基于 Twitter 的流感疫情可视化系统①

张振华[1,2], 吴开超[2]

[1](中国科学院大学, 北京 100190)
[2](中国科学院计算机网络信息中心, 北京 100190)

摘 要: 介绍了当下社交媒体的蓬勃发展以及基于社交媒体数据的多领域应用, 阐述了利用 Twitter 数据对流感疫情监测的优势和可行性, 着重论述了基于 Twitter 的流感疫情可视化系统的详细设计和具体实现, 并讨论了系统的意义和应用方向.

关键词: 社交媒体; Twitter; 流感疫情监测; 支持向量机; 分布式系统

Flu Epidemic Visualization System Based on Twitter

ZHANG Zhen-Hua[1,2], WU Kai-Chao[2]

[1](University of Chinese Academy of Sciences, Beijing 100190, China)
[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This work introduces the booming development of social media and associated applications on various domains, demonstrates advantages and feasibility of applications on flu epidemic detection on the basis of Twitter data. Specifically, this paper puts emphasis on describing the detailed design and concrete implementation of corresponding visualization system. Finally, it discusses the significance and potential applications of the system.

Key words: social media; Twitter; flu epidemic detection; SVM; distributed system

近年来, 社交网站、微博、博客、论坛等社交媒体产品层出不穷, 无线互联网的发展和智能手机的风靡更是为社交媒体的繁荣提供了契机. 融合多媒体、游戏、基于位置服务等多元素特征的社交媒体被众多用户当成日常工具, 成为他们获取新闻时事、人际交往、自我表达、社会分享及社会参与的重要媒介[1]. 据中国互联网信息中心(CNNIC)发布的《第 33 次中国互联网络发展状态统计报告》[2]称, 2013 年我国约有 2 亿 8 千万网民使用微博, 约占网民总数的 45%. 到 2013 年下半年, Facebook 作为全球最大的社交媒体, 有超过 10 亿的注册用户和 3.5 亿的月活跃用户[13]; 2009 年上线的新浪微博在短短的 4 年内注册用户也突破了 5 亿, 每天产生超过 1 亿的新微博[14].

Twitter 是 2006 年上线的微博产品, 发展至 2013 年下半年, 已拥有约 10 亿的注册用户和 2.5 亿的月活跃用户, 每天用户创建的推文(Tweets)多达 5 亿条[15]. 用户的高参与度和平台的高活跃度使得 Twitter 能够在一定程度上反应现实世界的特征[3], 尤其是平台表现出的对特殊事件的敏感性[4](如 2014 年巴西世界杯前 15 天, 世界杯相关的推文就达到 3 亿条[16]), 使得 Twitter 成为科研人员借助社交媒体研究现实事件特征的首选对象之一.

1 相关系统研究

从 2009 年起, 借助于 Twitter 平台数据研究现实事件特征的案例不胜枚举, 涉及的领域包括公共卫生、公共安全、旅游管理、政治事件、自然灾害防治等, 其中部分介绍了基于 Twitter 数据应用系统建立的方法模型和具体实现. 在政治事件研究方面, Wang 等[4]利用人工标注为 4 类的 17000 条推文建立基于朴素贝叶斯

① 基金项目:美国国家自然科学基金(NSF)(0846655,1047916);中国科学院计算机网络信息中心所级专项(CNIC_PY_1406);DNSLAB 开放基金 (DNSLAB-2013-D-U)

