基于大数据和人工智能技术场景融合的新型传染病预 警预测系统研发

一、立项的背景和意义

自 2019 年 12 月新型冠状病毒 (COVID-19) 在武汉爆发以 来, 截止到北京时间 2 月 15 日 24 时, 全国累计报告确诊病例 68500 例 [1]。除了病毒的传染性强等客观因素外,我国的传 染病预警系统没有讲行及时有效的预警,也是此次疫情如此严 重的原因之一。传染病的监测预警是公共卫生领域的重要组成 部分。高效的传染病自动预警信息系统能够早期发现传染病爆 发的异常情况并对其进行预警,以帮助有关决策部门尽早采取 相应的防控措施 [2]。如图 1 所示,我国目前所采用的疾病预 防控制系统于 2004 年开始投入使用,通过各医院的医疗 HIS 系统将传染病电子病例报告上传,然后通过四级公共卫生交换 平台上报,并最终到达系统的核心即传染病信息报告管理系统 (National Notifiable Diseases Reporting System, NNDRS) [3]。该系统采用固定阀值法和时间模型法对传染病疫情情况进 行计算和监测并向基层医院和疾控中心发送预警信息。然而该 套系统在此次疫情防控中, 没有起到很好的作用, 随着国家信 息化建设的进步,以及大数据和人工智能等技术的发展,建立 基于大数据和人工智能技术的新型传染病监测预警预测系统是

十分有必要的。新型的智能化传染病预警预测系统,将不仅仅针对此次新型冠状病毒,而是可以针对所有传染病,尤其是传染性较强的呼吸道类传染病进行精准的预警与预测,做到大数据和人工智能技术与疫情预警预测的场景融合。新型的预警预测系统将对当前的预警预测系统进行有效的补充与升级,帮助流行病学家更快和更准确地发现和追踪传染病,并提升政府对传染病的防控能力。

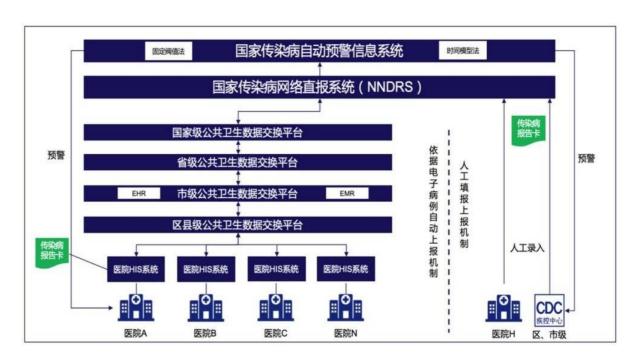


图 1. 国家传染病预警结构图

二、国内外研究现状和发展趋势

目前,在大数据与人工智能与传染病的预测预警领域,国内外已经开展了一些研究,这里就比较知名的国内外研究情况进行简单梳理。例如,在国内,中国平安与重庆疾控中心于2017

年进行了合作,建立了国内首个基于大数据和人工智能的疾病预测模型与疾病筛查模型。具体来说,该模型是通过重庆 2000万电子病例,以及天气数据,百度搜索数据以及新浪微博数据等所建立的自适应 AI 模型,将机器学习算法用在了疾病的预测领域。目前,该系统已经用于重庆疾控中心的季节性流感预防工作上,属于地方式的疾病预防预测系统平台 [4]。在国外,也有一些组织或人工智能初创公司在该领域做出了研究。比如,加拿大公司 BlueDot 通过人工智能与大数据技术来追踪传染病。BlutDot 每天会分析 65 种语言约 10 万篇文章,并追踪100 多种传染病的爆发情况,用到自然语言处理技术以及机器学习技术。该公司通过其 AI 监测平台于 2019 年 12 月份率先预测和发布了武汉出现传染疫情的新闻,引发了国内外新闻媒体对 AI 预测疾病领域的关注 [5]。

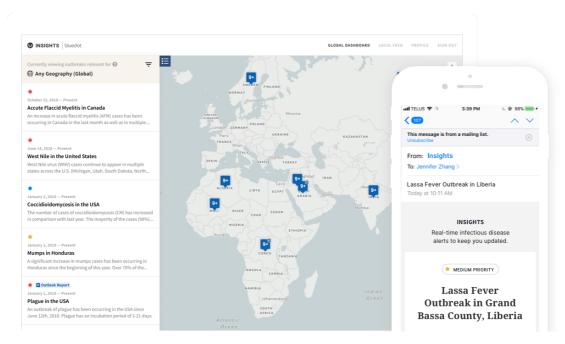


图 2 BlueDot 对其用户发送疾病预警示意图

除了 BlueDot 外,HealthMap 是由美国波士顿儿童医院的研究员、流行病学家和软件开发人员等组成的团队所建立的一个可免费访问的自动化电子信息系统。该系统可对全球的传染病爆发情况进行监测,组织和可视化。HealthMap 的用户包括美国的州和地方公共卫生机构,世界卫生组织(WHO),美国疾病控制和预防中心以及欧洲疾病预防和控制中心 [6]。

