

引用格式: 刘淑涵, 王艳东, 付小康. 利用卷积神经网络提取微博中的暴雨灾害信息[J]. 地球信息科学学报, 2019, 21(7): 1009-1017. [Liu S H, Wang Y D, Fu X K. Extracting rainstorm disaster information from microblogs using convolutional neural network[J]. Journal of Geo-information Science, 2019, 21(7): 1009-1017.] DOI: 10.12082/dqxxkx.2019.180701

利用卷积神经网络提取微博中的暴雨灾害信息

刘淑涵¹, 王艳东^{1,2,3*}, 付小康¹

1. 武汉大学测绘遥感信息工程国家重点实验室, 武汉 430079; 2. 地球空间信息技术协同创新中心, 武汉 430079;
3. 东华理工大学测绘工程学院, 南昌 330013

Extracting Rainstorm Disaster Information from Microblogs Using Convolutional Neural Network

LIU Shuhan¹, WANG Yandong^{1,2,3*}, FU Xiaokang¹

1. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; 2. Collaborative Innovation Center for Remote Sensing, Wuhan University, Wuhan 430079, China; 3. Faculty of Geomatics, East China University of Technology, Nanchang 330013, China

Abstract: Nowadays social media has played an increasingly significant role in disaster management, thanks to its real-time nature and location-based services. When a disaster happens, a large number of images and texts with temporal and geographic information quickly flood in the social media network. Complementary to the traditional disaster management, social media could provide a lot of dynamic, nearly real-time disaster information to researchers. Current studies place more emphasis on using machine learning to deal with social media disaster data. Yet, in many cases deep learning has a better performance in automatic feature extraction than the traditional machine learning, and it can be used to extract and classify disaster information from social media. This paper focused on a method of extracting the disaster information from social media data using Convolutional Neural Network (CNN). To obtain the word vector corresponding to social media texts, a corpus of disaster events by using social media data was trained by word2vec model. Then, the vectorized microblog sentences and their corresponding disaster categories were used as input to the multi-classification model, which is based on convolutional neural network. After training and optimization, we used this model to extract disaster information from a large number of social media data streams. For an experiment, we combined Sina Weibo API and web crawler, and got over twenty thousand microblog texts with the theme of "Beijing Heavy Rainstorm" happened in 2012. Besides the irrelevant texts, we divided the data into seven categories. The topic classification model of rainstorm disaster information was built and trained based on a small number of tagged Sina Weibo data. The experimental results achieved the F-value of over 80% and the precision of over 90%, proving the validity of applying the model to our dataset. Moreover, this model when used to classify the data on Beijing's

收稿日期: 2018-12-28; 修回日期: 2019-03-25.

基金项目: 国家重点研发计划项目(2016YFB0501403); 国家自然科学基金项目(41271399); 测绘地理信息公益性行业科研专项经费项目(201512015)。[**Foundation items:** the National Key Research Program of China, No.2016YFB0501403; the National Natural Science Foundation of China, No.41271399; China Special Fund for Surveying, Mapping and Geo-information Research in the Public Interest, No.201512015.]

作者简介: 刘淑涵(1996-), 女, 湖北十堰人, 硕士生, 研究方向为地理时空数据分析与挖掘。E-mail: liush96@whu.edu.cn

*通讯作者: 王艳东(1972-), 男, 湖南岳阳人, 博士, 教授, 研究方向为城市大数据时空分析计算。E-mail: ydwang@whu.edu.cn

rainstorm in 2016 newly crawled from Weibo also had a good performance. According to the different rainstorm emergency topics classified by model, we carried out the deep mining of time series and spatial features to detect the phases of disaster development. Through visualization and statistical analysis, it was found that the time series analysis of disaster was consistent with the development of actual disasters, indicating the effectiveness of the CNN-based method in monitoring Beijing rainstorm. The study shows that using deep learning to extract disaster emergency information from social media is effective and feasible, which provides a new approach to real-time disaster emergency management.

