DOI:10.3880/j.issn.1003-9511.2021.02.015

基于社交媒体的暴雨灾情信息实时挖掘与分析
——以 2019 年“4·11 深圳暴雨”为例

黄 晶[1],李梦晗[1],康晋乐[1],曹 阳[2],曾庆彬[3]

(1.河海大学商学院,江苏 南京 211100;2.丹阳市建设工程质量监督站,江苏 丹阳 212300;
3.深圳市水务规划设计院股份有限公司,广东 深圳 528200)

摘要:在暴雨天气过程中,实时获取详细的暴雨灾情信息,尤其是灾害带来的影响和后果,对灾情的及时响应和评估有着重要的意义。结合隐含狄利克雷分布(LDA)主题挖掘方法和支持向量机(SVM)分类算法,构建基于社交媒体的暴雨灾情信息挖掘模型,收集 2019 年“4·11 深圳暴雨”的微博数据,识别出 4 类微博文本主题,并进一步对“灾情信息”这一主题进行二次挖掘。结果显示,所构建的暴雨灾情信息挖掘模型能够准确识别出微博文本中蕴含的灾情信息,提取了深圳暴雨事件带来的交通影响、人员伤亡、建筑物倒塌、积水、停电、停水 6 种灾情信息;暴雨灾情信息挖掘和分析的结果能准确反映灾害的发展状况,在时间上,挖掘出的灾情信息实时反映了灾害的发展过程,在空间上,微博数量集中的地区与灾情发生地点保持一致,基于社交媒体的灾情信息实时获取方法能够为政府开展应急救援及救灾部署工作提供实时基础信息支持。

关键词:暴雨灾情信息;社交媒体;信息挖掘;文本分类
中图分类号:X43 文献标志码:A 文章编号:1003-9511(2021)02-0086-09

随着气候变化与人类活动的双重影响,近年来我国自然灾害的发生频率和强度有所增大[1]。应急管理部发布的 2019 年全国自然灾害基本情况显示,2019 年我国自然灾害以暴雨、台风、干旱、地震、地质灾害为主,共造成 1.3 亿人次受灾,直接经济损失达 3 270.9 亿元。暴雨是我国主要的气象灾害之一,在其发生、发展过程中,往往诱发一系列的次生、衍生灾害,有时比原生灾害的危害更大,不仅造成巨大的经济损失,也对人民群众生产生活带来不利影响。如暴雨灾害一般会引发积水内涝、雷击、电杆倒折、墙体倒塌等,进一步会导致人员伤亡,严重的城市交通瘫痪、大面积停水停电、房屋损毁等后果,直接影响了城市运行、居民生活和生产,造成了巨大的损失。尽管政府及有关部门已制定相关预案,实现灾前的紧急预警与灾害的损失评估,但却无法对灾情信息尤其是灾害带来的影响和后果进行即时监测与分析。因此,实时获取灾情详细信息尤其是灾害带来的影响和后果,对灾情的评判和应急救灾措施的部署有着重要的意义。随着互联网的普及,社交媒体作为新兴的灾害数据源已得到广泛应用,通过社交媒体反映暴雨灾情及其发展特征已经成为指导应急响应、救灾部署行之有效的手段之一[2]。

由于社交媒体中 80% 以上的信息是文本信息,大多数研究使用文本挖掘技术对灾害事件相关文本进行采集与识别,具体过程分为文本预处理、文本特征提取、分类模型构建等。文本的预处理过程要求对文本进行分词,常用的分词工具有 Jieba、ICTCLAS、SnowNLP 等,其中 ICTCLAS 是张华平等[3]研发的,2009 年更名为 NLPIR,是目前广泛使用的汉语分词开源系统。文本特征提取包括特征词选择与特征词量化,常用特征词选择的方法有卡方检验(CHI 统计)、信息增益(IG)、互信息等,常见特征词量化的方法有词频-逆文档(TF-IDF)、信息熵等,其中 TF-IDF 法因其简单快速且结果较符合实际情况而得到广泛应用。姚春华等[4]采用 TF-IDF 法计算不同类别和长度文档的特征权重,提出一种网络文本信息情感分类的方法;梁春阳等[5]采用卡方检验对台风“莫兰蒂”相关微博文本的词汇进行特征词

基金项目:国家自然科学基金重大研究计划重点项目(91846203);国家自然科学基金青年项目(71601070)
作者简介:黄晶(1986—),女,讲师,博士,主要从事管理科学与系统工程、水灾害风险管理与应急管理研究。E-mail:j_huang@hhu.edu.cn

选择,再使用 TF-IDF 法进行特征词量化。对于文本分类模型的构建,现有研究大多结合主题挖掘方法与分类算法。在主题挖掘方法中,隐含迪利克雷分布(LDA)主题挖掘方法已经被证明是一种非常有效的方式,Chen 等[6]采用 LDA 主题挖掘方法对 Twitter 文本中隐含的主题进行挖掘,发现了其与流感发病状态的关联;张连峰等[7]应用 LDA 主题挖掘方法进行建模,挖掘出微博舆情中的关键节点。常见的分类算法有决策树、支持向量机、朴素贝叶斯、随机森林、逻辑回归等。由于支持向量机算法具有全局最优、结构简单、推广能力强等优点,近年来被广泛使用[8]。白华等[9]采用支持向量机(SVM)作为文本分类算法,识别和分类与地震相关的社交媒体数据,开发了高效的灾害事件即时监测系统;夏梦南等[10]使用支持向量机作为分类算法,对微博中的情感进行分类。

国内外相关研究证明,由于其自身的时间和地理空间属性,社交媒体数据能够应用于灾害事件的实时监测和趋势预测[11-12]。基于社交媒体数据进行灾情信息的挖掘时,通常对灾害相关的文本数据进行主题分类。如梁春阳等[5]结合 LDA 与 SVM 构建主题分类模型,将台风“莫兰蒂”相关微博文本分成“预警信息”“灾情信息”“无关信息”与“救援信息”4 类主题;王艳东等[13]结合 LDA 与 SVM 构建应急主题分类模型,将北京与暴雨相关的微博数据分为“交通状况”“天气预报”“灾情信息”“损失与影响”“救援信息”“内涝原因”6 类主题。虽然上述研究对灾情相关的关键词进行了提炼和划分,识别了隐藏在微博文本数据中的灾害相关主题,但是并不能反映详细的灾情信息,尤其是灾害导致的次生、衍生灾害,无法为政府应对突发灾害事件时的应急响应和救援部署提供信息支撑。

