基于微博舆情数据的震后有感范围提取研究*

曹彦波，吴艳梅，许瑞杰，张方浩

（云南省地震局，云南 昆明 650224）

摘要：提出了基于微博舆情信息的震后有感范围快速判定技术框架，构建了微博舆情数据的获取方法和技术流程。根据中国地震烈度表和地震现场工作调查规范，将微博特征词与地震灾情速判指标进行关联匹配，建立微博灾情信息分类指标体系，通过自然邻点空间插值方法将离散分布的微博灾情点转化为连续分布的灾情面，形成震后灾区有感范围的时空变化特征分布图，辅助灾情快速判定。以 2014 年景谷 6.6 级地震为例，进行探索和实践。结果表明：在震后 1～2 h 内，微博用户活跃度高，信息量大且丰富，对信息充分挖掘有助于对灾情的宏观把握，对救灾决策部署有一定的参考意义，弥补了传统获取技术时效性差、数据量少、覆盖面小等问题。

关键词：微博舆情数据；灾情判定；有感范围提取

中图分类号：P315.941 **文献标识码**：A **文章编号**：1000－0666(2017)02－0303－08

0 引言

地震发生后，灾情信息的快速获取、处理、分析和研判是各级党委政府、各级抗震救灾指挥部成员单位部署抗震救灾工作，派遣救援力量、调配救灾物资的关键环节，尤其是震后 2 h 的黑箱期内，如何快速判定灾区影响范围灾情时空分布、震害规模、强度等是地震应急灾情快速获取及服务的关键（聂高众等，2012）。目前，在震后有感范围确定方面，主要有以下几个途径：一是通过“三网一员”、政府、地震部门应急人员电话、传真，网站灾情填报等方式获取灾情，绘制有感范围图；二是根据烈度衰减模型快速计算生成地震影响场来预估灾区范围和强度（汪素云等，2000；王景来，宋志峰，2001；张方浩等，2016a）；三是基于智能手持采集终端（PDA、12322、IOS/Andrio 手机端 APP 等）获取地震信息，生成有感范围分布图（郑黎辉等，2012；陈维锋，2014；章熙海等，2014）；四是通过网络爬虫在网站上获取灾情信息，通过地址匹配、空间定位解析后插值生成有感范围分布图（帅向华等，2009，2013；胡素平，帅向华，2012；杨天青等，2016）。在实际地震应急中，上述几种途径在信息获取的时效性、获取效率、信息量、空间范围上存在一定的局限性，短时间内都难以全面客观地反应灾区有感范围的强度和分布，“互联网＋”时代的来临为我们在震后灾情快速获取方面提供了一种新的解决思路。

根据中国互联网络信息中心（CNNIC）发布的《第 38 次中国互联网络发展状况统计报告》显示，截至 2016 年 6 月，中国网民规模达 7.10 亿，互联网普及率为 48.8%，手机网民规模达 6.56 亿，微博客用户 2.42 亿。从统计数字可以看出，随着移动互联网技术的飞速发展，数量众多的个人成为信息传播的重要载体。相对于手机信令、浮动车、微信等数据，以新浪微博为代表的新兴社交平台具有实时性、互动性、强扩散、空间分布广泛性等特点，微博数据可以在互联网上被免费、公开地获取（廉捷等，2011；刘经南等，2014；仇培元等，2016）。尤其是在破坏性地震发生后数小时内，大量与地震相关的信息发布并广泛传播，汇集形成海量数据，包括用户账号、发布时间、经纬度坐标、博文、图片、微视频、关注热点等，这些数据中包含有地震灾情信息，如震感、人员伤亡、房屋破坏、生命线工程破坏、地震地质灾害等（王松等，2014；何宗宜等，2015；徐敬海等，2015）。通过对这些微博“大数据”进行充分挖掘、分析、表达和应用，能客观地反映灾情时空演变规律，辅助地震灾情快速研

* **收稿日期**：2017－01－06.

基金项目：中国地震局震灾应急救援司专项课题《云南地震公共服务平台研发》和《基于微博位置信息的地震灾害速判方法研究》共同资助.