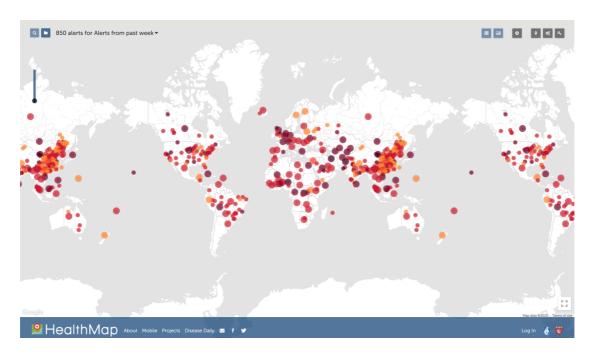


图 3 HealthMap 网站截图

国内目前的传染病预警预测系统需要各地方疾控中心或地方医院将传染病情况层层上报,并通过国家传染病预警系统统一发送预警信息,因此预警信息不及时等情况会发生[3]。从国内外现状分析中可以发现,目前国外的大型公共卫生组织如WHO,美国疾控中心以及重庆疾控中心等,都已经对智能化的

疾控系统进行了尝试。因此,基于大数据与人工智能技术场景融合的新型传染病预警预测系统的研发可以作为第三方软件平台对现有的国家预测预警系统进行很好的补充,而且也很符合当前的市场发展趋势。

产品市场容量分析

截止到北京时间 2 月 15 日 24 时,根据国家卫健委的官方数 据,在此次新型冠状病毒中,我国累计确诊病例 68500 例;现 有疑似 8228 例:累计死亡 1665 例 [1]。有经济学家预计,在 疫情的影响下,我国一季度的经济将遭受重创,其中仅仅餐饮 行业在第一季度就起码有 5000 亿元的损失 [8]。而旅游服务行 业也完全遭到冻结,游客纷纷取消出游行程,航空公司、酒店 等退票退单的订单量巨大 [10]。根据相关专家预测此次疫情至 少对全国旅游业造成 3 万亿的损失 [11]。疫情也对电影行业造 成巨大损失。2019年的春节档票房为58亿元人民币左右,而 今年电影院全部撤销春节档,影院没有任何收入 [12]。在交通 运输上,疫情发生后的交通运输渠道旅客人次大幅下滑,春节 黄金周 (除夕-大年初六)出行人数同比 2019 年下降 63.9% [11]。根据交通运输部的数据,截止到2月5日,春运旅客运 输量为 13.18 亿人次,同比下降 35%,公路客运 10.7 亿人 次,同比下降36.3%,预计春运后半程的客运总量为4亿人 次,同比下降约70%[16]。我国2019年第一季度的交通运

输、仓储和邮政业的经济增加值 (GDP)为 9386.6 亿元 [15]。 因此,按照交通运输减少三分之一来估算并忽略邮政业的影响,粗略估计今年第一季度交通运输业受疫情损失约为 2000 亿元。而由于疫情防控需要,许多企业目前还未恢复生产,预计中国 6300 万中小企业短期内或将出现资金流动问题,如果维持当前不能复工的现状,大多数中小企业只能维持三个月 [13]。可见此次爆发的疫情已经从多个方面对我国造成了严重的影响。

在 SARS 疫情爆发后的 2003 年下半年,中央拨款建立新的传染病与突发公共卫生事件网络直报系统 [9]。从中国疾控中心搭建这套新的网络直报系统,至 2014 年 4 月 1 日正式运行,整套系统共耗资 11 亿元 [14]。就市场容量来分析,在新冠病毒(2019-nCoV)的疫情结束后,由于此次疫情的预警不力,预计我国会继续加强在公共卫生以及疫情预警预控方面的支出, 对传染病预测预警系统进行升级以加强我国在疫情预警预测的能力。目前基于大数据与人工智能技术场景融合的传染病预测预警系统还属于新兴技术领域,因此本课题所研发的产品的市场前景是很巨大的。

三、项目预期目标

(一)项目研究内容

1. 项目名称

基于大数据和人工智能技术场景融合的新型传染病预警预测系统研发

2. 研究内容

关于基于大数据和人工智能技术场景融合的新型传染病预警预测系统研发,具体研究内容包括:

(1) 通过网络爬虫等技术对传染病(新冠肺炎,H1N1, SARS)等关键信息进行大数据搜集。爬虫技术拟使用 python来实现数据搜集。通过对目前比较流行的网络爬虫技术框架进行比较后,分析并尝试了 Requests 库, BeautifulSoup 库和Scrapy 框架,然后决定使用 Scrapy 框架进行本项目爬虫部分的开发(https://scrapy.org/),Scrapy 的运行截图如图 4 所示。

```
| Computation | Design Console | TERMINAL | Problem | Pr
```

图 4 使用 Scrapy 后台爬取截图

本项目研究所使用的大数据获取渠道包括:世界卫生组织 (WHO),国家疾病预防控制中心,省级、市级的疾病预防控制中心,专业的医疗期刊杂志等,预计可能共需要研究数百万 篇相关文章来满足训练深度学习模型的需要。目前确定的数据来源为:

- 世界卫生组织 (WHO): https://www.who.int/
- 中华人民共和国国家卫生健康委员会(NHC):
 http://www.nhc.gov.cn/wjw/xwdt/list.shtml
- 中国疾病预防控制中心 (CCDC):
 http://weekly.chinacdc.cn/en/zcustom/currentVolum
 e/1
- 中国香港政府新闻网:
 https://www.news.gov.hk/eng/categories/health/index.html
- 新英格兰医学杂志 (NEJM): https://www.nejm.org/
- 柳叶刀 (Lancet): https://www.thelancet.com/
- 美国医学会杂志 (JAMA): http://jama.ama-assn.org/
- 英国医学期刊 (BMJ): http://www.bmj.com/

本项目除了通过爬虫去搜集数据外,还考虑使用"百度舆情"SaaS 作为数据源的补充。百度舆情 SaaS

(https://cloud.baidu.com/product/byapi.html) 服务覆盖超过 8000 万家网站,提供自定义数据源的订制数据服务,可考虑通过订阅该 API,丰富数据来源。



图 5: 百度舆情 SaaS

百度舆情 SaaS 的数据提供订阅 API 能够覆盖百度搜索、百度贴吧、微博、微信、新闻、论坛、博客等全网数据源; 日采集数据源达 1 亿条以上。本项目可提供关键词如 "SARS"," 肺炎"或"咳嗽"等与传染病有关的关键词,通过百度舆情 SaaS来获取来自新闻、百度搜索等大量有效数据。



图 6: 百度舆情 SaaS 产品功能截图

在收费标准方面,百度舆情 SaaS 是根据关键词数量进行收费,最低关键词购买数量为 10 词,最低购买年限为 1 年,最低使用费用为 **1 年 30000 元人民币** (参考图 7)。

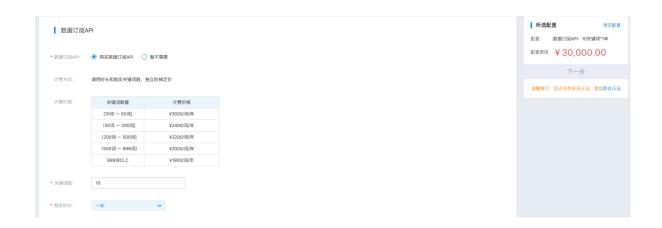


图 7: 百度舆情 SaaS 产品收费情况截图

(2) 对获取到的多源大数据进行集成并且将处理后的大数据存

入分布式数据库。通过爬虫搜取到的文章大多数都是非结构化的数据,通过分词,分句,命名实体识别等自然语言处理技术 (NLP) 可将大数据进行有效的预处理,并转化为统一的格式存入分布式数据库如 MongoDB 中, 以有利后续进行于大数据分析。由于大数据的来源广泛,数据量巨大,因此需要克服在大数据采集过程中对多源数据的处理以及对相关数据的存储等难点问题。

