摘　要

科学技术的发展使民众随时随地访问网络、跨时空沟通交流不再是空想。各类社交平台的产生与流行拓展了用户的交流渠道，不同于传统的面对面对话，跨地区即时共享信息成为当代用户的主流选择。网民在这些平台上通过转发、评论和点赞等操作来发表自己的观点，这些蕴含发布者真情实感的信息经过网络的传播后极易汇聚成网络舆情。一旦网络舆情的发展超过合理可控范围会造成舆情治理困境，影响网络空间治理，威胁社会稳定。

本文以数据挖掘为手段，结合空间统计方法为用户情感的时间演化做补充，利用时空分析方式更好地为舆情的合理控制、引导提供参考价值。首先，文章确定以“河南暴雨”事件为案例，通过爬取新浪微博上官方微博发布的微博内容、微博评论及评论者个人信息为原始数据进行实验。其次，在时序分析中根据 Pearson 系数确定阶段划分的标准为微博评论量，并将舆情周期确定为爆发期、波动衰退期和消亡期三个阶段。再次，从用户情感、热点话题实现对网络舆情的时序分析，从用户评论的数量分布探寻用户关注度的分布特征，从用户评论情感值分布是否存在明显的舆情情感聚集空间，根据这三方面探寻网络舆情情感态势演化的时空特征来解读舆情演化的内在机理，以求得出普遍规律。

研究发现：（1）“河南暴雨”事件中用户情感演变、舆情主题演变及用户情感分布具有明显的阶段性。用户情感倾向总体以正向情感为主，但在特殊时间点有出现情感均值、正向情感比重降至最低点的情况。（2）不同阶段内舆情主题演变呈现出明显的主题关联规律，具有连续性和继承性。（3）就空间角度而言，用户情感分布具有全域覆盖性和全民参与性，存在明显的聚集特征。表现为在我国西北方向存在明显的用户情感冷点区域，情感热点区域则多位于东部地区。基于研究结论，本文提出网络舆情引导可以从传播主体、时空变化特征两个方面进行，为有关管理部门在舆情研判、分析预测和合理引导等方面提供辅助决策支持和依据。

关键词：微博舆情；舆情分析；情感倾向；时空演化

Abstract

The development of science and technology makes it no longer a dream for people to access the Internet anytime and anywhere and communicate across time and space. The emergence and popularity of various social platforms have expanded the communication channels for users. Different from traditional face-to-face conversations, instant information sharing across regions has become the mainstream choice for contemporary users. Netizens express their opinions on these platforms through operations such as forwarding, commenting, and liking. These information containing the true feelings of the publishers are easily aggregated into online public opinion after being disseminated on the Internet. Once the development of network public opinion exceeds a reasonable and controllable range, it will cause the dilemma of public opinion governance, affect the governance of cyberspace, and threaten social stability.

This paper uses data mining as a means, combined with spatial statistical methods to supplement the time evolution of user emotions, and uses spatiotemporal analysis methods to better provide reference value for the reasonable control and guidance of public opinion. First of all, the article determines to take the "Henan Rainstorm" incident as a case, and conduct experiments by crawling the Weibo content, Weibo comments and commenters' personal information published by the official Weibo on Sina Weibo as the original data. Secondly, in the time series analysis, according to the Pearson coefficient, the standard for determining the stages is the amount of Weibo comments, and the public opinion cycle is determined into three stages: the outbreak period, the volatility recession period, and the extinction period. Thirdly, the time series analysis of online public opinion is realized from the user emotions and hot topics, the distribution characteristics of user attention are explored from the number distribution of user comments, and whether there is an obvious public opinion emotion aggregation space from the distribution of user comments sentiment value, according to these three aspects. The temporal and spatial characteristics of the emotional situation evolution of network public opinion are used to interpret the internal mechanism of the evolution of public opinion in order to obtain general laws.

The research found that: (1) The evolution of user emotions, the evolution of public opinion topics and the distribution of user emotions in the "Henan Rainstorm"

event had obvious stages. The user's emotional tendency is generally dominated by positive emotions, but at special time points, there are cases where the average value of emotions and the proportion of positive emotions drop to the lowest point. (2) The evolution of public opinion topics in different stages shows obvious thematic correlation laws, which are continuous and inherited. (3) From the perspective of space, the distribution of user emotions has global coverage and participation of the whole people, and has obvious aggregation characteristics. It is manifested that there are obvious user emotional cold spots in the northwest of China, and emotional hot spots are mostly located in the eastern region. Based on the research conclusions, this paper proposes that online public opinion guidance can be carried out from two aspects: the main body of communication and the characteristics of temporal and spatial changes, providing auxiliary decision support and basis for relevant management departments in public opinion research and judgment, analysis and prediction, and reasonable guidance.

Key words: Weibo Public Opinion; Public Opinion Analysis; Emotional Inclination; Space-time Evolution

目 录

第 1 章　绪论

1.1　研究背景与研究意义

1.1.1　研究背景

一方面，信息技术水平近年来不断提高，移动互联网与智能移动终端的普及打破了时空界限，为民众提供更加广阔、便捷的交流渠道。另一方面，中央持续推进网络强国建设战略，各类新兴媒体如雨后春笋般应运而生，对经济发展、社会稳定和文化传播产生重要的影响。除了最基本的沟通交流功能，互联网平台的虚拟性、即时性为网民在线社交赋予了更高的灵活性。文字不再是唯一的信息媒介，图片、音频、视频等丰富了用户的观点表达形式。据中国互联网络信息中心发布的第 49 次《中国互联网络发展状况统计报告》显示：截止至 2021 年 12 月，我国网民规模达 10.32 亿，较 2020 年 12 月增长 2175 万人，其中手机网民占比高达 99.7%；互联网普及率升到 72.4%[1]。可见移动互联网带来了新媒体时代全民传播的发展趋势。在这种发展趋势中，关于用户情感和行为倾向的数据迅速汇聚，使得网络舆论借助平台以爆炸式的传播状态进行传播，极易形成网络舆情。

衣食住行贯穿于居民的日常生活，与之息息相关的公共安全事件往往能够获得居民较高的关注度，针对该类事件的讨论易成为网络舆情的热点问题。近些年屡次发生的“公交事故”造成了不小的社会影响，例如 2020 年 7 月 7 日的“贵州公交坠湖”事件，最终造成了 21 人死亡、15 人受伤的悲痛结果。后经当地公安局发布通知证实，该事故是已故司机因生活不如意而蓄意发起的一次行为。事件一经发布，迅速占据各大平台首页，引发公众的广泛关注与激烈讨论，并以信息、网页分享的形式被大量转发。新浪微博上“贵州公交坠湖司机蓄意报复社会”这一词条更是高达 15.6 亿的阅读次数。舆论狂潮中不仅有对司机的强烈谴责，更引发了网民对公交公司是否存在对雇佣司机心理健康监管不到位的思考，不少民众对于公众出行安全表示忧虑。由此可见，不同平台用户的评论、转发行为在互联网上相互影响、感染，一旦负面情绪或不当言论持续增长，容易造成衍生舆情迭代，政府公信力缺失的局面，给网络空间安全治理和国家稳定造成极大的威胁。因此，以网民情绪为切入点，对用户情感进行时空分析，可帮助政府部门及时应对负面情绪，实现情绪预警，对避免群体性行为出现具有重要的理论和现实意义。

本文以“河南暴雨”事件为例，选择新浪微博为数据来源，以博文内容、博文下方的评论和发表评论的用户个人信息为采集对象，利用算法对已处理的数据进

行信息挖掘，积极探索用户情感的时空演变规律。为相关部门能规范市场管理、及时控制和管理舆论、加强舆情监督和维护社会大环境基本稳定提供理论支持和建议。

1.1.2 研究意义

现实生活中突发公共事件的不可预测性、复杂性和未知性往往伴随着大面积破坏，加速了相关事件报道的迅速传播。这在一定程度上引发民众内心的恐慌与不安，影响网络舆情演变走势，进而对民众日常生活工作造成不小影响，甚至会威胁社会稳定。此外，突发公共事件所引发的网络舆情不仅存在时间差异性，也存在空间差异性。若不能针对两项差异性来制定相应的管理政策，则由区域用户负面情绪所导致的群体行为[2]和次生危机在所难免，为和谐社会的构建带来阻力。因此，本文以 2021 年 7 月的“河南暴雨”事件为例，通过对具体案例中网民情感进行挖掘研究，探索舆情传播在不同阶段内所产生的主题变化与网民情感变化，发现舆情演变的规律和特征，避免次生舆情迭代产生新的舆情危机。此外，本文通过提取、分析网民对该事件的关注程度与所表现的情感特征来探索全国范围内的舆情事件区域空间分布特性和舆情聚集性特征，为相关部门预警策略、应对策略和引导策略的制定提供思路和方向。

（1）理论意义

新冠疫情袭卷全球，在国际范围内造成重大的生命财产损失，如何更好地做好疫情防控管理，减少疫情反复出现次数成为各国的首要任务。各国学者对疫情的关注度居高不下，因此国内外关于突发公共事件网络舆情的研究多以公共卫生事件[3,4,5]为目标对象，有关自然灾害事件的研究相对较少。其次，网络舆情的相关研究成果大多集中在舆情管理[6,7]和舆情演变机理[8,9]上，缺乏以“情感”为特征切入点对突发公共事件进行分析的研究。其中，事件不具有情感，而“情感”作为人类特有的特征能够通过行为、语言等表现出来。事件的情感演化分析，实际上是“用户情感”为主体的演化分析。本文通过对“河南暴雨”事件进行实例挖掘研究，在时间层面上实现对事件传播过程中用户情感的识别，挖掘舆情传播各阶段主题演化规律，同时结合暴雨事件本身分析主题演化的原因，为监管部门有效制定策略提供思路；在空间层面上结合用户个人信息获取不同地域用户对“河南暴雨”事件的关注度，通过评论内容进行用户情感值计算，探索用户情感与用户个人信息在全国范围内的空间差异。最后，根据用户情感在时空分析中所呈现的规律进行网络舆情情感演化研究，作为网民情感演化的有利补充，为未来自然灾害型突发公共事件的舆情应对提供引导性的思路与方法。

