

基于文本挖掘技术的新冠肺炎舆情分析

陈顺丽^{1,2} 胡钰^{1,2}

¹武汉大学遥感信息工程学院 武汉 430079 ²广州城市信息研究所有限公司 广州 510665

摘要: [目的/意义]分析“新冠肺炎”事件影响下,相关网络舆论情况及民众的情绪变化特征,探究文本挖掘技术在舆情研究中的作用和意义。[方法/过程]利用python网络爬虫技术爬取了人民日报官方微博下“每日疫情通报”话题的评论数据,并对采集的舆情信息进行量化处理和词频分析,从热度变化、地域分布、情感趋势等方面进行文本信息挖掘,并对分析结果进行了可视化处理。[结果/结论] 1、“新冠肺炎”事件引起了短期内的爆发性网络舆论,相关舆情随事件发展时间线呈现出较为明显的情绪波动变化;2、时空下的文本语义分析技术实现了从纷繁复杂的网络文本数据中挖掘出有价值的信息,为舆情的研究提供新的视角与思维扩散点。

关键词: 网络爬虫; python 文本挖掘; 情感分析; 新型冠状病毒

分类号: TP391

1 引言

随着新媒体体系的迅速成长,网络用户数量也越来越多,据中国互联网信息中心(CNNIC)在2019年8月30日发表的第44次《中国互联网络发展状况统计报告》显示:截至2019年6月,我国网民规模达8.54亿,互联网普及率达61.2%,我国手机网民规模达8.47亿,网民使用手机上网的比例达99.1%。^[1]随着互联网时代的高速发展,网络媒体已成为民众获取外界信息与发表自身观点的最主要载体。借助于互联网技术,人们可以随时随地地获取资源信息,也可借助微信、微博、短视频、网络直播等互联网平台发表个人见闻和对当前社会热点事件的观点、态度、意见、情感等民意取向从而形成突发公共事件中不容小觑的网络舆情。

2020年初,一场“新冠肺炎”席卷全国,新型冠状病毒感染的肺炎是一种急性感染性肺炎,具有很强的传染性。自2019年年底发现到2020年年初快速扩散到全球数十个国家和地区,引起疫区大众严重的心理恐慌。社交媒体的流行,迅速让全国人民目睹了这场灾难。民众通过各种渠道关注着疫情的情况,也通过各种网络平台吐露自己的心

声;新浪微博作为最大的网络社交平台之一,既是获取疫情信息的主要来源,也是民众发表自己的观点的网络重要平台,通过分析新浪微博上疫情相关热点话题的评论可以看到大家在疫情发展变化下的心里活动变化及微博舆论走向。为了分析民众对“新冠肺炎”这一热点话题的舆论焦点及舆论情感变化,探讨官方微博在舆论引领中发挥的作用,笔者通过网络爬虫技术爬取人民日报官方微博中“每日疫情通报”下的评论数据,然后基于中文文本挖掘技术,对采集所得的舆情信息进行量化处理和分析,从舆情热度变化、词频热度分析及情感倾向等方面进行舆情信息分析。

2 相关研究

网络舆情,是指“在一定的社会空间内,通过网络围绕社会事件的发生、发展和变化,民众对公共问题和社会管理者产生和持有的社会政治态度、信念和价值观”。^[2]近年来,网络舆情信息挖掘也受到国内外学者的关注,相关学者从特征词提取、情感分析、舆情可视化及舆情传播特征分析等各个方面

作者简介: 陈顺丽,女,硕士研究生, E-mail: 1277216146@qq.com; 胡钰,女,硕士研究生, E-mail: 2680383499@qq.com.