为充分挖掘微博中蕴含的灾情详细信息,实现对暴雨灾害事件带来的后果和影响的实时监测,本文以 2019 年的“4·11 深圳暴雨”为例,结合 LDA 主题挖掘模型和 SVM 分类算法,构建暴雨灾情信息挖掘模型,将微博文本分为 4 个主题并对“灾情信息”这一主题进行二次分类。通过将时空分布结果与现实灾情对比分析,验证该模型的可靠性,从而为暴雨灾害的应急管理与救灾部署提供决策支持。

1 数据来源与方法

1.1 研究区域及灾害事件选择

研究区域为位于广东省中南沿海地区的深圳市,该市属亚热带海洋性气候,降水情况各地区差异很大,容易出现局地性的洪涝灾害和短时雷雨大风天气。深圳市是中国经济特区,下辖福田区、罗湖区、盐田区、南山区、宝安区、龙岗区、龙华区、坪山区、光明区 9 个行政区和大鹏新区 1 个新区,其中福田区和罗湖区是深圳市商业中心,为最繁华区域。2019 年深圳市 GDP 总量 26 927.09 亿元;截至 2018 年末,常住人口 1 302.66 万人,常住人口城镇化率全国排名第一。此外,深圳市作为中国“最互联网”城市,互联网普及率位居全国第一,截至 2017 年末,网民渗透率高达 87.1%。

2019 年 4 月 11 日晚,受冷暖气流交汇影响,深圳市出现冰雹、大风、雷暴和强降雨等强对流天气,全市平均雨量 40.6 mm,其中罗湖区平均雨量 65.0 mm,10 分钟最大雨量达 39.2 mm。短时极端强降水导致深圳市多个区域突发洪水,罗湖区、福田区等区域受灾严重,造成 11 人死亡。暴雨天气一直持续,18—20 日再次迎来一轮强降雨,造成南山区一简易住房土墙坍塌,2 人被困。除了人员伤亡外,此次持续的暴雨洪涝灾害还造成部分地区建筑物倒塌、道路积水等灾情,引起了人们的广泛关注。

1.2 数据的获取和预处理

基于目前国内最受欢迎的社交媒体平台——新浪微博,利用网络爬虫,收集了从 2019 年 4 月 11 日 0 时到 4 月 23 日 0 时,以“深圳暴雨”为关键词的微博文本数据,去除其中非常短(少于 4 个字)和重复的微博文本,最终留下 10 015 条微博数据,其中有 1 321 条带有位置信息,有 1 044 条定位在深圳市。

1.3 模型与方法

结合 LDA 主题挖掘方法和 SVM 分类算法,构建暴雨灾情信息挖掘模型。首先,利用 LDA 主题挖掘方法,通过计算文本集合的离散词语共现频率找出隐藏在文本集合中的相关主题,同时输出对应主题的词汇分布。基于已发现的主题,利用 SVM 分类算法训练已有的文档样本(主题和主题的词汇分布)。当有新的微博文本进入时,通过该模型的判断确定该微博文本的主题类别,从而实现微博文本的实时分类。然后,利用相同的处理方式对“灾情信息”这一文本类别进行二次分类。最后,将分类结果可视化。相关步骤如图 1 所示。

1.3.1 中文分词

使用 NLPIR 大数据搜索与挖掘实验室研发的 NLPIR-ICTCLAS 汉语分词系统(http://www.nlpir.org)对微博数据进行分词,结合暴雨灾害领域知识,对暴雨灾害特征词汇进行补充,形成适用于暴雨灾害的分词词典,如“特大暴雨”“雷阵雨”“暴风雨”“大雨”“中雨”“暴雨红色预警”“暴雨橙色预警”等。去除广告、部分标点符号和介词等无关词汇,去

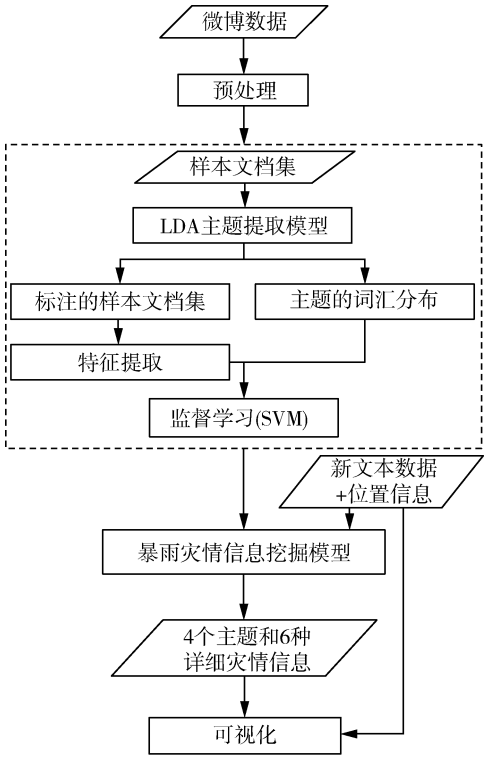


图1　基于社交媒体的暴雨灾情信息挖掘流程

除常见且缺乏价值信息的停用词,实现对文本数据的有效处理。

1.3.2　隐含主题的挖掘

LDA 主题挖掘方法由 3 层贝叶斯模型构成,文本层、主题层、单词层。基本思想是每条文本由多个主题以多项式分布构成,每个主题又由多个单词以多项式分布构成,而多项式分布的先验概率分布为狄利克雷分布,依此法可以生成一个包含多个文本的数据集。由于它是一种无监督学习算法,不需要事先对文档进行标记,输入的参数仅包括用户构建的语料库以及用户设定的主题个数,输出结果为文档-主题矩阵,即文档在各个主题上的概率分布,和主题-词矩阵,即主题在各个词汇上的概率分布,最后人工对分出的主题类别进行归纳与相似主题的合并。本文基于 Java 编程语言,利用 Phan 学者发布的开源 Gibbs 采样代码(https://github.com/hankcs/LDA4j),实现基于 LDA 主题挖掘方法的主题挖掘。

