

基于实时技术的大气颗粒物在线分析系统^①

潘争光^{1,2}, 赵 奎², 王鸿亮², 王俊霖³

¹(中国科学院大学, 北京 100049)

²(中国科学院 沈阳计算技术研究所, 沈阳 110168)

³(大连理工大学 软件学院, 大连 116024)

摘 要: 针对环境监测中, 难以实时在线处理海量颗粒物数据的问题, 提出了一种基于实时技术的大气颗粒物在线分析系统, 实现了颗粒物统计、浓度变化、来源解析等功能. 该系统利用实时数据库来实时采集、存储海量大气数据, 解决了环境监测中数据的海量问题; 同时, 引入自适应共振神经网络算法和逻辑回归模型进行数据分析, 成功降低数据规模, 提升数据分析速度. 实践表明, 该在线分析系统能在合理时间内得到准确的分析结果, 具有重要的实际意义.

关键词: 实时技术; 海量数据; 聚类分析; 在线分析; 环境监测

Atmospheric Particle Online Analysis System Based on Real-Time Technology

PAN Zheng-Guang^{1,2}, ZHAO Kui², WANG Hong-Liang², WANG Jun-Lin³

¹(University of Chinese Academy of Sciences, Beijing 100049, China)

²(Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

³(School of Software Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: For environmental monitoring, the existing online analysis system is difficult to deal with massive atmospheric particle data. In this paper, we propose an atmospheric particle online analysis system based on real-time technologies, which aims to achieve atmospheric particle statistics, concentration change and the source analysis. The system adopts real-time databases to realize real-time capturing, stores massive atmospheric particle data, and solves the massive data problem in environmental monitoring. Besides, to accelerate data analysis and reduce data scale, the system adopts the ART-2a neural network algorithm and logistic regression model. The experiment results prove that the online analysis system could get accurate analysis result within a reasonable time. Besides, the experiment demonstrates the practical significance of our system.

Key words: real-time technology; massive data; clustering analysis; online analysis; environmental monitoring

1 引言

当前, 大气环境污染问题日益严重, 大气中各种细小颗粒物对人体健康极为有害. 速度开展环境状况及影响分析, 已经是摆在科研人员面前的紧迫任务. 传统的环境数据分析方法是以人工的方式进行数据采集、手动或半自动化地进行数据分析、最后以静态的表格形式存储, 整个过程费时费力; 再者我国经济处于高速发展的阶段, 环境数据更新迅速, 这就对环境数据分析方法提出了时效性要求; 另外一方面, 随着现

在环境监测手段的多样化、自动化, 比如质谱仪每分钟采集到数百的质谱数据, 如何高效、自动化地存储分析数据显然十分必要^[1,2].

本系统利用实时数据库来采集、存储、管理海量颗粒物数据, 再通过相应的数据库接口, 提供给上层数据分析系统, 上层数据分析系统再从海量大气信息中, 通过各种数据分析方法, 实现海量大气颗粒物的统计谱图、粒径分布、浓度变化曲线图、自动命名、来源解析等分析功能, 实时、在线地提取有效信息, 助力

① 基金项目: 国家水体污染控制与治理科技重大专项(2012ZX07505003)

收稿时间: 2016-04-26; 收到修改稿时间: 2016-06-21 [doi:10.15888/j.cnki.csa.005532]

环境中心的空气污染监测。

2 技术概述

2.1 实时技术与实时数据库

实时技术具有低延迟、快速反应、实时处理的特点,实时数据库技术自上世纪 80 年代开始飞速发展,其作为数据库系统的一个分支,是传统数据库技术与实时技术结合的产物,它具有高实时性、高数据吞吐量等特点,是各种信息监测系统的基础,广泛应用在工业生产现场的信息采集、装置监控、历史数据管理^[3],已经在各行业的实时在线系统中扮演重要角色。

2.2 大气颗粒在线分析方法

现行多种大气颗粒物数据分析方法,比如颗粒物总体分析和单颗粒分析^[4]。总体分析是以采集样本的总体作为研究对象,通过 X 射线荧光光谱或者中子活化分析法来对颗粒物中的元素进行检测,分析出样本整体表现出的光谱或化学特征,从而得到总体数据;单颗粒分析法以单个空气颗粒为分析单位,利用空气动力学和光学相关知识,能够对单个颗粒的粒径和化学成分进行分析,精确程度较高,但会产生海量离子、谱图数据。