收稿时间:2014-07-01;收到修改稿时间:2014-08-11

方法的分类器, 利用此分类器判断从 Streaming API 获得的每条新推文对于 2012 年美国总统大选中每个候选人的态度, 并在一定时间段内对候选人支持数作聚合计算, 从而实时地捕捉并可视化出民众对不同候选人支持态度变化的动态; 在自然灾害预防方面, Sakaki 等[5]利用基于推文关键字、推文字数以及推文上下文等特征的分类器侦测事件的发生, 提出一种概率时空模型确定事件发生的地理中心, 并基于此方法构建地理快速侦测系统, 通过实验此系统可以准确报道 93% 发生在日本震级三级以上的地震(数据来自官方地震预测机构 JMA), 并且侦测需要的时间明显小于官方机构; 在公共安全方面, MacEachren 等[6]建立了地理位置可视化分析系统 SensePlace2, 利用不同主题关键字从 Twitter API 获取感兴趣的推文, 在提取推文的地理位置信息的同时对其作全文索引用以支持可视化系统的高效查询, 并阐述了 SensePlace2 在犯罪管理上的应用. 在 Twitter 数据应用的众多领域中, 针对公共卫生方面的研究最多. Achrekar 等[7]通过从 Twitter 和 Facebook 的数据建立在线的流感侦测和疫情发展预测系统 SNEFT, 利用文本挖掘提高方法计算流感疫情的准确度(与官方疾控中心公布数据比较), 并借助 ARX 回归模型建立流感疫情发展趋势的预测模型. Dredze 等[8]介绍了基于 Twitter 的 geolocation 系统 Carmen, 利用推文内容和用户注册的信息将每条推文与地理位置相关联, 并阐述其在公共卫生监控方面的应用.

2 系统介绍

2.1 背景简介

作为全球最受欢迎的微博平台, Twitter 吸引了数亿来自全球各地的用户在平台上分享经历、见解等. 尤其在美国等互联网发达区域, Twitter 用户的分布与人口的分布正相关, 因此 Twitter 用户群可以被认为是现实人口的一个样本, 其行为在一定程度上蕴藏着现实世界的时空特征.

流感是一种极度危害公共卫生安全的传染性疾病, 流感的爆发每年在全球约造成 300 万到 500 万人罹患重疾, 并致 25 万到 30 万人死亡[8]. 而流感主要通过人与人的接触而传播[9], 因此, 对流感爆发区域的快速侦测可以更好地协助决策者尽快作出预警, 隔离疫情较重区域, 并调控必要的医疗资源.

2.2 系统的优势和可行性

当下流感疫情主要通过医院和医学实验室对病例汇总上报来统计, 这样的数据统计往往有较大的时延, 不利于疫情的动态快速侦测; 而社交媒体(如 Twitter)数据是实时的, 通过对实时数据的分析计算得出的流感疫情可以较好地规避传统上报方式监测的数据滞后性, 并且可以作为对当前流感疫情判断的参考.

Twitter 是一个开放的平台, 用户可以随意分享见闻经历(包括身体状况), 受益于 Twitter 用户的高参与度, Twitter 数据有可能隐含流感疫情爆发的线索. ARAMAKI 等[10]通过对 Twitter 数据的分类证明了由 Twitter 得出的流感疫情与官方疾控机构公布的疫情有很高的相关度.

2.3 系统功能

基于 Twitter 的流感疫情可视化系统主要包括(a)流感疫情的区域可视化、(b)流感患者的移动模式可视化和(c)可视化配置三个模块(见图 1).

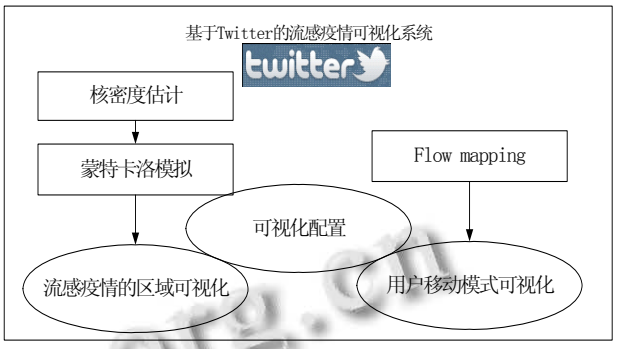

图 1　流感疫情可视化系统功能模块

(a) 流感疫情的区域可视化模块 将由 Twitter 实时数据计算出的元组{地理位置, 流感疫情值}通过核密度估计平滑后经由蒙特卡洛模拟得出被划分出来的各个子区域中疫情估计, 并将各个区域的疫情估计值以热度图的方式结合 Bing 地图展示出来, 分析者可以很直观地分析疫情的分布.

