

# 新冠疫情数据可视化平台的设计与实现

李泽沁<sup>1</sup>, 巩如悦<sup>1</sup>, 宋秉健<sup>2</sup>

(1. 广东东软学院 计算机学院, 广东 佛山 528225; 2. 暨南大学 智能科学与工程学院, 广东 珠海 519070)

**摘 要:** 新型冠状病毒感染疫情数据高效率利用一直是值得研究的问题, 到目前为止, 还没有一个完美的可视化工具能满足医疗学科对疫情的要求。文章基于 Django 网络框架, 借助 ECharts 完成疫情数据可视化处理和界面的交互功能, 通过爬虫获取疫情数据并建立数据库, 完成网页端可视化界面多元化疫情数据展示的构建, 可视化内容涵盖各省市的确诊数、死亡数等, 疫情数据可视化平台的搭建为人民群众和医疗从事人员了解疫情提供技术支持。

**关键词:** 信息采集与过滤; 网络信息抽取; 新型冠状病毒; 可视化

**中图分类号:** TP311; TP391

**文献标识码:** A

**文章编号:** 2096-4706 (2023) 09-0157-05

## Design and Implementation of a Data Visualization Platform for the COVID-19 Epidemic Situation

LI Zeqin<sup>1</sup>, GONG Ruyue<sup>1</sup>, SONG Bingjian<sup>2</sup>

(1.School of Computer Science, Neusoft Institute Guangdong, Foshan 528225, China;

2.School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China)

**Abstract:** The efficient utilization of novel coronavirus pneumonia epidemic situation data has always been a problem worth studying. So far, there has been no perfect visualization tool that can meet the requirements of medical disciplines for the epidemic situation. This paper is based on Django network framework, completes the visualization processing of epidemic situation data and the interactive function of the interface with the help of ECharts, obtains epidemic situation data via crawler and establishes a database, and forms the construction of a diversified epidemic situation data display in the visual interface on the web terminal. The visualization contents cover the number of confirmed cases and deaths in various provinces and cities. The establishment of the epidemic situation data visualization platform provides technical support for the broad masses of people and medical practitioners to gain an insight into the epidemic situation.

**Keywords:** information collection and filtering; web information extraction; novel coronavirus; visualization

## 0 引言

新型冠状病毒感染 (Corona Virus Disease 2019, COVID-19) 自 2019 年末爆发以来, 在全世界范围内快速蔓延。截至本文写就之时, 国内疫情虽得到有效控制, 但局部地区仍出现疫情反弹现象, 政府官网、主流媒体等实时公布疫情动态, 疫情数据、全民舆情关注等信息每天更新。因信息更新速度快、周期短, 海量的疫情数据如雨后春笋般涌现, 信息数据骤然剧增, 这些数据只有被合理解读与展示, 才能将数据利用最大化。因此, 全国各地为了科学化地管理新冠疫情数据, 不断推出各种数据管理平台, 但都存在疫情数据散乱、模块独立、集成度低等问题。针对上述问题, 本文提出利用爬虫技术实现对新冠疫情数据的爬取和可视化展示, 方便人民群众和科研医务人员更好地了解新冠疫情实时动态, 并为预测疫情的发展趋势提供参考。

## 1 疫情可视化平台研究

### 1.1 系统结构框架设计

在传染病毒和大量人口流动双重条件影响下, 新冠疫情

防控面临重大挑战, 人类社会受到前所未有的健康挑战, 应用数据分析等方法成为防控疫情的一项重大工作。本系统的设计目的在于将散乱无序、数据量庞大的疫情数据以高效、可视化的方式展现出来。疫情数据的收集以每日更新的疫情数据为主, 字段含省累计确诊数、省累计死亡数、省累计治愈数、省累计疑似病例数、市累计确诊数、市累计死亡数、市累计治愈数、市累计疑似病例数, 数据搜索设计可通过搜索定位框查看全国各省市的疫情数据。本系统设计的主体思想以 Django 作为服务器框架, 选定特定 URL 发送 Requests 请求, 获取清洗出初始数据, 将数据储存在 MySQL 数据库完成初步工作, 在结合了 ECharts、Scrapy 等工具从而实现疫情可视化平台的设计。各模块相互协同工作对数据进行深度统计, 在探索新冠疫情发展趋势层面有着重大意义。本文可视化平台的构建由四大模块组成, 分别为: 数据采集、数据筛选、数据存储和数据可视化, 如图 1 所示。

