大数据抗击新冠肺炎疫情的实践与思考

刘跃文
西安交通大学

关键词：大数据　新冠肺炎风险预测　隐私保护　计算效率

2020年春节，新冠肺炎（COVID-19）疫情的突然爆发让大家感到措手不及。与2003年非典型性肺炎（SARS）疫情相比，新冠病毒的感染能力更强，也更加狡猾;而且2020年整个社会的流动性超大幅度提升，这使得病毒的传播的速度更快、范围更广。统计数据表明，2002年全年航空旅客客运量为8600万人次[1]，而2019年为6.6亿[2]，翻了7.7倍；2002年全年铁路旅客客运量为10.6亿人次[1]，高铁尚未建成，而2019年全年铁路旅客发送量为36.6亿[3]，翻了3.6倍，高铁也已成为主要的客运方式。这都为新冠病毒在非常短时间内扩散到全国提供了极其便利的条件。

幸运的是，信息技术（如移动互联网、云计算、大数据等）的发展，为防控新冠疫情提供了强有力的武器。特别是大数据的发展，为准确地计算传染风险、评估防控策略、追溯感染人群等都提供了良好的数据基础和可行性。新冠肺炎疫情爆发期间，笔者及团队成员参与了云南省的大数据抗击疫情工作，与合作单位一起，迅速研发了一套可以一秒钟计算感染风险的大数据模型和系统，也研发了可以快速追溯感染人群的扫码系统，有效地帮助了云南省的疫情防控。

大数据抗疫情：外防输入、内防扩散

疫情传播期间，全国各地人员流动量极大且结构复杂，交叉感染风险极高，加之云南是冬季旅游的热门地区，如果不能有效地控制云南的疫情，就有可能会影响到全国的疫情防控。

疫情防控初期的核心工作是“外防输入”，当时的主要方法是利用旅行大数据，筛查之前去过疫源地（湖北）的人员并输出名单。筛查过程中发现了一些问题：有很多人在乘坐公共交通工具的过程中和来自疫源地的人，甚至是确诊病患接触过，但自己根本不知情；还有些人近期去过疫源地或接触过来自疫源地的人，但是自己觉得无所谓或者干脆不承认。这两个问题在基层一线摸排和核查时特别突出。基层核查人员只能依据身份信息找到武汉人、湖北人。而这种方法并无法及时获取相关数据并形成综合风险判断。为了解决基层反馈的问题，我们基于2015年发表的一项科研成果[4]，快速地搭建了一个新冠肺炎感染风险预测模型。这个模型基于个人的旅行数据，分析其是否到过疫源地、是否与疫源地人员接触过、是否与已感染病患接触过等多项指标，利用贝叶斯方法，计算其感染新冠病毒的可能性指数,并预警高风险人员。模型训练完成后，我们又研发了新冠肺炎感染风险实时计算及预测系统。该系统可以通过扫描身份证、网页查询、批量计算等方式投入实际应用，计算一个人感染风险的时间不到一秒钟，从而快速地锁定风险。

2月10日之后，随着各地逐步开始复工复产，疫情传播方式发生了一定的变化，即从“输入型”

传播向“聚集型”扩散转变。如果公共场所人员流动大幅增加，新冠疫情聚集型扩散的风险就会增大。当时媒体上频繁出现“紧急扩散”寻找公共场所中接触人员的新闻，这样的情况促使我们设计了一种能快速准确地追溯公共场所内接触人员的系统，变“紧急扩散”为“主动找回”。经过反复研讨，我们研发了一个基于微信小程序的追溯接触人员的轻量化系统“扫码抗疫情”。市民在进入和离开公共场所时，用微信扫描二维码便可登记其停留信息。一旦有人被确诊，就可以通过分析后台积累的扫码数据，快速找回确诊病患在公共场所中可能接触的所有人员。其系统示意如图 1 所示。

图 1　扫码抗疫情系统示意图

上述两个系统都得到了广泛的应用和认可。计算新冠感染风险的系统被部署在数万台终端上，核查数百万人次，发现 500 多名感染高风险人员。注册“扫码抗疫情”小程序的公共场所数量超过 100 万，扫码用户数超过 2000 万，扫码量达到了 5 亿多人次，找回病患接触人员 200 余名。

新冠疫情期间的大数据研发不仅是科研工作，而且是将科研成果迅速转化为战斗力的“实战”工作。在工作过程中，我们对很多问题进行了反复的思考和探讨，包括模型选择的策略、数据采集的策略、计算效率提高的策略等。正是因为对这些问题的深入讨论和妥善解决，才使得大数据抗击疫情的工作得到有效开展和快速推进。

模型选择：“准确性”与可解释性的权衡

在进行新冠肺炎感染风险预测的过程中，遇到的第一个问题就是模型的选择。在开发过程中，能够从样本中采集到的特征较多，包括是否为疫源地人（如湖北人、武汉人）、是否到过疫源地（如是否从湖北、武汉乘坐飞机、火车抵达云南）、是否与疫源地人员接触过（如是否与湖北、武汉人乘坐相同航班、火车等）、是否与已确诊病患接触过（如是否与已确诊病患乘坐相同航班、火车等）、出行次数、出行天数、旅行城市数、年龄等多项指标。能够采集的样本主要是当时云南已经确诊的数十名新冠肺炎病患。

