



开放科学
(资源服务)
标识码
(OSID)

自然灾害事件主题发现与演化研究 ——以知乎自然灾害讨论为例

杨海林¹ 田军¹ 李莉芳²

1. 西安交通大学管理学院 西安 710049;
2. 伦敦国王学院生物统计与健康信息学系 伦敦 SE 8AF

摘要: [目的/意义] 我国是自然灾害最为频发的国家之一, 网络上的灾害讨论信息容易引发负面舆情影响, 危害社会稳定, 研究自然灾害事件的舆情主题演化规律能协助政府制定舆论管理策略。[方法/过程] 本文以两次暴雨洪涝灾害“2021年7·20河南暴雨”及“2020年中国南方洪涝灾害”事件为背景, 采用 Correlated Topic Model (CTM) 主题模型分析方法对知乎用户所发布信息进行建模分析。具体而言, 本文结合主题强度和主题传播广度两方面分析用户在暴雨灾害中的注意力转变, 并深入探究了用户参与情况对主题传播的影响, 进而推演灾害期间主题传播内容与演化规律。[结果/结论] 研究得出了两次自然灾害事件的主题传播内容与演化过程情况, 为未来自然灾害事件中政府的舆情应对和处理提供了参考。

关键词: 自然灾害事件; 主题演化; 知乎; 主题模型; 舆情传播

中图分类号: G350

A Study on the Topic Detection and Evolution Analysis of Natural Disaster Event: Taking the Natural Disaster Discussion in Zhihu as an Example

YANG Hailin¹ TIAN Jun¹ LI Lifang²

1. School of Management, Xi'an Jiaotong University, Xi'an 710049, China;
2. Department of Biostatistics and Health Informatics, King's College London, London SE 8AF, China

Abstract: [Objective/Significance] Natural disasters such as rainstorms and flooding happen frequently in China. During such disasters, public opinions disseminating quickly on social media which may trigger negative effects on social stable. Extracting the discussion themes of the public during natural disaster can help the government agencies formulate efficient disaster

基金项目 国家自然科学基金资助项目“政府应急管理能力成熟度评估研究”(71171157)。

作者简介 杨海林(1998-), 硕士生。研究方向为应急管理与应急决策、应急舆情管理; 田军(1964-), 博士, 教授, 研究方向为应急管理、供应链管理与管理信息系统; 李莉芳(1989-), 博士, 副研究员, 研究方向为危机健康信息管理, 应急舆情管理。

引用格式 杨海林, 田军, 李莉芳. 自然灾害事件主题发现与演化研究——以知乎自然灾害讨论为例[J]. 情报工程, 2023, 9(4): 40-53.

management strategies. [Methods/Processes] The study applied one rainstorm ("July 20, 2021 rainstorm in Henan Province") and one flooding ("2020 flood in Southern China") Zhihu datasets to investigate the main topics. Specifically, Correlated Topic Model (CTM) was applied to examine the topics in the posts of Zhihu users. Specifically, this study combines the two aspects of topic intensity and topic communication breadth to analyze users' attention changes during rainstorms, and explores the impact of user participation on topic communication, and then deduce the content and evolution patterns of topic communication during disasters. [Results/Conclusions] We uncovered the topic content and evolution patterns of topic communication during disasters and provided practical insights to the government agencies.

Keywords: Natural disaster events; Topic evolution; Zhihu; Topic model; Public opinion dissemination

引言

在自然灾害事件爆发后，公众在社交媒体平台上参与各类热点话题的讨论。知乎、微博和推特等社交媒体平台成了公众对突发灾害事件发表包含情绪、态度、意见与想法等的关键渠道。2021年7月，河南省郑州市接连遭遇特大暴雨，此次特大自然灾害造成全省150个县（市、区）、1663个乡镇、1453.16万人受灾，直接经济损失高达1142.69亿元^[1]。在此次自然灾害的形成与发展过程中，各类话题在社交媒体上引发了广泛的关注与讨论，这些话题体现了公众的关注与诉求。但是网络空间上的信息交流密集且复杂，如何从海量信息中提取出能体现公众核心诉求的信息至关重要。因此，政府有必要通过了解公众在突发事件中关注的主要话题，及时传递政府信息，妥善平息舆论风波，安抚公众的紧张或恐慌情绪，从而保障网络空间和现实社会的和谐与稳定^[2,3]。

知乎作为中文社区最大的社会化问答平台，其月活跃用户规模已超过一亿^[4]，其定位用户群体是愿意分享专业的知识、经验、见解的社会人群。知乎平台采用问答化的长文本交流方式，内容创作者希望自己的观点

或看法能被他人所认可，这使得对问题发表的评价注重原创性、多元性和科学性，能够发现有别于微博、推特等短文本社交媒体平台的观点态度^[5]。知乎平台上所发表的信息质量更高，提取这些信息不仅能够反映自然灾害事件中公众的核心诉求^[6]，也能识别部分专业人员对突发事件的解决方法^[7]。因此，本文将聚焦知乎用户的关注主题分析，通过了解知乎话题来解析自然灾害事件期间公众关注焦点的演化规律。

本文基于知乎平台上用户发表内容，挖掘关于两次洪灾的自然灾害事件的舆情话题演化规律。本文想要了解并解决以下几个问题：

- （1）在自然灾害事件舆情传播过程中，公众对哪些主题产生了关注，所关注的主题内容产生了怎么样的变化？
- （2）高质量用户的参与情况对主题的传播和分享有什么影响？
- （3）不同时期的同类型自然灾害事件的主题演化过程中，有哪些相似与差别？

