基于微博数据的应用研究综述*

刘晓娟 尤 斌 张爱芸

(北京师范大学 管理学院 北京 100875)

摘 要 微博数据已经成为学术界重要的数据来源,近年来国内外学者开始利用大量开放的微博数据进行社会科学、企业营销、医疗卫生、政府建设等多方面的研究。从数据来源及获取方式、数据选择及其处理办法和研究结果的应用三个方面对国内外研究情况进行了总结,并阐述了目前研究存在的局限性。

关键词 微博 大数据 Twitter 新浪微博 可视化 数据集

中图分类号 G203 **文献标识码** A **文章编号** 1002-1965(2013)09-0039-07

Review on the Data Used in Researches of Microblogs

Liu Xiaojuan You Bin Zhang Aiyun

(School of Management, Beijing Normal University, Beijing 100875)

Abstract The data of microblogs has been an important source for academics. Recently scholars use a vast amount of data of microblogs for the study of social science, business marketing, health-care and governmental development. This paper aims at introducing the research status in three aspects: data source and collection, data selection and processing, the application of research findings. At last it indicates some limitations of the existing study.

Key words microblog big data twitter sina weibo visualization data sets

0 引 言

大数据时代的到来,引起了各界学者广泛的关注。哈佛大学社会学教授加里·金说:"这是一场革命,庞大的数据资源使得各个领域开始了量化进程,无论学术界、商界还是政府,所有领域都将开始这种进程。"IBM公司把大数据概括成了三个V,即大量化(Volume)、多样化(Variety)和快速化(Velocity)。这些特点也反映了大数据所潜藏的价值(Value),或许可以认为,这四个V就是大数据的基本特征[1]。

微博是近年来新兴的一种网络服务,如国外的Twitter、国内的新浪微博、腾讯微博等,截至2012年12月底,Twitter以及新浪微博的注册用户都突破了5亿,用户生成的数据量之巨大也可想而知。微博使用简单便捷、信息丰富、传播速度快、更新迅速,符合大数据的基本特征,被认为是大数据时代的典型代表。同时,微博用户产生的大量数据也成为大数据环境下社会研究的重要数据来源,2010年4月国会图书馆与Twitter达成合作协议,允许访问Twitter自2006年上线以来的Twitter消息,截至2013年1月,国会图书馆已收集了1700亿条,这些数据将归入国会图书馆的"Web收藏"(Web capture)项目[2],微博数据也正式成为历史的见证。

目前国内外关于微博的研究已经大量开展,本文将对这些研究从数据来源、获取方式、数据选择及其处理办法以及研究结果的应用几方面进行综述,以期对下一步的研究提供指导和参考。

1 数据来源及获取方式

1.1 数据来源 通过对以往研究的回顾可以发现,基于微博的研究,其研究数据多来源于Twitter和新浪微博,网易微博稍次之。通过归纳比较,可以发现研究者选取Twitter和新浪微博进行研究的原因主要有两个:

收稿日期:2013-05-21 修回日期:2013-07-01

基金项目:国家社会科学基金项目"基于网络计量方法的热点WEB空间研究(09CTQ028)"的成果之一。

作者简介:刘晓娟(1980-),女,博士,副教授,研究方向:信息计量,信息可视化;尤 斌(1988-),女,硕士研究生,研究方向:网络计量,信息可视化;张爱芸(1990-),女,硕士研究生,研究方向:信息检索,网络计量。

a. 用户数量多,活跃度高。Twitter 是由 Jack Dorsey 于 2006 年 3 月创建,经过多年的发展,在 2012 年 7 月 31 日巴黎分析公司 Semiocast 发布的报告中称,Twitter 注册用户量已经超过 5 亿,约相当于 Facebook 用户总数的一半,是用户人数第二多的社交网站,这其中仅有 1.418 亿用户来自于美国,可见大多数的用户来自于美国以外的市场。新浪微博作为国内最大的微博服务平台,截至 2012 年 12 月底,注册用户数突破 5 亿,日活跃用户数达到 4 620 万。美国投资公司 T. H. Capital 分析师也指出新浪微博是"无可取代的",在中国微博市场上拥有"独占"地位[3]。因此,大部分研究学者都会选择 Twitter 和新浪微博的用户数据作为研究的数据来源。

b. 提供开放 API,数据获取便利。Twitter 很早就为研究者们提供了开放的 API 服务,方便研究人员进行数据获取,因此国内诸多社交媒体平台如新浪微博、腾讯微博等也顺应趋势纷纷开放了 API,为第三方开发者和研究者提供了上百种 API 接口和各主流语言的 SDK,研究者和开发者可以通过 API 获取用户的个人信息、好友信息、照片、群组等大量数据来进行应用的开发和调查研究工作,为研究提供了极大的便利和帮助。