判，服务政府应急救援决策。

本文根据微博舆情数据特点和传播特性，研究如何利用微博舆情数据分时段提取地震有感范围，并以 2014 年景谷 6.6 级地震为例进行应用检验。

1 研究技术框架

当破坏性地震发生后，首先根据地震三要素信息，通过微博 API 调用、关键字检索、网络爬虫、专业地理抓取等技术手段，实时获取微博用户发布的信息，信息主要来源于新浪、腾讯、网易、人民网等主流网站微博用户，对这些信息进行存储管理，形成结构化的数据库。其次，对微博数据进行解析、去重，提取有效信息，包括微博发布时间、博文内容、图片、空间经纬度坐标等，并对核心博文内容进行中文分词、清洗等挖掘处理，提取与地震灾情相关的特征词，根据相关标准和规则对微博数据与地震烈度判定的描述性信息进行关联匹配，建立微博地震灾情信息分类表。最后，以微博灾情节点为基础进行空间插值，将离散分布的灾情点转化为连续分布的灾情有感范围图，描述灾情时空演变规律，辅助灾情研判。具体研究技术框架如图 1 所示。

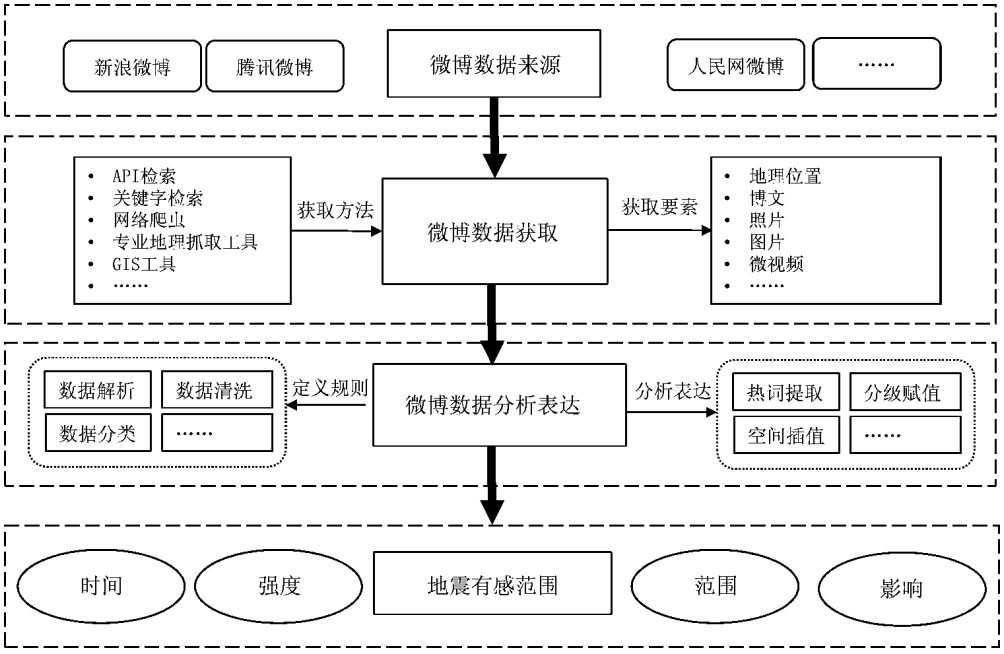


图 1 技术框架

Fig. 1 The technical framework

2 微博数据获取

微博数据获取方式有网络爬虫技术和调用微博官方 API 接口两种途径，基于网络爬虫技术获取信息的基本流程是通过设定入口 URL 地址，按照一定的爬行策略将网页内容保存，并提取网页中有效地址作为下一次爬行的入口 URL 地址，直到爬行完毕。由于地震灾情信息抽取和空间定位要求，该方式信息获取效率不高，空间地理位置还需通过地名规则、地址匹配技术进行解析获取，另外，多次访问有账号屏蔽风险（廉捷等，2011）。因此，本文以当前用户基数较大的新浪微博为例，注册认证后获取调用新浪微博的 API 权限，通过调用相关 API，解析服务器返回的 JSON 数据文档获取微博信息，该方式微博信息获取时效性高，数据格式清晰，便于数据的存储和解析。微博数据获取技术流程如图 2 所示。

3 微博数据分析表达

3.1 微博数据分析处理

面对海量的微博信息“大数据”，为提高数据挖掘效率和准确率，需对原始数据进行解析、去

重，提取微博的发布时间、内容、图片、经纬度坐标等有效信息，并对核心博文内容进行中文分词、清洗等挖掘处理，滤掉一些频繁出现而意义又不大的词，比如“的”“就”“是”“和”等语气助词、副词、介词和连词，提取与地震灾情相关的特征词、热词，对微博灾情信息进行分类和编码，具体流程如图 3 所示。

对微博信息进行数据挖掘完成后，建立微博与地震灾情信息分类映射是微博灾情可视化表达的关键环节。通过对 2014 年云南地区 70 余次地震新浪微博博文内容进行解析，提取主体特征词，从结果分析看，震后与地震相关的内容，主体集中于人的反应、器物反应方面，约占统计的 70% 以上，房屋破坏、人员伤亡、生命线工程破坏的信息以及地震地质破坏方面的较少。依据《中国地震烈度表》《地震现场工作第 3 部分——调查规范》等相关标准，将微博信息与地震灾情描述性信息进行关联匹配，建立了微博灾情分类表（表 1）（曹彦波等，2010；张方浩等，2016b）。

