

文章编号:1672-3961(2020)05-0118-09 DOI:10.6040/j.issn.1672-3961.0.2019.371

# 基于微博数据的台风“山竹”舆情演化时空分析



张岩<sup>1,2</sup>, 李英冰<sup>1\*</sup>, 郑翔<sup>3</sup>

(1. 武汉大学测绘学院, 湖北 武汉 430079; 2. 武汉大学测绘遥感信息工程国家重点实验室, 湖北 武汉 430079; 3. 武汉大学信息管理学院, 湖北 武汉 430072)

**摘要:** 将情感分析模型、动态演化模型、话题聚类模型、网络社团模型结合地理可视化技术应用到台风的灾害评估中。将微博情绪与台风灾害联系起来, 从情感值与讨论热度两个角度入手, 根据台风“山竹”相关话题的 25 798 条微博数据, 完整的展示本次事件网络舆情的演化过程, 通过隐含狄利克雷分布(latent dirichlet allocation, LDA)主题模型挖掘用户关注话题, 发现台风登陆事件与湖南收费站事件对公众情绪的消极影响; 抽取台风“山竹”相关微博中蕴含的地理位置信息, 建立广东省 21 个城市的网络社团模型, 检验用户情绪、城市词频、用户位置、网络节点活跃度等指标探测受灾城市的能力; 根据广东省 38 个气象站点的 24 h 最大降雨数据进行空间插值。降水主要集中在广东南部地区, 阳江市发生特大暴雨, 引发了严重的洪涝灾害, 其情绪值也是最低的。

**关键词:** 自然语言处理; 空间分析; 社团分析; 新浪微博; 公共安全

**中图分类号:** TU984.116; X43      **文献标志码:** A

**引用格式:** 张岩, 李英冰, 郑翔. 基于微博数据的台风“山竹”舆情演化时空分析[J]. 山东大学学报(工学版), 2020, 50(5): 118-126.

ZHANG Yan, LI Yingbing, ZHENG Xiang. Spatial and temporal analysis of network public opinion evolution of typhoon “Mangkhut” based on Weibo data[J]. Journal of Shandong University(Engineering Science), 2020, 50(5): 118-126.

## Spatial and temporal analysis of network public opinion evolution of typhoon “Mangkhut” based on Weibo data

ZHANG Yan<sup>1,2</sup>, LI Yingbing<sup>1\*</sup>, ZHENG Xiang<sup>3</sup>

(1. School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, Hubei, China; 2. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan 430079, Hubei, China; 3. School of Information Management, Wuhan University, Wuhan 430072, Hubei, China)

**Abstract:** Internet public opinion was the sum of the public various emotions, attitudes and opinions on related topics. This paper applied the sentiment analysis model, dynamic evolution model, topic clustering model, network community model and geographic visualization technology to the typhoon disaster assessment. This research fully illustrated how the public opinion on typhoon “Mangkhut” evolved by analyzing the 25 798 Weibo related from the two perspectives of emotional value and discussion heat. By utilizing LDA clustering method the negative impacts of typhoon “Mangkhut”的 landing and “Hunan toll station event” on public sentiment were found. After collecting the geographical location information of those typhoon “Mangkhut” related Weibo, a network-community model of 21 cities in Guangdong province was established and the model tested the ability to explore the affected cities through such indicators as users’ sentiments, city word frequency, users’ location, and network node activity. Spatial interpolation was performed based on the 24 h maximum rainfall data from 38 meteorological stations in Guangdong province. Precipitation was mainly concentrated in the southern part of Guangdong. Heavy rains occurred in Yangjiang City, which caused severe flooding and the lowest emotional value.

**Key words:** natural language processing; spatial analysis; community analysis; Sina Weibo; public safety

收稿日期:2019-07-12; 网络首发时间:2020-02-22 15:39:59

网络出版地址: <http://kns.cnki.net/kcms/detail/37.1391.T.20200221.1529.004.html>

基金项目: 国家重点研发项目计划(2018YFC0807000)

第一作者简介: 张岩(1997—), 男, 河南临颍人, 博士研究生, 主要研究方向为时空大数据. E-mail: sggzhang@whu.edu.cn

\*通信作者简介: 李英冰(1972—), 男, 湖北房县人, 博士, 副教授, 主要研究方向为时空大数据. E-mail: ybli@sgg.whu.edu.cn

## 0 引言

实时产生的社交媒体数据对研究自然灾害的演变与影响有着相当重要的作用,一方面可以基于社交媒体数据监测用户的情感波动,从而挖掘灾区人民的话题关注与情感态度;另一方面获取灾区最新信息,融合多源信息合理的划分地区风险等级,进而制定更为合理的救援方案。台风是海上丝绸之路经济带受影响最为严重的自然灾害<sup>[1]</sup>,从灾害发生期间产生的大量相关微博数据中挖掘出有用的信息,结合地理可视化技术,依据多个维度(实时情感、舆情热度、话题关注、地理分布、社团划分、气象条件等),将台风灾害过程全景式完整还原,既有利于总结台风灾害规律,弥补抗灾过程中的短板与不足,又可以在台风灾害发生过程中,实时的掌握灾区灾害动态与大众舆论风向,为应急决策提供可靠参考。

国内外对于社交媒体数据在灾害过程中的应用研究大部分基于文本数据,忽视文本背后的地理信息与文本所蕴含的区域社团信息。Ning 等<sup>[2]</sup>利用神经网络进行推特文本危机监测,从大量推文中自动识别出最重要和最有价值的推文;Nair 等<sup>[3]</sup>利