1.2 **数据获取方式及其比较分析** 微博中数据丰富并且数据量巨大,因此在对微博数据研究时,应选择合理的数据获取方式,从而为研究提供便利,本文将数据获取方式分为以下三种——基于官方 API 进行获取、通过网络爬虫爬取微博页面和直接利用开放的数据集。

1.2.1 数据获取方式。a. 基于官方 API 开发的系统。微博服务提供商为了使微博服务和应用更加多样化,更具有吸引力,选择了向应用开发者和研究者提供开放应用程序接口,即 Open API(Open Application Programming Interface),方便研究人员和开发者获取数据。Open API 指利用 SOAP、JavaScript 等实现网站互联的一系列技术,开发者通过 Open API,便能够以程序的方式访问网站的数据和平台。Twitter 目前支持以下的四种数据返回格式:XML、JSON、RSS、Atom,用户可以在每次请求时使用不同的请求方法返回对应特定格式的数据。新浪微博的 API 返回的数据格式为 JSON 和 XML。JSON 的格式简单,易被用户读懂,如{"created_at":"Wed Jan 02 12:16:18 +0800 2013","id":3529996356152943}。JSON 较 XML 有更高的稳定性[4],且现在网络上也有很多对 JSON 进行解析的工具供人们使用,如 BeJSON[5]。

开放 API 的提供也成为了微博走向市场和学术领域的重要途径。Rui H 利用 Twitter 提供的 API 开发了一个收集用户信息的程序,获取了一系列用户的个人信息,用以研究用户的行为特征[6], Lee R 等利用 Twitter 提供的 API 开发了一个地理微博监视系统来监视和利用大规模地理微博的密度,以此来确定某个特定区域的群体行为[7],Cheong M 利用 Hashtags. org(基于 API 开发的可自动追踪用户标签的程序)来观察 Twitter 的在线行为在真实社会中是如何反映的[8]。国内也有很多学者利用新浪微博提供的 API 获取数据进行研究,杨成明以用户"生活月刊"作为用户数据采集的起点,分别采集其"粉丝"和"关注",再采集"粉丝"的"粉丝"及"关注"的"关注",层层递进,采集时间从 2010 年 3 月 1 日 0 时 0 分到 3 月 8 日 9 时 40 分,共采集到 649 006 个用户后,人工中止采集程序,再从用户性别、地域、影响力等多个角度揭示当前微博客用户的行为特征[9]。通过调研发现,这类微博数据获取方式一般都是在官方 Open API 的基础上,根据研究目标进行二次开发,使获取数据更加便捷和准确。

b. 通过网络爬虫爬取微博页面。通过网络爬虫爬取微博数据一般是指通过 HTTP 协议,模拟浏览器向服务器发送请求,对返回的网页进行解析,从中抽取出相应的微博数据[10]。该方法几乎适用于任何微博数据的获取,并且不像利用官方 API 获取数据一样会受到微博运营商权限开放范围的限制。不足在于其稳定性差,微博运营商可能会调整 HTTP 请求的参数设置和返回的 HTML 页面的格式,这样极有可能导致微博数据的无法获取和解析。这种方式需要定期监测网络爬虫的运行情况,根据需要对程序进行及时更新。

c. 开放的数据集。随着 Web2.0 的发展,信息公开和资源共享显得越来越为重要,越来越多的学者将自己获取的语料库和数据集以不同的开放程度进行公开,供其他人开发和使用。利用已有的数据集,可以免去预处理的过程,研究者可以根据研究需求对数据集进行修改后加以利用,既提高了研究效率,又节省了研究经费。Culotta A 在利用微博数据推导精确流感比率模型的研究中[11],采用了 O'Connor 的舆论调查研究中所用的数据集的子集,该数据集利用 Twitter API 抓取了 2008 和 2009 年的十亿条微博[12]。

开放数据集是获取研究数据的比较便捷的途径,近年来计算机的超强存储和检索能力使得超大规模数据集公开平台开始出现,如 GetTheData[13] 和 Datamob[14]。GetTheData 是一个 Q&A 模式的网站,研究者可以使用网站提供的工具或者第三方工具来进行数据清洗或者数据集的可视化,其中问题和答案可以修改和完善,还可以用相关的关键字标记任何问题,方便将来再次访问网站,并进行材料积累。Datamob 旨在使用简单的方式来利用公共数据源,目前在该网站上