对网络舆情分析展开研究,并取得了一定的研究成果。

Wu 等^[3]提出词频——逆文档频率 (term frequency-inverse document frequency, TF-IDF), 以此反映各个单词不同的特征重要性并提高了分类效果。林江豪等^[4]使用 one-hot 编码方法将微博文本表示成向量, 并利用贝叶斯分类器进行分类。侯敏等^[5]基于微博数据, 提出利用短语情感词典分析话题型微博的情感倾向。赵妍妍等^[6]将微博数据中的表情符号提取并转化成词语种子, 然后结合现有的情感词典, 构建了更完善的情感词典。钮成名等^[7]利用改进的循环神经网络方法进行中文微博情感分析, 有效提高微博文本情感分类的正确率。陈涛等^[8]通过定性分析搜索引擎关注度的时间序列和搜索引擎用户的空间分布, 研究舆情话题强度随时间的变化以及舆情话题的受关注区域。王雪猛等^[9]基于倾向分析构建网络舆情分析模型, 并提出五级情感色彩划分方法, 划分民众情感倾向立场, 实现网络舆情预警功能。周义棋等^[10]通过对新浪微博舆情信息进行量化处理和词频分析, 从热度变化、地域分布、情感趋势等方面分析舆情信息, 分析民众对巴黎圣母院大火的舆论关注情况及舆论走向, 探讨官方微博在舆论引领中发挥的作用。HAN 等^[11]以新浪微博为载体, 研究了 2018 寿光水灾相关舆情信息的时空分布及其内容, 并探讨了寿光水灾期间的公众情绪; Allan J^[12]提出话题检测与跟踪(TDT)是了解社交媒体热点话题及其演变过程的重要手段。W Ding 等^[13]提出了半监督狄利克雷—霍克斯方法(SDHP)来处理 Twitter 中的主题检测和跟踪问题。CODY 等^[14]采集 Twitter 上含“气候”一词的推文, 通过情感分析探讨了公众情绪随气候和自然灾害的变化情况; KARAMI 等^[15]提出了基于 Twitter 的舆情分析框架, 用于了解灾害发生时的公众情绪, 可为决策提供参考; DOAN 等^[16]和 SHIBUYA 等^[17]分别搜集了 Twitter 和 Facebook 信息, 分析了 2011 年日本地震和海啸后公众的情绪。以往的相关研究都为舆情分析提供必要的工具。本文基于对相关文献的学习总结, 利用中文文本挖掘

技术对“新冠肺炎”事件在新浪微博平台上的所引起的网络舆情焦点及舆情传播过程中存在的表现特征做分析和探讨。

3 研究方法

3.1 数据采集

目前获取微博数据的方法众多, 主要包括官方 API(Application Programming Interface) 调用、网络爬虫及公共数据集下载三种方法^{[18][19]}。通过程序调用官方的 API, 用户可以很方便地得到想要的数 据, 但由于微博平台对其 API 的访问有设置权限, 并且有访问次数的限制, 若要获取大量的数据一般通过 API 经过多次的访问才能得到。故本次研究中 Python 编程为基础, 使用 Python+Selenium+Chrome 编写爬虫程序, 通过模拟用户登录爬取研究所需的微博评论数据。

3.2 舆情文本信息挖掘

利用 Python 中 jieba 分词技术、TF-IDF 技术及情感分析技术等, 对所采集的数据进行文本信息挖掘及可视化分析研究。首先, 对爬虫采集的数据进行数据清洗及分词等预处理, 并过滤常见的停用词; 然后, 使用 TF-IDF 进行特征词提取; 情感分析方面, 本文基于情感词典与 SVM 结合的方法进行评论文本的情感倾向分析。支持向量机 (Support Vector Machine, SVM) 模型^[20]于 1995 年由 Vapnik 等人提出, 是一种有监督学习方法, 以结构风险最小化为原则, 是一种具有很好泛化能力的分类工具, 可以用来进行预测, 分类和异常检验, 作为一种有监督学习方法, SVM 常被用于二分类问题。最后, 基于时间序列, 将微博舆情的情感态度变化按时间序列进行可视化处理, 更清晰地反映网民的态度随疫情事态发展的变化趋势。

