DOI:10.13203/j.whugis20190054 文章编号:1671-8860(2020)05-0691-08

基于社交媒体共词网络的灾情发展态势探测方法

王艳东[1,2,3] 李萌萌[1] 付小康[1] 邵世维[4] 刘 辉[4]

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
2 地球空间信息技术协同创新中心,湖北 武汉,430079
3 东华理工大学测绘工程学院,江西 南昌,330013
4 武汉市自然资源和规划信息中心,湖北 武汉,430014

摘 要:对灾害发生过程中产生的社交媒体数据进行主题演化探测和分析可以反映灾情的发展态势。提出了一种基于共词网络社区演化进行灾情发展态势感知的方法,首先依据词频-逆文档频率方法筛选出与主题相关的关键词汇,基于关键词的共现关系,构建以关键词为节点的社交媒体共词网络,结合模块度最优化思想,对社交媒体共词网络进行主题社区探测;然后在验证主题探测的基础上,基于时间窗口划分,对相邻时间窗口的主题社区进行演化类型判别,进而得到主题社区演化的结果;最后以2012年"7.21北京特大暴雨"灾害事件为例,利用该方法对收集到的含关键词的微博数据进行主题演化分析。实验结果表明,该方法能够很好地反映主题的演化过程,并能进一步揭示灾情的发展态势,帮助应急管理者了解灾害的发展过程,从而辅助管理者在合适的时间采取相应的应急响应措施。

关键词:社交媒体;共词网络;灾情态势;主题挖掘;主题演化

中图分类号:P208 **文献标志码**:A

随着互联网的蓬勃发展,社交媒体应用已成为人们创建和分享各类信息的主要平台。用户自发的含有全球定位系统(global positioning system,GPS)位置的社交媒体数据可以用来分析地理现象,因此又被称为人类传感器[1]。当灾害发生时,通过挖掘社交媒体数据可以检测灾害事件的影响程度[2-4]、分析灾害事件空间分布规律[5-7],这对灾害的探测和应急都具有重要的意义。文献[8]使用Twitter中带有地理位置的数据,以2014年Napa地震为研究案例,结合文档主题生成隐含狄利克雷分布(latent Dirichlet allocation,LDA)和时空分析的模型,得到话题在空间上的热点分布,从而有效地识别地震足迹及发生重大损失区域;文献[9]利用微博数据研究了2012年北京暴雨事件,使用LDA方法提取应急主题,探寻突发事件随时间的发展趋势,并利用空间聚类分析对主题下的微博进行聚类,从而发现突发事件在空间上的分布规律和异常情况,为应急响应提供决策支持;文献[10]使用Twitter中带有地理位置的数据研究台风"海燕",基于用户的多种交互关系构建网络并探测主题社区,进一步对主题社区的Twitter文本进行空间上的聚类,从而发现灾害应急中的紧急事件及其具体位置。

了解灾情的发展阶段和发展态势对灾害应急具有重要意义。社交媒体的高时效性为灾情发展态势的探测提供了新的思路。灾情的发展阶段和发展态势可通过社交媒体中用户的行为[11-13]或者发布的话题变化[14-15]来进行探测。社交媒体中用户的行为和话题的变化可通过网络分析方法得到,文献[15]以微博为节点,根据微博之间共同拥有词汇的关系构建网络识别主题,并进行主题的时序分析,从而识别暴雨的发展阶段;文献[16]基于用户的提及和转发关系构建网络,研究日本地震和海啸发生前后社区的演变规律,分析了突发事件前后用户行为的变化,有助于态势的感知。然而,以文本,即一条微博或推

收稿日期:2019-05-14

项目资助:国家重点研发计划(2016YFB0501403);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)。

第一作者:王艳东,博士,教授,主要从事城市大数据挖掘与城市功能区空间结构研究。ydwang@whu.edu.cn

文为节点的网络，每条文本只存在一个时间点；以用户为节点的网络，用户在不同时间段的重叠度不高，两者在用于反映灾情发展态势上均有一定的局限性。

基于此，本文提出基于共词网络社区演化感知灾情发展态势的方法。首先根据词频-逆文档频率自动地筛选出与灾情相关的关键词汇，构建基于关键词共现关系的社交媒体共词网络，结合模块最优化思想对网络进行主题社区探测。然后进行时间窗口划分，探测出各时间窗口主题社区，并基于微博关键词在不同时间段可重复出现的特性，以社区演化的方式揭示灾害中社交媒体的主题演化过程，探测灾情发展阶段和态势。最后以2012年“7.21北京特大暴雨”灾害事件为例，利用所提出的方法对收集到的微博数据进行主题演化分析。实验结果表明，该方法能够很好地反映主题的演化过程，并能进一步揭示灾情的发展态势。