列出了 227 个数据源、165 个应用程序和 66 个资源,分为 67 个标签,多由政府和公共机构发布。斯坦福大学的网络分析平台(SNAP)提供了来自 Slashdot、Epinions、LiveJournal、维基百科、亚马逊、Twitter 和 Meme-Tracker 的数据集[15]。

数据堂[16]和中国爬萌[17]是国内专业的科研数据共享服务平台。研究者可以在数据堂上传自己收集的数据,也可以发布自己的数据需求。数据堂目前已有 166 万科研人员共享 42 492 组科研数据集,涵盖学科超过 20 个,它对分散在各个领域的数据进行收集、加工、整理,通过统一的平台提供数据检索、数据下载服务,致力于为国内外高等院校、科研机构、研发企业及相关科研人员提供科研数据支持。中国爬萌,以众包的形式为高校学生和科研人员抓取所需要的互联网数据。目前是以微博数据为主,包括博主基本信息、关注关系、微博消息数据等。

虽然国内基于微博的定量研究越来越多,而且也有一定的数据开放平台支持,但是真正利用开放数据集的研究还较少,大部分研究都是利用研究者自己开发的系统或者应用,进行数据获取,相对来说,国外的数据集开放平台的发展就比较完善,而且利用情况也相对较好。

1.2.2　数据获取方式比较。目前对于微博数据的应用研究中,数据获取方式以第一种方式为主。这些数据获取方式基本都能满足研究者的需求,但是它们之间也存在一些差异。

从适用性角度来看,利用 Open API 开发系统受到了官方发布范围的限制,研究人员只能在权限范围内对开放接口进行利用,例如新浪微博 API 并不对普通用户开放指定主题获取数据的接口,这为数据获取带来了一定的困难,使该种方法在某些情况下难以胜任;利用开放的数据集进行数据获取,受到以往研究的限制,只能研究以往出现的数据集或者其子集中的内容,可能会与研究目的、研究者需求有所偏差;而利用网络爬虫爬取微博页面适用性则相对较好,由于它可以爬取微博页面中的所有数据,对于任何研究比较适用。

从易用性来讲,使用开放的数据集应该是最容易使用的,对于研究者自身的计算机等其他方面的知识掌握要求较少;而通过 Open API 开发的系统虽然对编码有一定要求,但是在已经开放的 API 基础上进行二次开发,根据研究需求进行修改,相对来说比较容易;对于大多数非计算机专业研究者来说,通过网络爬虫爬取微博页面的获取方式略显困难,由于该种方式对计算机知识要求较高,一般研究者需要自己设计爬虫程序以控制爬行范围,而且对于后续数据的处理会显得相对较繁琐。

从稳定性来说,使用开放的数据集是相对比较稳定的,它们都是在以往的研究中经过实践验证过的;由于官方发布的 API 接口不会轻易改动,因此利用 Open API 开发系统的方法也比较稳定;相比之下,利用网络爬虫获取数据的稳定性则显得相对较差,可能会造成数据无法获取或无法解析的情况。

2　数据选择

根据研究目标,选择合适的数据集是数据分析的重要基础。在以往的基于微博数据的研究中,研究者一般以两种方式进行数据的选择:第一,针对指定主题或者用户筛选信息;第二,随机获取用户数据如用户 ID、粉丝数、关注数等,获取之后再对数据集进行筛选处理。

2.1　**指定主题或者用户**　当研究者使用微博中的数据开展社会现象的调查分析或者进行用户行为分析等与实际生活密切联系的研究时,通常采用指定主题或者用户的方式进行数据选择。根据研究内容需要,研究者通常会采用规定时间段来限制数据量。在数据选择过程中也存在着随机选择的过程,即在指定时间内指定主题后随机获取用户的信息和行为。宋恩梅等在研究中选取了“时尚”标签下排名前 50 位的用户,统计其粉丝数、关注人数和用户之间相互关注情况。利用社会网络分析方法揭示新浪微博特定标签圈的网络结构特征[18]。在基于微博的研究中,指定主题或者用户的数据选择方式颇受研究者欢迎,因为这种方式可以使研究对象的范围更为专一,更切合研究目的,研究结果也更为准确。