1.3.3　特征词的抽取

采用卡方统计量对词汇进行特征选择,根据词汇卡方值筛选每个文本类别的特征词汇:

$$x^2(t,c) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{1}$$

式中:N 为训练样本集文档总数;A 为一个类别中,包含某个词的文档数量;B 为在一个类别中,排除该类别,其他类别包含某个词的文档数量;C 为在一个类别中,不包含某个词的文档数量;D 为在一个类别中,不包含某个词也不在该类别中的文档数量。

1.3.4　特征矩阵的构建

因卡方检验在文本特征选择时会产生过分夸大低频词的现象,为此,在选择特征词汇后,使用词频-逆文档算法对特征词汇进行特征量化:

$$w_k^i = tf_{ki} \cdot \lg \frac{N}{N_{ki} + 1} \tag{2}$$

式中:w_k^i 为单词词汇的权重;tf_{ki} 为第 i 个文档单词 k 的词频;N_{ki} 为包含第 i 个文档单词 k 的文档个数;N 为语料库中文档的总数。

1.3.5　分类器选择

基于标注好的文档样本集,采用 SVM 方法进行训练。为了验证模型分类的精准度,将选取的训练样本划分为 M 部分,其中 $M-1$ 部分作为模型的训练样本,剩下的一部分作为模型参数确定的检验样本,利用检验样本来验证 $M-1$ 部分数据分类结果的精度。

准确率是分类问题中最简单直观的评价指标,是分类正确的样本占总样本个数的比例:

$$R = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{3}$$

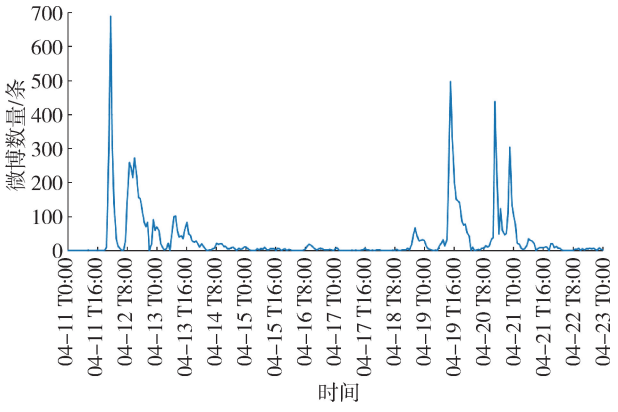
式中:n_{correct} 为被正确分类的样本个数;n_{total} 为总样本个数。

2　结果与分析

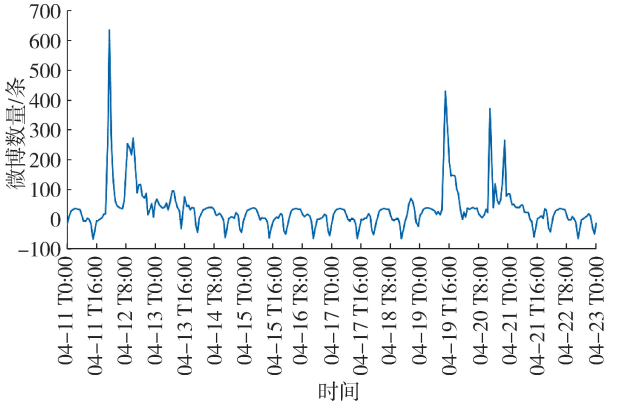
2.1　总体时空分布

根据采集到的“深圳暴雨”微博数据,分析其随时间变化的趋势(图 2(a))。从微博的总体趋势图来看,数据集中在暴雨暴发后的 3 天内,然后开始慢慢平息,淡出社交媒体的热门话题。图 2(a)中每天的凌晨都会出现最低点,然后数据开始上升,呈现循环波动。因此,假设此时间序列有多期的按天循环波动趋势,利用季节性趋势分解来进一步分析灾害事件情况。季节性趋势分解后的微博文本时间序列如图 2(b)所示。可见 4 月 11—12 日微博数量波动幅度最大,在 11 日的 23 时达到峰值,并且维持在较高的水平。23 时 15 分,全市取消暴雨黄色、大风黄色预警,微博数量有减少趋势,但此后的 4 月 12、13 日“深圳暴雨”依然是比较热门的话题。4 月 18 日下午微博数量再次上升,在 4 月 19—20 日有较大幅度的波动,并在 19 日 14 时达到最大值,之后有减少趋势。但 4 月 20—21 日微博数量再次出现不同寻常的波动。

此次深圳暴雨事件引起了全国各地的广泛关注,但主要集中在深圳市及其周边城市。为给城市



(a) 时间序列的总体趋势



(b) 时间序列的季节性调整序列

图2 “4·11 深圳暴雨”时间趋势季节性分解

暴雨的灾情管理提供决策支持,着重分析深圳市内各区域的微博数量分布情况。图3 为深圳暴雨微博数量的分区空间分布,可见微博数量主要集中在深圳市西部,其中南山区最多,其次是宝安区、福田区。据深圳市气象局报道,在此次暴雨灾害事件中,南山区、宝安区均遭遇了严重的强降雨和冰雹侵袭。

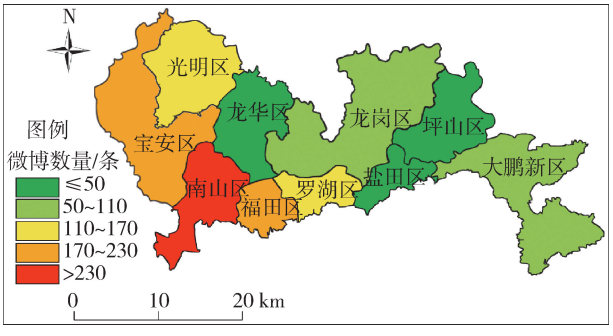


图3 “4·11 深圳暴雨”微博数据区域空间分布

2.2 主题分类结果

经过重复实验,确定初始主题的最佳数量为25。但因该主题模型是一种无监督学习算法,故人工对25 个主题进行归纳与相似主题的合并,最终确定4 个主题,分别为“天气状况”“灾情信息”“救援信息”和“无关信息”。对“灾情信息”这一主题实施二级分类,最终确定了“交通影响”“人员伤亡”“建筑物倒塌”“积水”“停电”和“停水”6 种灾害影响类型。经验证,本文所构建的暴雨灾情信息挖掘模型初次分类的准确率为88.0%,二次分类的准确率为82.7%。

