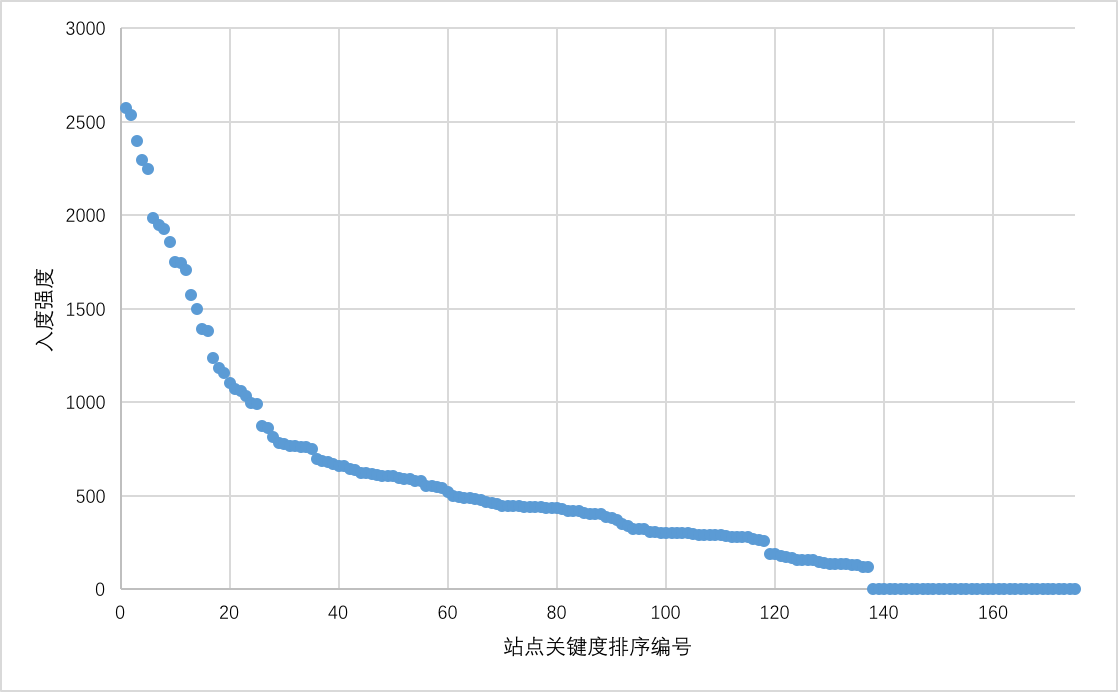


图5-2 空气质量监测站入度强度

由图5-1和图5-2可知，节点入度强度的增大，尽管部分节点出现异常，但大部分站点的NodeRank值也随之增大，两者基本符合正相关系，因此证明了NodeRank算法的正确性。

## 5.4 本章小结

本章主要是在大气污染物传播网络图的基础上对网络中的脆弱点进行了挖掘。首先，对大气污染物传播网络中的节点进行了分析，并列举了常见的几种关键节点评价指标；其次，简述了传统PageRank算法，并根据构建的大气污染物传播网络，对传统PageRank算法中的节点数和转移概率两方面进行了改进，最终提出了算法NodeRank,并详细描述了算法NodeRank的实现过程；最后，给出了试验结果并对试验结果进行了分析，并与实际各站点污染物超标率进行了对比，与各站点的入度强度分布进行了对比，发现它们分布情况十分接近，因此证明了挖掘出的脆弱点的有效性。

结 论

随着我国经济的飞速发展，大气污染物问题越来越引起了人们的重视，而大气污染物传播有着明显的区域性特点。因此，有效的防治大气污染物传播，改善区域内的空气质量就显得尤为重要。本文在京津冀区域内空气质量监测站以及气象监测站历史数据的基础上，做出了以下研究：

（1）提出了一种基于大气污染物传播机理的污染物传播模型。由于无法获取详尽的污染源清单，因而无法采用传统的气体扩散模型来分析污染物的传播规律。因此，分析大气污染物的传播机理，筛选出影响污染物传播的主要因素，如风、降雨量、海拔、水平距离等因素。并以此为依据，构建大气污染物传播模型，其值为大气污染物传播代价。