如何利用这些数据构建一个判断是否感染新冠病毒的模型？这似乎是一个非常传统的分类问题。我们尝试了各种解释性较强的机器学习方法，如决策树方法、逻辑回归方法等。尽管这些方法的模型训练结果都还不错，但在对模型的可解释性进行分析的过程中，出现了一系列不易解释的参数。例如，决策树会反复对某一个特征进行判断，逻辑回归的某些系数的正负号与预期不同。考虑到样本量较小，我们认为这些情况可能是由于过拟合造成的。如果过拟合的模型投入实际应用，可能会造成误判，风险较高。

为了解决这一问题，我们使用了一个更为简单的、解释性更强的方法[4]，该方法发表于国际期刊《决策支持系统》（*Decision Support Systems*）上，是一个用于好友推荐的解释性极强的方法，即计算一个人“可能被认识”的概率。我们将这一逻辑迁移到抗击新冠肺炎的场景下，变成计算一个人“可能被感染”的概率。简单地说，如果观察到证据 A 时，其患新冠肺炎的概率为 P(新冠 | 证据 A)；观察到证据 B 时，其患新冠肺炎的概率为 P(新冠 | 证据 B)；假设 A 和 B 是相互独立的，那么根据概率公式，感染新冠肺炎的概率为 P(新冠)=1-[1-P(新冠 | 证据 A)]*[1-P(新冠 | 证据 B)]。这一方法有若干好处：一是容易计算，只需要分别统计单一证据对应的新冠肺炎感染概率即可；二是容易扩充，如果新增证据 C，只需要在公式后面乘以 [1-P(新冠 | 证据 C)]；三是容易解释，计算结果即为感染新冠肺炎的概率，可

以直接根据概率判断风险高低。

这一方法的使用使得模型的计算结果易读易用。我们在实际应用中计算了两个分指数，即疫源地感染指数和二次传播感染指数，用于评估其病毒感染的来源。我们还根据概率得分的高低，划分了高、较高、中、较低、低 5 个风险等级，进一步提高结果的可读性。此外，还在系统中提供了概率得分的一些关键证据，帮助基层核查人员理解计算结果。我们认为，只有当模型的计算结果 100% 可解释时，才能放心地将结果应用在实际的核查工作中，否则可能会给很多人造成不必要的困扰。

少就是多：数据采集与隐私保护

在扫码抗疫情小程序的研发过程中，大家对于要采集的信息项进行了讨论，出现了若干种不同的意见。一种意见是应采尽采，采集每个人的身份证号、姓名、手机号、家庭住址、近期到访的城市等；另一种意见是能不采就不采，只采集手机号，只要确保能联系到就可以；还有一种意见较为折中，采集身份信息，如身份证号、手机号、姓名等。经过我们的反复讨论，最终决定采取第二种方案“能不采就不采”，只采集手机号码。当一个人首次扫码时，需要填写手机号并短信验证，之后只需扫码即可，不需要填写任何信息。

这其中的原因，一方面是对隐私数据的顾虑，我们不希望占有过多的隐私数据，采集的隐私数据越多，整个系统隐私保护的压力就越大，数据被滥用的风险就越高。另一方面，根据多年来的大数据

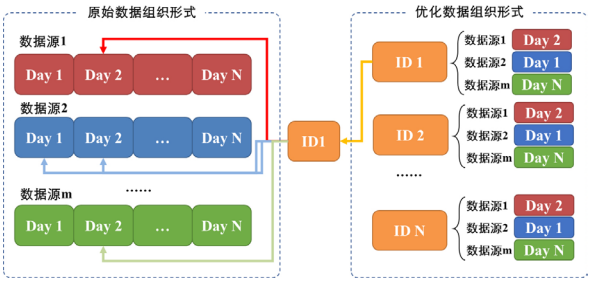

图 2 数据组织形式优化

工作经验，我们深知数据量和数据质量之间的矛盾。当要求数据量时，数据质量就很难得到保证；而当要求数据质量时，就不能过于苛求数据量。更直白地讲，“少就是多”。在新冠疫情采集数据的工作中，采集的字段越少，对大家生活的打扰越少，大家才更愿意参与到扫码中，数据量才能得到保证。相反，如果采集的字段较多，大家可能会有隐私泄露的担忧，不愿意配合扫码；而且当采集字段较多时，填入的数据有可能是假数据，会对后续的大数据分析产生干扰。

隐私数据的采集和使用需要非常小心。扫码抗疫情系统在互联网上运行，每一分钟都有被滥用甚至被恶意攻击的风险。小程序上线后第 4 天，就在后台系统的某台服务器上发现了木马，我们在第一时间更换了服务器，没有造成任何损失。这一事件为数据的安全性敲响了警钟。为了保护抗疫情数据不外泄，在专业的安全团队的帮助下，我们对系统做了全面的防护。尽管付出了一定的代价，但是非常值得。

