DOI:10.13203/j.whugis20140804 文章编号:1671-8860(2016)03-0290-08

基于社交媒体的突发事件应急信息挖掘与分析

王艳东[1] 李 昊[1] 王 腾[1] 朱建奇[1]

1 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

摘 要:社交媒体越来越多地被看作是随人们移动的传感器,感知周围发生的事件。当突发事件发生时,大量含有位置信息的文字迅速地充斥整个社交网络。本文探讨突发事件应急信息挖掘与分析的一种新思路。基于社交媒体,建立实时应急主题分类模型,从大量、实时的文本流中快速提取、定位应急信息;针对不同主题,利用统计分析和空间分析方法,探寻突发事件的时间趋势和空间分布,为应急响应提供决策支持。

关键词:社交媒体;突发事件;趋势分析;空间分析;数据挖掘

中图法分类号:P208 文献标志码:A

随着大数据时代的到来,与时空相关的数据挖掘成为当前GIS领域研究的热点。手机数据[1-2]、GPS轨迹数据[3-5]、智能卡数据[6]等都可以作为挖掘相关知识的数据源,特别是以人类为中心的相关活动规律。近几年来,各种社交网站发展迅速,例如国外的Twitter、Facebook以及国内的新浪微博等。越来越多的人愿意在社交媒体平台上发表自己的看法,这也使得社交媒体数据成为了一种重要的数据源,可以反映现实世界的各种社会活动[7-8]。将社交媒体数据与不同的领域知识相结合可以研究和挖掘不同的信息,如结合传染性疾病领域知识,社交媒体数据可以研究疾病的当前状况、疾病的传播模式[9],甚至对疾病的发展趋势做出预测[10]。结合城市规划相关知识,社交媒体数据可以研究不同区域的活动模式[11]、检测城市的用地类型[12]以及评价公共服务设施的合理性等。结合社会学知识,社交媒体数据可以研究探索人类不同的活动模式[13-14]。结合市场决策相关知识,社交媒体数据可以用于研究企业的营销策略、产品传播模式等[15]。结合灾害应急的相关知识,社交媒体数据可用于检测灾害事件的发生[16]以及了解事件发生的状况[17]等。

本文旨在探讨如何利用社交媒体数据,在突发事件发生的时候,通过挖掘与分析,得到有价值的应急信息以辅助决策者作出应急响应,合理分配应急资源。本文基于目前国内最热门的社交媒体平台——新浪微博(一种允许用户通过互联网公开发布并可以及时更新的简短文本形式),建立了基于新浪微博文本数据的应急主题分类模型,从实时、大量的文本流中快速分辨、定位突发事件的实况、救援等应急信息。针对不同主题,从新浪微博数据量和空间属性出发,探寻突发事件随时间的发展趋势并分析可能的影响。同时,利用空间聚类分析找出突发事件的空间分布规律和异常区域,为应急决策提供依据。

1 应急信息的主题分类与定位

在突发事件中,社交媒体数据蕴含着大量的主题、时空等应急信息。通过对实时、海量的应急信息进行分类,能够识别出事件实况、救援、事件影响等主题信息,有利于了解突发事件的状况。

针对微博数据文本简短、主题多样、高时效性、丰富的空间信息等特征,本文提出了一种基于社交媒体文本流的应急主题实时分类与定位模型,其流程如图1所示。

应急信息实时甄别与定位模型是实现对实时获取的微博数据进行分类与定位,本文将主题模型[18](latent dirichlet allocation,LDA)与支持向量机[19](support vector machine,SVM)结合起来。首先利用主题模型发现隐藏在已经获取的微博数据文本中的主题;然后,在这些已发现主题的基础上,

收稿日期:2014-10-29

项目资助:国家自然科学基金(41271399);测绘地理信息公益性行业科研专项经费(201512015);高等学校博士学科点专项科研基金(20120141110036)。

第一作者:王艳东,博士,教授,主要从事城市大数据分析计算相关研究。ydwang@whu.edu.cn

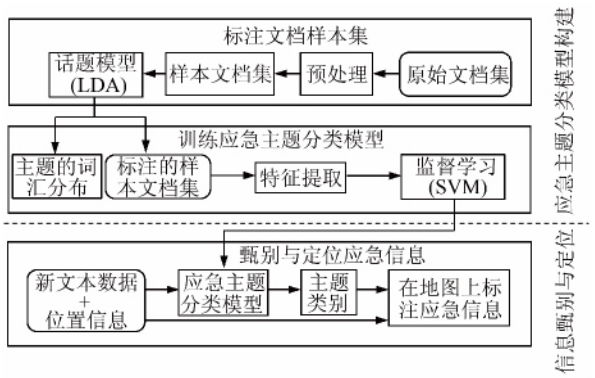


图 1　应急信息实时甄别与定位

Fig. 1　Identifying and Locating Real-time Emergency Information

采用支持向量机进行监督学习,这样,对每一条新获取的微博数据可以利用 SVM 进行实时分类。另一方面,微博数据由于文本短且存在稀疏性,大量的转发而导致重复度高,且口语化明显,因此,模型构建程中需要先进行数据预处理,应急信息实时甄别与定位模型的具体流程如下。