Key words: convolutional neural network; Sina Weibo; short text classification; rainstorm disaster; information extraction

***Corresponding author:** WANG Yandong, E-mail: ydwang@whu.edu.cn

摘要:从社交媒体中挖掘灾害应急信息,能够有效帮助传统灾害管理获取实时、主题丰富的灾害信息,从而成为灾害应急管理的新手段。得益于深度学习在自动特征提取上的成就,本文研究了一种利用卷积神经网络对社交媒体中的灾害应急信息进行自动实时提取与分类的方法。首先,利用社交媒体数据和 Word2vec 模型,构建与灾害类事件相关的语料库并获得相应的词向量;其次,将词嵌入文本和相应的灾情类别作为卷积神经网络的输入,经过多分类学习得到分类模型,用以提取近实时灾害信息。以 2012 年“7.21 北京特大暴雨”事件为案例,通过分类模型获得常见灾情类别的暴雨灾害社交媒体信息。该模型在测试集上的精度达到了 90% 以上,并且将模型运用于新爬取的 2016 年暴雨数据集上也得到了较好的表现,说明该模型在近实时自动提取灾害信息方面具有可行性。在对 2012 年分类结果进行时空分析结果表明,通过社交媒体获得的暴雨灾害主题信息符合灾害发展的规律,说明了利用深度学习提取社交媒体数据中的灾害应急信息的有效性和可行性,能够为实时灾害应急管理提供新的思路。

关键词:卷积神经网络;新浪微博;短文本分类;暴雨灾害;灾害信息提取

1 引言

据联合国减灾署在 2016 年国际减灾日发布的报告显示,全球约有 135 万人过去 20 年在 7000 多起灾害中死亡,灾害每年对世界造成的经济损失高达 3000 亿美元^[1]。灾害信息的获取事关防灾减灾救灾决策的科学性,能够为灾害预警、损失评估、救灾决策、恢复重建提供可靠的信息支撑^[2]。灾害现场信息的获取,更是应急灾害管理的关键环节^[3]。卫星遥感观测技术能够客观准确地获取地面真实状况,比如利用多源卫星遥感可以提取暴雨灾情的时空动态信息^[4],但卫星遥感在获取灾害信息时受空间分辨率、运行周期的影响较大,不能满足灾害信息获取实时性的要求。

社交媒体,以其信息传播的实时性和基于位置服务的高速发展,成为灾害应急信息获取的新来源^[5-8]。过去,主要由官方政府等单方面调控灾害,反馈效率低下;而现在,有了社交媒体作为补充,能够让灾害信息在第一时间以真实、原始的方式展现在公众面前,并能为灾害管理者提供决策支持^[9-10]。以日本地震为例,Sakaki 等^[11]利用支持向量机算法(Support Vector Machine, SVM)对日本地震相关的

推特文本进行分析,使用概率模型(Probabilistic Model)对地震进行事件探测和位置估计,最终实现了比官方更快速地传播灾害信息。Wang 等^[12]结合文档主题生成模型(Latent Dirichlet Allocation, LDA)和 SVM 算法,基于暴雨主题内容对社交媒体文本流进行分类。而 Zuo 等^[13]从时间、空间、语义 3 个方面阐述了基于地理标记社交媒体数据分析的方法,用于感知社会事件的影响。

在社交媒体大数据环境下,深度学习技术的飞速发展使得其在灾害应急管理中应用也越来越广泛^[14]。Zhang 等^[15]基于深度信念网络(Deep Belief Nets, DBN)提出了一种中国紧急事件识别模型,之后提出了一种动态监督的 DBN,提高了识别性能并有效地控制了训练时间。Li 等^[16]利用卷积神经网络(Convolutional Neural Network, CNN)对从微博中抓取的数据进行分类训练,并将模型运用于深圳和武汉暴雨事件的监测上。深度学习能够通过多层非线性处理单元自动学习社交媒体文本的句法语义特征,替代了传统的人工特征获取。在利用深度学习对社交媒体文本进行提取的研究中,由于短文本的特殊性^[12],这些研究对模型自动、近实时提取新文本特征的能力的应用研究是不够的。

本文研究了一种利用卷积神经网络对社交媒体中的灾害应急信息进行提取与分类的方法,对暴雨灾害主题进行了更加细致的划分,实现了对社交媒体暴雨灾害信息的自动近实时提取。基于新浪微博,利用卷积神经网络构建了多分类模型,实现从大量的社交媒体文本流中提取出天气、交通、救援等暴雨相关灾害信息。基于由暴雨文本构成的语料库,对从微博获取的数据进行词嵌入处理,作为神经网络的输入层;通过卷积神经网络进行训练和优化,将训练完成后的神经网络用于新的暴雨信息的主题分类。同时,从相对数量、空间属性2个方面,对不同的暴雨灾害主题进行可视化和统计分析,探究暴雨事件的灾害发展时间趋势以及空间分布特征。在测试集上的分类精度达到了90%,在新的社交媒体数据集上的应用也有较好的表现,说明了利用深度学习自动提取近实时社交媒体数据中的灾害应急信息具有有效性,能够为实时灾害应急管理提供新的思路。

