

引用格式:梁春阳,林广发,张明峰,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. [Liang C Y, Lin G F, Zhang M F, et al. Assessing the effectiveness of social media data in mapping the distribution of typhoon disasters[J]. Journal of Geo-information Science, 2018,20(6):807-816.] DOI:10.12082/dqxxkx.2018.180022.

社交媒体数据对反映台风灾害时空分布的有效性研究

梁春阳¹, 林广发^{1,2,3*}, 张明峰^{1,2,3}, 汪玮杨¹, 张文富¹, 林金煌¹, 邓超¹

1. 福建师范大学 地理研究所, 福州 350007; 2. 福建省陆地灾害监测评估工程技术研究中心, 福州 350007;
3. 海西地理国情动态监测与应急保障研究中心, 福州 350007

Assessing the Effectiveness of Social Media Data in Mapping the Distribution of Typhoon Disasters

LIANG Chunyang¹, LIN Guangfa^{1,2,3*}, ZHANG Mingfeng^{1,2,3}, WANG Weiyang¹, ZHANG Wenfu¹, LIN Jinhuan¹, DENG Chao¹

1. Institute of Geography, Fujian Normal University, Fuzhou 350007, China; 2. Fujian Provincial Engineering Research Center for Monitoring and Assessing Terrestrial Disaster, Fuzhou 350007, China; 3. Research Center for National Geographical Condition Monitoring and Emergency Support in the Economic Zone on the West Side of the Taiwan Strait, Fuzhou 350007, China

Abstract: When a disaster occurs, a large number of images and texts with geographic information quickly flood the social network, which provides a new data source for timely awareness of disaster situations. However, due to the regional variation in the number of social media users and characteristics of information diffusion in cyberspace, new problems have risen in the mode analysis of spatial point processes represented by the check-in data. Examples are the correlation between check-in point density and disaster location density, spatial relation between check-in points or spatial heterogeneity of point pattern and associated influences. In this study, we took Typhoon No.14 in 2016 as an example and collected Sina Weibo data between September 14 and September 17, 2016 using keywords “Typhoon” and “Meranti”. We classified the Weibo texts using Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) algorithms and constructed a disaster database containing relevant check-in information. In addition, considering the spatial heterogeneity of Weibo users, we proposed a weighted model based on user activity at the check-in points. Using the global autocorrelation statistics Moran’s I as an indicator, we compared the check-in data before and after adding weights and discovered obvious spatial autocorrelation of the check-in data in real geographical locations. We tested our model on Weibo data with keyword “rain” and “power failure”. The results show that a series of maps generated by our model is able to reflect the typhoon disaster spatio-temporal process trends.

Key words: social media; typhoon disaster; spatial analysis; data mining; spatial autocorrelation

***Corresponding author:** LIN Guangfa, E-mail: GuangfaLin@qq.com

收稿日期 2018-01-02; 修回日期: 2018-04-08.

基金项目 国家重点研发计划重点专项(2016YFC0502905); 福建省公益科研院所专项(2015R1034-1); 福建省测绘地理信息局科技资助项目(2017JX03)。 [**Foundation items:** National Key Research and Development Program of China, No.2016YFC0505905; Non-profit Research Projects of Fujian Province, No.2015R1034-1; Development Foundation of Surveying, Mapping and Geoinformatics of Fujian Province, No.2017JX03.]

作者简介 梁春阳(1993-), 男, 硕士生, 研究方向为自发地理信息与应急管理。E-mail: peps8696@163.com

*通讯作者 林广发(1970-), 男, 副教授, 主要从事地理信息系统应用研究。E-mail: GuangfaLin@qq.com

摘要 当灾害事件发生时,与之相关的社交媒体数据不断产生,其中包含了丰富的灾情信息和签到地理位置信息,这为灾情态势的及时感知提供了一种新的数据源,但是因社交媒体用户量的地区差异及网络空间中信息传播模式的特点,给社交媒体签到数据所代表的空间点过程的模式分析带来了一些新的问题,如签到点密度与实际灾害点事件密度之间的对应关系、签到点之间的空间关系、点格局的空间异质性及其影响因素等。本文以2016年14号台风“莫兰蒂”为例,以“台风”和“莫兰蒂”为关键词,在新浪微博平台上采集了2016年9月14-17日的微博数据,使用文档主题生成模型(Latent Dirichlet Allocation, LDA)和支持向量机(Support Vector Machine, SVM)对微博文本进行分类,构建了含有签到位置信息的灾情点事件数据库。在此基础上,针对社交媒体用户分布的空间异质性提出了一种基于签到点用户活跃度的加权模型。以全局自相关统计量Moran's I 为指标,对加权前后的签到微博数据进行对比,发现这些在社交网络中产生的签到微博数据在现实地理空间中存在明显的空间自相关性;基于“雨”、“停电”等关键词,利用上述加权处理后的微博数据库进行灾害制图,结合真实灾情资料进行时空对比分析,结果表明系列图谱能够反映台风灾害的时空过程趋势。

关键词 社交媒体;台风灾害;空间分析;数据挖掘;空间自相关

1 引言

随着社交媒体的兴起,其丰富的数据吸引着各个领域的研究者,研究内容也呈现出多样的发展趋势,因该平台具有实时性和基于位置服务的特点,使其成为灾害应急管理的研究热点之一。例如,在2013年30号台风“海燕”登陆后,许多第一手资料及统计信息来自社交媒体,包括新闻信息、救援信息的发布^[1]以及灾害情绪释放等;2013年雅安庐山县地震,微博成为各种信息的主要传播平台,官方机构也将微博作为一个重要的信息发布平台,另外,微博对地震的反应几乎是即时的,远快于权威地震监测机构的发布^[2],震后9 min内,地震相关网络视频便开始不断涌现。许多学者通过对社交媒体中灾害事件的传播机制和作用的探讨,指出社交媒体可以以更快、更准确的方式,使民众、企业和非政府组织等社会力量参与到政府主导的灾害应急管理当中,且政府可以利用社交媒体传播灾害的相关信息,引导并整合民间力量,辅助政府决策^[3-4]。