1 研究述评

研究社交媒体平台信息主题演化的方法有

很多, 概率话题模型近年来被广泛使用, 相比于传统的统计学文本分析方法, 其对数据集的适应性强, 所涵盖内容广, 主要有 LDA, CTM, PDM 等模型方法。LDA 主题模型是目前应用最为广泛的概率话题模型^[8,9]。例如王璟琦等^[10]通过空间自相关理论和 LDA 模型, 提取“旅游”话题演化的时空规律。姜金贵等^[11]利用 LDA 主题模型探究“红黄蓝虐童”事件中主题的传播和民众情绪变化对舆情走向的影响。王子涵等^[12]利用 DMM 主题模型对微博平台上“韩国部署萨德系统”事件话题的演化进行分析。李月^[13]采用后离散时间法结合 LDA 主题模型, 对突发公共卫生事件中公共政策主题的演化规律进行了研究。刘国威等^[14]采用 LDA 主题模型和词项共现网络探究了微博平台上“南海仲裁案”事件的周期演化规律。

在突发应急事件的主题演化研究中, 许多学者利用社交媒体平台的信息优势, 也对各类突发应急事件的舆情主题演化进行了分析。例如王艳东等^[15]对微博平台上的暴雨应急事件信息挖掘了舆情主题的时间与空间分布。吴小兰等^[16]以 H7N9 事件的用户关系网络为对象探究了微博社区用户特征和主题演化情况。安璐等^[17]围绕埃博拉疫情探究微博各阶段热点主题和用户行为的相关关系。张兆阳等^[18]探究了新冠疫情爆发时期微博平台上的讨论话题和用户行为模式。而在自然灾害应急事件中, 李纲等^[19]在“哈维”飓风的 Twitter 短文本数据集下对自然灾害的主题演化进行研究, 筛选出受灾地区和非受灾地区不同分类的话题信息, 并探究了其在不同时间阶段的主题差异。李绍攀等^[20]对微博平台上台风“山竹”与“利奇马”的话题热

度与舆情情感变化展开分析, 此外还通过数据结果与地区情况对城市灾害损失进行了估计。

从上述研究来看, 现有文献多数对如微博、Twitter 这类短文本信息多的社交媒体平台展开舆情主题演化研究, 但自然灾害这类突发应急事件, 还缺乏如知乎这类长文本信息平台的舆情主题演化分析。不同平台由于其用户角色、行为特征等方面有所差异, 对知乎平台上的舆情主题进行分析可以补充相关研究^[21]。

针对突发应急事件主题研究, 主要集中于事件中各个阶段的热点主题内容, 时间和空间层面上传播过程分析, 以及主题和用户行为模式与情感特征等关系。这类研究鲜有依据挖掘文本的类型筛选适合的主题模型方法进行分析。且通常分析单一案例下主题演化情况, 未能充分反映同一类型灾害事件下主题演化的机理。此外, 在用户行为模式上学者们主要关注点赞数、评论数等因素, 而平台的高质量用户参与情况也会对主题传播产生影响。本文以知乎平台讨论为例, 研究自然灾害事件主题演化规律, 采用针对性方法建模, 并结合两个案例的数据类比分析得出结论。

2 研究原理与设计

2.1 研究过程设计

基于本文拟解决的问题, 设计图 1 所示的研究思路与流程, 具体步骤如下:

(1) 以知乎平台的关键词话题下的问答、文章推文为数据源抓取文本数据。

(2) 进行去除文本数据中的噪声、分词并去除停用词等预处理工作。

(3) 通过 CTM 主题模型确定主题数量，获取各主题的概率分布。

(4) 根据模型结果，在时间序列上进行主题演化分析，从而探究自然灾害事件中公众的

话题关注点和话题关注度变化规律。在主题内容演化分析过程中，本文探究的内容为：①主题强度演化；②主题传播广度演化；③高质量用户（意见领袖）参与情况对主题演化的影响。

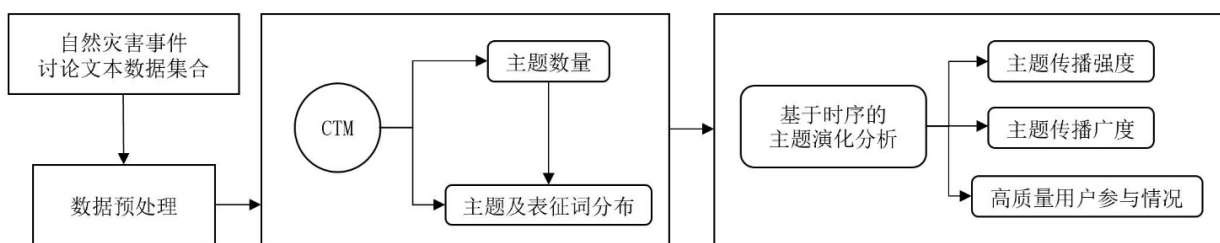


图1 实验流程图

2.2 CTM模型原理

本文的主题生成方法采用 Blei 与 Lafferty 等^[22]于2006年提出的CTM（Correlated Topic Model）主题模型。模型的核心思想是用逻辑斯蒂正态分布来代替LDA主题模型中的狄利克雷分布^[23]，引入协方差矩阵来描述语料库中主题之间的潜在相关性。其 K 维多元分布的指数型分布概率由以下的等式表示：

$$P(z|\eta) = \exp\{\eta^T z - a(\eta)\} \quad (1)$$

在公式(1)中， z 可以取 k 的值，并由一个 k 维向量表示。均值参数化或单纯形的映射函数和这种对数化过程表示为：

$$\eta_i = \log \frac{\theta_i}{\theta_k} \quad (2)$$

在逻辑斯蒂-正态分布（Logistic Normal distribution）下，假设 η_i 服从正态分布（Normal distribution）并使用公式(2)的倒数，然后将其定义到均值参数化或单纯形上。那么在CTM主题模型中生成文档的过程如图2所示^[21]。