1 基于共词网络社区主题演化探测

在科技情报领域，共词网络被认为是能够呈现科学认知的结构，即基于文章中关键词对的共现关系构成网络，通过该网络可以表达科学知识领域的结构[17]。本文提出一种利用共词网络来挖掘灾害中的应急主题，探测主题社区的演化，从而感知灾情发展态势的方法。该方法的具体架构如图1所示，包括构建共词网络、主题社区探测和主题社区演化探测。图1中以两个时间段t_1、t_2为例，首先分别将t_1、t_2内的微博表达成多个关键词，如微博1表示为w_1、w_3、w_4、w_6 4个关键词，并根据词汇的共现关系构建共词网络G_1和G_2；然后对G_1、G_2分别进行社区探测，得到对应的社区C_1^1、C_1^2、C_1^3和C_2^1、C_2^2、C_2^3，并识别出各社区对应的主题；最后通过计算t_1、t_2两社区的包含量，判定社区之间的演化关系。

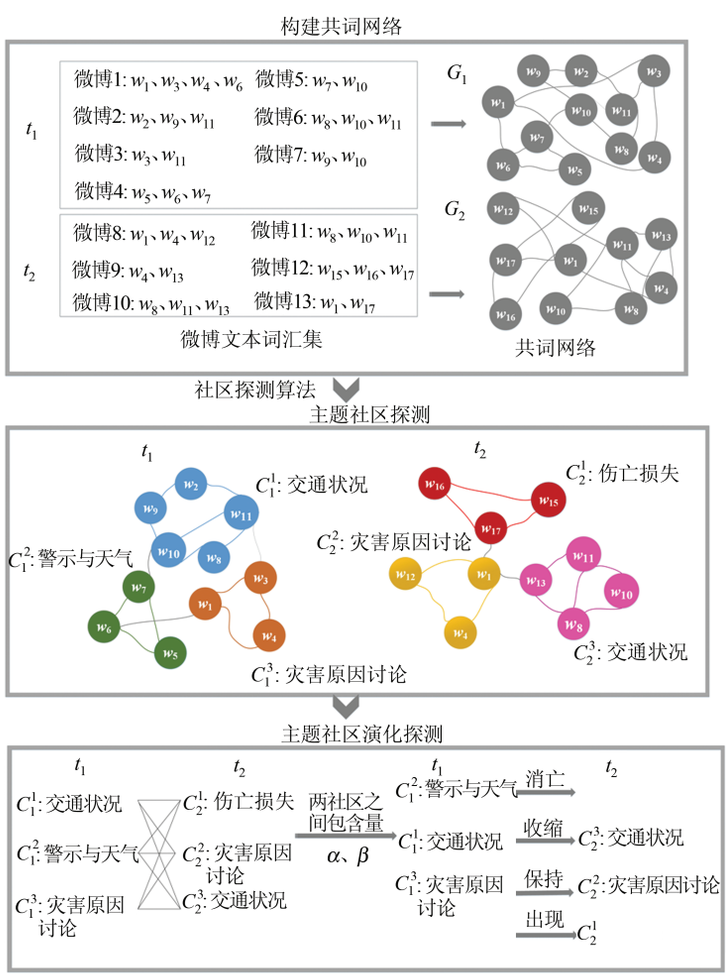


图1 利用共词网络从社交媒体数据中探测灾害应急主题演化示意图

Fig.1 Schematic Diagram of Detecting the Evolution of Disaster Emergency Topics of Social Media Data Based on Co-word Network

1.1　主题关键词筛选

本文利用词频-逆文档频率(term frequency - inverse document frequency,TF-IDF)[18]方法从社交媒体数据中提取主题关键词。如果某个词或短语在一篇文章中出现的频率高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类,以此为依据选取主题关键词。根据所有文档中词汇的TF-IDF分布,使用头尾打断法(head/tail break)[19]自动筛选能够反映主题信息的关键词集合。频率计算如下:

$$f_{i,j}=\frac{n_{i,j}}{\sum n_{r,j}}\times\log\frac{\|D\|}{1+\left\|\{j_{w_i\in d_j}\}\right\|} \tag{1}$$

式中,$n_{i,j}$为词汇w_i在文档d_j中的出现次数;$\sum n_{r,j}$为文档d_j的总词汇数;$\|D\|$为数据集中的文档总数;$\left\|\{j_{w_i\in d_j}\}\right\|$为包含词语$w_i$的文档数目。

1.2　共词网络构建及表示

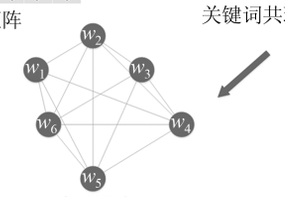
共词网络是指根据共现关系构建的一种网络。首先判断任意两个关键词在某一篇文档中是否共同出现,然后统计这种共现关系在所有文档中出现的次数并构建关键词共现矩阵,依据共现矩阵生成共词网络。本文设计的社交媒体共词网络构建流程如图2所示。图2中d表示文档(本文对应微博),w表示主题关键词。文档-关键词矩阵中,若数字为1,则代表该文档中包含对应的关键词,即该关键词在文档中出现过,如w_3、w_4都在d_1、d_4出现过;关键词共现矩阵的对角线代表的是每个关键词在所有文档中的词频,即有多少个文档包含该关键词;其他位置则表示这对关键词在所有文档中共现的次数,即同时包含这对关键词的文档数,如w_3、w_4这对词在d_1、d_4两篇文档中同时出现,因此共现矩阵中w_3、w_4对应值为2。根据关键词共现矩阵可构建共词网络。本文中网络的节点表示关键词,网络的边表示两个关键词至少在一条微博中存在共现关系。