2 研究方法、数据来源及处理

2.1 研究框架

灾害爆发时,社交媒体会产生大量的不同主题的带有时间、空间属性信息的数据。利用深度学习,对社交媒体进行信息甄别与分类,可以挖掘灾害信息以达到监测灾害事件的目的。本文使用的基于深度学习的社交媒体暴雨灾害信息分类研究方法如图1所示。在此框架中有数据获取与处理、文本特征提取、模型构建与训练3个层次。为了更深入地验证此框架在实际应用中的有效性,本文对暴雨灾害信息进行了时空分析,以体现本文的研究

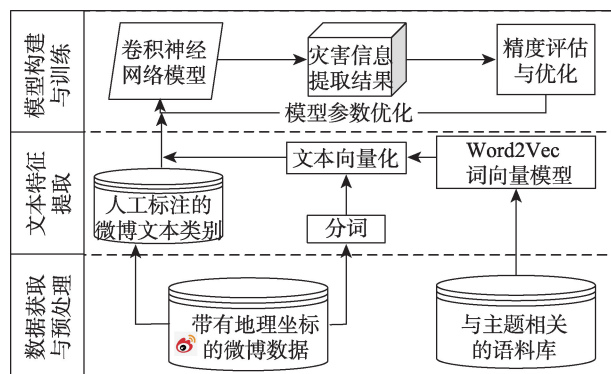


图1 利用卷积神经网络提取微博暴雨灾害信息的研究框架

Fig. 1 Research framework of extracting rainstorm disaster information from microblogs using CNN

在灾害管理中的意义。

2.2 数据获取与预处理

获取适合研究主题的数据是研究的基础。本文利用新浪微博,结合微博API和网络爬虫的方式,获取特定主题的微博数据。带有地理坐标的新浪微博数据可以被认为是地理空间中的一个点,有助于进行空间分析,从而适合作为本研究的实验数据。而其余不带地理坐标的数据,可以被用作构建与主题相关的语料库。语料库是与主题相关的微博文本的集合,是进行词向量训练的基础。

依据特定关键字抓取的数据,存在与研究内容关系不大或者语义类别丰富的情况,所以需要对文本进行筛选和分类。人工标注是有监督分类学习中常用的添加标签的方法,其优点在于人对文本语义的理解比算法好,有准确度高特点。通过对微博文本进行阅读,确定多个常见灾害信息主题,并通过人工标注使得每一条训练文本都有特定的主题。

2.3 微博文本特征获取

文本矩阵化,是对社交媒体数据进行特征提取重要的部分。这里有2个步骤:①利用与主题相关的语料库,生成词向量空间;②基于词向量空间实现样本文本的向量化。

2.3.1 词向量空间的构建

预训练的词向量在短文本分类任务中是一个重要的因素^[17-18]。Word2vec实现了从文本到向量的映射,满足了神经网络需要结构化数据作为输入的需求。本文使用Word2vec移植在Python语言Gensim模块上的开源C包来生成词向量空间。将符合条件的文本和停用词输入Word2vec工具,通过优化后的训练模型生成词向量空间。词向量空间中存储了微博中的单词及其对应向量。词向量空间生成的思路是:

(1)读取暴雨主题相关的语料库,解析出其中的文本内容,采用Jieba分词对其进行分词。Jieba分词结合了机械分词和统计分词两种分词算法,能够将连续文本划分为独立词汇。

(2)对分词后的文本进行去除停用词操作,得到语料库。停用词是文本中没有实际语义、对研究帮助甚微的词的集合。例如:“啊”等语气词、无法被解析的表情符、“这些”等功能词。本文结合暴雨微博文本生成了本实验的停用词集合,尤其剔除了

