哈尔滨工业大学深圳校区

毕业设计（论文）中期报告

题　　目　新冠肺炎疫情患者的时空数据收集与分析

姓　　名　段 裕

学　　号　180110704

学　　院　计算机科学与技术学院

专　　业　计算机科学与技术专业

指 导 教 师　冯 山 山

日　　期　2022 年 03 月 13 日

目　录

1　课题主要研究内容及进度

1.1　课题主要研究内容

本课题将从以下几个方面开展研究：

（1） 依托网易、腾讯、百度等的疫情大数据平台，收集自2020年03月01日至2022年03月01日中国各省、各地级市的新冠疫情本土新增确诊数据。

（2） 依托深圳市卫生健康委员会门户网站、深圳市人民政府门户网站，收集自2020年08月17日深圳市本土新增新冠肺炎确诊病例的报告数据。

（3） 对收集到的中国各地级市的新冠疫情本土新增确诊数据和深圳市病例的详细数据进行简单的统计分析，使用条形图、折线图、中国地图等等多元化的图表对数据的不同方面进行展示。

（4） 探讨研究Hawkes时空点过程模型的起源发展、方法原理和应用实例。

（5） 以中国各地级市的新冠疫情数据为数据集，训练一个Hawkes时空点过程模型，并尝试对模型进行评估。

1.2　进度介绍

目前，已经完成的工作包括：

（1） 编写网页爬虫收集和清洗自 2020 年 03 月 01 日至 2022 年 03 月 01 日中国各省、各地级市的新冠疫情本土新增确诊数据。

（2） 编写网页爬虫收集和清洗自 2020 年 08 月 17 日深圳市本土新增新冠肺炎确诊病例的报告数据。

（3） 初步完成了中国各省、各地级市的新冠疫情本土新增确诊数据的统计分析和可视化，主要包括全国疫情实时新增和累计数据、全国疫情新增和累计趋势折线图、全国疫情实时新增确诊地图、疫情最严峻的地级市条形图（依据单日新增本土确诊病例数）等。

（4） 初步了解了 Hawkes 模型的工作过程，初步运行了基于 python+pytorch 的 Hawkes 模型项目示例。

目前，还需要完成的任务包括：

首先对新冠疫情本土新增确诊数据集地进一步处理，以对接模型训练的输入，例如转为 nump 专用文件格式；其次是深圳市本土新增新冠肺炎确诊病例的报告数据进行统计分析和可视化；最后需要继续构建和训练 Hawkes 模型。

2　已完成的研究工作及结果

2.1　新冠肺炎疫情数据爬虫设计

2.1.1　爬虫的一般概念与工作流程

如图 2-1-1 所示，网络爬虫(Web Crawler)是从指定的初始 url 队列开始，对于每一个队列中的 url，通过网络资源请求获取到网络页面，并结合网页解析技术提取有效信息同时把过滤得到的新 url 添加到初始 url 队列尾并继续爬取页面的自动化过程，这一过程在达到用户指定的结束条件时终止。

