

哈爾濱工業大學(深圳)

毕业设计（论文）开题报告

题 目：新冠肺炎疫情患者的时空轨迹数据收集
与分析

专 业 计算机科学与技术专业

学 生 段 裕

学 号 180110704

指导教师 冯 山 山

日 期 2021-11-19

哈尔滨工业大学（深圳）教务部制

1. 课题来源及研究的目的和意义

- 课题来源:

2020 年年初,湖北省武汉市爆发新冠肺炎疫情,新冠肺炎疫情是新中国成立以来发生的传播速度最快、感染范围最广、防控难度最大的一次重大突发公共卫生事件,对中国是一次危机,也是一次大考。在本次疫情战役中,全国上下一心贯彻“坚定信心、同舟共济、科学防治、精准施策”总要求,用一个多月的时间初步遏制了疫情蔓延势头,用两个月左右的时间将本土每日新增病例控制在个位数以内,用 3 个月左右的时间取得了武汉保卫战、湖北保卫战的决定性成果,疫情防控阻击战取得重大战略成果。根据以上事实,结合导师意见,本人确定该选题。

- 研究目的:

本次疫情爆发突然、蔓延迅速、波及范围广、持续时间长,目前全国范围内部分地区仍有零星的病例出现,同时仍有传染扩散的风险,形式依然严峻,不容乐观。课题收集自 2020 年 2 月疫情全面爆发以来近一年每日全国各省市的新增病例人数,围绕新冠患者在感染前后的时空轨迹分布,收集数据并整理成为相性良好的数据库系统,结合传染病学原理探讨个人时空运行轨迹、各类疫情防控措施例如隔离等在疫情扩散中的消极积极作用,预期还将达到疫情预测预警能力。

- 研究意义:

以计算机科学与数据分析的视角观察过去一年半全国疫情爆发、全民抗疫的过程,力求科学理性的探讨个人行动轨迹、防疫措施在疫情传播的影响,为国家为人民以后的重大公共卫生事件的应对提供微薄的参考。

2. 国内外在该方向的研究现状及分析

- 国内研究现状:

目前,国内关于新冠肺炎疫情的研究主要集中在医学学科,包含内科学、临床医学、护理学等等,涉及到的主题有新冠病毒性状分析与药物研发、新冠肺炎诊疗方案、新冠肺炎与其它并发症的关系、中医药在新冠肺炎防治中的作用、新冠肺炎的遗传机制等等,该方面期刊论文发表活跃度则在 2020 年持续走高,在今年 2021 年也十分活跃。

在涉及到与计算机学科交叉的领域则可供参考的期刊文献学术论文等则相对减少,其中涉及到的领域包括大数据分析病毒溯源、大数据

分析疫情走势预测、疫情数据信息管控机制、基于区块链的智能隔离监控系统等等，总体来说，国内关于新冠疫情患者大数据分析方面的研究十分活跃。

- 国外研究现状：

依据 Nature、The Lancet 等等学术期刊网站，可以了解到国外目前针对新冠肺炎疫情的研究主要针对新型冠状病毒药品研发、蛋白分子生物学、深度学习量化分析疾病严重程度等等。在新冠疫情人口流动分析、患者感染数据评价等相关领域，例如使用地区解析的元种群模型来模拟病毒的传播，结合迭代贝叶斯推理算法——集合调整卡尔曼滤波器——它吸收了每个县日常病例的观察结果，以及县之间的人口流动，模型拟合能够估计病死率 (CFR) 和感染死亡率 (IFR)，从而具备一定的疫情高峰预测能力。总的来说，在疫情人口时空轨迹分析这个方面，国外目前所做的研究较为具体，可供参考的范例较多。

- 课题现状分析：

近年来，随着计算机科学技术的不断发展，大数据分析、人工智能等等学科开始涌现并不断为各个领域的科学研究提供强有力的方法和武器，数据越来越成为一种关键的战略资源，在本次疫情防控中大数据分析也起到了重要作用。总体来说，新冠肺炎疫情患者的时空轨迹数据收集与分析这一课题贴近生活，贴近社会，贴近科学前沿，具有较好的科学理论意义以及实践应用意义，但另一方面，国内外在这一方面这一领域的研究方兴未艾，还没有成熟的理论模型以及成熟的数据分析系统，这对于学生的学科交叉分析实战能力是一个不小的考验。

