线上降雨灾情检测系统设计与应用

黎洁仪[1,2],梁之彦[2,3],范绍佳[2,4],梁家鸿[1]
(1. 广州市突发事件预警信息发布中心,广东 广州 511430;
2. 广东省环珠江口气候环境与空气质量变化野外科学观测研究站,广东 广州 510275;
3. 广州市气象台,广东 广州 511430;
4. 中山大学 大气科学学院,广东 广州 510275)

摘 要:结合社会媒体大数据获取城市降雨灾情数据和开展灾害风险评估是一种新的可行途径。但互联网数据量大,有效处理数据是工作中的难点。为此提出利用社交媒体数据,并基于降雨专业词汇、广州地区语言特色、支持向量机算法以构建降雨灾情文档分类模型。同时根据数据采集与预处理、降雨灾情文档分类模型、灾情权重分级和热点分析的流程设计了广州线上降雨灾情检测系统。该系统采用 B/S 架构,利用 WEB 与 GIS 技术,实现了灾情应用管理、风险告警、数据分类、数据过滤、数据采集的功能。实际运行效果表明,系统利用机器学习算法解决了大量数据处理效率低下的问题,同时通过灾情热点分析结合利用气象雷达、自动站观测数据进一步提高灾情提取的准确度,以自动检测和评估降雨雨情、灾情的状态是可行的,在灾情收集业务应用上具有一定的参考价值。

关键词:数据挖掘;机器学习;灾情提取;文本分类;社交媒体
中图分类号:TP319 **文献标识码**:A **文章编号**:1673-629X(2022)08-0191-06
doi:10.3969/j.issn.1673-629X.2022.08.031

Design and Application of Online Rainfall Disaster Detection System

LI Jie-yi[1,2],LIANG Zhi-yan[2,3],FAN Shao-jia[2,4],LIANG Jia-hong[1]
(1. Guangzhou Emergency Early Warning Release Center,Guangzhou 511430,China;
2. Guangdong Provincial Observation and Research Station for Climate Environment and Air Quality Change in the Pearl River Estuary,Guangzhou 510275,China;
3. Guangzhou Meteorological Observatory,Guangzhou 511430,China;
4. School of Atmospheric Science,Sun Yat-sen University,Guangzhou 510275,China)

Abstract:Combining social media with big data to obtain rainfall disaster data and carry out disaster risk assessment is a new feasible way. However,the Internet has a large amount of data,and it is difficult to deal with data effectively. Based on the social media data, professional vocabulary, Cantonese and support vector machine algorithm, rainfall disaster text categorization model is established. Guangzhou online rainfall disaster detection system is designed based on the processes of data collection and preprocessing, rainfall disaster text categorization model,disaster weight classification and hot spot analysis. The system uses B/S architecture,WEB and GIS technology to realize the functions of disaster application management,risk alarm,data classification,data filtering,and data acquisition. The actual operation result shows that the system uses machine learning algorithm to solve the problem of low efficiency of a large number of data processing,and it is feasible to further improve the accuracy of disaster extraction by analyzing the hot spot of disaster and using the observation data of meteorological radar and automatic station,so as to automatically detect and evaluate the state of rainfall and disaster. It has a certain reference value in the application of disaster collection business.

Key words:data mining;machine learning;disaster information extraction;text categorization;social media

0 引 言

近年来中国城市化发展迅速,城市内涝和交通拥堵等“城市病”也随之而来。特别是降雨造成的交通拥堵、涵洞和道路积水等,严重影响了城市的运行,甚至威胁到公众的人身安全。针对这类“城市病”,及时准确的雨情和内涝灾情信息能在城市应急处置和防灾

收稿日期:2021-08-15 **修回日期**:2021-12-17
基金项目:广东省科技计划项目(科技创新平台类)(2019B121201002);广州市科技计划项目(201803030014)
作者简介:黎洁仪(1984-),女,硕士,高级工程师,研究方向为气象防灾减灾和公共服务、信息服务。