4 新冠肺炎舆情分析

4.1 数据源

新浪微博作为中国最大的微博平台, 具有用户数量庞大, 数据开放程度相对较高, 低门槛、即时性、交互性强和传播迅速等特

征，成为舆情的发源地和聚集地。因此笔者基于新浪微博平台，利用文本挖掘技术对“新冠肺炎”的相关微博舆情展开研究。关于原始数据的采集主要考虑两个原因：一是若对所有以“新冠肺炎”为关键词的微博内容都进行抓取并研究，则数据量过大，且杂乱无章，超出研究的人才、物力与时间所限，不便于研究。二是“人民日报”的官方微博作为国内一大新闻信息源，受到的关注度较高，且从2020年1月21号开始每天持续进行每日疫情速报，故本次研究爬取了人民日报官方微博下2019年1月23日至2019年3月18日之间每日疫情速报下的评论共计20余万条作为研究基础进行相关舆情分析。具体爬取数据包含当天疫情通报发布内容、发布时间、微博转发数、评论数、点赞数，及当前微博下所有评论信息，其中又包括每一条评论用户名称、用户ID、用户链接、评论内容及评论发表时间。

4.2 数据处理及技术路线

对爬取获得的微博评论数据进行去噪去重、分词及去停用词等预处理后入库，再使用词频分析、TF-IDF 特征词提取及基于SVM的情感倾向分析方法进行建模分析，最后将所得结果通过数据可视化方法展现。主要流程如图1所示。

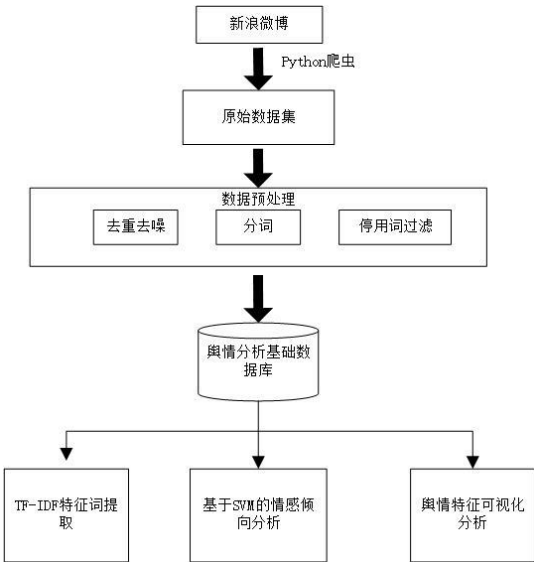


图1 微博舆情分析技术路线图

(1) 文本清洗及分词处理

在进行方法验证之前，需对原始评论数据进行预处理。爬取的原始数据导出为CSV

文件，包括人民日报微博下疫情相关话题主页的每日评论数和转发量及每日详细评论信息。每日评论详细数据中包括评论者用户名、用户id、评论内容、评论发表时间等字段。其中内容字段中为文本舆情分析数据，但由于爬取到的数据中存在诸如空白、无效评论信息等一些无用数据，所以首先对采集的数据进行清洗，在尽量保证数据完整的前提下，确保采集数据的有效性。然后，从数据源特征出发，对表格中的数据进行评论文本信息的提取，最后采用python中的结巴分词工具对提取的语料集进行分词和去除停用词等处理。

(2) 特征词提取及文本向量化

网络平台用词日新月异，若直接使用结巴分词对数据进行处理，可能会导致分词不准确问题以至于影响后面分析结果，故本文在对“知网情感词典”和“台湾大学简体中文极性词典 NTUSD”进行研究的基础上，结合人工添加新词的方法构建于适用于本次疫情舆情分析的用户自定义词典（表1）和情感词典（表2）。其中，用户自定义词典主要新增了一些新兴词汇及结巴自动分词可能引起歧义的部分词汇，而情感词典主要包括正负情感词集，用于辅助舆情情感倾向分析。

表1 部分新增词汇示例

| 新增词汇 |
|---|
| 重危、牛逼、抗疫、市长、洁厕灵、灭活、急一急、新冠肺炎、上网课、挺过去、靠谱、肿么了、冲鸭、居住史、营销号、骗赞、养号、自媒体、健康卡、带节奏、掰扯、拉黑、憋批、飘零、绿码、防护镜、奥利给、输入性、官宣、熟食物、生食物、心塞、逼格、接地气、猴赛雷、吃土、傻白甜、狗带、然并卵 |