现有研究中,社交媒体数据被广泛应用于灾害事件的实时检测和趋势预测等方面。白桦等^[5]利用支持向量机(SVM)识别地震相关的社交媒体数据,构建了Sina Weibo Incident Monitor系统,实现了基于社交媒体平台的地震灾害事件检测;Murzintcev等^[6]利用社交媒体的话题标签来过滤与事件无关的社交媒体数据,降低了训练文本分类模型成本,实现灾害相关信息的快速采集;Chen等^[7]分析了推文内容和流感发病状态的联系,基于隐含狄利克雷分布主题模型(LDA)推断社交媒体用户的生物状态,对流感的爆发时间进行预测。

上述研究大都聚焦于使用自然语言处理及文本挖掘技术对灾害事件相关的文本进行采集与识

别,未能充分利用社交媒体数据自身的时间和地理空间属性进行细粒度的灾情挖掘^[8-9]。对此,国内外学者使用GIS的空间分析方法论证了社交媒体签到点的分布模式和现实灾情空间分布之间的关联性。王艳东等^[10]采用基于密度的聚类算法DB-SCAN分析了“7·21”北京特大暴雨相关“交通状况”和“灾情”主题微博签到点的空间分布与现实灾情分布的关联;徐敬海等^[9]根据微博文本中的灾情描述词赋予签到微博权重,并对签到微博进行插值形成灾情分布图,应用于地震灾情的快速提取;陈梓等^[11]采用ArcGIS工具的热点分析对台风“达维”相关的签到微博点进行分析,发现签到微博的热点分布明显集中于强降雨地区;Bakillah等^[12]结合谱聚类与变密度聚类算法VDBSCAN对台风“海燕”袭击菲律宾后的相关推特数据进行地理社区检测,应用于识别和定位灾后发生的事件。但由于此类研究未考虑用户分布的空间异质性,导致分析结果偏向于人口聚集的地区^[9-13]。

为消除社交媒体用户分布的空间异质性,充分挖掘社交媒体灾情签到数据与现实台风灾情之间的关联性,本文在综合利用地图数据、灾情签到数据和签到点的日常签到数据的基础上,评估了因用户分布的空间异质性对数据分析的影响,提出了一种基于签到点用户活跃度的用户分布加权模型,并结合现实灾情资料进行实验验证。

2 数据源与灾害数据库的构建

2.1 数据采集与预处理

本文基于中国用户最多的社交网站之一——新浪微博,采集的微博数据属性字段包含文本内容、用户名称、发送时间、是否原创和签到位置用于

构建微博属性数据库。空间数据来自于全国县级行政区划矢量地图。灾情资料数据来自政府公告、新闻报道和相关文献中综合的受灾情况,用于对由微博数据反映的灾情分布和事件趋势进行验证。

(1) 微博数据的获取和去冗余

基于模拟登陆的方法,构建新浪微博数据抓取系统,在数据入库时采用用户ID和微博发送时间构成唯一标识,防止数据冗余。以“莫兰蒂”和“台风”为关键词采集了从2016年9月14日0时至9月18日0时,共96 h的台风“莫兰蒂”相关微博共17万余条,其中为用户原创并含有签到位置信息的微博数据共27 218条,以此作为待分类的微博数据集。

(2) 受台风影响区域的确定

为确定台风影响区域,人工整理了灾情资料数据,根据灾情资料显示受台风莫兰蒂影响区域分布于福建、浙江、江西东北部、江苏中南部、安徽东南部和上海等地区。将这些地区以市为单位统计微博签到条数,去除签到量较少的城市,并以全国县级行政区划的矢量地图为底图对受台风影响的地区以市为单位进行矢量化,作为待研究区域。

(3) POI签到量的获取

新浪微博官方每日会对POI的签到统计量进行更新,同样基于模拟登陆的方法,每日定时采集POI的签到次数,获得研究区域内POI一段时间内签到量的观测序列,为签到点的用户活跃度计算提供数据支撑。

2.2 微博文本数据隐含主题挖掘

采集的社交媒体数据蕴含不同类型的灾害应急信息。为确定文本中包含的类别,本文采用Blei等^[14]提出的隐含狄利克雷分布主题模型(Latent Dirichlet Allocation, LDA)来发现隐藏在微博文本中的主题,对2016年14号台风“莫兰蒂”相关微博进行主题分类,得到20个主题类别,但因该主题模型是一种无监督学习算法,故采用人工对20个主题类别分析完成主题的归纳与相似主题的合并,确定文本数据的待分类类别。通过整合20个主题类别最终确定“预警信息”、“灾情信息”、“无关信息”与“救援信息”4个类别,作为签到微博文本待分类类别。

2.3 微博文本数据分类

2.3.1 样本集的选择

为对微博文本分类,人工筛选800条微博构成训练文档集。首先,根据LDA模型整合的主题类

别,对800条样本类别标注;然后,利用中科院计算所研发的ICTCLAS分词组件^[15]对样本集进行分词,由于分词的准确性影响分类器的性能,因此对台风灾害特征词汇进行补充,建立针对台风灾害的词典,并去除广告,部分标点符号和介词等无关词汇;最后,因微博文本较短,表达较口语化,并含有丰富的符号,故本文将表情符号引入分词组件的词库,提高短文本的维度,以弥补微博文本的稀疏性。