（2）现实意义

突发公共事件因其特有的偶然性、不确定性和不可预测性，发生后往往具有报道传播力强，影响范围广的特点，此时若不加以重视，则可能造成较大的社会危害。疫情刚有好转，夏季暴雨袭来，河南省降雨量大幅度超过往年降雨量值，造成了重大人员伤亡与经济损失。此事一经发生，网络社交平台上关于“暴雨救援”、“人员伤亡”以及“祈祷”的观点信息大幅度爆发，给疫情防控增加了不少的网络舆情监控压力。因此本文选取“河南暴雨”事件中重要时间节点的网民评论加以分析，了解网民用户的情感倾向、关注焦点，以及用户情感的整体区域分布特点，多角度对事件进行综合剖析，以求探寻该类事件的时空演化特征，以此为政府及相关监管部门提供科学合理的舆情引导对策，提升政府公信力，构建和谐社会。

1.2 国内外研究现状

1.2.1 国外研究现状

笔者在 Springer Link、Science Direct、Emerald、EDS 资源发现系统和 Web of Science 等数据库中，以“Network Public Opinion” “Sentiment Analysis” “Network Public Opinion and Sentiment Analysis” “Network Public Opinion and Spatial Correlation” 和 “Network Public Opinion and Sentiment Analysis and Spatial Correlation”为检索关键词进行检索。将与检索关键词关联度不高的文献筛选以后，得到国外文献检索结果[①]（见表 1.1）。

笔者根据已获得的国外检索结果，经过对相关文献的筛选、整理与阅读，发现国外学者关于网络舆情、情感分析和空间相关性这三方面的研究重点具体如下。

① 检索日期为 2022 年 3 月 1 日

表 1.1　国外文献检索结果①（单位：篇）

检索词 \ 数据库	Spinger Link	Science Direct	Emerald	EDS 资源发现系统	Web of Science
Network Public Opinion	292	465	5	4975	485
Sentiment Analysis	16880	3291	82	1514390	16443
Spatial Correlation Analysis	652	13676	0	4095	506
Network Public Opinion and Sentiment Analysis	72	71	1	409	35
Network Public Opinion and Spatial Correlation	20	2	0	94	9
Network Public Opinion and Sentiment Analysis and Spatial Correlation	15	0	0	54	0

（1）网络舆情研究现状

相比较国内对于网络舆情的研究起始时间，国外的相关研究可向前追溯到18世纪。最早由卢梭提出“公众意见（public opinion）”的概念，不少学者在该概念的引导下开启了舆情的相关研究。发展到20世纪初“舆情”的概念被广泛使用，此后网络舆情研究取得可观的成果：Sarah[10]先后通过爬虫获取Twitter用户关于“俄克拉荷马州大火”和“红河洪水”的推文，并结合用户发文所在地理经纬度、发文当天天气状况和用户关注的热门话题等进行建模，评估舆情危机状况，识别态势感知信息。Ratkiewicz[11]将内容分析、网络拓扑结构分析相结合，构建了一个基于机器学习的文本信息分析预测系统，目的在于实现政治选举的内容预测。通过对Twitter用户所发布的相关文本内容进行实证后发现准确率高达96%，肯定了所选用算法的准确性，证明该系统的有效性。Sudha Verma等人[12]则以自然语言处理技术为手段对Twitter用户所发布的文本实现了主客观信息进行分类，在此基础上计算出热点事件的高频词，用于对网络舆情事件的危机评估。分类器的计算准确率超过80%，这证明再次遇到相类似情况时实验中已构建的分类器能够很好的对该类事件进行有效处理。Truphi M等人[13]通过调用API接口，利用

① 备注：针对不同的外文数据库，笔者选择了不同的检索匹配方式。Spinger Link数据库中以“with all of the words”方式进行检索词检索；Science Direct数据库中以“Title, abstract or author-specified keywords”方式进行检索词检索；Emerald数据库中以“title”方式进行检索；EDS资源发现系统中以“关键词”方式进行检索词检索，Web of Science数据库中以“topic”方式进行检索词检索。

Hadoop 技术实现了用户的情绪归类与预测，针对性修改营销策略，达成了营销策略的改进目标。K Zheng[14]立足于公共安全领域的现实需要，提出一种基于ICTCLAS 的网络舆情热点信息自动检测法，并通过实例证明该方法具有较大的实用性和可靠性。

（2）情感分析研究现状

情感(Sentiment)一词最开始起源于心理学领域，心理学界认为情绪和情感都是人们对客观事物所持有的态度，二者所不同的是情绪更倾向于个人的原始生理需求，而情感则更倾向于社会需求。网络舆情并不具备“情感”这一要素，对网络舆情进行情感分析，指的是对不同类型的社交平台中已公开表达的数据文本进行挖掘并提取有效的观点数据的研究[15]，泛指对带有主观性文本的进行分析、处理、归纳和推理的过程。

情感分析研究最早可以追溯到上个世纪 90 年代，1995 年 Rosalind Picard 教授融合了计算机科学、认知科学与心理学的学科知识点，创新地提出“affective computing”概念，为情感分析的研究热潮打下了扎实的理论基础。另一方面，借助着科学技术水平，网络社交平台的交流效率极大提高，为进行情感分析研究提供了现实基础；海量的文本数据迅速积累，为情感分析提供初始数据基础，越来越多的学者投身于情感分析的研究领域，并取得不斐的成果：Alam 等人[16]要求志愿者以面部表情的形式分别表现出“Happiness、Surprise、Sadness、Fear、Disgust、Anger”等六种情绪，目的在于探析 10 个不同文化国家的人民在情绪表达中的共性与差异。T Wilson[17]研究发现若不考虑语义环境，脱离上下文地去考虑词语极性，则此时词语一律表现为中性，这一做法极为不妥。因此他带领团队成员通过比较多种机器学习算法，实现了对积极、消极和上下文极性的特征识别，从而实现对文本数据进行更贴合实际情况的分类。Dey[18]设准确性、召回率和精密度为评价标准，利用两种机器学习算法对两个不同类型的文本数据进行情感分析。Soelistio[19]在文本挖掘的众多应用领域中选取了数字报纸为媒介，利用朴素贝叶斯分类算法对文章作者的情感态度进行倾向性研究，经过不断更新迭代，可以持续获得关于某政治家的最新情绪态度。Agarwal[20]和 Vishal Kharde[21]分别基于 Twitter 平台的用户社交数据进行情感分析，探索数据背后的深层信息含义。

（3）空间相关性分析研究现状

1969 年 Tobler 提出“地理学第一定律”，即“任何事物都是空间相关的，距离近的事物的空间相关性大”。该定律的提出逐渐拉开了地理现象中空间相关性和异质性特征研究的帷幕。Patrino 等[22]利用空间统计与莫兰指数研究乔治亚西北部区域发电站站点间距离、风向和地形的空间相关性，揭示了风暴路径与目标地区降雨的关联。Russell G[23]以视觉检查和空间自相关分析为手段，对同一地区

农业面积、牧场面积和森林面积三种不同空间复杂度的数据集进行实证研究，以发现不同数据生成地图的误差，从而进行土地覆盖图的精度评估。S.Shadkhoo[24]在加州地震发生后对地震数据进行空间相关性研究，并得出研究结论。可见国外的空间相关性分析多聚集在地理、环境等自然研究领域。

1.2.2 国内研究现状

本研究在中国知网、中文科技期刊（维普）和万方数据知识服务平台等数据库中进行文献检索①（见表 1.2）。检索主要按照主题进行，关键词确定为“网络舆情”“情感分析”“空间相关性分析”“网络舆情 and 情感分析”“网络舆情 and 空间相关性”和“网络舆情 and 情感分析 and 空间相关性”。

表 1.2 国内文献检索结果（单位：篇）

数据库 / 检索词	中国期刊全文数据库	中国硕士学位论文全文数据库	中国博士学位论文全文数据库	中文科技期刊（维普）	万方数据知识服务平台
网络舆情	15457	2570	88	10356	24075
情感分析	8141	3829	273	2799	205436
空间相关性分析	581	118	10	19127	49737
网络舆情 and 情感分析	437	194	19	77	1623
网络舆情 and 空间相关性	1	0	0	0	52
网络舆情 and 情感分析 and 空间相关性	0	0	0	0	9

笔者通过对检索到的文献进行整理与分析，发现国内学者的研究主要集中在以下方面：

（1）网络舆情研究现状

随着大数据和云计算技术的飞速发展，国内越来越多的学者开始关注网络舆情技术和建模这一领域。任立肖[25]通过梳理多种模型的发展历程，证实复杂网络对于网络舆情模型的研究具有重要意义，认为目前已有的演化模型大多忽略网络有向性，未来的努力方向是在目前的基础上构建实现新的有向网络演化模型；冯兰萍[26]指出网络舆情演变是“政府行为”和“主流情绪”的共同作用成果；刘继[27]认为网络信息在社群空间中的传播演化已经由用户“一对一单传播”发展到一个用户面向多个用户群的“N+M”型多传播机制；李燕凌等[28]由“黄浦江死猪”案件得知事件的演变同时受媒体、政府、公众者、生产者等因素影响，需要对关

① 检索日期为 2022 年 3 月 1 日

键时间节点重点把握。袁国栋[29]引入卷积神经网络并构建演化模型，较好地拟合出六个阶段内舆情整体的演化规律。由此可见，国内外网络舆情研究多着眼于应用研究，缺乏对传播过程的演化机制研究。

（2）情感分析研究现状

情感分析又被称为倾向性分析，是指公众对个体、事件等特定目标的主观情感判断，通常以社交媒体平台为载体，借助文本、声音、图像和视频为表现形式来传播情感、表达个人看法和意见，并使用特定的情感符号、语言格式和不同的语言功能来呈现和增强情感表达。随着互联网的日益发展，在线交流变得频繁和快捷。如何收集公众意见，整合和分析公众的态度和情感倾向，加强对互联网的监控变得愈发重要。虽然国内的情感分析起步较晚但仍取得众多成果，从内容上看主要包括主客观分析、多情感分类，和文本情感极性分析。

