DOI:10.13203/j.whugis20180225　　文章编号:1671-8860(2018)12-2287-08

一种基于共词网络的社交媒体数据主题挖掘方法

王艳东[1,2,3]　付小康[1]　李萌萌[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
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

摘　要:对社交媒体所包含文本数据的深入挖掘,有利于有效地进行后续的时空分析。提出了一种新的基于共词网络的社交媒体数据主题挖掘方法,依据词频-逆文档频率分析,自动筛选出与主题相关的关键词汇,基于微博间是否包含相同的关键词汇,提出构建以微博为节点的共词网络,并结合 Louvain 社区探测算法进行文本主题挖掘。所提出的方法是一种无监督方法,且具有不需要指定聚类数目的优点。实验表明,该方法在主题挖掘表现上,准确率和召回率均优于常用的文档主题生成模型。以收集的 2012 年北京暴雨期间包含关键词的微博为例,利用提出的方法对微博数据集进行挖掘和时空分析,结果表明所提方法在实际应用中的有效性。

关键词:共词网络;社交媒体;Louvain 社区探测;主题挖掘

中图分类号:P208　　**文献标志码**:A

社交媒体是一种重要的时空数据源,又被称为"社会传感器[1]",它们在揭示人类移动和活动规律[2-3]、探索社会经济模式[4]、流感预测[5-7]、灾害探测[1,8-9]、选举结果预测[10-11]和发现城市格局[12]等方面得到了广泛的研究。对社交媒体所包含文本数据的主题挖掘,有利于有效地进行后续的时空分析。如文献[13]通过研究 Twitter 中的信息内容,预测了某种疾病的发生时间和发生地点;文献[14]通过人工分析,将飓风桑迪发生时所产生的 Tweets 划分成不同主题,并分析了它们的时序分布以了解灾害所处的阶段,进而辅助应急管理人员在灾害管理中采取有效的行动。这些研究表明,主题挖掘是利用社交媒体有效地进行地理现象研究的重要手段。

目前,社交媒体的主题挖掘多采用人工分析和机器学习的方法。然而,这些方法通常需要大量的人工分析和标注工作,无法快速高效地完成主题挖掘任务。大量科研工作者尝试通过无监督的主题模型来自动地完成社交媒体数据的主题挖掘任务。其中,最为主要的是使用文档主题生成(latent dirichlet allocation,LDA)模型以及基于 LDA 改进的无监督模型从社交媒体文本数据中挖掘主题[15-16]。如文献[15]以暴雨灾害事件为例,结合 LDA 模型和支持向量机算法,提出了基于社交媒体文本数据的灾害应急信息分类框架。但文献[17]指出,LDA 模型并不适合微博短文本的情形。文献[18]则讨论了两种可能的 LDA 模型改进方式,并指出了使用 LDA 模型对 Twitter 数据进行分类的挑战。

本文提出了一种新的基于共词网络的社交媒体数据主题挖掘方法。该方法根据词频-逆文档频率分析,自动筛选出与主题相关的关键词汇;基于微博间是否包含相同的关键词汇,提出了构建以微博为节点的微博共词网络;并结合 Louvain 社区探测算法进行文本主题分类和挖掘。本文所提出的方法是一种无监督方法,且具有不需要指定聚类数目的优点,具有很好的实用性。以 2012 年北京暴雨期间包含关键词的微博数据为例,实验表明所提出的方法在主题挖掘表现上,准确率和召回率均优于常用的 LDA 模型。利用本文提出的方法对北京暴雨数据集进行挖掘和时空分析,结果表明了该方法在实际应用中的有效性。

1　基于共词网络的主题挖掘方法

共词网络方法主要应用于文献计量学领

收稿日期:2018-10-15

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。ydwang@whu.edu.cn

域[19]，通常根据文献集中关键词在同一篇文献中共同出现的次数，构建以关键词为节点的共词网络，来分析该文献集所研究的各主题之间的关系。本文基于共词网络方法的思想，提出了一种新的社交媒体数据主题挖掘方法，如图1所示，该方法包含如下3个层次：