### 1.2 数据采集

国家卫生健康委员会网站会发布当日各省 (自治区、直辖市) 的疫情通报情况<sup>[1]</sup>, 需要获取上述网页中不同省市的新增确诊病例、新增疑似病例、新增出院病例、新增死亡病例及统计出以上四项累计报告病例。数据采集的方法使用 Scrapy 框架对网页数据进行爬取, 选择 MySQL 数据库进行

收稿日期: 2022-12-19

基金项目: 广东大学生科技创新培育专项

资金资助项目 (pdjh2022b0607)

数据的存储<sup>[2]</sup>。网络爬虫模块向Web服务器发送HTTP请求,通过选取初始URL策略,对网页发送爬取请求,获取相应的数据存储在文本文件中,同时,规定爬取的时间规则。获取疫情数据的技术路线,也可通过丁香医生或腾讯网、阿里健康等网站的API地址获取开源的疫情数据源。实验系统采用Python 3.10为编程语言,以Chrome浏览器为例,发出请求,同时查看浏览器的user-agent,使用Requests请求时,设置user-agent伪装浏览器,由于字符串格式不方便数据进行分析,在网页预览中发现数据为类似字典的JSON格式,JSON数据标准化;解析JSON数据;以xls格式输出到本地,建立不同数据的列表元素名称,如表1所示。

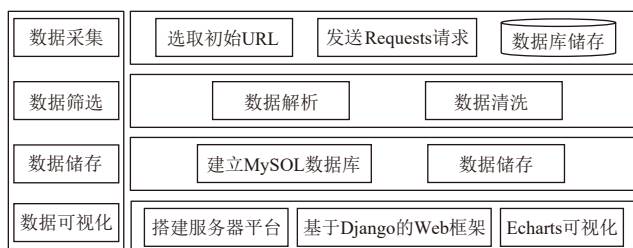


图1 技术路线

表1 数据类型定义

名称	数据类型
Today	当日疫情数据
Total	各省当日累计数据
NULL	空数据
ID	各省邮政编码
Name	各省名称
Update Time	更新时间
Last Name	省下级数据

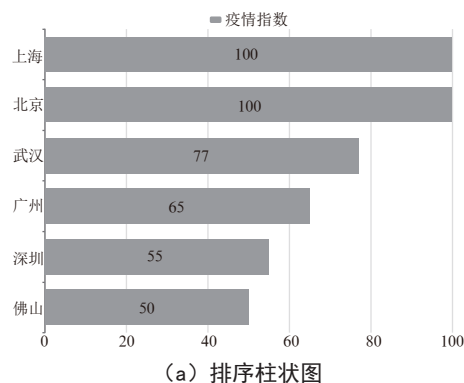
### 1.3 数据清洗

在信息爬取阶段会面临着数据冗余等问题,数据清洗广泛应用于大数据之中,数据清洗旨在识别和纠正数据中的噪声,以尽量减少数据噪声对数据分析结果的影响。数据中的噪声点主要包括不完整的数据、冗余数据、冲突数据和错误数据<sup>[3]</sup>。抓取疫情数据并处理筛选数据的工作中,存在省市疫情数据时间维度上的杂乱,造成数据分析上的噪声,因此需要单独处理各属性疫情数据进行数据聚类,分词识别等尝试性融合,形成一个每次可以处理的全部属性,然后仅运行一轮即可处理各元组全部属性的问题<sup>[4]</sup>。

