·综述·

互联网搜索数据与流感预警

李锐 王增亮 张志杰 高杰 姜庆五

【关键词】 流感预警；实时监测；互联网；流行病学

Review on studying the early warning of influenza based on the data from Internet *LI Rui, WANG Zeng-liang, ZHANG Zhi-jie, GAO Jie, JIANG Qing-wu. Department of Epidemiology, School of Public Health, Fudan University; Key Laboratory on Public Health Safety, Ministry of Education; Laboratory for Spatial Analysis and Modeling, School of Public Health, Fudan University; Center for GIS Training, School of Public Health, Fudan University, Shanghai* 200032, China

Corresponding author: ZHANG Zhi-jie, Email: epistat@gmail.com

This work was supported by grants from the National Natural Science Foundation of China (No. 81102167, 81172609), *the National Science and Technology Mega Projects of China* (No. 2012ZX10004-220, 2008ZX10004-011), *A Foundation for the Author of National Excellent Doctoral Dissertation of China, FANEDD* (No. 201186) *and Specialized Research Fund for the Doctoral Program of Higher Education, SRFDP* (No. 20110071120040).

【Key words】 Early warning of influenza; Real time monitoring; Internet; Epidemiology

流感监测预警是其预防控制的重要内容，也是各国长期以来重要的研究领域。目前对人群流感的监测主要依赖传统监测手段，包括各级医疗机构、疾病预防控制中心和流感样病例监测哨点医院协作，其中由医疗机构诊断并报告流感临床诊断病例和确诊病例。流感样病例监测哨点医院是根据各地区情况和哨点医院的选择原则设定的监测哨点，而流感流行病学和实验室监测网络又是由各级疾病预防控制中心和具备资质的流感实验室组成。整个流感监测体系已较为完善，但仍有不足：首先由于获取数据的方式是“定时抽样，每周汇总”，必然存在数据结果相对滞后；其次该监测手段耗费大量人力物力，且各实验室检测和逐级上报的过程亦显繁琐；最后，该监测手段获取的数据来源单一，无其他来源数据的比对修正。而使用实时互联网进行数据分析获取人群流感传播的方法，在实时性上有较大进步，但一直因准确性不足，并未普及。鉴于根据互联网数据的流感预警研究可能对今后流感，乃至其他相关传染病的预防控制有重要意义，以下进行系统分析和综述。

1. 基于搜索信息的季节性流感预警：2008年Google公司开发了“谷歌流感趋势”(Google Flu Trends)的软件[1,2]，并监测美国的流感。该方法利用Google其巨大用户搜索数据(认为网络用户及其家人出现流感相关症状时可能采取搜索相关关键词的行为)，综合2003－2007年每日用户进行搜索的关键词和频率，构建了一个流感预测模型，并将该模型与美国疾病预防控制中心发布的流感监测结果比对和修正。结果显示，Google在大西洋中部地区的监测数据与美国疾病预防控制中心在同一地区监测数据的相关系数约为0.90，且比传统监测系统提前1～2周发现流感发病高峰[3]。“谷歌流感趋势”后经改进[4]，研究人员利用Google Insights for Search[5]所下载的数据和Google提供的方法，对比2004－2009年西班牙国家流行病中心网站的数据[6]，并进行Spearman分析，对关键词“gripe”的结果显示，相关系数最高达到0.70，并且整体数据比传统公布的结果提前2周时间。同时作者还对“aviar”、“tos”、“neumonía”等关键词做了相同分析，结果类似。值得注意的是，使用关键词“varicela”时相关系数最高可达0.96，但在实时性上表现并不理想，比传统公布的结果滞后约1周。