1) 对社交媒体原始文本集合进行去掉重复、广告和低频词汇等预处理;另外,微博中的表情符号传达了重要的语义信息,纳入词典,同时弥补了微博的稀疏性。

2) 利用 LDA,通过计算文本集合的离散词语共现频率来找出隐藏在文本集合中的应急主题;同时输出对应的主题的词汇分布。

3) 基于已发现的主题,利用 SVM 训练已有的训练样本(应急主题和主题的词汇分布)。当新微博文本进入时,通过模型的判断确定该微博文本的主题类别,从而实现微博文本的实时分类。

4) 另外,如果该条微博包含 GPS 位置信息,可以对微博进行地理标注,使用经纬度进行定位。

5) 微博数据由大众参与,实时性极高,能在第一时间反应突发事件的状况和发展态势。因此,每间隔一定时间,新获取的微博文本将被纳入 LDA 主题模型,更新分类模型、词汇分布。

突发事件以"7·21"北京特大暴雨为例:2012 年 7 月 21 日至 22 日 8 时左右,北京遭遇 61 年来最强暴雨及洪涝灾害。据北京市政府公布数据显示,此次暴雨造成房屋倒塌 10 660 间,160.2 万人受灾,经济损失 116.4 亿元。我们实时地收集了从 7 月 20 日 0 时到 8 月 10 日 24 时,以"北京暴雨"为关键词的 706 835 条微博,其中包含 GPS 位置信息的微博有 26 050 条,有 10 988 条位于北京。初始时,取出截止 7 月 21 日 24 时的 79 723 条文本作为应急主题分类模型中 LDA 的训练样本,得到样本文档各自的主题分布和所有主题各自的特征词汇分布,其中部分主题如图 2 所示。

#Topic 4		#Topic 17		#Topic 18		#Topic 27		#Topic 29		#Topic 33	
机场	0.060572	积水	0.096204	预警	0.042271	救援	0.077722	死亡	0.132123	造成	0.033764
小时	0.052066	排水	0.039654	降雨	0.040495	遭遇	0.060422	遇难者	0.049918	损失	0.029508
滞留	0.027527	严重	0.037401	地区	0.040141	袭击	0.04877	灾	0.039559	特大	0.026503
地铁	0.022227	交通	0.023929	最大	0.035745	加油	0.036526	确认	0.034913	车辆	0.023031
旅客	0.021267	瘫痪	0.018862	雨量	0.023859	最强	0.033091	身份	0.032775	情况	0.022429
坐	0.014818	路段	0.017363	小时	0.021661	[话筒]	0.028187	遇难	0.032523	涉水	0.017626
火车	0.014723	无	0.01619	部分	0.020818	无法	0.027953	人数	0.030909	车主	0.015034
真实	0.014431	中心	0.015243	预计	0.019912	致	0.027324	特大	0.029438	进行	0.014099
晚点	0.014421	雨水	0.014202	达到	0.019417	消防	0.02534	致死	0.023085	导致	0.013442
站	0.014053	道路	0.014158	降雨量	0.019362	严重	0.024981	致	0.022143	保险公司	0.013184
没人	0.013371	导致	0.013339	持续	0.015673	现场	0.022374	公布	0.022075	险	0.011235
线	0.011875	立交桥	0.011919	气象台	0.015664	求	0.020939	灾害	0.021528	汽车	0.011104
百年一遇	0.011562	路面	0.011039	全市	0.015511	调派	0.019004	溺水	0.021184	理赔	0.01055
影响	0.010943	长	0.009032	城区	0.015369	情况	0.018881	发现	0.02105	灾害	0.010264
列车	0.010692	图	0.008984	橙色	0.014722	前往	0.018615	发生	0.018821	人员伤亡	0.010012
公交	0.010651	专家	0.008719	蓝色	0.014547	遭	0.018542	房屋	0.018143	及时	0.009236
首都机场	0.01056	市政	0.007733	黄色	0.014382	被困	0.017739	名单	0.017801	引起	0.00783
停运	0.010138	报告	0.007567	平均	0.013896	扩散	0.017288	死者	0.014983	水灾	0.007754
问	0.009949	部门	0.00735	泥石流	0.013396	淹没	0.015447	倒塌	0.014968	受损	0.007695
到达	0.009792	恭喜发财	0.007043	影响	0.013292	人员	0.014371	触电	0.014137	发动机	0.00677
乘客	0.009504	河	0.006357	信号	0.013012	警力	0.013492	雷击	0.013532	自然灾害	0.006449
...	...	...	...	...	...	...	...	...	...	...	...