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
w_1	0	1	1	0	1	0	0	1
w_2	1	0	0	0	1	1	1	0
w_3	1	0	0	1	0	0	0	0
w_4	1	1	0	1	0	1	1	0
w_5	0	1	0	1	1	0	1	0
w_6	1	1	1	1	1	0	0	0

文档-关键词矩阵

	w_1	w_2	w_3	w_4	w_5	w_6
w_1	4	1	0	1	2	3
w_2	1	4	1	2	2	2
w_3	0	1	2	2	1	2
w_4	1	2	2	5	3	3
w_5	2	2	1	3	4	3
w_6	3	2	2	3	3	5

关键词共现矩阵



共词网络

图2　共词网络构建流程

Fig.2　Co-word Network Construction Process

1.3　主题社区探测

一个社区内的词汇联系越紧密,其对应的文本则拥有越相似的主题。因此本文通过对词汇的社区划分,实现对应文本的类别划分,并根据社区中的词汇的主题来解释社区的主题。本文基于Louvain算法[20]研究社交媒体共词网络社区的探测方法。该算法是一种基于模块度最优化思想的算法,模块度的计算如下:

$$Q=\frac{1}{2m}\cdot\sum_{w_1w_2}\left[A_{w_1,w_2}-\frac{k_{w_1}\cdot k_{w_2}}{2m}\right]\cdot\delta(c_{w_1},c_{w_2}) \tag{2}$$

$$\delta(c_{w_1},c_{w_2})=\begin{cases}0, c_{w_1}\neq c_{w_2}\\ 1, c_{w_1}=c_{w_2}\end{cases} \tag{3}$$

式中,m表示网络中边的总数量;A_{w_1,w_2}表示节点(关键词)w_1、w_2之间的连边权重;k_{w_1}、k_{w_2}分别表示所有与关键词节点w_1、w_2相连的边的权重之和;c_{w_1}、c_{w_2}分别表示节点w_1、w_2所属的社区。

Louvain算法通过两个步骤快速优化模块度:(1)把网络中的每个节点(关键词w)当作一个社区,对每个节点w,依次尝试将其分配到每个邻居节点所在的社区,并计算每次分配前与分配后的模块度变化ΔQ,找到模块度变化最大的邻居节点。对所有节点重复该过程,直到所有节点不再变化。(2)处理第一阶段的结果,将一个个小社区归并为一个超节点来重新构建网络。重复执行上面两个步骤,直到整个网络的模块度不再发生变化。ΔQ的计算如下:

$$\Delta Q=\left[\frac{\sum\text{in}+2k_{w,\text{in}}}{2m}-\left(\frac{k_w+\sum\text{tot}}{2m}\right)^2\right]-\left[\frac{\sum\text{in}}{2m}-\left(\frac{\sum\text{tot}}{2m}\right)^2-\left(\frac{k_w}{2m}\right)^2\right] \tag{4}$$

式中,$\sum\text{in}$表示关键词节点w邻居社区内边的权重之和;$\sum\text{tot}$表示关键词节点w邻居社区内所有与其相连的边的权重和;k_w表示关键词节点w的所有边的权重和;$k_{w,\text{in}}$表示社区内所有与关键词节点w相连的所有边的权重和;m表示网络中边的总数量。

1.4　社区演化探测

社区演化是对一系列时间窗口下,相邻时间段网络社区之间变化的探测,可能发生的事件包括保持、收缩、生长、分裂、融合、出现、消亡。由

于同一词汇在不同时间段的不同社区仍然可能重复出现，即以关键词为节点构建的网络社区节点具有可重叠性，本文基于文献[21]中提出的社区演化算法设计了对社交媒体共词网络进行主题社区演化探测的方法。该方法使用包含量来判别社区的演化。取前两个时间段的社区C_1^1和C_2^1，计算C_1^1社区在C_2^1社区中的包含量$I(C_1^1,C_2^1)$，以及C_2^1社区在C_1^1社区中的包含量$I(C_2^1,C_1^1)$。根据$I(C_1^1,C_2^1)$对应阈值α以及$I(C_2^1,C_1^1)$对应阈值β判断社区的演化类型。C_1^1社区在C_2^1社区中的包含量定义为：

$$I(C_1^1,C_2^1)=\frac{\left\|W_{C_1^1}\cap W_{C_2^1}\right\|}{\left\|W_{C_1^1}\right\|}\cdot\frac{\sum\limits_{w\in(W_{C_1^1}\cap W_{C_2^1})}P_{C_1^1}(w)}{\sum\limits_{w\in(W_{C_1^1})}P_{C_1^1}(w)} \tag{5}$$

式中，$W_{C_1^1}$、$W_{C_2^1}$分别为社区C_1^1、C_2^1中的关键词集合；$P_{C_1^1}(w)$表示词汇节点w在主题社区C_1^1中的重要程度，可由节点度、节点中介性[22]等来计算。根据该社区演化算法，单个社区随时间可能发生的演化类型如表1所示。