“暴雨”等在研究中没有反应灾情的词汇。

(3)Word2vec依据语料库,利用连续词袋(Continuous Bag-of-Words, CBOW)和Skip-gram模型进行训练,得到词向量空间。Mikolov在文章^[19]中提出了CBOW和Skip-gram两种模型,这也是Word2vec工具中主要运用的2个模型。研究表明,CBOW模型在大语料库(大于一亿词)上具有更好的训练效果^[20]。故在实验中,对大语料库使用CBOW算法,小语料库使用Skip-gram来训练模型。

2.3.2 文本向量化

文本向量化处理,通过捕捉有意义的语义和句法规则可以缓解数据稀疏性问题,并将这些词映射到实数的向量便于CNN模型的后续处理。为了满足神经网络的输入,需要将训练集的文本转换为矩阵。将每一条微博文本单独分词,从词向量空间中获取所分单词的对应向量组成矩阵,这时每一个矩阵对应着微博中的一条文本。不包含在词向量空间中的单词,其词向量由随机初始化得到^[16]。如果一条微博文本有 n 个词汇,每个词汇是一个 d 维的向量,句子中词汇对应的词向量从上到下依次排列,那么就构成了一个 $n \times d$ 的矩阵,矩阵形式如图2所示。

2.4 卷积神经网络模型构建与训练

2.4.1 卷积神经网络模型的构建

本研究使用CNN对微博文本中的暴雨灾害应急信息进行识别和分类。选择CNN的原因在于:① CNN基于数据的局部相关性来提取特征,而微博短文本也具有这个特性;② 卷积层的局部连接特性使得其参数较少,减轻了计算负担。

CNN的核心是几层具有像修正线性单元(Rectified Linear Unit, ReLU)或者双曲正切(tanh)这样的非线性激活函数的卷积,它的基本组成结构是输

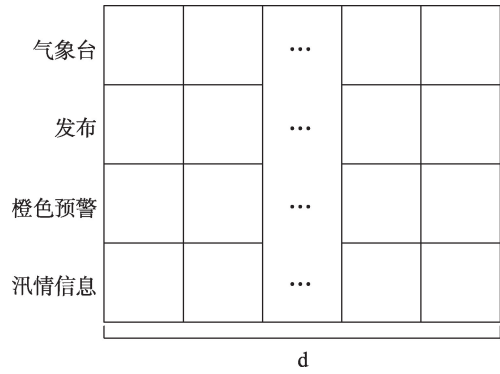


图2 一条微博文本对应的矩阵形式

Fig. 2 An example of a sentence matrix

入层、一组或几组卷积层和池化层的组合,以及带有分类器的全连接层用以输出^[21]。本文构建了如图3所示的模型,使用了基于Tensorflow的keras深度学习框架进行实现。由于训练集的限制,这里没有使用卷积层和池化层的多轮叠加,也没有使用融合层的操作,只使用了最基础的模型结构。从实验结果来看,简单的模型结构一定程度上可以提高训练效果,避免过拟合问题。

模型实现了由社交媒体文本到文本类别的映射。对于暴雨灾害相关的微博文本,输入层将每一条微博文本的单词嵌入到低维向量,其变量 x 和 y 分别是暴雨灾害文本和对应的灾情主题。本文使用了预训练的词向量,即2.2节对应的内容,直接将句子对应的矩阵输入到了卷积层中。在程序实现时,针对微博文本长短不一的问题,本文依据文本的最长长度构造定长的矩阵,长度不足的微博文本采用零填充的方式补齐。

卷积和池化是文本灾害信息提取的关键步骤。首先是使用特定窗口的过滤器对嵌入的单词向量进行卷积。因为矩阵行表达的是一句完整的信息,所以过滤器的宽应与词向量维度一致,每次