2.3 颗粒物聚类分类方法

单颗粒分析法在分析大气颗粒物时,会产生海量数据,因此需要利用数据挖掘技术来进行高效的、自动化的数据分析。

数据挖掘中有多种聚类算法,适用于质谱仪数据的聚类算法多采用基于密度的聚类方法,如 K-means 算法、模糊 c 均值、ART-2a(自适应共振神经网络)等,综合考虑大气数据特征与算法效率特征,本系统采用的是 ART-2a 算法。

3 系统设计

3.1 系统需求分析

本系统旨在实现一个基于实时技术的大气颗粒物在线分析系统,主要包含以下功能子模块:数据来源配置、化学成分分析、颗粒物来源解析、查询统计等模块^[5]。具体功能划分如图 1 所示。

3.2 架构设计

本系统采用 C/S 架构,在架构设计上分为表示层、逻辑层和数据层,如图 2 所示。

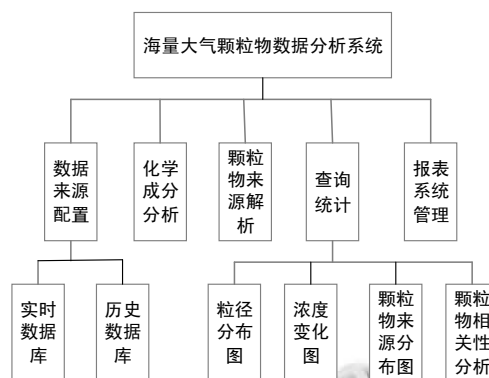


图 1 功能模块图

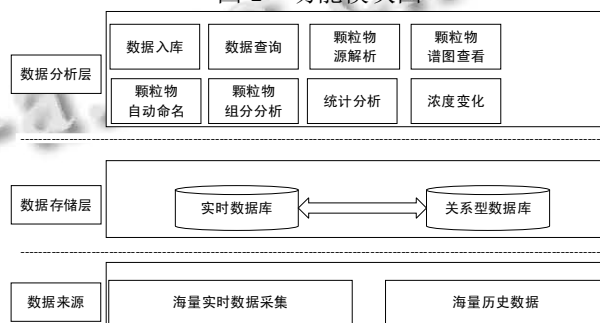


图 2 系统架构图

3.2.1 数据服务层

数据服务层是系统的基石,在本系统中,采用某品牌实时数据库与 MySQL 数据库结合的方式来构建数据采集、存储和管理系统,解决了海量大气颗粒数据监测中的实时性需求和海量数据存储需求,为上层的业务逻辑层提供高速的数据服务。

3.2.2 业务逻辑层

业务层是各种数据分析过程的具体实现,是整个系统的核心部分,包含众多的数据分析过程,如数据导入导出、颗粒物聚类、颗粒物自动命名、统计分析、浓度变化曲线图、颗粒物来源解析等。

图 3 是对空气颗粒物进行来源解析的流程图。

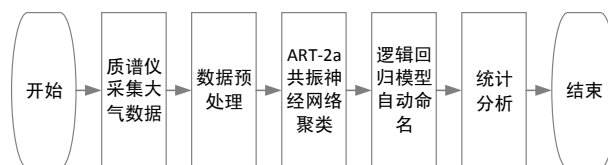


图 3 源解析流程图

4 系统实现

本系统基于 Visual Studio 2010 平台,采用 C++ 语言,使用某品牌实时数据库系统做实时数据采集和存

储,使用 MySQL 数据库来存储数据分析结果和存储本系统的管理数据^[6]。

本系统具有多种优点: (1)界面友好、操作方便; (2)可以对海量大气数据进行快速分析; (3)对分析结果以图表的形式进行展示,并可以对结果进行导入导出; (4)对数据分析的参数可以动态配置; (5)可以对颗粒物聚类结果实现自动命令,高度自动化,节省人力物力; (6)提供管理员等多角色管理功能。

下面主要介绍数据服务层、数据预处理、数据聚类分类分析、数据统计等四个关键模块的实现。

4.1 数据服务层的实现

实时数据采集和历史数据都可以通过相应的 API 或者数据库管理系统来完成,其提供 C/C++二次开发 API 让客户很方便地对实时采集和历史数据进行增删查改,如 goh_get_archived_values 可获取单个标签点一段时间内的存储数据, goh_get_single_value 获取某标签某时间点的数据, goh_update_value 可修改某标签某时间段的数据值, goh_remove_values 可删除某标签一段时间内的数据值, goh_get_cross_section_values 可获取批量标签点的数据值。