(b) 用户移动模式可视化模块 利用 Twitter 数据流中计算得出的各个区域之间人员(或流感患者)流动情况使用制图学经典的 Flow mapping 方法[11]可视化在 Bing 地图上. 通过用户移动的 Flow mapping 可视化可以很直观地获知哪些区域之间的人员流动较大; 当发现疫情较重的区域 A 至疫情较轻区域 B 有较大流量(特别是这样的流量大部分由患流感用户造成)时, 这

样的人员移动就很可能成为区域 B 疫情变重的原因, 从而可以帮助分析者更好地理解流感传播的原因, 并协助相关策略的指定.

(c) 可视化配置模块 主要控制(a)和(b)可视化的具体选项, 如是否显示流感患者的移动、多源和单源 Flow mapping 可视化的切换等

3 系统框架设计

本系统是基于 B/S 模式的, 流感疫情的可视化结果和移动模式可视化结果分别作为附着在 Bing 底图的两个图层. 在进行流感疫情可视化和移动模式可视化计算前, 需要从 Twitter 数据流通过基于分布式架构(Storm)的实时处理模块计算出各地点的流感疫情值以及区域之间人员的移动流量, 并持久化至数据库(以推文创建时间为分片键的分布式 MongoDB)以便应用查询使用. 系统用户做可视化分析时, 可以根据时间段从分布式数据库中高效查询出(a)各地点流感疫情值, 进行核密度估计及蒙特卡洛模拟计算, 并将模拟的结果(各区域的疫情值)以热度图的方式展示到 Bing 地图上; (b)各区域之间人员的流动, 并实施 Flow mapping 算法计算出代表移动统计规律的对象, 并通过 HTTP 协议返回给页面用 OpenLayers 可视化. 因此, 系统的模块框架设计如下图 2, 数据流图设计如图 3.

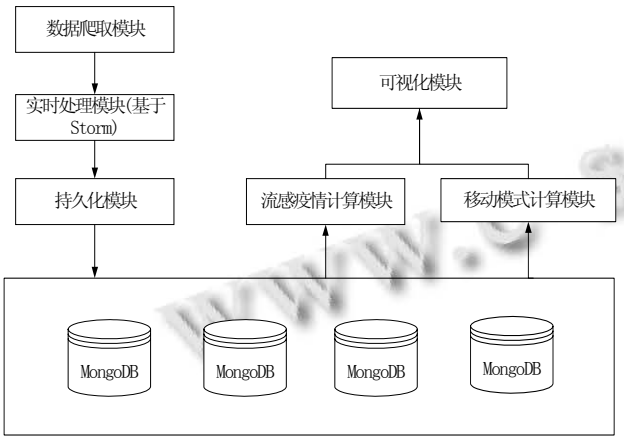

图 2 系统的模块设计

4 系统模块实现

本系统以美国大陆(不包括夏威夷和阿拉斯加)作为研究区域, 将其在不同空间精度下划分成多个子区域(栅格), 从而支持在不同空间精度下对各区域流感疫情的可视化.

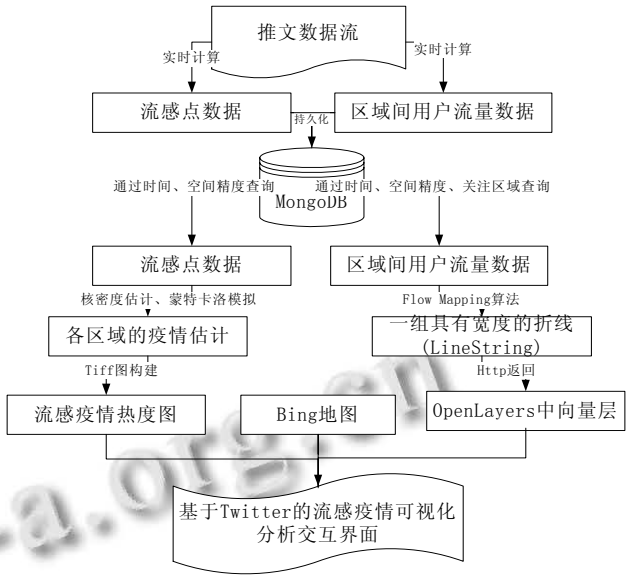

图 3 系统的数据流设计

4.1 数据爬取模块

Twitter 平台具有相对开放的数据政策, 尤其是 Streaming API, 用户可以通过维护长期的会话从平台中获取用户实时产生的推文数据. Twitter4j 是封装 Twitter 数据 API 的第三方库, 它将调用数据 API 所必须面临的 HTTP 页面访问、JSON 返回的解析作了成熟的封装. 系统的数据爬虫将美国大陆的地理边界作为参数传递给 Streaming API, 从而获取在美国大陆发出的具有地理位置信息的推文.

