基于爬虫技术的社会化观测数据获取及应用

王书欣 陈元昭 张舒婷

如何获取及利用网络中蕴含的大量社会观测数据成为新媒体时代气象观测发展的重点。深圳市气象局利用爬虫技术获取数据，并通过机器学习方法对数据进行筛选过滤，建立了一套高扩展性、高效性和低成本的气象社会观测数据的采集系统，快速获取、筛选、分析和提取有价值的、多样化的气象相关的社会观测数据，并对数据加以分析应用，为预报员进行公众服务、大城市气象灾害风险预警提供支撑。

DOI：10.3969/j.issn.2095-1973.2019.03.038

传统地面综合气象观测是当前对天气进行预测的一种重要手段，但随着社会经济的快速发展及计算机网络技术的不断完善，各大领域的数据量都飞速增加，使人们进入到大数据社会时代，移动网络成为公众获取天气信息的主要渠道，同时也成为信息发布的参与者之一，微博则是最具影响力的传播途径。在这一背景下，为使微博中大量社会数据更好地为气象部门提供服务，就必须完善社会化观测数据获取方法。因此探讨基于爬虫技术基础之上的社会化观测数据与获取具有重要意义。

目前，尽管国内气象部门尚未有基于爬虫技术的数据获取技术，但国内外大量专家学者针对网络爬虫技术开展了大量的研究工作。基于以往研究，深圳市气象局首度尝试建立基于爬虫技术的社会化观测数据获取平台。本文将着重基于爬虫技术探讨社会化观测数据获取及应用，以打破传统气象观测壁垒，开展多源观测数据的在线融合并推进气象观测的社会化，弥补观测空缺，利用获取数据分析热度和情感，使社会化数据可以在天气预报服务中得以利用。

1 社会化观测数据获取平台建设

随着大数据时代的到来，气象部门需建设一套高扩展性、高效性和低成本的气象社会观测数据的采集系统，通过快速获取、处理、分析和提取有价值的、多样化的气象相关的社会观测数据，以满足当今大数据环境下对于文本、图片等数据的采集、存储、分析及可视化需求，为社会提供更优质的服务。

为此，搭建基于爬虫技术的社会化观测数据获取平台，可以完成数据获取，数据过滤以及数据分析三部分工作。通过爬虫技术获取包括冰雹、龙卷等现有探测设备无法精准捕捉的中小尺度天气现象，其形式包括文字、图片、视频等。由于爬取到的数据有大量的重复、过时甚至是虚假信息，需要对其进行过滤，最终将可用的数据进行气象实况监控和公共服务舆论情感分析，后文将详细阐述这部分工作。

平台包含数据获取模块、数据存储模块和结果分析展示模块（图1）。基于气象社会观测信息的特点，采用分布式数据采集技术，并存储在数据库中，通过建立机器学习、深度学习的模型对数据进行计算和分析，得到统计信息，最终通过可视化的界面展

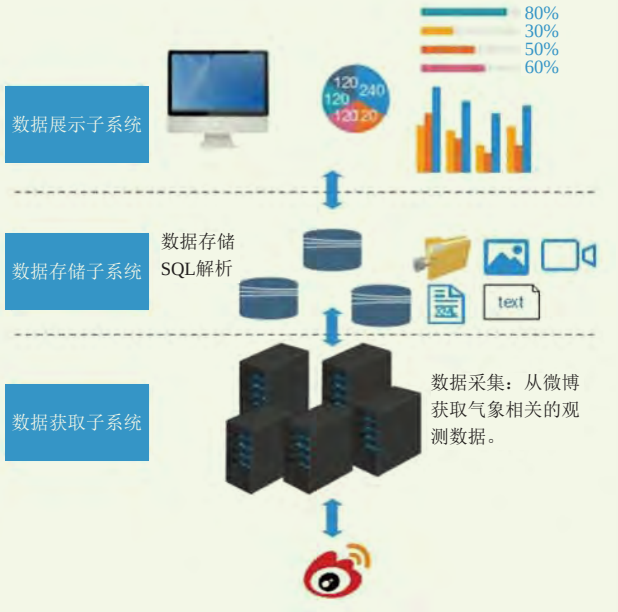

图1 社会化观测数据获取平台设计模型

收稿日期：2018年11月30日；修回日期：2019年4月14日

示。三个模块分工明确，下层向上层提供可靠服务，最终构成整个完整的平台。

2 社会化观测数据来源

深圳天气微博建立8年，截至2018年11月，粉丝187万人，仅2018年阅读量达10亿次，转评次数超过50万次，超强台风山竹话题讨论量过100万次，在如此庞大的信息库中存在着海量的社会自发上传至社交媒体的观测信息，与传统气象监测如自动站、雷达、卫星数据不同，社会观测数据虽不能精确测量各种气象要素，但可以监测到包含冰雹、龙卷等罕见无法监测的天气现象，以及积水、滑坡等气象部门无法掌握的衍生灾害实况，这些数据有效地对传统气象观测数据进行补充，通过收集该信息的发布时间、发布地点以及相关内容包括图片、视频等信息，扩大气象数据观测网，共实现爬取冰雹、大风、暴雨、雷电、龙卷5种气象类信息，以及积水、洪涝、滑坡3种灾害影响类信息。