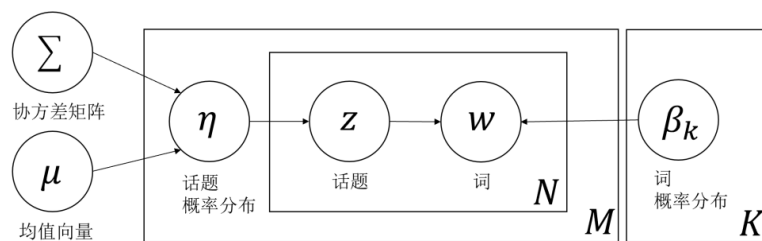


图2 CTM模型文档生成过程

本文所研究内容基于自然灾害事件，有别于长周期的突发公共卫生事件，发展变化周期通常在1-2个月，主题内容在时间序列上较为密集，且主题数量一般在60个以下。

根据Li等基于NIPS数据子集的实验，在主题数量较小情况下，CTM主题模型与其他静态主题模型相比模型对数似然函数值最大，如图3所示，函数值越大时模型困惑度越小，

表明模型的聚合效果更佳^[24]。知乎内容具有文本长和相互关联性强的特点，而 CTM 主题模型在一个大类主题的细分类型情境下能

够实现优于其他方法的分类效果^[25]。本文基于话题模型库 TOMOTOPY 实现 CTM 主题模型生成。

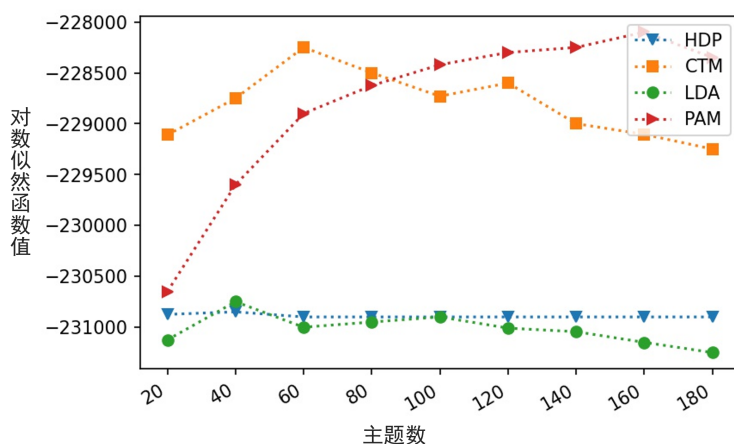


图3 模型不同数量主题下的对数似然函数值比较

2.3 数据获取与预处理

在 Python 编程环境下，爬取两次暴雨洪涝灾害事件的数据集作为样本。2021 年 7 月 12 日至 2021 年 9 月 5 日期间“2021 年 7·20 河南暴雨”及其相关关键词下问题贴回答、发布文章、话题精华中的全部讨论信息为事件 A。2020 年 6 月 25 日至 2020 年 8 月 19 日期间“2020 年中国南方洪涝灾害”及其相关关键词下问题贴回答、发布文章、话题精华中的全部讨论信息为事件 B。这些信息包含语料内容、发布时间、点赞数、评论数等，将包含以上信息的每条语料称为数据文档。事件 A 共计文档数 32542 条，事件 B 共计文档数 4154 条。数据文档按时间分布情况见图 4，按照舆情生命三周期理论划分为发展期，高潮期，消散期，符合客观实际^[13]。

本研究使用 pandas、jieba 等工具包实现分词和文本清洗等预处理，并将语料库按照一定比例重复，以提高主题模型的建模效果。最终

事件 A 获得 26007 条数据文档，事件 B 获得 4012 条数据文档。

2.4 主题数量确定

有监督主题模型假设了主题数量是一个给定的值，而数据集的主题数量事先未知，需要利用判别方法来确定最佳主题数量以保证结果的合理性。本研究通过困惑度 (Perplexity)^[27,28]和主题一致性 (Topic Coherence)^[29]确定主题数量。

困惑度是现在判断主题模型最佳主题数量最常用的指标，其计算公式为^[30]：

$$Perplexity(D_t) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right\} \quad (3)$$

其中， D_t 表示数据文档集， M 表示数据文档数量， d 表示文档编号 ($d \in D_t$)， N_d 表示单个数据文档的词数， w_d 表示单个数据文档里的词， $\log p(w_d)$ 表示数据文档中词 w_d 对数概率函数值。