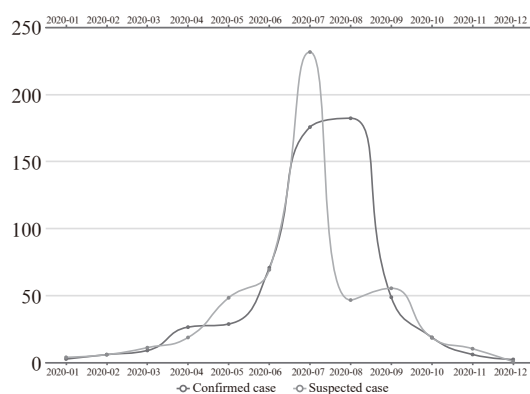
### 1.4 疫情数据可视化

ECharts是百度开源的数据可视化图表库,PyEcharts是一款将Python与Echarts结合的数据可视化工具,借着良好的交互性,精巧的图表设计,得到了众多开发者的认可。Python在所有编程语言里面,代码量最低,非常易于读写,遇到问题时,程序员可以把更多的注意力放在问题本身上,而不用花费太多精力在程序语言、语法上。从海量数据中寻找规律是非常困难的,通过对疫情数据分析和可视化处理,得益于直观的图表,可以大大降低理解数据的难度,帮助人们更快地理解数据<sup>[5,6]</sup>。如图2所示,本文通过折线图、柱状图、地图等形式对疫情数据进行可视化处理,

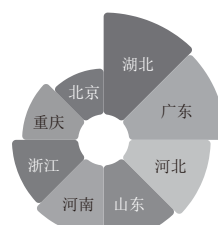
实验所采用的开发工具及相应扩展库的版本为Python 3.10、PyEcharts 1.7.1,使用Py-Charm 2022.1.1作为语言编辑器,PyEcharts可视化的步骤可以分为:导入数据包;准备数据库;生成图表;对图表进行处理,绘制折线图、热力图以及柱状图。



