·研究快报·

利用网络爬虫技术分析我国活禽贸易与H7N9禽流感病毒传播的关系

卢珊　陈晨　于伟文　王海印　杜鹏程　阚飙　徐建国

【导读】 采用网络爬虫技术分析2013年中国各省份通过活禽贸易携带H7N9禽流感病毒活禽的可能传播方向和范围，并预测疫情发展趋势。数据分析显示，有18个省份存在高感染风险，其中13个省份截止2014年2月已有报告病例。预测5个无感染风险的省份迄今未报告病例。

【关键词】 人感染H7N9禽流感；暴发；网络数据

Investigating geographical spread of the human infection with avian influenza A (H7N9) virus by online knowledge analysis of the live bird trade with a distributed focused crawler

Lu Shan[1], *Chen Chen*[1], *Yu Weiwen*[1,2], *Wang Haiyin*[1], *Du Pengcheng*[1], *Kan Biao*[1], *Xu Jianguo*[1]. 1 *State Key Laboratory for Infectious Diseases Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing* 102206, China; 2 *College of Software Engineering, Beijing University of Aeronautics & Astronautics*

Corresponding author: *Xu Jianguo*, Email: xujianguo@icdc.cn

This work was supported by grants from the National Natural Science Foundation of China (No. 81290345) *and the National Key Program of Mega Infectious Diseases* (No. 2011ZX10004-001).

【Key words】 Human infection with avian influenza A(H7N9); Outbreak; Internet data

流行病学调查显示，约80%的人感染H7N9禽流感病例有明确的活禽市场暴露史[1]。显然，这些病例是在活禽市场通过某种传播机制而被感染。在流行地区、流行期间关闭活禽市场可以防止新感染病例出现，是控制疫情的有效手段[2]。可是，携带H7N9禽流感病毒的活禽一般不表现出明显的临床症状，可能会通过活禽贸易而扩散，导致疫区扩大。因此控制疫情发展，必须要控制携带H7N9禽流感病毒的活禽贸易。但采用常规方法通常很难获取活禽贸易的数据，特别是在疫情流行期间。“人肉搜索”和“分布式聚焦爬虫技术”是使用网络大数据解释研究人群行为学的方法[3]。本研究组曾使用该项技术，分析活禽交易信息，为推测H7N9禽流感病毒通过活禽贸易的传播和分布，提供了有益的信息[4]。

首先利用“人肉搜索”和“分布式聚焦爬虫技术”获得活禽市场交易的海量网络数据，收集所有相关公共网站含义明确的交易信息，分析这些数据在网络出现的频率，评估各省份之间活禽交易情况和交易量。使用基于内容的链接(context-based connection, CC)和基于信息的链接(information-based connection, IC)分别定义以上两种方式获得的网络数据，并分别在“省”和“市”两个层次上评估活禽交易活动。CC数据是由人工搜索大量网络数据根据内容直接获取。建立搜索关键词和共计367个城市名字的索引库，并根据这些关键词获得与城市名称及活禽贸易相关的网页。所有查询结果中需包含城市之间连接信息和源网站或互联网协议地址(IP地址)以备抽选单独验证，确保其准确性。最后利用程序比较网页内容，删除重复的查询或者冗余的网页，进而通过比较分析这些非冗余的网页信息，若在一个页面中搜索到多个(≥2个)城市，则定义相应城市间存在活禽贸易联系。所有城市之间的活禽交易相关网络信息采用2人独立查询以保证数据完整可靠。共搜索得到与211个城市相关的315个网页，建立了591项对应这些城市所在省份的CC信息并纳入后续研究。