1)微博文本集表示层次。在该层次，原始微博文本通过分词后变成“微博文本-微博词汇-主题关键词”的形式。其中，微博词汇通过分词得到，主题关键词则是微博词汇中能反映主题信息的关键词汇(如图1标注红色的词汇)。

2)基于共词的网络表示层次。在该层次，微博文本集通过微博间是否包含相同关键词形成了“网络-节点、边”的形式。其中，节点对应微博文本集表示层次的微博文本，边则表示微博间包含了相同的主题关键词。

3)社区和主题表示层次。在该层次，构建的微博共词网络被表示成了“网络-社区、主题-节点、边”的形式。通常包含相同词汇越多的微博其主题也越相似，因此，可通过探测网络中紧密连接的团体来实现社区的发现，并通过社区内包含的词汇来解释社区的主题。

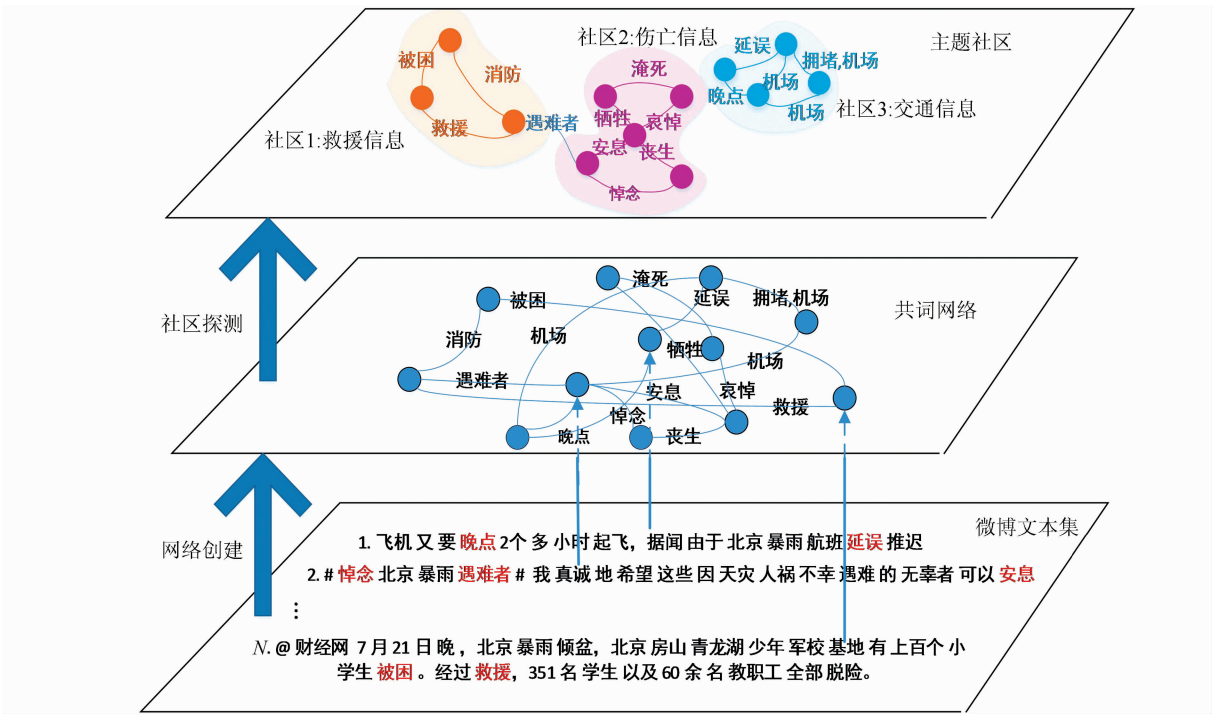

图1 基于共词网络的社交媒体数据主题挖掘示意图

Fig. 1 Topic Mining Scheme of Social Media Messages Based on Co-word Network

1.1 微博文本集表示

本文提出将原始微博文本表示为“微博文本-微博词汇-主题关键词”的微博文本集形式，该过程包含两个步骤：首先获取原始的微博文本，并对微博文本进行分词；然后在划分的词汇中筛选出能够反映微博主题信息的关键词。