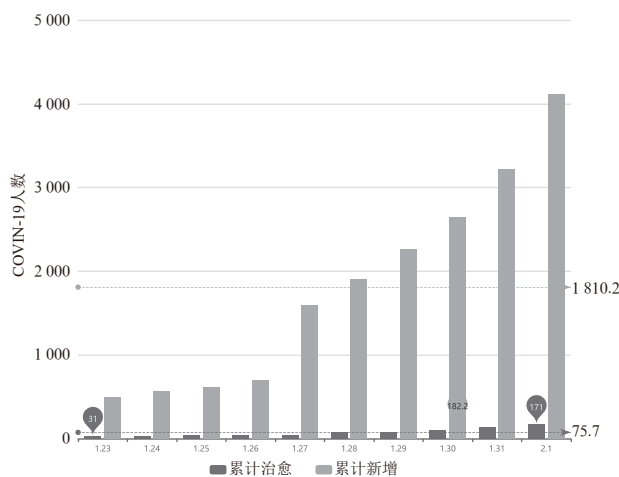
(a) 排序柱状图



(b) 多X轴曲线图



(c) 南丁格尔玫瑰图



(d) 组合型示例图

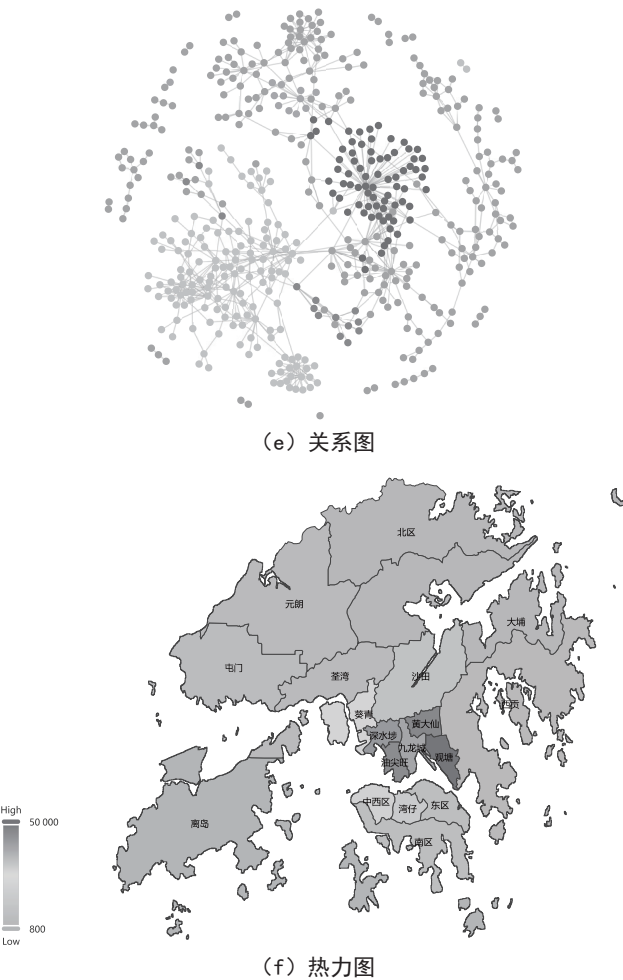


图 2 ECharts 可视化方式示例图

苏格兰工程师、政治经济学家威廉·普莱费尔最先在《The Commercial and Political Atlas》和《The Statistical Breviary》中使用了折线图、条形图和饼状图等图表表达方式用于传达复杂的经济数据<sup>[7]</sup>。折线图在数据统计中广泛应用,在统计疫情数据在省/市疫情动态和时间维度关系上有利于展示信息的波动情况,折线图还可以应用在分析疫情数据治愈率/病死率/重症率等在时间序列上的规律特征,有助于揭示疫情的拐点。柱形图常常应用于不同类别事物的具体事件数据对比,可以帮助数据阅读者快速对比数据的大小规律,疫情数据可视化处理过程中可通过柱形图表对立城市/省/全国/海外疫情数据进行多维度对比。热力图采用颜色深度来显示该区域的数值,直观地展示了国内外疫情的增长趋势,可用于分析疫情的空间传播,控制传染源。词云用于对出现频率较高的“关键词”予以视觉上的突出,可对疫情数据中的关键字按照出现次数进行展示。

2 疫情可视化平台设计与实现

2.1 系统总体框架

本系统采用四层架构设计模式,具体可分为应用前端、服务终端、数据库、运行基础,如图 3 所示,主要实现疫情数据的爬取和疫情数据的分析并可视化两大主要功能。运行基础是系统的硬件设备和软件系统的集成基础支持,通过获

取的疫情数据资源储存至已经建立的 MySQL 数据库中,基于系统的需求,完成后端的数据提取与封装。Django 是一个用 Python 语言编写的开源 Web 开发框架,可以快速地实现网站的快速开发和维护,因其语言的兼容性和使用的灵活性受到广泛的关注。本系统的服务层利用 Django 搭建网页框架平台, Ajax 技术实现前后端的交互<sup>[8]</sup>,在 ECharts 中丰富多样化的可视化图库支持下实现全国各个疫情数据资源的可视化。