2.3.2 文本特征选择与分类器参数寻优

本研究采用卡方检验^[16]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,因卡方检验在文本特征选择时会产生过分夸大低频词的作用,为此,在特征选择后,再使用词频-逆文档(Term Frequency- Inverse Document Frequency, TF-IDF)算法^[17]对特征词汇进行特征量化。

在文本分类领域中常见的分类算法有朴素贝叶斯和支持向量机(SVM),它们在执行文本分类的准确率问题上,学者普遍认为在召回率和准确率上SVM算法有较大优势^[18]。SVM是由Corinna等^[19]提出的一种机器学习方法,建立在统计学习理论的VC维理论和结构风险最小原理基础上根据有限的样本信息在模型复杂性和学习能力之间寻求最佳折衷。由于选择不同的核函数可以构造不同的SVM,其识别性能也不同,故本文采用较适用于文本分类的线性核函数对样本集进行训练得到分类模型。C(惩罚因子)是线性核函数必备的参数,其取值的好坏直接影响分类精度,本文采用K-fold Cross Validation交叉检验的方法确定C的取值^[20],即将选取的训练样本划分N部分,其中N-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N-1部分数据分类结果的精度,不断改变C来获取更高的分类精度。该模型在训练样本集的准确率为87.2%,说明SVM算法可以有效地对微博文本进行分类,最终得到13 088条含有签到位置信息的灾情微博。

3 用户分布加权模型的构建与检验

3.1 用户分布加权模型的构建

由于微博签到量的空间分布受用户分布影响^[9-13],若不考虑用户分布的空间异质性,会主观地认为签到量集中的位置与灾情发生或蔓延的方向相关。为消除用户分布差异对分析的影响,本文采

用新浪官方每日更新的POI签到量来刻画地区间社交媒体用户的分布差异并进行如下推导:在现实情况中当一次灾害事件发生时,由于签到点之间信息传播相互影响,可表达为一种联合概率链的形式(式(1)), P_{checkin} 为灾害发生时签到点集合出现的概率,但该种情况下,由于缺乏先验数据导致签到点之间的条件概率无法准确计算,研究中假设签到点之间相互独立,将式(1)简化为式(2)。

$$P_{\text{checkin}} = P(C_1)P(C_2|C_1)P(C_3|C_2, C_1), \dots, P(C_n|C_1, C_2, \dots, C_{n-1}) \quad (1)$$

$$P_{\text{checkin}} = \prod_{i=1}^n P(C_i) \quad (2)$$

$$P(C_i) = \frac{N_i}{T} \quad (3)$$

$$T = \sum_{i=1}^n N_i \quad (4)$$

式中: $P(C_i)$ 为位置*i*发生签到的概率,其中 N_i 为正常情况下未发生突发事件时位置*i*的签到量, T 为每个签到位置的签到量的求和。当灾害发生时,不同的微博用户会出现在相同位置进行签到的情况,于是将用户在相同位置签到的情况进行合并,即由式(2)合并相同项后得到式(5)。之后,将式(4)代入式(5)并对两侧取对数,得到式(6)。

$$P_{\text{checkin}} = P(C_1)^{n_1} \cdot P(C_2)^{n_2} \cdot \dots \cdot P(C_i)^{n_i} \quad (5)$$

$$-\ln(P_{\text{checkin}}) = \sum_{i=1}^m [n_i \cdot \ln(T/N_i)] \quad (6)$$

式中: n_i 和 N_i 项都对应单一签到点的属性值,但由于台风灾害影响范围广,本研究将以市级城市作为研究粒度,因此 n_i 、 N_i 和 T 的指代含义发生了改变。其中, n_i 为城市*i*的台风灾情签到微博的总和, N_i 为城市*i*的签到量, T 为研究范围内所有城市的签到量的求和。此时计算式(6)中的 N_i 项即可完成权重的计算。

为计算 N_i 项采集了2017年7月10-16日由新浪微博官方每日更新的城市签到次数,部分城市的签到次数如图1。由于社交媒体用户的签到行为可视为独立事件^[21],另在社交网络中分布着海量的签到点,即某微博用户在某签到点进行签到的行为是一种小概率事件记为 P_c ,而微博的用户为较大的群体记为 N_{user} (式(7)),以上情况符合泊松概率分布的特征,因此城市每日更新签到量的概率模型满足泊松分布,式(7)取极限后得到泊松分布的概率密度函数(式(8))。基于此本文采用极大似然估计的方法对构造的似然函数 $l(\lambda)$ (式(9))求解,完成城市每

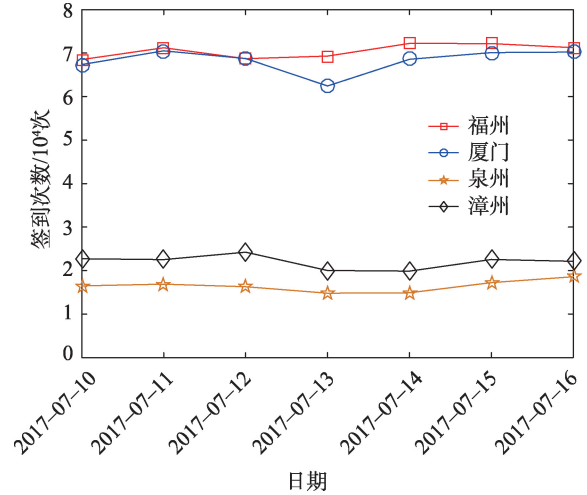


图1 部分城市每日签到次数统计图

Fig. 1 The statistics of daily check-in times in some cities
日更新签到量的参数估计。其中, x_{id} 为 N_i 项在*d*天的观测值, λ 为城市*i*每日更新签到量的极大似然估计值。

$$\lim_{N_{\text{user}} \rightarrow \infty, P_c \rightarrow 0} C_{N_{\text{user}}}^k P_c^k (1 - P_c)^{N_{\text{user}} - k} \quad (7)$$

$$f(X=k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (8)$$

$$l(\lambda) = \prod_{d=1}^n \frac{\lambda^{x_{id}}}{x_{id}!} e^{-\lambda} \quad (9)$$