通过对文献梳理发现目前的情感分析的研究方法主要分为两种：基于情感词典的分类方法和基于机器学习的分类方法。深度学习的分类方法作为机器学习的分支被应用广泛。

①基于情感词典的分类方法主要以文本中带有情感倾向的情感词为基础。根据现有的情感词典或构建新的情感词典，计算文本中情感词与关联信息来达到文本情感极性分析的目的，因此构建和选择情感词典是此类情感分析方法的工作重点。王志涛等[30]整合大规模微博数据后扩充出专属于微博领域的情感词典，并在考虑句子规则的前提下构造出新的情感计算方法。栗雨晴[31]考虑到民众乐于中英文搭配的表达习惯，构建出一个基于双语词典的情感分析模型，能够有效捕捉群体深层意见。李钰[32]首先选择并评估了 Word Net、How Net 和大连理工等经典的情感词典，对它们进行了整合、筛选，构成基础情感字典，最后引入扩展情感词、表情符号、程度副词等最终构成四个情感词典。②基于机器学习的方法主要通过支持向量机(SVM)、朴素贝叶斯、最大熵马尔科夫模型等成熟的文本分类方法来监督、训练大量已处理的文本数据，并提取其中的情感关键词和主题等情感表达特征的文本，该方法的重点和难点在于利用技术手段提取文本中的情感特征。唐慧丰[33]通过实验比较了四种文本分类方法：中心向量法、KNN 算法、贝叶斯分类算法、支持向量机算法，得出综合信息增益(IG)和支持向量机(SVM)的文本分类方法对情感特征提取的效果最好。李寿山等[34]提出将支持向量机(SVM）、最大熵(Maximum Entropy)、朴素贝叶斯(Navie Bayes)、随机梯度下降(SGDA)四种文本分类方法叠加到常用的单一文本分类方法中，叠加后的综合算法在文本情感特征提取方面的效果均优于四种算法单独进行提取。基于深度学习的方法主要是基于词向量模型，采用深度学习的某些模型对文本的句子和篇章的关联信息中表达的情感进行学习，达到情感分析的目的。相关的深度学习模型包括长短记忆网络、

卷积神经网络等。朱晓霞等[35]结合 TF-IDF 和 K-means 聚类方法，通过人工标注，利用情感词典和相关信息计算情感极性值。后又构建一种基于主题-情感挖掘模型的情感分类方法。实验结果表明，与实验中的基准模型相比，该方法的分类准确率提高 14.24%。胡西祥[36]基于词向量模型，对比朴素贝叶斯模型和 LSTM、CNN 模型的文本情感极性分析结果，结果显示深度机器学习模型准确率优于机器学习模型。

（3）空间相关性分析研究现状

国内的空间相关性研究成果中，武春燕等[37]利用 GIS 系统对山东省卫生资源空间分布状况进行空间相关性分析，结果表明省内的卫生资源分布呈现由东向西的倾斜趋势，西部地区最低，表现为较高程度的结构失调。同样是山东省，吴桢[38]和苗晓颖[39]分别率领团队，探索海底鱼类资源、蔬菜种植格局是否存在空间异质性特征。任志远[40]利用 ESDA 方法，以两期人口普查数据为基础，分析了陕西省 10 年间人口分布变化特征。吴桐[41]构建了以旅游发展效率、环境质量、交通基础和公共服务输出的旅游发展质量评价指标体系，结合熵值法、莫兰指数和障碍度诊断分析法，综合分析、比较影响西部地区旅游业发展的具体障碍因素。施益强[42]通过厦门市区二氧化硫、二氧化氮、PM2.5 和 PM10 的数据指标对当地的污染浓度进行空间探索，对厦门市区行政区的调整提出更为合理的建议。综上可知，空间相关性探究的技术和手段均已成熟，在旅游、经济、城市发展、环境等各方面取得可观的成果。空间相关性理论作为理论基础，ESDA、ARCGIS 平台提供技术手段，为网络舆情发展的空间分析提供了可行性与现实意义。因此，本文借助 Arcgis10.8 平台中的空间统计分析工具，对网络舆情发展过程中是否存在空间相关性进行探究，如果存在相关性，又表现为怎样的特征。

1.2.3 国内外研究述评

综上所述，许多学者对网络舆情及其发展进行了跨学科、多角度的探讨，特别是充分探讨了用户情感交流和社交平台舆论演变的规律。这为本文后续实验中的用户情感识别、用户情感特征识别和情感的时序变化提供了丰富的理论和技术支持。尽管已有的研究文献硕果累累，然而仍存在以下不足：第一，国内外学者多以技术探究为主，缺乏对网络舆情整体演化规律的理论性描述；第二，国内外空间自相关研究多着眼于自然科学领域，较少涉及人文社科领域；第三、网络舆情的相关研究鲜少涉及空间相关领域。在当前“地球村”中，信息扩散迅速、网络影响增长快，结合时序分析的网络舆情空间分析研究更具有必要性和现实意义。

1.3 研究思路与研究方法

1.3.1 研究思路

舆情发展是一个动态变化的过程，网民在不同阶段内的关注点、情感和关注群体都会发生动态改变，对舆情发展的动态演化过程研究和对可能发生舆情事件的演变进行预警具有指导意义。本文在确定待研究的目标事件后，首先通过网络爬虫算法获取文本数据，通过相关算法提取出相关的重点舆情主题和用户情感特征，其中用户情感特征指的是事件曝光后网民对于该事件的观点、看法以及情感态度的总称，它包含用户情感倾向特征和地域特征。其次对事件情感的演化展开时序分析，分析情感倾向变化规律，对重点舆情主题演化和用户情感演化进行原因的深度挖掘；最后根据全国网民对于舆情传播过程中的关注程度变化与空间差异特征为相关部门制定引导对策提供依据。

具体研究思路是：①选取 2021 年 7 月“河南暴雨”事件为研究案例，收集新浪微博上关于该事件的有关文本，并对其进行整理与分类；②对采集获取的原始数据进行清洗、规范化处理后得到样本数据，调用 Python 中的第三方库 SnowNlp，得到情感倾向性概率分布；③通过 SPSS24.0 验证微博转发数、评论数和点赞数三者之间是否具有关联性，从而确定一个指标为舆情传播指标，根据该指标绘制特定时间区间内的评论文本传播趋势图。并结合生命周期理论，实现网络舆情传播阶段的划分；④对事件中网民用户的情感演化、热点舆情话题演变进行时序分析；⑤基于用户情感特征中的地域特征，深入剖析了各阶段内不同省份用户情感状态、热点舆情话题和评论的空间聚集性，探索了用户情感的空间分布特征，为向相关部门提供针对性的舆情回应和舆情引导建议提供支持。

1.3.2 研究方法

（1）文献研究法。本文对国内外关于网络舆情、情感分析和空间关联的相关文献进行了检索和深入分析，梳理三者的概念，理解了三者之间的联系，在了解当今研究热点以及技术的前提下，选择合适的技术手段来实现对应的实验目标。

（2）案例分析法。本文选取焦点事件作为研究案例，进行实证研究，通过对案例的综合分析，得到突发公共事件网络舆情中网民（微博用户）的情绪演变特征，并进行时空分析，为政府部门应对此类事件提供决策支持。

（3）量化分析法。本文通过量化“河南暴雨”事件发生后新浪微博发布的全部数据，不仅获取官方微博的发布内容、发布时间，还爬取了评论文本内容和发表评论的用户个人信息（包含评论用户 ID、评论发表时间 、评论用户年龄和评论用户所在地区），通过数据挖掘方法和信息可视化方法提供了更直观的结果，有利于研究网络舆情情感的时空演变。

1.4 研究内容与论文创新点

1.4.1 研究内容

本文主要以新浪微博为载体，对突发公共事件微博舆情的情感演化进行研究，内容划分为六章，主要包括：

第一章：绪论。对论文选题的研究背景和研究意义进行阐述，并总结国内外相关研究文献，提出研究思路和创新点。

第二章：概念界定与理论基础。对突发公共事件网络舆情、微博舆情的相关概念与特点进行梳理，并介绍了生命周期理论、沉默螺旋理论、议程设置理论和空间自相关等理论基础，阐述了常见的情感分析技术方法、关键词抽取算法和话题挖掘算法等数据挖掘方法。

第三章：以“河南暴雨”事件为例的情感时空演化分析。本章在确定了研究案例后，获取自事件发生至平息后新浪微博上的相关网民评论文本和发文用户个人信息两份原始数据，经过分析处理得到实验所需的样本数据。首先对样本数据中的评论文本进行清洗与预处理，将处理过后的干净数据利用 Python 中的 SnowNlp 库实现对干净数据的情感倾向分析；其次，根据 Pearson 相关系数验证微博转发量、评论量和点赞量之间是否具有相关性，而后根据网民评论文本绘制每日数量变化趋势图，结合生命周期理论实现舆情传播阶段的划分；最后，围绕关键词抽取方法、话题识别方法和空间相关性统计方法展开说明。

第四章：“河南暴雨”事件情感时空演化结果。依照第三章的研究框架，从用户情感、热点话题实现对微博舆情的时序分析；从用户评论的数量分布探寻用户关注度的分布特征；从用户评论情感值分布是否存在明显的舆情情感空间，根据这三方面探寻微博舆情情感态势演化的时空特征，解读舆情演化的内在机理，以求得出普遍规律。

第五章：突发公共事件中微博舆情引导对策。从网络舆情传播主体、时间分析结果和空间分析结果三个角度为突发公共事件网络舆情的引导、管控提供策略支持。

第六章：研究结论与展望。总结概括本研究所做的具体研究，提出自身研究的不足之处，为今后相关的研究立下基础。

1.4.2 论文创新点

（1）从空间分析的角度进行网络舆情演化研究

互联网的实时性、海量性以及“一对一、一对多、多对一、多对多”的多向性、互动性传播方式，使得网络舆情数据具有较强的时效性。加上因网络舆情控制不得当而引发的网络舆情危机成为了近年来的新涌现话题，针对网络舆情的演化、

引导成为各级政府部门的重点关注点。然而，目前关于网络舆情的空间演化研究较少，多集中于时间序列分析角度。本文结合空间统计方法，引入时间与空间两个因素进行全面分析：时间上分析舆情传播过程中主题演化、用户情感倾向时序变化；空间上根据用户所在地区挖掘用户间关注程度、情感分布特征，分区域把握舆情态势，并剖析深层次原因，针对性制定调节、引导、防护策略。