- (3) **在算法层面,通过神经网络构建预警预测模型**。可利用长短期记忆单元(LSTM)构建递归神经网络来对传染病进行定量预警[7],或者将传染病动力学模型、图结构、神经网络以及传染病固定阈值法[2]等相结合,来对传染病发展趋势进行预测。
- (4) 在软件与服务层面,通过构建可视化应用软件,实现人工 智能模型与疫情预警预测的场景融合。具体来说,在场景融合 上,做到可视化与交互性两个方面。在可视化上,利用前端技 术 (HTML, CSS, JavaScript) 和后端技术 (NodeJS), 以及数 据可视化库如: ChartJs (https://www.chartjs.org/), Echarts (https://www.echartsjs.com/en/index.html) 和 Leaflet (https://leafletjs.com/) 等来搭建一个响应式的疫情预警预测可视化 网站。网站应可以部署在移动端, PC 端以及控制中心大屏幕端 等。并针对不同端,进行相应的优化以达到最优的数据表现方 式。普通用户应可以使用移动端和 PC 端,并得到疫情的预警 信息以及基本的疫情预测信息,但是无法进行场景融合的模拟 操作。而针对疾控中心等政府职能机构用户,允许其模拟不同 疫情防控措施后疫情的发展效果,具体包括以下四个措施:城 市居民自我隔离,切断城市交通,城市封闭化管理,城市军事 化管理。通过允许政府相关机构在软件上模拟在四个不同防控 措施下的疫情扩散情况,以让有关部门作出最优的疫情防控决 策。

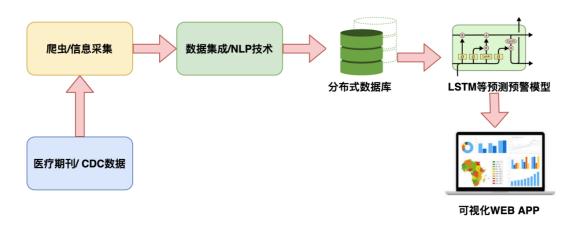


图 8 研究内容的简化流程图

3. 关键技术

本项目的关键技术包括三项,首先通过自然语言处理技术,从与传染病相关的非结构化大数据中集成可用的有效数据,并建立分布式数据库。数据的集成部分首先会将爬虫抓取的文章进行分词,分句以及实体命名识别,然后通过建立并训练深度学习语言模型,将输入的文章与新闻进行处理,并将有效的文章筛选出来,并输出文章中对应传染病名称,地点,时间等有效的结构化数据信息。第二项关键技术是在算法层面利用神经网络构建预测预警模型。例如在疫情预警方面可以使用基于LSTM的模型,并使用上一步所创建的数据库里的数据对模型进行训练。在疫情预测方面,使用基于传染病动力学模型 SEIR与神经网络所结合的模型或使用基于多图卷积的图结构所构成的模型。使用历史真实数据如 2003 年的 SARS 数据和现在的

新冠病毒 (2019-nCoV) 的数据来训练预测模型,用均方根误 差 (RMSE) 来评估预测率以验证模型的精确程度。第三项关 键技术为构建疫情预警预测可视化软件。软件为 web 应用程 序,应用场景为手机端、PC 端或控制中心大屏幕端等。软件开 发采用的架构为前端后端分离,前端后端之间用 JSON 来进行 通信。此外,前端采用 Echarts 和 Leaflet 等 JavaScript 库来 作为图和表的数据展示容器。图和表包括散点图,折线图,以 及用于地理可视化的地图, 热力图等。数据库拟采用分布式数 据库 MongoDB, 利用分布式设计,来提高系统的响应能力, 增强系统的性能。此外,不同用户拥有不同的权限,普通用户 拥有得到基本预警信息的权限以及基本的传染病预测信息的权 限。而针对高级别用户如疾控中心和政府相关管理部门,允许 这些用户能够通过本软件实现场景融合, 做到模拟城市居民自 我隔离, 切断城市交通, 城市封闭化管理, 城市军事化管理这 四个不同疫情防控措施下疫情的防控效果。

4. 主要创新点

我国目前的传染病监测预警系统是基于 NNDRS 系统的,其预警是通过各级各类的医疗卫生机构进行传染病病例上报,并需要各地的医疗机构再进行预警信号核实和现场调查确认等响应 [3]。在目前的国内外疫情预测预警研究和系统中,还没有相应的研究将大数据、人工智能技术与疫情防控做到场景融合。不

同于中国平安等相关研究使用的医疗健康数据,本项目的一大创新是使用了自然语言处理等人工智能技术,在一些杂志与新闻报道等大数据中去筛选有效数据信息, 并通过训练后的深度神经网络模型来预警疫情。本项目可对目前的监测预警系统进行智能化的补充,通过 AI 智能分析提前预测传染病的发生情况。此外,本项目做到了疫情防控的场景融合,针对特定用户如政府决策层,允许这些用户利用本软件对防疫四个措施(城市居民自我隔离,切断城市交通,城市封闭化管理,城市军事化管理)进行模拟,并通过软件可视化展现给用户采取不同措施后疫情防控所达到的效果。而针对普通用户,虽然他们不能进行防疫措施模拟,但依然可以获得预警信息以及对当前疫情的预测信息。