获得CC数据后，利用“分布式聚焦爬虫技术”从244个公共新闻网站(如新浪、搜狐、雅虎等)以及56个论坛和微博批量获取网络数据信息。搜索的关键

DOI: 10.3760/cma.j.issn.0254-6450.2014.03.002

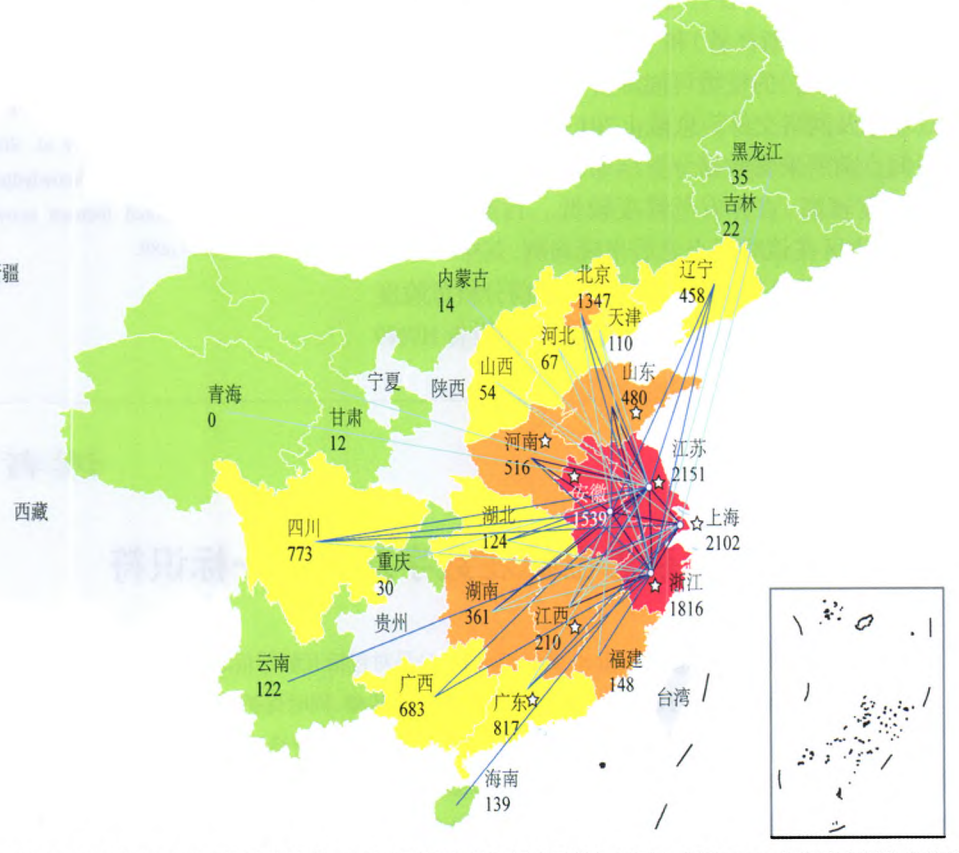
基金项目：国家自然科学基金(81290345)；国家科技重大专项(2011ZX10004-001)

作者单位：102206 北京，中国疾病预防控制中心传染病预防控制所 传染病预防控制国家重点实验室 感染性疾病诊治协同创新中心(卢珊、陈晨、王海印、杜鹏程、阚飙、徐建国)；北京航空航天大学软件学院(于伟文)

通信作者：徐建国，Email：xujianguo@icdc.cn

词使用的格式为“城市名”和“活家禽/家禽/鸽/鹌鹑/迁徙鸟/鸡/鸭/鹅”和“来自/交易/来源/运输/旅游/供应/市场/屠宰场/加工厂”。对符合要求的网页去除冗余信息，然后分析2个城市相关网页数并计算两城市间IC，再分析有CC数据支持或无CC数据支持的IC数据分布。相对于省际间的联系，城际联系由于缺乏数据或连接较少，很难精确分析这些城市的交易环节，尤其是在我国西部地区。噪声率的计算方法：①首先通过相关关键词搜索得到粗略数据；②再利用更具体、相关的关键词，搜索和生成更精确的数据集；③第一和第二次采集的数据分别采用不同的错误率计算。为排除来自网站的噪声信息，基于CC数据评估IC数据的假阳性率和假阴性率、设定临界值，最终确定10和20分别作为“省”和“市”的临界值。利用CC和IC两种数据支持的网页数量计算和比较每两省间的相关性，其感染风险采用2个参数予以评估：①与疫情暴发区域有潜在联系的省和市；②包含关键词和2个省市名称的搜索数量。

本研究共分析了835 635个网页，最终获得2 943个与搜索关键词相关的来自不同省市的网页。其中，上海地区及其相关网页有922条与活家禽交易信息，网站内容检索显示上海市活禽主要来自江苏、浙江、安徽省。利用精确关键词的方式，估计这些查询的噪声为45.17%。因此可以确定收集的数据较为准确，并可去除噪声的影响，用来估计禽类交易的动向。2013年4月6日上海市政府下令关闭农贸市场，该举措改变了活禽贸易的方向和活跃度。因此，将2013年的H7N9禽流感疫情分成两波，即4月5日以前为第一波，4月5日后为第二波（图1）。4月5日以前的网络信息分析表明，上海市活禽交易主要来自江苏、浙江、安徽等省，在搜索到的682个网页中，84%（571个）的查询结果源自上述三省。该结果与上海市政府官方声明该地区有80%的活禽交易来源于上述三省相一致。



注：根据查询活禽交易信息的CC和IC数据拓扑网络计算绘制，显示省际活禽贸易频率及其与预测和实际H7N9禽流感病例分布的关联。红色为2013年第一波发生H7N9禽流感病例的4个省份（江苏、安徽、浙江、上海）；橙色为第二波发病的6个省份（北京、山东、河南、湖南、江西、福建）；黄色为与第一波发病省份中>2个具有活禽交易网络信息的8个省份（辽宁、河北、天津、山西、湖北、四川、广西、广东），认为通过活禽贸易发生H7N9禽流感疫情的危险度较高；绿色为仅与第一波发病省份中1个有交易信息的8个省份（黑龙江、吉林、内蒙古、甘肃、青海、重庆、云南、海南），认为危险度较低；灰色区域的8个省份（新疆、西藏、宁夏、陕西、贵州、香港、澳门、台湾）未搜索到与第一波发病省份的活禽交易信息，认为没有危险性。两省间连线表示存在活禽交易，颜色由深至浅表示交易记录条数由多到少；各省名称下方数字为搜索得到的该省活禽交易信息总条数；☆表示该省活禽市场样本H7N9禽流感病毒检测阳性