根据上述假设和推导,本文将式(6)中 $\ln(T/N_i)$ 项记为城市*i*的微博用户签到活跃度,用该项对相应的城市灾情微博签到量进行加权,消除微博用户分布的空间异质性,并采用自然间断法对用户签到活跃度进行划分(图2),权重最低的一类城市分别是南京、合肥、苏州、上海、杭州、广州和深圳,该类城市相对图中其余城市经济较发达,信息化水平高,与客观事实相符,式(6)中的 n_i 为城市*i*的灾情签到微博数量,其空间分布如图3所示。

3.2 空间自相关检验

空间自相关是检验某一地理实体的属性值是否显著地与相邻空间属性值相关联的重要指标,分为正相关和负相关,正相关表明某单元的属性值变化与其邻近空间单元具有相同变化趋势,负相关则相反^[22]。本文采用空间自相关分析来研究灾情签到数据的空间分布特征,探究在低空间摩擦的社交网络中产生的灾情签到量在现实地理空间的分布模式,挖掘灾情签到点个数与其空间位置的相互作用程度。常用的全局空间自相关指标为Moran's *I*,其计算结果由研究粒度的大小、空间权重规则的定义和邻域范围共同决定^[23-24]。

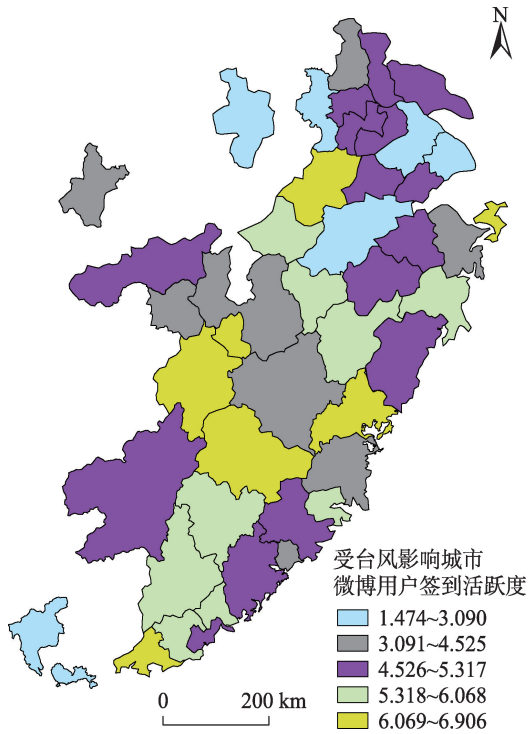


图2 微博用户签到活跃度分布图

Fig. 2 The distribution of microblog user check-in activities

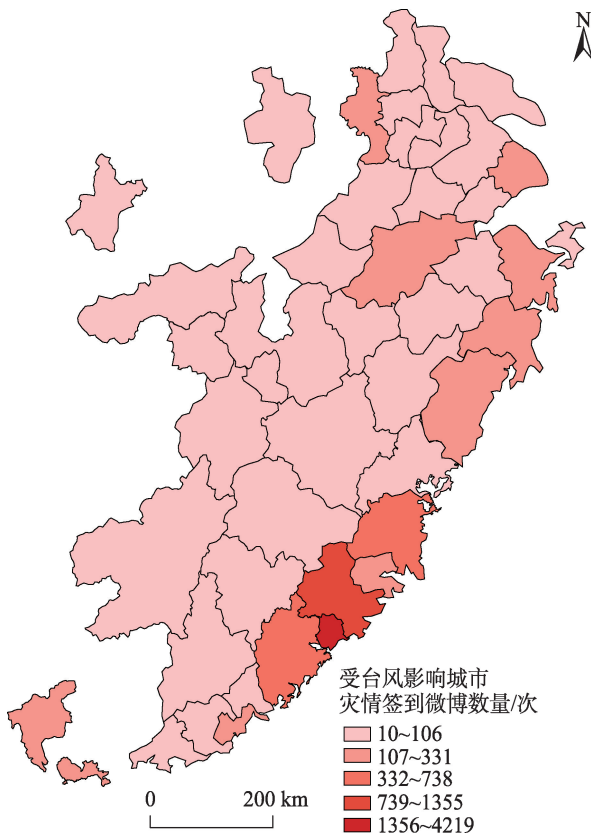
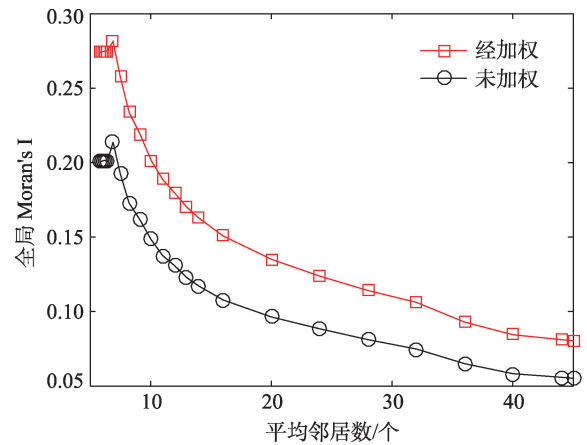
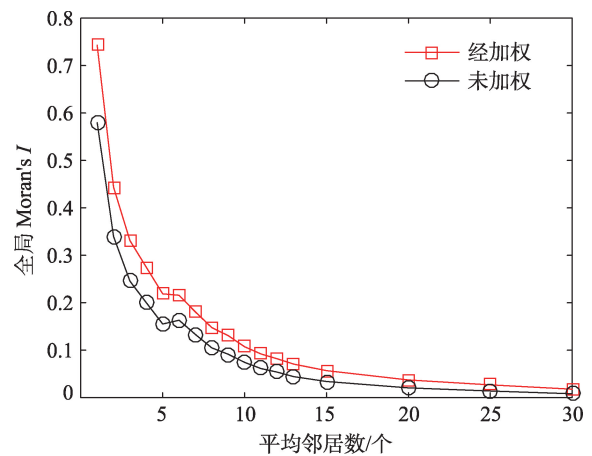