微博文本集表示的关键在于如何筛选出主题关键词。本文使用词频-逆文档频率(term frequency-inverse document frequency，TF-IDF)[20]来度量一个词汇在文档中的重要程度，TF-IDF可按式(1)计算：

$$f_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\lg\frac{D}{j:t_i\in d_j} \tag{1}$$

式中，$n_{i,j}$为该词在一条文档(即微博d_j)中出现的次数；$\sum_k n_{k,j}$则为在一条微博d_j中所有词汇出现的次数之和；D表示数据集中的微博总数；$j:t_i\in d_j$表示包含词语t_i的微博数目。

本文结合分析文档中词汇的TF-IDF分布，提出使用头尾打断法(head/tail breaks)[21]自动筛选出能够反映主题信息的关键词集合。文献[22]表明文档中词频的分布多符合重尾分布，重尾分布是自然界中一种常见的分布，可以使用头尾打断法对其进行分类。重尾打断分类法将所有数据依据算术平均值分为两部分，选择头部(高于平均值的部分)并继续迭代过程，直到头部的数据不再是重尾分布。

1.2 基于共词的网络表示

基于共词关系，本文提出将微博文本集表示成“网络-节点、边”的网络形式，其中节点表示微

博文本,边表示微博间包含了相同的主题关键词。其网络表示可通过表 1 的定义描述。

表 1　微博共词网络中的概念定义

Tab.1　Conceptions Table of the Microblog Co-word Networks

序号	定义	解释
1	u,v	表示网络的节点,即微博
2	$U=\{i\},V=\{j\}$	表示微博 u,v 所包含的词汇集合,其中 i、j 表示词汇
3	$E(u,v)$	表示网络中连接节点 u,v 之间的边。当 $U\cap V\neq\varnothing$,且 $U\cap V\subseteq T$,$E(u,v)$ 存在
4	$T=\{t\}$	表示能够代表微博主题信息的关键词集合,其中 t 表示主题关键词
5	$N^{u,v}_{\text{co-words}}$	表示微博 u,v 中相同关键词出现的总次数
6	$N^{u,v}_{\text{total-words}}$	表示微博 u,v 所包含所有词汇的个数
7	$W_{E(u,v)}$	表示网络中边 $E(u,v)$ 的权重,$W_{E(u,v)}=\frac{N^{u,v}_{\text{co-words}}}{N^{u,v}_{\text{total-words}}}$

1.3　社区探测与主题挖掘

在基于共词的网络表示中,连接相对紧密的微博团体,通常包含相同关键词越多,其主题也越相似。基于此,本文提出将基于共词的网络表示成“网络-社区、主题-节点、边”的形式,通过探测网络中紧密连接的团体(即社区)来实现对微博的自动分类,并通过社区内包含的词汇来解释社区的主题。

为了使包含相同关键词的微博划分在一起,形成单一主题的社区,需要使划分的社区内部包含相同关键词的数目尽可能多,且使社区之间包含相同关键词的数目尽可能少。本文使用模块度来刻画这种社区的紧密程度,应用 LM(Louvain method)[23] 快速优化模块度以实现基于共词网络的社区探测。模块度可表示为:

$$Q=\frac{1}{2m}\sum_{uv}\left[W_{E(u,v)}-\frac{k_u k_v}{2m}\right]\delta_{uv} \tag{2}$$

$$\delta_{uv}=\begin{cases}0, c_u\neq c_v\\ 1, c_u=c_v\end{cases} \tag{3}$$

式中,$W_{E(u,v)}$ 表示节点(即微博 u 和 v)之间边的权重;k_u 和 k_v 分别表示所有与微博节点 u 和 v 相连的边的权重之和;$2m$ 表示网络中所有边的权重之和;c_u 和 c_v 分别表示微博节点 u 和 v 所属的社区;δ_{uv} 用来判断微博节点 u 和 v 是否属于同一个社区。LM 通过两个步骤快速优化模块度。首先,把网络中的每个节点(即微博 u)当作一个社区,对每个微博节点 u,依次尝试把它分配到它的每个邻居节点所在的社区,计算分配前与分配后的模块度变化 ΔQ:

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

式中,$\sum\text{in}$ 表示微博节点 u 的一个邻居社区内边的权重之和;$\sum\text{tot}$ 表示微博节点 u 的一个邻居社区内所有与其相连的边的权重之和;k_u 表示微博节点 u 的加权度;$k_{u,\text{in}}$ 表示社区内与微博节点 u 相连的所有边的权重之和;$2m$ 表示基于共词的网络中所有边的权重之和。记录 ΔQ 最大的邻居微博节点,如果 $\max\Delta Q>0$,则把微博节点 u 分配到 ΔQ 最大的邻居微博节点所在的社区,否则保持不变。把这个过程重复地应用到所有的微博节点,直到不会发生模块度的增加。然后,对第一个步骤得到的结果进行处理形成一个新的网络,将所有在同一个社区的节点压缩成一个新节点,社区内节点之间的边的权重转化为新节点的环的权重,社区间的边权重转化为新节点间的边权重。重复执行这两个步骤,直到整个网络的模块度不再发生变化。

2　案例研究与方法评估

以 2012 年“7·21 北京特大暴雨”事件作为研究案例,结合新浪微博 API 和网络爬虫的方式,收集了从 2012 年 7 月 21 日 6 时到 7 月 24 日 4 时,包含“北京暴雨”关键词的 389 168 条微博,其中包含 GPS(Global Positioning System)位置信息的微博有 16 759 条。

本文以包含 GPS 位置信息的微博作为实验数据,使用提出的模型进行主题挖掘,并和常用的 LDA 模型得到的结果比较进行评估。

2.1　基于共词网络的主题挖掘

对微博的文本使用 Jieba 分词工具进行分词(把常用的词汇加入到停用词),对得到的每一个词通过式(1)计算词汇的 TF-IDF 值。微博中词汇的 TF-IDF 值分布如图 2 所示,可以看出微博中大部分词汇的 TF-IDF 值都比较小,只有小部分词汇的 TF-IDF 值比较大,即微博中词汇的 TF-IDF 分布符合重尾分布。经过两次头尾打

断，发现头部不再符合幂律分布，选取最后一次头部的词汇作为主题关键词，共得到654个主题关键词。根据得到的主题关键词，利用共词关系构建出以微博为节点的共词网络，其包含14 949个节点，5 258 956条边。

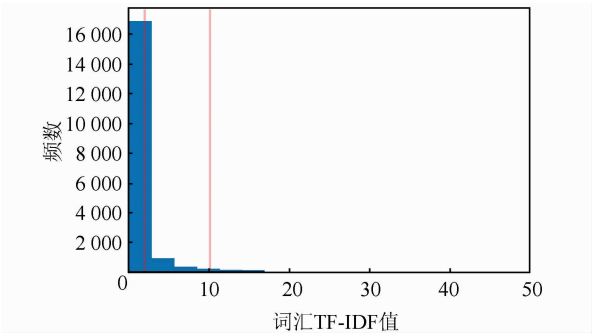

图2 微博词汇的TF-IDF分布

Fig. 2 Distribution of Words' TF-IDF in Microblogs

对构建的共词网络进行社区探测，得到了7个社区。由于部分社区内包含的微博主题并不唯一，本文在探测到的社区基础上，利用相同的方法对各个社区进行子社区探测，共得到48个子社区。为了解释说明各个社区的主题含义，计算出各个社区中关键词的频率分布，并将其可视化为词云图。图3表示其中8个社区的词云图，词汇的大小代表该词汇出现频率的高低。可以看出，45号社区中的微博存在大量与“伤亡”主题有关的讨论，27号社区则主要和“天气预警”主题有关，15号社区表示“提醒朋友”主题，47号社区则表示“正能量和祈祷”主题，6号社区表示“灾害原因讨论”主题，8号社区则表示“交通状况”主题。

