doi: 10.3969/j.issn.0490-6756.2020.02.031

基于微博数据的“新冠肺炎疫情”舆情演化时空分析

陈兴蜀[1,2], 常天祐[3,4], 王海舟[1,2], 赵志龙[3,4], 张 杰[1]
(1. 四川大学网络空间安全学院, 成都 610207; 2. 四川大学网络空间安全研究院, 成都 610065;
3. 四川大学吴玉章学院, 成都 610207; 4. 四川大学计算机学院, 成都 610207)

摘 要: 本文依托 2020 年 1 月 1 日至 2 月 29 日期间共计 6 万条新浪微博博文与 1.5 万条微博热门评论,基于分布式爬虫技术、分布式数据库系统、SnowNLP 情感分析模型以及 K-Means 文本聚类算法,对与“新冠肺炎疫情”相关的话题展开舆情分析,可视化地展现本次疫情事件中网络舆情的时空演化过程. 在时间维度层面,通过文本聚类与情感分析,发现网民对于此次肺炎疫情的态度大致经历了三个阶段,即起伏不定的紧张焦虑期、缓慢攀升的团结振作期以及波动很小的自信平稳期,总体上呈现积极大于消极、正面大于负面的情绪状态. 在空间维度层面,通过地理统计分析,发现疫情最严重地区网民评论人数最多,同时情感值也最低.

关键词: 新浪微博; 新冠肺炎疫情; 分布式爬虫; 情感分析; 文本聚类; 地理统计分析
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2020)02-0409-08

Spatial and temporal analysis on public opinion evolution of epidemic situation about novel coronavirus pneumonia based on micro-blog data

CHEN Xing-Shu[1,2], *CHANG Tian-You*[3,4], *WANG Hai-Zhou*[1,2], *ZHAO Zhi-Long*[3,4], *ZHANG Jie*[1]
(1. College of Cybersecurity, Sichuan University, Chengdu 610207, China;
2. Cybersecurity Research Institute, Sichuan University, Chengdu 610065, China;
3. Wu Yuzhang College of Sichuan University, Chengdu 610207, China;
4. College of Computer Science, Sichuan University, Chengdu 610207, China)

Abstract: Relying on 60 thousand blogs and 15 thousand hot blog reviews in Sina micro-blog from January 1st to February 29th in 2020, this article launches the analysis of the public opinion to the topic about novel coronavirus pneumonia based on distributed crawler technology, distributed database system, SnowNLP sentiment analysis model and K-Means algorithm. This analysis can show visually the spatial and temporal evolution process of Internet public opinion in the events of this epidemic situation. On spatial dimension, the netizens' attitude towards this pneumonia epidemic has roughly gone through three periods. The first period appears in the shape of bigger fluctuation which presents as tension and anxiety. The second period appears in the shape of rising slowly which presents as unity and excitement. The third period appears in the shape of slight fluctuation which presents as confidence and stability. On the whole, it shows the emotional state that positive is greater than negative and Optimism is greater than pessimism. On spatial dimension, we find that the area which has the most serious epidemic has the

收稿日期: 2020-03-03
基金项目: 四川省科技厅新型冠状病毒疫情防控科技攻关项目(2020YFS0007); 四川大学新冠肺炎应急项目(2020scunCoV 应急20012); 四川大学大学生创新创业计划(C2020109217)
作者简介: 陈兴蜀(1968-), 女, 教授, 博士生导师, 博士研究方向为云计算/大数据安全体系, 威胁检测, 开源情报分析. E-mail: chenxsh@scu.edu.cn
通讯作者: 常天祐. E-mail: 2965841037@qq.com

most comments and the lowest emotional value through geographical statistical analysis.

Keywords: Sina micro-blog; Novel coronavirus pneumonia; Distributed crawler; Sentiment analysis; Text clustering; Geostatistical analysis

1 引 言

2019 年 12 月 31 日,武汉市卫健委发布通报称,该市近期部分医疗机构发现接诊的多例肺炎病例与华南海鲜市场有关联. 这一通报引发了较为广泛的社会关注,舆论的关注点主要集中于“华南海鲜市场”、“肺炎”、“传染”等词. 2020 年 1 月 22 日,国务院新闻办公室举行新闻发布会,1 月 23 日,湖北省人民政府新闻办公室举行新闻发布会,介绍新型冠状病毒感染的肺炎防控工作的有关情况,舆情不断升温,出现多次舆情高峰. 在 2020 年 1 月 1 日至 2020 年 2 月 29 日期间,“专家称武汉不明原因的病毒性肺炎可防可控”、“钟南山指出新型冠状病毒具有传染性,已经出现人传人现象”、“各地医护人员驰援武汉”、“首个潜在治疗新冠肺炎药物获批上市”等成为新浪微博热议话题. 关于新冠肺炎疫情,网民不仅讨论热度高,而且持续时间长. 无论是核心话题,还是时空二维的动态分析方法,对“新冠肺炎疫情”舆情的研究,都具有一定的理论意义. 就现实层面而言,可视化地表达“新冠肺炎疫情”舆情的时空演化过程,不仅可以客观地展示民众对此次疫情在态度上的变化性,而且还可以形象地反映不同地区民众对此次疫情在情绪上的差异性,有助于各级党政机关及时而准确地掌握舆情动态、回应民众关切,从而提高应对能力. 此外,根据客观疫情与民众舆情的数理相关性,推理出未来疫情发展的时段走势与区域特征,可以为党政机关、企事业单位科学预判疫情趋势提供参考,从而根据不同时段情况,制定相应的疫情防控措施;针对不同区域情况,完善差异化防控策略. 因此,对“新冠肺炎疫情”舆情进行时空演化分析具有重要的应用价值.