用机器学习方法,针对 2015 年钦奈洪水灾害进行了推文分类并进行了网络意见领袖的识别;Alfarrarjeh 等<sup>[4]</sup>利用多源社交媒体数据与多种情感研究方法,对飓风桑迪和纳帕地震进行了地理情感分析;唐晓波等<sup>[5]</sup>提出微博热度的概念,并将其引入到 LDA 模型研究中;阮光册<sup>[6]</sup>将 LDA 模型与 HowNet 知识库结合进行微博评论的主题发现;Choi 等<sup>[7]</sup>对社会大数据的实时搜索进行了 Twitter 用户地理位置与灾情演化的分析;白华等<sup>[8]</sup>基于微博评论数据,开发了高效的灾害事件即时检测系统;陈梓等<sup>[9]</sup>论证了微博数据与台风灾情的联系;王心瑶等<sup>[10]</sup>分析了 H7N9 事件中微博舆情演变与走势;梁春阳等<sup>[11]</sup>利用 LDA 与支持向量机(support vector machine, SVM)进行微博关键字分类展示了台风灾害的时空过程趋势。

本研究在上述文章基础上,将情感分析模型、动态演化模型、话题聚类模型、网络社团模型结合地理可视化技术引入到台风灾害的评估中。充分挖掘了社交媒体数据蕴含的位置信息与情感信息,不仅将用户位置纳入考虑,而且检索文本中蕴含的地理信息,构建了城市舆情社团网络。论文方法流程图如图 1 所示。

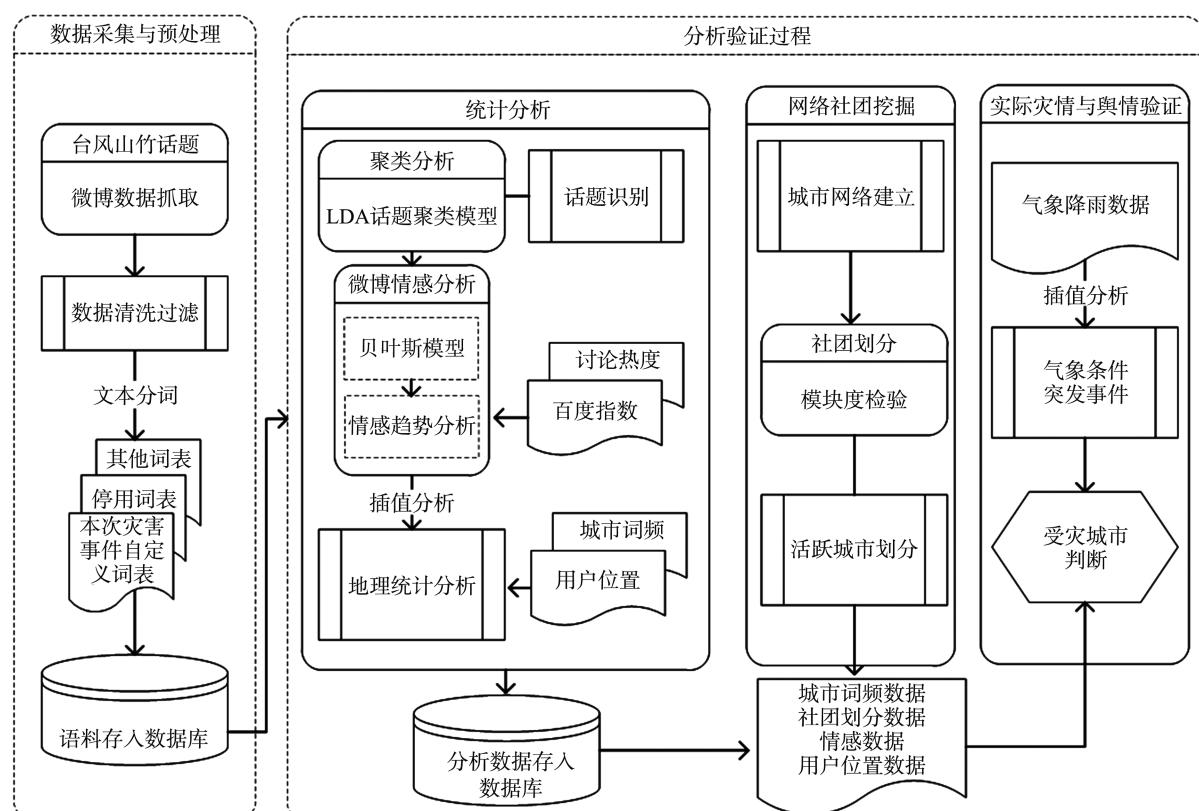


图 1 论文方法流程图

Fig.1 The workflow chart of the proposed method

# 1 研究区概况、数据源及研究方法

## 1.1 研究区概况

2018 年 9 月 7 日台风“山竹”在太平洋海面生成,9 月 14 日中央气象台发布黄色预警,9 月 15 日登陆菲律宾,9 月 16 日晚间在广东登陆,造成广东、广西、海南、湖南、贵州 5 省(区)近 300 万人受灾,在此期间公众通过互联网表达和传播了大量网络舆情信息。主要研究区域为受灾最为严重的广东省,地处  $20^{\circ}13'N \sim 25^{\circ}31'N$  与  $109^{\circ}39'E \sim 117^{\circ}19'E$  之间,下辖 21 个地级市,国土面积 17.97 万  $km^2$ 。在新浪微博上,有关“台风山竹”的讨论主要集中在 2018 年 9 月 14 日至 2018 年 9 月 21 日,因此选择此区间的舆情数据作为主要研究。

## 1.2 数据源

编写程序自动采集人民日报、中国新闻网、环球网、环球时报、中国新闻周刊、头条新闻等权威新闻媒体关于台风山竹话题微博的用户评论,获得评论数据、发表日期、评论用户编号、微博编号与评论用户昵称等内容,将其存入数据库。微博评论采集完毕后,利用获取到的评论用户编号,获取用户所在地、性别等信息,利用地址解析接口获取用户所在地的经纬度坐标。通过正则表达式过滤、字符串分割等方法提取准确的微博发表时间,然后对获取到的微博评论进行文本清洗,过滤掉微博表情以及特殊符号,删除空微博,删除信息不完整的用户,经过数据清洗后有 25 798 条有效数据。