图3 灾情签到微博数量分布图

Fig. 3 The distribution of disaster-related microblog's records with location information

本研究以市为空间单元,使用反距离(Inverse Distance)和K近邻(K Nearest Neighbors)2种空间权重矩阵计算 Moran's I ,其随平均邻居数增加的折线统计图如图4、5。在ArcGIS软件中因将反距离权重的阈值距离设置为空时,程序根据输入的最小邻居数 N_{\min} 与地理要素的分布确定阈值距离,以确保每个地理要素至少存在 N_{\min} 个邻居,该机制使反距离空间权重下的平均邻居数以5.87为初始值(表1)。

图4 反距离空间权重下的平均邻居数与 Moran's I 指数Fig. 4 The average neighbors and global Moran's I using inverse distance weight图5 K近邻空间权重下的平均邻居数与 Moran's I 指数Fig. 5 The average neighbors and global Moran's I using K nearest neighbors weight

从图4、5可发现:①随平均邻居数增加, Moran's I 都呈下降趋势,其中K邻近权重下 Moran's I 下降趋势较在反距离权重矩阵下更明显,说明地理要素之间随着距离增加属性值差异也在增加,若不随着距离去调整 W_{ij} 权值, Moran's I 下降更加迅速。②未加权时,灾情签到微博整体呈空间正相关。经

表1 反距离权重下的平均邻居数与阈值距离对应表
Tab. 1 The comparison between average neighbors and threshold distance using inverse distance weight

最小邻居数/个	平均邻居数/个	阈值距离/km
1	5.87	178.6
5	6.41	369.1
10	10.04	488.3
15	15	637.8
20	20	750.5
...

加权后,在平均邻居数 $5.87 < N < 7.48$ (阈值距离 $0: 178.6 \text{ km} < \theta < 401.1 \text{ km}$) 的反距离空间权重矩阵下 Moran's I 值稳定在 0.28 左右且在 $P < 0.05$ 的显著性水平上,主要表现为以市为基础单元时位置相近的城市,其灾情微博签到量的相似程度更为显著。

4 时空分析

从 3.2 节的结果来看,虽然灾情签到微博在低空间摩擦的网络中产生,但是在现实地理空间中并不是随机分布。本节将灾情微博按照不同灾情特征词进行划分,探讨灾情签到微博的时空变化过程并挖掘其与现实灾情分布之间的关联性。若存在关联可根据灾情微博数据的空间分布和内容来快速感知受灾情况,甚至预测灾害的发展方向,对灾害应急管理有重要意义。

灾情微博包含台风灾害所产生的不同灾情特征词汇如:雨、停电和滑坡等,根据词频统计结果来看雨和停电两个词汇也是台风“莫兰蒂”的高频词汇,并能体现当时台风情况的关键词。因此,本文使用“雨”,“停电”这两个高频词为检索条件对灾情点事件数据库进行模糊查询,其中包含“雨”和“停电”的灾情微博分别为 3042 条和 1236 条;按照微博生成时间将含“雨”的灾情微博划分到 6 个时间段内,并计算每个时间段内签到点的平均中心位置,为使每个平均中心点权重值相等,本研究不按固定时间窗划分时间段,而是采用分位数法,即保证每个时间段内的微博灾情签到点个数相同。

由于在 3.1 节提出的用户分布加权模型与灾情签到点时空分析所采用的空间粒度不一致,导致无法直接通过该模型得到的权值即微博用户签到活跃度,计算灾情签到点的中心位置。另因灾情点在空间中以二维点的形式表达,所以需将权值取整后才能应用于计算灾情中心点的位置,而取整方式的

不同也会对原始数据的空间分布产生影响。本文为解决上述问题,首先,将每个图斑经模型计算得到的权值赋给该图斑内的各个灾情点;然后,为保证签到点的原始空间分布将权值同等倍数放大至整数,逐个对二维灾情签到点在原位置进行扩充;最后,对扩充后的坐标集合求中心坐标(式(10)),式中 T_a 表示第 a 个时间段, x_i 和 y_i 是 T_a 内数据点 i 的坐标, n 是 T_a 内的灾情微博总量。

$$\begin{cases} \bar{X}(T_a) = \frac{\sum_{i=1}^n x_i}{n}, \\ \bar{Y}(T_a) = \frac{\sum_{i=1}^n y_i}{n} \end{cases} \quad (10)$$

图 6 为以“雨”检索词经模糊查询得到的不同时段经加权与未经加权的灾情中心点空间分布图。

在等量灾情签到微博数量下,微博生成最密集的时段为 15 日 0:11–6:10,虽然该时段在凌晨在线用户相对较少,但报道频率却最高并且经加权和未经加权的灾情点都集中于厦门市图 6(b)。此外,随时间的推移经加权处理的灾情点中心坐标在逐渐向浙江和江苏移动,如图 6(c)–(e)所示,这与收集到的浙江省气象台、江苏省气象台和中国气象直播平台发布的降雨预警信息和实况灾情报道十分吻合。而未经加权的灾情点中心坐标虽有向浙江省方向移动的趋势但并不明显。在图 6(f)中,经加权处理的灾情点中心坐标向台风登陆位置回弹,我们猜想可能是因为最后一个时段内雨情减弱且分布较为均匀,另根据中国天气直播平台 17 日 7 时 51 分报道,台风“莫兰蒂”17 日凌晨在黄海南部海域变性为温带气旋,中央气象台凌晨 2 时对其停止编号,这则报道也验证了我们的猜想。