2.2　**随机获取用户数据**　当研究者需要研究微博及其应用的结构特点、拓扑结构或者性能评估等有关微博自身理论或实践研究时,研究者通常会采用随机获取用户数据的方式。Rui H[6]、Java A[19]都是基于 Open API 构建系统来随机获取用户信息的数据集,包括用户名、位置、更新数量、粉丝数、关注数、账号创建日期和简短的自我介绍,然后对用户行为等关注点进行分析和研究。随机性获取用户数据信息的方式能对整个微博用户群体的数据的整体把握,可以得出较为准确和全面的结果,也更有利于微博自身的开发和发展。

3　数据处理及分析

3.1　**数据预处理**　数据预处理是指将获取的数据进行前期处理,满足后续进行分析的需求。处理质量的高低直接影响着研究结果的有效性和准确性。表 1 总结了对微博数据进行预处理的基本流程。

表1 微博数据预处理过程

处理步骤	描述
清洗数据	删除重复数据
	文本清洗,去除无用的空格及符号
	去除异构数据
格式转换	统一编码格式
	统一各类型属性格式
分词	去除停用词与无意义词
	对中文微博文本进行分词处理(英文文本无此项)
	词性标注
特征提取	抽取关键词、命名实体(时间、地点、人物等)、关注数、粉丝数等特征
	对抽取到的特征进行存储

早期的文本预处理研究多是针对普通网络信息(如新闻网页和博客等长文本信息)的处理,已有一定的研究成果,而针对短小的微博信息的研究相对较少。由于微博自身的特点,预处理过程与传统长文本存在一定的差异:微博的文本内容限制在140字符之内,包含信息量小,虽然简洁但是不便处理;微博使用的随意性较大,文本信息多口语化和碎片化,传统长文本多具有连贯性;传统长文本可利用词频统计识别主题,而微博则无法识别。

在文本的预处理方面,由于中英文语言差异,在处理过程中需要分词工具对中文微博信息进行分词处理。国内一般利用武汉大学虚拟学习团队开发的ROST CM软件对文本进行分词与词频统计[20],或者利用中科院开发的开源中文分词工具ICTCLAS[21]对微博文本和用户简介等较长的文本进行分词操作并提取关键词。国外在对Twitter数据进行处理时,自然语言工具包(Natural Language ToolKit,NLTK)[22]是经常被使用的一个流行模块:它提供了大量的用于各种文本分析的工具,包括常见度量的计算、信息提取和自然语言处理。卡耐基梅隆大学开发了基于Java的Twitter分析工具Twitter NLP[23],它可以提供快速而强健的分词和词性标注功能。

对数据进行表1所示的处理之后,根据研究目标可能还需要对数据集进行再处理,比如在情感倾向分析研究中,根据极性对数据进行分类[24]或者对有情感倾向的词语进行抽取[25]。常用的方法有利用情感词典进行分类和采用机器学习的方法,使用朴素贝叶斯、支持向量机作为分类器进行分类。目前常用的分类策略是将微博信息中用户的情感倾向划分为正面情绪、负面情绪及中性情绪三种类型,与传统文本不同的是微博文本简短且包含信息量少,如何准确提取微博中的关键信息是目前情感分析领域的热点。中文常用的情感词典有HowNet和NTUSD[26],国外也有很多情感词典如OpLexicon[27]。

3.2 **数据分析方法** 数据分析阶段的主要任务是在微博数据的基础上对其进行特征提取和分析研究。一般采用以下方法:社会网络分析方法、数理统计方法和数据挖掘方法。

3.2.1 社会网络分析方法。社会网络分析方法主要利用网络拓扑关系图来反映社会结构之间的关系和属性。研究的对象是社会整体和社会结构,而不是个体。该分析方法更能从整体上把握微博的总体特征和用户交互情况,在以往的研究中也证实了在微博中运用社会网络分析方法是可行并且相对成熟的。

目前典型的社会网络分析软件主要有UCINET、Pajek、Gephi。这些工具能够将微博这个复杂网络中的个体和个体之间的相互关系抽象成结点、线以及方向,来测量个体与他们所处的网络社区之间的关系和连结,并对分析结果进行可视化。研究者可以利用这些工具对微博网络的数据信息构建的矩阵进行分析,对社区网络分析、核心用户挖掘、微博交流网络特征、微博社区网络交流结构等进行研究。

3.2.2 数理统计方法。数理统计方法是社会科学研究中一种常用的定量分析方法,该方法通过基于微博用户的基本信息数据和关注数据,利用统计学方法对其中一些参数以及参数间的相关关系进行统计分析,得出数据分布特征,如探索用户行为特征、核心用户及用户间关系、地域特征等。