救援中发挥至关重要的作用。但传统的降雨灾情收集往往基于现场调查,而恶劣的天气非常不利于实时了解现场状况。

随着互联网和社交媒体的发展,微博、微信等具有广泛参与性和实时性的社交资讯,在灾情提取中表现出良好的应用前景,国外已有在突发事件、灾情收集等方面的应用[1-3],国内也有利用社交媒体进行自然灾害应急管理[4]、地震应急[5]、台风灾情[6]、城市内涝[7]、大风[8]等方面的灾情挖掘和分析。多方研究表明社交媒体大数据在及时提取和分析灾情中具有充分可行性。但如何从海量数据中快速提取对应的信息是灾情采集的难点,作为拥有庞大用户的社交媒体,每天都能产生大量的数据。人工分类存在成本高、效率低的缺陷,而机器学习已实现在文本分类[9-10]、目标识别方面的应用[11-12];基于机器学习如神经网络、K 近邻、决策树、支持向量机(support vector machine,SVM)等方法在文档信息提取和识别上已取得良好的效果,同时有实验显示:SVM 在文档分类识别的精度上存在优势[13-17]。基于以上背景,该文利用微博、微信收集到的带地理位置信息的社交媒体数据,运用自然语言处理及机器学习技术,设计高效的线上降雨灾情检测系统,从而充分利用社交媒体雨情和灾情动态资讯,为自动挖掘第一手降雨情报提供技术支持,以提高大城市降雨灾害应急管理和处置的效率。

1 线上降雨灾情检测流程设计

1.1 总体流程设计

线上降雨灾情检测的总体流程设计如图 1 所示。

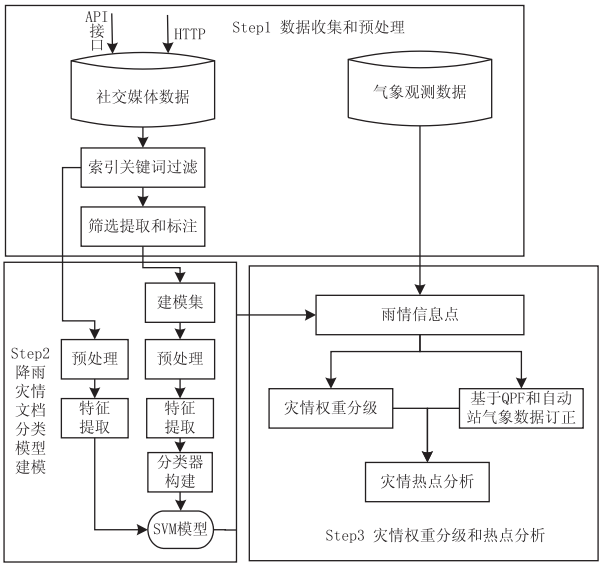

图 1 线上降雨灾情检测的总体流程

线上降雨灾情检测的总体流程包括:数据采集和预处理、降雨灾情文档分类模型建模、灾情权重分级和热点分析。

1.2 数据采集和预处理

(1)降雨灾情提取索引关键词目录。通过 ICTCLAS(中国科学院计算技术研究所开发的汉语词法分析系统)的语义分析,根据广州地区灾害性天气时期社交媒体的高频词和关键词,以及本地用语特征,挑选了雨情相关的(雨、大雨、暴雨、冰雹、落狗屎)及积水相关的(水浸、大水、积水、淹、涝、涨水、洪水、水灾)索引关键词共 13 个。

(2)社交媒体数据采集。从新浪微博的开发者平台接口获取位置微博数据,通过自建微信灾情上报 HTML5 页面获取微信数据,以匹配到任意索引关键词的文档信息数据构建成降雨灾情检测基础信息库。

(3)数据标注及建模集划分。从降雨灾情检测基础信息库通过人工筛选提取部分文档信息组成降雨灾情文档分类模型的建模数据,按照雨情信息和其他信息的二分类要求对文档信息条目进行人工标注。共标注了雨情信息(766 条)和其他信息(234 条)数据,再按 70% 为训练集、30% 为测试集的比例从中随机分割数据构成建模集。

(4)数据预处理。对降雨灾情检测基础信息库数据进行去重去噪,包括剔除重复数据、特殊符号及错别字过多或字数极少的文档;采用百度停用词表和 ICTCLAS 分词工具进行去停用词、去特殊字符、分词等预处理。在文档预处理中,由于广州地区用词有一定的特色词汇和易混淆的词条,如“落汤鸡”(粤语的“淋湿”)、“落狗屎”(粤语的“下大雨”)等,为此制定了专用分词词库供文档预处理使用。专用分词词库的来源为:高频词、热门人名和地名、粤语特色词汇、指示性灾情词等。部分词条如图 2 所示。