目前平台数据主要来源于新浪微博，但此技术同样可应用于微信、各大门户、新闻网站以及各政府部门网站，从而获取气象信息和其影响信息，此项工作未来将逐步开展。

3 社会化观测数据获取、筛选与分析

3.1 基于爬虫技术的社会化观测数据获取

网络爬虫技术是互联网搜索功能中一项基本技术，其在中国最成功的应用就是百度搜索引擎，通过一传十、十传百的裂变搜索方式，实现信息的网状获取，该技术的优点在于信息获取速度快、内容全。为此引入网络爬虫技术来获取新浪微博中社会观测数据，并按照一定的预设关键词、地域、时间等阈值进行自动识别、抓取气象相关信息的程序和脚本。

基于网络爬虫技术获取社会化观测数据的方法主要包括：基于第三方软件或者第三方微博数据集的方法、基于新浪公开API的方法和网络爬虫抓取的方法。通过使用爬虫技术中通用的Scrapy爬取方式，可以同时发送多条爬取请求，同步进行信息爬取，最大化增进爬取效率。由于采集到的微博数据并非都是描述冰雹、龙卷风、大风等发生信息的数据，需要采用文本分类技术，将实际含有上述关键词的文本识别出来，同时记录其相关的图片、视频、网页链接等信息。

所获取的信息同时需判断以下几个条件以便进行后期分析：1）记录的信息与灾害性天气相关；2）如果与灾害性天气相关，有明确的发生气象现象的位置或时间；3）记录的信息与舆情是否相关。

目前，基于新浪微博平台的数据进行爬取的数据可在5 min内完成，但结合深圳实际天气情况以及工作需要将爬取频率保持1次/h，鉴于目前雷达数据的更新频率为6 min，在恶劣天气下，也可后台更改爬取频率为1次/6 min。

3.2 无效数据过滤筛选

通过爬虫技术获取的文本信息存在大量失真、失效、无用甚至是广告数据，为保证数据的可用性，需对其进行过滤筛选。通过机器学习方式，使用支持向量机（SVM）模式进行数据分类与回归分析，由预报员人工判别给定的多组社会数据训练实例，将训练实例分类标记为有效、无效两类，通过机器不断学习，使SVM模型成为非概率的二元线性分类器。当出现新的实例时，SVM模型将其进行分类为有效或无效其中一类。经过大量数据训练，机器将过滤筛选后的数据推送至前段展示，预报员仍可手动调整信息类别，通过不断增加训练实例，形成正反馈机制，不断优化筛选模型。

3.3 重复微博数据过滤

利用爬虫技术采集筛选后的微博数据仍存在大量的重复数据从而影响分析结果，选用simhash算法去重可以高效地将爬虫系统每日数以千万级的数据进行去重合并，通过对文档关键词进行拆分并整理成关键词集合，对比不同文档关键词集合相似度，去除重复数据。

目前对广东省范围内的冰雹、大风、暴雨、雷电、龙卷5种气象类信息，以及积水、洪涝、滑坡3种灾害影响类信息进行爬取，共获取到53900条数据，其中气象类数据48538条，灾害影响类5362条，由图2逐日数据结果展示可以直观获知深圳4—9月重大灾害天气发生时间，如8月底持续季风低压降水和9月15—16日超强台风山竹影响，对于4月前汛期深圳无强对流、回南天等高影响天气这种反例也有明显表现。同时可以获知社会数据获取强度，对强天气过程、衍生灾害进行准确识别。

4 主题词热度分析和情感分析应用

4.1 主题词热度分析

气象中所关注的热度，是市民在一段时间内所关注的某一天气类型、灾害信息或是相关话题，我们提取其关键字作为热度的主题词。传统的基于词频分析的主题模型不能从语义中进行分析，而将微博热度作为计算基数的LDA主题模型则是将评论数、转发数纳