常用的数据分析工具有分析统计软件SPSS,其功能非常强大,提供了11种类型136个函数,具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。研究者通常使用SPSS,以微文数、粉丝数和关注数等数据作为控制变量做相关的分析,探索这些参数之间存在的关联性,以及相关性强弱的程度。

3.2.3 数据挖掘方法。数据挖掘是采用自动或半自动的智能方式,利用关联分析、聚类分析、分类、预测、时序模式和偏差分析等技术,对大量数据进行分析并做出归纳性的推理,得到数据信息的趋势和相关性,从中挖掘出隐含的、先前未知的并有潜在价值的信息。常用的数据挖掘软件有SPSS公司的Clementine和SAS Enterprise Miner。研究中可以利用聚类分析方法对微博短文本进行聚类,按话题或者地域等共同点对微博内容或者用户进行划分,然后在此基础上再根据研究目的进行计算和分析。也可以根据大量用户的标识以及其他用户信息,利用决策树来分析用户特征或者潜在联系等内容,以支持企业决策和个性化营销的可行性。

4 研究结果的应用

微博作为社交媒体的主流平台,是大众进行信息

沟通交流的渠道。研究者们感兴趣的是透过微博平台,利用微博提供的数据资源,来了解用户的行为,揭示当前社会现象,找到当前存在的问题和缺陷,并为社会发展提供参考和支持。因此基于微博的研究,应用也十分广泛,下面将从商业应用、科学研究、公共服务三个角度对其进行概括和归纳。

4.1　**商业应用**　微博用户数量巨大且分布广泛,因此国内外很多研究者对微博用户的行为特征进行了研究分析,以期通过利用用户行为特征分析的结果支持用户个性化推荐和辅助微博营销。

首先,实现用户个性化推荐是微博营销的关键,其次,对于企业和机构来说,获取核心用户是微博营销中推广活动的有效手段,核心用户的认可,有助于消息在一定范围内的迅速传播。何黎等人利用微博用户的基本信息数据和关注数据,通过数据挖掘技术对微博的用户特征及核心用户进行分析,挖掘出社区中的核心用户,并指出针对核心用户,进行个性化营销是可行的,可以改进产品和服务,增加企业效益[28]。陈渊等人以用户的关注人数、粉丝数和发布的微博数为标准对用户信息进行了定量分析,针对不同特征的用户群体提出了相应的标签推荐方法[29]。

面向微博的情感分析在商业领域也有十分重要的作用,如今互联网不仅成为人们发布信息的重要媒介,也是人们表达观点和情感的重要工具,企业和机构更应该抓住用户的情感特征以及感情倾向,对用户情感进行分析[30],了解用户心理和需求和对新产品、新品牌等的反应,以便灵活决策,实现个性化推荐。

国外还有学者基于微博进行股票预测。慕尼黑工业大学的两位学者对 Twitter 进行了较为细致的分析,他们筛选出提到标准普尔 100 指数中的公司的推文(比如 $ AAPL 代表苹果公司),分为“买入”、“持有”或“卖出”三类,并算出每支股票的看涨程度。结果发现,推文的总数和交易量,看涨程度和标准普尔 100 指数之间,都有密切关系[31]。麻省理工学院的张雪等人,从一个白名单 IP 获取了从 2009 年 3 月 30 日至 2009 年 9 月 7 日共六个月的微博,根据情绪词将推文标记为正面或负面情绪。结果发现,无论是像“希望”这样的正面情绪,还是像“害怕”、“担心”这样的负面情绪,其占总推文数的比例,都预示着道琼斯指数、标准普尔 500 指数、纳斯达克指数的下跌。研究者据此认为,只要是情绪的突然爆发,无论希望或担忧,都反映出人们对于市场的不确定性,因此能预测股市之后的走向[32]。

4.2　**科学研究**　随着微博的飞速发展,各领域的学者逐步重视微博的使用,将其视为自由、开放交流思想的平台,通过微博可以对科研成果进行评价、对所引用文献以及对教学方法进行讨论,因此利用微博数据,可以进行科研辅助,如发掘学科热点、学科领袖人物,改善微博学术交流。有学者以新浪名人堂的 514 位专家学者数据为调查样本,揭示了利用微博进行学术交流中的知名领袖人物和普通领袖人物,通过博文分析,进行了微博学术信息评价,从而为微博学术交流的完善提供建议[33]。Ross C 等学者将 Twitter 作为三个物理会议的学术平台,收集分析微博数据,强调了 Twitter 以一种新的方式增强了社区之间的相互影响,扩大了成员沟通和参与的程度,而且验证了数字人文社区中群体协作后得到的知识比独立研究更有价值[34]。盛宇以基于新浪微博的“数据挖掘”领域学科热点的研究作为实例,选取关键词“数据挖掘”,使用 ROST CM 软件在规定时间段内采集数据,基于微博发现数据挖掘领域热点,并进行热点跟踪和分析,发现该方法同传统热点分析结论有重合部分,但又可以反映出传统方法所无法反映出的最新热点[35]。