对4 个微博主题分别统计其微博条数及所占比例(表1)发现,“灾情信息”占“深圳暴雨”相关微博总量的68.5%,是公众关注的热点问题。其次是“天气状况”和“救援信息”,分别占19.3%和7.9%。

表1 微博不同主题分类统计

主题	微博数量/条	占比/%
天气状况	1 932	19.3
灾情信息	6 639	66.3
救援信息	791	7.9
无关信息	653	6.5

将“灾情信息”这一主题进行二次分类,6 类具体灾情信息及其微博条数统计结果如表2 所示。“人员伤亡”“积水”和“交通影响”是深圳暴雨期间最受关注的灾情,微博数量分别占44%,22.3% 和15.6%。其他具体灾情信息的微博数量占比均不足10%。

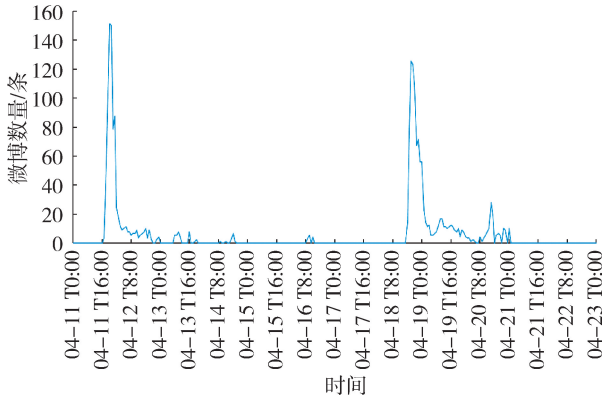
表2 6 种具体暴雨灾情信息分类统计

具体灾情信息	微博数量/条	占比/%
交通影响	1 033	15.6
人员伤亡	2 925	44.0
建筑物倒塌	624	9.4
积水	1 478	22.3
停电	478	7.2
停水	101	1.5

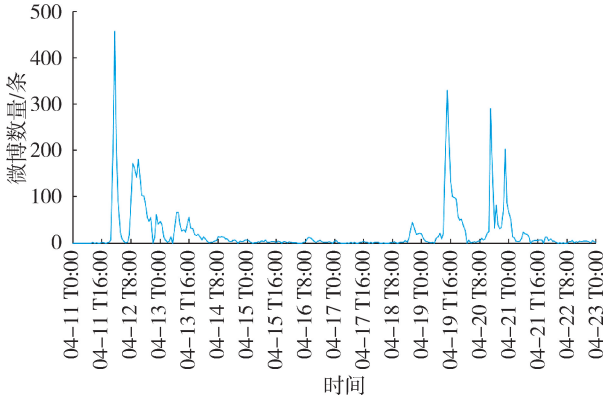
2.2.1 时间趋势

除“无关信息”外,其他3 类主题微博数量随时间的变化趋势如图4 所示。“天气状况”微博集中分布在两次强降雨发生前,即4 月11 日下午和4 月18 日晚上。“灾情信息”微博的时间趋势与总体时间趋势相似,主要集中在暴雨发生中以及暴雨发生后,并在4 月12 日12 时达到一个峰值,13 日之后微博数量逐渐减少。“救援信息”的微博数据主要集中在暴雨发生中以及暴雨结束后。

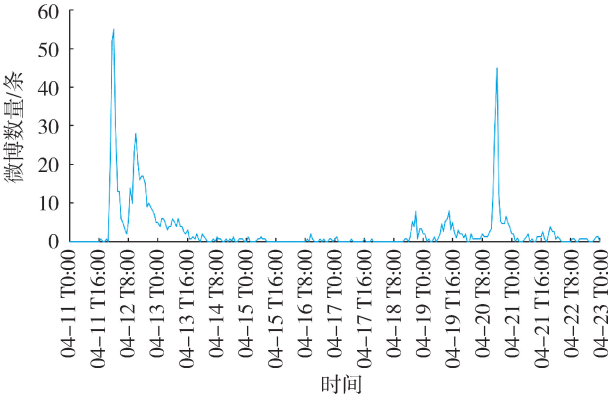
将“灾情信息”这一主题进行二次分类,分析“人员伤亡”“交通影响”“积水”“建筑物倒塌”“停电”“停水”6 种灾害影响信息的微博数量随时间的变化趋势(图5)。“人员伤亡”的微博数量存在两个峰值,分别在4 月11 日23 时和4 月20 日14 时,其余时间相对平稳;“交通影响”的微博集中分布在4 月12—15 日以及4 月19—22 日,持续的暴雨对交通状况造成很大影响;“建筑物倒塌”的微博数量在4 月20 日16 时达到峰值;“积水”微博数量的时间变化趋势具有短时间内急剧上升、持续时间较长的



(a) 天气状况



(b) 灾情信息



(c) 救援信息

图 4　不同微博主题时间趋势

特点,持续时间在 3 ~4 天。“停电”和“停水”时间趋势相似,但“停电”的微博数量普遍多于“停水”,二者主要集中在 4 月 11 日、19 日。同时发现,不同类型灾害影响信息的微博数量与总微博数量增幅基本呈正相关,其中“人员伤亡”“交通影响”以及“积水”随时间变化增幅较大,表明这 3 类受关注度较高。