（2）案例特殊性

河南省遭遇极端强降雨，多个国家级气象观测站日降雨量突破有气象记录以来历史极值，百年一遇的大暴雨造成了巨大损失。自然灾害频发对于我国政府应急管理工作来说是一个挑战，不仅要投入大量人力物力资源抗灾救险，同时也要加强网络舆情的监测与分析。“河南暴雨”事件总体是“天灾”，具体有“人祸”，特别是发生了地铁、隧道等本不应该发生的伤亡事件。针对该案例的研究可拓宽本领域的素材范围。

第 2 章　概念界定与理论基础

2.1　相关概念界定

2.1.1　网络舆情

网络舆情作为社会科学、计算机科学和心理学等多学科的交叉领域，不同学者从不同学科角度对“网络舆情”的相关理论进行了阐述和补充。杜骏飞[43]提出网络舆情是民众关于社会现象的观点集合在互联网上的一种形式；刘毅[44]指出网络舆情是公众借由互联网表达和传播的，对与自身利益紧密相关的各种事物所持有的态度总和；王国华[45]认为网络舆情是各类事件推动刺激而产生的借助互联网平台传播的网民对于该社会现象所持有的不同意见、态度、情绪和行为倾向的集合；李昌祖[46]将网络舆情概括为作为舆情主体的民众对国家管理者所产生和持有的社会政治态度。由此可见，网络舆情的本质是民众关于社会问题的个人情绪，或意见所构成的集合，并以互联网为媒介进行传播。如今，随着通信技术的不断进步，互联网已然成为公众言论和舆论的重要传播载体。网络舆情的传播速度加快，影响范围扩大是必然结果，网络舆情的特点可被归纳为以下几点：

（1）匿名性。所有的在线网络平台上，公众注册账号后通过自己设置昵称来进行交流，用户间无法知道对方的真实身份，言论内容与社会责任短时间内难以实现匹配，相关从业人员的信息分析难度较大[47]。不排除居心不良之人通过不当言论的发表，制造并传播谣言，导致负面信息不断发酵至形成网络舆情危机。

（2）偏差性。网络舆情的匿名性影响了法律道德的约束效果，如果网民缺乏自律，就会直接导致言论的偏激与非理性，造成群体盲从与冲动。很多网民面对现实生活中的挫折，或者对社会问题的片面认识都会通过网络加以宣泄。

（3）指向性。网络舆情的发展演化过程中，可发现受网民认可度高的观点基本趋于一致，反映了大众的普遍认知。该类看法、观点多由意见领袖所发表，具有较强的指向性。此时若存在不同群体对同一事物发表了不一致的看法，则易出现极端言论从而造成舆情危机。

2.1.2　微博舆情

微博舆情作为网络舆情的重要组成部分，是反映民意的舆情在微博平台的集中体现[48]，在一般舆情具有的普遍特征之外，微博舆情还具有一些自身独有的特点。具体如下：

（1）传播主体多元化。首先，由于使用微博平台的用户不受年龄层次、性别区别、工作种类的限制，在用户上呈现出平民化的特点。只要用户拥有移动设

备并连接到网络，就可以参与微博上的各种讨论。这就造成微博舆情在真实性和可靠性上存在不足。其次，由于网民用户的教育背景、社会地位和成长环境等不尽相同，因此他们在事件发生后所传递的观点和做出的行为往往存在差异。同时网络平台特有的虚拟性和匿名性赋予了用户更高的自由化和随意性，不少人直接转发消息，无形中充当了谣言的散播者。在这种情况下，如果政府部门无法及时介入，进行相应的引导，就极可能会造成重大的社会问题。

（2）传播速度即时化。一旦微博舆情发生，微博平台提供的转发功能就会为信息的迅速传播提供基础，实现信息的同步更新，爆炸式的舆情传播也会因此发生。这种传播可以在极短的时间内扩散到极其广泛的范围，由此对平台受众产生影响，甚至会影响其价值观的树立。

（3）社会效应扩大化。在“互联网+”不断发展的社会背景下，越来越多的民众拥有了较强的互联网意识，也更加依赖网络平台来实现信息的获取与传播。微博作为当今中国最为主要的社交媒体之一，受到众多民众的青睐，产生许多的信息传播行为。当某一舆情事件发生，往往会在微博平台上迅速扩散并引发广泛讨论，引发高度的民众关注度，微博舆情因此会产生广泛的社会效应。微博舆情有利有弊，如果可以正视并加以引导，就可以尽快安抚民众情绪，迅速解决舆情事件，并通过这样的方式提高民众对于政府部门的信赖度。反之，就会扩大民众的负面情绪，催生群体性事件，引发更为严重的舆情事件，对于社会稳定造成不良影响。

（4）意见领袖重要化。微博平台的意见领袖指的是当舆情事件发生后，能够对该事件的发生、扩散和演化过程中造成重大影响的用户。这些意见领袖能够在特定的领域中形成超越常人的影响力，可以辐射到庞大的用户群体。微博意见领袖的影响力可以通过现实世界衍生，因为微博平台可以提供实名认证的功能，帮助普通用户辨别意见领袖的真实身份。除此之外，舆情事件当事人、明星群体和具有较多粉丝基量的普通用户也可以在事件中形成强大的能量。正是由于微博平台具有的广泛的用户群体和快捷的传播速度，在目前的网络平台中占据着极为重要的作用，微博舆情也因此具有日趋强大的影响力。基于此，政府部门需要及时介入微博舆情，通过技术手段形成日常的监测机制，在舆情事件发生后进行正确引导，控制虚假舆情的扩散。

2.2 相关理论基础

2.2.1 生命周期理论

生命周期的概念最早产生在心理学领域，指人或者家庭由出生、成长、衰老、

生病到死亡的全过程，后来逐渐演变成泛指事物从产生到生长，直至最后消亡的整个过程。与大多数生命体的生命周期相似，网络舆情事件也会经历一个从产生到灭亡的过程。因此生命周期理论作为认识网络舆情的重要理论之一，往往成为了研究者们划分网络舆情生命周期的参考标准之一。并且由于对演化阶段划分没有定论，不同的研究间存在着一定的差异，表现为研究成果从“三阶段”到“六阶段”不等（见表 2.1），这为本文后续研究的案例——“河南暴雨”事件网络舆情情感演化研究中的阶段划分提供理论基础。

表 2.1　网络舆情阶段划分

代表人物	阶段划分
刘国威	酝酿期-爆发期-衰退期[49]
于兆吉	孕育期-传播期-扩散期-衰退期[50]
梁冠华	潜伏期-发生期-持续期-恢复期[51]
张宇	初显期-爆发期-波动期-降温期-长尾期[52]
任凯	潜伏期-扩散期-爆发期-波动消退期-衰退期[53]
崔鹏	酝酿期-爆发期-扩散期-反复期-消退期-长尾期[54]
曹天阳	征兆期-高峰期-持续期-恢复期-复发期-平稳期[55]

2.2.2　沉默螺旋理论

20 世纪 70 年代的德国社会学家伊丽莎白.诺尔.诺伊曼从社会心理学角度首次提出“沉默的螺旋”假设（The Spiral Silence）。该假设认为“社会民众出于对被孤立的恐惧，而保持沉默，对多数或优势意见产生趋同感”，这说明多数观点会更加得势，而少数观点会更加沉默，从而形成一个“螺旋式”的传播过程[56]。在新媒体环境下，由于用户可匿名性、传播主体多元化等多种特征，越来越多的用户不再局限于仅仅做信息的接纳方，转而选择在网络平台发声，打破了“优势意见”为主流的信息传播局面[57]。因此，沉默的螺旋理论受到理论界的质疑,不少学者提出不同看法。刘建明[58]认为人作为具有能动性的社会主体，常常以反沉默螺旋方式发表意见；黎勇[59]指出反沉默螺旋假说能够帮助大众理解和认识“沉默螺旋”及舆情演变中一些难以解释的现象，“舆情反转”就是具体表现之一。然而近年来，网民通过转发、点赞等方式对他人观点表示支持。这样一来，热度高的

观点往往排在前面，用户首先接触的就是这些信息，即主流意见。因此，尽管民众由“不知道、无所谓和不感兴趣”转变为了“我知道、我想说和我感兴趣”，但由于双方话语权的悬殊，意见领袖举足轻重，民众观点越难以广泛传播，“沉默的螺旋”越奏效，尤其是在强调集体主义的国家[60]。总的来说，沉默螺旋理论为政府针对性引导舆情发展提供重点，为引导策略的提出提供了理论基础。政府应当辩证地看待“沉默的螺旋”和可能出现的“反沉默的螺旋”，通过正确的措施让其各自发挥作用，推进舆情的发展。

2.2.3 议程设置理论

议程设置理论的萌芽始于 20 世纪中期，在拉扎斯菲尔德和默顿所提出的“大众媒体具有地位赋予功能”的概念中，认为大众媒体的报道具有一定的地位性，可以通过报道使得社会问题、组织问题、私人问题和社会运动等各类问题引发广泛的关注，同时这种关注度的产生是合法且合理的。在新媒体广泛运用的社会大背景下，信息接收者多元化的特点十分明显，民众不再是单一的接收群体，他们在信息传播中拥有了更强的参与度，也更加乐于主动参与，自身便担当起传播者的角色。在这种情况下，议程设置的概念便得到延伸和优化。和以往的议程设置相比，新的议程设置受到新媒体的影响而呈现传播主体和传播渠道多元化、传播效率即时化的新特征。

（1）传播主体。在新媒体的社会大背景下，社交平台层出不穷，网络自身具有的便捷性使得民众拥有了更多可选择的发声渠道，这一新特点改变了传统模式下发声渠道单一的特征，每一位民众都可以成为信息发布的主体，可以成为信息传播主体。新的议程因此也会具有传播主体多元化的特征。（2）传播渠道。和以往传统的议题传播渠道不同，传统模式下的传播渠道主要是报纸、电视和广播，而在当今，相同主题的信息可以通过各类网络平台快速传播，这种传播方式呈现出高效即时的特点。（3）传播效率。在传统媒体盛行的时代中，出于传统媒体话语权较大的特点，信息发布的主观性较强，发布的信息并不是非常完整的，议题设置也比较滞后。而在新媒体的背景下，议题可以借助互联网的力量通过不同渠道发布，信息受众因此可以广泛地参加讨论，由此极大地提升议程的传播效率。