图 2　相关主题的特征词汇分布

Fig. 2　Distribution of Word Features Within Related Topics

有了标注好的文档样本集(应急主题和词汇),就可以利用 SVM 进行训练。同时,为了验证模型分类的精准度,我们随机将文档样本集分为 5 组,其中 4 组作为训练集,1 组作为测试集,经交叉验证,该模型准确率为 87.5%,这说明此模型适合对微博数据进行应急信息分类。

利用上述应急信息实时甄别与定位模型,将"7·21"北京特大暴雨所有微博进行主题分类,最后得到了 40 个主题类别。分析发现其中有一些很有意义的主题,把这些主题挑选出来并进一步归纳,将相似的主题合并,以便作进一步的分析。例如,从图 2 中的主题的特征词汇分布来看,Topic 29 和 Topic 33 都是讨论暴雨事件所引起的损失与影响方面的内容,将这两个 Topic 的内

容合并为一个有关暴雨事件损失影响的大主题。这样,最终得到“交通状况”、“天气预报”、“灾情信息”、“损失与影响”、“救援信息”、“内涝原因”等6个应急信息相关的主题。同时,基于该模型,实现了微博应急决策支持原型系统。

2 突发事件的时空分析

2.1 突发事件趋势分析

人们除了关注突发事件的应急信息,还想了解突发事件的发展趋势。微博数量随突发事件的时间发展而产生波动,通过分析这些变化可以发现事件总体趋势,有助于提前作出应急对策;另外,人们讨论的主题会随着事件的发展而不断发生变化,利用不同主题微博数目的变化来分析事件发展过程,有助于了解突发事件的发展规律。

2.1.1 总体趋势

研究表明[20],社交媒体数据量与事件的发展存在一定的关系。通过分析“北京暴雨”事件的相关微博数据,可发现突发事件趋势与微博数量有明显的相关关系。图3展示了暴雨事件中每小时的微博总数随着时间的变化趋势。从微博的总体数量趋势图来看,数据集中在暴雨爆发后的一周内,然后开始慢慢地平息,淡出了社交媒体的热门话题。图3中每一天的微博曲线有个最低点,为凌晨4时左右,然后数据开始上升,到达峰值有波动,慢慢下降到另一天凌晨4时左右,呈现循环波动。鉴于此,我们假设认为此时间序列有多期的按天循环波动趋势,利用季节性趋势分解[21]来进一步分析事件情况。