2.2.2　空间分布

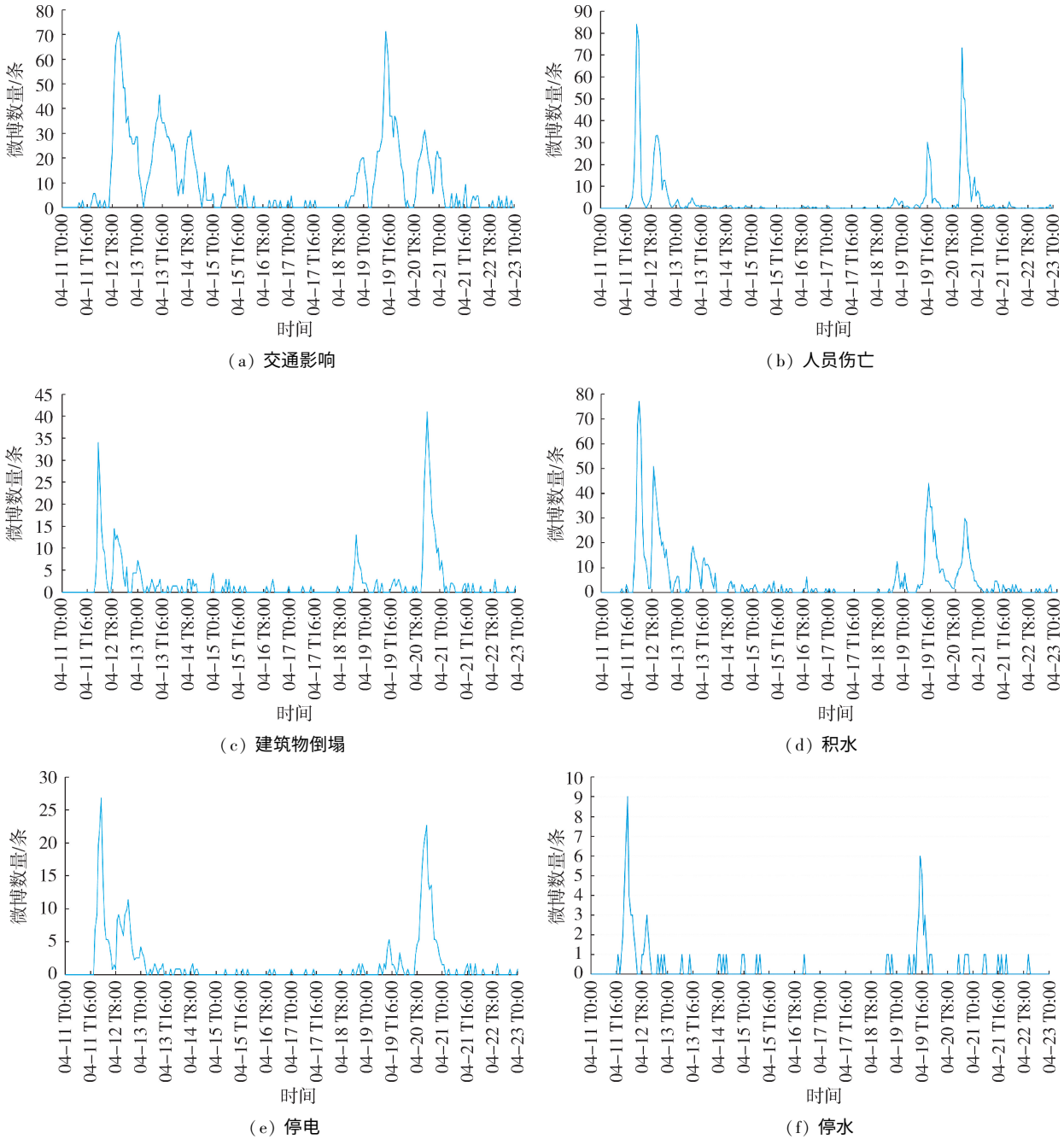
不同主题微博数量的空间分布情况如图 6 所示,3 个主题的空间分布情况各有不同。“天气状况”的微博数据集中分布在宝安区、光明区和龙华区;“灾情信息”的微博数量显著高于其他两个主题,且集中分布于深圳市西部的宝安区和南山区,以及位于深圳市中心的福田区和罗湖区;“救援信息”的微博数量相对较少,主要分布在福田区、南山区和罗湖区。

详细灾情信息的微博数据空间分布如图 7 所示。“交通影响”微博数据集中分布于宝安区,以及宝安区周边区域;“人员伤亡”和“建筑物倒塌”的微博数据空间分布情况类似,都集中在深圳市的西南部,包括南山区、福田区以及罗湖区。“积水”微博数据集中分布于降雨较严重的地区;“停电”微博数据分布较为广泛,其中在人口聚集区,即福田区和罗湖区分布较多;“停水”相关微博数量虽然较少,但仍可以看出数据集中分布在深圳市西北部。

3　讨　论

本文基于新浪微博,对“深圳暴雨”的微博数据进行了挖掘,结合 LDA 主题挖掘方法和 SVM 分类算法,构建暴雨灾情信息挖掘模型,将微博文本分为“天气状况”“灾情信息”“救援信息”和“无关信息”4 个主题,分类精度为 88%,并在此基础上有效识别了 6 种具体灾情信息,二次分类精度为 82.7%,与已有研究相比,本文提出的暴雨灾情信息挖掘模型对深圳暴雨的微博文本数据分类准确度较高。如王艳东等[13]构建了应急主题分类模型,将“北京暴雨”相关微博文本分为 6 个主题,总精度为 87.5%。韩雪华等[14]通过构建主题提取与分类模型,识别了 2018 年寿光洪水期间的公众情绪,一次分类和二次分类准确率分别为 89%、78%。

在一次分类的基础上,进一步对“灾情信息”这一主题进行二次分类,得出“交通状况”“人员伤亡”“建筑物倒塌”“积水”“停电”“停水”6 种灾情详细信息,可以更加全面、细致地反映暴雨灾害带来的次生、衍生灾害的实时发展状况。而现有大多数研究利用文本挖掘方法对微博文本进行主题分类,一般都是一次分类,如王艳东等[13]利用应急主题分类模型将“北京暴雨”微博数据分为“交通状况”“天气预报”“灾情信息”“损失与影响”“救援信息”“内涝原因”6 个应急主题;梁春阳等[5]将台风“莫兰蒂”相关微博分成“预警信息”“灾情信息”“无关信息”与“救援信息”4 个类别。也有研究对分类后的主题进行二次加工,韩雪华等[14]对 2018 年寿光洪水期间的公众情绪进行了二次分类;王艳东等[13]也曾采用聚类算法对“交通”“灾情”主题的微博数据进行了聚类,虽然反映了灾情的空间分布特征,但灾情类别单一,不能完全反映灾害带来的各种详细灾情和影响。