法国的研究人员利用Google提供的方法和数据，分析了2004－2009年法国流感样疾病、水痘、肠胃炎的流行情况，并对比法国监测网络(French Sentinel Network)提供的数据，分别获得0.82、0.90和0.78的相关系数[7]。该研究进一步验证了Google提出方法的有效性及其扩展性。研究中还对比讨论了使用关键词序列“grippe-aviaire-vaccin”和关键词“grippe”所得到的相关系数差异(分别为0.82和0.34)，提出联合使用多关键词所得到的结果在相关性上大大高于使用单一关键词。Carneiro和Mylonakis[8]也将该方法扩展至监测西尼罗病毒、呼吸道合胞体病毒感染和禽流感等，其结果与Google流感趋势的图表大体一致，即在相应时间出现了搜索峰值。但该作者并未对具体数据进行定量分析。此后，Ortiz等[9]引入更多变量，对2003年9月28日至2008年5月17日不同地区不同季节的流感发病情况利用Google的方法计算后验证，使用“谷歌流感趋势”、CDC Outpatient ILI

DOI: 10.3760/cma.j.issn.0254-6450.2013.01.023

基金项目：国家自然科学基金(81102167, 81172609)；国家科技重大专项(2012ZX10004-220, 2008ZX10004-011)；全国博士学位论文作者专项资金(201186)；高等学校博士学科点专项科研基金(20110071120040)；复旦大学曦源项目(122401)

作者单位：200032 上海，复旦大学公共卫生学院流行病学教研室 公共卫生安全教育部重点实验室 空间分析与建模实验室 GIS培训中心

通信作者：张志杰，Email：epistat@gmail.com

Surveillance Network 和 US Influenza Virologic Surveillance System的数据对比。结果显示，“谷歌流感趋势”与后两者数据的相关系数分别为0.72（95%*CI*：0.64～0.79）和0.94（95%*CI*：0.92～0.96）。并对2003—2008年历年的数据对比，其相关系数为0.67～0.98。作者还比较了不同地区的相关系数，认为地域因素对相关性的影响较大，包括该地区的人群数量、对互联网依赖程度及当地新闻媒体对人群搜索行为的影响。

以上研究仅停留在初步的应用阶段，主要表现在利用Google的方法和数据并提出模型进行验证和应用，对不同地区不同时间进行横向和纵向对比，但结果并不理想，其中人口数量、结构和当地经济发展水平以及互联网普及程度均影响其结果，并在疾病暴发峰值前的一段时间内，Google的预警可能有不同程度的“虚假峰值”，即在Google的数据上显示出有一个流感暴发的小高峰，而实际此高峰却并未发生，这可能是疾病早期由于人群恐慌心理和媒体对该疾病的报道，导致该时间段内在互联网上集中搜索相关信息。而多关键词联用可较为有效解决数据本身波动过大的问题。尽管该方法在各研究中均体现出良好的实时性和准确性，但由于其算法不完善，只能作为辅助的监测手段。

2. 基于搜索信息的流感大流行预警：2009年H1N1流感大流行时首次评估“谷歌流感趋势”在非季节性流感监测方面的准确性，结果表明其相关性低于预测季节性流感。2011年Cook等[10]更新该模型和方法，增加了构建模型的数据量，其重点为流感样疾病并发症，新模型更强调流感感染的情况。并分别在4个不同时期，即H1N1流行前期、夏季H1N1流行期、冬季H1N1流行期和H1N1全流行期（pre-H1N1, summer H1N1, winter H1N1, and H1N1 overall）评估新旧两种模型对美国门诊流感样疾病监测网络（ILI Net, the U.S. Outpatient Influenza-like Illness Surveillance Network）的吻合程度[11]，计算两模型和ILI Net的相关性和均方根误差，还对比了查询数量及类型等，发现两模型在H1N1流行前期和全流行期的相关性较好，而旧模型在流行前期低估了H1N1流感的活动，不利于提供早期预警。新模型在单一流行时间内的相关性也好于旧模型（夏季H1N1流行期，$r_1=0.95>r_2=0.29$）。