落汤鸡	淹没	雨水节气	风雨兼程
落狗屎	龙吸水	周冬雨	挥汗如雨
落雨大	雨点	雨神	冰雹预警
水浸街	谷雨	暴雨预警	雷雨大风预警

图 2 专用分词词库的部分词条

(5)降雨灾情检测信息特征向量提取。将预处理后的降雨灾情检测信息表示成能表征文档语义的词语序列。对词语序列进行类别关联度计算和特征词提取,得到每个降雨灾情检测信息文档的特征向量。

1.3 降雨灾情文档分类模型建模

(1)文档类别特征提取。该文将降雨灾情检测信息文档 m 表示为 $m=\{t_1,t_2,\cdots,t_m\}$,t_i 表示文档中的特征词,利用 CHI 计算特征词与类别之间的关联度,提取出关联程度高的特征词作为文档的特征向量[18]。特征权重则利用 TF-IDF 计算,其公式为:

$$\mathrm{TF-IDF}(t_k,d_j)=\mathrm{TF}_{kj}\times\log\left(\frac{N}{n_k}\right) \quad (1)$$

式中,TF – IDF(t_k,d_j) 为特征词 t_k 在文档 d_j 中的特征权重值,TF_{kj} 为特征词 t_k 在文档 d_j 中的词频,N 为文档总数,n_k 为包含 t_k 的文档数。

(2)降雨灾情信息 SVM 分类器构建。该文选择 SVM 算法实现降雨灾情文档信息分类,并以 LIBSVM 软件包构建二分类模型,即判断文档信息若属于雨情相关,则为雨情信息文档,反之,则为其他信息文档。SVM 通过寻找最优超平面,并使该平面在分割各类数据时,让各类数据离超平面的间隔达到最大,以确保其泛化能力达到最好,因此具有很好的分类效果[19-20]。其学习策略是训练一个超平面,并在处理高维度问题时引入核函数,以适用于文档分类。其最优超平面的求解如下:

$$\min_{w,b,\zeta} \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{l} \zeta_i \tag{2}$$

$$\text{Subject to } y_i[(w^{\mathrm{T}}\varphi(x_i) + b)] \geqslant 1 - \zeta_i, \zeta_i \geqslant 0 \quad (i = 1,2,\cdots,l) \tag{3}$$

式中,l 是数据样本总数,$x_i, i = 1,2,\cdots,l$ 是数据,y 是类别且 $y_i \in (1,-1)$。$\varphi(x_i)$ 把 x_i 映射到高维空间,同时引入了惩罚因子 C 和松弛变量 ζ,则有决策函数如下:

$$f(x) = \mathrm{sgn}[\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b] \tag{4}$$

式中,K 为映射到高维空间的核函数。

同时以精确率 P(Precision)、召回率 R(Recall)和 F1 值作为模型评价指标。其中精确率 P 评价的是某个类别的分类是否有更多的正确数,召回率 R 评价的是某个类别的数据是否多数被正确分类。对于二类的分类问题,其分类结果如表 1 所示。

表 1　文本二分类结果

	正类样本	负类样本
分类判断为正	TP	FP
分类判断为负	FN	TN

则:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

F1 值是精确率 P 和召回率 R 的调和平均值,综合反映整体指标。

$$\mathrm{F1} = \frac{2PR}{P + R} \tag{7}$$

利用建模集对核函数类型 t、惩罚因子 C、核函数参数 g 取值进行调试。其中三种核函数 t 比较结果如图 3 所示:可见线性核函数分类效果最优,其 F1 值明显高于多项式核函数、RBF 核函数。而多项式核函数的高召回率是因为该模式判定其中一类为无,所以不选用。另外对于 SVM 来说,通过调整惩罚因子 C,可以更有效地解决其在分类中数据集的“偏斜”问题,“偏斜”是指参与分类的两个类别或多个类别的样本数据量差异很大。采用了交叉验证法,得出最优的 C 和 g 参数组合为(8,0.5)。优化后召回率和 F1 值分别提高了 2.08%、0.37%,表明类别内的分类正确性有提高,这在一定程度上提高了模型应对数据集“偏斜”的准确性和适应性。