同上,以“停电”检索词经模糊查询得到的不同时段经加权与未经加权的灾情中心点空间分布图(图 7)。

图 7 中经加权与未经加权的灾情签到微博中心点位置都集中于厦门地区,说明台风“莫兰蒂”对登陆点厦门的影响程度相对于其他地区较严重,同样也通过福建省 2016 年气候公报和厦门日报发布的内容验证了这一事实。

此外,本研究在 6 个时间段分别对含“雨”和“停电”特征词的灾情微博,经用户分布加权模型处理后使用邻居数为 1 的 K 近邻空间权重计算得到的全局 Moran's I 值进行统计,以此对数据在空间中

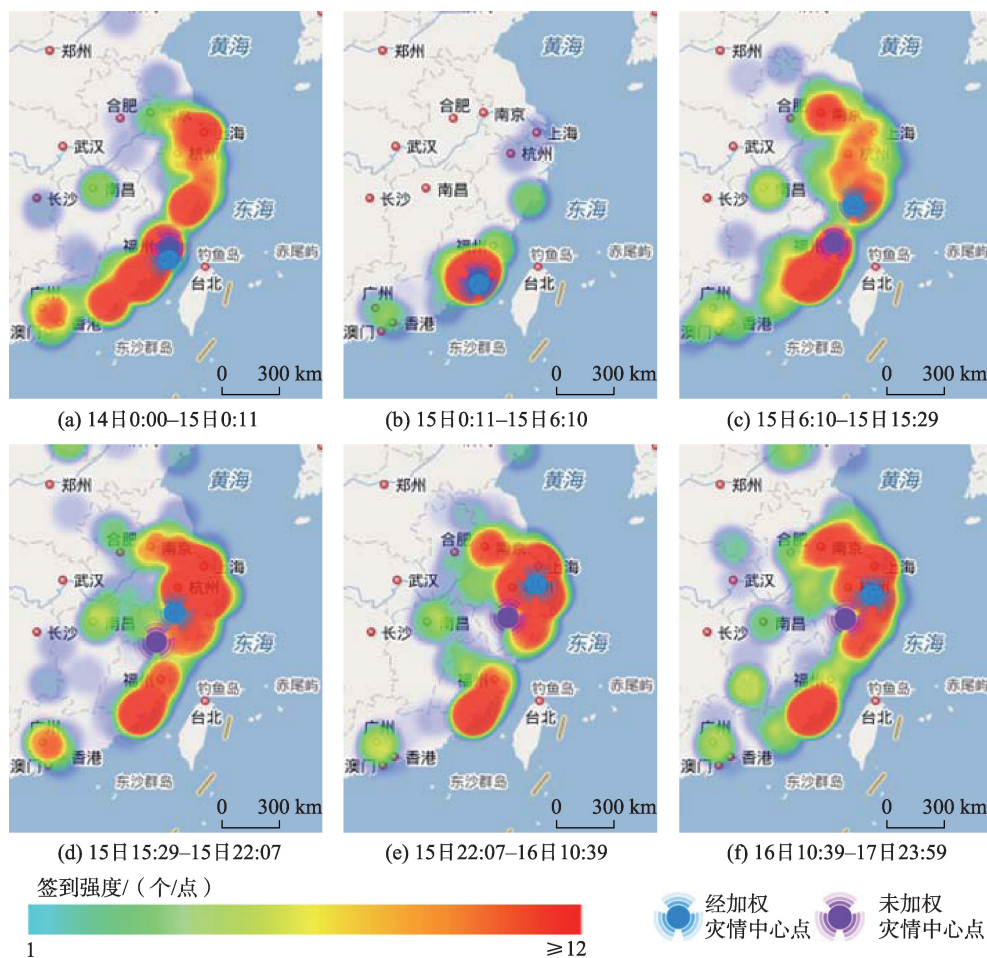


图6 以“雨”为关键词进行模糊查询的时序图

Fig. 6 The time series of maps with fuzzy query using "rain" as keyword

的相似程度测度(图8),时段划分方法同样采用分位数法。在14日0时至15日5时,含有“雨”和“停电”特征词的灾情签到微博其Moran's I 值均上升,说明签到数据在空间中呈聚集状态,台风的登陆时间为15日凌晨3时5分也正是在该时间区间内;含有“雨”特征词的灾情微博数据其峰值出现在15日6:10-15:29,而此时降雨在向浙江、江苏扩散的过程中,并且不同地区之间受降水的影响其生成的微博量相关性较强;含有“停电”特征词的灾情微博其Moran's I 峰值出现在台风登陆后的最近一个时段即15日3:49分至15日5:11分,此后迅速下降,观察数据属性值发现,含“停电”特征词的灾情微博数据主要集中于厦门和泉州,其余地区大都为0,因此厦门和泉州2个地区含“停电”特征词的灾情微博属性值的差异对Moran's I 影响较大。15日11:31分后,泉州地区含有“停电”特征词的灾情微博量的下降趋势明显大于厦门,从而导致Moran's I 迅速下降,

也说明厦门地区受灾情况较其他城市严重并且持续时间较长。

5 结论与展望

本文探索了一套社交媒体灾情签到数据的处理与分析方法,对2016年14号台风“莫兰蒂”灾情签到点的时空演变特征进行挖掘,并结合现实灾情资料进行了对比分析。对于如何消除社交媒体用户的空间分布差异,提出了一种基于签到点用户活跃度的用户分布加权模型,并采用时空分析的方法将灾情签到点和现实灾情资料进行对比分析,结果表明,经用户分布加权模型处理后的灾情签到点数据能更好的体现台风灾害的时空变化过程。空间自相关分析表明,经用户分布加权模型处理后,在市级城市空间粒度下,通过反距离空间权重矩阵在阈值距离为178.6 km $<\theta<$ 401.1 km时可使Moran's I