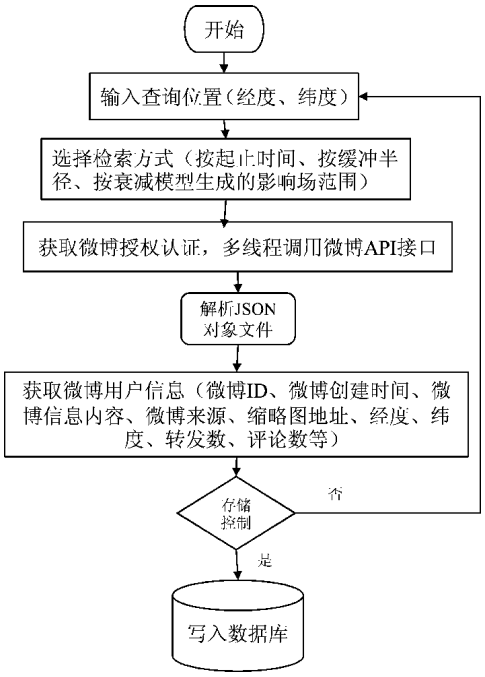


图 2　数据获取技术流程

Fig. 2　The data acquisition process

表 1　微博灾情信息分类表

Tab. 1　The information classification of microblog disaster

灾情分类	灾情描述						
	有感	轻微区	轻度区	中度区	重灾区	极重灾区	巨灾区
人的反应	有震感	有震感，有感觉，梦中惊醒	震感明显，多数人有感、惊醒、骑行有感	震感强烈，惊慌多数人惊逃，站立不稳，骑行不稳	害怕，摇晃颠簸，行走困难	很害怕，站不稳、坐不稳，跌倒	非常害怕，骑行摔倒，抛起、颠簸、栽倒
器物反应	吊灯摇晃	门、窗作响，悬挂物明显摆动	杯子中水振荡、悬挂物或树枝明显摆动、器皿碰撞作响	悬挂物剧烈摇摆或损坏坠落、书物掉落、轻家具移动	多数家具移动、部分翻到、树干摇动、树枝折断	器物翻倒、树干折断、衣柜等重家具和放置稳当的家具翻倒	器物损毁、家具和电器损毁、树木倒塌
房屋破坏	无	个别破坏，数间，屋架颤动，掉灰，微细烈缝，掉瓦	基本完好，数十间，墙体开裂，梭、掉瓦，填充墙体开裂	轻微破坏，数百间，屋架倾斜、脱榫，墙体开裂，梭、掉瓦，填充墙开裂	中等破坏，数千间，局部倒塌，开裂明显，X 型裂缝贯通，填充墙局部倒塌	严重破坏，数万间，结构严重破坏，较严重的水平或“X”型贯通裂缝	毁坏，十万间以上，普遍倒塌
人员伤亡	无	几无死亡：0 人，受伤：0 – 10 人	个别死亡：1 ~ 4 人，受伤：11 ~ 50 人	少量死亡：5 ~ 20 人，受伤：51 ~ 100 人	较多死亡：21 ~ 99 人，受伤：101 ~ 500 人	重大死亡：100 ~ 500 人，受伤：501 ~ 1 000 人	特别重大死亡：大于 500 人，受伤：大于 1 000 人
其他（生命线震害、地震地质灾害、救援行动等）	无	几无；个别轻微受损等	轻度；路面轻微开裂；生命线设施轻微受损；零星落石、滑坡等	中等；路面开裂；生命线设施轻微变形、开裂；少量土石滑落、个别滑坡点等	较大；路基变形；生命线设施开裂、变形、受损；大量崩滑、滑坡、泥石流等	重大；路基下沉；生命线设施断裂、损坏、局部垮塌；大规模山体崩塌、滑坡、泥石流等	特大；路基扭曲；桥梁涵洞垮塌；生命线设施中断、泄露、损毁、垮塌、坍塌；巨大规模的山体崩塌、滑坡、泥石流