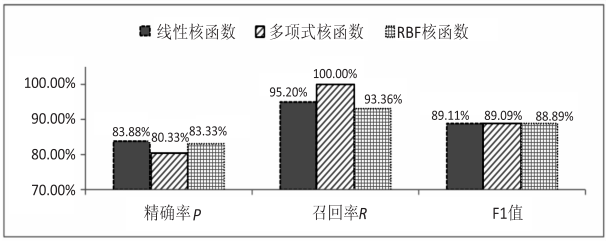

图 3　三种核函数比较

1.4　灾情权重分级和热点分析

(1)灾情权重分级。基于索引关键词的权重值,计算每个雨情信息文档的权重等级。考虑社交媒体内容常带有较浓的人的主观感受和情感状态,当人在感受到更危急情况时,会倾向于用更严重的描述,或者描述的词汇量更多。而这些往往表示该地点的雨情或灾情更重,因此根据这点把索引关键词赋予不同的权重值,如把“雨”设置为 1 的权重值,把具有明显灾情指示性的词语(如大雨、落狗屎、水浸、大水、积水、淹、涝、涨水)设置为 2 的权重值,把显示严重和紧急灾情的词语(如暴雨、冰雹、洪水、水灾)设置为 3 的权重值。具体如:当雨情信息文档出现“雨”为关键字,如“下雨了,很讨厌下雨啊”则该信息点的权重等级为 1;当出现“大雨、暴雨”为关键字,如“上午落大雨,晚上大暴雨”则该信息点的权重等级为 5。

(2)灾情热点分析。分析每个目标雨情信息点及其邻近位置中的每个雨情信息点的灾情权重等级。要成为具有显著意义的灾情信息热点,该雨情信息点的权重等级应具有高值,且被其他同样具有高权重等级值的雨情信息点所包围。

2　系统设计与实现

2.1　总体设计

该系统采用 B/S 的框架结构:包括应用管理层、风险告警层、数据分类层、数据过滤层、数据采集层五个层级。后台应用程序将文档信息数据通过采集、过滤、分类、插值等加工处理后存入数据库中,前台 WEB 平台运用 GIS 地图技术展示处理后的降雨灾情信息。线上降雨灾情检测系统的总体框架如图 4 所示。

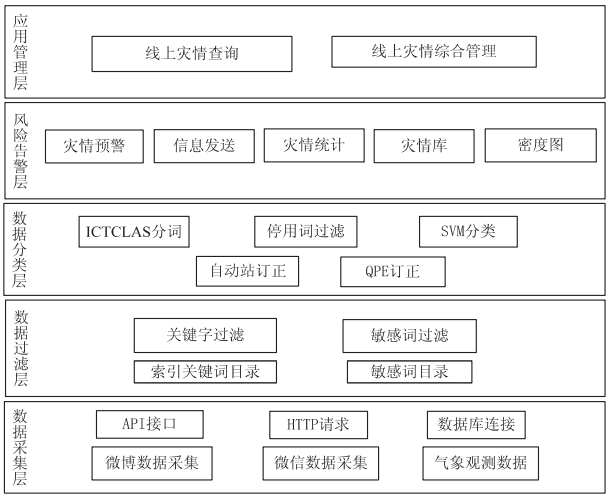

图 4 系统总体框架

2.2 系统功能

五个层级的具体功能如下：

(1)应用管理层：提供线上灾情查询和综合管理功能。其中灾情查询用于实时展示降雨灾情情况，将以灾情检测流程处理后的信息以 GIS 方式直观地展示出现降雨灾情的时间、地点、热点分析等。系统管理员可通过灾情综合管理对系统和模型的相关参数进行查询和操作，包括：微博数据管理、微信数据管理、分词专用词库、索引关键词目录、敏感词目录、索引关键词权重值等的查询和配置。