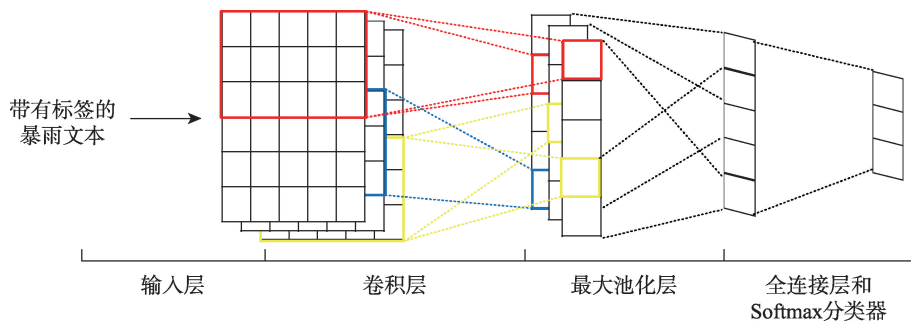


图3 用于暴雨文本分类的卷积神经网络结构

Fig. 3 Structure of the convolutional neural network for text classification

窗口覆盖连续的几个单词。这里不考虑边界问题,基于过滤器特殊的形状,得到了一个宽度为1的特征图。其次是对卷积层的结果执行最大池化操作。研究表明,最大池化操作的效果通常比平均池化的效果要好^[22]。取特征图中值最大的元素,该元素被认为是在卷积之后的最重要的特征。

最后,通过全连接层和 Softmax 分类器获得分类结果。Softmax 函数的输出是单条文本属于每一类灾害主题的概率分布,取概率值最大的一类作为最终的分类类别。由此,通过少量标注的数据,便可实现利用分类模型,自动完成大量社交媒体文本的灾害主题分类。

2.4.2 模型精度评估依据

在训练集上学习得到初步的模型,通过交叉验证方法,在测试集上评价模型的测试误差,并计算精度,以选出最合适的分类模型。

本实验的分类性能通过精确度(*precision*)、召回率(*recall*)、F值(*f1_score*)3个指标进行衡量。精确度是指在确定的总阳性中真阳性的比例(真阳性和假阳性总和)。召回率是真阳性与真阳性和假阴性之和的比例。F值综合衡量精度和召回率。精确度、召回率和F值的定义由式(1)–(3)给出。

$$precision = \frac{TP}{TP + FP} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$f1_score = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

式中:TP、FP和FN分别表示真阳性,假阳性和假阴性的数量。

3 实验与结果

3.1 实验数据描述

以2012年“7.21北京特大暴雨”灾害事件作为研究案例。2012年7月21日至22日8时左右,中国大部分地区遭遇暴雨,其中北京周边地区遭遇61年来最强暴雨及洪涝灾害。根据北京市政府举行的灾情通报会的数据显示,此次暴雨造成160.2万人受灾,经济损失达116.4亿元,其暴雨降雨量之多、强降雨历时之长、局部洪水之严重实属历史罕见。因此,结合微博API和网络爬虫,从微博中以“北京暴雨”为关键字获取了从2012年7月20日6时到8月10日24时共26 059条带有时间和位置信息的数

据。数据形式包含微博ID、用户名、用户ID、发布时间、微博文本和发布位置等信息。

基于先验知识和相关文献的总结^[23],剔除掉微博中与北京暴雨事件不相关的内容,将剩余暴雨文本分为天气预警、交通信息、提醒朋友、灾害原因讨论、正能量祈祷、伤亡受灾、救援信息7个灾害相关信息主题。基于这些主题,人工阅读标记了2870条文本数据作为CNN模型的训练样本,并将训练样本随机划分为了训练集和测试集以便构建误差最小的模型。因为随机划分训练集和测试集可能导致某些类别遗漏,以致训练效果较差,因此在处理训练集和测试集时,从已标记数据中按各类别比例取出了75%的数据集作为训练集,25%的数据集作为测试集。

3.2 精度评估结果

神经网络训练过程中最重要的步骤之一是优化。通过在训练集上不断反复尝试,尽力去寻求最合适的参数和模型结构,以提高实验精度。在本实验中,采用加入Dropout层和正则化加权来控制过拟合问题,采用网格优化策略来选择合适的参数。通过对模型进行精度评估,来调整和确定最终的分

类模型。精度评估是基于测试集进行的。基于评估标准的分类精度对比如图4所示。从图中可以看出,利用深度学习对微博中的暴雨灾害信息进行提取具有可行性。F值处于精确度和召回率之间的折衷,数值达到83%,分类器具有一定的分类效果。整体分类精度高达90%,比召回率高,说明分类器识别正确文本的能力仍有待提高。

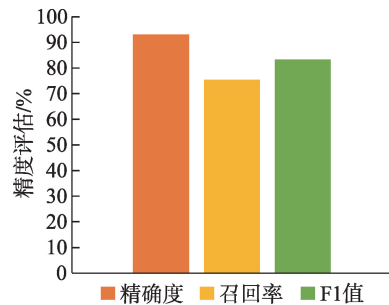


图4 微博暴雨灾害文本分类精度评估

Fig. 4 Evaluation of the classification experiment using weibo rainstorm datasets