3. 主要研究内容

新冠疫情的持续蔓延给人类生活和社会经济带来了严重的影响，准确预测各地区的疫情趋势对防控疫情而言至关重要。现有研究主要基于经典的时间序列预测模型和传染病模型，鲜有考虑疫情在传播过程中呈现出的地区关联复杂和时序依赖性强的特点，限制了其疫情预测的性能。为此，针对新冠疫情的预测任务，提出了一种时空注意力驱动自编码器框架，一方面通过引入空间注意力机制捕捉疫情感染序列间的动态空间关联性，另一方面利用时间注意力机制挖掘疫情序列中复杂的时序依赖性，以此实现对不同地区的新冠病毒传播趋势的准确预测。具体来说，在模型的编码器端，通过融合了空间注意力机制的 LSTM 网络，关联目标地区与其他地区的疫情序列，并从中提取该区域近期疫情的时序特征。在模型的解码器端，将时间注意力机制引入基于 LSTM 网络的解码器中，通过捕

捉疫情序列的时序依赖性推测未来的疫情变化。

4. 研究方案

课题主要围绕中国各县市自 2020 年 2 月新冠肺炎疫情爆发以来单日新增病例数的统计数据，构造一个便于检索和数据模型分析的数据库系统，在构造好的数据库系统上对新冠患者的时刻轨迹数据进行模型量化分析，结合疫情防控的重大事件与传染病医学原理分析疫情扩散模式、评测防疫措施的可行政与科学性。

数据研究方案如下：

- 数据可信度调研、数据收集、数据清洗，这一步总结为数据的抽取，是为了保证能够收集到正确可靠的数据；
- 数据的再清洗、数据仓库的建立、数据后台的建成与运作，这一步总结为数据的存储，具体是要先把数据再次校验，除去不完整和有明显错误的数，然后把它存储起来，并建立后台服务系统对数据进行维护；
- 数据的处理、数据建模、数据可视化，这一步是把海量的数据中最为关键的特征提取出来并形成科学的结论，这是最后一步，也是最为关键性的一步，同时，由于关注点的不同，提取的特征也会不同，数据中会有许多的信息可供挖掘，所以这一个步骤是反复的、无止境的。

5. 进度安排，预期达到的目标

2021 年 10 月 20 日~2021 年 10 月 25 日：与毕业设计导师一同商量毕业设计的选题；

2021 年 10 月 26 日~2021 年 11 月 08 日：初步了解和收集关于应用时空轨迹数据挖掘技术应用与疫情防控相关的科研成果，完成任务书

2021 年 11 月 12 日~2021 年 11 月 22 日：进一步搜集国内外各大期刊关于新冠肺炎患者时刻轨迹分析的最新科学研究论文或期刊的学术成果，进一步讨论关于数据分析、数据建模的基本流程与技术手段，进一步理解关于人口流动、药物非药物防控措施在病毒性疫情的传播过程中的影响，进一步加深对于课题学科交叉思想的理解认识，明确毕业设计的目标和方向，完成毕业设计开题报告；

2021 年 11 月 22 日~2021 年 12 月 01 日：准备毕业设计演讲稿，参与毕业设计开题答辩；

2021 年 12 月 02 日~2022 年 02 月 01 日：收集和整理全国各省市单日新冠患者数据；

2021 年 02 月 02 日~2022 年 04 月 01 日：将数据清洗好并建立查询检索接口，

做好数据分析、模型建立和可视化工作；

2021 年 04 月 02 日及以后：总结研究成果，撰写毕业论文。

6. 课题已具备和所需的条件、经费

已具备的条件：相关人员等

所需的经费、条件：学术期刊论文库的检索与免费下载等

7. 研究过程中可能遇到的困难和问题，解决的措施

1、数据来源可能是第一个要解决的问题，首先要可信，二是要易于采集

2、其次是数据量庞大，如何清洗和存储将是第二个问题

3、关键的问题还在于如何把数据利用好，挖掘其中的理论内涵

解决的措施包括与导师定期沟通讨论、检索电子期刊或查阅学习纸质文献等

8. 主要参考文献

【1】曹文洁,刘杰,张本龚.大数据与人工智能技术在 COVID-19 疫情防控中的应用分析[J].武汉纺织大学学报,2021,34(01):10-14.