表1 单个社区随时间演化过程示例

Tab. 1 Example of Single Community Evolution Process over Time

时间	演化类型	社区
t_1	出现	C_1^1
t_2	生长	C_2^1
t_3	分裂	C_3^1,C_3^2
t_4	收缩-保持-出现	C_4^1,C_4^2,C_4^3
t_5	合并	C_5^1
t_6	消亡	

2 北京暴雨灾害演化实验分析

本文以2012年“7.21北京特大暴雨”事件作为研究案例。首先对整个研究时间区间进行主题挖掘和空间分析，验证本文方法在主题挖掘上的可靠性；然后在此基础上，将时间窗口设置为2 h，利用本文方法分析主题的演化过程，揭示灾害的发展阶段和态势。

2.1 实验数据收集与处理

2012-07-21—2012-07-22，北京及其周边地区遭遇了61年来最强暴雨及洪涝灾害，基础设施、居民正常生活均受到重大影响，且有79人遇难。降雨可以分为两个阶段，第一个阶段发生在21日10:00—20:00，主要特点是短期降雨强烈、强度变化幅度明显；第二个阶段发生在21日20:00至22日04:00，降水逐渐平缓，降雨强度显著降低[23]。本实验以“北京暴雨”为关键词，收集2012-07-20 00:00—2012-08-10 24:00的微博数据共706 835条，其中包含GPS位置信息的微博有26 050条。由于微博数据和只含有GPS位置的微博数据均为某种程度的抽样数据，且含有GPS位置的微博数据具有较少的噪声，因此本文主要以含有GPS位置的微博数据来进行实验和案例分析。本实验中，取2012-07-21 06:00—2012-07-24 04:00的带有GPS位置的11 779条微博作为本文的实验数据，该时间区间包含降雨的两个阶段及暴雨后的一段时间。本实验的时间区间内每小时的微博分布曲线如图3所示。

实验首先对原始微博集合进行去重、中文分词、去停用词并保留表情符号等预处理，得到每一条微博的词汇表达，从而得到微博对应的词汇集合。然后根据所有微博词汇的TF-IDF值的分布，进行两次头尾打断后，选取头部的共计815个词汇作为实验中的主题关键词，使用主题关键词汇以及获得的微博词汇集合构建关键词的共现关系矩阵后，将关键词对作为两个节点、关键词对共现的微博文本数作为该对节点边的权重，构建共词网络。最后使用Louvain算法对网络进行两次社区探测，共得到17个社区。

2.2 主题挖掘结果评估

为了评估本文方法在主题挖掘上的有效性，随机抽取550条微博，通过人工标注与灾害应急相关的类别形成验证集。在灾害应急信息分类的研究中，文献[24]提出了有利于提高灾害态势感知的14类应急信息主题类别，在此基础上，根据暴雨突发事件的特点，本文通过筛选和合并，确认挖掘5类灾害应急信息：救援、伤亡损失、警示与天气、交通状况、灾害原因讨论。

本文基于纯度计算[10]，使用准确率和召回率两个参数来评价主题分类的结果，具体步骤如下：

1)确定各社区主题。计算各个社区内各主题的样本数，找到样本数最多的主题，对应的样本数记为N_{true}^c，计算纯度，并根据纯度阈值来确定各社区的主题。纯度定义如下：

$$\text{Purity}=\frac{N_{\text{true}}^c}{N^c} \tag{6}$$

式中，N^c表示一个社区中总的样本个数。若该社区纯度大于设定的阈值，则认为该社区主题为最大样本数对应的主题；若纯度小于设定阈值，则认为该社区没有被识别出主题。

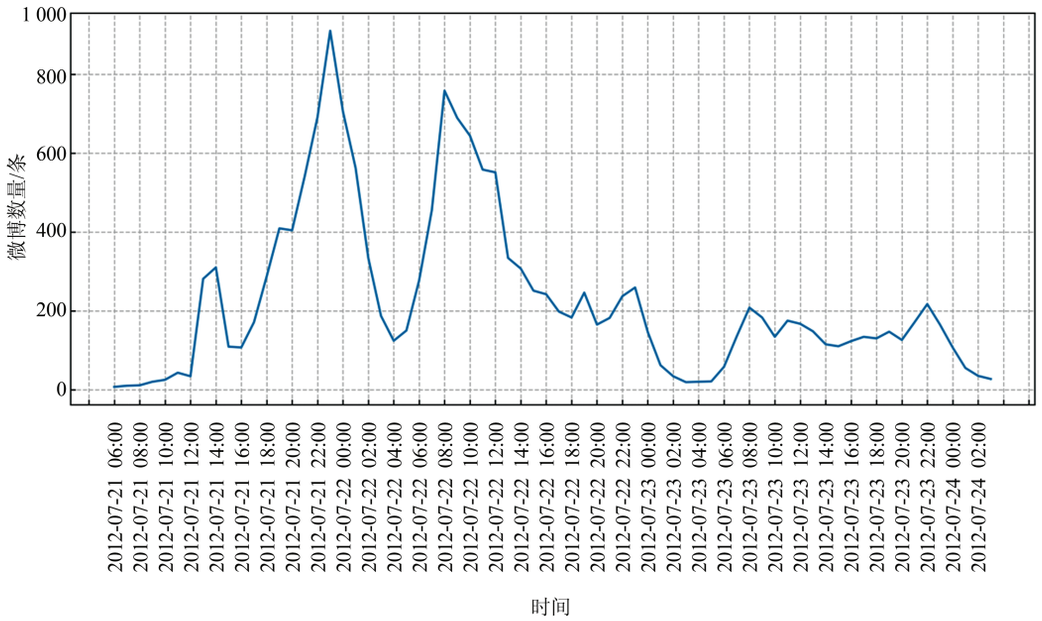


图 3　微博分布曲线

Fig.3　Weibo Distribution Curve

2)计算准确率和召回率。召回率是模型分类正确的样本数与验证集中样本总数的比值:

$$\text{recall}=\frac{\text{relevant\_retrieved\_samples}}{\text{authoritative\_samples}} \quad (7)$$

式中,relevant_retrieved_samples表示模型分类正确的样本数,即识别出的主题与验证集中主题一致的样本数;authoritative_samples为验证集的样本总数。

准确率表示方法分类正确的样本与分类识别到主题的样本总数的比值:

$$\text{precision}=\frac{\text{relevant\_retrieved\_samples}}{\text{discovered\_samples}} \quad (8)$$

式中,discovered_samples表示分类识别出主题的样本总数,即所有主题社区样本总数。

针对§2.1中得到的17个社区,设置不同纯度阈值,分别计算不同纯度下的准确率和召回率,结果如表2所示。由表2可知,当纯度为0.7时,灾害应急主题挖掘的准确率可以达到80%以上;而当纯度为0.8时,灾害应急主题挖掘的准确率可达到90%。因此可以说明本文提出的方法在主题挖掘上具有较好的准确率结果。为了在得到较高准确率的同时保证被识别出主题的微博数量尽可能多,本文选择纯度为0.7时得到的主题挖掘结果进行后续的分析。

2.3　主题空间分布模式分析

分析主题的空间分布情况可以发现有效的灾害应急信息,有助于相关部门了解灾害事件中的灾情空间分布规律。本文选用的数据均为带有GPS位置的微博数据,每条微博可看成一个点状的地理实体。本文以主题交通状况为例,研究了该主题下的微博在空间上的分布模式,该主题下的微博共有1 473条。通过核密度分析,得到该主题下微博的空间分布如图4所示。由图4可以看出,在区域1、2、3,即北京首都国际机场、北京西站及北京站出现了热点区域,这是由于北京暴雨造成火车、飞机的延误、晚点、取消等,导致很多乘客滞留在这些地方。除北京外,在国内很多地方甚至国外也有一些微博的分布聚集情况,这是因为其他地方到北京的航线或火车也受到了北京暴雨的影响。

表 2　基于共词网络的社交媒体应急信息挖掘方法的准确率和召回率

Tab. 2　Accuracy and Recall Rate of Social Media Emergency Information Mining Methods Based on Co-word Network

纯度	准确率	召回率
0.3	0.63	0.42
0.4	0.63	0.42
0.5	0.74	0.32
0.6	0.74	0.32
0.7	0.82	0.23
0.8	0.90	0.11

2.4　主题演化探测

本文通过微博中灾害相关主题的演化过程来反映灾害的发展阶段。主题演化过程是通过

基于时间窗口的社区演化方法进行探测。目前已有以1 h为时间间隔对微博进行灾害主题时序变化的研究[9,15],以此为基础,根据本文实验数据的情况,为了保证每个时间窗口内有足够的微博数目、分析结果具有统计意义,又能以较细时间粒度反映主题演化过程,本文将时间窗口设置为2 h,将2012-07-21 06:00—2012-07-24 04:00这段时间划分为35个子时间区间。对各子时间区间内的微博数据分别进行共词网络构建、社区探测、主题发现及主题演化探测,得到不同时间段下的社区主题及社区演化事件。