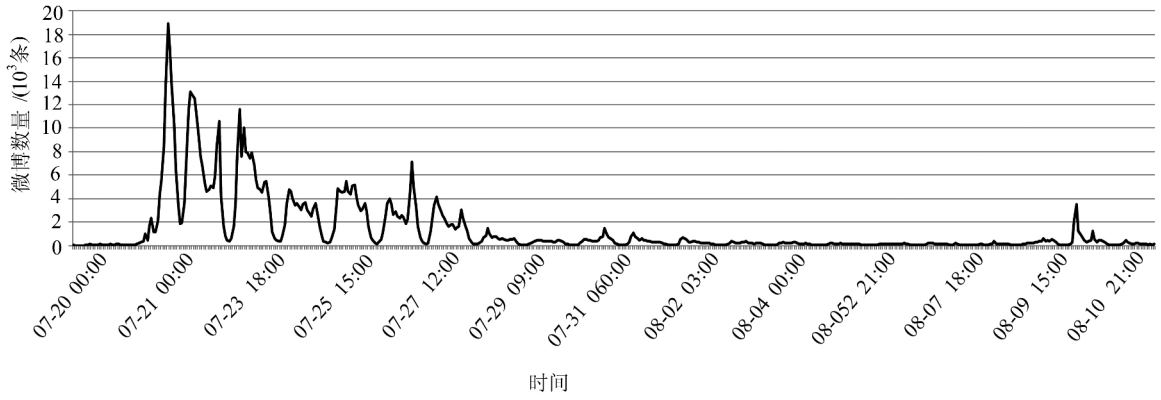


图3 “7·21”北京特大暴雨时间趋势

Fig. 3 Temporal Trend of “7·21” Beijing Heavy Rainstorm

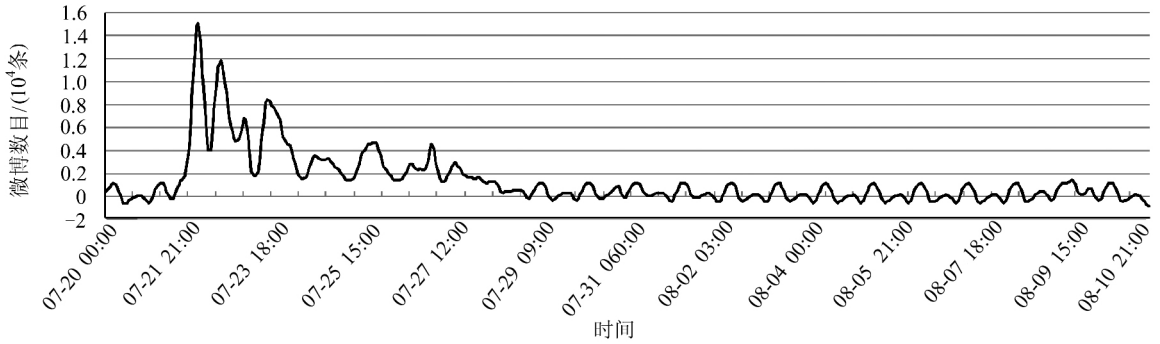
季节性趋势分解将某一时刻的趋势被分解为3个不同的部分,包括趋势循环、季节性因子以及误差项,如式(1)所示:

$$x_t = T_t + S_t + R_t \tag{1}$$

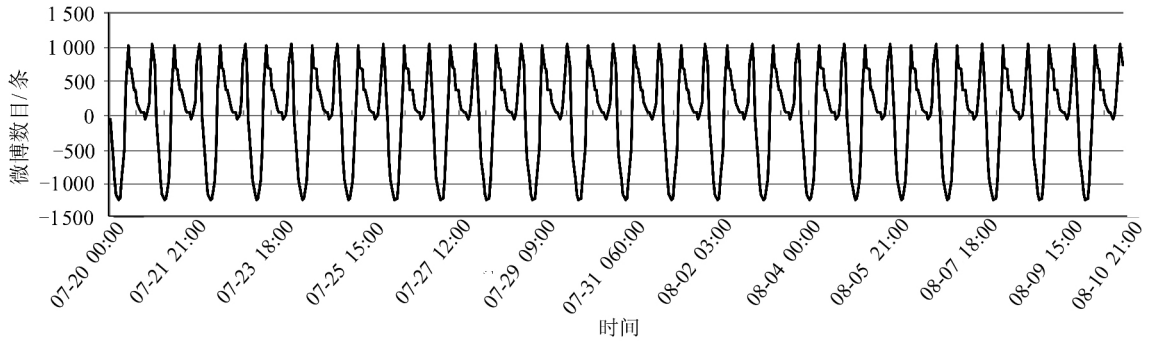
式中,x_t 代表原始数据;T_t 代表总体趋势;S_t 代表季节性因子;R_t 代表误差项。

本文采用季节性趋势分解对北京暴雨相关微博数据量的波动趋势进行分析,结果如图4所示。图4(a)表示总体趋势图,反映了北京暴雨微博数量的总体变化趋势;图4(b)表示季节性因子图,反映了北京暴雨微博数量的周期性变化的部分,从图中可以看出微博数量最低点出现在每天凌晨4时左右,而在每天上午9时和晚上10时会出现峰值,这很好地反映了微博活动的周期性趋势;图4(c)表示季节分解误差图,反映了微博数据量的波动趋势中的随机因子,体现出一些偶然因素所导致的微博数据量的波动。为了将季节性因素从时间序列中分离出去,以便观察又暴雨事件本身对微博数量的影响,我们进一步作了季节性调整;图4(d)是经过季节性调整后的序列,从图中可以得知,北京暴雨相关微博数据量的趋势在暴雨爆发前和暴雨过后呈现规律的周期波动,这是一种正常态。在暴雨发生时,微博的数量激增,微博数量表现出很大的波动,7月21~7月23日微博数据量波动幅度最大,在21日的22时(图中 A 点)达到最大值,并且维持在较高的水平上。此后的7月24、25日两天北京暴雨依然是比较热门的话题,从曲线上看,表现为在较高的水平上周期波动,然而,在26日(图中 B 点),微博数据表现出异常,在短期内数据激增,与通常的模式相异,经查阅资料,可以发现在26日当天,北京市气象局和一些网络电商发布了暴雨预警,引起人们对北京暴雨的再次关注。但是,当天北京市只是降了小雨,并没有带来危害的暴雨发生,所以这次暴雨事件也就慢慢的趋于平缓,并没有在社交网络上引起持续热议。因此,北京暴雨事件的发展趋势