(2)风险告警层：提供灾情预警、告警信息发送、灾情统计、灾情库、灾情密度图等功能。灾情预警主要是根据城市网格化管理的要求生成各个网格的告警信息。告警信息发送用于对网格管理员、值班人员的信息发送，支持微信、短信等渠道发送。灾情统计可通过区域纬度统计当前生效的所有降雨灾情数据。灾情库将某个时段的降雨灾情数据进行汇总形成灾情事件库，并支持外部附件和数据的上传导入。灾情密度图可查看和统计某个时段内的降雨灾情空间密度分布情况。

(3)数据分类层：提供分词、停用词过滤、模型分类功能。分词主要利用 ICTCLAS 和专用分词词库将文档信息进行分词操作。停用词过滤是剔除停用词表中的用词。模型分类是通过 SVM 最优模型对采集到降雨灾情检测基础信息库的文档信息进行分类处理。

(4)数据过滤层：提供关键词过滤、敏感词过滤。关键词过滤根据索引关键词目录将与降雨信息或灾情状况无关的文档信息进行剔除过滤。敏感词过滤主要针对社交媒体数据来源于用户的主动上传，没有进行严格的审核和排除，内容可能存在一些敏感的风险词汇，可根据用户设置的敏感词目录将包含了当中关键字的数据进行剔除。

(5)数据采集层：通过新浪微博开放 API 接口获取实时位置微博数据。通过 HTTP 请求的方式，从微信服务号自建开发的 API 接口定时获取降雨灾情上报数据。通过数据库连接的方式定时获取相关的气象数据。

3 关键技术

3.1 多线程采集

微博数据采集通过 API 接口实现对某个位置周边动态数据的获取。其流程如图 5 所示。先获取经纬度点数据集，然后启动线程池，过程中针对多个经纬度点，需采用多线程技术，对每个点启动一个线程单独去执行，再通过微博应用授权码 accessToken 以获取数据。

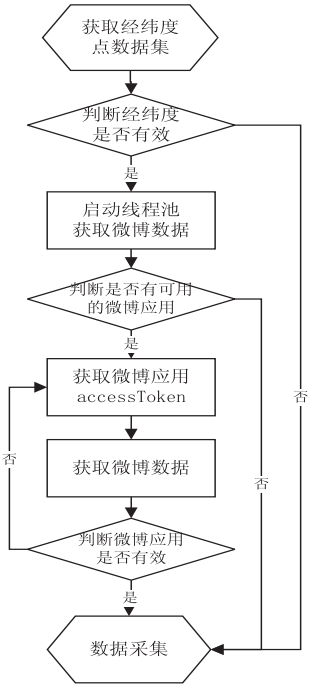

图 5 微博数据采集流程

3.2 数据过滤入库

数据过滤均基于可动态更新的词目录。在信息数据入库时，将预处理后的数据遍历关键词和敏感词目录进行过滤，然后通过逆地址编码接口获取该条数据的地理位置信息，最后入库。

3.3 网格管理

通过灾情综合管理功能模块配置城市网格管理信息和网格告警的权重等级，包括行政区域、街道、网格名称、网格员姓名、网格员联系方式、网格范围(经纬度坐标)等信息，同时支持行政区域、街道或者单个网格的告警阈值设置。当社交媒体数据经过采集、过滤、分类等流程处理后，若信息点的灾情权重等级达到网格告警阈值时，则自动生成告警信息，并可选择向指定人员发送告警信息，同时将告警动态展示在线上灾情查询模块上。

3.3.1　利用 QPE 及自动站数据订正

每 6 分钟定时获取最新时次的天气雷达定量降水估测(QPE)产品资料,通过 IDW 库函数将格点形成的 QPE 产品采用反距离加权方法插值到相应的雨情信息点上。若 QPE 产品插值后得到的雨情信息点降水量未达到设定的阈值,则认定该雨情信息点为误报,取消其告警。

每 5 分钟获取最新时次的自动站分钟数据,并实时计算各站点 5 分钟滑动雨量和 1 小时滑动雨量数据,通过 NEAR 函数将自动站数据采用距离最近法关联到相应的雨情信息点上。若自动站附近无雨情信息点,则订正结束;若有,则判断该自动站 5 分钟滑动雨量和 1 小时滑动雨量是否达到设定的阈值,如都未达到则认定该雨情信息点为误报,取消其告警。