4.2 实时处理模块

实时快速生成的具有地理位置特征的推文流包含了推文的 Id、内容、地理坐标 (可能为空)、地理边界、创建时间、用户 Id 等信息, 需要经过如图 4 的预处理后才能计算图 3 中的需要持久化至数据库的两项内容.

图 4 推文的预处理步骤

4.2.1 基于 SVM 的流感推文分类器

每一条推文内容都可以提取出若干关键词(去掉 the, a, an 等停用词)作为特征值形成向量. 本系统利用人工标注好是否代表流感案例的 5052 条推文作为训练样本(其中 4473 条被标注为不代表流感案例, 579 条被标注为流感案例), 针对整个训练集推文的关键词形成向量空间(计 4450 维), 并按照 LibSVM 要求的格式

对每条推文计算{特征在向量空间的 index:特征所对应的权重}, 最终借助于 5 折交叉验证在网格空间(2^{-8}~2^{8})中最终确定了最优的模型参数为 c=4.0 和 gamma=0.025, 此时的判断准确率为 93.7%, 并将此训练出的模型保存至文件结合向量空间文件供后续分类使用.

4.2.2 流感点数据计算

核密度估计计算的输入是某一时间段{location、流感数目}这样的流感点数据. 每条推文都可以解析出创建时间、推文内容、地理坐标信息、地理边界信息, 当地理坐标为 null 时, 推文的位置可以用地理边界的中心点估计, 推文的内容可以利用训练好的 SVM 模型结合向量空间文件判断出是否代表流感案例. 这样, 流感点数据的计算流程如图 5 所示.

4.2.3 区域间用户流量计算

用户在区域 A 的某位置发了推文后又移动到区域 B 的某位置发了推文, 这样就构成了本系统所识别的一次区域移动. 在推文信息流中, 通过解析每个推文的位置从而判断 Twitter 用户发文所在的区域, 如果后续又出现该Twitter用户的推文, 计算推文对应的区域, 并且更新该用户的最近发文区域. 每 7 天对于各区域之间的流量作一次统计, 并将其持久化至数据库中. 区域间用户流量计算流程如图 6 所示.