用 MySQL 数据库可以存储数据分析后的结果,方便数据的存储、查询、导入、导出。

4.2 数据预处理模块

数据预处理阶段,主要是针对实时采集的颗粒物数据,从中提取到系统数据分析模块所需要的电离离子的峰高、峰面积、相对峰面积等数据,并根据需求选择一个指标作为后续处理的基准。

4.3 颗粒物聚类分析

本系统通过 ART-2a 自适应共振神经网络算法来将相似的颗粒物聚集到同一个分组中,ART-2a 神经网络是一种无监督的矢量分类器,能有效地处理大数据集和高维数据集,大大降低数据的规模。并且当某个数据点与当前存在的所有分类都没有达到预设的相似度时,ART-2a 为其自动产生一个新的类别,而不影响其它已经存在的颗粒物聚类,因此该算法很适用于质谱仪数据聚类分析^[7]。

ART-2a 算法的流程如下:

- 1). 利用数据矩阵,随机初始化输入向量;
- 2). 对输入向量进行归一化处理;
- 3). 计算输入向量与已存在的感知器进行相似度计算,即向量内积计算;

4). 若相似度达到阈值参数,则该颗粒物属于该分组,并进行共振,更新感知器位置;若未达到相似度阈值,则自动产生新类;

5). 将所有的颗粒重复上面的步骤,并进行多轮迭代直到分类结果稳定。

聚类效果如图 4 所示。

分组名称	颗粒数量
No_1_二毛(1)	560
No_1_二毛(2)	417
No_1_二毛(3)	378
No_1_二毛(4)	332
No_1_二毛(5)	325
No_1_二毛(6)	290
No_1_二毛(7)	286
No_1_二毛(8)	275
No_1_二毛(9)	242
No_1_二毛(10)	218

图 4 聚类效果示意图

4.4 颗粒物自动命名

本模块给出了基于逻辑回归模型的分系统,来实现对颗粒物自动命名。主要思路是:以离子信息的峰高、峰面积、相对峰面积等参数作为特征值,通过训练样本来调整分类器的参数,在训练样本充足的情况下,会得到相应颗粒物的回归模型,随着后期用户的不断反馈,进行在线学习,可以进一步更新模型参数,提高分类精确度^[8]。

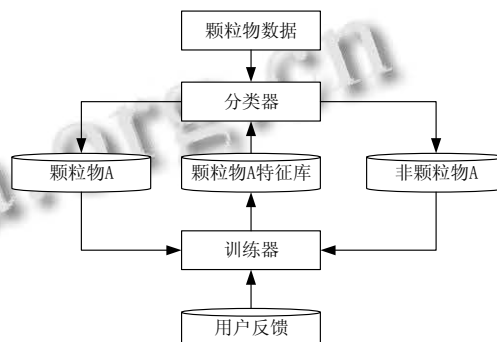


图 5 回归模型

逻辑回归模型用于二分类问题,大气中常见的有七种颗粒:元素碳(EC)、钠钾(NaK)、钾(K)、矿物质(M)、重金属(HM)、大分子有机物(HOC)、有机碳(OC)。因此,我们需要对此七种颗粒物分别建立模型,需要相应的回归模型。在对实际数据进行分类测试时,只需分别对每个颗粒物进行二分类即可。

自动命名结果如图 6 所示。

种类	数量	百分比
EC	127	0.06%
OC	55749	24.76%
HOC	172	0.08%
ECOC	69	0.03%
HM	225	0.10%
DUST	11	0.00%
K	168756	74.96%
NaK	22	0.01%

图6 自动命名示意图

4.5 颗粒物数据统计

为更加宏观直观地显示空气数据，需要对分析后的数据进行统计，以便制定应对方案，助力环境保护。图7是颗粒物粒径分布图；图8表示颗粒物浓度变化曲线图；图9表示颗粒物来源分布饼状图。