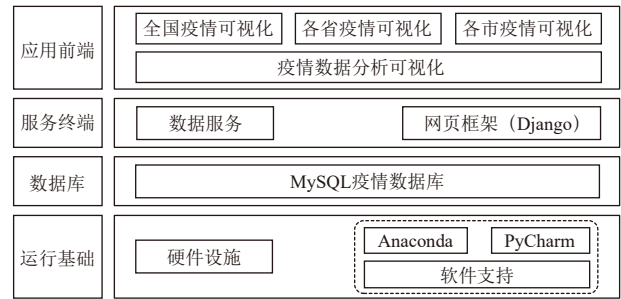


图 3 系统总框架图

2.2 系统需求分析

用于增强认知所采用的计算机辅助抽象数据,其直观的表达方法能够为数据观察者理解信息提供便利。疫情数据可视化是指通过数据分析与可视化设计等手段把疫情信息直观易懂地展现在大众面前,需要根据需求对可视化设计进行优化<sup>[9]</sup>。综合考虑新冠疫情相关数据可视化设计的问题,在搜集已有互联网平台上发布的优秀可视化案例并进行对比分析的基础上,重点对新冠疫情数据的内容信息、呈现形式以及交互形式等方面进行可视化可用性分析。计算机性能与数据存储疯狂而令人难以置信的发展正改变着科学,知识、监控、自由、读写能力、艺术等一切可被表达成数据或者基于这些表达的事物,大数据正影响人们未来的生活<sup>[10]</sup>。百度、丁香医生和网易等互联网公司纷纷发布了各自专属的疫情数据可视化平台。如图 4 所示<sup>[11-13]</sup>,疫情伊始这些平台就扮演了重要角色,给民众和医护人员带来了言简意赅的数据理解。大数据具有数据规模庞大 (Volume)、种类繁多 (Variety)、更新迅速 (Velocity) 3 个典型特点<sup>[14, 15]</sup>,依据以上理论基础,本文搭建了一个轻量级的疫情数据可视化平台,该平台可高效快捷地对系统海量疫情信息进行快速集成和图表可视化处理,其中系统数据处理能力涵盖了基本的数据收集模块,优化数据分析能力,提升数据使用者对可视化图表的交互体验。

2.3 系统功能设计与实现

2.3.1 全国疫情数据可视化

全国疫情数据追踪可视化方式通过热力地图加以颜色区分展示出来,通过数据绑定可实现移动鼠标到指定省份会显示各省疫情数据,进而实现数据的多层次展示。如图 5 所示,除湖北省外浙江省的确诊人数最多,其次为河南,为了使可视化效果更佳,在前端设计上加入鼠标悬停发生事件的功能,当鼠标悬停到特定省份板块上会弹出浮动框显示本省份的确诊、新增、死亡等疫情数据,丰富界面功能,数据来自国家卫生健康委员会官方网站<sup>[1]</sup>。