2.3 数据挖掘技术

2.3.1 情感分析技术方法

情感作为心理学用词，主要包含道德情感和价值情感。《心理学大辞典》中将其定义为“情感是人对客观事物是否满足自己的需要而产生的态度体验”。生

活中的情感表面上是人的喜怒哀乐，实际上是个人生理感官、心理和精神层面的表达。

情感分析指的目的在于通过对公众的观点、意见进行分类，方便有关部门有针对性、个性化地引导和管理网络舆情，维护网络秩序以构建清朗的网络空间，从而反作用于社会安定。根据待处理文本的粒度差异，情感分析可按从小到大分为词汇、句子和文章三个等级的情感分析。

最早使用的情感分析方法为基于情感词典的方法。通过对常见情感词的收集、归纳与整理，人工构建出情感词典，后期将实验文本与词典匹配，判断文本的情感极性。然而直接使用情感词典的分析方法具有耗时耗力、依赖性强的弊端，因此众多学者在传统情感词典的研究基础上进行了不同程度的改进。

Pang[61]利用朴素贝叶斯、最大熵和支持向量机三种常见的机器学习方法对电影评论进行分类分析。但单纯的机器学方法效果仍具有局限性，表现为需要通过人工标记来识别句子是否有主题相关特征。为避免人工标注带来的主观依赖，作为机器学习分支的深度学习方法被引入网络舆情研究领域。该类学习方法通过使用神经网络模型的方法对数据进行学习，自动提取大量文本特征，能够提高算法有效性。Massa Baali[62]通过构建深度卷积神经网络模型实现对阿拉伯语言推文的较好分类，且效果优于传统机器学习方法。2018 年末，谷歌推出以 transformer 为核心的 Bert 模型，可有效解决 RNN 模型中顺序依赖、无法并行计算的问题。该模型能够更好地获得文本特征，对文本进行分类，因此该模型被广泛推广。

除了机器学习、情感词典和深度学习等情感分析方法，SnowNlp 作为一个专门处理中文的类库，能够通过调用库直接实现中文分词、情感分析、文本分类、词性标注和文本相似度计算等操作。贝叶斯模型是 SnowNlp 库中情感分析的核心代码的基本模型，该模型作为一种来源于数学领域的生成式模型，其基本思想是通过先验概率得出后验概率，以达到分类目的[63]，学者们基于贝叶斯模型的诸多成果验证了贝叶斯模型的有效性[64,65,66,67,68,69]。本文通过调用 SnowNlp 库，完成评论文本的分类，实现对用户情感倾向的有效识别。

2.3.2 关键词抽取

关键词抽取作为文本挖掘领域的一个分支，是文本检索、文档比较、摘要生成和文档分类的基础性工作，按照有无人工标注训练样本可分为：有监督抽取方法、半监督抽取方法和无监督抽取方法三种。由于人工标注成本较高，因此无监督的关键词抽取方法成为研究人员的首要选择。无监督关键词提取方法主要分为：基于统计特征的关键词提取、基于词图模型的关键词提取和基于主题模型的关键词提取三种。三种关键词提取算法的区别在于：（1）基于统计特征的关键词提取

算法的核心思想是利用文本中词语的统计信息从文本文档中提取关键词。（2）对于基于词图模型的关键词提取，首先要构造就文档的语言网络图，然后对语言网络图进行分析，寻找到图中具有重要作用的词或短语作为关键词。（3）第三种提取算法主要是通过主题模型中关于主题分布的性质来提取词语。常见的关键词提取算法有：TF-IDF 算法、TextRank 算法、基于 Word2Vec 的词聚类算法、信息增益提取算法、互信息提取算法和卡方检验关键词提取算法等。TF-IDF 算法因为其简单易行，且分类效果好而成为最常见的关键词抽取算法。因此本文选择 TF-IDF 进行关键词抽取，获取用户在舆情演化过程中的关注焦点。

2.3.3 话题挖掘

如果说关键词分析能够从细粒度的词语角度刻画网络舆情演变的普遍性特征，热点话题的识别更能够揭示舆情的动态传播变化。媒体在事件发生后相继跟进，以报道的形式传递信息，民众选择性地围绕报道发表自己态度。此时，在媒体、民众和事件本身这三个主体间形成了一个闭环互动，即事件的相关信息不断发酵后经由媒体传递给民众，民众的意见发表促进相关热点话题的产生与迁移，对事件的舆论走向产生影响。因此，对影响力高的话题进行识别与挖掘逐渐成为了研究的重点。

话题检测与追踪（Topic Detection and Tracking ,TDT）技术发展过程中，最先由 Blei[70]提出 LDA 模型，将每一条文本转换成包含多个“Word of Bag”的词频向量。作为一种发展较为成熟的话题聚类算法，LDA 算法被广泛应用于网络舆情研究内，以实现舆情变化过程中的话题识别与追踪[71,72,73]：刘玉文[74]在 LDA 模型中融合进位置参数，完成了话题地域特征的识别与演化研究；陈阳[75]通过 UR-LDA 模型实现对联系人关联关系出发，更好地处理用户关系数据；庄穆妮[76]利用改进后的 LDA-ARMA 模型实现了对“香港修例”风波的舆情预测。

2.4 空间自相关分析

空间自相关的概念来自于时间序列的自相关，描述的是空间中某一位置与其相邻位置上同一变量的相关性，即分析空间事物的分布是否具有关联性，以空间自相关统计量来衡量相关程度。在舆情空间内，不同区域的用户情感相互影响。因此，本章节内基于空间自相关的理论基础，将不同地理区域内的用户情感值作为该空间区域的舆情情感属性，分析不同空间区域间的舆情情感的分布关联性。

Moran’s I 统计量和 Geary’s C 统计量作为空间自相关领域中的两大基础统计量，其目的在于通过比较邻近面积单元的值来统计是否具有相关性。当值相似时，则表明两个面积单元间存在较强的正相关；若值不相似，则说明存在负相关。本

文采用 Moran's I[①]统计量来验证全局空间自相关。

莫兰指数统计量在计算过程中需要假设区域中有 n 个面积单元，记第 i 个单元上的值为 y_i，$\bar{y}$ 为 y_i 在 n 个总面积单元中的均值，则莫兰指数计算公式如（1）：

$$I=\frac{n}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\left(y_i-\bar{y}\right)\left(y_j-\bar{y}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}} \qquad \text{式（1）}$$

右边的 $\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\left(y_i-\bar{y}\right)\left(y_j-\bar{y}\right)$ 是一个协方差，邻接矩阵 W 和 $\left(y_i-\bar{y}\right)\left(y_j-\bar{y}\right)$ 相当于规定 $\left(y_i-\bar{y}\right)\left(y_j-\bar{y}\right)$ 对相邻的单元进行计算。若 i、j 两个单元处于相邻位置，则 y_i 和 y_j 同号，即 I 值为正，否则 I 值为负。莫兰指数的有效范围是（-1，1）。当 I 为 0，表明二者在空间上不相关；当 I 小于 0，则二者存在相关性，且为负自相关，反之则为正自相关。

莫兰指数统计量分为全局莫兰指数和局部莫兰指数，公式（1）是用来分析有没有空间自相关性存在的全局莫兰指数，局部莫兰指数依据公式（2）来探测异常值或聚集范围。

$$I_i=\frac{y_i-\bar{y}}{\frac{1}{n}\sum(y_i-\bar{y})^2}\sum_{j\neq i}^{n}w_{ij}\left(y_i-\bar{y}\right) \qquad \text{式（2）}$$

I_i 代表目标区域中第 i 个地区的局部莫兰指数，W_{ij} 依旧为邻接权重矩阵。将 I_i 的计算结果可视化，可得到莫兰散点图（见图 2.1）。

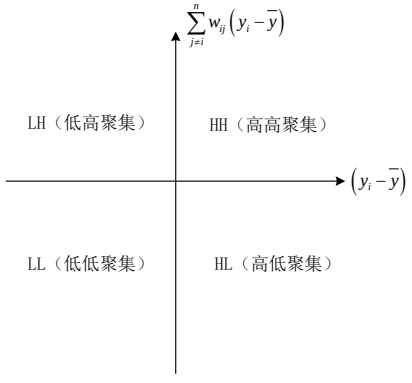

图 2.1　莫兰散点图

HH 聚集区代表当前单元值为高，且周边单元值也为高；LL 聚集区代表当前单元值为低，且周边单元值也为低；LH 聚集区代表当前单元值为低，但周边单元值为高；HL 聚集区代表当前单元值为高，但周边单元值为低。即：一三象限说明空间相似值聚集，二四象限则表明空间异常。

①Moran's I：官方称为莫兰指数，由澳大利亚统计学家帕克.莫兰于 1950 年提出。本文后续将“Moran's I”统一记为“莫兰指数”。

第 3 章　以“河南暴雨”事件为例的情感时空演化分析

由于网络具有隐匿性、开放性和匿名性，网民面对突发事件时容易受到虚假信息的干扰而失去理智，最终选择错误的方式在社交平台上传播负面情绪，负面情绪互相感染后并传播，引起舆论波动，最终导致负面舆情大规模爆发。因此，有关网络舆情的演化分析成为不少学者的关注重点，基于社交平台上用户评论的舆情分析，有助于政府机关尽早掌握群众的关注重点，更好地构建“群策群力、群议群定”的立体化、多层次全民交流网络，促进相关舆情应对和舆论引导工作由事后管理、被动管理向事前管理、主动管理的全面转变[77]。