表2 部分情感词汇示例

| 词集 | 极性 (标识) | 情感词 |
|----|------------|-----|
|----|------------|-----|

| | | |
|-------|-----------|---|
| P_set | 正向 (2) | 希望、期待、安全、真实、正常、安好、感谢、支持、有效、负责、平平安安、致敬、实事求是、强大、伟大、同意、愿意、免费、积极、稳定、准确、喜欢、乐观、国泰民安、团结、精准、快乐、感恩、欢迎、严谨、珍惜、一视同仁、英勇、祝愿、善良、幸福 |
| N_set | 负向 (0) | 害怕、可怕、严重、辛苦、恐慌、麻烦、吓人、担心、恐怖、厉害、不行、危险、难受、揪心、强制、难过、严峻、封闭、心痛、心疼、掉以轻心、困难、着急、触目惊心、歧视、失望、绝望、寒心、痛心、抱怨 |

基于所建词典对评论文本进行分词及特征词提取，生成关键词词云，并采用 TF-IDF 对特征项进行权重计算，综合否定副词及程度副词的强度计算特征项的情感值，将特征权重与特征项的情感值加权融合，最后再以空间向量的形式表示该文本，应用于 SVM 模型，进行疫情评论情感分析。

4.3 微博舆情分析及可视化

(1) 疫情舆论热度趋势分析

通过“疫情通报”下每日评论数和转发量统计，可看到随着“新冠肺炎”事态变化下相关舆论热度变化趋势。如图 2 所示，日评论数最多时高达 7 万余条，单日转发量最多时高达近 5 万次。1 月 23 号—26 号及 2 月 13 号日评论量都达 1 万以上，评论数高峰出现在 1 月 25 和 2 月 13 号；转发峰值则出现在 1 月 25 号、2 月 13 号和 3 月 11 号三天。1 月 23 日武汉宣布“封城”，进行严厉的防疫措施，限制公共交通出入，全国的目光都投向湖北武汉，新冠病毒真正进入全民关注的焦点，在接下来的两天里疫情舆论便达到高峰。2 月 11 号到 13 号新冠肺炎新增确诊病例急速增加，11 号到 12 号更达到日增长量最高峰，单日新增确诊病超过 1 万，截至 2 月 13 日 0 时全国疫情新增确诊 15152 例，再度引起了激烈的网络反响，2 月 13 号当天疫情微博评论及转发数都达到峰值。之后舆论反响降温了一段时间，直到 3 月 11 号再次出现转发量和评论峰值，查看当天的评论数据发现“海外疫情”、“境外输入”等

话题词汇增多，明显随着后期国内疫情逐渐好转，及海外疫情的逐步严峻，海外疫情开始成为讨论的热点话题。

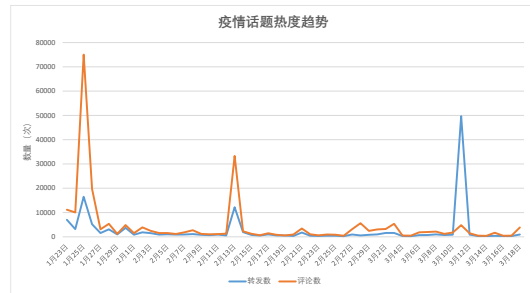


图 2 疫情舆论热度趋势图

(2) 疫情舆论热词统计分析：

综合所有日期评论信息，进行文本分词、词频统计处理得到舆论热词统计图（图 3），并通过关键词提取疫情评论中最重要的 200 个词，利用 WordCloud 制作词云图，matplotlib 库用来可视化词云图，得到如图 4 所示词云图。

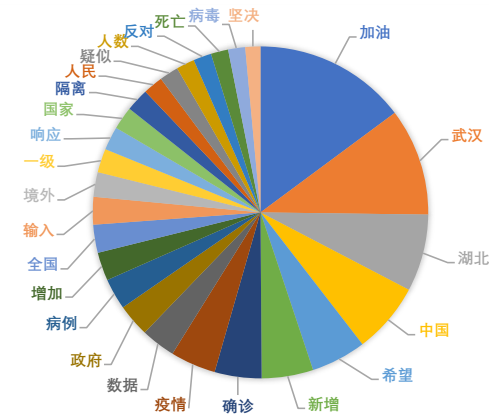


图 3 疫情舆论热词统计分析



图 4 疫情舆论热词分析——词云

从热词分析中看到排名前 3 的词为：“加油”，“武汉”，“湖北”；评论热词反映了疫情舆论的焦点，以及群众对疫情尽快好转的期盼；武汉作为此次疫情爆发的主要地区，成为评论中的高频词汇，位居评论热词第二，而湖北则紧随其后。