为方便话题聚类分析,将获取到的微博文本进行分词。加载哈工大停用词典过滤掉对文本分析没有帮助的语气词、助词与标点符号<sup>[5]</sup>,将搜狗词典、盘古词典、腾讯词典、百度词典以及本次台风灾害自定义词典加载到结巴分词工具进行分词。

## 1.3 研究方法

### 1.3.1 舆情话题聚类

LDA 模型是由 Blei 等<sup>[12]</sup>提出的生成主题概率模型,通常用来对大规模文档数据进行建模。LDA 模型在以往的词语-文章结构中加入了主题,成为词语-主题-文档 3 层结构。其思想来源于一个基本假设:文档是由多个隐含主题构成,这些隐含主题由若干个特定特征词构成。

LDA 模型的优点在于具有清晰的内在结构且采用无监督方法进行训练,适合对大量数据进行分类处理。其对文本信息的主题建模如图 2 所示。

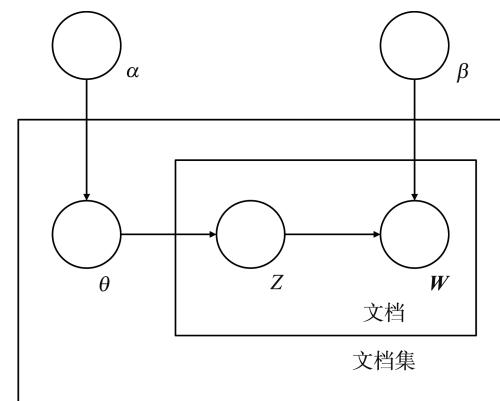


图 2 LDA 模型主题建模示意图  
Fig.2 Interval diagram of LDA model

在文档集中,参数  $\alpha$  反映潜在主题之间的相对强弱,  $\alpha$  越大, 文档包含的主题更多, 反之包含的主题更少。 $\beta$  表示为所有潜在主题的概率分布,  $\beta$  越大, 主题包含的单词越多, 反之包含的单词越少; $\theta$  表示在目标文档中, 潜在主题的比重; $w$  是目标文档的词向量表示, $z$  则表示该文档分配在每个词项上的潜在主题的个数。假设  $m$  是一个潜在主题,  $w_i$  是文档中的第  $i$  个词语, 则  $w_i$  属于  $m$  的概率

$$P(w_i) = \sum_{m=1}^k P(w_i | z_i = m) P(z_i = m), \quad (1)$$

式中: $P(w_i | z_i = m)$  表示词语属于潜在主题  $m$  的概率,  $P(z_i = m)$  表示  $m$  是文档的主题概率。

在实际情况中,LDA 模型精确解算较为复杂,采用了最常用的吉布斯采样方法进行参数估计<sup>[13]</sup>,输入分词后的微博数据,参数  $\alpha = 0.5$ ,  $\beta = 0.1$ , 迭代次数 300 次,聚类数  $K = 10$ 。最佳聚类数的选取可以通过词汇被选中的概率或者困惑度<sup>[14]</sup>来进行计算。

根据计算结果,从每个主题选出 7 个出现频率较高的词语,不同主题之间的区分较为明显。本次台风事件中热议的话题有:台风命名与山竹除名的讨论、港珠澳大桥扛过 17 级大风、工作人员坚守岗位、救援队返程被卡湖南收费站、俄航起飞、学校停课、台风预警等。

### 1.3.2 情感分析

情感分析是指针对用户评论文本进行有效的分析与挖掘,识别其情感趋势,也就是对于某件事情或某个人物持有“赞同”或者“反对”态度进行分类,从中提取出用户的感情极性实际上是一种二分类问题。情感分析技术常常用在用户评论分析与决策,舆情监控与信息预测等领域。

微博文本情感信息挖掘方法一般分为两类,一种是基于情感词典的方法,需要加载丰富的情感词

表,根据文本中所包含的正向情感词和负向情感词的个数来进行打分<sup>[15]</sup>。情感词典的方法需要兼顾语料所处的实际语境,并且很难顾及文本的顺序以及句子的隐含语境。国内最早由朱嫣岚等<sup>[16]</sup>基于知网知识库 HowNet 中一定数量的基准词对(贬义词,褒义词)对单词打分,基准词对中的贬义词表示为 kp,褒义词表示为 kn,来计算单词的语义倾向值,假设有 k 对基准词,计算公式为

$$O(w) = \sum_{i=1}^k S(kp_i, w) - \sum_{j=1}^k S(kn_j, w), \quad (2)$$

式中: $O(w)$  代表单词打分结果, $S(kp_i, w)$  代表单词与第  $i$  个贬义词相似度, $S(kn_j, w)$  代表单词与第  $j$  个褒义词相似度。

另外一种是基于机器学习的方法,需要大量人工标注好的语料库作为训练集,通过提取文本特征(词袋法、词嵌入法),构建分类器(朴素贝叶斯、支持向量机、最大熵、卷积神经网络、长短期记忆网络)实现情感分类。提取文本特征需要将文本向量化方便后续处理,词袋法是最常用的文本特征向量表示方法,可以将文本内容投影到高维空间。词袋法是以单词为最小处理单元,将文本拆分构建成一个词典,且词典内每一个单词都有唯一的索引,假设 $\{w_1:1, w_2:2, w_3:3, \dots, w_n:n\}$  是一段文本构成的一个词典, $w$  代表不同的单词,则一句话  $w_1 w_2 w_3 w_1$  可以用一个  $n$  维向量表示成 $(2, 1, 1, 0, \dots, 0)$ ,以单词出现次数代表单词权重。这样的特征提取方法简单可靠,但是存在空间浪费并且无法保留单词顺序信息。除此之外还可以采用词频(TF)、逆向文件频率(IDF)、卡方统计、信息增益、构建  $N$  元模型等方法提取文本特征。