图3 社区中关键词词云图示例

Fig. 3 Examples of Keyword Clouds in Communities

2.2 LDA主题挖掘

使用LDA模型进行主题挖掘需要指定聚类主题的个数，本文采用困惑度[24]来优化该参数。通常困惑度越小，说明聚类质量越好。图4显示了不同主题数目情况下的困惑度，可以看出，随着主题数目的增加，困惑度减小，并在主题数为40时开始收敛。为了避免主题数过多，本文最终将LDA模型的聚类主题个数设置为40。

由于LDA模型所得到的结果是微博隶属于各个主题类别的概率，为了在验证集上计算准确率和召回率并与基于共词网络的模型进行比较，将微博所属概率最大的主题类别作为该微博的主题类别。

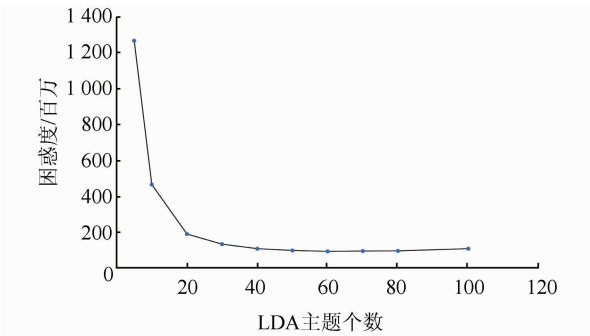

图4 LDA不同聚类个数下的困惑度

Fig. 4 Perplexity of Different Cluster Numbers of LDA Model

2.3 方法评估

基于共词网络的主题挖掘模型和常用的

LDA 主题模型均属于无监督聚类方法,为了对比两种聚类方法,本文构建了验证集,并结合纯度[25]的方法计算和对比两种模型的准确率和召回率,步骤如下:

1)构建验证集。通过随机抽取和人工标注构建了 600 条微博验证集,其中包含了 7 类与暴雨灾害应急相关的主题,分别为祈祷、救援、伤亡、天气预警、提醒朋友、交通状况和灾害原因讨论。

2)计算聚类簇内各个主题的样本数。找到样本数最多的主题,其对应的样本数记为 N^c_{true}。

3)计算聚类簇的纯度。根据聚类簇内各个主题的样本数,计算该聚类簇的纯度:

$$P^c = \frac{N^c_{\text{true}}}{N^c} \tag{5}$$

式中,N^c 表示该聚类簇中总的微博个数。

4)识别聚类簇的主题含义。当聚类簇的纯度大于给定阈值,则聚类簇所代表的主题为簇内样本数最多的主题,否则该聚类簇被标记为不代表任何主题的聚类簇。

5)计算准确率和召回率。召回率、准确率可依次定义为:

$$R = \frac{|\ TP\ |}{|\ AC\ |} \tag{6}$$

$$P = \frac{|\ TP\ |}{|\ DC\ |} \tag{7}$$

式中,TP 表示模型分类正确的样本数;AC 表示验证集中所有类别的样本数;DC 表示模型分类能识别到主题的样本数。

通过计算得到本文所提出的方法和 LDA 方法在不同纯度阈值情况下的准确率和召回率,如表 2 所示。可以看到,在相同的纯度下,使用基于共词网络的社交媒体主题挖掘方法的准确率和召回率均高于传统的 LDA 方法,且在纯度为 0.5 的情况下,准确率达到了 0.8 以上,说明本文所提出的方法可以达到比较高的准确率。