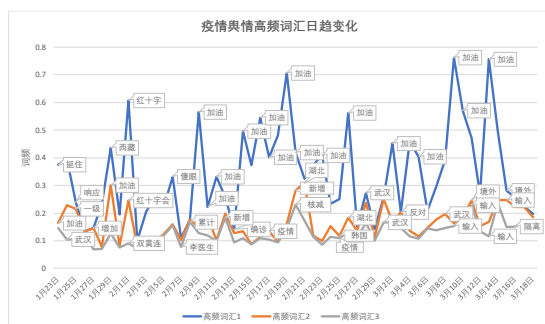


图5 疫情高频词汇日变化

通过“疫情舆情高频词汇日趋势变化”统计图（图5）可以看到疫情舆论焦点随时间的变化趋势。前期国内疫情发展是大家热议的话题，相关热词“加油”、“武汉”、“湖北”、“红十字”等词多次出现，后面随着2月下旬开始，新冠肺炎在全球蔓延，“境外”，“输入”等词频增高，说明网络舆情的话题热点更多转向了国外疫情的发展及“防止境外输入”等话题。

（3）舆论热点地区分析

通过对评论中出现的所有国内地名进行统计，获得地名的词云，并以各地区在评论中的出现次数（词频）为依据绘制了各省份关注度的分布图（图6）。可以看到，经过两个多月的疫情的发展，全国各地都进入了大家关注的视野，其中作为疫情爆发中心区域，武汉和湖北舆论热度最高，随着疫情的发展，其周边地区安徽、浙江等地也出现在热评中。1月29号，作为最后一块净土的西藏“沦陷”，首次出现了一例新增确诊病例，从而成为新的热议焦点。



图6 国内热点地区分布

（4）舆情情感分析：

通过对疫情评论做时间序列上的舆情情感分析，获取疫情事态变化下网民相应的情感变化。

如图7所示，每日情感正负情感分值为对单日所有评论进行SVM情感分析获取正负情感分值后再取均值的结果。通过情绪走势，可以明显看到2020年1月23日至3月18日时间段内，网民情绪随疫情事件发展的起伏变化。前期肺炎疫情发酵，舆论总体情感偏向负面，随着国内疫情逐渐好转，网民情感也逐渐向积极方向转变。

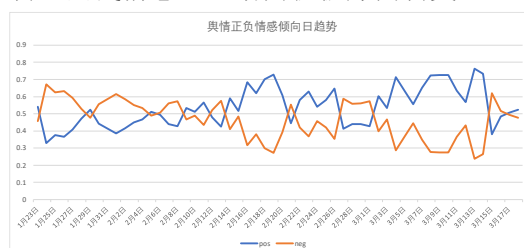


图7 舆情情感日趋势变化

通过对不同时间段内情感高频词汇的提取，获得情感热词阶段性变化（表3）。高频情感词汇的变化更直接地体现了人们在这段时间内的心里活动特征和随疫情发展在不同时间段的情感变化。