（a）交通影响　（b）人员伤亡

（c）建筑物倒塌　（d）积水

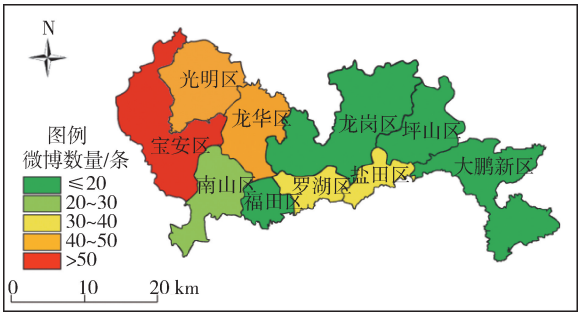
（e）停电　（f）停水

图 5　6 种灾情信息的时间趋势

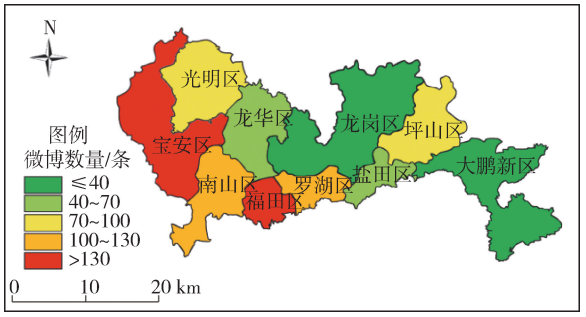
在时间上，挖掘出的灾情信息能够反映暴雨灾害的实时发展过程。深圳市气象局分别在 4 月 11 日晚和 4 月 19 日上午发布暴雨预警，因此微博数量在 11 日 23 时和 19 日 14 时达到两个峰值。“人员伤亡”灾情的微博数量分别在 11 日 23 时和 20 日 14 时达到峰值，4 月 11 日 22 时 10 分发生工人被冲走事件，1 小时内关于“人员伤亡”微博达到峰值；20 日 14 时两人被土墙埋困，救出后经抢救无效死亡，同一时刻再次引起了公众的关注。

在空间上，“深圳暴雨”微博的空间分布在一定程度上与灾害严重程度有关。微博数量较大的地区主要集中分布在深圳市西部南山区、宝安区和福田区。据深圳市气象局报道，在此次暴雨灾害事件中，南山区、宝安区均遭遇严重强降雨和冰雹侵袭。这是由于在受灾严重区域人们对灾害的关注程度更高，微博数量更多。但是福田区受灾并没有南山区、宝安区严重，其微博关注数量也相对较高。这主要是由于福田区是深圳市中心，人口密度大。因此，除了灾害的严重程度外，微博数量也与人口密度等其他因素相关。有研究发现社交媒体数据与人口密度、网络接入程度等因素具有高度相关性，可能会在一定程度上对微博数据的空间分布结果带来一些困扰[15]。

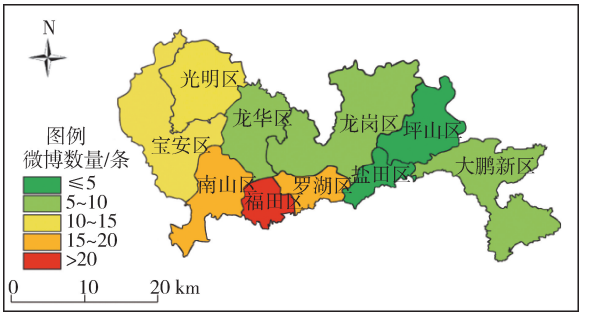
详细灾情信息微博数据的空间分布与该灾情严重的地区基本保持一致。深圳市气象局于 4 月 11 日晚发布预警，深圳市西北部将有严重暴雨来袭，因