Wilson等[12]和Baker等[13]针对2009年3月29日至10月4日H1N1流感在新西兰的流行情况，将“谷歌流感趋势”和哨点监测以及健康统计的数据对比，结果显示Google能很好预测峰值并比传统监测数据提前约1周时间，且对峰值的估计较为准确。该作者虽未给出具体相关性系数，但图表数据与传统数据契合度好。从结果可见除峰值预测上Google较传统数据提前1周，但在早期暴发的预测上表现并不理想，甚至不如利用当地卫生医疗服务热线进行预测及时。作者认为这是由于当地的生活习惯，导致出现流感相关症状时居民普遍使用电话而非互联网方式进行查询。

Valdivia等[14]评估了“谷歌流感趋势”和一些欧洲国家的哨点医生网络的相关情况，使用欧洲疾病预防控制中心流感监测网络[15]、世界卫生组织[16]、法国[17]、西班牙[18]、德国[19]以及瑞典[20]的数据，利用软件Stata 9.1进行Spearman分析，其结果覆盖大部分欧洲国家。在整个H1N1流感大流行期间，德国的相关性系数最高（$r=0.94$），波兰最低（$r=0.72$），其他国家依次分布在该区间内。值得注意的是，在流行前期即2009年3月23日至8月30日的相关性系数普遍低于后期（2009年8月31日至2010年3月28日）。在大多数国家中，采用“谷歌流感趋势”的峰值预测多与SPN（sentinel physician networks）的峰值在同一周内出现或提早1～2周，仅有保加利亚的峰值比传统监测方法获得的峰值滞后1周。

Hulth和Rydevik[21]利用“谷歌流感趋势”的原理，建立“GET WELL”的服务，从“www.vardguiden.se”网站上获取用户的搜索记录，并自动导入到相关数据库，通过数据分类和分析，得出流感预测的图表，再以邮件的形式发送给用户。为了评估该方法的准确性，进行了一次用户访谈，抽样调查用户的反馈情况。结果显示用户（大部分为流行病学者）对该服务的准确度普遍满意，并认为该服务能提高科研效率。同时还进行了定量评估，2009年5月至2010年5月H1N1流感大流行期间该系统与传统监测结果的相关系数为0.90，在峰值的准确性上好于Google，即后者在一定程度上低估了流行中期的疾病高峰。Scarpino等[22]使用了基于地理信息的“报告者模型”，使得数据结果能在样本量较小的情况下较为准确，在一定程度上弥补了Google基于大量数据模型的弊端。

综上所述，Google所提供的模型和方法是一个有价值的流感监测工具。开展的研究均基于Google提供的数据，但并未提供原始数据信息，因而其他研究人员无法自行调整算法以针对不同国家地区的具体情况，也无法针对不同类型的流感修正模型的准确度。

3. 展望：上述研究表明利用互联网信息对流感的监测可作为传统监测手段有效补充。并且随着方法的不断改进，在准确性上有了较大提升，能够综合分析整合除了用户搜索数据之外的其他信息（如对在线医疗网站的访问记录、论坛上对流感讨论情况及相关微博的信息等）。随着人们对互联网依赖程度不断提高，互联网可用于流感监测的数据将更多，也将更利于其监测。

传统的监测方法只是单方面对人群进行监测，同时也因数据处理和发布不及时，未有效利用人群对流感的反馈，进而制定相应的应急策略。而利用互联网信息的流感监测，不仅能监测流感暴发，还能够及时甚至即时了解人群对流感爆发的关注程度及其防疫需求，从而制定更为有效的防疫措施和控制人群的恐慌心理。

但该研究领域还存在一些问题，如数据获取和分析过程仍需要耗费大量人力，且目前还没有自动化的方法；数据集中在少数搜索引擎提供商手中，普通研究人员对该方法的扩展有限；我国的相关研究处于空白，未见类似研究进展；数据准确性有待提高。为此笔者已开展相关研究，如利用编写代码的方式将数据获取和分析自动化并提供GUI界面的程序开放使用，试图利用Google Insights Search获得较为原始的