4.3　**公共服务**　随着社会媒体的发展,微博已经对现实社会和人们的日常生活产生了重要的影响,尤其在网络舆情控制和卫生健康方面。

网络舆情监控是指通过对网络各类信息汇集、分类、整合、筛选等技术处理,再形成对网络热点、动态、网民意见等实时统计报表的一个过程[36]。如果一则信息在微博上被跟从者过万的知名博友转帖,在 10 分钟之内就很有可能形成全国性的舆论热点,其信息传播的速度和广度上都非常惊人,因此,微博可以说是当前十分优秀的网络舆情监控平台之一。熊祖涛提出了一种基于 Web 文本信息抽取方法,以及基于文本抽取的舆情分析技术,通过分析微博文本和微博用户的数据,确定热点话题、敏感核心用户,从而对于敏感事件的舆情态势进行预警[37]。孙帅等以“北京发布”这一政务微博,在“7 · 21”北京特大暴雨灾害事件期间发布的相关微博及其评论为数据样本,通过分析微博时间响应情况、微博内容响应情况与微博交互响应情况,对突发事件下政务微博信息传播规律的真实图景进行了描述,并结合该个案特征,提出了利用政务微博导控突发事件网络舆情的若干建议[38]。同时政府机构也可以在微博平台上对人民实施舆论引导,通过分析微博用户行为等数据,利用用户之间的相互关系及其发微行为,评价其在微博网络中的影响力,找出网络中的主干节点,对核心用户时刻关注,控制舆论导向和传播,为新政策的实施奠定群众基础,为微博网络舆情传播的监控提供帮助。微博情感分析也可以用于异常或突发事件监测,北航软件开发环境国家重点实验室的先进网络分析研究小组(GANA)对收集到的发布于 2011 年的近 7000 万条微博进行情感分析,根据各类

情绪比例的波动,他们提出一个快速的异常点发掘算法,并发现2011年全年发生的一些典型的异常或突发事件,均可以被有效地检测出来,甚至可以观察事件带来影响的持续时间等,有效地论证了基于微博情感理解实现事件监测的可信性[39]。

在卫生健康方面,社交网络系统可以对疾病的传播进行辅助追踪和检测。Xiang J等在研究过程中监测了5种流行病,提出了一个传染病爆发传播检测系统(EOSDS),确保能够充分利用Twitter可检索的丰富的数据,关注疾病爆发情况[40]。Sadilek和他的团队在2010年经过一个月时间对在纽约城中63万个用户通过GPS定位分析了440万的Twitter消息,通过编写一个机器学习的算法来分辨Twitter消息哪个是来自于健康的人,哪个消息是来自于生病的人,基于此绘制出了一天流感发病的热度地图,准确率高达90%[41]。第一个通过分析社交内容来为消费者提供疾病传播跟踪信息的网站Sickweather的创始人Graham Dodge,曾在分析完Facebook和Twitter上的1700万条状态和微博后,指出了如何通过社交网络追踪美国境内疾病的传播情况,如今,通过Sickweather已经可以跟踪多种疾病在世界各地的传播情况[42]。虽然微博信息的复杂性为数据分析带来了一定的困难,但是通过微博信息判断疫情的扩散方式、途径,并基于此寻找控制措施仍是今后研究的主要趋势之一。

5 总结和存在问题

国内外微博的发展为学术界带来了新的研究课题,为研究者们提供了大量的数据信息和广泛的研究范围。但是由于微博的结构和特点,在研究中仍然存在一些局限。本文总结了微博研究的局限性和国内外基于微博研究的异同点。