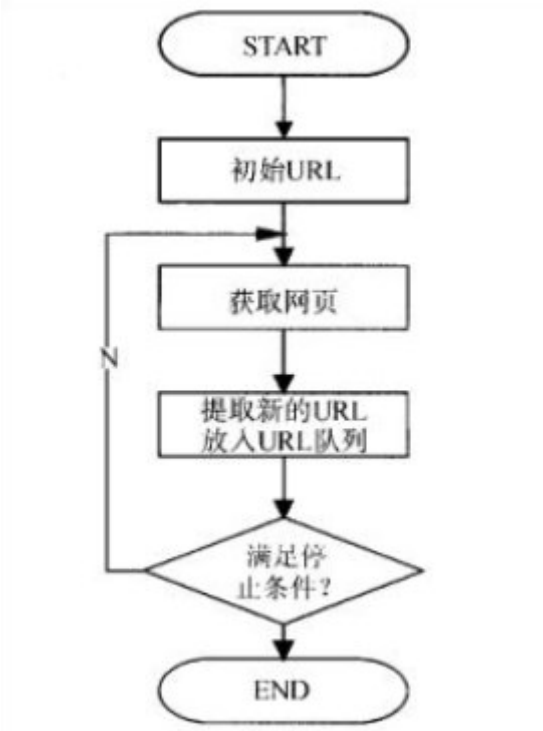

图 2-1-1 爬虫工作的基本流程

与通常的程序不同，爬虫程序是一种网络程序，需要考虑网络的波动性与服务器的负载能力，例如，需要编写异常处理模块以应对请求页面超时等问题，需要在连续获取页面时插入一段随机等待时间，需要为网络请求函数封装请求头部，并随机更换等等，这些措施一定程度上可以避免被识别为恶意攻击者以及本机 IP 受到封锁等情况。

其次，针对数据信息的存储问题，不难看出，采集的主要是文本数据，于是可以采取类似表格的 csv 文件（文本文件）进行存储，利用 python 的 pandas 模块

csv 文件进行读取存储和处理。

最后也是最关键的一步是数据清洗，这一步工作基于爬取模块存储的原始信息 csv 表格文件，将感兴趣的数据信息进行抽取，同时还要将一些非结构化的信息转为结构化的信息，例如对于提取出的地点信息，应当结合有关地图 API 得到经纬度的位置信息，对于日期信息，应该确定起止日期，并将起止日期以及之间的所有日期转为距起始日期的时间间隔等等。如图 2-1-2 所示，利用网络爬虫获取数据信息的一般工作流程概括如下：