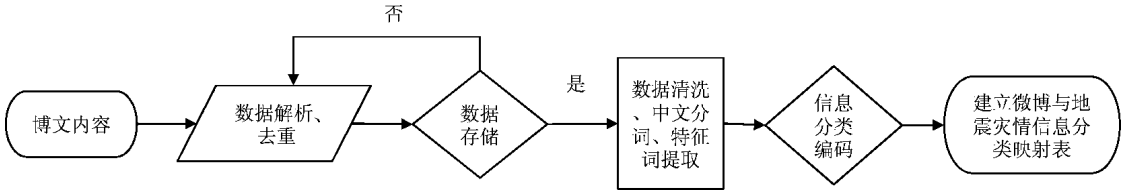


图 3 微博数据挖掘流程

Fig. 3 Searching flow of microblog data

3.2 微博数据可视化表达

微博数据的空间可视化表达是实现分析灾情时空演变规律的基础，震后获取到的微博灾情数据往往是在地理上分布不规则的离散数据，为了能够更直观地了解地震灾情时空分布特征，需以这些微博数据点为基础进行空间插值。常用的空间插值算法有反距离加权插值法（Inverse Distance to a Power）、克里金插值法（Kriging）、最小曲率插值法（Minimum Curvature）、样条函数插值法、Shepard 插值法和自然邻点插值方法（Natural Neighbor Interpolation）等（高洋，张健，2005）。本文将采用自然邻点插值方法来处理高度离散化分布的不规则微博灾情节点，通过插值拟合来描述灾情空间尺度上的变化特征。

由于震区绝大部分微博用户群体是一般公众，个人震感不一，对灾情的描述也有差别，为方便对微博灾情节点进行空间插值，使拟合出来的有感范围更接近实际，基于中国地震局工程力学研究所提出的“等震线长短轴半径与烈度对应经验关系”，计算微博灾情位置距震中的距离，根据距离震中远近对微博数据进行分级赋值，共分 7 级：1 代表有感区，对应烈度表的Ⅳ度区；2 代表轻微区，对应Ⅴ度区；3 代表轻度区，对应Ⅵ度区；4 代表中度区，对应Ⅶ度区；5 代表重灾区，对应Ⅷ度区；6 代表极重灾区，对应Ⅸ度区；7 代表巨灾区，对应Ⅹ度及以上（表 2）。

表 2 微博灾情信息分级表

Tab. 2 The information grading of microblog disaster

灾情分级	灾情描述	烈度	震感描述	参考长轴半径/km				
				$7.8 \leqslant M \leqslant 8.0$	$7.5 \leqslant M \leqslant 7.7$	$6.8 \leqslant M \leqslant 7.4$	$6.0 \leqslant M \leqslant 6.7$	$5.2 \leqslant M \leqslant 5.9$
7	巨灾区	≥Ⅹ	山河巨变	18 ~ 39	6 ~ 19	—	—	—
6	极重灾区	Ⅸ	剧烈	39 ~ 78	16 ~ 31	12 ~ 15	—	—
5	重灾区	Ⅷ	激烈	95 ~ 116	79 ~ 94	33 ~ 75	5 ~ 28	—
4	中度区	Ⅶ	强烈	120 ~ 161	96 ~ 116	45 ~ 94	16 ~ 39	4 ~ 14
3	轻度区	Ⅵ	中等	259 ~ 500	140 ~ 220	97 ~ 132	41 ~ 86	16 ~ 39
2	轻微区	Ⅴ	轻度	—	—	—	—	—
1	有感区	Ⅳ	有感	—	—	—	—	—

4 微博数据应用实践

2014 年 10 月 7 日 21 时 49 分 39 秒，云南省普洱市景谷傣族彝族自治县发生 6.6 级地震，震中位于（23.4° N，100.5° E），震源深度 5.0 km。这是进入新世纪以来云南省发生的震级最大的一次地震，影响范围广，引起较多的微博用户关注。笔者通过调用新浪 API，以本次地震震中位置为中心，150 km 为搜索半径，数据采集时段为震后 24 h。共获取到 1 231 条微博信息，经过清洗筛选后剩余 281 条与本次地震相关的灾情信息，包括人的反应信息 178 条，器物反应信息 56 条，房屋破坏信息 15 条，其他信息 26 条。发布这些信息的微博用户地理位置上主要分布在普洱市、临沧市、西双版纳州 3 个州（市）的 16 个县（区），震中附近区域震感强烈，微博活跃用户群体主要集中在永平镇、距离震中较近的景谷县城威远镇以及人口密集的普洱市主城区（图 4）。