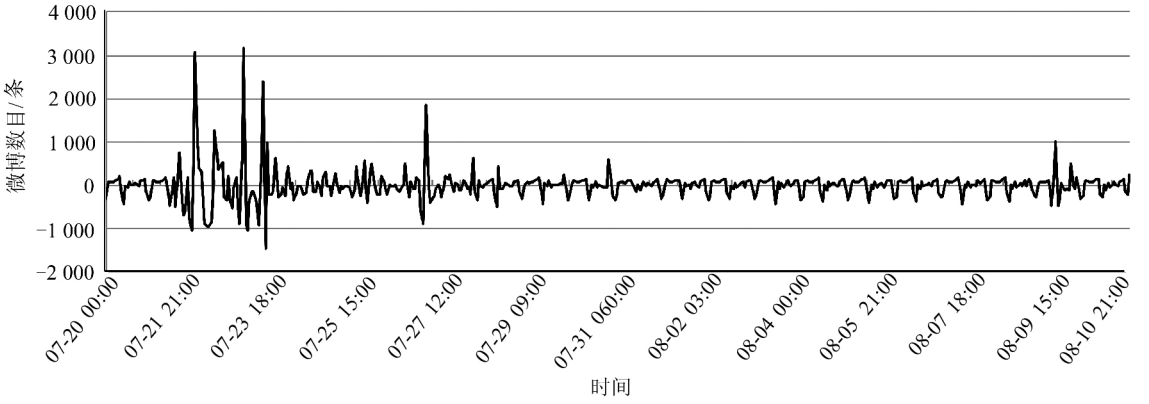
与微博数量的变化趋势存在明显的相关性,可以通过研究微博数量时间序列的季节性趋势来探索相关事件的总体波动趋势、周期模式等不同发展模式,以帮助人们更全面地了解事件的发展过程。



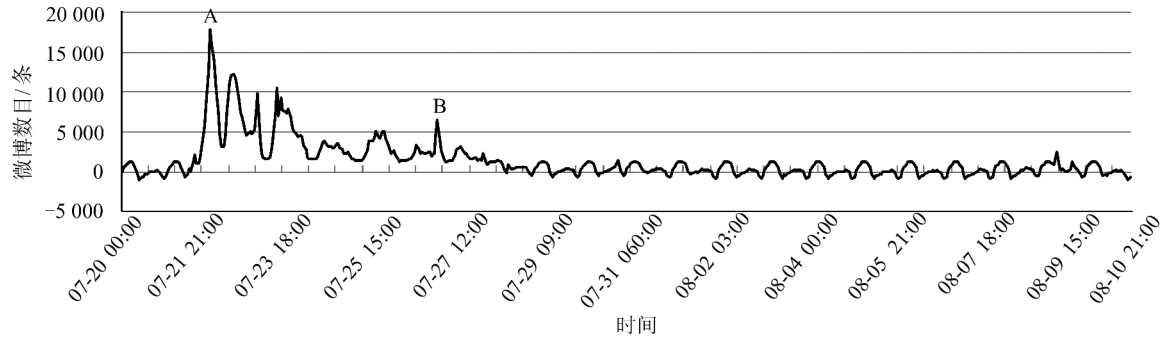
(a) 北京特大暴雨微博数量时间序列的总体趋势



(b) 北京特大暴雨微博数量时间序列的季节性因子



(c) 北京特大暴雨微博数量时间序列的季节分解误差



(d) 北京特大暴雨微博数量时间序列的季节性调整序列

图 4　"7.21"北京特大暴雨时间趋势季节性分解

Fig. 4　Seasonal-trend Decomposition of Temporal Trend of "7.21" Beijing Heavy Rainstorm

2.1.2 主题趋势

人们对应急事件的关注侧重点随着事件发展不断发生变化，而主题直接对应人们关注点。因此，探索主题分布有助于了解应急事件的发展过程。为精确刻划不同主题在时间维度上的变化，我们分别统计“灾情信息”、“天气报道”、“损失影响”3个主题在7月21日6时至24日4时内每个小时内微博数占对应小时内微博总数的比例。如图5所示，在暴雨来临之前，即7月21日的15时之前(图5中A点)，微博主要以天气为主，此时可能由于气象局的暴雨预警和天气变化引起的关注；然后在暴雨袭来的时候，给人们的下班出行带来阻碍，暴雨天气持续，并且产生了严重的灾情，人们关注的重点是灾情本身和极端的恶劣天气，在暴雨中也会有一些损失和影响的关注，但并没有占主要部分，在22日的19时(图5中B点)，暴雨已经慢慢退去的时候，带来的巨大损失和影响开始成为微博讨论的热门话题，主要应为灾后的损失评估和思考。结合整个“7·21”北京特大暴雨事件过程来看，“天气报道”、“灾情信息”、“损失影响”三个主题分别对应暴雨事件的前期、中期和末期。因此，在一个事件发生时，微博中不同主题变化能够反映应急事件发展的不同阶段，通过分析微博的主题，有助于进一步从不同的角度来分析事件的发展。