表3 情感热词阶段性变化

| 日期 | 高频情感词统计 |
|-------------|---|
| 1月23日-1月27日 | 害怕、希望、重视、恐慌、歧视、需要、致敬、注意、支持、难过、关注、抱歉、着急、谢谢、揪心、担心、接受、期待、心痛、难受 |
| 1月28日-2月2日 | 希望、恐慌、关注、害怕、重视、肯定、支持、需要、注意、致敬、心痛、难受、教训、担心、揪心、心疼、难过、寒心、失望、原谅、 |
| 2月3日-2月7日 | 希望、害怕、揪心、关注、心疼、恐慌、心痛、重视、难过、愤怒、难受、接受、需要、担心、致敬、掉以轻心、痛心、感谢、叫嚣、珍惜 |
| 2月8日-2月12日 | 希望、快乐、揪心、致敬、难受、祝愿、期待、注意、质疑、关注、掉以轻心、珍惜、担心、开心、集中、支持、谢谢、心碎、心酸、重视 |
| 2月13日-2月17日 | 希望、恐慌、接受、需要、害怕、肯定、震惊、致敬、感谢、谢谢、关注、质疑、揪心、掉以轻心、感恩、开心、心疼惶恐、担心、寒心、期待 |
| 2月18日-2月22日 | 希望、致敬、谢谢、关注、向往、开心、快乐、教训、重视、期待、尊重、心急火燎、恭喜、掉以轻心、怀念、盼望、担心、需要、疑惑、尴尬 |
| 2月23日-2月27日 | 希望、掉以轻心、感谢、期待、庆幸、关注、开心、难受、谩骂、珍惜、揪心、心疼、害怕、重视、后悔莫及、致敬、谢谢、心碎、可怜、感谢 |
| 2月28日-3月1日 | 希望、需要、抑郁、关注、歧视、期待、谢谢、嫌 |

| | |
|-------------------|--|
| 月 2 日 | 弃、嘻嘻哈哈、不以为然、教训、同意、通过、欢迎、支持、认可、愿意、供奉、致敬、妄想 |
| 3 月 3 日-3 月 8 日 | 希望、期待、惨痛、关注、掉以轻心、歧视、感谢、害怕、担心、将心比心、注意、掉以轻心、看不起、期望、盼望、恐慌、焦虑、感谢、谢谢、欢迎 |
| 3 月 9 日-3 月 13 日 | 希望、期待、担心、注意、迎接、关心、开心、教训、接受、支持、掉以轻心、冷嘲热讽、感谢、关注、致敬、悲哀、允许、珍惜、佩服、开心 |
| 3 月 14 日-3 月 18 日 | 希望、期待、感谢、愿意、关注、致敬、珍惜、欢乐、焦急、接受、需要、担心、掉以轻心、歧视、肯定、允许、感恩戴德、接受、难受、委屈 |

5 总结

本次研究基于数据挖掘技术，利用 python 爬取“新冠肺炎”微博评论，并通过词频、词云图、情感倾向性分析进行舆情分析，探讨了此次疫情事件引起的社会舆论效应及在此次突发公共卫生事件中人们情感倾向变化特征，发现：

1) 民众对“新冠肺炎”事件高度关注，2 月 25 日单日评论达 7 万余条，体现了舆情信息在短时间内的爆发性。

2) 随着“新冠肺炎”的爆发、蔓延，微博关于“新冠肺炎”的舆情信息也从疫情爆发中心武汉逐渐向全国扩散。

3) 舆论的主流倾向由前期的担忧，恐惧、埋怨等消极情绪，随国内疫情事态的好转继而转向希望、开心、感激等积极情绪。

4) 官方微博账号能够引导舆论的走向。在热点事件当中，官方微博的积极引导，有助于民众舆论朝着正能量的方向发展。

文本挖掘技术可以帮助人们从纷繁复杂的网络文本数据中挖掘出有价值的信息，实现对舆情事件的宏观把握与深度解读，为今后舆情的研究提供新的视角与思维扩散点。于政府机构而言，通过微博舆情的信息监测，也可以对热点话题进行监测和预警，实时掌握微博舆情，从而能够在必要时，对舆情导向做一个控制，进而维护社会的稳定。

参考文献

[1] 于朝晖.CNNIC 发布第 44 次《中国互联网络发展状况统计报告》[J]. 网信军民融合,2019(09):30-31.

[2] 曾润喜.网络舆情管控工作机制研究[J].图书情报工作,2009,53(18):79-82.

[3] WU HC, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions [J]. Acm Transactions on Information Systems, 2008, 26 (3): 55-59.

[4] 林江豪,阳爱民,周咏梅,等.一种基于朴素贝叶斯的微博情感分类[J].计算机工程与科学,2012,34(9):160-165.

[5] 侯敏,滕永林,李雪燕,等.话题型微博语言特点及其情感分析策略研究[J].语言文字应用,2013(2):135-143.