秒级响应：计算效率与后台设计

要服务抗疫情实战，快速计算和响应是一个不可避免的核心问题。在研发新冠感染风险计算系统时，我们发现，在原有的数据组织形式下，只获取一个人的全部旅行大数据相关信息就需要十多分钟，更不用说计算其同行数据、计算概率得分并推送结果的时间。如果在高速公路口，基层核查人员扫描身份证后要等待十多分钟才能获取风险指数，这样可能会使基层人员较长时间暴露在风险中，也可能会使高速公路上排起长队，影响整个通行情况，这显然是不可行的。

因此，必须要优化数据的获取和计算时间。为了解决这一问题，我们分析了数据存取现状（如图 2 中左图所示）。我们发现旅行大数据的存储方式是按照数据来源（如火车、公路、飞机等）分类存储的，每个数据源每日一张表。因为不知道要查询的 ID 数据在哪个数据源、在哪天中出现，就不得不遍

历所有的数据来获取该 ID 的信息，这就会花费大量的时间。针对这一问题，我们采用了“空间换时间”策略。如图 2 右侧所示，我们重新组织了所有的旅行大数据，以 ID 为索引，将相同 ID 的所有数据串到一起，使并发检索 1000 个 ID 数据的时间控制在 1 秒之内。通过这样的优化，扫描身份证后获得概率得分的时间就会大幅度缩短，实现了秒级响应，解决了基层核查人员扫描身份证后等待时间过长的问题。

扫码抗疫情系统上线后，从开始扫码到验证通过的时间大约为 4 秒。为了尽量减少对人们生活的影响，我们力求缩短这一时间，主要做了三方面工作：一是采取弹性负载均衡服务器将扫码请求发送到多个二级反向代理服务器上，再通过分布式数据库等技术提高后台处理的效率（如图 3 所示）；二是将接收扫码数据的生产库与分析扫码数据的分析库分离，生产库中只保留一天的数据，从而提高数据库的性能；三是优化扫码时的信息处理流程，减少不必要的环节，将扫码结果第一时间反馈到客户端。通过这三方面的改造，扫码验证时间从 4 秒缩短到了 1 秒，基本上做到了“即扫即走”，不需要等待，用户体验得到了大幅度改善。按照每天 2000 万次的扫码量来计算，每次扫码减少 3 秒，每天就可以减少 6000 万秒的等待时间，节省的时间是相当可观的。

最后九十公里

新冠肺炎疫情期间，很多工程师被困在各地不能返回工作岗位，我们不得不既承担设计师（数据分析及建模）的角色，又承担工程师（技术实现及优化）的角色。这次宝贵的经验使我们认识到，工程实现并不是最后一公里，而是最后九十公里。从一个理论上可用的模型出发，到完成计算、研发应用系统，并满足一定的性能要求，还有很长的路要走。

在抗疫期间完成的这些系统，都是十万火急同时也要精益求精的。新冠感染风险评估系统从模型设计到系统上线只用了 4 天时间，而扫码抗疫情小程序从设计到上线只用了 2 天时间。整个研发过程充满了各种挑战，仅是确诊患者名单就因为各种原因修订了十多回。扫码抗疫情系统上线后，有非常多的环节出现意想不到的问题，导致运行极度不畅。然而，疫情当头，所有人都顶住了压力，成为了抗疫情的“即战力”。大数据抗疫情的工作中，我们取得了不错的成绩，也积累了很多实战的经验。而这些经验对于以后开发类似的应急项目都会有一定的借鉴价值。

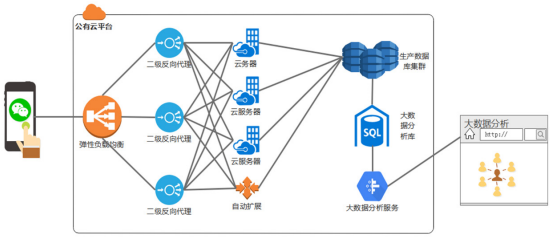

图 3 扫码系统的后台设计（云服务）

致谢：本文中的研究工作得到了国家自然科学基金应急项目 NSFC 62041206、国家自然科学基金面上项目 NSFC 71871179 的资助。特别感谢云南省公安厅科信处在疫情期间的共同奋战与全力支持。 ■

参考文献

[1] 中华人民共和国国家统计局（编），中国统计年鉴 2003（总第 22 期），北京：中国统计出版社，2003.7

[2] 中国民航网，“2019 年中国民航运输旅客 6.6 亿人次”，http://www.caacnews.com.cn/0103/0103/202001/t20200106_1288984.html

[3] 交通运输部，“2019 年铁路完成客运量 36.6 亿人次”，https://baijiahao.baidu.com/s?id=1667084939794752893

更多参考文献：http://dl.ccf.org.cn/cccf/list

刘跃文
西安交通大学副教授。主要研究方向为大数据管理、社交网络、电子商务等。
liuyuewen@mail.xjtu.edu.cn