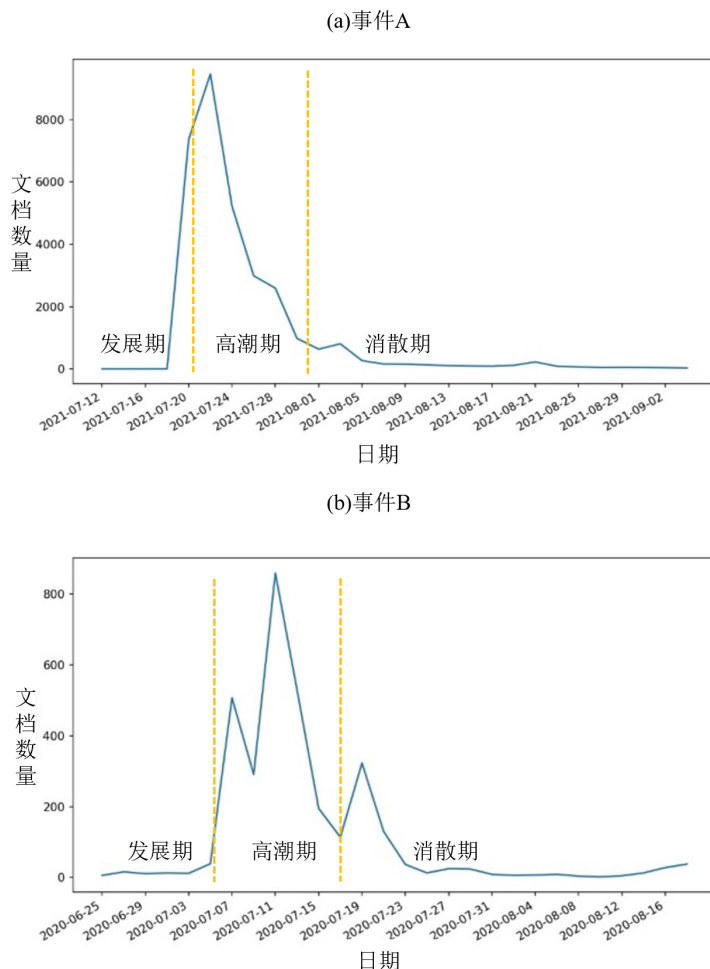


图4 舆情数据时间分布

基于困惑度在许多情况下能够获取较好结果,但由于所选取的种子数不同,主题分布会发生变化,模型可能无法得到解释性足够好的主题内容。主题一致性评价通过单词的共现频率统计数据来获取得分,确定主题数量。主题一致性得分越高,表示各主题的意义可读性越好。其常用的四种算法为 c_uci , u_mass , c_npmi 和 c_v , 其中 c_uci , c_v , c_npmi 方法需要单独的文本数据集进行计算,而 u_mass 是基于原本的语料库信息生成主题一致性得分^[31]。本文采用 `tomotopy` 库 `coherence` 类中的 `U_mass` 预设函数计算最佳主题数的选择范围。设 $D(v)$ 为词类型 v 的文档频率(即包含至少一个类型 v 标记的

文档数量),并且 $D(v, v')$ 为词类型 v 和 v' 的文档共现频率(即包含一个或多个类型为 v 标记和至少一个类型为 v' 标记的文档数量),则 u_mass 的主题一致性得分计算公式为^[31]:

$$Coherence(t; V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \quad (4)$$

其中 $V^{(t)} = (v_1^{(t)}, \dots, v_M^{(t)})$ 是主题 t 中 M 个最可能单词的列表。对数函数分子用单位 1 避免其函数值取 0。

计算事件 A 和事件 B 不同主题数下主题模型困惑度和主题一致性评分,结果如图 5 所示。综合考虑两指标下主题数量最佳区间,确定事件 A 主题数 14 个,事件 B 主题数 6 个。