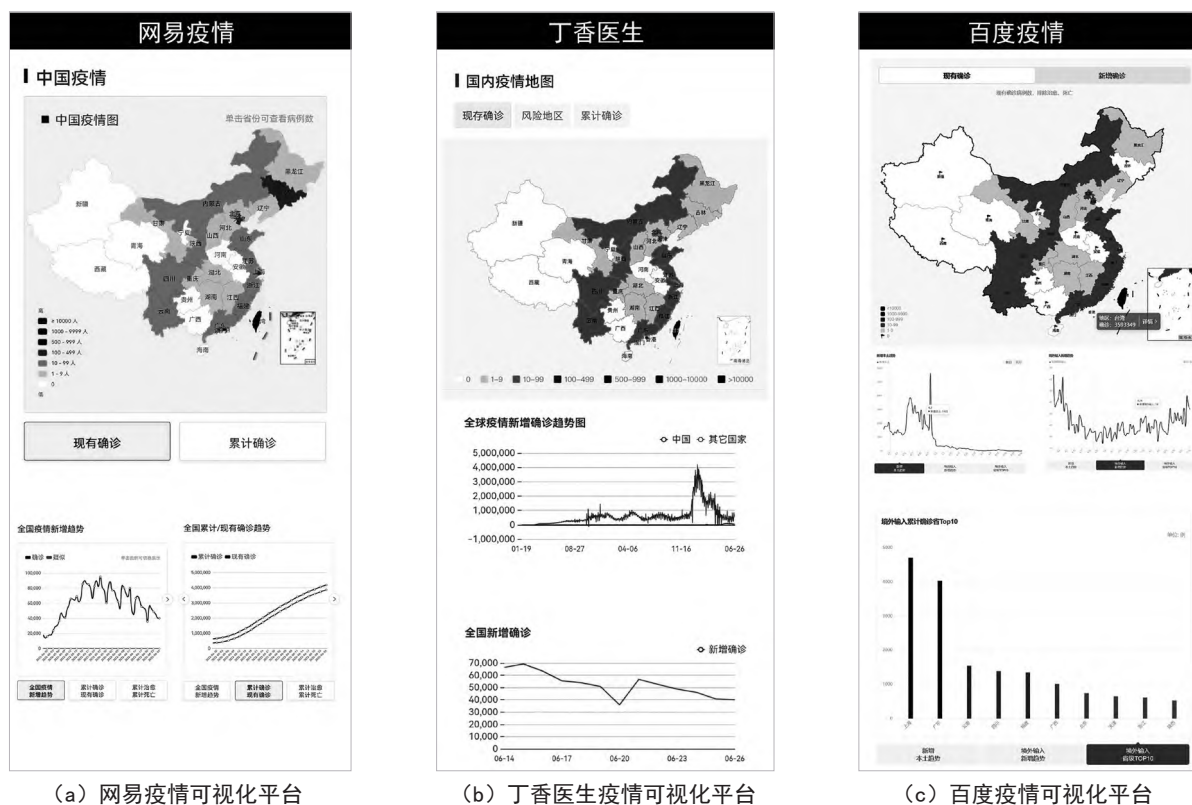


图 4 多方平台疫情可视化平台案例

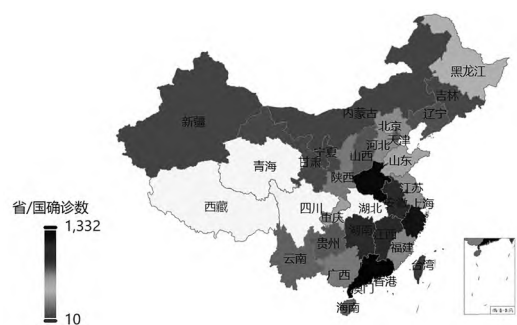


图 5 全国疫情热力图(除湖北)

### 2.3.2 各省疫情数据可视化

各个省疫情数据可视化方式通过热力地图的形式展示，在全国热力地图的基础上加上数据下钻的功能模块。显示效果如图 6 所示，下钻可以理解成增加维度的层次，从而可以由粗粒度到细粒度来观察数据，对热力地图情况分析时，可以沿着层次维度更细粒度的观察数据，在可视化领域中广泛应用。数据下钻的实现思路是获取鼠标点击的参数跳转至指定的 JSON，本实验中定义两层数据维度，数据下钻到省级热力地图直观地展示了疫情的增长趋势，可用于分析疫情的区域空间的传播。

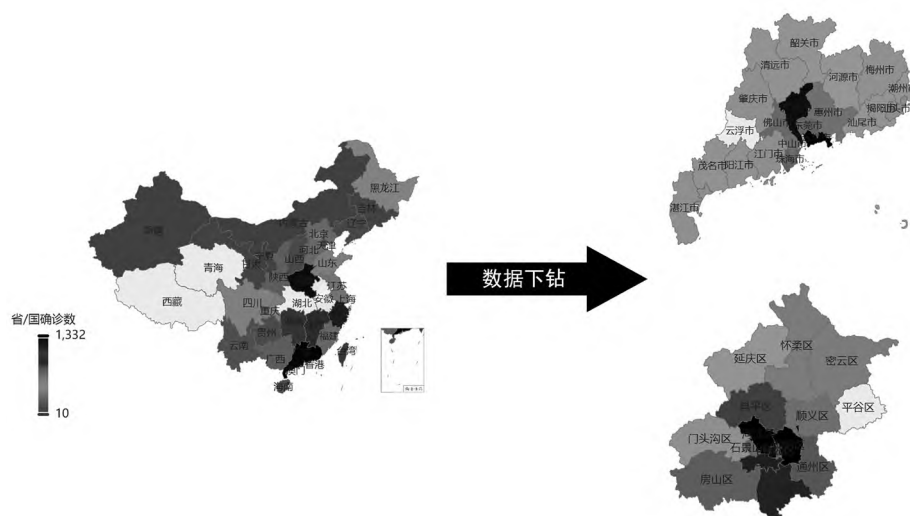
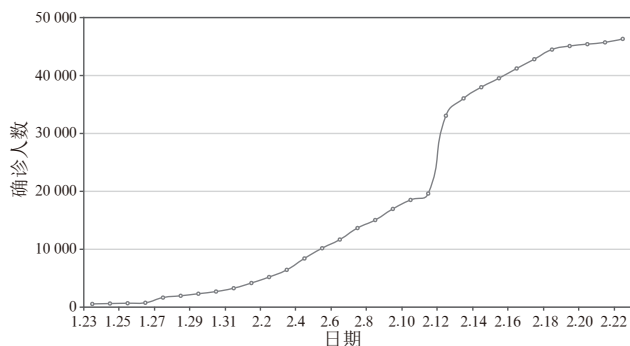