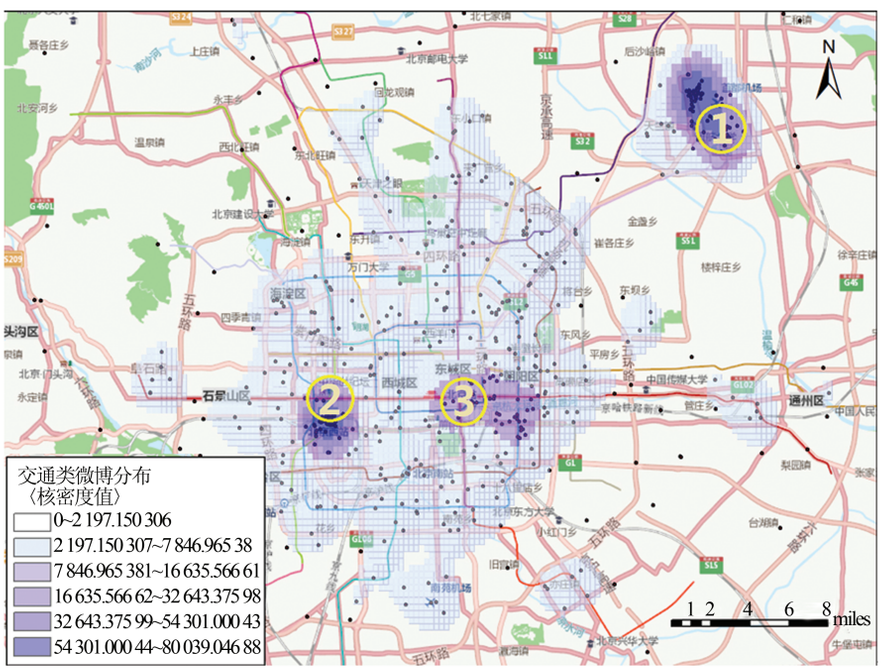


图4 交通状况主题空间分布

Fig.4 Spatial Distribution of Traffic Condition

根据暴雨的降雨过程,本文选取了4个时间片段,使用桑基图对主题演化过程进行展示。如图5所示,每个长方形为一个社区,长方形之间的连接表示社区间节点即主题关键词的流动,长方形的高度代表该社区与其他社区共有的词汇数量。图5(a)中的时间片段主要在暴雨发生前及暴雨初期,可以看出主要的主题为警示与天气及交通信息。警示与天气最早出现在暴雨发生前,即2012-07-21 06:00—08:00,在这之后该主题社区发生了一系列分裂、合并、生长等事件,出现越来越多天气预警主题的社区;而在08:00—10:00,交通信息主题社区也开始出现,10:00后开始降雨,交通信息类主题社区不断生长,这反映了降雨对交通的直接影响。图5(b)中的时间片段是暴雨的第一阶段的结尾及第二阶段的开始,此时警示与天气仍然是热点话题,直到2012-07-21 20:00,警示与天气主题社区一部分分裂出两个灾害原因讨论主题社区。图5(c)为暴雨灾害后期,暴雨已停止,灾害原因讨论主题社区生长,并在2012-07-23 08:00后逐渐分裂出伤亡损失主题的社区,之后伤亡损失主题继续生长。图5(d)为暴雨结束的第二天,即2012-07-24 00:00后,交通主题社区保持原状,灾害原因讨论主题社区生长且主题变为伤亡损失,随后出现新的伤亡损失主题社区以及救援信息,且都随时间发生生长、分裂、融合等事件。

通过以上分析可以看出,通过本文所提出的方法可以得到各个灾害主题的演变过程,不同的灾害主题会出现在不同的暴雨阶段中。因此,可利用不同灾害主题的演变过程反映和探测暴雨的发展阶段和态势,从而对应急管理者制定应急措施提供参考。

3 结 语

社交媒体被人们广泛使用,其中充斥着包含时间、空间及语义信息的社交媒体数据,通过挖掘社交媒体数据,可以探测出灾情的发展阶段及发展态势,从而为相关应急机构提供重要的决策参考。本文探讨了基于社交媒体共词网络探测灾情发展态势的一种新思路,利用共词网络从

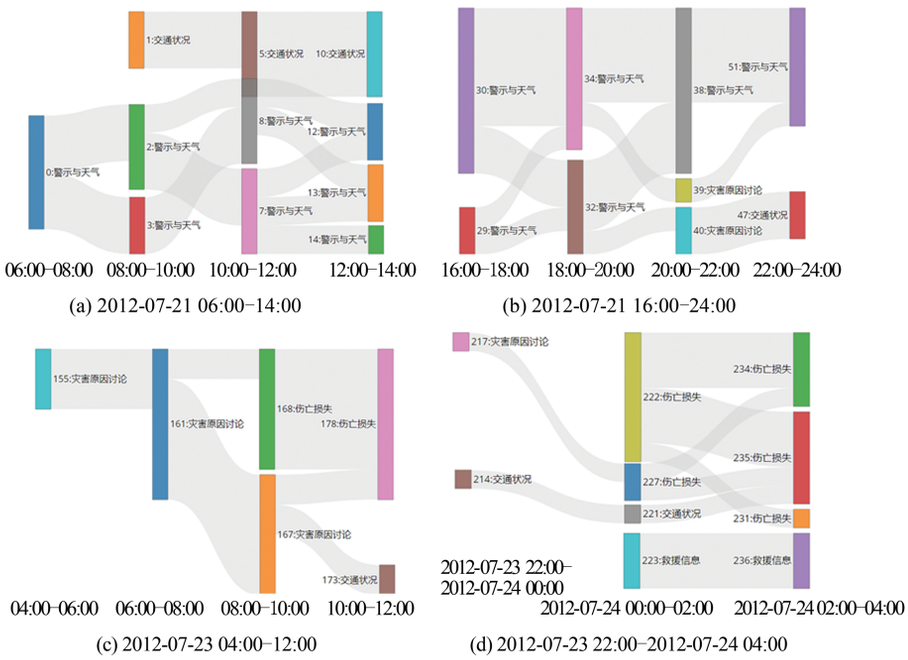


图 5　4个子时间区间的主题社区演化桑基图

Fig.5　Sankey Diagrams of Topic Community Evolution in Four Sub-time Intervals

社交媒体文本数据中挖掘主题社区,并基于微博关键词在不同时间段可重叠的特性,以社区演化的方式,探测主题社区的演化过程,继而探测灾情发展阶段和态势。将该方法应用于2012年北京暴雨事件,结果证明了该方法在应急主题挖掘上的有效性。通过划分时间区间,得到本文关注的5类应急主题随时间的演化过程,通过该演化过程能够探索人们在灾害不同时期的关注热点,并进一步探测灾情的发展态势,从而帮助应急管理者了解灾害的发展过程,辅助管理者在合适的时间采取相应的应急措施。