表 2　基于共词网络的社交媒体数据主题挖掘方法与 LDA 主题模型的准确率和召回率对比

Tab. 2　Precision and Recall of the Co-word Based Topic Mining Method and LDA Topic Model

纯度	基于共词网络的社交媒体数据主题挖掘方法		LDA 模型	
	准确率	召回率	准确率	召回率
0.3	0.66	0.60	0.55	0.42
0.4	0.70	0.56	0.58	0.35
0.5	0.81	0.46	0.69	0.23
0.6	0.84	0.41	0.73	0.18
0.7	0.88	0.38	0.80	0.08
0.8	0.94	0.27	0.83	0.07
0.9	0.97	0.23	1.00	0

3　模型应用分析

为了验证所提出的主题挖掘方法在实际应用中的有效性,利用该方法对北京暴雨微博数据集的挖掘结果作进一步的时空分析,以展示其在灾害应急响应中的价值。

3.1　时序分析

2012 年北京暴雨事件中,降雨可以分为两个阶段:第 1 个阶段发生在 21 日 10 时至 20 时,其主要特点是短时雨强大、强度变化波动显著(图 5 的灰色区域),第 2 个阶段发生在当日 20 时后至 22 日 4 时,降水逐渐平缓,雨强显著减小(图 5 的绿色区域)[26]。

本文对 2012 年北京暴雨前后不同主题的微博数目进行了时序分析。图 5 显示了每个小时属于各个类别的微博比率随着时间的变化趋势。可以看出,在暴雨来临之前,天气预警、祈祷和交通状况主题开始出现,并占据较大的比例。这是由于在暴雨来临之前,天气预警影响了航班,导致较多的乘客滞留,他们开始祈祷路途顺利,因此也导致了微博中有关祈祷和交通状况主题的增加。交通信息出现在整个降雨过程中,且在第 1 个阶段降雨时更多一些,反映了暴雨对交通状况的影响更加直接。祈祷信息在暴雨的第 2 个阶段开始增多,并持续了一段时间,反映了随着灾情的发展,人们对自己所处的环境开始慢慢有所认知,进而开始关注灾情状况并祈祷安全度过。救援信息则在暴雨发生一段时间之后增多,主要是由于有关“群众车内被溺死”的新闻所引发的大量讨论所导致的。有关伤亡、提醒朋友、灾害原因讨论的微博比率在灾害的整个过程中变化不大。

3.2　空间分布模式

本文选取了与灾情密切相关的“交通状况”主题进行了空间分析。“交通状况”主题包含 3 360 条带有 GPS 信息的微博。图 6 反映了“交通状况”主题下微博数据的空间分布,可以看出该主题下的微博在空间上形成比较明显的聚集分布。图 6 中共出现 5 处聚集点。其中,A 为首都机场,B 为北京火车站西站,C 为北京站,D 为北京南站,它们反映了暴雨影响航空以及铁路交通,导致了大量旅客滞留机场以及火车站。通过对图 6 中 E 处热点范围内的微博进行探索,发现在 E 处发生了严重的交通拥堵。