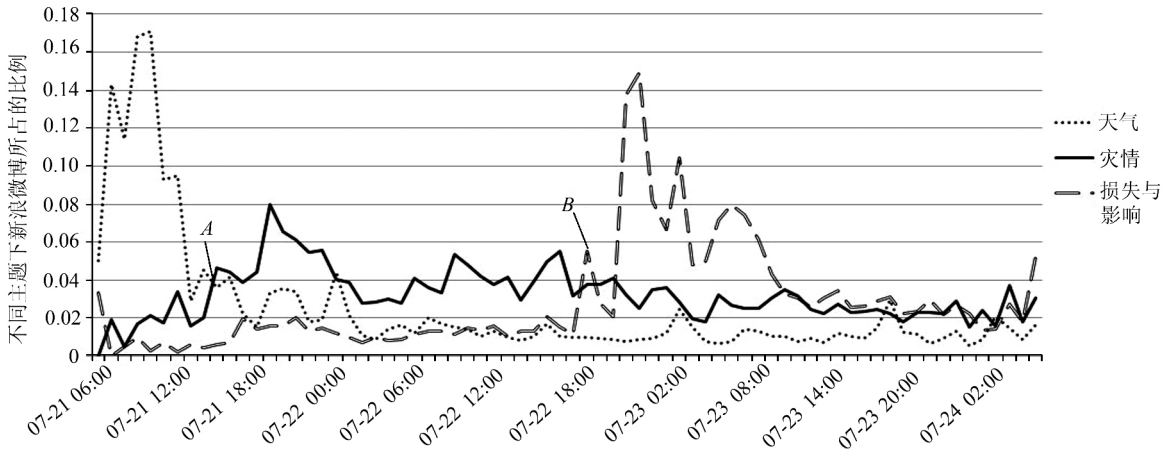


图5 不同主题下微博比例随时间变化曲线

Fig. 5 Trend of Microblogging Under Different Topics Over Time

2.2 探寻空间分布模式

由于包含GPS信息的微博数据具有丰富的位置信息，每条微博可看成一个点状的地理实体。通过对这些点状实体进行空间聚类分析，可揭示突发事件在空间上的热点区域和分布规律。

2.2.1 发现突发事件空间分布

为了探索突发事件的空间分布模式，通过对带有GPS信息的微博点进行聚类，以揭示突发事件在空间上的分布规律，有助于决策者了解突发事件的空间格局，作出正确的应急决策。本文在传统的贪心聚类算法基础上，设计了多层次贪心聚类算法。改进后的算法适应地图缩放，进行多层次的聚类，建立每层次聚类阈值与地图缩放层级的映射，计算聚类簇及其空间范围。随着地图的缩放，可得到不同空间范围下的聚类效果，从而获得不同空间尺度时突发事件空间分布模式。算法步骤如下：

1) 初始时默认地图缩放层级为0，将所有微博点看成一个聚类簇。

2) 根据地图缩放层级递增顺序，计算下一层级层次的聚类阈值。

3) 依次取出上一层级的每个聚类簇，根据步骤2的聚类阈值对该聚类簇重新聚类：依次取出聚类簇内的微博点a，计算其与各个新聚类簇的距离。如果距离小于聚类阈值，将a加入这个聚类簇；否则，形成一个新的聚类簇。

4) 从最低层级至最高层级，重复步骤2)、步骤3)，形成各个层级的聚类簇，计算聚类簇的凸包范围，并存储到树结构中。

图6为2012年7月21日13时～14时带有GPS信息的暴雨微博的聚类结果，在交通高峰时段，暴雨微博点沿地铁呈线状分布，这与人们的正常空间活动相符合。另一方面，微博点在地铁1号线的中段(城区)和东段(通州区)聚集较多。这一点与北京市气象台14时发布的暴雨黄色预警(预计通州、城区大部分地区未来3 h雨强将超过30 mm/h)非常吻合。这说明微博数据的分布不但能够反映出人口的聚集模式，还在一定程度上能够反映暴雨事件的实况信息。

图 6　暴雨微博点聚类图

Fig. 6　Clustering Map of Microblogging about Beijing Heavy Rainstorm

2.2.2　不同主题下微博的空间分布

各主题下的微博不仅具有时间分布规律性，还在空间上有着明显的分布模式。本文选取了“交通”和“灾情”两个主题的微博数据进行了空间聚类分析，“交通”主题包含 1 056 条带有 GPS 信息的微博，“灾情”主题包含 470 条带有 GPS 信息的微博。采用基于密度的聚类算法 DBSCAN[22] 分别对这两个主题的带有 GPS 的微博数据进行了聚类分析，选择聚类半径 400 m、最小邻域点数 4 作为聚类的参数。