【2】郑玲,肖淳健,陈菲,肖永红.基于区块链和物联网技术的新冠病毒感染密切接触者隔离智能监控系统的设计[J].中华医院感染学杂志,2020,30(08):1141-1146.

【3】闫雨蒙,李博,李泽宇,连博,苏祥飞,王天园,李萍,王强,程金莲,杨忠奇.新型冠状病毒肺炎临床研究和专家意见数据分析的思考与建议[J].中国中药杂志, 2020,45(7):1526-1530.DOI:10.19540

【4】漆翠芳,杨力仁,杨子轩,尚丽,谢桂兰,王瑞奇,王珊珊,杨文方.影响新型冠状病毒肺炎省际传播与发展的因素:基于 30 个省市的数据分析[J].西安交通大学学报(医学版),2020,41(5):757-763.DOI:10.7652

【5】魏剑浩,郭倩,李海聪,郜梦露,徐盛,戚伟强,凌云,韩伟靖,刘一力,朱召芹.上海地区 328 例新型冠状病毒肺炎患者实验室数据分析[J].检验医学,2020,35(8):778-783.DOI:10.3969/j.issn.1673-8640.2020.08.009.

【6】Zhou, Y., Hou, Y., Shen, J. et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 6, 14 (2020).

【7】Pan, F., Li, L., Liu, B. et al. A novel deep learning-based quantification of serial chest computed tomography in Coronavirus Disease 2019 (COVID-19). Sci Rep 11, 417 (2021).

【8】Sen, P., Yamana, T.K., Kandula, S. et al. Burden and characteristics of COVID-19 in the United States during 2020. Nature 598, 338–341 (2021).

【9】Levin, R., Chao, D.L., Wenger, E.A. et al. Insights into population behavior during the COVID-19 pandemic from cell phone mobility data and manifold learning. Nat Comput Sci 1,

588–597 (2021).

【10】 Davis, E.L., Lucas, T.C.D., Borlase, A. et al. Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. *Nat Commun* 12, 5412 (2021).

【11】 Leung, K., Wu, J.T. & Leung, G.M. Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing. *Nat Commun* 12, 1501 (2021).

【12】 Marcia C. Castro, Sun Kim, Lorena Barberia, Ana Freitas Ribeiro, Susie Gurzenda, Karina Braga Ribeiro, Erin Abbott, Jeffrey Blossom, Beatriz Rache, Burton H. Singer. Spatiotemporal pattern of COVID-19 spread in Brazil. Castro et al., *Science* 372, 821–826 (2021) 21 May 2021

【13】 鲍昕,谭智一,鲍秉坤,徐常胜.基于时空注意力机制的新冠疫情预测模型[J/OL].北京航空航天大学学报:1-11[2021-11-18].

【14】 郭文璇,房昊.人口流动、传染病蔓延应急管控与中国新冠疫情防控经验[J].中国经贸导刊(中),2021(10):47-52.

【15】金亚亚,毛晓蒙,王霞,宋頔,刘明.新冠疫情的空间分布及影响因素——基于中国病例数据的动态分析[J].陇东学院学报,2021,32(05):16-23.

【16】 栾绍童. 时空轨迹数据挖掘研究[D]. 国防科学技术大学,2019.

【17】 郑浩泉,何浩奇,刘伽椰,等. 时空轨迹数据存储方法研究[J]. 南京师大学报（自然科学版）,2017,40(3):38-44. DOI:10.3969/j.issn.1001-4616.2017.03.006.