数据构建模型分析。结合我国国情，使用百度或其他中文搜索引擎进行研究。随着数据量的增加以及软件算法的改进，其数据的准确性能够得到一定程度的提高；如果有条件结合数据挖掘技术对互联网信息结合语义分析，能够得到更多信息对数据的准确性进行修正。

流感监测中利用互联网技术提供有力的辅助手段也能在一定程度上弥补传统监测预警体系的不足，节约资金提高效率，随着我国互联网生态环境的逐渐建立，也为该项研究提供了较好的条件，并将发挥越来越重要的作用。

参 考 文 献

[1] http://www.google.com. [2012-05-01]

[2] http://www.google.org/flutrends/. [2012-05-01]

[3] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. Nature, 2009, 457 (19): 1012-1015.

[4] Valdivia A, Monge-Corella S. Diseases tracked by using Google Trends, Spain. Emerg Infect Dis, 2010, 16: 168.

[5] Google Insights for Search, 2009. [2009-08-02]. http://www.google.com/insights/search/#.

[6] Boletín Epidemiológico Semanal, 2004-2009. [2009-08-02]. http://www.isciii. es / jsps / centros / epidemiologia / boletines Semanal.jsp.

[7] Pelat C, Turbelin C, Bar-Hen A, et al. More diseases tracked by using Google Trends. Emerg Infect Dis, 2009, 15: 1327-1328.

[8] Carneiro HA, Mylonakis E. Google Trends: a Web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis, 2009, 49: 1557-1564.

[9] Ortiz JR, Zhou H, Shay DK, et al. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. PLoS One, 2011, 6: e18687.

[10] Cook S, Conrad C, Fowlkes AL, et al. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One, 2011, 6: e23610.

[11] CDC-Seasonal Influenza (Flu) -Flu Activity & Surveillance. [2011-07-28]. http://www. cdc. gov/flu/weekly/fluactivitysurv. htm

[12] Wilson N, Mason K, Tobias M, et al. Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: the New Zealand experience. Euro Surveill, 2009, 14(44): pii19386.

[13] Baker MG, Wilson N, Huang QS, et al. Pandemic influenza A (H1N1) v in New Zealand: the experience from April to August 2009. Euro Surveill, 2009, 14(34): pii19319.

[14] Valdivia A, López-Alcalde J, Vicente M, et al. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks-results for 2009-10. Euro Surveill, 2010, 15(29): pii19621.

[15] European Centre for Disease Control and Prevention (ECDC). European influenza surveillance network (EISN). Stockholm: ECDC. [2010-04-01]. http://www.ecdc.europa.eu/en/activities/surveillance/EISN.

[16] EUROFLU. WHO/Europe influenza surveillance. Copenhagen: WHO Regional Office for Europe. [2010-04-01]

[17] Réseau Sentinelles France. Situation Epidémiologique en France métropolitaine. Paris. http://websenti.b3e.jussieu.fr/sentiweb.

[18] Red Nacional de Vigilancia Epidemiologica. [2010-04-01]. http://vgripe.isciii.es/gripe.

[19] Robert Koch Institute (RKI). Berlin: RKI. [2010-04-01]. http://www.rki.de.

[20] Smittskyddsinstitutet (SMI). Stockholm: SMI. [2010-04-01]. http://www.smittskyddsinstitutet.se.

[21] Hulth A, Rydevik G. Web query-based surveillance in Sweden during the influenza A (H1N1) 2009 pandemic, April 2009 to February 2010. Euro Surveill, 2011, 16(18): pii19856.

[22] Scarpino SV, Dimitrov NB, Meyers LA. Optimizing provider recruitment for influenza surveillance networks. PLoS Comput Biol, 2012, 8(4): e1002472.

（收稿日期：2012-09-22）

（本文编辑：张林东）