如今,国内外学者对于舆情的研究大多是基于文本数据,而忽视了文本背后的时间信息与地理位置信息,很少将两者相结合来对舆情进行研究. 丁杰等人[1]将网络新闻及论坛、BBS 上的帖子依关键词搜索,并依“事件”分类,让管理者通过“阅读时间”了解正在发生或者已经发生的事件,并且自动持续追踪事件发展的功能,以协助管理者快速且完整地了解事件全貌,并且采用网页清理技术来减少数据量. 洪小娟等人[2]构建在. NET 平台下基于 Entity Framework 模型的网络舆情检测系统的 C/S 和 B/S 框架体系,系统应用马尔科夫链实现计算未来发展估计. 李然等人[3]介绍了文本情绪分析在不同场景下应用,整理归纳了文本情绪分析的主流方法,并对其进行了细致的介绍和分析对比. 陈兴蜀等人[4]基于发现的热点话题,提出了基于在线 LDA(OLDA)话题模型的论坛热点话题演化跟踪模型(HTOLDA),从而更加有效地对论坛中的热点话题进行演化跟踪.

情感分析作为舆情研究中极为重要的一部分,近些年不断改善,与之相关的应用研究也在蓬勃发展. 明弋洋等人[5]针对基于关键词字符匹配和粗粒度情感分析方法的传统不良信息检测方法准确率低的问题,提出一种基于短语级情感分析的不良信息检测方法. 胡思才等人[6]根据经典的特征选择方法在中文情感评论文本中应用的缺陷和不足,提出了一种改进的中文情感特征选择方法. Isa Maks 等人[7]为情感分析与观点挖掘提出了词典模型,该模型包括与观点挖掘和情感分析相关的语义类别的分类,并提供了用于识别态度持有者和态度的极性以及描述文本中涉及的不同行为者的情感的方法. Nazan Öztürk 等人[8]针对叙利亚难民危机问题,运用情感分析和文本挖掘,对 Twitter 上公众谈论内容进行分析,发现英语报道和土耳其语报道情绪的差异性. 凌海彬等人[9]提出一种多特征融合的图文微博情感分析方法,将对情感具有很好指示作用的内容特征和用户特征与微博句子进行融合,设计特征层和决策层融合的方法,将文本和图片情感分类模型进行融合.

本文的贡献在于,将 jieba 分词模型、SnowNLP 情感分析模型、K-Means 文本聚类算法、地理可视化技术引入到新冠肺炎疫情的舆情研究中,充分挖掘了文本的时间与空间特点,不仅分析了情感变化的时间曲线,而且将用户位置纳入研究,发掘了检索文本中的地理信息,并将情感分析与用户地理信息结合,获得不同地区人们对于新冠肺炎疫情的情感值. 最后,综合时空分布特点,演绎出了对于未来疫情发展的科学预测,为未来的灾害应对提供较完备的思路与方法.

2　研究方法与手段

2.1　方法流程

本文围绕“新冠肺炎疫情”话题对微博数据进行抓取，将数据进行文本清洗后存入数据库. 再通过 K-Means 文本聚类算法进行话题分类，之后借助 SnowNLP 对提取到的话题进行情感趋势分析和地理统计分析，最后，通过地理分布数据与新闻报道相结合判断出疫情较为严重的地区. 论文方法流程图如图 1 所示.