本文采用了朴素贝叶斯情感分类方法,设输入空间  $X \in \mathbf{R}_n$  为  $n$  维向量的集合,输出空间为标记好的集合  $Y = \{c_1, c_2, \dots, c_k\}$ ,输入特征向量输出类的标记,加载已经标注好的训练集  $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 。假设联合概率  $P(X, Y)$  为独立分布,则文本向量属于情绪类别  $c_k$  的概率是

$$P\left(\frac{Y=c_k}{X=x}\right) = \frac{P\left(\frac{X=x}{Y=c_k}\right) P(Y=c_k)}{\sum_k P\left(\frac{X=x}{Y=c_k}\right) P(Y=c_k)}. \quad (3)$$

本研究调用了百度自然语言处理情感分析接口,上传标注好的数据进行训练。每次请求返回情感分类的结果(积极、消极、中性),以及分类的置信度,积极类别的概率与消极类别的概率,并以积极类别的概率作为情绪指数。接口返回参数结构如

表 1 所示。

表 1 情感倾向分析接口返回参数  
Table 1 Emotional tendency analysis interface returns parameters

参数	说明	描述
log_id	uint64	请求唯一标识码
sentiment	int	表示情感极性分类结果,0:负向,1:中性,2:正向
confidence	float	表示分类的置信度,取值范围[0,1]
positive_prob	float	表示属于积极类别的概率,取值范围[0,1]
negative_prob	float	表示属于消极类别的概率,取值范围[0,1]

### 1.3.3 地理统计分析

地理统计是分析含有地理属性数据的统计方法<sup>[17]</sup>。现实中的大部分事物都与位置有关,对于点数据,可以采用频率统计或者插值分析方法,从有限的数据点上得出任意点的数值,进行空间上某个属性连续分布的展示。对于面数据,通过空间相关性研究,发掘事物的空间分布格局和背后的产生原因。

反距离加权法(inverse distance weighted, IDW)插值是空间分析中常用的插值方法,其优点在于可以将所有数据点都纳入考虑,且数据点对插值点的影响随着距离的变化而改变,本文采用欧氏距离公式作为距离函数。反距离插值的权重函数

$$w_i = \frac{h_i^{-p}}{\sum_{i=1}^n h_i^{-p}}, \quad (4)$$

式中: $P$  为幂参数, $n$  为数据点个数, $h_i$  代表插值点与数据点  $i$  的欧氏距离。插值点的属性计算公式为

$$Z(x, y) = \sum_{i=1}^n w_i \cdot Z(x_i, y_i). \quad (5)$$

### 1.3.4 网络社团模型

社团也可称之为聚类、群等,社团结构是复杂网络的一个重要拓扑结构特征<sup>[18]</sup>。城市舆情网络社团是一组由城市点位与有向共现城市词链接组成的集合,城市间联系紧密且存在社团化或群组化的结构。社团结构发现是指根据共现城市词链,将城市节点一个个划分到社团中的过程,社团内部节点存在某种特质。常用的社团结构发现算法包括图分割理论、Louvain 算法、GN 算法、Newman 快速算法等,常用于社交软件联系人自动推荐中。

本研究采用 Louvain 算法进行社团划分:其优点在于算法效率很高,可以将所有边缘节点统一纳入考虑且获得的社团结构具有层次性<sup>[19]</sup>。该算法的划分评判基于模块度这一指标,模块度越大代表

划分结果越好,社团结构越明显,模块度数学上的定义如下:

$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j), \quad (6)$$

式中: $A_{i,j}$ 代表节点*i*与*j*之间的边的权重; $k_i$ 代表节点的度(节点的弧尾条数加弧头条数); $m$ 代表复杂网络中节点的总数; $C_i$ 代表节点为*i*的社团,当 $C_i = C_j$ 时 $\delta$ 函数为1,否则为0。在随机情况下,节点*i*与节点*j*之间的边数为 $\frac{k_i k_j}{2m}$ 。Louvain 算法具体流程如下:

(1) 将每一个节点看作一个独立的社团,初始社团数目与城市数目相同。

(2) 遍历任意一个节点*i*,考虑其邻居*j*,通过从节点*i*所属社团移除节点*i*,然后将其加入属于节点*j*的社团,计算模块度的变化并进行比较,将节点*i*放入模块度增加最大的社团,若无法找到模块度收益为正的节点*j*,则保持节点*i*原有社团。

(3) 重复步骤(2),对于所有节点都执行此过程,直至达到模块度局部最大值,即没有任何节点可以提高网络模块度,社团结构不再发生改变。

(4) 对步骤(3)得到的社团结构进行压缩,将原有社团压缩成新节点,社团内部节点权重转化为新节点环权重,原社团之间边权重转化为节点之间边权重。

(5) 重复步骤(1)直至社团结构不再发生改变。

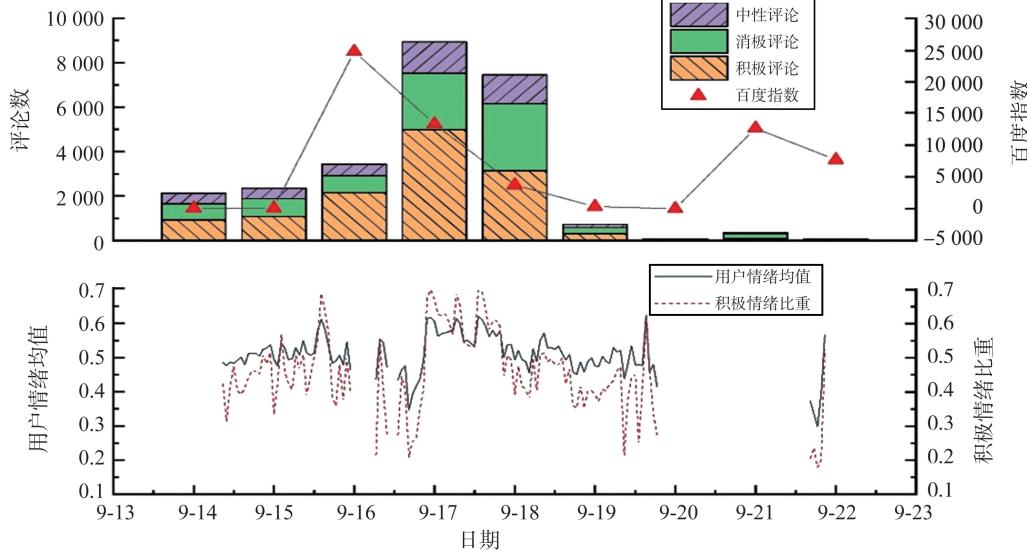


图 3 台风“山竹”期间用户情感趋势与百度指数趋势

Fig.3 User emotion trend and Baidu index trend during typhoon “Mangkhut”