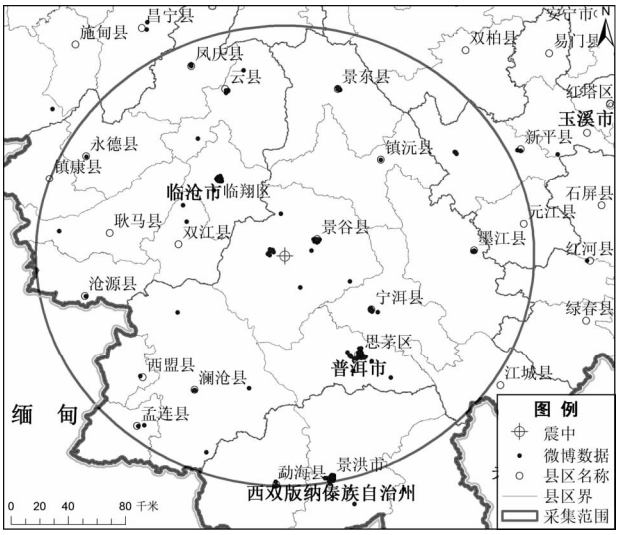


图 4　微博灾情数据空间分布

Fig. 4　Spatial distribution of microblog disaster data

从灾情数据分类结果来看，在震后 24 h 内，灾情描述信息以人的反应和器物反应为主，占总信息条数的 83%，而房屋破坏、地震地质、生命线破坏等情况描述较少，不到 20%，主要因为微博用户群体以一般公众为主，博文的内容主体集中在微博用户自身所处环境的感觉、表情、心情和身边器物反应的描述，对于其他如房屋破坏、地震地质，生命线破坏等比较专业的灾情描述不是很多（图 5）。

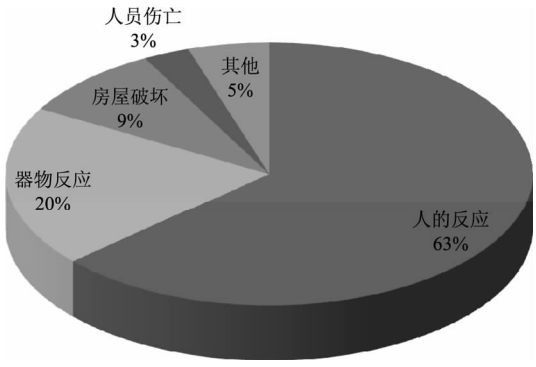


图 5　微博灾情数据分类

Fig. 5　Classify of microblog disaster data

从震后 24 h 内的微博灾情数据分时段统计结果看，大量的信息集中在震后 2 h，共发布 145 条，占总条数的 50% 左右，第一条微博信息发布于 2014 年 10 月 7 日 21 时 56 分 32 秒，也就是震后 3 min，发布的内容为“就在 1 分钟前，地震了，好恐怖［泪］，这个美丽的地方又地震了”，地理位置是（101.043 5°N，23.0588 8°E），距离本次震中 64 km（图 6）。对获取到的数据进行分析挖掘，提取特征词库，拟合形成了震后 10 h 灾情演变时空特征分布图（表 3，图 7）。

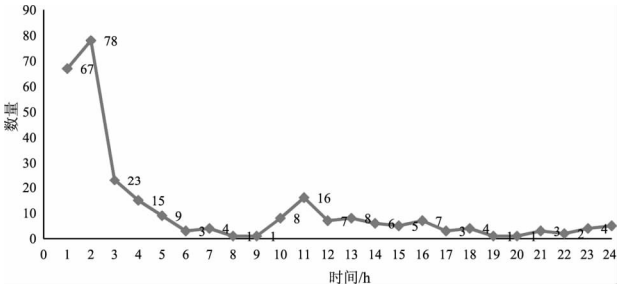


图 6　震后 24 h 内微博灾情数据分时段统计

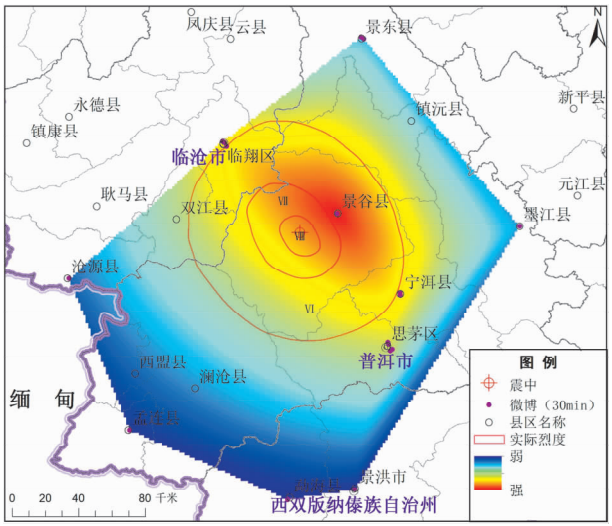
Fig. 6　Statistics of microblog disaster data in 24 h after the earthquake