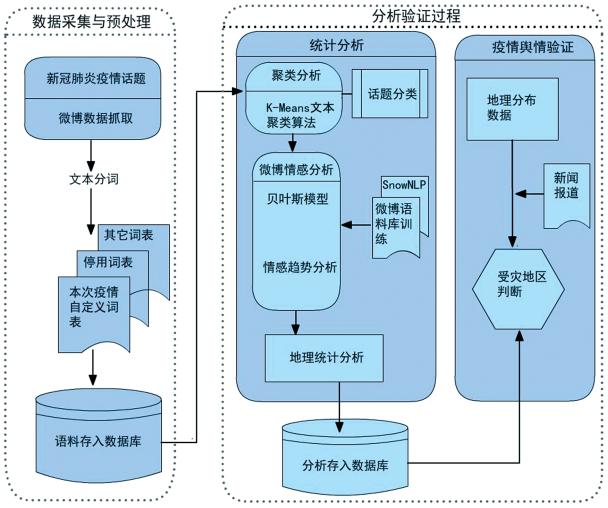

图 1　论文方法流程图
Fig. 1　Flow chart of paper method

2.2　数据采集

2.2.1 *数据获取*　微博已经成为人们生活中重要的信息来源，而新浪微博作为全球最大中文社交网络平台，它的活跃用户已超过 4 亿，而且数据开放程度较高，思想观点表达较为丰富，故此次研究利用新浪微博平台，以自然语言处理的方法对“新冠肺炎疫情”的相关微博展开研究. 在获取数据的方法中，使用 Python 语言开发的 Scrapy 开源框架，引入 Redis 开源框架来实现多机分布式爬虫[10]. Scrapy-Redis 的分布式策略是 Slaver 端从 Master 端获取任务(Request 和 URL)进行数据抓取，在抓取过程中将新产生的 Request 提交给 Master 端处理，Master 端只有一个 Redis 数据库，负责将未处理的 Request 去重和任务分配，将处理后的 Request 加入待爬队列，并且存储爬取的数据，如图 2 所示. 值得一提的是，Scrapy-Redis 爬虫框架中的 Duplication Filter 组件利用 Redis 中队列的不重复性，巧妙地实现了带爬取 URL 队列的不重复. 接下来在模拟微博登录的过程中，针对构造访客 Cookie 的方案设计实现了高可用代理池模块，进一步提高了数据采集效率[11].

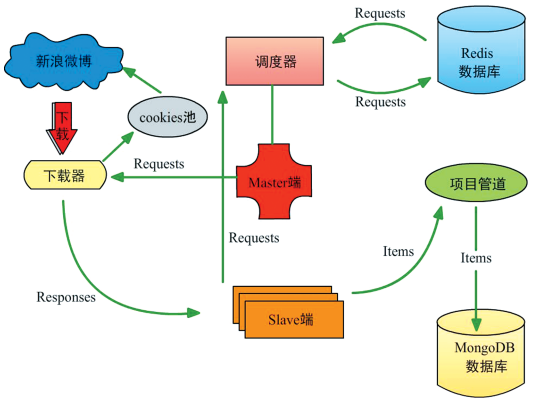

图 2　Scrapy-Redis 爬虫框架
Fig. 2　Scrapy-Redis crawler framework

自 2019 年 12 月 31 日武汉卫健委第一次发布“通报”称发现 27 例肺炎病例，至 2020 年 2 月 3 日全国新冠肺炎确诊病例突破两万人，再到 2 月 29 日国务院联防联控机制新闻发布会上国家卫健委新闻发言人通报疫情情况称：“截至 2 月 28 日 24 时，据各省(区、市)和新疆生产建设兵团报告：现有确诊病例 37 414 例，其中重症病例 7 664 例，累计治愈出院病例 39 002 例.”在这期间，关于新冠肺炎的微博讨论较为集中. 需要说明的是，关于此次肺炎的通报在名称上存在一个变动过程，2019 年 12 月 30 日～2020 年 1 月 10 日称不明原因肺炎、不明原因病毒性肺炎，2020 年 1 月 11 日～2 月 7 日称新型冠状病毒感染的肺炎，2 月 8 日，国务院联防联控机制新闻发布会通报了国家卫健委关于新冠病毒感染的肺炎暂命名为“新型冠状病毒肺炎”，简称“新冠肺炎”的信息，因此，以“肺炎”为关键词进行搜索，可以更全面地抓取到各时段关于本次疫情的议题. 本文对 2020 年 1 月 1 日至 2020 年 2 月 29 日之间(共计 60 天)以“肺炎”为关键词的新浪微博进行时段分割，以每日为一小段，共计 60 小段，每一小段按热度高低抓取微博 1 000 条，共计 6 万条微博博文，组成本文进行数据研究的基础单位. 每条微博的抓取内容包括微博博文、时间戳、用户 ID、点赞数、转发数和评论数等.

2.2.2　*数据存储*　由于从网页爬取的数据每次都默认存储在 Redis 的数据库中，每次启动 Redis 库时，都会将本机之前存储的数据加载到内存中. 如果数据量较大，则内存消耗会比较严重. 因此研究选择持久化的数据库 MongoDB 来进行数据的存

储,并利用MongoDB的服务器进行远程连接与数据共享,实现了分布式数据库系统的搭建.需要处理数据时,再利用Python中的pymongo第三方库与MongoDB的数据库进行对接,根据对数据所属的数据库和文档进行操作,从而达到对微博信息的查重操作,以便进行下一步的自然语言处理.

2.3 研究方法

2.3.1 文本聚类 在文本分词方面,文章采用jieba分词对采集到的文本进行分词工作,jieba分词采用了基于Trie树结构实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(DAG),再采用动态规划查找最大概率路径,找出基于词频的最大切分组合.对于未登录词,采用基于汉字成词能力的HMM模型,使用Viterbi算法,得到一个概率最大的BEMS序列,按照B打头、E结尾的方式,对待分词的句子重新组合,就得到了分词结果.但是由于jieba分词词库的本身有限性,难以处理一些较难的语句.本研究利用jieba分词提供的load_userdict()函数对jieba的分词语料库进行了优化,用《四十万汉语词库》、《四十万搜狗大词库》、《网络用语词库》等词库对jieba分词进行了分词训练,从而增加其分词的精确性.再通过TF-IDF算法构造文本向量矩阵,由于文本向量矩阵极为稀疏,因此利用pca降维,构造出较为稠密的矩阵.最后通过K-Means的文本聚类算法,得到话题聚类的结果[12,13].