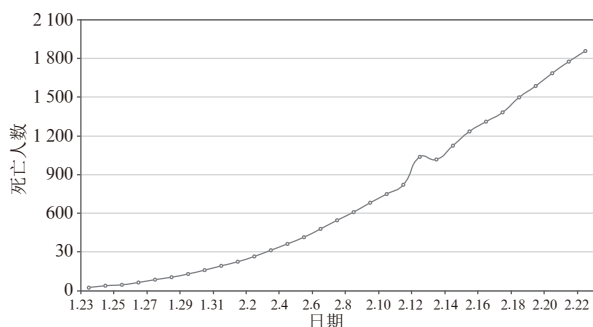
图 6 数据下钻(节选)

### 2.3.3 各市级疫情数据可视化

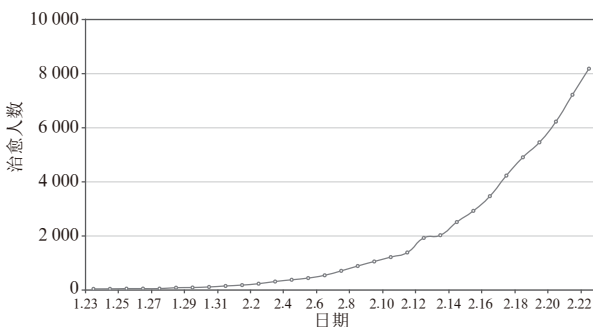
市级疫情数据可通过折线图统计方式可视化, 日期时间为横轴, 累计疫情数据为纵轴, 渲染效果如图 7 所示。折线图统计的方式主要包括累计确诊、累计死亡、疑似人数、治愈人数, 从图可以看出 2020 年 2 月 14 日附近武汉市累计确诊人数剧增, 折线图统计方式有利于快速判断疫情发展趋势, 数据来自国家卫生健康委员会官方网站<sup>[1]</sup>。



(a) 武汉市累计确诊人数折线图



(b) 武汉市累计死亡人数折线图



(c) 武汉市累计治愈人数折线图

图 7 疫情数据可视化

## 3 结 论

图表可视化数据已经成为人们日常生活中理解信息的重要

途径, 数据量的急剧变化, 加快了计算机科学技术对大数据处理方法的的发展, 从而更加高效、直接地对大规模数据流进行快速的数学模型构建、数据分析、图表绘制效率。新型冠状病毒可视化系统有效地应对大规模数据难以短时间规划构建等问题, 在用户应用层面, 方便人们随时进行对疫情数据的查看, 根据每日疫情数据呈现的结果也可推断出我国疫情的控制情况以及预测出疫情未来的发展趋势。通过可视化系统及时准确评估疫情传播、根据每日疫情情况了解疫情实时形势, 分析疫情发展趋势、防控措施有着重要意义。