微博舆情的相关研究中，郭耿在分析微博舆情网络结构特征的基础上提出复杂网络模型，并将模糊理论与其结合，二次改进后重新提出模糊相似度观点演化模型，并通过实证分析得出结论用户对于热门发现微博的关注度对舆情传播有影响[78],；姚翠友通过构建社会事件的舆情演化模型并以“于欢案”进行仿真模拟，研究了应当如何更好地发挥舆情引导和社会管理作用[79]；张柳构建了基于贝叶斯的用户评论情感分析模型，并就“里约奥运会中国女排夺冠”话题进行演化研究，探寻不同地区用户对同一话题的不同态度[68]；张雷基于 LDA 模型，构建了主题模型，能够精准识别舆情演化特征，扩大高校师德舆情分析的研究内容[80]；冯兰萍以数据挖掘为手段，以“天津大爆炸事故”为例，分析不同阶段下的网民情感与微博舆情热点变动，提出相应的引导策略[81]。由此可见，微博舆情的情感演化研究多通过计算机技术，实现对具体案例的实证分析，从而提出精准措施。与此同时，金城所提出的空间分析研究方法[82]为空间演化分析提供借鉴意义。这为后文以“河南暴雨”事件为例的微博舆情情感时空演化分析提供研究思路。

3.1　舆情演变时空分析研究框架

本章节以数据挖掘技术、空间自相关理论为基础，构建情感时空演化模型，探寻了解突发公共事件中用户情感演化及规律特征，为政府进行舆情引导提供参考价值，使相关部门能够及时做出有效应对。

首先通过 Python 爬取“人民日报”微博下“河南暴雨”事件的相关微博，并根据评论热度获取热度靠前的网民评论，同时获取并采集网民评论中发表评论的用户信息。针对网民评论信息，结合 jieba 分词、自定义词表和百度停用词表进行相关处理备用。其次，根据全部评论数据中的评论数、点赞数、转发数，通过 Pearson 相关系数分析，确定演化阶段划分指标，并结合生命周期理论实现演化周期的阶段划分。再次，针对各个阶段内的文本评论数据、用户个人数据实现内

容分析、每日情感分析和地理空间分析。最后，从基于 TF-IDF 的关键词统计、基于 LDA 的热点话题统计、基于朴素贝叶斯的情感倾向统计和基于空间自相关的用户个人信息统计四个方面直观展示挖掘突发事件舆情的演化规律和特征。

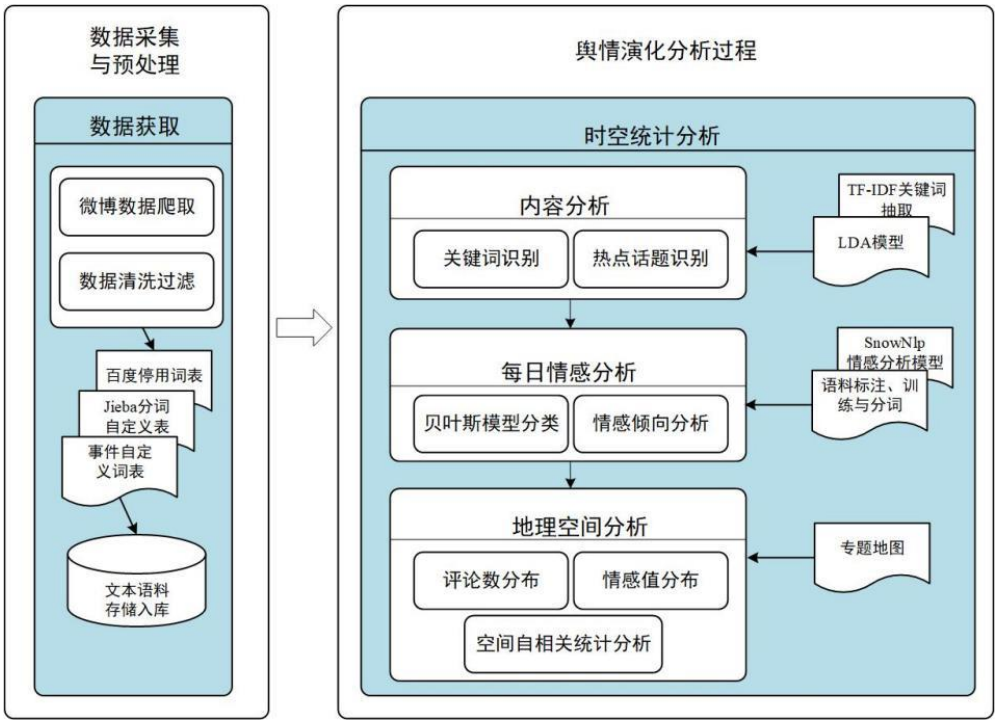

图 3.1　网络舆情演变时空分析研究框架

3.2　数据获取与处理

3.2.1　事件选取

自 2021 年 7 月 19 日开始，河南郑州市、巩义市、登封市等多个城市出现强降雨，多地范围内出现内涝，并伴发小范围山洪灾害。相关事件报道一出，该事件瞬间成为社会极其关注的突发热点事件，网民不再持有观望态度，充分行使自己的表达权力，舆情热度逐步攀升，最终在 7 月 21 日到达顶峰。出于平台代表性和事件典型性两方面的考虑，本文决定选取微博来开展实例研究。

（1）平台代表性。微博（Micro-blog）作为一种通过用户关注机制来实现简短实时信息分享的广播式社交平台，它允许用户通过多种方式完成信息的传播、互动。2009 年“新浪微博”内测版一经推出，受到我国网民的喜爱。而新浪成为国内首家提供微博服务的门户网站，多年来保持着较猛的发展势头。自此国内各大公司发展微博网站，形成“百花齐放”的局面。然而自 2014 年 11 月起，网易微博、腾讯微博和搜狐微博相继关闭，新浪微博一枝独秀。因此，选取新浪微博平台（以下简称微博）作为数据获取渠道具有较强的代表性，具体表现为几点：

①影响力强。微博官方平台于 2021 年发布的《微博 2020 用户发展报告》显示：截止至 2020 年 9 月，月活跃用户已达 5.11 亿，日活跃用户达 2.24 亿，并且同时向海外用户开放，在世界范围内具有较强的影响力。

②功能丰富。第一、微博由最开始仅允许用户将短文本向外传播给公众或者特定接触群体，发展到以“文字+视频”的分享形式发表观点看法，多元化的交流形式提升了用户使用满意度。第二、微博不仅设有网页端，还有移动端，能够充分利用用户日常生活中碎片时间，帮助用户实现便捷互动。第三、微博中的“高级搜索”设定功能，能够根据用户需求实现信息精准化搜查，迅速定位到目标内容，具有省时省力，使用效率高的优点。

③蕴含丰富的数据价值。用户在使用过程中，会相应地产生三类数据：一是结构化数据，即依照用户注册账号时所提交的个人信息所产生的数据；二是用户在使用过程中分享的文字、音频、图片、视频等发布信息、用户互动（点赞、转发和评论）信息所产生的半结构化数据；三是针对用户使用过程中留下的能够被发掘价值的非结构化数据，例如分享观点中所蕴含的情感观点、用户发表文本中的现实痕迹等。以上三种数据为数据的开发挖掘提供了较大的利用空间，具有较为明显的现实价值。

（2）事件典型性。媒体平台特有的媒体属性使微博成为“舆情之源”具有必然性，用户的信息发布、交流传递经过网络传递最终会汇聚形成网络舆情，一旦负面舆情大面积爆发，将会对现实社会造成重大影响。选择该事件进行研究主要有以下原因：第一、“河南暴雨”事件作为一起非人为的突发公共事件，相比较起因是人为操作失误的“2013 年八宝煤矿”事件而言，具有更大的非可控性。第二、在官方将“郑州地铁 5 号线”事件定性为责任事件之前，网友态度发生较大转变。其中，网民对媒体产生不信任感的原因很大部分来源于部分媒体或大 V 发布待求证消息，甚至不少网民由抨击媒体上升到抨击政府。根据以上原因，选择“河南暴雨”事件为案件，探究突发公共事件发生后用户情感传播的演化规律和时空变化，为政府及相关部门舆情引导提供可行性参考意见。

3.2.2 数据获取方法与工具

网络信息不但增长速度快，而且具有复杂性、多样性、多元化的特点，这对数据获取提出了巨大的挑战。搜索引擎所提供的信息并不能十分契合需求，需要用户进行二次筛选操作。为更快速、精准的对目标信息实现获取，传统的信息获取方法亟待转型升级，网络爬虫技术应运而生。网络爬虫(Web Crawler)遵循特定规律，通过程序或脚本实现网页内容的自动爬取。爬取数据时首先选取有用的链接将其添加到待抓取的 URL 队伍当中；其次以一定的策略选取接下来要抓取的

队列，重复操作；最后当满足系统所设要求时停止。

作为信息采集手段的网络爬虫遵循信息识别、信息采集、信息存储的工作流程，因此普通的网络爬虫由页面研究模块、数据库、采集模块、URL 队列及任务抓取几个部分组成[83]。根据覆盖面不同可将网络爬虫分为通用爬虫和聚焦爬虫，后者指的是将目标设为某一主题相关，以主题为中心向外发散的小覆盖面爬虫。尽管网络爬虫能够实现目标信息的准确定位和获取，但是仍需要使用者对计算机编程语言有一定的熟悉度。因此，各式各样的爬虫软件相继出现在市场上，目的在于以更简单的手段实现信息获取，例如八爪鱼采集器、后羿采集器、Goseeker 采集器等。尽管现成软件更加便捷，但获取的数据相对而言却不够灵活，较为死板。因此，为更精准地获取原始数据，本文采用基于聚焦爬虫的数据获取方法，搜集和存储目标事件的初始信息。

3.2.3 数据采集结果

2021 年 7 月 17 日以来，河南省出现极端强降雨天气，气象局多次发布预警表明该次降雨具有持续时间长、降雨量大的特点。这次降雨引起了全国人民甚至海外华人的重大关注，各地网民在微博积极参与讨论，其中“#河南暴雨互助#”这一话题甚至高达 170.4 亿的点击数。考虑到数据来源的代表性，本文选择“人民日报”为数据来源，通过“高级搜索”功能选定时间 2021 年 7 月 19 日至 8 月 6 日，以目标时间区间内“人民日报”下有关河南暴雨的微博为爬取对象，获取该条微博的内容 ID、发布时间、发布内容、转发数与点赞数。同时，爬取该条微博下热度靠前的前 1000 条评论（评论数不足 1000 的则获取全部评论），爬取信息包括：评论用户 ID、评论发表时间、评论内容、评论用户年龄、评论用户性别及评论用户所在地区等信息。将微博内容、评论信息和用户个人信息综合存储在一张表格上，命名为“河南暴雨事件原始数据.csv”文件。最终爬取总数量为 31002 条，部分样本数据如图 3.2。