2.3.2 情感分析 情感分析,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程.互联网(以新浪微博为代表的博客、论坛以及社会服务网络)上产生了大量的用户参与的、对于诸如人物、事件、产品等有价值的评论信息,从中提取的用户信息属于一分为二的极性分析,即“赞同”和“否定”.

对于微博博文的内容,研究采用Python的第三方库SnowNLP进行情感分析,SnowNLP的情感判断过程是:首先,读取已经分好类的文本neg.txt和pos.txt,再对所有文本进行分词、去停用词,从而计算每个词出现的频数.通过贝叶斯定理计算正面负面先验概率p(pos)和p(neg),对要进行判断的文本分词,计算每个词的后验概率p(词|neg)和p(词|pos),最后,选择计算出的概率较大的类别(正或负).可以看出,这个算法最重要的就是语料库的选择与分词.由于SnowNLP自身提供的语料库具有滞后性与局限性,研究特地采用微博文本语料库对其进行训练,并将其分词函数seg()替换成了训练好的jieba分词函数cut().利用这种方法,可以对每一条微博进行情感分析,得到每天的平均情感数值,再对每天的平均情感数值进行基于时间的排序,用matplotlib进行绘制,从而更加清晰、形象化地反映网民对于新冠肺炎疫情的态度变化趋势.

2.3.3 词云生成 用wordcloud和matplotlib对清洗过的所有时间段微博文本进行高频词的统计,并生成词云和高频词排序表,更直观地展示疫情发展不同阶段人们最为关心的话题的变化,并分析总结出人们的心态变化.

2.3.4 地理统计分析 地理统计又称空间统计,是对一定区域内的地理要素的数量、种类等情况进行汇总,反映地理要素的空间分布情况[14].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.

研究采用常规统计分析的方法,通过统计1月1日~2月29日来自全国34个省级行政区的网民在微博上发表评论的数量,通过情感分析计算出各个地区网民的平均情感数值,再结合数据可视化工具pyecharts,将全国各省级行政区的微博评论人数图与微博评论情感值图绘制出来.

在舆情演化的时空分析研究领域,如今的研究大多停留在自然灾害和局部区域分析统计这两个方向.张岩等人[15]抽取了台风“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.曹彦波[16]基于新浪微博数据,通过数据清洗、分类和挖掘,分析2018年8月13日和14日云南省通海县2次5.0级地震舆情信息时空演变规律.本研究的创新之处在于将时空分析的研究视角转向了公共卫生安全事件领域,并在空间上扩大到了对全国范围的舆情进行分析,最后结合全国各个省级行政区域情感分析数值,对疫情严重程度不同的地区做了疫情与舆情的数理相关性研究.

3 研究结果

3.1 新冠肺炎疫情微博用户情感趋势

SnowNLP情感分析得到的数值在0到1之

间，当结果大于 0.5 时，情感较为积极，越接近 1，情感越正面；当结果小于 0.5 时，情感较为消极，越接近 0，情感越负面. 研究对每天的微博情感分析结果做了取平均值(图 3)处理，根据图 3 的数据显示，从 1 月 1 日～2 月 29 日期间内，网民对于“新冠肺炎疫情”的态度整体上是趋于正面的. 在这 60 天中，有 55 天的情感分析数值大于 0.5，即趋于正向，只有 5 天的数值小于 0.5，即趋于负向. 从曲线走势来看，这段时间网民对“新冠肺炎疫情”的态度大致可以分为三个阶段. 第一阶段是 1 月 1 日～1 月 20 日，此段时间内网民情绪波动较大，较不稳定，处于正负面情绪交替出现状态；第二阶段是 1 月 20 日～2 月 3 日，此段时间网民情绪由负面向正面转化，且逐步上升；第三阶段是 2 月 3 日～2 月 29 日，这段时间网民的情绪稳定在正面的、积极的状态，波动很小. 我们运用话题聚类的方法，采用了 TF-IDF 算法构造文本向量矩阵，再利用 pca 降维，构造出较为稠密的矩阵，并采用 K-Means 聚类算法分出话题文档. 最后，结合实际疫情状况，对新浪微博用户这三个阶段的态度变化做了分析和推断.