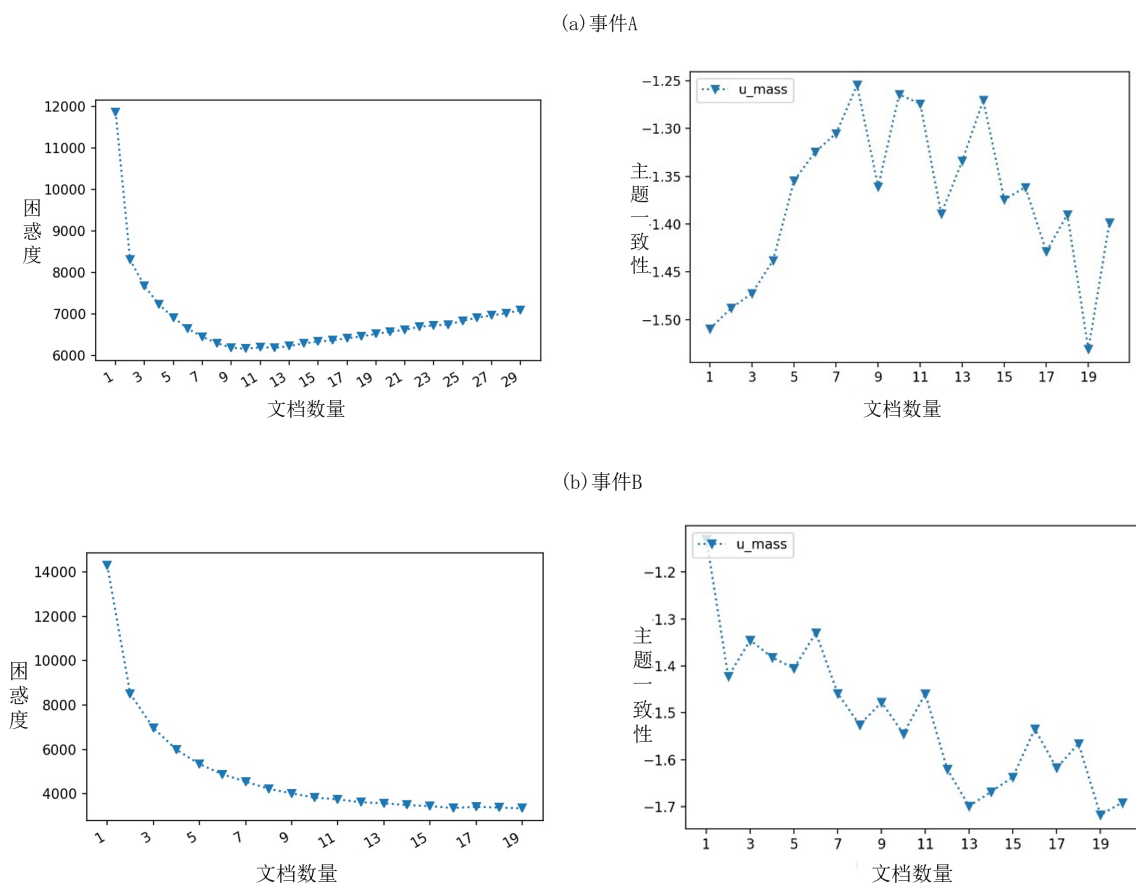


图5 困惑度和主题一致性得分曲线

3 实验分析

3.1 主题发现

按照上述主题数量,提取事件讨论的主题内容及各数据文档的主题分布概率,并根据表

征词命名各主题并数字编号。基于态势感知理论^[32,33],可以按照社会环境、建设环境、物理环境、非态势感知四种类别归纳主题。主题及其前十个表征词举例见表1,各个主题分类见表2。

表1 部分主题和前十个表征词分布概率

主题	表征词
A0 鸿星尔克捐款	鸿星、企业、品牌、买、网友、穿、营销、捐、支持、消费
A1 地铁停运和遇难	地铁、停运、情况、城市、地方、严重、问题、大雨、积水、希望
A2 暴雨救援工作	救援、救援队、信息、情况、洪水、希望、出现、安全、物资、紧急
A3 祈福与加油	加油、情况、洪水、城市、应急、希望、发生、人民、物资、安全
.....	
B0 洪灾原因	洪水、长江、江西、原因、造成、城市、已经、高、问题、影响
B1 高考因洪灾延期	高考、洪水、暴雨、长江、地方、考试、城市、水、考生、中国
B2 抗洪救灾工作	抗洪、洪水、江西、已经、暴雨、工作、泄洪、一定、城市、防汛
.....	

表 2 事件主题分类

信息类型				
社会环境		建设环境	物理环境	非态势感知
事件A	A0 鸿星尔克捐款	A7 人员、车辆等被困被淹情况	A1 暴雨发展情况	A3 祈福与加油
	A2 暴雨中警告与注意事项信息		A9 暴雨和疫情情况	A5 灾情中的性别话题
	A4 暴雨救援工作	A10 城市内涝与排水系统讨论		A8 暴雨中的不满情绪与歧视现象
	A6 明星志愿救援行为			A12 灾害调查情况
	A11 地铁停运与遇难			
事件B	A13 企业捐款与救助信息			
	B1 高考因洪灾延期	B3 城市排水与水利工程讨论	B4 洪灾发展情况	B0 洪灾原因
	B2 抗洪救灾工作		B5 洪灾与疫情情况	

3.2 主题强度演化

主题发布强度反映了网友对灾害中事件信息的关注程度。假设整个事件的持续时间的日期集合为 T ，每个数据文档的发布日期为 t_d ，则有 $t_d \in T$ 。通过 CTM 主题模型，可以得出每个数据文档中各主题分布概率 p_d^i ，那么根据事件

发生的第 t 日话题 i 的总概率分布，单日单个主题

题的强度可以表示为：

$$H_t^i = \sum_{d=x}^M p_d^i(t_x = t) \tag{5}$$

根据各主题在第 t 日的强度 H_t^i ，归纳出每日主题强度排名前 50% 的主题，可以得到主题发布强度随时间分布（各阶段）的演化情况见表 3。

表 3 事件主题强度演化

事件 A

强度排序	阶段				
	发展期	高潮期		消散期	
	07/12-07/21	07/22-07/24	07/25-07/28	07/29-07/31	08/01-09/05
1	A1*	A0	A0	A9	A12
	51.272**	311.730	101.384	36.151	20.048
2	A9	A13	A13	A8	A9
	48.605	267.980	100.731	19.724	19.944
3	A10	A11	A6	A11	A11
	48.564	262.823	98.091	35.754	19.929
4	A3	A5	A5	A10	A8
	48.259	260.206	96.725	35.187	19.724
5	A11	A6	A10	A5	A10
	47.897	252.556	96.532	35.140	19.533
6	A8	A10	A9	A7	A13
	47.465	252.315	95.533	34.432	18.914
7	A2	A12	A8	A1	A1
	46.853	249.981	94.990	34.012	18.725

事件 B

强度排序	阶段				
	发展期	高潮期		消散期	
	06/25-07/06	07/07-07/09	07/10-07/13	07/14-07/16	07/17-08/19
1	B4*	B1	B4	B4	B3
	1.114**	47.016	28.731	11.671	8.654
2	B3	B3	B0	B0	B5
	1.093	43.147	28.433	11.308	8.340
3	B2	B0	B5	B3	B2
	1.043	42.196	27.525	10.992	7.980