（2）提出了一种基于图矩阵的大气污染物关键传播路径的方法。首先，分析污染物传播约束条件，并在此基础上结合污染物传播模型，获得污染物传播路径序列，并将所有序列映射到污染物传播网络，经验证该网络符合复杂网络的特征。然后，统计出周期内污染物超标的城市，以这些地区作为起点对污染物传播网络进行深度优先遍历，得到污染物传播的路径序列。最后，将每天路径序列映射到图矩阵中，在此基础上，根据预先设置的频繁阈值和重复次数，筛选出符合要求的边，并将这些边拼接起来即为污染物关键传播路径。此方法时间复杂度较低，而且挖掘出的路径具有代表性，也基本符合京津冀地区的地理特征和气象规律。

（3）提出了一种基于PageRank算法改进的适用于空气污染物传播网络的NodeRank算法。NodeRank算法不仅考虑了站点之间的传播代价，而且对PageRank算法中的转移概率做出了改进，充分考虑了邻居站点的重要性。最后通过对每个站点进行评分排序，最终得出的排名靠前的站点即为污染物传播网络中的脆弱点。并将站点排序结果与实际各站点的污染物超标率和入度强度进行对比，发现有很好的同步性。

由于时间和精力有限，本文虽然取得了一些研究成果，但是仍然存在值得改进的地方，因此下一步的工作将主要从以下方面进行研究：

（1）由于本文只是研究了京津冀地区的污染物关键传输路径以及脆弱点，没有对其他区域进行研究，因此，为了进一步验证本文提出的污染物传播模型的广泛性以及适应性，接下来将着重研究其他区域、特殊地形下的污染物传输规律，来进一步完善污染物传播模型。

（2）大气污染是一个极其复杂的过程，不仅仅与污染源有关，大部分的来源污染物的二次污染，也就是污染物之间发生化学反应，再次生成大气污染物。因此，还需要从化学原理上对大气污染进行分析，才能更加准确的找到防治大气污染的有效方法。

参考文献

1. 戴红. 康复医学[M]. 北京: 人民卫生出版社, 1998: 3-10.
2. Krebs H I., Palazzolo J, Dipietro L, et al. Rehabilitation Robotics: Performance-Based Progressive Robot-Assisted Therapy[J]. Autonomous Robots, 2003, 15(1): 7-20..

攻读硕士学位期间承担的科研任务与主要成果

**（一）参与的科研项目**

**（二）发表的学术论文**

1. ×××,×××.并联2-RRR/UPRR踝关节康复机器人机构及其运动学[J]. 机器人, 2010, 32(1):6-12 .(EI收录号: 20101212786168)

致 谢

值此论文完成之际，谨向给予我指导、关心和帮助的老师、朋友、同学及家人表示衷心的感谢。

深深感谢我的导师黄国言教授，黄师严谨的治学态度、渊博的专业知识、执着追求的敬业精神和高尚的人格品质都使我受益匪浅。在攻读硕士学位期间，从课题的研究到论文的撰写都得到了导师全面、认真的指导。在学习遇到困难时，黄老师给予了耐心的解答和帮助，还提出了很多宝贵的意见和建议，使我渐渐的熟悉了研究内容。从导师那里我不仅学到了丰富的专业知识和科学的研究方法，更重要的是学到了一丝不苟的工作态度和谦和的为人处世原则。在此，谨向培养我的恩师致以最衷心的感谢和最诚挚的敬意！

感谢组里的郝晨谦、贾越洋、田相敏、刘新倩、何洪豆、齐聪雅等同学，他们在学术研究和项目实践等诸多方面给与我支持和帮助。在大家的共同努力下，实验室不仅是共同科研、一起学习进步的场所，更如一个温暖有爱的大家庭，互帮互助，互相激励，我们在实验室度过的每一天都温馨而充实。

感谢燕山大学对我七年的培养，在这里，我从一个懵懂无知的高中毕业生，成长为一名遇事有担当的青年。在这里，我遇到了人生的良师，结交了人生的挚友，学习到了将来工作中所需的诸多专业技能，在燕山大学浓厚的学术氛围和优美的校园环境中积累了宝贵的知识和精神财富

感谢我的家人，多年来对我无私的帮助和鼎力的支持，不管我遇到任何困难，都无条件的支持我、鼓励我、信任我。

最后，再次感谢所有关心和帮助过我的人们！