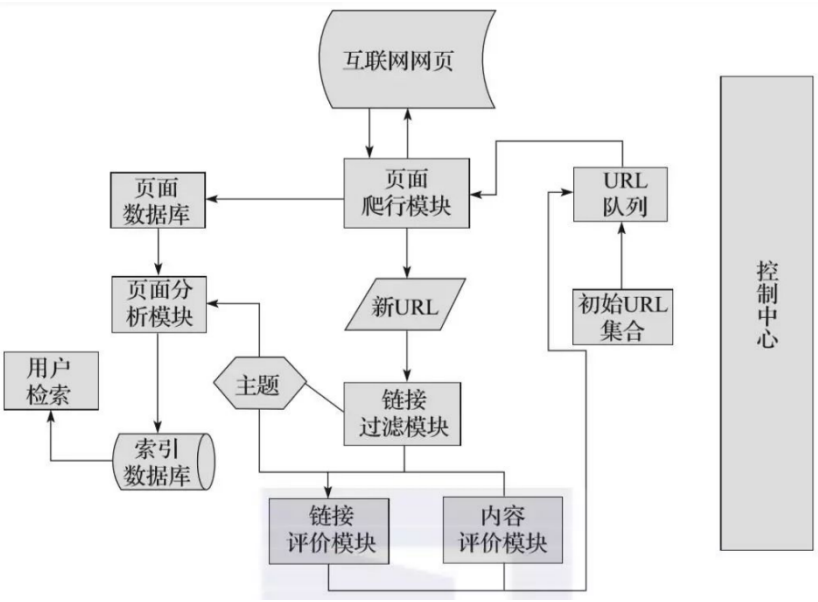

图 2-1-2 利用爬虫获取网络数据的一般流程

2.1.2 深圳市卫健委病例数据的爬取

如图 2-1-3 所示，首先从深圳市人民政府疫情通报发布页爬取疫情通报，主要包括文章发布时间、文章标题以及文章内容，其中文章内容以字符串列表的形式存放，按照 HTML 中正文内容安排依次存放其中每一个文本标签(p 标签)中的文字，这样处理有利于后续病例信息的提取。

```
publicDate,passageTitle,passageContent
2022-04-01,2022年4月1日深圳市新冠肺炎疫情情况,"['3月31日0-24时，深圳新
2022-03-31,2022年3月31日深圳市新冠肺炎疫情情况,"['3月30日0-24时，深圳
2022-03-30,2022年3月30日深圳市新冠肺炎疫情情况,"['3月29日0-24时，深圳
2022-03-29,2022年3月29日深圳市新冠肺炎疫情情况,"['3月28日0-24时，深圳
2022-03-28,2022年3月28日深圳市新冠肺炎疫情情况,"['3月27日0-24时，深圳
2022-03-27,2022年3月27日深圳市新冠肺炎疫情情况,"['3月26日0-24时，深圳
2022-03-26,2022年3月26日深圳市新冠肺炎疫情情况,"['3月25日0-24时，深圳
```

图 2-1-3 深圳市人民政府疫情情况通报（部分）

如图 2-1-4 所示，接下来是将每一个本土确诊病例的详细信息从通报中分离出来，主要包含确诊日期、病患性别、病患年龄以及常居地点。这一步首先依据是否含有类似“本土确诊”等字样过滤出含有本土确诊病例详细信息的通报，然后再过滤得到病患的详细信息。

```
confirmDate,patientGender,patientAge,homeAddress
2022-03-31,女,45,罗湖区莲塘街道鹏兴花园六期
2022-03-31,男,7,罗湖区莲塘街道鹏兴花园六期
2022-03-31,男,45,罗湖区莲塘街道鹏兴花园六期
2022-03-31,女,11,罗湖区莲塘街道鹏兴花园六期
2022-03-31,男,25,宝安区石岩街道兰姜新村十区
2022-03-31,男,30,宝安区石岩街道兰姜新村十区
2022-03-31,男,20,龙岗区布吉街道凤尾坑新村
2022-03-31,女,31,光明区马田街道福华西三巷
```

图 2-1-4 2022 年 3 月 31 日的病例详细情况

如图2-1-5所示，最后一步是要把非结构化的常居地转为可处理的结构化数据，也即经纬度，这里利用了百度地图 API，通过查询得到每个地点的经纬度。

```
confirmDate,patientGender,patientAge,homeLng,homeLat
2022-03-31,0,45,114.17449499876464,22.5670368207619
2022-03-31,1,7,114.17449499876464,22.5670368207619
2022-03-31,1,45,114.17449499876464,22.5670368207619
2022-03-31,0,11,114.17449499876464,22.5670368207619
2022-03-31,1,25,113.96341081077756,22.69237324693926
2022-03-31,1,30,113.96341081077756,22.69237324693926
2022-03-31,1,20,114.1252135093508,22.59775948374267
2022-03-31,0,31,113.87841646718188,22.78600559401392
```

图 2-1-5 2022 年 3 月 31 日的病例详细情况（结构化数据）

2.1.3 省市级疫情实时数据的爬取

本部分的数据采集主要依赖于疫情实时播报平台，此时网页更像是一个静态的容器，装入的数据是实时更新的，无法按照爬取网页然后利用 HTML 解析器来解析静态网页的方式获取数据。因此，需要考虑一种动态的方式，从请求网页资源开始，监视服务器端回复的所有格式的文件和数据包，通常数据存放在 json 格式的文件中。通过浏览器的开发者工具，抓取目标 json 数据包的 api 地址，获取和分析更加格式化的 json 数据。

经过数据对比分析，本人选择网易的实时发布平台作为爬取对象，将其中有关中国各省市级的病例 json 数据爬取下来，目前累计收集清洗数据量约为 32 万条。如图 2-1-6 所示，为香港 2022 年 3 月 13 日的疫情情况。可以看到，3 月 13 日香港单日新增确诊 8163 例，需要得到严格的管控，另外毗邻香港的深圳市单日新增确诊 60 例，形式依然严峻，从更大的时空范围看，疫情的爆发的确具有时间和空间的局部性，而基于疫情数据的时空点过程模型的研究，正是把这一种局部性用更形式化的语言加以表征。