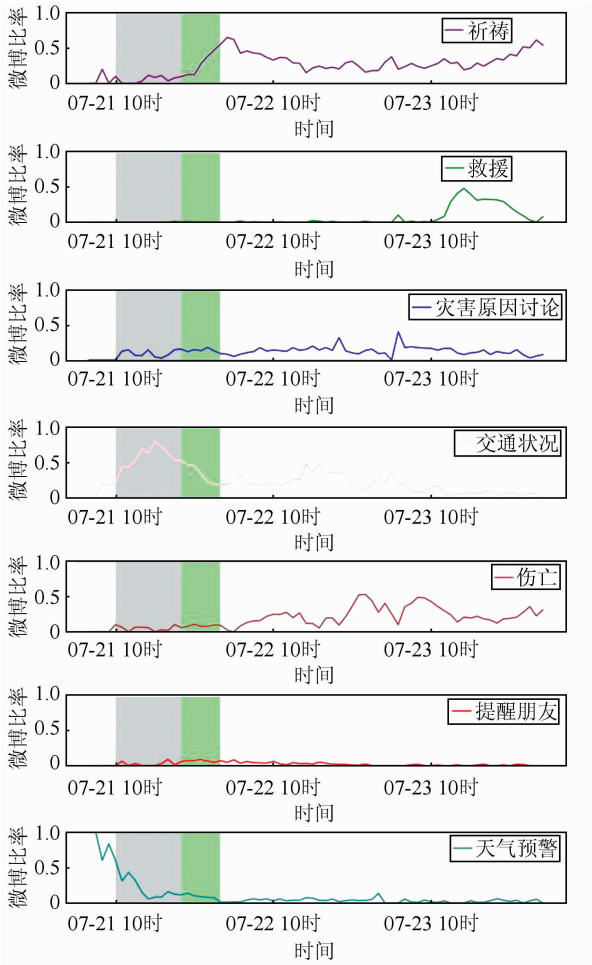

图5 不同应急主题的微博比率随时间的变化

Fig. 5 Ratios of Microblogs Under Different Topics over Time

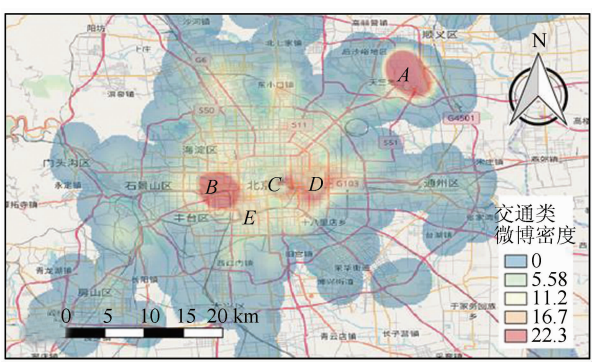

图6 交通类应急主题的空间分布

Fig. 6 Spatial Distribution of Microblogs Under Traffic Topic

4 结 语

社交媒体中通常包含了丰富的时间、空间和文本语义信息，可以用来揭示不同类型的地理现象。对社交媒体所包含的文本数据的深入挖掘，有利于更有效地进行后续的时空分析。本文探讨了在社交媒体短文本的情况下，如何有效地挖掘主题信息，提出了一种新的基于共词网络的社交媒体数据主题挖掘方法。本文所提出的方法是一种无监督方法，不需要指定聚类数目，具有很好的实用性。实验表明，该方法在准确率和召回率上也均优于传统的LDA模型。研究发现，该方法在北京暴雨这一灾害案例中，能够有效地分析事件的某一个方面而去除其他噪音的影响，因而在灾害发生时，能够获取更有价值的应急信息。该研究为如何有效地挖掘社交媒体中所包含的文本主题信息提供了一种新的思路，可用于准实时或灾后灾情的相关评估。由于使用LM社区分割算法得到的结果中，微博只能被划分到单个主题，忽略了一条微博可能包含多个主题的情况，未来可采用重叠社区分割算法进行改进。

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, NY, USA, 2010

[2] 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): e86026

[3] Caverlee J, Cheng Z, Sui D Z, et al. Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media[J]. *IEEE Data Eng Bull*, 2013, 36(3): 33-41

[4] 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

[5] 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): e237-1-e237-13

[6] Salathé M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control[J]. *Plos Computational Biology*, 2011, 7(10): e1002199

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

[8] Wang Yandong, Jing Tong, Jiang Wei, et al. Modeling Urban Air Quality Trend Surface Using Social Media Data[J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(1):14-20(王艳东, 荆彤, 姜伟,等. 利用社交媒体数据模拟城市空气质量趋势面[J]. 武汉大学学报·信息科学版, 2017, 42(1):14-20)