参　考　文　献

[1] Sakaki T, Okazaki M, Matsuo Y. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors[C]. The 19th International Conference on World Wide Web, New York, USA, 2010

[2] Power R, Robinson B, Moseley A. Comparing Felt Reports and Tweets About Earthquakes[C]. International Conference on Information and Communication Technologies for Disaster Management, Vienna, Austria, 2016

[3] Wang Yandong, Ruan Shisi, Wang Teng, et al. Rapid Estimation of an Earthquake Impact Area Using a Spatial Logistic Growth Model Based on Social Media Data[J]. *International Journal of Digital Earth*, 2019, 12(11): 1-20

[4] Bai Hua, Lin Xunguo. Sina Weibo Disaster Information Detection Based on Chinese Short Text Classification[J]. *Journal of Catastrophology*, 2016, 31(2): 19-23 (白华, 林勋国. 基于中文短文本分类的社交媒体灾害事件检测系统研究[J]. 灾害学, 2016, 31(2): 19-23)

[5] Kryvasheyeu Y, Chen H, Obradovich N, et al. Rapid Assessment of Disaster Damage Using Social Media Activity[J]. *Science Advances*, 2016, 2(3): e1500779

[6] Arthur R, Boulton C A, Shotton H, et al. Social Sensing of Floods in the UK[J]. *Plos One*, 2018, 13(1): e0189327

[7] Liang Chunyang, Lin Guangfa, Zhang Mingfeng, et al. Accessing the Effectiveness of Social Media Data in Mapping the Distribution of Typhoon Disasters [J]. *Journal of Geo-information Science*, 2018, 20(6): 807-816 (梁春阳, 林广发, 张明锋. 社交媒体数据对反映台风灾害时空分布的有效性研究[J]. 地球信息科学学报, 2018, 20(6): 807-816)

[8] Resch B, Florian U, Havas C. Combining Machine-Learning Topic Models and Spatiotemporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment [J]. *Cartography and Geographic Information Science*, 2018, 45(4): 362-376

[9] Wang Yandong, Li Hao, Wang Teng, et al. The Mining and Analysis of Emergency Information in Sudden Events Based on Social Media[J]. *Geomatic and Information Science of Wuhan University*, 2016, 41(3): 290-297 (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)

[10] Bakillah M, Li R Y, Liang S H L. Geo-Located

Community Detection in Twitter with Enhanced Fast-Greedy Optimization of Modularity: The Case Study of Typhoon Haiyan[J]. *International Journal of Geographical Information Systems*, 2015, 29(2): 258-279

[11] Mukkamala A, Beck R. Social Media for Disaster Situations: Methods, Opportunities and Challenges [C]. 2017 IEEE Global Humanitarian Technology Conference, San Jose, USA, 2017

[12] Li L, Zhang Q, Tian J, et al. Characterizing Information Propagation Patterns in Emergencies: A Case Study with Yiliang Earthquake [J]. *International Journal of Information Management*, 2018, 38(1): 34-41

[13] Kim J, Hastak M. Social Network Analysis: Characteristics of Online Social Networks After a Disaster [J]. *International Journal of Information Management*, 2018, 38(1): 86-96

[14] David C C, Ong J C, Legara E F T. Tweeting Supertyphoon Haiyan: Evolving Functions of Twitter During and After a Disaster Event [J]. *Plos One*, 2016, 11(3): e0150190

[15] Wang Yandong, Fu Xiaokang, Li Mengmeng. A New Social Media Topic Mining Method Based on Co-word Network [J]. *Geomatic and Information Science of Wuhan University*, 2018, 43(12): 2 287-2 294(王艳东,付小康,李萌萌. 一种基于共词网络的社交媒体数据主题挖掘方法[J]. 武汉大学学报·信息科学版, 2018, 43(12): 2 287-2 294)

[16] Lu X, Brelsford C. Network Structure and Community Evolution on Twitter: Human Behavior Change in Response to the 2011 Japanese Earthquake and Tsunami[J]. *Scientific Reports*, 2015, 4(1): 6 773

[17] Wang Xiaoguang. Formation and Evolution of Science Knowledge Network(Ⅰ): A New Research Method Based on Co-word Network[J]. *Journal of the China Society for Scientific and Technical Information*, 2009, 28(4): 599-605(王晓光. 科学知识网络的形成与演化(Ⅰ):共词网络方法的提出[J]. 情报学报, 2009, 28(4): 599-605)

[18] Lu Yonghe, Li Yanfeng. Improvement of Text Feature Weighting Method Based on TF-IDF Algorithm [J]. *Library and Information Service*, 2013, 57(3): 90-95(路永和,李焰锋. 改进TF-IDF算法的文本特征项权值计算方法[J]. 图书情报工作, 2013, 57(3): 90-95)