对于“交通”主题，如图 7 所示，得到了 27 个聚类中心，对应图中的每一个圆，其中圆的大小表示该聚类中心所包含点的个数。从图中可以看出，这些聚类中心主要分布在北京西站、北京站和首都国际机场。事实上，北京西站、北京站、首都国际机场是北京对外交通的枢纽，受强降雨影响，大量列车晚点，同时航班大面积延误，造成大量旅客滞留。

图 7　“交通状况”主题微博的空间聚类

Fig. 7　Clustering Map of Microblogging Under the Topic of “Traffic”

对于“灾情”主题，如图 8 所示，得到了 4 个聚类中心，这 4 个聚类中心都分布方框所在的区域，这说明在这个区域很可能有重大灾情发生。根据北京官方在北京暴雨事件过后所公布的信息来看，在此次暴雨事件中北京地区核心城区有一人遇难，而这位遇难者正是在这个区域(图中五角星所在的位置)。

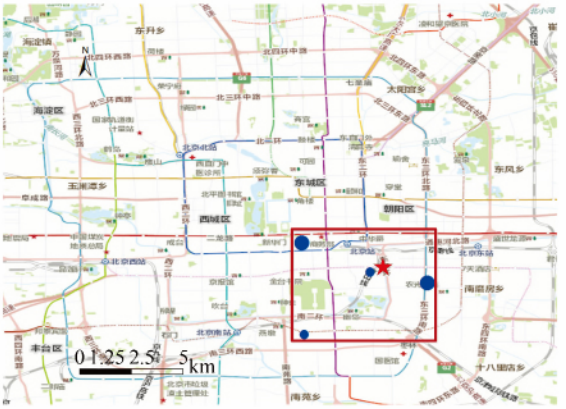


图 8　“灾情”主题微博的空间聚类

Fig. 8　Clustering Map of Microblogging Under the Topic of “Disaster Information”

综合上面两个主题的空间聚类分析来看，结合主题的微博数据聚类分析，能够很好地反映与当前主题密切相关的空间分布状况。在突发事件发生的时候，通过结合主题的空间分析可以针对性地分析事件的某一个方面而去除其他噪音的影响，从而获取更有价值的应急信息。

3　结　语

社交媒体正处于生机勃勃的良好时期，日益增长的跨学科研究人员结合不同的领域解决一系列挑战性的问题。同样，对于灾害应急响领域，社交媒体有着重大的意义。社交媒体是一个大众参与的平台，其中包含的信息具有很强的时效性，当一个突发事件发生时，往往在社交媒体中充斥着与事件相关的各种信息，如果及时获取其中有价

值的信息，将有利于决策者及时作出决定，有效地分配应急资源。

尽管目前社交媒体数据也存在一些问题，比如社交媒体数据与人口在地理上的分布特征具有高度相关性，可能会在一定程度上对分析的结果带来一些困扰。另外，社交媒体数据包含了大量的垃圾信息，还存在一定的片面性，这些都会影响数据的可信度。但是由于数据量的庞大以及众多用户的参与，社交媒体数据仍然应该成为一种重要的数据源。

本文探讨了突发事件应急信息挖掘与分析的一种新思路，对于实时获取的微博数据，根据其主题进行分类，有效地识别出有价值的应急信息，并加以定位。在此基础之上，进一步考虑时空分析，探索突发事件的发展趋势和空间分布规律，有利于发现深度挖掘事件的发展规律，帮助决策者合理调配应急资源以及时处理与应对突发事件。

社交媒体仅是突发事件应急信息来源之一，突发事件的其他属性数据也非常重要，如洪涝灾害中的实时天气、地形数据。今后将进一步结合多方面的突发事件大数据进行协同挖掘分析研究。

参 考 文 献

[1] Gao S, Liu Y, Wang Y, et al. Discovering Spatial Interaction Communities from Mobile Phone Data [J]. *Transactions in GIS*, 2013, 17(3): 463-481

[2] Chen Jia, Hu Bo, Zuo Xiaoqing, et al. Personal Profile Mining Baesd on Mobile Phone Location Data[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 734-738(陈佳，胡波，左小清，等. 利用手机定位数据的用户特征挖掘[J]. 武汉大学学报·信息科学版，2014,39(6): 734-738)

[3] Scholz R W, Lu Y. Detection of Dynamic Activity Patterns at a Collective Level from Large-volume Trajectory Data[J]. *International Journal of Geographical Information Science*, 2014, 28(5): 946-963

[4] Ren Huijun, Xu Tao, Li Xiang. Driving Behavior Analysis Based on Trajectory Data Collected with Vehicle-mounted GPS Receivers[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 739-744(任慧君，许涛，李响. 利用车载GPS轨迹数据实现公交车驾驶安全性分析[J]. 武汉大学学报·信息科学版，2014，39(6): 739-744)