用两种方法表示用户情绪,一种是情感分类后积极情绪评论、消极情绪评论与中性情绪评论所占所有评论的比例,第二种利用情绪指数来进行表示。如果以小时作为时间分辨率进行分析,为避免偶然误差,只选择每小时微博数目大于 10 的时段进

## 2 结果与分析

### 2.1 “山竹”台风事件微博用户情感趋势

以天作为时间分辨率,以微博数目作为讨论热度<sup>[20]</sup>,对评论数据进行分析,分析结果如图 3 所示。可以发现 9 月 13 日之前,尚未确定台风是否会影响我国时,山竹话题几乎没有用户讨论,这段时间属于网络事件传播的潜伏期。15 日台风登陆菲律宾,由于基本确定台风会从我国登陆,舆情热度急剧攀升,当天共有 2 357 条评论,事件进入发酵期。16 日中央气象台发布台风红色预警,下午 17 时台风登陆广东,话题讨论爆炸式增长,此时百度指数达到最高点,山竹相关话题浏览量达到了 24 749 次。在 9 月 17 日(星期天),台风山竹的微博讨论达到顶峰,事件进入爆发期,当天共计有 8 984 条讨论,其中积极评论远多于消极评论。随着 17 日台风山竹的退散和工作日的开始,话题讨论逐渐消失,在 21 日有一个小高峰,百度指数为 12 678,此段时间获取到 353 条微博评论,消极评论居多,此段时间属于舆情事件的反复期。22 日之后话题讨论完全退散,事件进入消亡期。

微博话题讨论与百度指数相比则稍显滞后性,体现了媒体对于公众注意力的引导,本次台风事件属于大众媒体导入型事件而非网络首发型事件<sup>[10]</sup>。

行绘制。其中纵轴代表评论用户的平均情绪指数和用户积极情绪的比重。由图 3 可知,9 月 16 日晚与 9 月 21 日,用户情绪有两个明显波动。

16 日 17 时两项指标分别达到当日最低点,情绪均值为 0.346 11,积极情绪比重为 0.208 33。21 日 19

时达到整个事件最低点,情绪均值为0.308 85,积极情绪比重为0.166 67。结合“山竹”在16日17时前后,由强台风级在广东省江门市台山沿海登陆的报道可以判断,台风的登陆对微博用户评论存在明显的负面影响。此外,无论哪种方法都可以看到21号用户情绪最为消极。查询话题分类结果并结合话题发生的时间,不难推断出21号发生的“台风救援队返程被卡湖南收费站”事件,对舆情有着很明显的负面影响。应急管理中心需要对此类事件做好预案工作,在灾害来临时以及灾害来临后的一段时间内,选择适当的舆情干预节点来管控网络舆情的波动。一旦侦听到公众情绪下滑,应急管理中心或相关部门应及时干预,如发布灾害通报、及时辟谣等,避免公众受到恶意引导,从而降低社会治理成本。

对两种情绪分析方法进行相关性检验,用户的平均情绪指数和用户积极情绪比重具有很强的相关性,相关系数为0.927 1。用户情绪均值指标更稳定,更容易检测异常点。

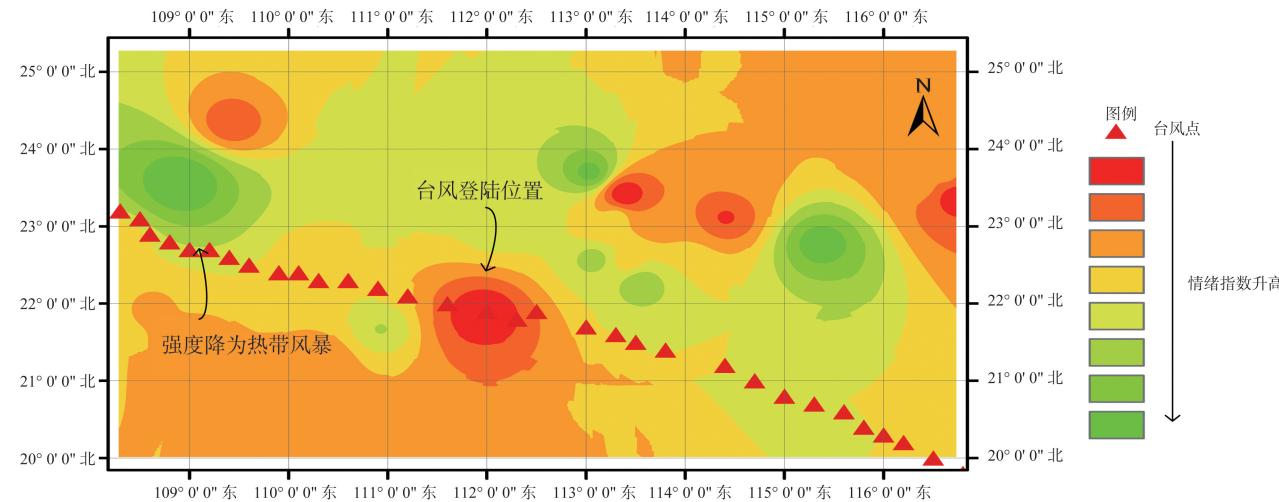


图4 台风“山竹”话题微博情绪插值图  
Fig.4 Sentiment interpolation map of typhoon “Mangkhut” topic Weibo