表 3　景谷地震微博灾情特征词库

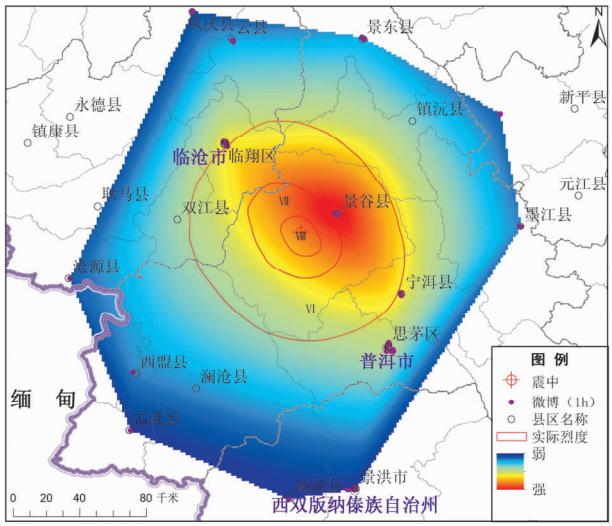
Tab. 3　Thesaurus of microblog disaster characteristic of Jinggu earthquake

灾情分类	特征词
人的反应	摇晃、头晕、站不稳，震感强烈，地震持续时间长，余震，震感，感觉地震了，摇晃强烈且时间长，震感超级强烈，摇晃的很厉害，余震不断，紧张，担心，委屈，忧愁，有惊无险，好想哭，吓哭了，吓死了，太害怕了，恐怖，害怕，吓到了，人心惶惶的，吓得不敢睡了……
器物反应	路灯摇晃，灯和窗帘都摇晃了，床在摇摆，玻璃窗发出声音，桌子响，门晃动，凳子摇晃，床在摇，床桌子稍微一动……
房屋破坏	宿舍楼整栋都在晃，房屋摇晃，房屋不是很严重，房屋没有影响，房子开裂，房屋受损，影响严重，房子在晃，教学楼成危楼，四周墙体裂了，卫生院房屋受损……
人员伤亡	有人受伤，治疗，23 名伤员，5 名重伤，重伤患者，十多名伤者，救护车、医疗器械，周围没什么伤亡……
其他	石头滚下来，路开裂了，消防队，救援，道路通畅，睡帐篷，物资，帐篷，棉被，24 个基站停电，通信保障车……

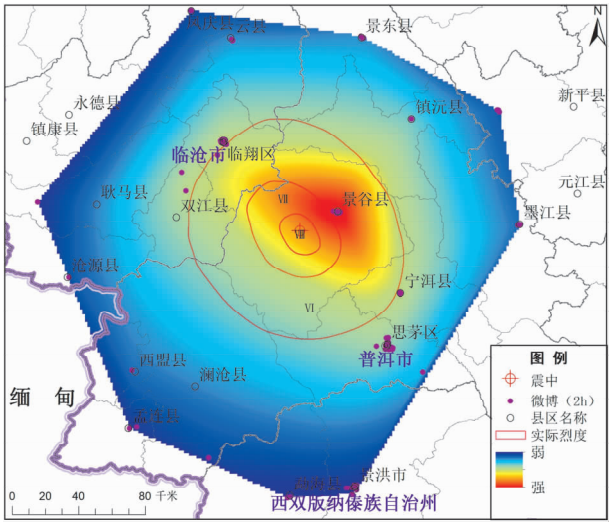
从景谷分时段灾情有感范围图分析得知，在震后 30 min 内，由于震后极震区短暂通信中断和信号拥堵，震中永平镇附近无微博信息，信息少部分来源于景谷县城威远镇的微博用户，大部分