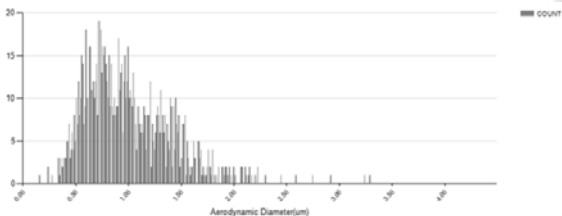


图7 颗粒物粒径分布图

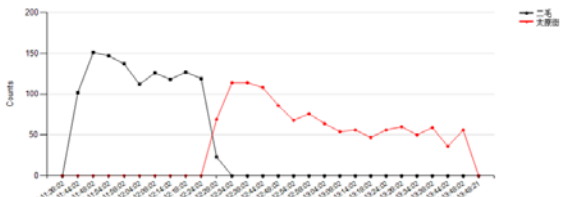


图8 颗粒物浓度变化曲线图

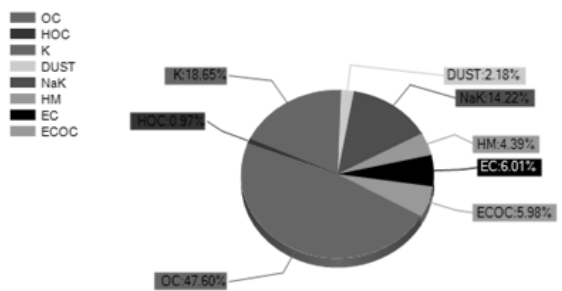


图9 颗粒物来源分布饼状图

根据颗粒物来源分布饼状图，可以进一步对颗粒物来源进行解析，判断出可能的污染源，如：生物质燃烧、工业、尾气、燃煤等。

5 实验分析

本系统测试服务器配置如下：Intel Xeon

E4-1235v2 CPU、64G 内存、Windows Server2008(64 位)，然后于辽宁大学、沈抚新城等位置获取 12W、22W、32W 组颗粒物测试数据，分别测试数据预处理、聚类分析、自动命名等模块的运行时间。

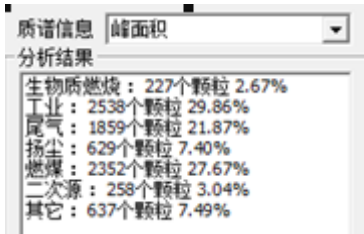


图10 来源分析结果

本文中采用 MySQL 数据库为系统优化前的速度，采用实时数据库后为系统优化后的速度，下面是两者在各阶段的处理性能，数据规模单位为 W(万组)，时间指标单位为 min(分钟)。

表1 系统性能测试

数据	预处理		聚类分析		自动命名		总时间		
	前	后	前	后	前	后	前	后	加速
12	12	8	30	19	4.5	3	46.5	30	35.5%
22	23	17	56	29	6.1	4	85.1	50	41.3%
32	30	20	82	51	7	5.3	119	76.3	35.9%

由上表可以看出，本系统在未优化前，时间分别为 46 分、85 分、119 分，对比于半自动化或者手动数据分析，整个系统分析速度提升数十倍以上；同时，采用实时技术的优化策略后，本系统的运行时间为 30 分、50 分、76 分，系统运行速度提升 35%。

6 结语

在本文中，一种基于实时技术的大气颗粒物在线分析系统被实现，该系统采用实时数据库来解决大气监测中的海量数据和数据存储的实时性两大问题，上层数据分析中采用了 ART-2a 自适应共振神经网络算法和逻辑回归模型等数据分析方法来降低数据规模，提升数据分析速度，成功地处理海量大气颗粒物数据。

实验数据表明，本在线分析系统可以有效地处理大气监测中的海量数据问题，同时整个系统的运行速度在期望范围内，满足环境监测中的实时性需求。在实际测试中，分组数据的正确性能达到 80%；同时，采用实时数据库优化后，相比基于 MySQL 的传统分析系统能提升 35% 的速度。由此可见，本系统对于自动化分析大气颗粒数据具有重要实际应用意义和价值。

参考文献

- 1 尹洧.大气颗粒物及其组成研究进展(上).现代仪器,2012, 18(2):1-5.
- 2 张莉.基于单颗粒气溶胶质谱信息的分类方法研究及其应用[硕士学位论文].上海:上海大学,2013.
- 3 翟明玉,王瑾,吴庆曦,等.电网调度广域分布式实时数据库系统体系架构和关键技术.电力系统自动化,2013,37(2): 67-71.
- 4 杨新兴,尉鹏,冯丽华.大气颗粒物 PM_{2.5} 及其源解析.前沿科学,2013,7(2):12-19.
- 5 王丹.辽宁省大气环境监测数据分析系统研究[硕士学位论文].沈阳:东北大学,2009.
- 6 Yin YF, Gong GH, Han L. Air-combat behavior data mining based on truncation method. Journal of Systems Engineering and Electronics, 2010, 10: 827-834.
- 7 李法运,陈亮.基于改进 BP 网络的网络论坛热点主题挖掘.计算机系统应用,2016,25(3):113-118.
- 8 曹占峰,刘海涛,张启伟.智能统计分析系统.计算机系统应用,2015,24(7):41-45.