(*表示主题标号, **表示主题强度具体数值)

事件 A: 在发展期, 暴雨发展情况 (A1) 和暴雨和疫情情况 (A9) 两个主题强度最高。在整个高潮期, 主题强度最高的三个主题依次为鸿星尔克捐款 (A0), 地铁停运与遇难 (A11) 和企业捐款与救助信息 (A13), 同时, 灾情中的性别话题 (A5), 明星志愿救援行为 (A6) 两个主题也有较高主题强度; 在消散期, 灾害调查情况 (A12) 主题强度最高。期间平均主题强度最高的两个主题依次为鸿星尔克捐款 (A0) 和企业捐款与救助信息 (A13), 平均值分别为 100.481 和 93.366;

事件 B: 在发展期, 洪灾发展情况 (B4) 主题强度最高; 在高潮期, 高考因洪灾延期 (B1) 主题强度最高, 其次是洪灾发展情况 (B4); 在消散期, 城市排水与水利工程讨论 (B3) 主题强度最高。期间平均主题强度最高的主题是高考因洪灾延期 (B1), 平均值为 18.497。

3.3 主题传播广度演化

主题传播的广度反映了网友在灾害中对事件信息传播与分享的意愿。在知乎平台上, 回答和文章的点赞数量越高, 则越有可能会被平

台推送至更多数量的其他用户推荐主页上, 由于知乎的回答和文章不显示阅读量, 其点赞数量的高低表示了信息传播范围的广狭。点赞行为通常代表用户对某一特定主题观点的肯定态度, 为放大特定主题的传播效果, 取概率分布最大的主题为所传播主题, 通过每条数据文档点赞数 v_d 和各主题分布概率 p_d^i , 主题传播广度可以表示为:

$$S_t^i = \sum_{d=x}^M v_d \max_{i \in N} (p_d^i) (t_x = t) \quad (6)$$

根据各主题第 t 日的传播广度 S_t^i , 归纳出每日主题传播广度排名前 50% 的主题, 可以得到主题传播广度随时间分布 (各阶段) 的演化表, 见表 4。

事件 A: 在发展期, 地铁停运与遇难 (A11) 传播广度最高, 其次是暴雨发展情况 (A1) 和暴雨救援工作 (A4)。在整个高潮期, 鸿星尔克捐款 (A0) 传播广度最高, 同时, 灾情中的性别话题 (A5), 明星志愿救援行为 (A6) 也有较高传播广度; 在消散期, 灾害调查情况 (A12) 传播广度最高。期间平均主题传播广度最高的两个主题依次为鸿星尔克捐款 (A0) 与灾情中的性别话题 (A5), 平均值分别为 224.0 与 116.5;

事件 B：在发展期，洪灾原因 (B0) 传播广度最高；在高潮期，高考因洪灾延期 (B1) 传播广度最高，其次是抗洪救灾工作 (B2)；在消散期，城市排水与水利工程讨论 (B3) 传播广度最高。期间平均主题传播广度最高的主题是高考因洪灾延期 (B1)，平均值为 17.6。

表 4 主题传播广度演化

(a) 事件 A

广度排序	阶段				
	发展期		高潮期		消散期
	07/12-07/21	07/22-07/24	07/25-07/28	07/29-07/31	08/01-09/05
1	A11*	A0	A0	A11	A12
	73.4**	811.3	213.5	48.0	26.5
2	A1	A5	A6	A5	A11
	66.4	352.7	135.0	46.0	25.0
3	A4	A6	A5	A1	A13
	58.2	259.3	132.5	42.7	21.0
4	A2	A1	A11	A7	A3
	57.6	233.3	92.5	40.7	20.5
5	A3	A3	A1	A8	A0
	57.0	217.3	88.0	40.0	20.0
6	A8	A13	A4	A0	A5
	46.0	202.7	87.5	37.3	19.5
7	A9	A2	A13	A9	A7
	45.2	198.7	80.5	36.7	17.0

(b) 事件 B

广度排序	阶段				
	发展期		高潮期		消散期
	06/25-07/06	07/07-07/09	07/10-07/13	07/14-07/16	07/17-08/19
1	B0*	B1	B0	B2	B3
	2.0**	44.0	33.0	14.3	12.8
2	B2	B0	B2	B3	B5
	1.4	21.3	29.3	13.3	8.5
3	B3	B3	B1	B4	B0
	1.2	20.3	28.8	12.0	7.3

(*表示主题标号，**表示主题传播广度具体数值)

3.4 高质量用户对主题演化的影响

由于每个数据文档不仅包含点赞数、评论数，还包含了用户的被关注数和获取赞同数等

信息，可根据这些信息观察用户群体如何影响自然灾害事件主题的演化过程。本文选取所有被关注数大于 10000，被点赞数大于 10000 的用户，记为高质量用户。这些高质量用户一般为

知乎社区的意见领袖,是指在人际关系网如社区群体中,能经常为他人提供有效的信息、建议或个人观点,同时对他人的决定和行为施加影响的活跃人群^[34]。将每个主题中符合筛选条件的回答用户人数记为高质量用户参与数。通

过计算两个自然灾害事件案例中主题的点赞数、评论数以及高质量用户参与数之间的 Spearman 相关系数,分析整个事件中三种数据之间是否存在两两相关关系。使用 SPSS 25.0 软件,得到的计算结果如表 5 所示:

表 5 点赞数、评论数与高质量用户参与数的相关性分析

Spearman Rho N = 20		点赞数	评论数	意见领袖参与数
点赞数	相关系数	1.000	.811**	.880**
	Sig.	.	.000	.000
评论数	相关系数	.811**	1.000	.647*
	Sig.	.000	.	.012
意见领袖参与数	相关系数	.880**	.647*	1.000
	Sig.	.000	.012	.