针对每个类别计算其数量和精度如表1所示。从2012年北京暴雨相关灾害信息分类结果来看,数量上,天气预警信息被识别出的数量较少,正能量

表1 2012年和2016年北京暴雨相关灾害信息精度评估结果

Tab. 1 Evaluation of the various types of rainstorm disaster information in Beijing in 2012 and 2016

类别	2012年北京暴雨				2016年北京暴雨			
	数量/条	精确度/%	召回率/%	F1值/%	数量/条	精确度/%	召回率/%	F1值/%
正能量祈祷	129	89.9225	89.2308	89.5753	26	71.2308	40.7347	51.8296
交通信息	75	97.3333	83.9080	90.1235	24	91.6667	62.8571	74.5763
伤亡受灾	24	99.9999	82.7586	90.5660	4	74.9999	42.8571	54.5454
灾害原因讨论	45	95.5556	91.4893	93.4783	26	84.6154	73.3333	78.5714
提醒朋友	54	88.8889	77.4193	82.7586	5	80.0000	57.1428	67.0000
天气预警	3	99.9999	6.9767	13.0434	3	66.6666	6.2499	11.4286
整体	330	93.0303	75.4299	83.3107	88	81.6818	39.6904	53.4222

祈祷信息则相对较多,认为通过微博可以有效进行正能量舆论氛围的传播。从精确度、召回率和F1值来看,大部分都在80%以上,说明大部分的灾害主题都能被正确识别,说明本文使用的方法具有较好的分类效果;而天气预警类别的召回率表现比较差,分析发现,有较多被错误识别的天气预警信息,这可能与天气预警微博文本本身的复杂性有关:天气预警信息可能还包含交通信息的预报、提醒注意安全等内容;另外,因测试集中救援信息类数量较少,没有被识别出,故没有计算其准确度和召回率。

为了对模型自动近实时的特性进行验证,从新浪微博中又爬取了以“北京暴雨”为关键词的从2016年7月17日到7月24日的微博,并从中抽取600条数据进行人工标注。将这600条数据运用于灾害信息提取模型中进行分类,其精度评估结果如表1所示。可以看到,整体精度达到了80%,说明该分类器具有自动提取文本特征的能力;结合具体文本来,大部分类别的精度处于70%左右及以上,通过阅读文本发现,人工标注的结果中有398条不相关内容,因此提取主题信息的数量较少;另外,2016年的数据存在文本缺失的情况,导致部分救援信息类别没有被识别出来,以及大部分类别召回率较低。但从精确度来看,该模型在实时自动提取灾害信息方面具有可行性。

4 灾害应急信息时空分析

4.1 应用描述

为了验证本文的方法在实际灾害应用中的有效性,利用该方法对北京暴雨微博数据集的挖掘结果做进一步的时空分析。这部分实验所涉及的数据来源于3.1节中除去样本数据集的剩余23189条

微博数据。利用模型对该数据集进行灾害信息的分类,剔除其中与暴雨无关的微博数据,针对不同主题的灾害信息,本文对其进行了时序上的趋势分析和空间上的热点分析。不同主题的灾害信息数量统计如图5所示。

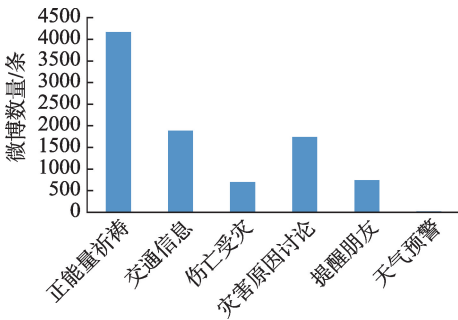


图5 2012年北京暴雨相关灾害信息的微博数量
Fig. 5 Number of microblogs related to Beijing rainstorm disaster information in 2012