图 2-1-6 香港 2022 年 3 月 13 日的 json 格式数据预览

2.2 历史数据统计分析

2.2.1 疫情实时地图

本模块的定位是一个简洁的数据展示系统，前端采用 flask 框架，结合 html+css 进行布局，利用 js 和 ajax 等技术进行前后端交互。业务逻辑是借助 pymsql 模块编写查询语句从数据库查询数据，返回前台进行渲染和展示，主要利用了 echarts 工具进行图像绘制。包括全国疫情地图、全国单日新增和累计趋势、确诊和治愈数量

最多的省市，展现形式主要为地图、条形图和折线图等。

系统界面截图如下：

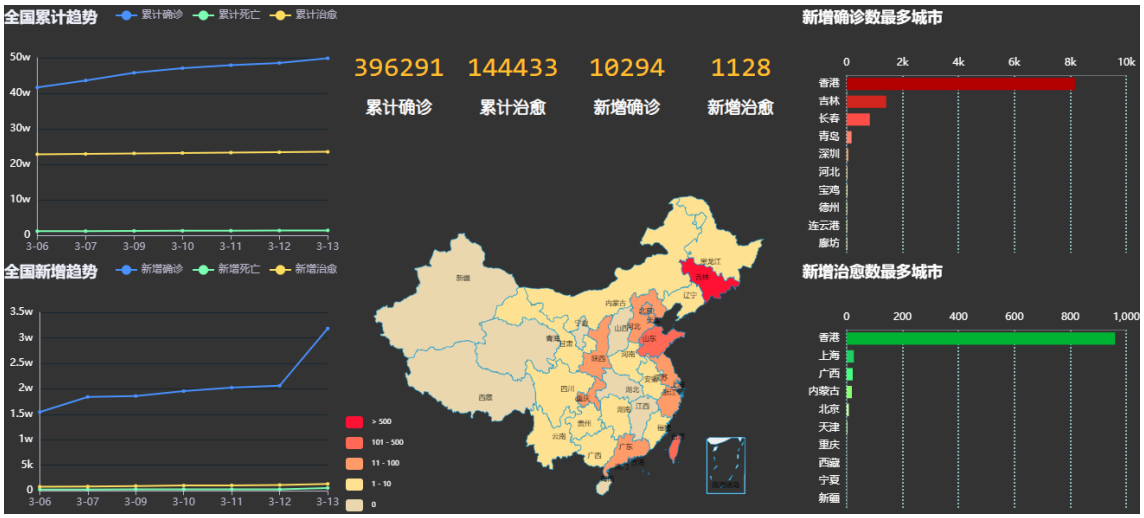

图 2-2-1 新冠疫情实时播报平台

如图2-2-1所示，展示界面左边是两张折线图，左上方是近一周全国累计趋势，其中的三条折线从上至下依次是累计确诊、累计治愈、累计死亡的趋势，左下方则为近一周全国新增趋势，其中的三条折线从上至下依次是新增确诊、新增治愈、新增死亡的趋势，可以看出，无论是累计趋势还是新增趋势，确诊所在折线均远高于其它折线，对比来看，累计确诊数字和新增确诊数字均有较大增幅，形式不容乐观。

区域中部上方精确展示了截至 3 月 13 日全国累计确诊、累计治愈、新增确诊、新增治愈的数字，中部中下方用一张中国地图展示中国各个省级行政区划实时的新增病例情况，新增病例数越多颜色越深。