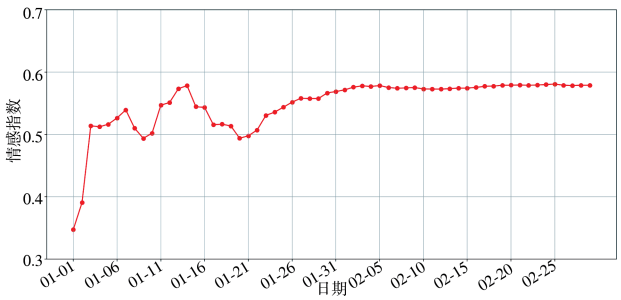

图 3　每日微博博文 SnowNLP 情感分析数值(平均值)
Fig. 3　Daily tweet SnowNLP sentimental analysis value (average)

疫情发展的第一阶段为 1 月 1 日～1 月 20 日之间波动较大的紧张焦虑期. 根据文本聚类，发现“武汉病毒性肺炎患者，发病 10 多天曾以为是感冒”与“华南海鲜批发市场休市整顿卫生，现场张贴官方公告”居于微博热议话题前两位，从图 3 可以看出，2020 年 1 月 1 日的情感分析的结果呈现最低值，约 0.34，说明面对突如其来的未知疾病风险，普通民众出现了强烈的担忧、恐惧等负面情绪反应. 然而，2020 年 1 月 3 日，情感数值攀升至 0.52，这一从负向到正向的急剧变化的原因何在？文本聚类显示，1 月 3 日居于微博热议榜首的话题是“武汉市卫健委通报病毒性肺炎情况”，具体内容包括“截至 2020 年 1 月 3 日 8 时，共发现符合不明原因的病毒性肺炎的诊断患者 44 例，其中重症 11 例”，“初步调查表明，未发现明显的人传人证据，未发现医务人员感染”. 由于感染肺炎人数较少，且未确定肺炎的种类与肺炎是否具有传染性，大多数网民认为肺炎不会人传人，紧张的心理得到很大的缓解，所以情感数值急剧上升. 1 月 9 日，微博热议话题位居前列的是“不明原因肺炎病原体为新型冠状病毒”，处于对冠状病毒可能具有传染性的担忧，情感数值出现第二个低点. 1 月 13 日，微博热议的话题是“专家称武汉不明原因的病毒性肺炎可防可控”，情感数值达到峰值，接近 0.6，说明网民的情绪转向放松. 由于新冠肺炎感染人数大量增加，民众关切度随之提高. 1 月 20 日，国家卫健委高级别专家组组长钟南山院士接受白岩松采访时指出“它(新型冠状病毒)具有传染性，已经出现人传人现象，同时医务人员也有传染，要提高警惕”，钟南山的分析、判断、提醒与建议引起人们的高度重视，“钟南山肯定新型冠状病毒肺炎人传人”成为当日微博热议话题，从图 3 可以看出，从 1 月 19 日～1 月 20 日，情感数值从正向直接降为负向. 1 月 20 日，情感分析的结果出现第三个低点，说明大多数网民意识到问题的严重性，情绪出现急剧变化，陷入警惕、担忧、焦虑状态. 总体看来，1 月 1 日～1 月 20 日情感分析数值呈现起伏不定的特征，且峰值与谷值都出现在这一时段，说明网民情绪波动较大.

疫情发展的第二阶段是 1 月 20 日～2 月 3 日之间缓慢爬升的团结振奋期. 如图 3 所示，经历了 20 日、21 日两天的负值，到 22 日之后，转变为正值，并持续上升 5 天，这是因为自 1 月 21 日开始，微博话题集中在“各地新增新型肺炎病例”、“抗击疫情”、“中国有信心打赢新型肺炎疫情攻坚战”等方面，其中 24 日的热门话题 “各地医护人员驰援湖北”、“各级政府有序开展防控措施”、“抗艾滋药物对新型肺炎治疗有效”更是给网民带来极大的鼓舞. 1 月 31 日，网民围绕“上海药物所、武汉病毒所联合发现：双黄连可抑制新型冠状病毒” 、“韩红爱心驰援武汉”、“武汉市金银潭医院 20 名新型冠状病毒肺炎患者集体出院”展开热议，又将舆情推向一个小的高潮. 如图 3 所示，1 月 31 日～2 月 3 日，情感数值在短暂平稳后继续攀升，情感数值接近峰值. 这说明在第二时段，关于新型冠状病毒感染的肺炎疫情，网民在态度上是积极向上的，在情绪上由担忧、焦虑转向团结、振作.

疫情发展的第三个阶段是 2 月 3 日～2 月 29 日波动很小的自信平稳期. 在 2 月 3 日的湖北省新闻发布会上,有关专家就疫情的相关状况答疑解惑,“无需担忧无症状患者”、“无临床证据支持双黄连可以预防和治疗新冠肺炎”等信息再获聚焦,而“同舟共济,共抗疫情! 武汉加油! 中国必胜”、“爱心守望、众志成城”等口号成为网民共同的发声,各个地区的医护人员驰援武汉取得优秀的战绩,为网民增加了抗击肺炎疫情取得胜利的信心. 2 月 13 日,新华社发布快讯,应勇任湖北省委书记,王忠林任武汉市委书记,网民对疫情防控新局面的打开充满期待. 2 月 16 日,“首个潜在治疗新冠肺炎药物获批上市”成为微博热议话题,2 月 29 日,“世卫组织专家:如果我感染了,希望在中国治疗!”居于微博话题榜首,这些都体现了网民对于中国打赢疫情防控阻击战持有坚定的信心,也为“中国力量”感到自豪. 此外,2 月中下旬,微博话题还多次谈论到日本、伊朗、韩国等其他国家的新冠肺炎确诊病例,这体现了在肺炎疫情面前,中国网民对其他疫情国家的关心. 从图 3 可以看出,随着全国各省的新增患病人数逐渐减少,疫情趋于稳定,情感数值表现为连续 27 天的正向平稳趋势. 2 月 3 日～2 月 29 日这一时段情感分析数值稳定在 0.56 以上,说明网民对于“新冠肺炎疫情”的态度积极而平稳.