4.2 时序分析

通过统计每小时各个类别的数量和占比,可以得到各时段灾害响应的情况,有助于了解灾害的发展规律。本文对7月20日到7月24日的数据进行了时序分析。因数据问题,分类结果中没有预测出救援信息。如图6所示,图6(a)显示了暴雨发生前后不同主题微博的数量变化,图6(b)显示了暴雨发生前后不同主题微博占比的变化。通过分析不同主题微博的数量随时间变化的趋势,可以探测出暴雨灾情发展的基本情况和规律,而每小时每个类别的微博占比反映了人们在灾害响应的不同阶段对不同类别的关注度差异。北京暴雨的大范围降雨主要可以分为2个时段:图中灰色区域是21日10时-20时,这个时间正是强降雨爆发时段,短时降雨量非常大,而绿色区域表示21日20时-22日4时,这个时间降水量有下降趋势,但仍有大雨^[24]。

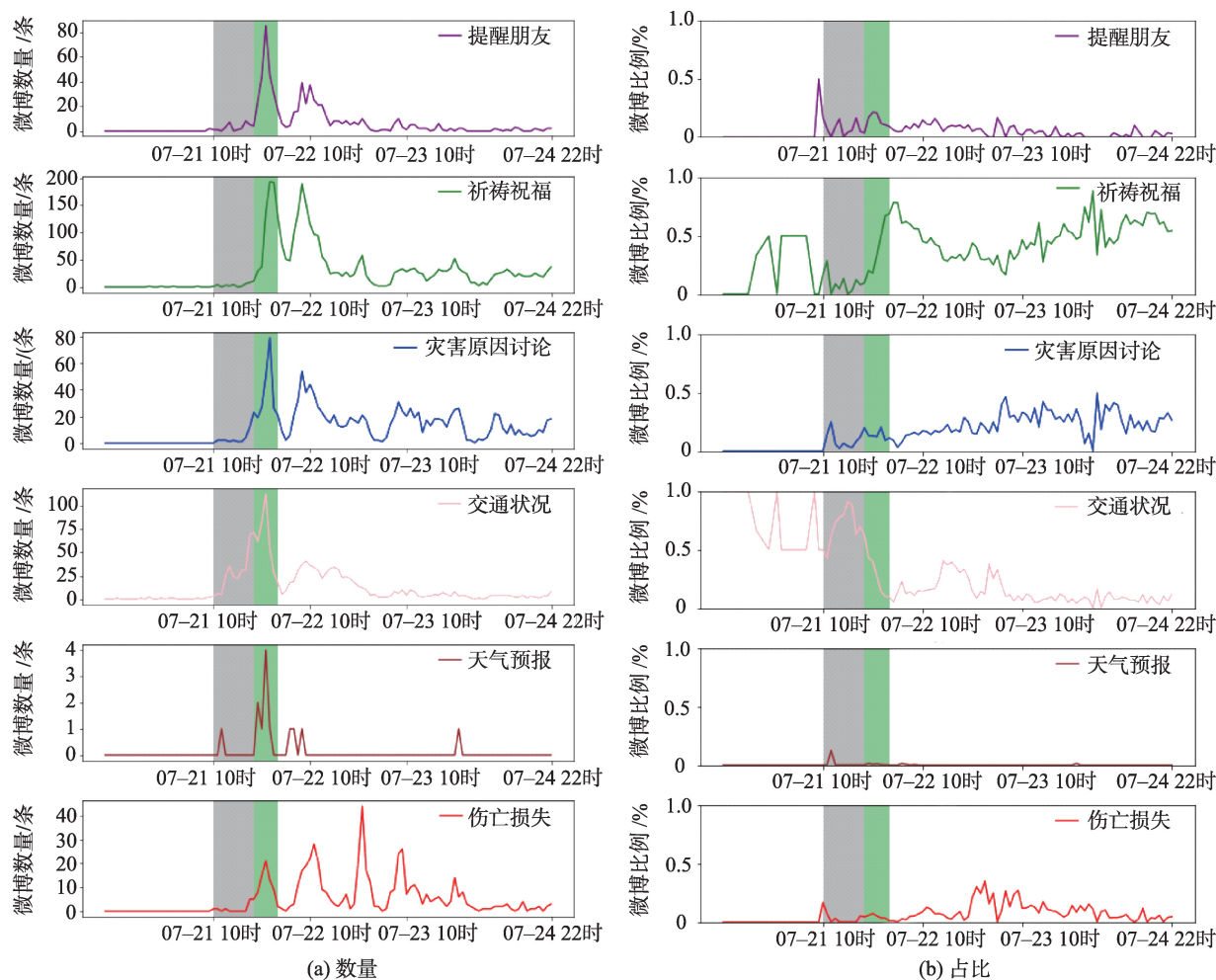


图6 2012年北京不同主题暴雨灾害信息数量和占比随时间变化曲线

Fig. 6 Trends of the different rainstorm disaster information over time in Beijing in 2012