微博发布时间	微博内容id	微博内容	微博评论数	微博转发数	微博点赞数	评论内容	评论者id	评论者性别	评论者年龄	评论者地址	评论日期
2021/8/2 18:18	KrC300mhL	【#国务院成立郑州特大暴雨灾害调查组#】国务院决定成立调查组，由应急管理部牵头，相关方面参加，对河南郑州“7·20”特大暴雨灾害进行调查。调查组聘请专家为调查工作提供技术支撑。调查组将依法依规、实事求是、科学严谨、全面客观地对灾害应对过程进行调	1368	1070	13536	查一查为啥好好的路天天挖，还要建花坛……	同学Heari	男	22	河南 郑州	2021/8/2 18:20
2021/8/2 18:18	KrC300mhL	【#国务院成立郑州特大暴雨灾害调查组#】国务院决定成立调查组，由应急管理部牵头，相关方面参加，对河南郑州“7·20”特大暴雨灾害进行调查。调查组聘请专家为调查工作提供技术支撑。调查组将依法依规、实事求是、科学严谨、全面客观地对灾害应对过程进行调	1368	1070	13536	好好查，尤其是地铁。	小丫头jwn	女	29	山西	2021/8/2 18:20

图 3.2 爬虫爬取数据示例（部分）

3.2.4 数据清洗

尽管网络爬虫具有较高的灵活性与准确性，但它的本质是机械性、重复性地实现信息的采集与存储。没有“记忆”的网络爬虫并不明白这个数据刚刚已经采

集过，同时它也没有“智商”，不具备识别网民评论是否与事件相关的能力。因此对已获取信息进行数据清洗是不可或缺的一步。

（1）数据清洗。①爬取的数据中存在因网络问题而爬取失败所产生的残缺数据，也有因网络卡顿而获取的重复数据，因此利用 Python3.8 版本中的 pandas 包实现对重复、残缺数据的删除；②针对目标领域的爬取对象，爬虫只拥有爬取能力，无法识别评论的事件相关性。因此，针对评论中诸如各类广告、纯标点和无关网址等无效信息，利用正则表达式进行替换处理。目的在于使得处理过后的文本更加干净，以便有效减少实验时间，提高实验效率。

（2）中文分词。清洗过后的数据具有更高的纯净性，为后续实验提供了数据基础。由于中英文的文本表达方式不一，中文文本则需要分词后再进行情感分析。常见的分词工具包括 jieba 中文分词、NLPIR 等，针对文本大小的不同和使用环境的不同，这些分词工具的分词结果都存在差异性。jieba 分词工具包作为 Python 的第三方内置库，主要具有以下优点：第一，全模式分词模式、搜索引擎模式和精确分词模式能够多方位地满足用户需求。其中，精确模式精确度最高，全模式可迅速扫描并输出待分词文本中的所有词语，搜索引擎模式则在精确模式的分词结果上去除冗余词语，达到更高的效果。第二，jieba 可以区分简体和繁体文字，可以实现对繁体文字语句的识别与分词操作。第三，为完善分词结果，jieba 还支持用户在内置词语库中的基础上添加自定义词典，提高分词效率。第四，jieba 直接支持 TFIDF、TextRank 关键词提取算法，使用方便快捷。第五，可以进行并行分词，效率快。因此，本文中利用 jieba 工具包对清洗过后的数据进行分词。

（3）去停用词处理。对文本进行分词后，发现语句中充斥着大量“的”、“得”和“似的”等没有实际意义的词语，这类词语被称为停用词。本文以哈工大停用词表、百度停用词表、中文停用词表为基础，通过合并、去重操作得到新的停用词表，并加入微博中常出现的网络流行语达到扩充词表的目的，以完善去停用词效果，为后文关键词抽取提供便利。

3.3 演化阶段划分

本小节以用户评论为文本数据进行文本情感挖掘，对传播规律进行时空分析，首先需要对演化阶段进行划分。用户的主要微博使用行为包括：发布微博、用户之间互相进行博文转发、博文评论、博文点赞等。其中，孟吉杰[84]研究认为转发数和点赞数作为信息传播行为的一个量化指标，能够有效反映用户的参与程度。因此，文本通过计算评论、转发和点赞这三种行为的相关性来剖析用户互动行为对网络舆情演化的影响因素。并选择其中一个指标，结合生命周期理论对“河南暴雨”事件的微博舆情进行演化阶段划分。

爬虫爬取了“人民日报”官方微博下指定时间段内的相关博文（共 49 条）及各博文下排名前 1000 的一级评论（不足 1000 条则取全部评论）。通过 SPSS24.0 计算转发数、评论数与点赞数三个指标之间的相关性（见表 3.1），根相关性结果确定分析传播趋势的指标。

Pearson 相关系数的取值范围是[-1,1]，若相关系数大于 0，说明两个变量之间呈正相关关系；相关系数小于 0，说明两个变量之间呈负相关关系；相关系数为 0，表明两个变量之间无关联。表 3.1 显示评论数与转发数、点赞数之间的相关性超过 0.7，并且点赞数与评论数之间系数达 0.84，说明三者之间具有高度的正相关性。故本文确定以评论数作为舆情阶段划分的主要变量，将每天 24 小时划分为 4 个单位长度时间，对已爬取的数据绘制出评论数趋势图（见图 3.3）。

表 3.1　三项指标的相关分析

		微博评论数	微博转发数	微博点赞数
微博评论数	Pearson 相关系数	1	.725	.831
	双侧 Sig.		.000	.000
	N	49	49	49
微博转发数	Pearson 相关系数	.725	1	.278
	双侧 Sig.	.000		.053
	N	49	49	49
微博点赞数	Pearson 相关系数	.831	.278	1
	双侧 Sig.	.000	.053	
	N	49	49	49

依照图中目标时间区间内微博评论数量的变化，可以发现该事件在整个传播中呈现出明显的传播爆发、波动减少、缓缓消亡的趋势。结合突发公共事件网络舆情传播的生命周期理论，本文将舆情阶段划分为爆发期（2021 年 7 月 19 日-2021 年 7 月 21 日，记为 T_1 阶段）、波动衰退期（2021 年 7 月 22 日-2021 年 7 月 24 日，记为 T_2 阶段）和消亡期（2021 年 7 月 25 日-2021 年 8 月 6 日，记为 T_3 阶段）。通过比对知微事见官网关于“河南暴雨”事件的微博平台传播趋势（如图 3.4），可知两个图的传播趋势相近，这表明根据评论量绘制出的图 3.3 是有依据的。

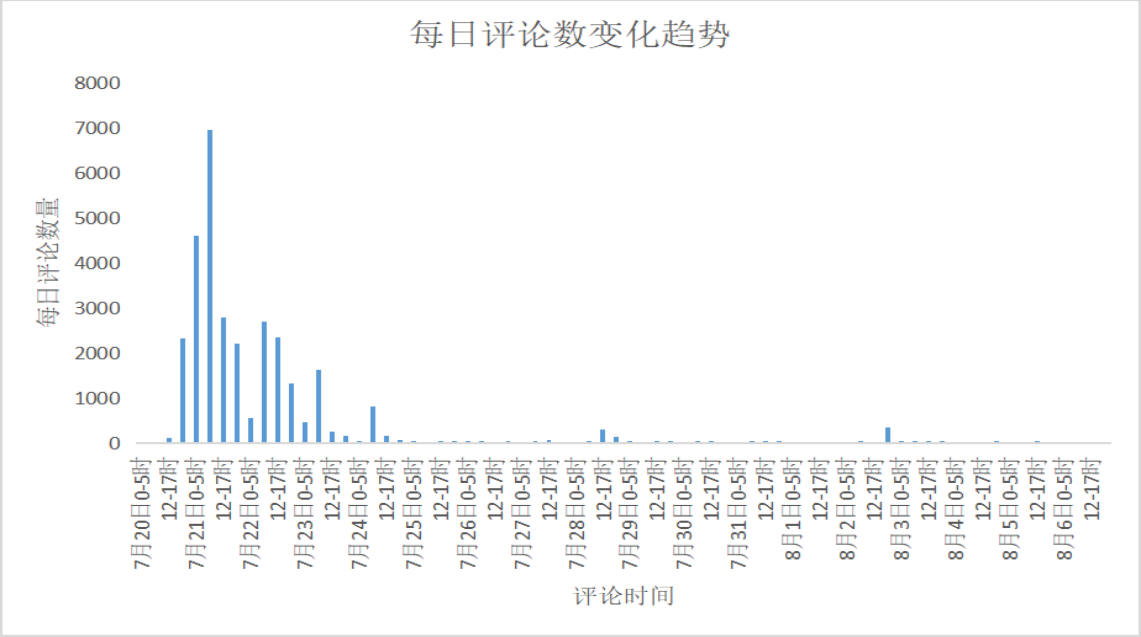

图 3.3　每日评论变化趋势图

（1）爆发期（T_1 阶段）。该阶段内关于该事件的热议铺天盖地，各方媒体都争相参与宣传报道，网民的目光皆集聚在这方面，最后演变成全民参与讨论的现状。舆情的爆发需要能够短时间内调动用户讨论积极性或者刺激用户情绪的导火索。尽管河南大面积的降雨一直牵动着全国人民的心，但直到 7 月 20 日郑州地铁 5 号线及周边区域发生严重积水现象，18 时左右积水冲破屏障进入正线区间，导致 5 号线一列列车被洪水围困，大量乘客被困于地铁内无法脱身。乘客被困后，诸如祈福、救援消息扩散的消息数量迅速上升，网络舆情关注度处于最高水平。同时地铁事故发生当天 12 名乘客不幸遇难，网络上的舆论顿时将主管部门推至风口浪尖。因此本文将 7 月 19 日至 7 月 21 日记为爆发期。