除了统计参与讨论用户所在地,还统计了广东省各地级市被提到的次数。依据广东省行政划分,统计每一个城市在本次事件语料库中出现的次数<sup>[21]</sup>,其中广州出现了277次,深圳出现了335次,此外还有珠海、湛江等城市被用户讨论超过了200次。如表2所示,深圳、广州、珠海、湛江、阳江、茂名、汕头等城市提及次数较多,大体呈现为南部城市出现频率较高,北部城市出现频率较低的趋势。即使考虑经济与人口的差异,通过地域词词频与该地户籍人口数量的比例归一化后,这些城市依然出现频率较高。语料库中出现频率较高的区域与本

## 2.2 “山竹”台风事件地理统计分析

为使微博舆情态势更加直观可见,对台风“山竹”事件进行地理统计分析,以某个区域用户发博量作为话题讨论热度。台风山竹的话题讨论涉及全国各地用户,广东用户讨论尤为活跃,不仅区域分布广,并且讨论热度高。所在地为广东广州的用户有935人,深圳市有712人,北京市有447人,成都市有344人,参与讨论达到200人以上的城市还有南宁、上海、杭州、西安、武汉与东莞,基本上都是经济繁荣、人口稠密的大城市或者本次台风受灾区域内的城市。

选取讨论较多的省份广东作为研究示例,以城市用户情绪均值作为区域情绪指数,其中阳江市情绪指数为0.484 037,汕头市为0.494 522,情绪指数最高的是汕尾市0.586 951与清远市0.580 898。为了结果更加直观,对其进行反距离加权法插值分析,效果如图4,呈现内陆情绪值高,沿海情绪值低的趋势,台风登陆位置的阳江市情绪指数最低,随着台风强度的降低,台风路径上的内陆城市情绪指数稍有上升。

次受灾最严重的区域存在一定的吻合。

## 2.3 城市社团挖掘

从获取到的微博文本中检索所在地为广东省用户(Source),统计其微博内容中含有城市信息的条目(Target),得到有向链接信息523条,再进行合并得到链接权重Weight。如某所在地广东河源市用户发博,“东莞这边雷刚打的巨响”,此时便建立一个有向链接,由河源指向东莞。将有向链接合并后得到100条边数据,21个点位数据(代表广东21个地级市)。利用Gephi软件Modularity模块进行社团检测,将21个城市分为了3个社团,同一社团内

用户之间的互动会比在不同社团之间更加频繁<sup>[22]</sup>,模块度计算结果为 0.161,由于广东省已经是高度划分的社团种类,城市之间联系密集,模块度不是很大。

用节点大小代表城市的度,即考虑到地区用户发微博的数量,又考虑到地区在此次台风期间被谈及的次数。由于生成的是有向图,顺时针代表图的连接方向,如广州深圳之间较粗的一条代表广州用户谈及深圳,较细的一条代表深圳用户讨论广州。其中广州用户 56 次提到广州,17 次提到湛江,深圳与阳江各涉及到 8 次。深圳用户有 82 次提到深圳,8 次提到湛江,广东省各个地区的网络社团关系如图 5 所示。

广州用户微博讨论范围几乎涵盖广东省所有地级市,且同一社团用户具有一定的地理相关性。将图 5 城市节点的入度与出度统一纳入考虑,节点活跃度较高的城市有湛江、茂名、阳江、江门、中山、珠海、佛山、深圳、东莞、广州、惠州、揭阳、潮州、汕头,基本上都是沿海城市。云浮、肇庆、韶关、清远、梅州、汕尾、河源等 7 个城市划为平静区域,平静区

表 2 台风“山竹”灾情验证  
Table 2 Verification of typhoon “Mangkhut” disaster

城市列表	情绪指数	发博人数/个	词频/次	情绪判断	发博人数判断	词频判断	节点活跃度判断
广州	0.536 3	935	277	正确	正确	正确	正确
深圳	0.536 9	713	335	正确	正确	正确	正确
佛山	0.531 4	154	57	正确	正确	正确	正确
东莞	0.524 3	202	76	正确	正确	正确	正确
江门	0.550 3	92	29	错误	正确	正确	正确
惠州	0.505 4	83	21	正确	正确	正确	正确
汕头	0.494 5	115	58	正确	正确	正确	正确
潮州	0.525 9	41	8	正确	错误	错误	正确
汕尾	0.587 0	34	15	错误	错误	正确	错误
茂名	0.544 6	87	65	错误	正确	正确	正确
阳江	0.484 0	75	65	正确	正确	正确	正确
清远	0.580 9	23	2	错误	错误	错误	错误
云浮	0.540 9	25	3	错误	错误	错误	错误
梅州	0.525 9	29	7	正确	错误	错误	错误
准确率/%				64.29	64.29	71.43	71.43

依据情绪指数判断准确率为 64.29% 略低于词频判断与地区活跃度判断的 71.43%,推测是由于地理尺度划分太大,某些城市的小部分区域而非大范围区域受到了灾害影响;不同地域情绪存在波动;用户情绪受其他因素的干扰;某些地区用户样本量过小进而放大了随机误差。除此之外,由于微博舆情与热点话题的传播相关度较大,情绪指数对于重灾区、灾情事件的判断效果较佳。在国家应急管理

域大都位于内陆或者广东东北部,远离本次台风灾害受灾最为严重的西南部,恰好躲开了台风路径。由此可以发现,微博用户活跃区域与灾情情况存在一定联系。

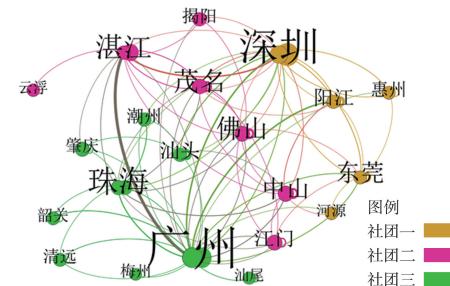


图 5 城市舆情网络社团图  
Fig.5 City lyrics network association map

#### 2.4 实际灾情与舆情验证

根据广东省民政厅的统计,广州、深圳、汕头、佛山、梅州、惠州、汕尾、东莞、江门、阳江、茂名、清远、潮州、云浮等 14 个城市受灾较为严重。利用情绪指标,用户讨论频度,词频指标以及综合考察入度与出度的节点活跃度判断指标,对广东省 21 个城市进行分析,以指标是否在前 14 位作为判断条件,见表 2。