3.2.2　灾情密度图

通过灾情密度图,可查看任意时间段内的降雨灾情数据在空间范围上的密度分布情况。首先根据设定时间范围,搜索出该时间范围内的分类处理后为雨情信息点的数据。然后再计算出每个雨情信息点的权重等级,其中权重值选取的是该点半径 2.5 公里(R)范围的所有雨情信息点权重等级(W)作为该点的密度权重,计算公式为:

$$\sum_{i=0}^{n} w_i/(\text{Math.PI} * R * R) \tag{8}$$

其中搜索半径 2.5 公里范围的雨情信息点时,需要计算两个经纬度点之间的球面距离,其算法实现如下:

```
Double:distance (lat1, lng1, lat2, lng2) {
    radLat1 = lat1 * Math.PI/180.0
    radLat2 = lat2 * Math.PI/180.0
    a = radLat1 - radLat2
    b = lng1 * Math.PI/180.0 - lng2 * Math.PI/180.0
    s = 2 * Math.asin(Math.sqrt(Math.pow(Math.sin(a/2),2) + Math.cos(radLat1) * Math.cos(radLat2) * Math.pow(Math.sin(b/2),2)))
    s = s * EARTH_RADIUS
    s = Math.round(s * 1000)
Return s
End distance
```

当完成每个雨情信息点的权重值计算后,通过热力图函数将数据插值到地图上。

4　应用案例

利用 2017 年 5 月 7 日广州多地突发特大暴雨过程进行应用案例的事后分析。系统首先通过正则表达式的方式将数据的 URL 地址("http://t.cn(.*?)")、@人(@(.*?))、话题(#(.*?)#)、特殊图片(\\? \\? \\? \\? +)等特殊符号剔除并进行关键词和敏感词过滤,再调用高德 API 接口 geocoder 获取对应的地理位置信息后将数据入库。在当日 0 时至 19 时降雨灾情检测基础信息库共采集到 1 159 条文档数据,进行人工标注后获得 617 条雨情信息。在雨情信息当中有 124 条含水浸、涨水、积水等相关资讯,证明在降雨灾情提取上,社交媒体数据能提供如水浸街道、易积水点等灾情实况,协助快速发现降雨造成的次生灾害,对灾情收集具有参考和实用价值。系统通过专用分词词库和 Library. Instance. NLPIR_GetKeyWords 函数提取关键字,并利用预存模型特征向量计算特征权重,通过 svm_scale. main({"-l","0","-u","1","-p",scalePath,featurePath}) 函数归一化处理后,再调用 svm_predict 函数分类,经气象数据订正后的自动分类结果与人工标注和灾情实况对比分析得出:热点分析显示(图 6)共有 32 个热点,其分布地点与广州市三防办通报的水浸受灾区域吻合(花都、增城、黄埔出现严重水浸,白云、天河、番禺多路段出现严重积水,多处交通中断受阻)。系统灾情检测的精确率为 79.6%,召回率为 96.82%,F1 值为 87.37%,证明该系统在重大灾害性天气中能实现雨情、灾情数据的自动识别和提取,提高社交媒体数据的利用效率,能一定程度上反映出广州地区雨情和灾情的真实状态。