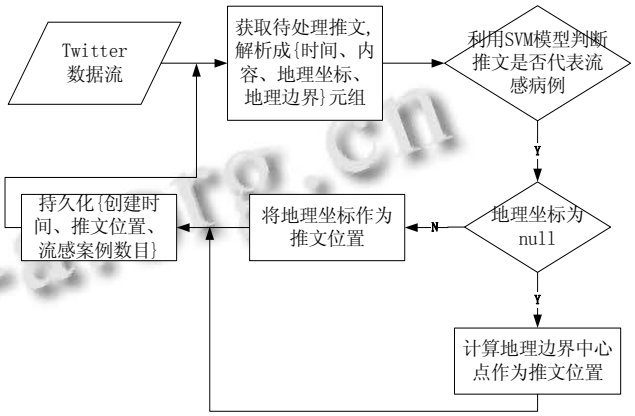

图 5　流感点数据的计算流程

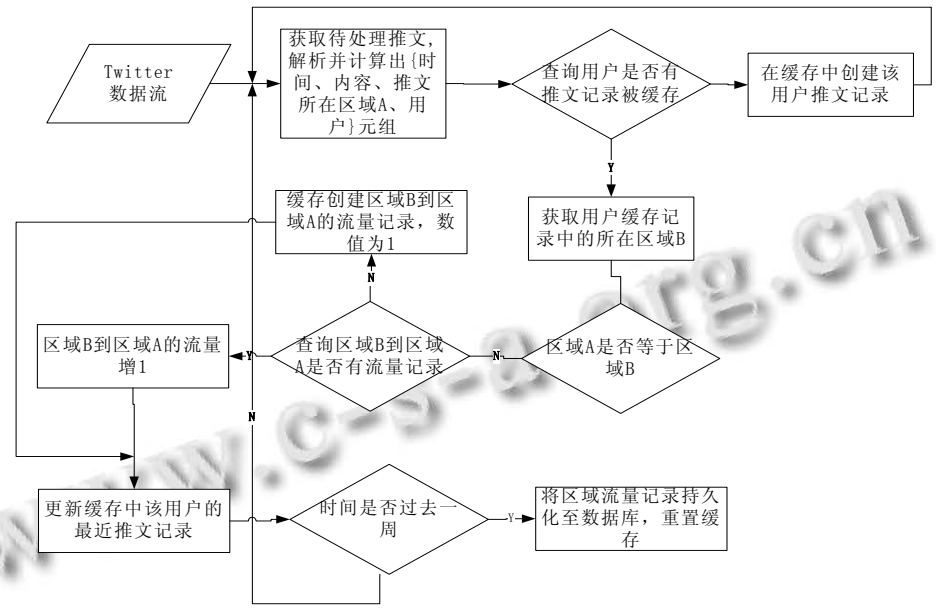

图 6　区域间用户流量计算流程图

4.2.4 分布式框架的应用

由于爬虫每秒钟平均收集到 25 条推文, 针对每条推文实施既定逻辑的计算速度要小于推文产生的速度, 这样就会导致数据的积压. 本系统利用 Storm 将推文的各个计算子任务 (如判断是否跟流感相关、计算地理位置、过滤等) 分配到不同节点的不同线程; 并且为了支持对历史数据的查询和系统的可扩展性考虑, 系统采用了基于推文创建时间槽为分片键的分布式 MongoDB 作为存储流感疫情点数据和各区域间用户移动流量数据.

4.3 持久化模块

利用 MongoDB 基于 Java 的 JDBC 库, 将疫情可

视化和移动模式可视化计算依赖的两类元组{时间槽、地点、流感数},{时间槽、空间精度、区域 Id、该区域到其它区域的流量、该区域代表点坐标}持久化至以时间槽分片的 MongoDB 集群.

4.4 流感疫情计算模块

通过浏览器交互界面, 系统用户可以选择感兴趣的时间段作为搜索条件高效地从 MongoDB 中得到流感点数据, 将流感点数据结合各区域的人口计算罹患流感的比率, 并采用自适应核函数的核密度估计平滑流感点数据, 采用随机标记的蒙特卡洛模拟计算出各个划分区域的流感疫情评价, 并最终形成 tiff 图缓存.

4.5 用户移动模式计算模块

通过空间精度、时间槽和区域 Id 可以在数据库中查询到在某段时间内从该区域到其它区域的用户移动流量统计, 利用基于导航数据的 Flow mapping 算法[12]可以在表示各区域人员移动特征的同时融入地理元素, 对于流感传播的分析更有意义. 用户移动模式计算模块的输入可以看成是一个加权图, 输出可以看成是一组不同权值的折线.

4.6 可视化模块

基于 OpenLayers 的可视化模块, 将流感疫情计算模块和用户移动计算模块的结果, 分别以 WCS 图层展示各区域流感疫情的热度图和 Vector 图层构建一组带方向加权折线来表示不同区域间用户移动的情况. 特别地, 展示患流感用户在不同区域间的移动情况有可能帮助分析流感传播, 理解流感传播的原因.

图 7 是 2013 年 9 月 3 日 10 点系统的截图, 表示从 2013 年 8 月 28 日 10 点到 2013 年 9 月 3 日 10 点的一周时间内, 美国大陆的流感疫情热度图以及从芝加哥到其它区域的用户移动情况. 从界面中可以很清晰地看出(a)基于 Twitter 数据分析出的高流感风险的区域, 当地图缩放时, 系统会展示不同空间精度下各区域的疫情情况, 使得对疫情的监控更加灵活和全面; (b)由芝加哥地区到其它地区的用户移动的总体情况, 当向某个区域的罹患流感用户移动流量较大时, 就有可能加重这个区域的疫情, 使得后续时间的疫情展示颜色更加趋近于红色; 反过来, 当发现某区域疫情变重时, 可以尝试通过该区域的用户移动特征寻找疫情加重的原因.