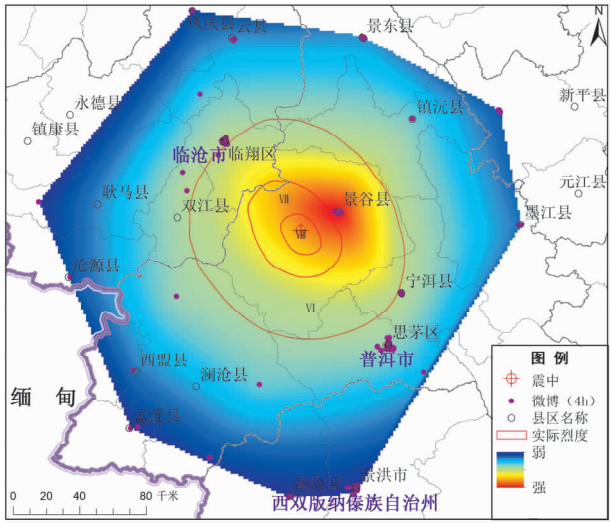
（a）震后 30 min



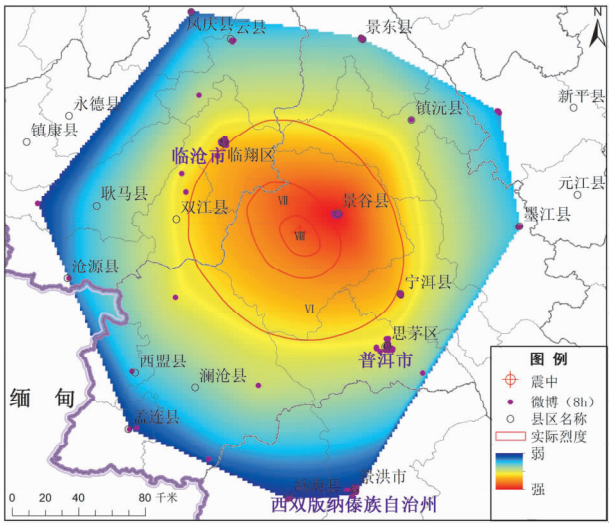
（b）震后 1 h



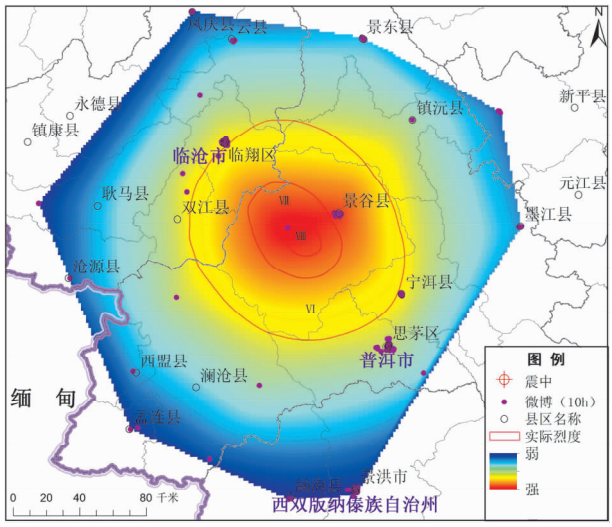
（c）震后 2 h



（d）震后 4 h



（e）震后 8 h



（f）震后 10 h

图 7 震后 10 h 微博灾情时空演变特征图

Fig. 7 The map of spatio – temporal evolution characteristics about Sina blog within 10 h after the earthquake

来自Ⅵ度区以外的临翔区和思茅区，景谷县城附近震感强烈，周边的景东、墨江、景洪、勐海、孟连、沧源有感，根据该图，可帮助决策部门在较短时间内把握宏观灾情的空间分布和重点救助区域的初步判断。震后 1 ~4 h，随着救援力量、救灾物资的投入，灾区社会秩序逐步恢复，灾民得到救助和转移安置，通信恢复正常，微博粉丝活跃度逐渐增加，信息发布量增多，有感范围增大，但震感较强烈的区域还是在震中附近。震后 8 h，随着救援行动的进一步深入，震中永平镇附近也发布有 10 多条相关的微博灾情信息，灾情进一步明朗，空间分布上主要集中在永平镇和威远镇一带，强有感区边界也较清晰明显。震后 10 h，微博灾情有感范围基本与实际地震烈度范围一致。

5　结语

“互联网 +”时代背景下，微博等新兴社交平台产生的“大数据”信息丰富、多变、复杂，充分挖掘利用这些数据，对震后有感范围提取，灾情快速判定提供了新的研究方法和技术实现途径。本文提出了基于微博舆情数据的震后有感范围快速判定的技术框架，详细论述了微博舆情数据的获取方法和技术流程，根据《中国地震烈度表》和《地震现场工作第 3 部分——调查规范》等规范，将微博主体特征词与地震烈度判定的描述性信息进行关联匹配，建立了微博地震灾情信息分类表，采用自然邻点方法将微博灾情节点通过插值拟合来描述地震有感范围时空变化特征。最后以景谷 6.6 级地震为例获取了震后微博灾情数据，对灾情数据进行了分析处理，生成有感范围时空演变图，对于决策部门震后快速把握灾情提供了一种可行和有效的途径。但在实际地震应用中，应将微博拟合结果与衰减模型烈度、仪器烈度、震源机制、破裂过程等信息进行对比分析和综合判定，提供更科学、合理的决策建议。

参考文献：

曹彦波，李永强，胡秀玉 . 2010. 地震现场灾情信息编码体系研究[J]. 地震研究，33(3)：344 – 348.