（a）天气状况

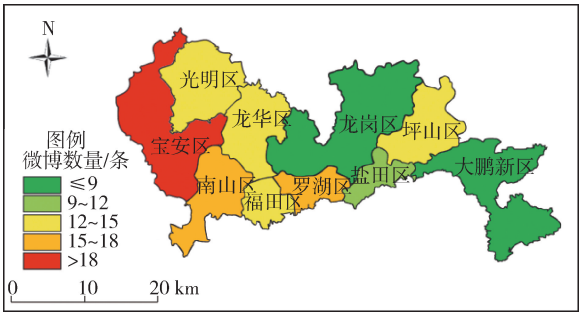


（b）灾情信息

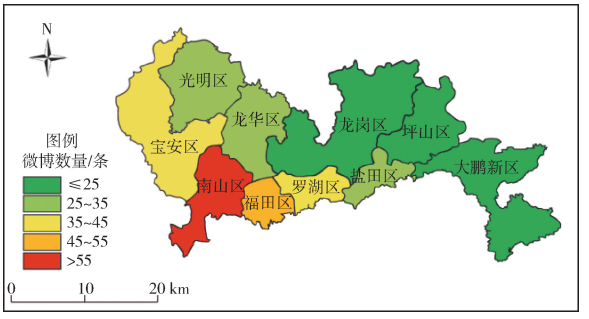


（c）救援信息

图 6　不同微博主题的空间分布



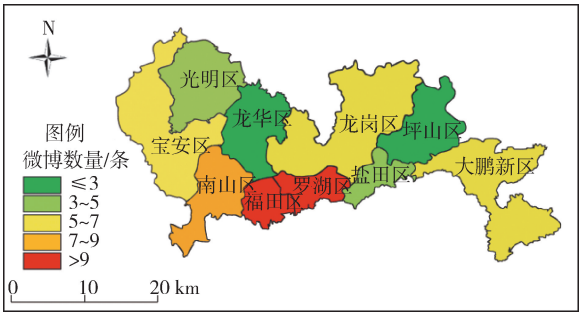
（a）交通影响



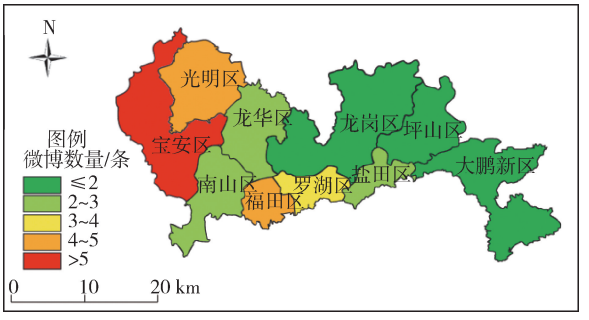
（b）人员伤亡

（c）建筑物倒塌

（d）积水



（e）停电



（f）停水

图 7　6 种灾情信息的空间分布

此,“天气状况”相关微博主要集中在深圳市西北部的宝安区、光明区和龙华区;针对暴雨带来的灾情状况,深圳市应急救援现场指挥部组织消防、应急等部门持续开展紧急搜救、伤员救治等工作,因此“救援信息”的微博主要集中在 11 日较严重的受灾区域,即福田区、罗湖区和南山区。“交通影响”的微博集中分布于宝安区,以及宝安区周边区域,这是因为宝安区是深圳市重要交通枢纽,且设有宝安国际机场;“人员伤亡”的微博主要集中于南山区、福田区和罗湖区,这 3 个区域受灾严重,均有人员伤亡事件发生。这些信息与官方发布的灾害评估报告基本一致。通过分析不同灾害信息的空间分布情况,可以获取灾情的空间分布特征和严重程度,为各区域的应急防范、救灾部署提供实时信息。

4 结 论

暴雨天气发生时,有效地获取社交媒体文本数据中包含的详细灾情信息,能够为暴雨灾情应急管理与救灾部署提供信息支持。本文结合隐含狄利克雷分布(LDA)主题挖掘方法和支持向量机(SVM)分类算法,构建基于社交媒体的暴雨灾情信息挖掘模型,并以 2019 年“4·11 深圳暴雨”的微博数据为例,识别出 4 类微博文本主题,并进一步对“灾情信息”这一主题进行二次挖掘。

a. 构建的基于 LDA 主题挖掘方法和 SVM 分类算法的暴雨灾情信息挖掘模型,对微博文本中蕴含的具体灾情信息的识别与分类效果显著,一次文本主题分类和二次灾情信息分类的准确率分别为 88% 和 82%。一次文本主题挖掘提取出“天气状况”“灾情信息”“救援信息”和“无关信息”4 个主题。对主题中“灾情信息”进一步挖掘提取出“交通影响”“人员伤亡”“建筑物倒塌”“积水”“停电”和“停水”6 种具体灾情信息。

b. 暴雨灾情信息挖掘和分析的结果在时间上能够反映暴雨灾害的实时发展过程。深圳市气象局分别在 4 月 11 日晚和 4 月 19 日上午发布暴雨预警,因此微博数量在 11 日 23 时和 19 日 14 时达到两个峰值。受到实际事件的影响,不同类型灾情详细信息的微博数量与总微博数量变化特征基本一致。其中,“人员伤亡”“交通影响”以及“积水”3 类灾情受关注度较高。

c. 暴雨灾情信息的空间分布与灾情发生严重的地区基本保持一致。微博数量较大的地区主要分布在南山区、宝安区和福田区,其中南山区、宝安区均遭遇严重强降雨和冰雹侵袭,对灾害的关注程度更高。具体的暴雨灾情信息与该灾情发生的地点保持一致。“天气状况”相关微博主要集中在受暴雨侵袭的深圳市西北部的宝安区、光明区和龙华区。“救援信息”和“人员伤亡”的微博均集中在受灾严重、发生人员伤亡事件的南山区、福田区和罗湖区。“交通影响”的微博集中分布于机场所在的宝安区及其周边区域。

本文构建的暴雨灾情信息挖掘模型,能够实现对社交媒体中蕴含的详细灾情信息的挖掘,为暴雨灾情的监测提供了新的思路。在此研究的基础上,可以进一步构建基于暴雨灾害的灾情信息实时监测系统,帮助气象灾害应急管理部门掌握暴雨灾害本身及其次生、衍生灾害的发展趋势,从而对应急管理计划、措施和预案进行实时调整和动态优化。实时精准的灾情信息是灾害风险管理的基础,相关政府部门应将社交媒体信息纳入灾害监测预警与风险管理中,加快推进大数据资源融合共享,实现灾害多源立体监测及全方位预警。

参考文献:

[1] 黄国如,罗海婉,卢鑫祥,等. 城市洪涝灾害风险分析与区划方法综述[J]. 水资源保护,2020,36(6):1-6.

[2] 曾大军,曹志冬. 突发事件态势感知与决策支持的大数据解决方案[J]. 中国应急管理,2013(11):15-23.

[3] ZHANG Huaping, YU Hongkui, XIONG Deyi, et al. HHMM-Based Chinese lexical analyzer ICTCLAS[C]// Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan: Association for Computational Linguistics, 2003:184-187.

[4] 姚春华,罗强,胥小波,等. 一种网络文本信息情感分类的方法[J]. 通信技术,2019(11):2757-2760.

[5] 梁春阳,林广发,张明锋,等. 社交媒体数据对反映台风灾害时空分布的有效性研究[J]. 地球信息科学学报,2018,20(6):807-816.

[6] CHEN Liangzhe, HOSSAIN K, BUTLER P, et al. Syndromic surveillance of flu on Twitter using weakly supervised temporal topic models[J]. Data Mining and Knowledge Discovery, 2016,30(3):681-710.

[7] 张连峰,周红磊,王丹,等. 基于超网络理论的微博舆情关键节点挖掘[J]. 情报学报,2019,38(12):1286-1296.

[8] 牛景太. 基于奇异谱分析与 PSO 优化 SVM 的混凝土坝变形监控模型[J]. 水利水电科技进展,2020,40(6):60-65.