5.1 使用微博数据进行研究的局限性 微博消息比较短小,往往在140个字以内,所包含的信息量偏少,往往造成分析样本较大;微博数据往往是海量的,杂乱无章的,名为"僵尸网络"的计算机程序总是会在像Twitter这样的网络上发布类似垃圾电子邮件的垃圾信息,如何过滤获得最有效的数据,是数据处理时的重要问题之一;微博语言不同于一般的书面语言,随意性较强,情绪化表达较多,表达中实褒暗贬,一词多义的情况多有出现,在一定程度上干扰了微博情感的识别,而且微博中经常流行一些新词,如果研究中使用传统的分词表或者更新不及时,会让分词结果产生偏差,从而对分析结果造成影响;基于微博开发的第三方应用具有很大的局限性,绝大多数应用功能都比较单一:不仅分析功能单一,而且无法根据研究者的需要进行调整和改变,影响研究结果;微博平台虽然给研究者提供了大量的用户数据,但实际的微博用户中存在着僵尸粉、死粉等无效用户,在数据处理过程中如何辨识这些用户对研究者来说是一个很大的难题;微博数据的获取接口受微博服务商的限制,以国内的新浪微博为例,其官方API对于用户的数据获取在获取频次、获取量、获取对象方面都作了比较严格的限制。

微博数据是对现实社会的人及组织行为的映射,虽然利用微博数据进行研究还有所局限,但是随着微博的广泛使用,其数据所具有的广泛性、代表性、实时性以及真实性是许多数据无法比拟的,同时微博构成的庞大社交网络,对于分析个人和群体行为模式也具有重要的参考价值,因此我们要肯定微博数据存在的利用价值,不断完善微博数据的处理技术,从而更高效地利用微博数据。

5.2 国内外基于微博研究的异同点 微博作为重要的社交网络,对于了解用户个人,了解社会群体都具有很大的帮助。因此对于微博数据的研究也成为了当前的研究热点。国内外虽然都以微博为研究对象,既有相同之处,也有相异之点。其相同之处主要表现在:第一,研究对象都是国内外数据量巨大且具有代表性的微博平台,国内大部分研究是基于新浪微博,国外研究则是基于Twitter;第二,研究者们的关注点类似,研究者们基本都关注微博用户行为特征和微博本身结构特征;第三,大多研究关注微博背后揭示的社会意义及其利用现状。

国内外对于微博的研究又有着细微的差别:国内研究者们更关注的是微博自身结构以及微博用户行为特征等,以观察其背后的意义,如利用微博行为进行舆情监测和微博营销;而国外的研究则比较广泛,不仅包括对Twitter本身性质的研究,还包括很多利用Twitter数据对城市特点,政府舆情,卫生健康等社会多方面的研究。究其原因,则可能与中文数据处理较英文复杂,难以快速准确处理有关,同时还有可能与国内外微博的发展历史不同、国内外国情不同、社会文化不同,使得对微博开发和利用有所差异,在未来的研究中可以进行深入地探讨。

参 考 文 献

[1] 曹 磊,陈薇娜,缪其浩,等. 大数据:数字世界的智慧基因[N]. 文汇报. 2011-11-08

[2] US Library of Congress to archive Twitter messages[EB/OL]. [2013-5-11]. http://phys.org/news190479285.html

[3] ChinaVenture[EB/OL]. [2013-4-15]. http://news.chinaventure.com.cn/47/20130222/110416.shtml

[4] 廉 捷,周 欣,曹 伟. 新浪微博数据挖掘方案[J]. 清华大学学报:自然科学版,2011(10):1300-1305

[5] BeJSON[EB/OL]. [2013-6-22]. http://www.bejson.com/

[6] Rui H, Whinston A. Information or Attention An Empirical Study of User Contribution on Twitter[J]. Information Systems and e-Business Management, 2012(10):309-323

[7] Lee R, Wakamiya S, Sumiya K. Urban Area Characterization Based on Crowd Behavioral Lifelogs over Twitter[J]. Personal and Ubiquitous Computing,2013,17(4): 605-620

[8] Cheong M ,Lee V. Twittering for Earth: A Study on the Impact of Microblogging Activism on Earth Hour 2009 in Australia[C]. Proceedings of the Second International Conference on Intelligent Information and Database Systems:Part II, 2010: 114-123

[9] 杨成明. 微博客用户行为特征实证分析[J]. 图书情报工作,2011,55(12):21-25

[10] 张国安,钟绍辉. 基于微博用户评论和用户转发的数据挖掘[J]. 电脑知识与技术,2012(27):6455-6456

[11] Culotta A. Lightweight Methods to Estimate Influenza Rates and Alcohol Sales Volume From Twitter Messages[J]. Language Resources and Evaluation,2013,47(1):217-238