（2）波动衰退期（T_2 阶段）。这一区间指的是事件舆情爆发后，民众接受事实后舆情呈现下降趋势，但仍旧会出现小幅度波动的阶段。在这一阶段内，由于相关部门采取了效果良好的补救措施或者事件有了新的发展，导致网民对于该事件的关注度逐渐减少，事件影响力减弱，但可能出现舆情反复。7 月 22 日至 7 月 24 日间，尽管鹤壁、巩义等地的受灾情况陆续公布，但全国各地派遣救援队驰援河南，丰富的救援物资被送达受灾地区，这极大程度上缓解了网民内心的焦虑，转移了网民的注意力，网民议论稍有减弱。

（3）消亡期（T_3 阶段）。衰退期指的是网络舆情的消亡期，该阶段由于事件的落幕，导致网民中只有极少部分的人依旧保持着关注，此时事件的舆情关注度和信息量都保持着极低水平，并基本没有太大起伏。

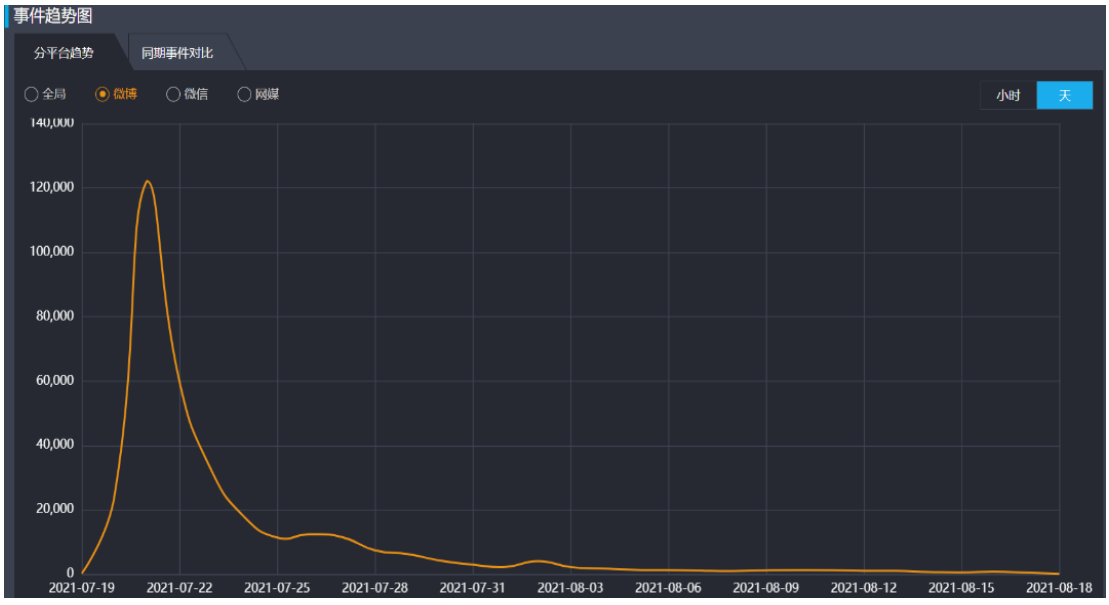

图 3.4　知微事见官网关于“河南暴雨”事件的微博平台传播趋势图

3.4　基于 TF-IDF 的关键词识别

词频-逆文件频率（TF-IDF）主要用来测评特定词语在当前文本集或者大规模语料库文件中的重要程度[85]。假设一个指定词语在单条文本中出现次数越高，但在整个数据库中出现次数越少，则表明该词语具有较高的重要性。其中词频（term frequency, TF）指一个给定词语在文本中出现的次数，逆向文件频率（inverse document frequency, IDF）指特定词语在整个语料库文件中的低频率。通常使用 TF-IDF 方法来过滤常见词语，而保留下重要的关键词。

在本小节中，定义 $D=\{D_1,D_2,...,D_i\}$ 为预处理过后的待分析微博文本构成的文本总合集，共有 i 条文本。$D_i=\{S_1,S_2,...,S_j\}$ 代表每条微博文本数据由 j 条语句组成，$d_{i,j}$ 表示 D_i 中的一个特定词语，则 $d_{i,j}$ 的词频（TF）计算公式为：

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \qquad \text{式（3）}$$

$n_{i,j}$ 表示词语 $d_{i,j}$ 在微博博文文本 D_i 中出现的次数，$\sum_k n_{k,j}$ 表示 D_i 中的全部词语数量。词条 $d_{i,j}$ 的逆文档频率（IDF）的计算公式为：

$$IDF=\log\frac{|M|}{|N\in D:d_{i,j}\in N|+1} \qquad \text{式（4）}$$

其中 $|M|$ 表示微博博文文本合集 D 中的文本总数，N 表示包含词 $d_{i,j}$ 的文档数。鉴于此时 N 在分式中充当着分母的角色，需要在分母加 1 操作以避免分式无意义。某个词语在当前文档的高词频率，与它在整个数据预料库中的低词语频率相乘得到的乘积结果，代表了该词语在整个语料库中的重要程度。本文定义微博博文文本中特征词 $t_{i,j}$ 表示特征词与在微博中的重要性，则 $t_{i,j}$ 计算如下：

$$t_{i,j} = TF * IDF \quad \text{式（5）}$$

用户对某条微博进行多次转发或评论操作时，该微博中的关键词词频也越高，即重要性越强。因此通过文本内容中的关键词权重，整理演化中各阶段的前 10 个关键词（见表 3.2）。

从提取的关键词中（表 3.2）可以看到，爆发期中的热点词有“加油”“平安”“挺住”等态度词，“河南”“郑州”等地点词和“转发”“扩散”等动词，表明该阶段内网友表达的是暴雨状态下自身的积极祈祷态度和力所能及的消息扩散行为；波动衰退期内的关键词相比较第一阶段而言，除了上一阶段的一系列正能量鼓舞词语，明显增加了“卫辉”“鹤壁”“辉县”三个县市地名，这表明此时网民的关注点转向了暴雨所带来的影响及当前阶段河南省内伤亡较严重的其他地区；消亡期内，“查查”“地铁”“追责”“花坛”等权重较大的词语都不约而同地指向了“郑州地铁 7.20”事件。此时降雨量下降，河南省内各县市区开始灾后定损与灾后重建工作，网民的关注点主要转向了导致事故灾难发生的原因公布，希望政府及相关部门能够给予公众一个合情合理的解答。

表 3.2　TF-IDF 关键词抽取结果

时间段	关键词	演化阶段
T_1	加油、平安、河南、平平安安、挺住、暴雨、转发、郑州、希望、扩散	爆发期
T_2	加油、新乡、河南、暴雨、平安、转发、卫辉、鹤壁、辉县、互助	波动衰退期
T_3	花坛、郑州、河南、感谢、查查、地铁、心善、暴雨、河南人、追责	消亡期

3.5　基于 LDA 的热点话题识别

挖掘出“河南暴雨”事件网络舆情传播周期中各阶段的主题，描述舆情事件主题演化规律，勾勒网民情感倾向变化的时序变化趋势，能够为舆情应对的决策与分析提供科学依据。前文详细描述了舆情演化阶段的划分根据，证明划分结果有效性，并将事件的整个发展过程划分为三个阶段。TF-IDF 关键词抽取出舆情传播各阶段的热词，反映出当前所处阶段的高频热词。但不难看出，尽管各阶段内都有权重较高的词语出现，但依旧概括不出主题。

作为话题挖掘的常见模型之一，LDA 主题模型的最明显特征在于能够将若干文档自动编码分类为一定数量的主题，极大程度上减少人为干预和负担，具有较好的效果。LDA 模型的运行需要先人为设定主题数量，然后得到每个主题词

下词语的分布概率以及文档对应的主题概率。该模型在训练过程中采用 Gibbs 采样，具体工作过程如图 3.5 所示：

（1）α 随机生成文档对应主题的多项式分布 θ

（2）θ 随机生成主题 z

（3）β 随机生成主题对应词语的多项式分布 φ

（4）综合主题 z 和主题对应词语分布情况 φ 生成词语 w

（5）循环(4)这一步骤，生成一个包含 m 个词语的文档

（6）最终生成 k 个主题下的 n 篇文档

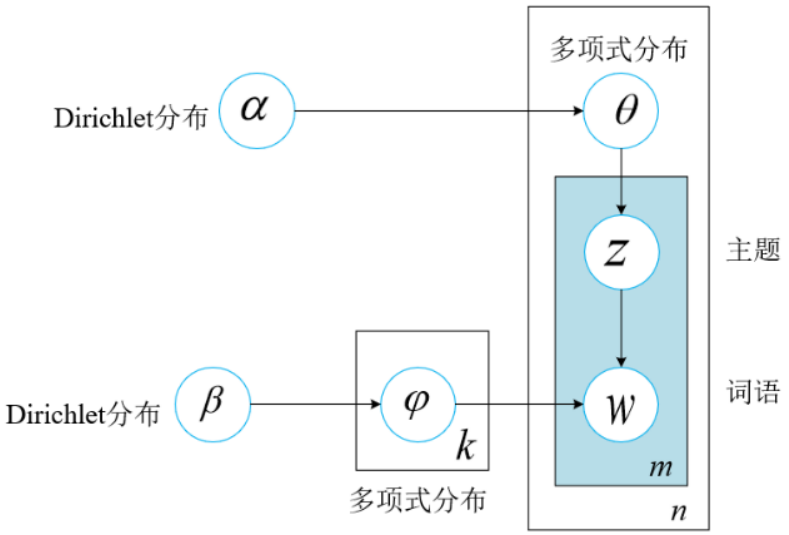

图 3.5　LDA 模型工作流程

因此，本小节将结合 LDA 主题模型和时序分析，进行时序主题挖掘，为政府及相关部门针对舆情事件施行目标监控、制定舆情预警决策提供理论支撑。

采用 LDA 主题模型进行主题聚类的过程中，首先需要依据用户发表评论的对应时间点，确定好各阶段内的目标文本集。其次引入 LDA 主题模型，通过计算困惑度来确定主题集合，当困惑度最低（见图 3.6）时对应的 topic 就是最优主题数 k，确定最优主题数后，选择排序前 10 的主题词作为该主题的特征词，并根据特征词进行主题概括（见表 3.3）。