(**.在0.01级别(双尾),相关性显著。*.在0.05级别(双尾),相关性显著。)

从表 5 可见,点赞数、评论数和高质量用户参与数之间存在相关关系,点赞数和评论数、高质量用户参与数之间的 Spearman 相关系数均在 0.8 以上,置信度均小于 0.01,两者的相关性很强;而评论数和高质量用户参与数之间的 Spearman 相关系数在 0.6 以上,置信度小于 0.05,评论数与高质量用户参与数之间的相关性稍弱。

4 结果与讨论

4.1 自然灾害事件中主题关注度和演化规律

总体而言,自然灾害事件主题的演化规律遵循灾害发展过程的演化逻辑,且两次事件在不同阶段主题具有一定相似性。在灾害发生期间,事件发展期用户对灾情的发展态势和救援情况较为关注,并且愿意去发表灾情相关的祈祷与祝福内容;而在事件爆发期,用户的关注

点转移到突发社会热点事件,有关管理部门对灾情各种问题的应对情况等,在此期间用户舆情信息发布数呈指数级增长;事件消散期中,用户关注应急处置工作的反思与总结,灾害调查情况等内容。此外,关注度更高的“2021 年 720 河南暴雨”,其话题数量在时间序列上更为集中。

4.2 主题类型与主题关注度情况分析

从两个案例类比来看,用户对于灾害事件中的社会环境类主题有着极高关注度,如主题“鸿星尔克捐款”与“高考因洪灾延期”。这两个主题的关键传播关键节点分别为爆发期的中期与初期阶段,都在灾害事件中引起了用户的广泛关注,其较高的主题强度和传播广度一定程度上推动了平台用户对整个灾害事件有较高讨论热度。政府管理部门应当在灾害爆发期时密切关注社会环境类事件,及时控制信息传播,防止舆情进一步扩散发酵。两个案例中,

用户讨论主题存在如“城市内涝与排水系统”“城市排水与水利工程”等相似内容,证明用户对于科普知识等内容存在关注与了解,而求助信息等内容难以受到关注。在社会化问答平台上,公众对专业的科普信息有着一定需求,平台应当通过一些激励措施鼓励高质量用户发布专业知识内容,帮助民众提升防范灾害能力。同时,应急机构也可以和平台展开合作,集中管理灾害期间求助信息的转发与扩散,推动建立灾情期间信息协调与共享机制,提高受灾地区用户信息的可见度,能够更有效地开展应急响应与处置工作。

4.3 高质量用户参与情况对主题传播的影响

在事件网络讨论中,社会化问答平台上的高质量用户的参与情况会对主题所讨论内容获取的点赞数产生很大影响,而点赞数高的主题也往往会在评论区有较多用户参与讨论,这致使了公众关注焦点转移,进而加快了主题传播的速度。平台需要重视和利用好高质量用户在灾害事件中的社会化网络功能,让高质量用户所发布专业知识、受灾需求等信息的传播更为有效。平台也应当做好这些用户的内容把关,与政府形成畅通的反馈渠道,及时阐述真实准确的事件情况,合理引导舆情的发展方向。

5 结论

本文以社会化问答平台知乎为例,结合舆情生命三周期理论,揭示了自然灾害事件舆情传播的演化规律。在理论层面,以往的各类网络舆情研究通常通过 LDA 等主题模型进行主题

提取以及困惑度确定主题数量,而本文根据平台内容特点选取 CTM 模型以得到更好的挖掘效果,并且采用困惑度和主题一致性相结合的方式对照分析,提升了主题数量确定方式的合理性,并且,引入一种适用知乎平台主题强度和主题传播广度的计算方法,可视化结果可以更直观反映主题的周期变化情况,为未来的突发事件舆情主题演化研究提供一定的方法参考。在实践层面,以往研究多局限于单个案例,本文对比分析前后两年的自然灾害舆情主题演化过程,研究发现在自然灾害事件发生后数天内舆情传播速度极快,且不同周期下舆情主题有较为明显的划分类型,更为全面地展示了用户对自然灾害这类型事件舆情关注的变化情况,有利于自然灾害事件下应急管理相关部门及时对舆情事件做出识别,引导和控制管理过程,为政府在社交媒体网络上对自然灾害事件的舆情应对提供决策支持。

由于本文只涉及不同时间维度上的两个案例,未来研究中可引入其他维度如空间维度进行多维度综合分析,也可以采取数据量更大或时间范围更广的案例进行深入研究,得出普适性更强的结论。

参考文献

- [1] 河南省人民政府新闻办公室.河南举行全省防汛救灾新闻发布会(第十场)[EB/OL].(2021-08-02)[2022-08-01].<http://www.scio.gov.cn/xwfbh/gssxwfbh/xwfbh/henan/Document/1710343/1710343.htm>.
- [2] 安璐,吴林.融合主题与情感特征的突发事件微博舆情演化分析[J].图书情报工作,2017,61(15):120-129.
- [3] 宋宁,刘婵君.过程与要素视角下的突发事件网络