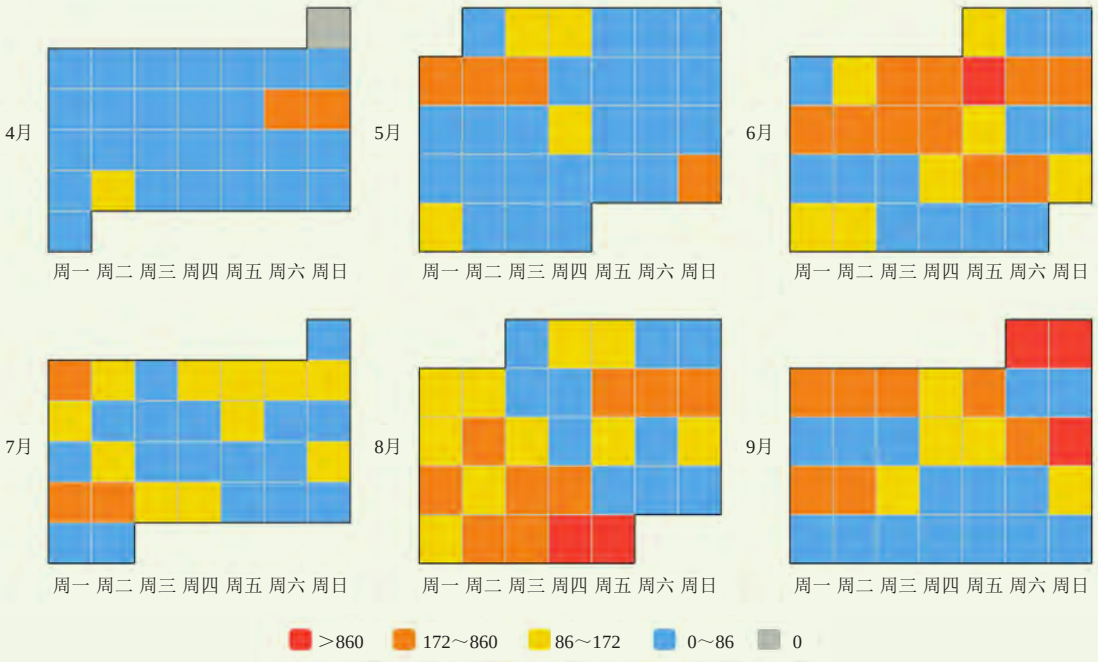

图2 2018年4—9月逐日获取数据结果展示（条）

入计算，获取微博主题热度分布，得到真实的高关注度数据信息，可供预报员更有针对性进行服务或是发现并处理舆情。

对于微博气象信息的挖掘，由于微博用户之间具有关注与被关注、转发与评论的关系，社会关系网庞大而复杂，常规的分析方法无法胜任。“深圳天气”微博信息构成的文本矩阵的稀疏性和高维度性，选择使用潜在狄利克雷分布的主题生成模型（LDA）来完成基于潜在语义分析的文本挖掘方法进行的微博主题的挖掘。

2018年6月6—8日受南海热带低压影响，广东大部出现大暴雨。通过统计2018年6月5—9日微博中社会化数据得出主题词热度如图3所示，经与天气实况以及预报员舆情监控对比看来，此次记录为真实有

图3 2018年6月5—9日主题词热度分析结果展示

效。统计显示最受网友关注的天气现象为台风、暴雨和大暴雨，深圳、江门和广州则为受影响关注度最高的地区，说明本次过程对珠三角地区的高密度人群影响更为显著；同时天气预警信息和高考信息也同样备受关注，在进行公众服务时应将其与其他高热度主题结合共同服务。

经过未来的长期主题词热度数据积累，可以总结出公众所真实关心、讨论的天气现象，从而根据需求，加大该天气条件下的气象服务力度。

4.2 情感分析

气象服务由于其必然存在的不准确性以及目前与公众所期望的精细化预报间的差距，气象部门经常陷入舆论风波，由于舆情信息的不能及时获取，往往不能正确地化解舆情。而基于爬虫技术获取到的数据中除大量的社会观测数据，其中还包含着社会的情感状态，包括正面积极鼓励的言语或是负面批评的指责，分析数据中的情感走向有助于更好掌握舆情动态，引导大众评论走向，为气象服务做出正面回应。

目前爬取3万条微博，84万条深圳天气微博的评论，人工对其中1万条评论进行情感定性，分为积极评价（pos）和消极评价（neg）以及中性，根据多元伯努利事件模型（NB）、支持向量机（SVM）、卷积神经网络（CNN）、循环神经网络（RNN）、霍普菲尔网络（HN）、深度霍普菲尔网络（Att+HN）、AVG函数、时间敏感网络（TSN） 8种机器学习方法