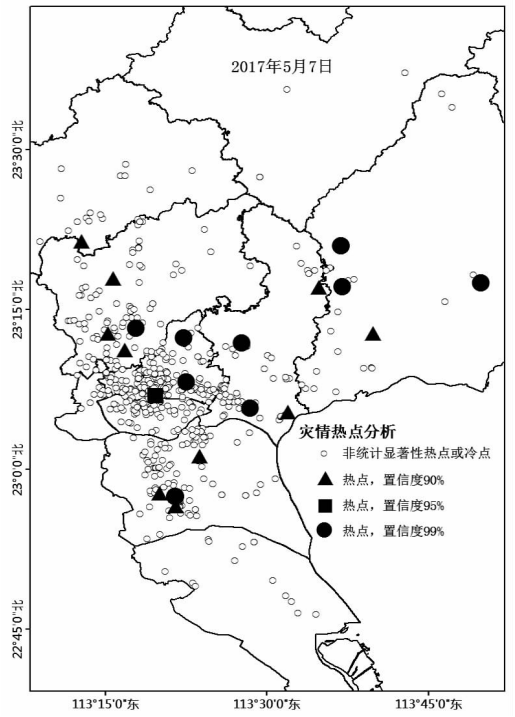

图 6　灾情热点分布

5　结束语

目前传统的气象灾情收集大多为灾情实地调查,在突发状况或恶劣天气下非常不利于及时了解实时灾情状况。而线上降雨灾情检测系统根据气象灾情收集

业务的需要,利用社交媒体大数据检测降雨雨情、灾情的发生发展状态,结合了社交媒体数据优势,同时以 SVM 机器学习算法解决大量数据处理效率低下的问题。在提取过程中,通过灾情热点分析结合利用气象雷达、自动站观测数据进一步提高灾情提取的准确度。因为社交媒体数据具有低成本、大数据、即时性的优势,所以结合 SVM 算法模型建成的线上降雨灾情检测系统,能为降雨灾情的收集提供有效的实时采集工具。

参考文献:

[1] LU X,BRELSFORD C. Network structure and community evolution on Twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami[J]. Scientific Reports,2014,4:6773.

[2] ZOOK M, GRAHAM M, SHELTON T, et al. Volunteered geographic information and crowdsourcing disaster relief: a case study of the haitian earthquake[J]. World Medical & Health Policy,2010,2(2):7-33.

[3] LEIDNER D E,PAN G,PAN S L. The role of IT in crisis response:lessons from the SARS and Asian Tsunami disasters[J]. Journal of Strategic Information Systems,2009,18(2):80-99.

[4] 邬柯杰,吴吉东,叶梦琪. 社交媒体数据在自然灾害应急管理中的应用研究综述[J]. 地理科学进展,2020,39(8):1412-1422.

[5] 苏晓慧,邹再超,苏 伟,等. 面向地震应急的自媒体信息挖掘模型[J]. 地震地质,2019,41(3):759-773.

[6] 邬群勇,裘钰娇. 微博数据位置信息反映台风灾情的有效性分析[J]. 测绘科学技术学报,2019,36(4):406-411.

[7] 吴先华,肖 杨,王国复,等. 基于微博大数据的城市内涝灾害的灾情及公众情绪研究—以南京市为例[J]. 灾害学,2018,33(3):117-122.

[8] 黎洁仪,梁之彦,范绍佳. 基于位置微博的大风信息提取及应用[J]. 气象科技,2019,47(5):879-884.

[9] 彭云建,欧善国,梁 进. 在线气象科普知识竞赛试题的自动组卷方法[J]. 计算机技术与发展,2021,31(5):209-214.

[10] 骆 聪,王 帅. 结合深度学习与词性标注的网页分类算法研究[J]. 计算机技术与发展,2018,28(8):71-74.

[11] 王新美,丁爱玲,雷梦宁,等. 基于 CNN 和 SVM 融合的交通标志识别[J]. 计算机技术与发展,2020,30(6):7-12.

[12] 姬晓飞,石宇辰. 多分类器融合的光学遥感图像目标识别算法[J]. 计算机技术与发展,2019,29(11):52-56.

[13] 王莉莉,杨鸿武,宋志蒙. 基于多分类器的藏文文本分类方法[J]. 南京邮电大学学报:自然科学版,2020,40(1):102-110.

[14] 段文影,饶 泓,段隆振,等. 基于 IA 参数寻优组合核的 SVM 文本分类研究[J]. 南昌大学学报:理科版,2018,42(3):289-292.

[15] 徐沛娟,李雄飞,惠 玥,等. 中文文本分类相关算法的研究与实现[J]. 吉林大学学报:理学版,2009,47(4):790-794.

[16] 马建斌,李 滢,滕桂法,等. KNN 和 SVM 算法在中文文本自动分类技术上的比较研究[J]. 河北农业大学学报,2008,31(3):120-123.

[17] VIII L,INTELLIGENZ K U,JOACHIMS T. Text categorization with support vector machines[J]. Fakultäten, 1999, LS8:137-142.

[18] 余伟中. 基于 VSM 的中文文本分类算法研究[D]. 南京:南京邮电大学,2018.

[19] 高亚波. 文本分类系统的设计与实现[D]. 北京:北京交通大学,2008.

[20] 陈蔼祥. 用于字符和数字识别的若干分类方法的比较研究:实验结果[J]. 计算机科学,2015,42(S1):102-106.