陈维锋，郭红梅，张翼，等 . 2014. 四川省地震灾情快速上报接收处理系统[J]. 灾害学，29(2)：116 – 122.

仇培元，陆锋，张恒才，等 . 2016. 蕴含地理事件微博客消息的自动识别方法[J]. 地球信息科学学报，18(7)：886 – 893.

高洋，张健 . 2005. 基于自然邻点插值的数据处理方法[J]. 中国科学院研究生院学报，22(3)：346 – 351.

何宗宜，苗静，彭将，等 . 2015. 结合微博数据挖掘的时空特征分析[J]. 测绘通报，(10)：60 – 64.

胡素平，帅向华 . 2012. 网络地震灾情信息智能处理模型与地震烈度判定方法研究[J]. 震灾防御技术，7(4)：420 – 430.

廉捷，周欣，曹伟，等 . 2011. 新浪微博数据挖掘方案[J]. 清华大学学报：自然科学版，51(10)：1300 – 1305.

刘经南，方媛，郭迟，等 . 2014. 位置大数据的分析处理研究进展[J]. 武汉大学学报：信息科学版，39(4)：379 – 385.

聂高众，安基文，邓砚 . 2012. 地震应急灾情服务进展[J]. 地震地质，34(4)：783 – 789.

帅向华，侯建盛，刘钦 . 2009. 基于地震现场离散点灾情报告的灾害空间分析模拟研究[J]. 地震地质，31(2)：321 – 333.

帅向华，胡素平，郑向向 . 2013. 地震灾情网络媒体获取与处理模型研究[J]. 自然灾害学报，(3)：178 – 184.

汪素云，俞言祥，高阿甲，等 . 2000. 中国分区地震动衰减关系的确定[J]. 中国地震，16(2)：99 – 106.

王景来，宋志峰 . 2001. 地震灾害快速评估模型[J]. 地震研究，24(2)：162 – 167.

王松，吴亚东，李秋生，等 . 2014. 基于时空分析的微博演化可视化[J]. 西南科技大学学报，29(3)：68 – 75.

徐敬海，褚俊秀，聂高众，等 . 2015. 基于位置微博的地震灾情提取[J]. 自然灾害学报，24(5)：12 – 18.

杨天青，席楠，张翼，等 . 2016. 基于离散灾情信息的地震烈度分布快速判定方法研究[J]. 地震，36(2)：48 – 59.

张方浩，和仕芳，吕佳丽，等 . 2016b. 基于互联网的地震灾情信息分类编码与初步应用研究[J]. 地震研究，39(4)：664 – 671.

张方浩，蒋飞蕊，李永强，等 . 2016a. 云南地区地震烈度评估模型研究[J]. 中国地震，32(3)：572 – 583.

章熙海，宋法奇，胡晓荣，等 . 2014. 基于 PDA 的地震灾情信息流动采集系统的设计与实现[J]. 地震，34(2)：131 – 137.

郑黎辉，黄声明，林岩钊，等 . 2012. 基于智能手机的地震灾情快速上报系统的设计与实现[J]. 国际地震动态，(6)：164 – 164.

Research about the Perceptible Area Extracted after the Earthquake Based on the Microblog Public Opinion

CAO Yanbo, WU Yanmei, XU Ruijie, ZHANG Fanghao
(*Earthquake Administration of Yunnan Province*, *Kunming* 650224, *Yunnan*, *China*)

Abstract

It is an effective way to obtain disaster information quickly after the earthquake, through the analysis and mining of the microblogs public opinion data, because microblog has the characteristics of real – time, interactive, strong diffusion, wide spatial distribution, and so on. Based on the microblogs public opinion, the technology framework was proposed for the earthquake felt area of fast determining, and the data access methods and technological processes were built. According to Chinese seismic intensity scale and post – earthquake field works for field survey, it sets up a micro – blogging earthquake disaster information classification system through micro – blogging feature associated with the earthquake rapid determinate target match. The perceptible area used for secondary disaster quick judgment is extracted after the earthquake, and secondary disaster quick judgment was achieved through the Natural Neighbor interpolation method of distributed micro – blogging disaster into a continuous distribution. This method has been applied in practice after the earthquake with magnitude 6. 6 occurred on October 7, 2014 in Jinggu county. It is concluded that fully mining microblog information is used for wide grasp of disaster and relief decisions within 1 to 2 hours after the earthquake, with the microblog user activity high, informative and rich. It makes up for the traditional technical problems with inefficient and little coverage of data.

Keywords: microblog public opinion; disaster determining; perceptible area extracting