[9] Yates D, Paquette S. Emergency Knowledge Management and Social Media Technologies: A Case Study of the 2010 Haitian Earthquake[J]. *International Journal of Information Management*, 2011, 31(1): 6-13

[10] 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

[11] Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment[J]. *Icwsm*, 2010, 10(1): 178-185

[12] 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, Illinois, USA, 2011

[13] Culotta A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages[C]. The First Workshop on Social Media Analytics, New York, USA, 2010

[14] Huang Q, Xiao Y. Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery[J]. *ISPRS International Journal of Geo-Information*, 2015, 4(3): 1 549-1 568

[15] Wang Yandong, Li Hao, Wang Teng, et al. The Mining and Analysis of Emergency Information in Sudden Events Based on Social Media[J]. *Geomatics 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)

[16] Duan Lian, Guo Wei, Zhu Xinyan, et al. Constructing Spatio-Temporal Topic Model for Microblog Topic Retrieving[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(2):210-213(段炼, 闵维, 朱欣焰,等. 基于时空主题模型的微博主题提取[J]. 武汉大学学报·信息科学版, 2014, 39(2):210-213)

[17] Zhao W X, Jiang J, Weng J, et al. Comparing Twitter and Traditional Media Using Topic Models[C]. Advances in Information Retrieval, Berlin, Garmany, 2011

[18] Kireyev K, Palen L, Anderson K. Applications of Topics Models to Analysis of Disaster-Related Twitter Data[C]. NIPS Workshop on Applications for Topic Models: Text and Beyond, Whistler, Canada, 2009

[19] Cheng Qikai, Wang Xiaoguang. A New Research Frame for Analyzing the Evolution of Research Topics Based on Co-word Network Communities[J]. *Library and Information Service*, 2013, 57(8): 91-96(程齐凯, 王晓光. 一种基于共词网络社区的科研主题演化分析框架[J]. 图书情报工作, 2013, 57(8): 91-96)

[20] 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)

[21] 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

[22] Sampson G. Word Frequency Distributions[J]. *Computational Linguistics*, 2002, 28(4): 565-569

[23] 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, DOI: 10. 1088/1742-5468/2008/10/P10008

[24] Heinrich G. Parameter Estimation for Text Analysis[EB/OL]. http://citeseerx. ist. psu. edu/viewdoc/download? doi=10. 1. 1. 74. 6555&rep=rep1 &type=pdf,2008

[25] 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

[26] Sun Jisong, He Na, Wang Guorong, et al. Preliminary Analysis on Synoptic Configuration Evolvement and Mechanism of a Torrential Rain Occurring in Beijing on 21 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变特征及成因初探[J]. 暴雨灾害, 2012, 31(3):218-225)

A New Social Media Topic Mining Method Based on Co-word Network

WANG Yandong[1,2,3] *FU Xiaokang*[1] *LI Mengmeng*[1]

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

Abstract: The in-depth exploration of the text data contained in social media facilitates efficient analysis of time and space. This paper proposes a new social media topic mining method based on the concept of co-word network and community detection. The method uses term frequency-inverse document frequency (TF-IDF) analysis to identify the key words of the messages automatically. Based on the problem whether the microblogs contain the same key words or not, we put forward the concept of microblog co-word network with microblog as the node. The network combined with the Louvain community detection algorithm is used to classify the microblogs into different clusters with topics. The proposed method is an unsupervised method. The advantage of this method is that there is no need to specify the number of clusters. Experiments demonstrate that the performance of the proposed method is better than the commonly used latent dirichlet allocation (LDA) model on both precision and recall. Taking the collected microblogs during the 2012 Beijing rainstorm as the case study, the method is used to conduct in-depth mining and time-space analysis of the microblogs dataset. The results demonstrate that the proposed method is effective in real world applications.

Key words: co-word network; social media; Louvain community detection; topic mining

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, No. 2016YFB0501403; the National Natural Science Foundation of China, No. 41271399; China Special Fund for Surveying, Mapping and Geoinformation Research in the Public Interest, No. 201512015.