5. 技术指标

在数据的处理与整合方面:对传染病相关的大数据进行集成,并建立分布式数据库来存储处理后的数据,数据库预计将存储数百万条相关数据。

在算法方面:通过输入数据库中的数据,来训练基于深度神经网络的语言模型,并对传染病进行精确的预警与预测。在疫情预警方面,因为自然语言处理、机器学习与疫情预警的交叉领域目前缺乏足够多的研究,还没有公认的基准指标,因此本项目将建立该领域的基准,为后续的相关学者研究提供参考。

在软件方面:建立功能性强的,用户体验好的疫情数据可视化web应用程序。在交互性上实现场景融合,做到对防疫四个措施(城市居民自我隔离,切断城市交通,城市封闭化管理,城市军事化管理)的有效模拟。同时,软件应能够运行在主流浏览器上,并针对手机端、PC端、控制中心大屏端等不同端进行对应的优化。例如,手机端和PC端主要针对普通用户,这类用户可以获得预警信息以及基本的疫情预测信息(疫情走势图等),但无法调整系统的参数设置,只能看到目前疫情的预警预测现状。而控制中心大屏端主要针对政府决策部门如疾控中心等,该类用户拥有更高级别的系统权限,被允许进行场景融合的四大防疫措施模拟,可以调整系统参数设置,来对四层防疫机制进行选择以选取最优的防疫措施。

6. 主要功能及应用领域

该项目的主要功能包括使用基于大数据和人工智能机器学习等技术,来实现对传染病防控的场景融合与预测预警。通过 web 程序将与预测预警相关的传染病数据进行可视化,通过场景融合,允许用户对不同的防疫措施进行模拟,以帮助用户对防疫措施进行最优选择以及对疫情发展趋势等作出更直观的判断。

本项目主要应用在疾病预防预控与公共卫生领域。

(三)成果指标

国际会议论文投稿 "Early-warning Infectious Disease Framework based on Machine Learning", International Conference on Knowledge Discovery and Data Mining (KDD 2020) Workshop.

Reference

- [1]http://www.nhc.gov.cn/xcs/yqfkdt/202002/4a1b1ec6c03548099de1c3aa935d04fd.
 shtml (Assessed on 16 February 2020)
- [2] 张洪龙, et al. "2016 年国家传染病自动预警信息系统运行情况分析." 疾病监测 33.2 (2018): 159-167.
- [3] 赵自雄, 赵嘉, and 马家奇. "我国传染病监测信息系统发展与整合建设构想." *疾病监测* 33.5 (2018): 423-427.
- [4] Su, Kun, et al. "Forecasting influenza activity using self-adaptive AI model and multi-source data in Chongqing, China." *EBioMedicine* 47 (2019): 284-292.
- [5] http://www.elecfans.com/rengongzhineng/1162099.html (Assessed on 16 February 2020)
- [6] Nelson R (2008). HealthMap: the future of infectious diseases surveillance? Lancet Infect Dis 8: 596.
- [7] 黄鹏. *基于机器学习的乙类传染病预测模型研究与实现*. MS thesis. 电子科技大学, 2019.
- [8] http://www.bjd.com.cn/a/202002/17/WS5e49f0e0e4b00949486813ea.html
 (Assessed on 18 February 2020)

[9] https://money.163.com/20/0209/07/F4U76MM300258105.html

(Assessed on 18 February 2020)

[10] http://finance.sina.com.cn/chanjing/gsnews/2020-01-31/dociimxxste7896790.shtml (Assessed on 20 February 2020)

[11] https://finance.sina.cn/china/cjpl/2020-02-08/detail-iimxyqvz1258307.d.html

(Assessed on 20 February 2020)

[12] https://www.bbc.com/zhongwen/simp/business-51355815 (Assessed on 20 February 2020)

[13] http://www.asianews.it/newszh/%E5%8C%97%E4%BA%AC%EF%BC%8C%E5%8F%97%E6%96%B0%E5%8
6%A0%E7%97%85%E6%AF%92%E5%BD%B1%E5%93%8D%E6%9C%80%E5%8
6%A0%E7%97%85%E6%AF%92%E5%BD%B1%E5%93%8D%E6%9C%80%E5%A4%A7%E7%9A%84%E4%B8%AD%E5%B0%8F%E4%BC%81%E4%B8%9A49276.html (Assessed on 20 February 2020)

[14] http://www.bjnews.com.cn/news/2020/02/04/684397.html

(Assessed on 20 February 2020)

[15] http://data.stats.gov.cn/tablequery.htm?code=AB0304

(Assessed on 21 February 2020)

[16] <u>http://cppcc.china.com.cn/2020-02/06/content_75679151.htm</u>

(Assessed on 21 February 2020)