[5] Huang L, Li Q, Yue Y. Activity Identification from GPS Trajectories Using Spatial Temporal Poi's' Attractiveness [C]. Proceedings of the ACM Sigspatial International Workshop on Location Based Social Networks, Chicago, USA, 2010

[6] Seaborn C, Attanucci J, Wilson N H M. Using Smart Card Fare Payment Data to Analyze Multi-Modal Public Transport Journeys in London [C]. The 88th Transportation Research Board Annual Meeting, Washington D C, USA, 2009

[7] Li L, Goodchild M F, Xu B. Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr[J]. *Cartography and Geographic Information Science*, 2013, 40(2): 61-77

[8] Tsou M H, Yang J A, Lusher D, et al. Mapping Social Activities and Concepts with Social Media (Twitter) and Web Search Engines (Yahoo and Bing): A Case Study in 2012 US Presidential Election[J]. *Cartography and Geographic Information Science*, 2013, 40(4): 337-348

[9] Signorini A, Segre A M, Polgreen P M. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the US During the Influenza a H1N1 Pandemic [J]. *PloS One*, 2011, 6(5): e19467

[10] Achrekar H, Gandhe A, Lazarus R, et al. Predicting Flu Trends Using Twitter Data[C]. 2011 IEEE Conference on Computer Communications Workshops, Shanghai, China, 2011

[11] Ferrari L, Rosi A, Mamei M, et al. Extracting Urban Patterns from Location-Based Social Networks [C]. The 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, USA, 2011

[12] Frias-Martinez V, Frias-Martinez E. Spectral Clustering for Sensing Urban Land Use Using Twitter Activity[J]. *Engineering Applications of Artificial Intelligence*, 2014, 35: 237-245

[13] Wu L, Zhi Y, Sui Z, et al. Intra-urban Human Mobility and Activity Transition: Evidence from Social Media Check-in Data[J]. *PloS One*, 2014, 9(5), doi:10.1371/journal.pone.0097010

[14] Liu Y, Sui Z, Kang C, et al. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-in Data[J]. *PloS One*, 2014, 9(1), doi:10.1371/journal.pone.0086026

[15] Hanna R, Rohm A, Crittenden V L. We're all Connected: The Power of the Social Media Ecosystem[J]. *Business Horizons*, 2011, 54(3): 265-273

[16] Sakaki T, Okazaki M, Matsuo Y. Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development[J]. *Knowledge and Data Engineering, IEEE Transactions on*, 2013,

25(4): 919-931

[17] Crooks A, Croitoru A, Stefanidis A, et al. Earthquake: Twitter as a Distributed Sensor System[J]. *Transactions in GIS*, 2013, 17(1): 124-147

[18] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *The Journal of Machine Learning Research*, 2003(3): 993-1 022

[19] Cortes C, Vapnik V. Support Vector Machine[J]. *Machine Learning*, 1995, 20(3): 273-297

[20] Nagel A C, Tsou M H, Spitzberg B H, et al. The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets[J]. *Journal of Medical Internet Research*, 2013, 15(10), doi: 10. 2196/jmir. 2705

[21] Cleveland R B, Cleveland W S, McRae J E, et al. STL: A Seasonal-Trend Decomposition Procedure Based on Loess[J]. *Journal of Official Statistics*, 1990, 6(1): 3-73

[22] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, USA, 1996

The Mining and Analysis of Emergency Information in Sudden Events Based on Social Media

WANG Yandong[1]　*LI Hao*[1]　*WANG Teng*[1]　*ZHU Jianqi*[1]

1　State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: Social media has played an important role in disaster emergency responses, which is increasingly being regarded as mobile sensors, perceiving events near human beings. When an emergency occurs, a large number of images and texts with geographic information quickly flood the social network. This paper presents a new method of mining and analysis of emergency information with a case study to analyze the Sina-Weibo text streams during and after the 2012 'Beijing Rainstorm'. The topic classification model of real-time emergency information is built, and the emergency information from real-time text stream are identified and located. Decomposition of seasonal components from the time series data is applied to explore the trend of the number of Sina-Weibo texts related to the 'Beijing rainstorm'. According to different topics, using statistical and spatial analysis, a possible spatial structure for distributing resources in response to emergencies is indicated. The study can help to understand how the emergency events are evolved and what are impacted by the events, which will benefit decision-makers by allowing timely decisions emergencies for effective mitigation efforts and better allocation of resources.

Key words: social media; sudden events; trend analysis; spatial analysis; data mining

First author: WANG Yandong, PhD, professor, specializes in Big Data analysis and calculation. E-mail: ydwang@whu. edu. cn

Foundation support: The National Natural Science Foundation of China, No. 41271399; the China Special Fund for Surveying, Mapping and Geoinformation Research in the Public Interest, No. 201512015; the Specialized Research Fund for the Doctoral Program of Higher Education, No. 20120141110036.