总的来说,网民对于“新冠肺炎疫情”的态度大致经历了焦虑紧张期、团结振作期与自信稳定期三个阶段,总体上呈现积极大于消极,正面大于负面的情绪状态. 由情感曲线的走势可以推测,未来一段时间网民将维持积极情绪,关于新冠肺炎疫情的舆情将会趋于平稳向好态势.

3.2 新冠肺炎疫情微博用户关注话题

为进一步了解在“新冠肺炎”疫情期间网民所讨论的主要话题,研究构建从 1 月 1 日～2 月 29 日期间微博场域中与“新冠肺炎疫情”相关的词频排序,为了更直观地展示讨论的主题,文章将排名前 15 的高频关键词制作成柱状图予以展示. 如图 4 所示.

如图 4 所示,出现频度最高的是“肺炎”一词,共出现了 65 617 次之多;此外还有“新型”、“武汉”、“疫情”、“冠状病毒”、“不明”、“感染”等词汇. 这些词汇都直接反映了网民对于此次疫情状况的关注,也体现了网民对此种新型病毒感染的肺炎的高度关注,表明引发此次疫情的罪魁祸首——新型冠状病毒的传染范围之广,网民对其关注度之高.

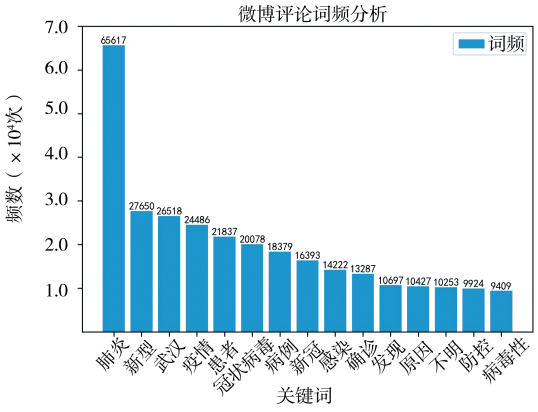

图 4 微博高频词统计柱状图

Fig. 4 Statistical histogram of high-frequency words on micro-blog

值得注意的是,关注度第三高的关键词是“武汉”,一方面,因为新冠肺炎疫情的早期扩散发生在武汉,尽管新冠病毒不一定源于武汉华南海鲜市场;另一方面,因为武汉是全国疫情灾害最严重的地区,大部分感染新冠肺炎的患者都集中在武汉. 武汉成为网民关注的焦点,与之相关的“患者”、“防控”等词汇高居前列,从侧面体现了网民对武汉人民的关心. 如何防控病毒,如何治疗患者成为网民关注的热点,同时也是防疫工作的重中之重.

总体来说,网民对于此次疫情的关注的话题是较为积极,高频词中并未出现任何消极的词汇,人们关注的话题往往是如何与病毒做斗争,如何抗击肺炎,拯救患者,支援武汉,这反映了我们国家“一方有难,八方支援”的团结精神.

在词云统计图中,词频是通过字体的大小进行分布的. 如图 5,我们可以看出,首先,“肺炎”、“新型”、“疫情”、“冠状病毒”等字体明显突出,说明关于此次疫情的话题是以新型冠状病毒肺炎为核心的. 其次,“抗击”、“救治”、“患者”、“全国”、“加油”等词汇在词云中也清晰地呈现出来,体现了网民齐心协力,众志成城,形成了共同抗击肺炎的决心以及战胜新冠肺炎病毒的信心. 再次,词云中出现了“钟南山”人名,这反映了网民对钟南山品德的敬仰以及对其贡献的肯定. 在此次疫情中,钟南山是较早奔赴武汉的医疗专家之一,在抗击疫情的过程中起到了重要的作用. 此外,词云中还出现了“日本”国家名,日本对我国抗击疫情所提供的帮助也体现在了热议的话题之中. 以上这些反映了关于新冠肺炎疫情的舆情中,微博网友大力弘扬的是正直善良、无私奉献、团结互助、睦邻友好等正能量.