[6] 赵妍妍,秦兵,石秋慧,等.大规模情感词典的构建及其在情感分类中的应用[J].中文信息学报,2017,31(2):187-193.

[7] 钮成明,詹国华,李志华.基于深度神经网络的微博文本情感倾向性分析[J].计算机系统应用,2018,27(11):205-210.

[8] 陈涛,林杰.基于搜索引擎关注度的网络舆情时空演化比较分析——以谷歌趋势和百度指数比较为例[J].情报杂志,2013,32(03):7-10+16.

[9] 王雪猛,王玉平.基于情感倾向分析的突发事件网络舆情预警研究[J].西南科技大学学报(哲学社会科学版),2016,33(01):63-66.

[10] 周义棋,田向亮,钟茂华.基于微博网络爬虫的巴黎圣母院大火舆情分析[J].武汉理工大学学报(信息与管理工程版),2019,41(05):461-466.

[11] HAN X H, WANG J L. Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang city flood in China [J]. International Journal of Geo-Information, 2019, 8(4): 185.

[12] Allan J. Introduction to topic detection and tracking[M]//Topic detection and tracking. Springer, Boston, MA, 2002: 1-16.

[13] Ding W, Zhang Y, Chen C, et al. Semi-supervised Dirichlet-Hawkes process with applications of topic detection and tracking in Twitter[C]//2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016: 869-874.

[14] CODY E M, REAGAN A J, MITCHELL L, et al. Climate change sentiment on twitter: an unsolicited public opinion poll [J]. Plos One, 2015, 10(8): e0136092.

- [15] KARAMIA, SHAH V, VAEZIR, et al. Twitter speaks: a case of national disaster situational awareness [J/OL]. Journal of Information Science.[2019-09-30].http://xueshu.baidu.com/usercenter/paper/show?paperid=1n760ck01t2e0480sx3b00w00r152009&site=xueshu_se.
- [16] DOAN S, VO B K H, COLLIER N. An analysis of twitter messages in the 2011 Tohoku earthquake[J]. Computer Science, 2011(91):58-66.
- [17] SHIBUYA Y, TANAKA H. Public sentiment and demand for used cars after a large -scale disaster: social media sentiment analysis with facebook pages [J /OL] . [2019-09-30]. http://xueshu.baidu.com/usercenter/paper/show?paperid=bf489713ed68e033708239266df6c4df&site=xueshu_se.
- [18] 谢松燕,刘肖依.基于微博平台的舆情分析研究综述[J].商,2016(14):206.
- [19] 汤露阳. 面向网络舆情分析的数据采集与管理方法研究[D].电子科技大学,2017.
- [20] Cortes C, Vapnik V. Support vector machine[J]. Machine learning, 1995, 20(3):273-297.

作者贡献说明:

陈顺丽: 论文中实验设计, 数据分析及论文的修改和定稿;

胡钰: 论文数据采集、数据预处理及论文初稿。

The Network Public Opinion Analysis of 2019-nCoV Based on Text Mining Technology

Chen Shunli ^{1,2} Hu Yu ^{1,2}

¹College of Remote Sensing Information Engineering, Wuhan University, Hubei 430079

²China Digital Cities Institution Co. LTD, Guangzhou 510665

Abstract: [Purpose/Significance] In order to analyze related network public opinion as well as the public's emotional characteristics caused by the influence of the novel coronavirus (2019-nCoV) and explore the role and significance of text mining technology in public opinion research. [Method/Process] Using Python web crawler technology to crawl the comment data of the topic of "daily report of epidemic situation" on the official microblog of The People's Daily, then doing the text vectorization as well as word frequency analysis on the collected data, and carried out text information mining from the aspects of heat change, regional distribution, emotional trend, etc. In the end, the analysis results of previous steps were visualized. [Result/Conclusion] The experimental results show: 1. The novel coronavirus pneumonia caused a sudden public opinion in the short term, and the related public opinion showed a more obvious emotional fluctuation with the change of the event. 2. The text semantic analysis technology combining time and space can mine valuable information from the complex network text data, and provide a new perspective and thinking diffusion point for public opinion research.

Key words: Web crawler; python text mining; emotion analysis; 2019-nCoV