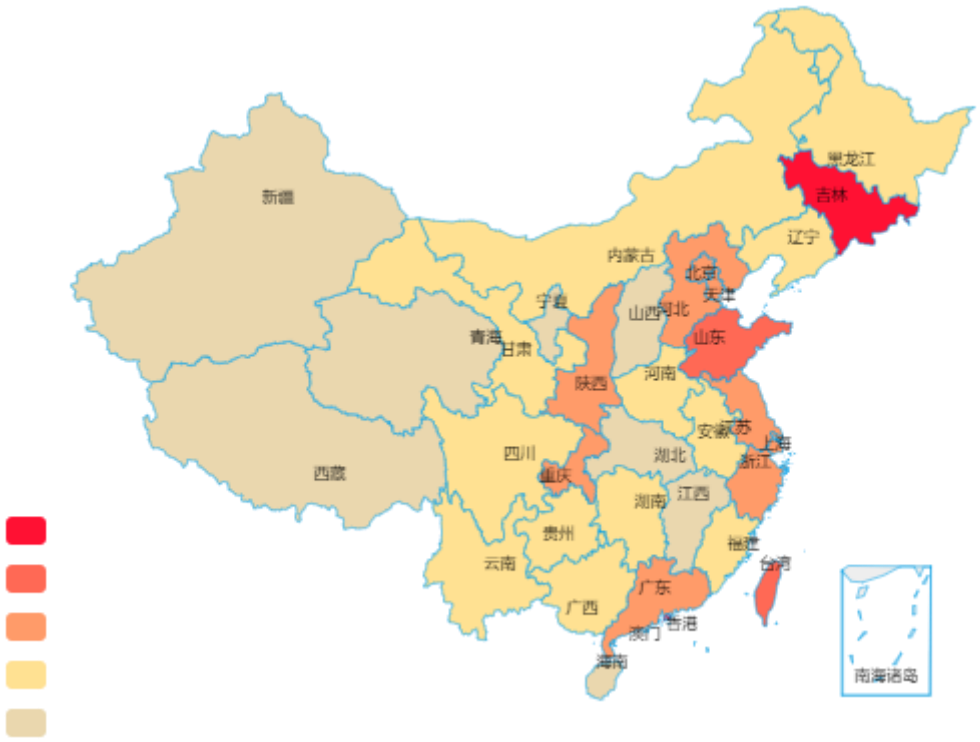

图 2-2-2　2022 年 03 月 13 日中国疫情实时地图

如图 2-2-2 所示，沿海多个省市疫情形式都比较严峻，包括吉林、山东、上海、浙江、广东等地区，而靠近内陆的省市疫情则相对缓和。

右边则为横向的直方图，从上到下分别表示新增确诊数、新增治愈数最多的十个省市，按照数量递减的顺序进行排列，可以看到香港单日新增确诊人数超过 8 千，疫情形式十分严峻。

2.3　时空点过程分析法

2.3.1　Hawkes 点过程模型的提出与适应性分析

Hawkes 点过程模型最早由英国杜伦大学(University of Durham)的 Hawkes 教授在 1971 年发表的论文 Spectra of some self-exciting and mutually exciting point processes 中提出，此后逐渐应用到各个研究领域，例如用于地震的位置和震级的数学建模、传染病学模型分析等众多领域。目前在国内还没有基于 Hawkes 点过程模型及类似方法对中国疫情时空数据进行数学建模的相关研究工作，因此本文呈现的技术方法和研究工作是具有一定创新性的。

2.3.2 Hawkes 点过程模型的分析

在 Hawkes 教授提出该模型之前，Bartlett 教授研究讨论了一种基于点系(point spectrum)的分析方法，该方法考虑了一种静态的点过程模型 N(t)，表示 0 时刻和 t 时刻之间事件(event)发生的累计数量，其中这一过程的强度函数