平台上查询,广东汕头、揭阳,9月 16 日出现船只险情,深圳、珠海、江门、湛江、阳江受灾情况较为严重。其中汕头情绪指数为 0.494 523, 揭阳为 0.530 609, 均被划入了受灾区域,通过情绪指数可以监测灾害过程中发生的某些突发事件。

根据美国国家海洋和大气管理局(National Oceanic and Atmospheric Administration, NOAA)全球气象数据,将本次台风过境期间(2018 年 9 月

16—9月17日)广东省38个气象站监测到的24 h内最大降雨量进行反距离加权插值,参考中央气象局降水划分等级标准进行分级设色。如图6所示,广东省大部分地区发生暴雨天气,西南部的阳江市24 h内降水总量超过250 mm,属于特大暴雨降水。

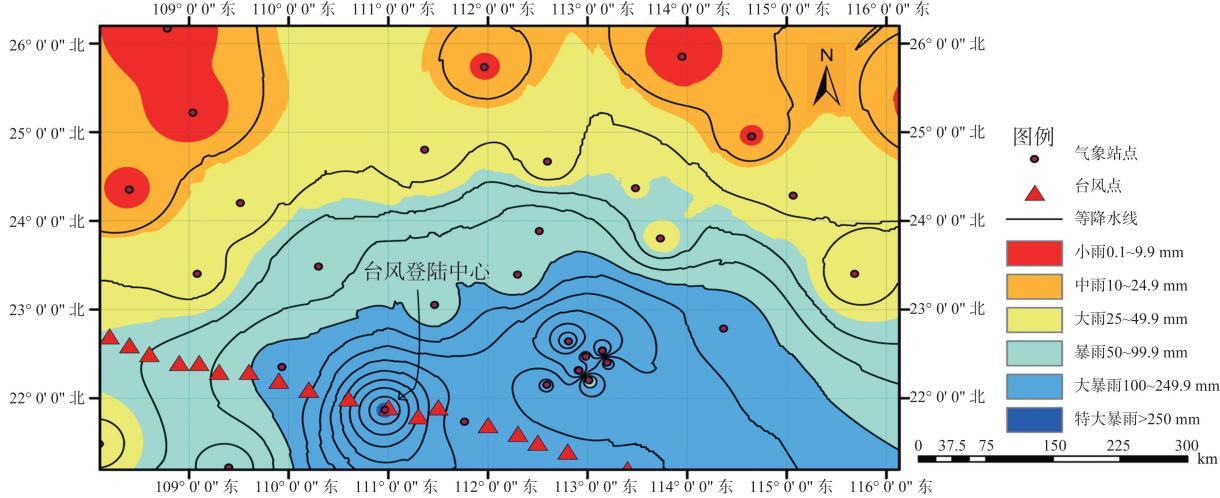


图6 台风过境期间24 h最大降水插值图

Fig.6 24-hour maximum precipitation interpolation map of typhoon transit

情绪指数对于重灾区的判断效果较佳,然而也出现了识别错误的情况,如韶关市并不是受灾特别严重的地区。可以通过改进灾害情绪识别方法、构建台风灾害语料库、构建情感词典、增加数据量等方法予以修正<sup>[23]</sup>。词频判断与节点活跃度判断的准确率达到71.43%,可以为受灾区域的判断提供较好的参考。用户讨论频度跟城市人口基数相关性较大,对于受灾区域判断效果不明显。

### 3 结论

(1)在微博数据文本清洗、位置获取、数据结构化等预处理的基础上,根据贝叶斯模型计算舆情信息中的情感值和情感比例。(2)百度指数与舆情热度随时间的对比分析,发现了本次事件微博舆情的滞后性。(3)采用LDA模型对舆情信息进行聚类分析,挖掘台风期间公众主要关注的话题,发现台风登陆事件与救援车辆被卡收费站事件对情感倾向具有显著的负面影响。(4)利用Louvain算法检测社团结构,对不同城市用户进行关联性分析和社团模块度检测,将广东省划分为活跃地区与平静区域,发掘了模块度为0.161的三个广东城市社团。(5)依据城市用户情感,城市降雨数据进行了插值分析,其中阳江市是用户情感值最低、降雨量最高的城市,也是本次台风事件中受灾最严重的城市。(6)利用城市词频、用户所在地、城市活跃度、城市

阳江市是台风山竹正面袭击的地点,情绪指数最低,其中阳江阳春市是整个台风过程中受影响最为严重的区域,遭遇了1981年以来最为严重的洪涝灾害,直接经济损失高达12.3亿元。

用户情感等指标进行了地理统计分析,检验了发掘受灾城市的准确率。

在台风灾害期间,可以利用实时产生的微博评论数据进行分析,结合降雨量、媒体报道、气象条件观测等多源数据,进而快速确定受灾区域。后续将进行热点话题的空间扩散与地域关联性研究<sup>[24]</sup>,挖掘意见领袖<sup>[25]</sup>在灾害事件中的话题传播作用,剖析不同种类用户在关注话题上的区别,以及网络谣言的产生与消亡过程。

### 参考文献:

- [1] 刘哲, 张鹏, 刘南江, 等. “一带一路”中国重点区域自然灾害特征分析[J]. 灾害学, 2018, 33(4):65-71.  
LIU Zhe, ZHANG Peng, LIU Nanjiang, et al. Characteristics of natural disasters in key regions of One Belt One Road initiative[J]. Disaster Science, 2018, 33(4):65-71.
- [2] NING Xiaodong, YAO Lina, WANG Xianzhi, et al. Calling for response: automatically distinguishing situation-aware tweets during crises [C]//The 13th International Conference on Advanced Data Mining and Applications. Singapore, Singapore: Springer, 2017.
- [3] NAIR M R, RAMYA G R, SIVAKUMAR P B. Usage and analysis of Twitter during 2015 Chennai flood towards disaster management[C]//7th International Conference on Advances in Computing & Communications (ICACC-2017). Cochin, India: Elsevier, 2017.
- [4] ALFARRARJEH A, AGRAWAL S, KIM S H, et al. Geo-spatial multimedia sentiment analysis in disasters [C]//The 4th IEEE International Conference on Data Science and Advanced Analytics 2017. Tokyo, Japan:

- IEEE, 2017.
- [5] 唐晓波,向坤. 基于 LDA 模型和微博热度的热点挖掘 [J]. 图书情报工作,2014,58(5):58-63.  
TANG Xiaobo, XIANG Kun. Hotspot mining based on LDA model and microblog heat[J]. Library and Information Service, 2014, 58(5):58-63.
- [6] 阮光册. 基于 LDA 的网络评论主题发现研究 [J]. 情报杂志,2014,33(3):161-164.  
RUAN Guangce. Topic extraction research of net reviews based on Latent Dirichlet Allocation[J]. Journal of Intelligence, 2014, 33(3):161-164.
- [7] CHOI S, BAE B. The real-time monitoring system of social big data for disaster management [M]. Berlin, Germany: Springer, 2015.
- [8] 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.  
BAI Hua, LIN Xunguo. Research on social media disaster incident detection system based on Chinese short text classification[J]. Disaster Science, 2016, 31(2):19-23.
- [9] 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.  
CHEN Zi, GAO Tao, LUO Nianxue, et al. Discussion on the effectiveness of social media reflecting the spatial and temporal distribution of natural disasters[J]. Surveying Science, 2017, 42(8):44-48.
- [10] 王心瑶,郝艳华,吴群红,等. 社交媒体环境下 H7N9 事件网络舆情演变与比较分析 [J]. 中国公共卫生,2018,34(9):1232-1236.  
WANG Xinyao, HAO Yanhua, WU Qunhong, et al. Evolution and comparative analysis of H7N9 event network in social media environment[J]. China Public Health, 2018, 34(9):1232-1236.
- [11] 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.  
LIANG Chunyang, LIN Guangfa, ZHANG Mingfeng, et al. Effectiveness of social media data to reflect the temporal and spatial distribution of typhoon disasters[J]. Journal of Earth Information Science, 2018, 20 ( 6 ): 807-816.
- [12] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(4/5):993-1022.
- [13] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. Proc Natl Acad Sci U S A, 2004, 101(Suppl 1): 5228-5235.
- [14] CAO Juan, TIAN Xia, LI Jintao, et al. A density-based method for adaptive LDA model selection[J]. Neuro Computing, 2009, 72(7):1775-1781.
- [15] 周咏梅,阳爱民,林江豪. 中文微博情感词典构建方法 [J]. 山东大学学报(工学版), 2014, 44(3):36-40.  
ZHOU Yongmei, YANG Aimin, LIN Jianghao. Construction method of Chinese Weibo emotional dictionary [J]. Journal of Shandong University ( Engineering Sci-
- ence ), 2014, 44(3):36-40.
- [16] 朱嫣岚,闵锦,周雅倩,等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报,2006(1):14-20.  
ZHU Yanlan, MIN Jin, ZHOU Yaqian, et al. Calculation of lexical semantic tendency based on HowNet[J]. Journal of Chinese Information Processing, 2006(1):14-20.
- [17] 杨振山,蔡建明. 空间统计学进展及其在经济地理研究中的应用 [J]. 地理科学进展, 2010, 29(6):757-768.  
YANG Zhenshan, CAI Jianming. Progress in spatial statistics and its application in economic geography research [J]. Progress in Geography, 2010, 29(6): 757-768.
- [18] NEWMAN M. Detecting community structure in networks [J]. European Physical Journal B, 2004, 38 ( 2 ): 321-330.
- [19] 李沐南. Louvain 算法在社区挖掘中的研究与实现 [D]. 北京:中国石油大学(北京), 2016.  
LI Munan. Research and implementation of Louvain algorithm in community mining[D]. Beijing: China University of Petroleum ( Beijing ), 2016.
- [20] 赵燕慧,路紫,张秋娈. 多类型微博舆情时空分布关系的差异性及其地理规则 [J]. 人文地理,2018,33(1): 61-69.  
ZHAO Yanhui, LU Zi, ZHANG Qiuluan. The differences of temporal and spatial distribution relationships of multi-type Weibo and their geographical rules [ J ]. Human Geography, 2018, 33(1):61-69.
- [21] 李静. 基于 LDA 的微博灾害信息聚合 [D]. 武汉:武汉大学, 2018.  
LI Jing. LDA-based microblog disaster information aggregation [D]. Wuhan: Wuhan University, 2018.
- [22] SUN Penggang, GAO Lin, HAN Shanshan. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks [J]. Information Sciences, 2010, 181(6):1060-1071.
- [23] 刘超然. 在线新闻网民评论情感倾向性分析及可视化研究 [D]. 哈尔滨:哈尔滨工业大学, 2018.  
LIU Chaoran. Online news netizens comment on emotional orientation analysis and visualization[D]. Harbin: Harbin Institute of Technology, 2018.
- [24] 杨腾飞,解吉波,李振宇,等. 微博中蕴含台风灾害损失信息识别和分类方法 [J]. 地球信息科学学报, 2018, 20(7):906-917.  
YANG Tengfei, XIE Jibo, LI Zhenyu, et al. Identification and classification of typhoon disaster loss information in Weibo[J]. Journal of Earth Sciences, 2018, 20 ( 7 ): 906-917.
- [25] 王袆珺,张晖,李波,等. 一种基于话题演化的意见领袖发现方法 [J]. 山东大学学报(工学版), 2016, 46(2): 35-42.  
WANG Weijun, ZHANG Hui, LI Bo, et al. A method for opinion leader discovery based on topic evolution[J]. Journal of Shandong University ( Engineering Science ), 2016, 46(2):35-42.