[9] 白华,林勋国. 基于中文短文本分类的社交媒体灾害事件检测系统研究[J]. 灾害学,2016,31(2):19-23.

[10] 夏梦南,杜永萍,左本欣. 基于依存分析与特征组合的微博情感分析[J]. 山东大学学报(理学版),2014,49(11):22-30.

[11] 陈梓,高涛,罗年学,等. 反映自然灾害时空分布的社

交媒体有效性探讨[J]. 测绘科学,2017,42(8):44-48.
[12] CROOKS A, CROITORU A, STEFANIDIS A, et al. Earthquake: Twitter as a distributed sensor system[J]. Transactions in GIS, 2013,17(1):124-147.
[13] 王艳东,李昊,王腾,等. 基于社交媒体的突发事件应急信息挖掘与分析[J]. 武汉大学学报(信息科学版),2016,41(3):290-297.
[14] HAN Xuehua, WANG Juanle. Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang City flood in China[J]. ISPRS International Journal of Geo-Information, 2019, 8(4):185-201.
[15] 王森,肖渝,黄群英,等. 基于社交大数据挖掘的城市灾害分析:纽约市桑迪飓风的案例[J]. 国际城市规划,2018,33(4):84-92.

(收稿日期:2020-05-27 编辑:胡新宇)

(上接第23页)

[2] 汪克亮,孟祥瑞,程云鹤. 环境压力视角下区域生态效率测度及收敛性:以长江经济带为例[J]. 系统工程,2016,34(4):109-116.
[3] 赵钟楠,张越,黄火键,等. 基于问题导向的水生态文明概念与内涵[J]. 水资源保护,2019,35(3):84-88.
[4] 张丛林,乔海娟,董磊华,等. 水生态文明制度体系框架研究[J]. 水利水电科技进展,2017,37(5):28-34.
[5] SCHALTEGGER S, STURM A. Ecological rationality[J]. Die Unternehmung, 1990,44(4):273-290.
[6] Eco-efficient leadership for improved economic and environmental performance[R]. Geneva: WBCSD, 1996: 3-16.
[7] Eco-efficiency[R]. Paris: Organization for Economic Cooperation and Development, 1998:1-2.
[8] 马骏,周盼超. 长江经济带生态效率空间异质性及其影响因素研究[J]. 水利经济,2019,37(6):8-12.
[9] 邓光耀. 基于污水排放量分配的中国水资源利用效率测算[J]. 水资源保护,2019,35(5):28-34.
[10] CHARNES A, COOPER W W, RHODES E. Measuring the efficiency of decision making units[J]. European Journal of Operational Research, 1978,6(2):429-444.
[11] ANDERSEN P, PETERSEN N C. A procedure for ranking efficient units in data envelopment analysis[J]. Management Science, 1993,39(10):1261-1264.
[12] 吴昊,车国庆. 中国地区生态效率的空间特征及收敛性分析[J]. 商业经济与管理,2018,38(5):50-61.
[13] 李雪松,张雨迪,孙博文. 区域一体化促进了经济增长效率吗? 基于长江经济带的实证分析[J]. 中国人口·资源与环境,2017,27(1):10-19.
[14] 侯孟阳,姚顺波. 1978—2016年中国农业生态效率时空演变及趋势预测[J]. 地理学报,2018,73(11):2168-2183.
[15] 陆砚池,方世明. 基于SBM-DEA和Malmquist模型的武汉城市圈城市建设用地生态效率时空演变及其影响因素分析[J]. 长江流域资源与环境,2017,26(10):1575-1586.
[16] TONE K. A slacks-based measure of efficiency in data envelopment analysis[J]. European Journal of Operational Research, 2001,130(3):498-509.
[17] TONE K. A slacks-based measure of super-efficiency in data envelopment analysis[J]. European Journal of Operational Research, 2002,143(1):32-41.
[18] CHUNG Y H, FARE R, GROSSKOPF S. Productivity and undesirable outputs: a directional function approach[J]. Environment Management, 1997,51(3):229-240.
[19] 蔡昉. 中国经济增长如何转向全要素生产率驱动型[J]. 中国社会科学,2013(1):56-71.
[20] 汪克亮,孟祥瑞,杨宝臣. 环境压力视角下中国省际生态效率的分解及收敛性[J]. 北京理工大学学报(社会科学版),2015,17(6):1-11.
[21] 卢二坡,杜俊涛. 环境策略互动与长江经济带的生态效率[J]. 陕西师范大学学报(哲学社会科学版),2018,47(6):68-78.
[22] 孟望生,邵芳琴. 黄河流域环境规制和产业结构对绿色经济增长效率的影响[J]. 水资源保护,2020,36(6):24-30.
[23] 刘云强,权泉,朱佳玲,等. 绿色技术创新、产业集聚与生态效率:以长江经济带城市群为例[J]. 长江流域资源与环境,2018,27(11):2395-2406.
[24] 卢丽文,宋德勇,李小帆. 长江经济带城市发展绿色效率研究[J]. 中国人口·资源与环境,2016,26(6):35-42.
[25] 孙欣,赵鑫,宋马林. 长江经济带生态效率评价及收敛性分析[J]. 华南农业大学学报(社会科学版),2016,15(5):1-10.
[26] 汪克亮,孟祥瑞,杨宝臣,等. 基于环境压力的长江经济带工业生态效率研究[J]. 资源科学,2015,37(7):1491-1501.
[27] 刘毅,周成虎,王传胜,等. 长江经济带建设的若干问题与建议[J]. 地理科学进展,2015,34(11):1345-1355.
[28] 崔木花. 长江经济带污染排放问题及情景规划[J]. 学术界,2015(4):218-227.
[29] 黄亮雄,安苑,刘淑琳. 中国的产业结构调整:基于三个维度的测算[J]. 中国工业经济,2013(10):70-82.
[30] 汪侠,徐晓红. 长江经济带经济高质量发展的时空演变与区域差距[J]. 经济地理,2020,40(3):5-15.

(收稿日期:2020-03-20 编辑:胡新宇)