[19] Jiang B. Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-Tailed Distribution [J]. *The Professional Geographer*, 2013, 65(3): 482-494

[20] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast Unfolding of Communities in Large Networks [J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 2 008(10): P10008

[21] Brodka P, Saganowski S, Kazienko P. Group Evolution Discovery in Social Networks [C]. International Conference on Advances in Social Networks Analysis and Mining, Kaohsiung, Taiwan, 2011

[22] Musial K, Kazienko P, Brodka P, et al. User Position Measures in Social Networks [C]. Social Network Mining and Analysis, Paris, France, 2009

[23] Sun Jisong, He Na, Wang Guorong, et al. Preliminary Analysis on Synoptic Configuration Evolvement and Mechanism of a Torrential Rain Occurring in Beijing on 21st July 2012[J]. *Torrential Rain and Disasters*, 2012, 31(3): 218-225(孙继松,何娜,王国荣,等. "7.21"北京大暴雨系统的结构演化特征及成因初探[J]. 暴雨灾害, 2012, 31(3): 218-225)

[24] Vieweg S. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness [C]. Sigchi Conference on Human Factors in Computing Systems, Atlanta, USA, 2010

A New Method to Detect the Development Situation of Disasters Based on Social Media Co-word Network

WANG Yandong[1,2,3] *LI Mengmeng*[1] *FU Xiaokang*[1] *SHAO Shiwei*[4] *LIU Hui*[4]

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

2 Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

3 Faculty of Geomatics, East China University of Technology, Nanchang 330013, China

4 Wuhan Natural Resources and Planning Information Center, Wuhan 430014, China

Abstract: The development trend of disasters can be perceived through mining and analyzing the topics

(下转第735页)

corresponding relationship of semantic features among entities, the concept of attribute feature entropy is introduced to calculate the weight values of different features, and then the overall semantic similarity of geographical entities is measured by synthesizing the multi-feature similarity. Finally, the model is applied to road entity matching. The road matching is realized by calculating the semantic similarity between entities. Meanwhile, the validity of the model is verified. The experimental results show that the semantic similarity measurement model based on multi-feature constraints can reasonably calculate the semantic similarity of geographical entities and improve the efficiency of semantic matching of geographical entities.

Key words: geographical entities; semantic similarity; attribute feature entropy; road matching

First author: ZHAO Yunpeng, PhD candidate, majors in multi-source vector spatial data fusion and mapping. E-mail:dpyk_zyp@163.com

Corresponding author: SUN Qun, PhD, professor. E-mail: sunqun@371.net

Foundation support: The National Natural Science Foundation of China(41571399,41801313).

引文格式:ZHAO Yunpeng, SUN Qun, LIU Xingui, et al. Geographical Entity-Oriented Semantic Similarity Measurement Method and Its Application in Road Matching[J]. Geomatics and Information Science of Wuhan University, 2020, 45(5): 728-735. DOI: 10.13203/j.whugis20190039(赵云鹏,孙群,刘新贵,等.面向地理实体的语义相似性度量方法及其在道路匹配中的应用[J].武汉大学学报·信息科学版,2020,45(5):728-735.DOI:10.13203/j.whugis20190039)

(上接第 698 页)

evolution of social media data in disasters. A method of studying the evolution of the topic communities based on the common word network is proposed, so the development trend of the disaster situation can be sensed. Firstly, according to the word frequency-inverse document frequency analysis, the key words related to the topics are selected and a social media co-word network with keywords as nodes is constructed. Thus, the topic community detection is performed on the social media co-word network based on the module optimization. Secondly, on the basis of verifying the topic detection, and the time window division, the evolution types of the topic communities in adjacent time windows are distinguished, and then the result of the topic community evolution is obtained. Finally, taking the 7.21 Beijing Heavy Rainstorm disaster event in 2012 as an example, the proposed method is used to analyze the collected microblog data. The experiment shows that the method can reflect the evolution process of the topics well. It can further reveal the development trend of the disaster, and help emergency managers understand the development process of disasters, so as to assist managers to take appropriate emergency response measures in appropriate time.

Key words: social media; co-word network; disaster situation; topic mining; topic evolution

First author: WANG Yandong, PhD, professor, specializes in the theories and methods of spatial data mining and urban spatial structure. E-mail: ydwang@whu.edu.cn

Foundation support: The National Key Research and Development Program of China(2016YFB0501403); the National Natural Science Foundation of China(41271399); China Special Fund for Surveying, Mapping and Geo-Information Research in the Public Interest (201512015).

引文格式:WANG Yandong, LI Mengmeng, FU Xiaokang, et al. A New Method to Detect the Development Situation of Disasters Based on Social Media Co-word Network[J]. Geomatics and Information Science of Wuhan University, 2020, 45(5): 691-699. DOI: 10.13203/j.whugis20190054(王艳东,李萌萌,付小康,等.基于社交媒体共词网络的灾情发展态势探测方法[J].武汉大学学报·信息科学版,2020,45(5):728-735.DOI:10.13203/j.whugis20190054)