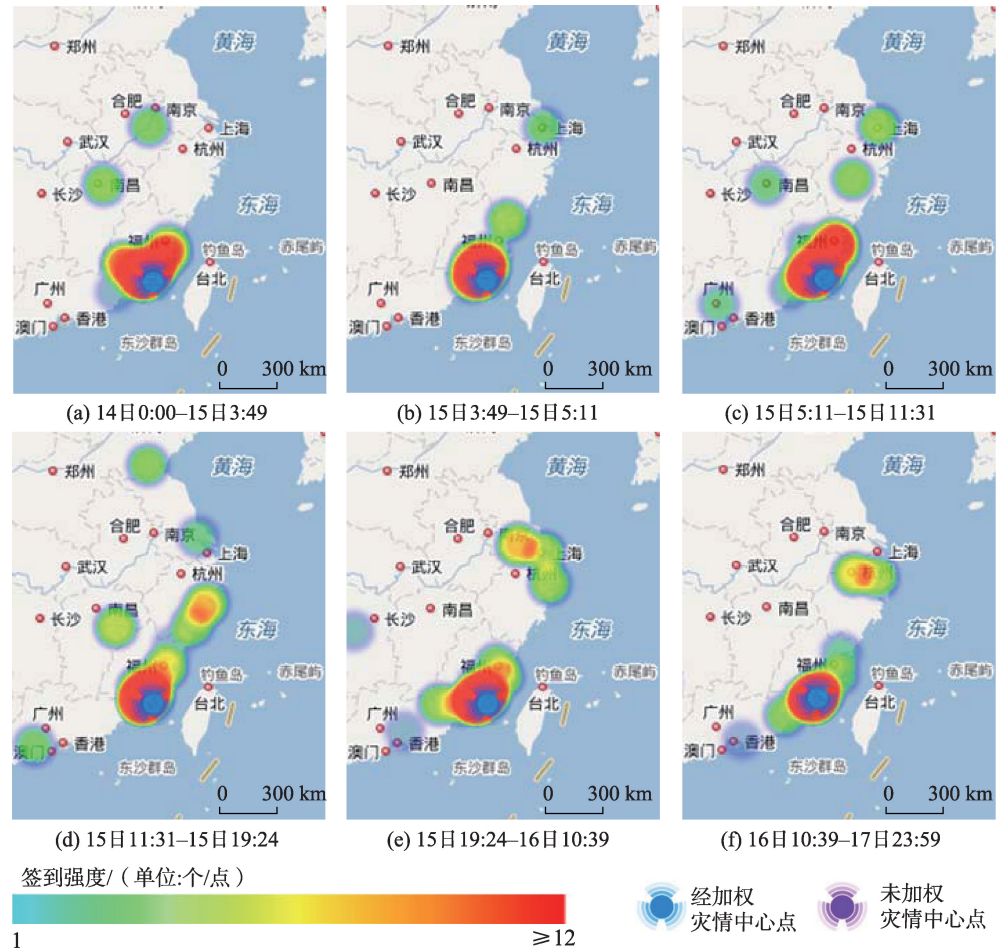


图7 以“停电”为关键词进行模糊查询的时序图

Fig. 7 The time series of maps with fuzzy query using "power failure" as keyword

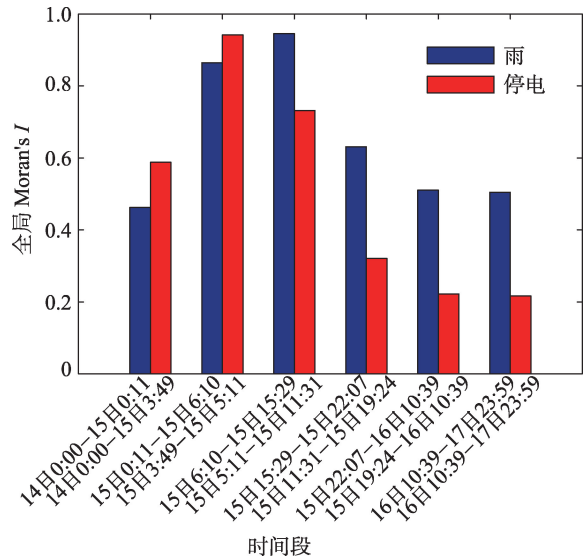


图8 含不同灾情特征词的微博在不同时间段
的Moran's I统计图

Fig. 8 Moran's I index of microblog's records with different
disaster keywords in different time period

稳定在0.28左右且 $P < 0.05$ 的显著性水平,说明灾情签到数据虽然在低空间摩擦的社交网络中产生,但在该空间粒度下灾情签到微博整体上存在显著的空间正相关性,为通过微博签到量的空间分布模式来感知灾情分布提供了理论依据。

但是,空间数据具有多粒度、多尺度特性,属性数据之间的关系常随空间尺度和单元划分方式的不同而发生变化^[25-26],本文仅在市级城市的空间尺度下进行空间自相关分析得到灾情签到数据的分布模式为正相关,下一步研究将对灾情签到数据在多种粒度大小和划分方式下进行综合分析,寻求最合适的表达粒度。

社交媒体数据的兴起为快速感知灾情信息提供了新的途径,但其中也包含了大量的冗余信息,并存在一定的片面性,与经过严谨科学实验得到的数据相比,社交媒体数据也是一种有偏数据,挖掘社交媒体数据需要更严谨和鲁棒的算法与模型^[27]。下一步将结合基于深度学习框架的Word2vec模