图1 我国省际活禽贸易与人感染H7N9禽流感病例分布的关联网络图（截止2013年5月7日）

因此假设上述三省和上海市的活禽均可能携带H7N9禽流感病毒，三省一市的活禽贸易方向可能就是H7N9禽流感疫情发展的方向。基于第一波感染省市活禽交易信息，构建了基于省份的H7N9禽流感疫情蔓延的拓扑图。有8个省份（吉林、甘肃、青海、内蒙古、海南、云南、黑龙江和重庆）与第一波中1个省份有关联，判定为可通过活禽贸易有较低的发生H7N9禽流感疫情危险性（其中只有青海省基于内容搜索的方式找到了链接）；有18个省份被确定为高风险[4]；新疆、宁夏、西藏、贵州、陕西、香港、澳门和台湾尚未发现与这些疫情暴发省份活禽交易相关信息。截止2014年2月，在18个预测为高风险省份中，13个已有病例报告，显示了较好的发病风险预测作用。活禽交易信息数据提示北京市H7N9禽流感病毒可能是由滁州（安徽省）或盐城（江苏省）传播而来，该假设与流行病学数据并不相悖。该两市病例的发生日期（2013年3月9日、4月8日）均早于北京疫情（4月11日）。此外，滁州H7N9禽流感疫情可能由杭州（浙江省）和上海传入，南昌（江西省）和福州（福建省）的疫情可能源自上海。本研究所用发病数据及网络交易信息截止2013年5月7日，基于该时间点前所采集数据分析结果表明，广东、广西省区危险度较高，吉林省危险度较低。目前广东、广西、吉林省区在该时间点之后出现病例，其中广东已报告90例、广西2例，吉林仅1例，与此前分析危险度基本相符。2014年初台湾和香港地区均报告H7N9禽流感病例，但台湾报告病例是来自江苏为输入病例，与活禽贸易无关；而香港报告的病例，可能与2013年春季活禽贸易亦无关，但也有可能与本研究在网络上较难获得港澳台地区活禽交易信息有关。

控制活禽贸易对预防发生H7N9禽流感病例和疫区的扩大具有决定性作用。目前活禽交易信息还无法通过传统的流行病学和微生物学方法获得。暂时关闭活禽市场可以防止出现新病例，但不足以控制疫区扩大。利用“分布式聚焦爬虫技术”获得和分析活禽贸易网络大数据，可为控制H7N9禽流感疫区扩大提供参考信息。需要强调的是，上述分析数据是通过互联网追溯获得，而非实验室数据，如何有效地使用和评价，还需更多的现场实践予以证实。

参 考 文 献

[1] Gao R, Cao B, Hu Y, et al. Human infection with a novel avian-origin influenza A (H7N9) virus[J]. N Engl J Med, 2013, 368(20): 1888-1897.

[2] Xu J, Lu S, Wang H, et al. Reducing exposure to avian influenza H7N9[J]. Lancet, 2013, 381(9880): 1815-1816.

[3] Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends[J]. Sci Rep, 2013, 3: 1684.

[4] Chen C, Lu S, Du P, et al. Silent geographical spread of the H7N9 virus by online knowledge analysis of the live bird trade with a distributed focused crawler [J]. Emerg Microb Infect, 2013, 2(12): e89.

（收稿日期：2014-02-25）

（本文编辑：张林东）

读者·作者·编者

中华医学会系列杂志已标注数字对象惟一标识符

数字对象惟一标识符（digital object identifier, DOI）是对包括互联网信息在内的数字信息进行标识的一种工具。

为了实现中华医学会系列杂志内容资源的有效数字化传播，同时保护这些数字资源在网络链接中的知识产权和网络传播权，为标识对象的版权状态提供基础，实现对数字对象版权状态的持续追踪，自2009年第1期开始，中华医学会系列杂志纸版期刊和数字化期刊的论文将全部标注DOI。即中华医学会系列杂志除科普和消息类稿件外，其他文章均需标注DOI，DOI标注于每篇文章首页脚注的第1项。由中华医学会杂志社各期刊编辑部为决定刊载的论文标注DOI。

参照IDF编码方案（美国标准ANSI/NISO Z39.84-2000）规定，中华医学会系列杂志标注规则如下：“DOI：统一前缀/学会标识.信息资源类型.杂志ISSN.****-****.年.期.论文流水号”。即：“DOI：10.3760/cma.j.issn.****-****.yyyy.nn.zzz”。

中华医学会系列杂志标注DOI各字段释义：“10.3760”为中文DOI管理机构分配给中华医学会系列杂志的统一前缀；“cma”为中华医学会（Chinese Medical Association）缩写；“j”为journal缩写，代表信息资源类别为期刊；“issn.****-****”为国际标准连续出版物号（ISSN）；“yyyy”为4位出版年份；“nn”为2位期号；“zzz”为3位本期论文流水号。

中华医学会杂志社