- 舆情演化规律研究综述[J]. 图书情报工作, 2016, 60(15): 140-147.
- [4] 新浪财经. 知乎财报出炉:2021 年营收 29.59 亿, 月活用户 1 亿[EB/OL]. (2022-03-04)[2022-08-01]. <http://finance.sina.com.cn/jjxw/2022-03-14/doc-imcwipih8455201.shtml>.
- [5] 沈波, 赖园园. 网络问答社区“Quora”与“知乎”的比较分析[J]. 管理学报, 2016, 29(5): 43-50.
- [6] 李纲, 陈璟浩. 突发事件情境下网络问答社区用户构成和行为分析——基于问题胶囊事件的实证研究[J]. 图书情报工作, 2013, 57(24): 95-100.
- [7] 王志英, 邓航宇, 王念新, 等. 问答社区信息安全突发事件应急知识传播模型研究[J]. 情报杂志, 2019, 38(10): 136-145.
- [8] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.
- [9] BLEI D M. Probabilistic topic models [J]. Communications of the ACM, 2012, 55(4): 77-84.
- [10] 王璟琦, 李锐, 吴华意. 基于空间自相关的网络舆情话题演化时空规律分析[J]. 数据分析与知识发现, 2018, 2(2): 64-73.
- [11] 姜金贵, 闫思琦. 基于主题和情绪相互作用的微博舆情演化研究——以“红黄蓝虐童事件”为例[J]. 情报杂志, 2018, 37(12): 118-123.
- [12] 王子涵, 王玉辉, 王雷, 等. 一种基于社交媒体的突发事件话题演化分析系统研究[J]. 信息网络安全, 2017(9): 98-102.
- [13] 李月. 突发公共卫生事件中公共政策主题演化研究——以国家中心城市官方微信为例[J]. 情报杂志, 2020, 39(9): 143-149.
- [14] 刘国威, 成全. 基于网络舆情生命周期的微博热点事件主题演化研究[J]. 情报探索, 2018(4): 11-19.
- [15] 王艳东, 李昊, 王腾, 等. 基于社交媒体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, 41(3): 290-297.
- [16] 吴小兰, 章成志. 基于突发事件特征网络的用户社区发现与社区主题演化研究——以新浪微博 H7N9 事件为例[J]. 情报理论与实践, 2017, 40(5): 94-98+60.
- [17] 安璐, 梁艳平. 突发公共卫生事件微博话题与用户行为选择研究[J]. 数据分析与知识发现, 2019, 3(4): 33-41.
- [18] 张兆阳, 王君领, 黄佳妮, 等. 重大突发事件期间微博主题与用户行为的关联演化研究[J]. 信息资源管理学报, 2021, 11(2): 28-38.
- [19] 李纲, 陈思菁, 毛进, 等. 自然灾害事件微博热点话题的时空对比分析[J]. 数据分析与知识发现, 2019, 3(11): 1-15.
- [20] 李绍攀, 赵飞, 周义棋, 等. 基于微博数据的台风灾害舆情分析与灾害损失估计[J]. 清华大学学报(自然科学版), 2022, 62(1): 43-51.
- [21] 陈苗苗, 安璐. 突发传染病情境下社会化问答平台用户角色形成及转变——以知乎平台为例[J]. 图书情报工作, 2022, 66(12): 68-81.
- [22] BLEI D, LAFFERTY J. Correlated topic models[J]. Advances in neural information processing systems, 2006(18): 147.
- [23] STEYVERS M, GRIFFITHS T. Probabilistic topic models[M]. Handbook of latent semantic analysis. London: Psychology Press, 2007: 439-460.
- [24] 余森森, 王俊丽, 赵晓东, 等. PAM 概率主题模型研究综述[J]. 计算机科学, 2013, 40(5): 1-74.
- [25] VAYANSKY I, KUMAR S A P. A review of topic modeling methods[J]. Information Systems, 2020(94): 101582.
- [26] 李纲, 海岚, 陈璟浩. 突发自然灾害事件网络媒体报道的周期特征分析——以地震和台风灾害为例[J]. 信息资源管理学报, 2015, 5(3): 18-24.
- [27] ARUN R, SURESH V, VENI MADHAVAN C E, et al. On finding the natural number of topics with latent dirichlet allocation: Some observations[C]// Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, 2010: 391-402.
- [28] WANG B, LIU S, DING K, et al. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence

- analysis: a case study in LTE technology[J]. Scientometrics, 2014, 101(1): 685-704.
- [29] NEWMAN D, LAU J H, GRIESER K, et al. Automatic evaluation of topic coherence[C]//Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010: 100-108.
- [30] ZHAO W, CHEN J J, PERKINS R, et al. A heuristic approach to determine an appropriate number of topics in topic modeling[C]//BMC bioinformatics. BioMed Central, 2015, 16(13): 1-10.
- [31] MIMNO D, WALLACH H, TALLEY E, et al. Optimizing semantic coherence in topic models[C]// Proceedings of the 2011 conference on empirical methods in natural language processing. 2011: 262-272.
- [32] VERMA S, VIEWEG S, CORVEY W, et al. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2011, 5(1): 385-392
- [33] 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 systems. 2010: 1079-1088.
- [34] HOLLANDER E P. Emergent leadership and social influence[M]. Washington: Washington University, The Social Science Institute, 1959.