$$\lambda = E\{dN(t)\}/dt$$

是一个常数，并且协方差密度函数

$$\mu(\tau) = E\{dN(t+\tau)\, dN(t)\}/(dt)^2 - \lambda^2$$

不依赖于时刻t。不同于这一静态过程，Hawkes 教授将过程强度函数修正为关于 t 的函数：

$$\lambda^*(t) = \mu + \alpha \sum_{t_i<t} g(t - t_i)$$

其中，μ表示背景强度，大于 0，一般情况下$g(t) = e^{-wt}$，表示衰减函数，其中w代表历史事件对当前影响的衰减指数，α理解为学习速率。进一步的，公式的后半部分是和的形式，、表明 Hawkes 过程是自激励的，其值依赖于直至时刻t的过程的历史，到目前为止，我们讨论了“单变量”霍克斯过程。然而，我们可以假设，每个事件的发生都有一个来自有限集合的离散标记或标签。在这种情况下，人们不仅可以将系统视为一个单一的随机过程，而且可以将其视为一系列相互作用或相互激励的时间点过程。如图 2-2-3 展示了不同类型的事件相互激励的过程。

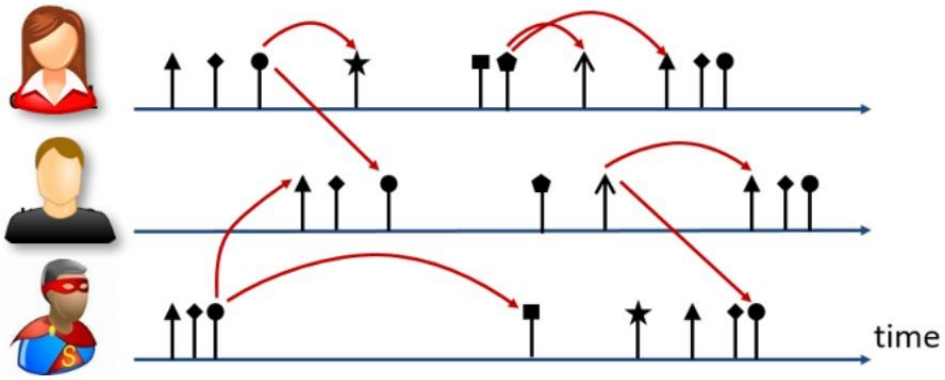

图 2-2-3 不同类型的事件相互激励的过程示意

3　后期拟完成的研究工作及进度安排

3.1　后期拟完成的研究工作

后续需要完成的工作主要是：

- 将深圳市自 2021 年 05 月 30 日有详细记录的第一例病例至 2022 年 03 月 14 日的病例数据进行清洗整合；
- 搜集有关时空点过程模型的论文、项目，尝试将模型应用于新冠疫情的病例数据分析；
- 整理研究结果，作出研究结论，撰写毕业论文。

3.2　进度安排

具体进度安排如下：

2022 年 4 月　1 日——2022 年 4 月 5 日：整理深圳市的疫情数据；

2022 年 4 月 6 日——2022 年 4 月 23 日：搜集时空点过程模型的相关论文会议、文档、视频、项目等介绍信息，尝试运行一个模型示例；

2022 年 4 月 24 日——2022 年 5 月 9 日：将时空点过程模型运行在新冠疫情数据集上，包括基于城市的大粒度数据集和基于病例的细粒度数据集，并对模型做出适应性调整和改进；

2022 年 5 月 10 日——2022 年 6 月 20 日：整理研究成果，撰写毕业论文。

4　存在的困难及解决方案

4.1　存在的困难

1. 目前有关时空点过程模型的资料还比较缺乏和杂乱，对该模型的理解不够清晰。

4.2　解决方案

针对目前存在的问题和困难，提出以下解决方案：

从多方渠道获取信息，先对模型有初步的理解与判断，一是搜寻论文期刊顶会等各类中外文学术网站上的论文、文章、演示文稿等；二是在各种不同的技术交流社区搜寻相关的技术类博客和项目等。在对模型需要进行进一步的理解时若有疑虑困惑及时询问导师。

5　论文按时完成的可能性

本课题目前已经完成了基本的数据收集工作。

目前正在进行深圳市细粒度的患者数据收集工作和关于时空点过程模型的资料搜集工作，将要进行新冠疫情时空点过程模型项目的运行与调整工作。

综上所述，论文能够按期完成，并取得一定的研究成果。