型^[28]度量词汇相似性,完善词汇近义词表,提高文本分类与数据检索精度并构建台风灾害描述本体,建立科学和严谨的社交媒体数据实时监测和动态展示系统。

参考文献(References):

- [1] Cool C T, Claravall M C, Hall J L, et al. Social media as a risk communication tool following typhoon Haiyan[J]. Western Pacific Surveillance & Response Journal Wpsar, 2015(Suppl 1):86-90.
- [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. [Zeng D J, Cao Z D. Big data solutions for emerging situation awareness and decision[J]. China Emergency Management, 2013(11):15-23.]
- [3] Terpstra T, Vries A D, Stronkman R, et al. Towards a real-time twitter analysis during crises for operational crisis management[C]. Proceedings of the 9th International IS-CRAM Conference, 2012(4):1-9.
- [4] Vieweg S, Hughes A L, Starbird K, et al. Microblogging during two natural hazards events:What twitter may contribute to situational awareness[C]. Proceedings of the SIGCHI Conference on Human Factors in Computing System. New York, USA: ACM, 2010:1079-1088.
- [5] 白桦,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. [Bai H, Lin X G. Sina weibo disaster information detection based on chinese short text classification[J]. Journal of Catastrophology, 2016,31(2):19-23.]
- [6] Murzintcev N, Cheng C. Disaster hashtags in social media[J]. International Journal of Geo-Information, 2017,6(7):204.
- [7] Chen L, Hossain K S M T, Butler P, et al. Flu Gone Viral: Syndromic surveillance of flu on Twitter using temporal topic models[C]. IEEE International Conference on Data Mining. IEEE Computer Society, 2014:755-760.
- [8] Wang Z, Ye X, Tsou M H. Spatial, temporal, and content analysis of Twitter for wildfire hazards[J]. Natural Hazards, 2016,83(1):523-540.
- [9] 徐敬海,褚俊秀,聂高众,等.基于位置微博的地震灾情提取[J].自然灾害学报,2015,24(5):12-18.[Xu J H, Chu J X, Nie G Z, et al. Earthquake disaster information extraction based on location microblog[J]. Journal of Natural Disasters, 2015,24(5):12-18.]
- [10] 王艳东,李昊,王腾,等.基于社交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,2016,43(3): 290-297. [Wang Y D, Li H, Wang T, et al. The mining and analysis of emergency information in sudden events based on social media[J]. Geomatics and Information Science of Wuhan University, 2016,43(3):290-297.]
- [11] 陈梓,罗年学,高涛.基于VGI的台风灾情评估研究[J].测绘与空间地理信息,2016,39(10):33-34. [Chen Z, Luo N X, Gao T. Research of typhoon disaster assessment based on VGI[J]. Geomatics & Spatial Information Technology, 2016,39(10):33-34.]
- [12] Bakillah M, Li R Y, Liang S H L. Geo-located community detection in Twitter with enhanced fast-greedy optimization modularity: The case study of typhoon Haiyan[J]. International Journal of Geographical Information Science, 2015,29(2):258-279.
- [13] 陈媛媛,高勇.利用社交媒体的位置潜语义特征提取与分析[J].地球信息科学学报,2017,19(11):1405-1414. [Chen Y Y, Gao Y. Extracting and analyzing latent semantic characteristics of locations using social media data[J]. Geo-Information Science, 2017,19(11):1405-1414.]
- [14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003,3: 993-1022.
- [15] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-Based chinese lexical analyzer ICTCLAS[C]. Proceedings of the 2nd SigHan Workshop, 2003:184-187.
- [16] Meesad P, Boonrawd P, Nuipian V. A Chi-Square-Test for word importance differentiation in text classification[C]. International Conference on Information and Electronics Engineering, 2011.
- [17] Jones K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972,28(1):11-21.
- [18] Colas F, Brazdil P. Comparision of SVM and some older classification algorithms in text classification task[J]. Artificial intelligence in Theory and Practice, 2006,217: 169-178.
- [19] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995,20:273-297.
- [20] Bengio Y, Gr Y. No unbiased estimator of the variance of K-Fold cross-kalidation[J]. Journal of Machine Learning Research, 2003,5(22):1089-1105.
- [21] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users;real-time event detection by social sensors[C]. Proceedings of the 19th International Conference on World Wide Web. New York,USA: ACM, 2010:851-860.
- [22] 陈彦光.基于Moran统计量的空间自相关理论发展和方法研究[J].地理研究,2009,28(6):1449-1463. [Chen Y G. Reconstructing the mathematical process of spatial autocorrelation based on Moran's statistics[J]. Geographical Research, 2009,28(6):1449-1463.]

- [23] 李双成,蔡运龙.地理尺度转换若干问题的初步探讨[J].地理研究,2005,24(1):11-18. [Li S C, Cai Y L. Some scaling issues of geography[J]. Geographical Research, 2005,24(1):11-18.]
- [24] 孟斌,王劲峰.地理数据尺度转换方法研究进展[J].地理学报,2005,60(2):277-288. [Meng B, Wang J F. A review on the methodology of scaling with geo-data[J]. Acta Geographica Sinica, 2005,60(2):277-288.]
- [25] Qi Y, Wu J G. Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices[J]. Landscape Ecology, 1996,11(1):39-49.
- [26] Jelinski D E, Wu J. The modifiable areal unit problem and implications for landscape ecology[J]. Landscape Ecology, 1996,11(3):129-140.
- [27] 吴志峰,柴彦威,党安荣,等.地理学碰上“大数据”:热反应与冷思考[J].地理研究,2015,34(12):2207-2221. [Wu Z F, Chai Y W, Dang A R, et al. Geography interact with big data: dialogue and reflection[J]. Geographical Research, 2015,34(12):2207-2221.]
- [28] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013(1):28-36.