[12] O'Connor B, Balasubramanyan R, Routledge B R,et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series[C]. Proceedings of the International AAAI Conference on Weblogs and Social Media,2010:122-129

[13] GetTheData[EB/OL]. [2013-5-20]. http://getthedata.org/

[14] DataMob[EB/OL]. [2013-5-20]. http://datamob.org/

[15] Stanford Network Analysis Project[EB/OL]. [2013-6-20]. http://snap.stanford.edu/data/

[16] 数据堂[EB/OL]. [2013-5-20]. http://www.datatang.com/

[17] 中国爬萌[EB/OL]. [2013-6-24]. http://www.cnpameng.com/

[18] 宋恩梅,左慧慧. 新浪微博中的"权威"与"人气":以社会网络分析为方法[J]. 图书情报知识,2012(3):43-54

[19] Java A, Song X D, Finin T,et al. Why we Twitter: An Analysis of a Microblogging Community[J]. Lecture Notes in Computer Science, 2009,5439: 118-138

[20] 李 如. 谣言在微博中的传播研究[D]. 安徽大学,2012

[21] Ictclas[EB/OL]. [2013-4-27]. http://www.ictclas.org/

[22] Russell M A. Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites[M]. Sebastopol: O'Reilly Media, 2011

[23] Twitter NLP and Part-of-Speech Tagging[EB/OL]. [2013-6-24]. http://www.ark.cs.cmu.edu/TweetNLP/

[24] 王 晶,朱 珂,汪斌强. 基于信息数据分析的微博研究综述[J]. 计算机应用,2012,32(7):2027-2029,2037

[25] 周胜臣,瞿文婷,石英子. 中文微博情感分析研究综述[J]. 计算机应用与软件,2013,30(3):161-164,181

[26] 蒋盛益,麦智凯,庞观松. 微博信息挖掘技术研究综述[J]. 图书情报工作,2012,56(17):136-142

[27] Souza M, Vieira R. Sentiment Analysis on Twitter Data for Portuguese Language[C]. Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language,2012:241-247

[28] 何 黎,何 跃,霍叶青. 微博用户特征分析和核心用户挖掘[J]. 情报理论与实践,2011(11):121-125

[29] 陈 渊,林 磊,孙承杰,等. 一种面向微博用户的标签推荐方法[J]. 智能计算机与应用,2011(3):21-26

[30] 谢丽星,周 明,孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报,2012(1):73-83

[31] Sprenger T O, Tumasjan A, Sandner P G, et al. Tweets and Trades: The Information Content of Stock Microblogs[EB/OL]. [2013-6-20]. http://ssrn.com/abstract=1702854

[32] Zhang X, Fuehres H, Gloor P. Predicting Stock Market Indicators Through Twitter - "I hope it is not as Bad as I fear"[EB/OL]. [2013-6-20]. http://www.ickn.org/documents/COINs2010_Twitter4.pdf

[33] 盛 宇. 基于微博的学术信息交流机制研究——以新浪微博为例[J]. 情报研究,2012(7):62-66

[34] Ross C, Terras M, Warwick C, et al. Enabled Backchannel: Conference Twitter use by Digital Humanists[J]. Journal of Documentation. 2011, 67(2): 214-237

[35] 盛 宇. 基于微博的学科热点发现、追踪与分析——以数据挖掘领域为例[J]. 图书情报工作,2012,56(8):32-37

[36] 网络舆情监控[EB/OL]. [2013-6-20]. http://baike.baidu.com/view/2013618.htm

[37] 熊祖涛. 针对微博数据的信息抽取与舆情分析[J]. 信息系统工程,2013(3):156-158

[38] 孙 帅,周 毅. 政务微博对突发事件的响应研究——以"7·21"北京特大暴雨灾害事件中的"北京发布"响应表现为个案[J]. 电子政务,2013(5):30-40

[39] 北 航:微博情感分析可用于异常或突发事件的监测[EB/OL]. [2012-04-11]. http://www.cnbeta.com/articles/182029.htm

[40] Xiang J, Soon A C, Geller J. Epidemic Outbreak and Spread Detection System Based on Twitter Data[C]. Proceedings of the First international conference on Health Information Science, 2012:152-163

[41] Twitter 是怎样预测我们会生病的?[EB/OL]. [2013-6-12]. http://www.cnbeta.com/articles/205491.htm

[42] Sickweather: Twitter 可以追踪疾病,能预测流行病爆发吗?[EB/OL]. [2013-06-18]. http://www.199it.com/archives/51516.html

(责编:白燕琼)