### 参考文献:

- [1] 中华人民共和国国家卫生健康委员会. 疫情通报 [EB/OL]. (2022-07-01) [2022-07-01]. [http://www.nhc.gov.cn/xcs/yqtb/list\\_gzbd.shtml](http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml).
  - [2] 王君, 马乐荣. 新冠肺炎疫情信息累积更新采集系统的设计与实现 [J]. 延安大学学报: 自然科学版, 2021, 40 (2): 96-99.
  - [3] 郝爽, 李国良, 冯建华, 等. 结构化数据清洗技术综述 [J]. 清华大学学报: 自然科学版, 2018, 58 (12): 1037-1050.
  - [4] 杨东华, 李宁宁, 王宏志, 等. 基于任务合并的并行大数据清洗过程优化 [J]. 计算机学报, 2016, 39 (1): 97-108.
  - [5] 裴丹丹. 基于 ECharts 的数据可视化实现 [D]. 北京: 北京邮电大学, 2018.
  - [6] 王竞俨, 侯彦东. 基于 PyEcharts 的第三次农业普查数据可视化分析 [J]. 河南科学, 2021, 39 (4): 517-523.
  - [7] 陈小巧. 西方信息图表的崛起 [J]. 设计, 2015 (10): 132-133.
  - [8] 李相霏, 韩珂. 基于 Flask 框架的疫情数据可视化分析 [J]. 计算机时代, 2021 (12): 60-63+68.
  - [9] 巩如悦. 基于本体的苹果病虫害垂直搜索引擎研发 [D]. 咸阳: 西北农林科技大学, 2017.
  - [10] DOCTOROW C. Big Data: Welcome to the Petacentre [J]. Nature, 2008, 455 (7209): 16-21.
  - [11] 网易新闻. 肺炎疫情实时动态 [EB/OL]. (2022-06-26) [2022-06-26]. [https://wp.m.163.com/163/page/news/virus\\_report/index.html](https://wp.m.163.com/163/page/news/virus_report/index.html).
  - [12] 丁香医生. 疫情实时动态 [EB/OL]. (2022-06-26) [2022-06-26]. <https://ncov.dxy.cn/ncovh5/view/pneumonia>.
  - [13] 百度. 疫情实时大数据报告 [EB/OL]. (2022-6-26) [2022-06-26]. [https://motion.baidu.com/activity/yiqing\\_predict/main](https://motion.baidu.com/activity/yiqing_predict/main).
  - [14] MCAFEE A, BRYNJOLFSSON E. Bigdata: The Management Revolution [J]. Harvard Business Review, 2012, 90 (10): 60-68+128.
  - [15] 沈恩亚. 大数据可视化技术及应用 [J]. 科技导报, 2020, 38 (3): 68-83.
- 作者简介: 李泽沁 (2001—), 男, 汉族, 广东湛江人, 本科在读, 研究方向: 电子信息工程; 巩如悦 (1990—), 女, 汉族, 山东滨州人, 助教, 硕士研究生, 研究方向: 智能信息化; 宋秉健 (2000—), 男, 汉族, 广东湛江人, 本科在读, 研究方向: 人工智能。

(上接 156 页)

### 参考文献:

- [1] 刘磊. 动车组单车数字化调试系统的设计与实现 [J]. 智慧轨道交通, 2022 (5): 53-59.
- [2] 吴艳鹏. 动车组单车调试工艺及发展趋势 [J]. 城市轨道交通研究, 2021 (2): 140-143.
- [3] 马博. 浅谈地铁车辆调试潜在风险控制对策研究 [J]. 科技经济市场, 2018 (2): 171-172.

[4] 吕奔. 地铁车辆调试工作探讨 [J]. 黑龙江科技信息, 2014 (27): 143.

[5] 史建飞. 电力机车数字化调试系统设计 [J]. 轨道交通装备与技术, 2021 (3): 29-31.

[6] 孙剑. 特种车辆制造数字化工艺协同设计能力建设 [J]. 新技术新工艺, 2020 (8): 1-5.

作者简介: 逯宏伟 (1987—), 男, 汉族, 河南郑州人, 工程师, 硕士研究生, 研究方向: 轨道交通车辆电气及调试工艺技术。