对3万条评论进行分析实验，通过计算情感定性准确率、相关系数以及方差确定学习方法，结果如图4。

Method	P_{pos}	P_{neg}	R_{pos}	R_{neg}	F_{pos}	F_{neg}
NB	0.7287	0.5929	0.7145	0.6097	0.7215(7)	0.6012(8)
SVM	0.7464	0.5864	0.6828	0.6597	0.7132(8)	0.6209(5)
CNN	0.7535	0.6044	0.7046	0.6619	0.7282(5)	0.6318(3)
RNN	**0.7591**	0.6064	0.7020	**0.6733**	0.7294(4)	0.6381(2)
HN	0.7326	0.5964	0.7152	0.6172	0.7238(6)	0.6066(7)
Att+HN	0.7333	**0.6370**	0.7713	0.5886	**0.7518(1)**	0.6118(6)
AVG	0.7577	0.6281	**0.7355**	0.6551	0.7465(2)	**0.6413(1)**
TSN	0.7496	0.6099	0.7172	0.6487	0.7330(3)	0.6287(4)

图4 微博评论数据的对比试验结果

统计表明，RNN、Att+HN 、AVG三种方法的分析结果更为准确，其中AVG方法的方差结果更优，准确率和相关性结果也名列前茅，综合考虑AVG方法更稳定更适合进行情感分析。

同样，对6月5—9日降雨过程进行情感分析，共获取15060条数据（表1），发现大部分评论感情色彩以中性为主，其中大部分是提出咨询，恶劣天气来临或正在影响时人们关注度大幅提高，过程结束关注度急剧下降，9日数据量仅为7日的1/10；恶劣天气下消极评价量、比例同步上升，而在恶劣天气最初影响时人们更愿意发表有感情色彩的言论，6日积极和消极评价占比总评价数为22.3%，相较其他日期上升3%～5%。预报员根据以上情感分析数据及时引导舆论。

表1 2018年6月5—9日情感分析结果展示

日期	positive	neutral	negative	total	pos_ratio	neu_ratio	neg_ratio
6月5日	162	1080	67	1309	0.123759	0.825057	0.051184
6月6日	519	2866	302	3687	0.140765	0.777326	0.081909
6月7日	683	4485	429	5597	0.12203	0.801322	0.076648
6月8日	565	3219	232	4016	0.140687	0.801544	0.057769
6月9日	78	360	13	451	0.172949	0.798226	0.028825

5 结论及展望

根据过去一年的平台建设与数据获取分析发现，基于爬虫技术来获取社会化观测数据可以有效地补充常规气象观测的不足，尤其是在冰雹、大风、暴雨等气象灾害发生时可以快速获取大量信息，并获取其带来的影响，加大舆情监控，为预报员进行公众与决策服务进行数据支撑。通过爬虫技术获取到的数据我们可以清楚获知灾害发生的种类、时间、时长与地点，并进行记录统计，为预报和决策服务提供支持；通过主题词热度分析，预报员可以清晰感知公众关注热点，并有针对性地开展公众服务；情感分析帮助预报员实时监控舆情，在恶劣天气或预报失误时，及时化解舆情。

未来爬虫技术在社会化观测数据将结合雷达与自动站实况进一步优化数据筛选结果，加大其真实可用性，并且获取途径将不仅限于新浪微博平台，深圳天气微信同样具有100万粉丝，年阅读量超过1000万次，各大新闻客户端如腾讯、网易、今日头条也具有极高的互动性，在上述平台开展社会化观测数据获取工作，可以进一步扩大数据来源。该技术也可运用到政府网站及其他类型网站中，以用于获取如河道、水位、浪潮等基础信息的更细，使决策服务技术得到更多数据支撑。

深入阅读

杨富莲, 2017. 地面综合气象观测能力提升对策. 科技与创新, 9: 47-48.
姜青山, 2018. 浅谈气象服务App的开发与应用. 科技风, (1): 124-124.
王杰, 2017. 基于微博大数据的舆情监测系统的设计与实现. 天津: 中国民航大学.
刘庆华, 覃茹芊, 2013. 探索区域气象观测站社会化保障的新模式. 气象研究与应用, 34(z2): 170-171, 173.
石磊, 2013. 新浪API与网络爬虫结合获取数据的研究与应用. 中国电子商务, (22): 58-59.
毛夏, 李磊, 江崟, 等, 2017. 深圳超大城市气象探测数据在科学研究中的应用. 广东气象, (6): 2-5.

（作者单位：深圳市气象局）