图 5　新冠肺炎疫情话题词云图

Fig. 5　Topic cloud chart of novel coronavirus pneumonia

3.3　新冠肺炎疫情地理统计分析

为使“新冠肺炎疫情”的舆情态势更加直观可见，研究抓取了从 1 月 1 日～2 月 29 日期间与“新冠肺炎疫情”话题相关的热门评论共 1.5 万条，评论包括评论内容、评论人所在地、评论时间等. 运用地理统计分析的方法，以某个地区用户所发热门评论数量作为此次疫情话题在该地区的讨论热度，并绘制了全国微博热门评论人数分布图（如图 6），“新冠肺炎疫情”的讨论涉及全国各地用户，所在地为湖北的用户讨论极为热烈，占评论总人数的 9.33%，这与湖北是全国新冠肺炎疫情的重灾区密切相关. 所在地为广东的用户占比为 8.84%，北京市为 7.87%，江苏省为 7.78%，参与讨论人数占比超过 5%人的行政区还有浙江、四川、山东、上海，这些大多是本次疫情蔓延比较严重的地区，其中还包括对新冠肺炎疫情不太严重但防控意识较强的地区.

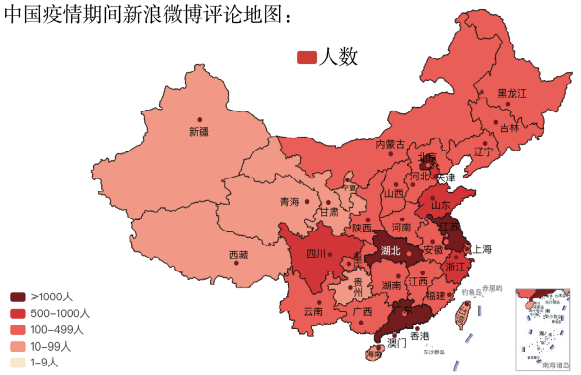

图 6　全国 34 个行政区微博热门评论人数分布图

Fig. 6　Distribution map of popular comments on microblog in 34 Administrative Regions of China

研究又对每个行政区的评论做了情感分析，以微博用户情绪均值作为该地区的情感指数，其中，湖北省的情感指数最低，为 0.525 6，代表心情很差（如图 7），西藏为 0.534 1，代表心情差，江苏为 0.598 8，安徽为 0.606 8，代表心情适中. 而新疆为 0.654 4，甘肃为 0.662 3，代表心情很好. 需要说明的是，由于西藏网民只占总评论人数的 0.19%，所以其心情较差应不具有代表性. 结合截至 2 月 29 日的全国新冠肺炎累计确诊总数的地图（如图 8）可以看出，疫情越严重的地方情感指数越低，即越负向，而疫情较轻的地方情感指数越高，即越正向. 因此语料库中出现率较高且情感指数较低的区域，与此次“新冠肺炎疫情”最严重的区域存在较高的吻合度.

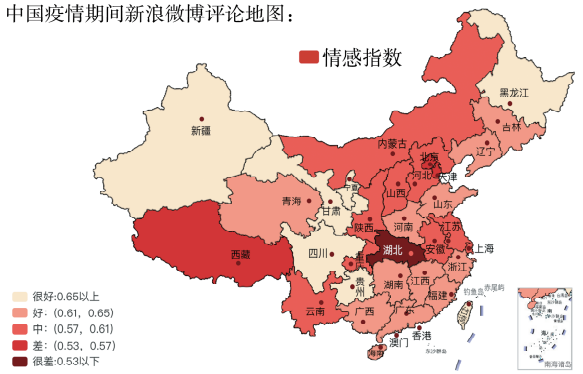

图 7　全国各行政区网民情感值分布图

Fig. 7　Distribution map of sentimental value of Internet users in all administrative regions of China

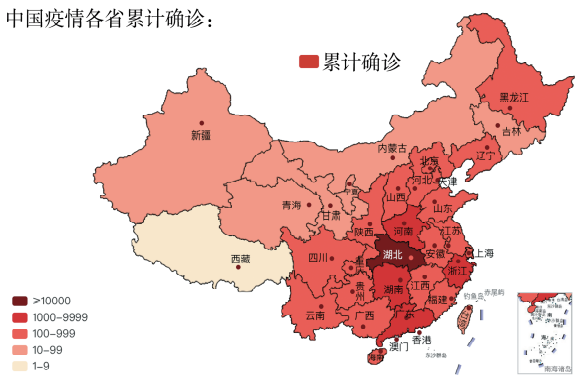

图 8　全国新冠肺炎累计确诊地图

Fig. 8　Cumulative quantity map of confirmed cases of novel coronavirus pneumonia in China

4　结　论

本研究使用 Scrapy-Redis 分布式爬虫对微博热门话题进行抓取，利用 MongoDB 分布式数据库对抓取的数据信息进行存储，使用 K-Means 算法对微博数据进行话题聚类，并优化 SnowNLP 对采集到的文本进行情感分析，最后通过地理统计分析将各地网民情绪与各地疫情状况相印证. 结果初步

揭示了网民对新冠肺炎疫情的态度变化趋势,总结了在不同时间段网民关注的多个话题,同时结合疫情中不同地理位置网民的心态以及评论人数,分析预测接下来疫情的时空发展趋势,研究发现的意义主要有以下四个方面:

(1) 网民对于"新冠肺炎疫情"的整体态度是向好的,虽然在前一阶段出现了较大波动,但是在此之后网民态度逐渐好转,并稳定在以积极情绪为主导的态势,并用文本聚类检测出网民情感波动的原因与肺炎疫情是否可防可控、新型冠状病毒会不会人传人等话题密切相关,而人们积极情绪的出现与疫情得到较好控制有关;(2) 通过高频词柱状图和词云,得出在疫情期间网民关注的话题大多与如何进行新冠肺炎疫情防控与全国共同抗击肺炎有关;(3) 通过地理统计分析,得出新冠肺炎疫情影响严重的区域,评论人数更多,情感更偏于负向;(4) 图 3 所显示的情感波动幅度与疫情态势基本成正相关的规律,即波动幅度越小,疫情态势越稳定向好,可以预判全国疫情发展走势将趋于稳定向好,疫情防控工作还要稳扎稳打. 由图 7 所显示的全国各行政区网民情感值与疫情严重程度基本成负相关的规律,即情感值越低,疫情越严重,可以推测湖北省的疫情形势在未来一段时间仍然会比较严峻,武汉依然是全国疫情防控、病患救治等工作展开的重点城市.

在疫情趋于平稳的阶段,可利用实时产生的评论数据进行分析,通过对不同所在地用户评论、微博的文本聚类与情感分析,结合新闻媒体报道,预测和定位到疫情蔓延的受灾区域的受灾情况以及当地感染人数的增减. 后续将进行热点话题的空间分布与地域性关联进行研究[17],同时根据挖掘意见领袖在灾害事件中的话题传播作用[18],剖析在网络舆情中,谣言的产生与传播过程. 并试图构建灾害情况与舆情的回归曲线方程,从而更好地通过舆论情报来判断特定地理位置的受灾情况.

参考文献:

[1] 丁杰, 徐俊刚. IPSMS: 一个网络舆情监控系统的设计与实现[J]. 计算机应用与软件, 2010, 27: 188.

[2] 洪小娟, 宗江燕, 于建坤, 等. 网络舆情监测系统的分析与设计[J]. 软件工程, 2019, 22: 37.

[3] 李然, 林政, 林海伦, 等. 文本情绪分析综述[J]. 计算机研究与发展, 2018, 55: 30.

[4] 陈兴蜀, 高悦, 江浩, 等. 基于 OLDA 的热点话题演化跟踪模型[J]. 华南理工大学学报: 自然科学版, 2016, 44: 130.

[5] 明弋洋, 刘晓洁. 基于短语级情感分析的不良信息检测方法[J]. 四川大学学报: 自然科学版, 2019, 56: 1042.

[6] 胡思才, 孙界平, 琚生根, 等. 基于扩展的情感词典和卡方模型的中文情感特征选择方法[J]. 四川大学学报: 自然科学版, 2019, 56: 37.

[7] Maks I, Vossen P. A lexicon model for deep sentiment analysis and opinion mining applications[J]. Decis Support Syst, 2012, 53: 680.

[8] Öztürk N, Ayvaz S. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis [J]. Telematics Inform, 2017, 35: 136.

[9] 凌海彬, 缪裕青, 张万桢, 等. 多特征融合的图文微博情感分析[J/OL]. 计算机应用研究. (2018-12-04) [2019-06-06]. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 12. 0929.

[10] 白鹤, 汤迪斌, 王劲林. 分布式多主题网络爬虫系统的研究与实现[J]. 计算机工程, 2009, 35: 13.

[11] 王培名, 陈兴蜀, 王海舟, 等. 多策略融合的微博数据获取技术研究[J]. 山东大学学报: 理学版, 2019, 54: 28.

[12] 袁逸铭, 刘宏志, 李海生. 基于密度峰值的改进 K-Means 文本聚类算法及其并行化[J]. 武汉大学学报: 理学版, 2019, 65: 457.

[13] 李健, 曹垚, 王宗敏, 等. 融合 k-means 聚类和 Hausdorff 距离的散乱点云精简算法[J]. 武汉大学学报: 信息科学版, 2020, 45: 250.

[14] 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: 757.

[15] 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/OL]. 山东大学学报: 工学版. (2020-02-21). [2020-03-11]. http://kns. cnki. net/kcms/detail/37. 1391. T. 20200221. 1529. 004. html.

[16] 曹彦波. 基于新浪微博的 2018 年云南通海 5. 0 级地震舆情时空特征分析[J]. 地震研究, 2018, 41: 525.

[17] 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: 906.

[18] 王祎珺, 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: 35.

引用本文格式:

中 文: 陈兴蜀, 常天祐, 王海舟, 等. 基于微博数据的"新冠肺炎疫情"舆情演化时空分析[J]. 四川大学学报: 自然科学版, 2020, 57: 409.

英 文: Chen X S, Chang T Y, Wang H Z, *et al*. Spatial and temporal analysis on public opinion evolution of epidemic situation about novel coronavirus pneumonia based on micro-blog data [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 409.