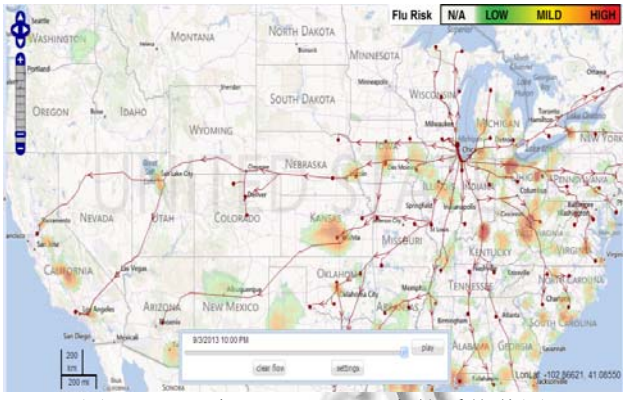

图 7 2013 年 9 月 3 日 10 点的系统截图

5 结语

本文具体阐述了基于 Twitter 的流感疫情可视化系统的建立动机、详细设计和具体实现. 基于 Twitter 的流感疫情可视化系统旨在借助最受欢迎的社交媒体之一 Twitter 开放的 Streaming API 获取的数据, 建立对美国大陆流感疫情的近实时监控, 以热度图的方式在多时空尺度下直观地展示各区域的流感疫情状态, 并采用经典制图学运动可视化算法 Flow mapping 可视化出各区域之间(罹患流感的)用户移动的情况, 为更好地理解用户的移动模式和流感传播的原因提供参考.

结合中文自然语言处理和新浪微博数据实现国内的流感疫情监测是本文的一个拓展方向.

参考文献

1 丁兆云,贾焰,周斌.微博数据挖掘研究综述.计算机研究与发展,2014,51(4).

2 中国互联网信息中心.第 33 次中国互联网络发展状态统计报告.2014.http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/t20140305_46240.htm.

3 Asur S, Huberman BA. Predicting the future with social media. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE. 2010, 1. 492–499.

4 Wang H, Can D, Kazemzadeh A, et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proc. of the ACL 2012 System Demonstrations. Association for Computational Linguistics. 2012. 115–120.

5 Sakaki T, Okazaki M, Matsuo Y. Tweet analysis for real-time

event detection and earthquake reporting system development. IEEE Trans. on Knowledge and Data Engineering, 2013, 25(4): 919–931.

6 Maceachren AM, Robinson AC, Jaiswal A, et al. Geo-twitter analytics: Applications in crisis management. 25th International Cartographic Conference. 2011. 3–8.

7 Achrekar H, Gandhe A, Lazarus R, et al. Online social networks flu trend tracker: A novel sensory approach to predict flu trends. Biomedical Engineering Systems and Technologies. Springer. 2013. 353–68.

8 Dredze M, Paul MJ, Bergsma S, et al. Carmen: A twitter geolocation system with applications to public health. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI). 2013

9 Clayton D, Hills M. Statistical models in epidemiology. Oxford University Press, 1993.

10 Aramaki E, Maskawa S, Morita M, eds. Twitter catches the flu: Detecting influenza epidemics using twitter. Proc. of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011. 1568–1576.

11 Wang S, Cao G, Zhang Z, Zhao Y, Padamahban A. A CyberGIS environment for analysis of location-based social media data. Location-based Computing and Services, 2013.

12 Padmanabhan A, Wang S, Cao G, et al. FluMapper: A CyberGIS application for interactive analysis of massive location-based social media. Concurrency and Computation: Practice and Experience, 2014.

13 http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/.

14 http://www.donews.com/net/201402/2711464.shtm.

15 http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/.

16 http://tech.sina.com.cn/i/2014-06-27/11499463261.shtml.