从图6(a)来看,大部分灾情主题的微博数量都随着暴雨事件的发生经历了大幅度起伏,在强降雨发生期间,微博数据均达到极值,说明从微博中提取的暴雨灾害主题的发展趋势与实际灾害阶段一致。除去伤亡损失类别,各个灾害主题在数量上都呈现出双峰的趋势。这表明,除了在强降雨发生时人们表现出对降雨的关注,在降雨结束后出现的严重灾害信息也影响着人们的生活。伤亡损失类信息在灾后呈现出波动的趋势,说明社交媒体上逐步出现了伤亡信息的传播。

而从图6(b)来看,在暴雨发生之前,航空公司发布预警,活跃在微博上的人开始关注空中交通信息,在暴雨发生时,铁路、公路、航空交通均受到不同程度的影响,故微博占比较高;在暴雨发生后期,也就是灾害恢复阶段,人们探讨灾害原因,吐槽北京排水系统的同时,灾后救援工作也随着伤亡损失

的出现紧张进行着;于是人们不断在微博上传播着正能量信息,祈祷受灾人员的平安,赞扬救灾人员的敬业,提醒朋友注意安全并宣传简单基础的救援策略;而微博作为一个社交平台,朋友之间的@行为(提醒朋友)发生在灾害的全过程中。从不同的灾害主题变化趋势中,可以探究出灾害发展的不同阶段,因此,通过分析微博的主题,有助于从不同的角度分析灾害的发展。

4.3 空间分布模式

带有GPS信息的微博具有位置信息,从而每一条微博可以被认为是一个具有类别的点状实体。图7中的地图显示了由红色点表示的这些灾害相关的新浪微博文本的空间分布。在北京市五环高速通道范围内和机场地区,确实可以观察到新浪微博文本的集中;而且不难发现,一些信息点沿着道路

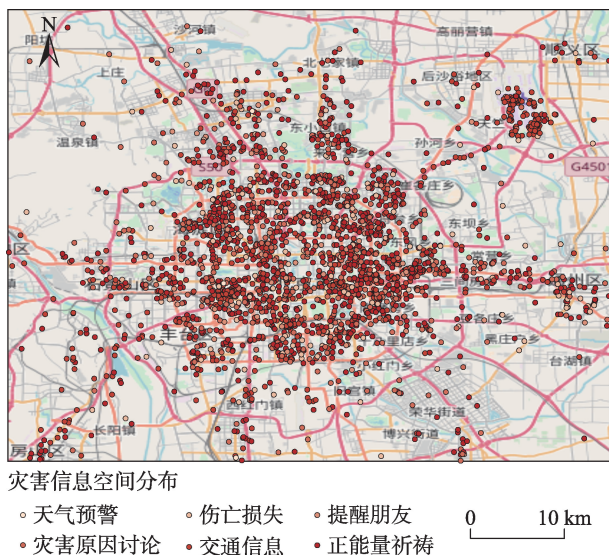


图7 2012年北京暴雨灾害信息空间分布

Fig. 7 Spatial distribution of the disaster information obtained from Beijing rainstorm in 2012

分布,猜测这些点可能反应了道路上的交通状况。

为了进一步探索新浪微博中的空间信息,从预测结果中抽取与交通状况相关的1889条文本,通过对交通类信息进行空间上的核密度分析,得到如图8的空间分布图。从图中可以看出,在空间上交通信息呈现出一种聚集的趋势,共有5处聚集热点区域。A为北京西站,B为北京南站,C为北京站和王府井商圈,D为CBD商圈,E为首都机场。这些热点地方表现出人群聚集的特点,符合突发事件现场信息可以第一时间通过社交媒体进行反馈的事实。而次要热点主要沿道路分布,一定程度上反映了暴雨对道路交通的影响情况,突出的有北京北站、热门商圈等人群聚集地和积水地段。简而言之,使用新浪微博数据集对北京暴雨事件的分类结

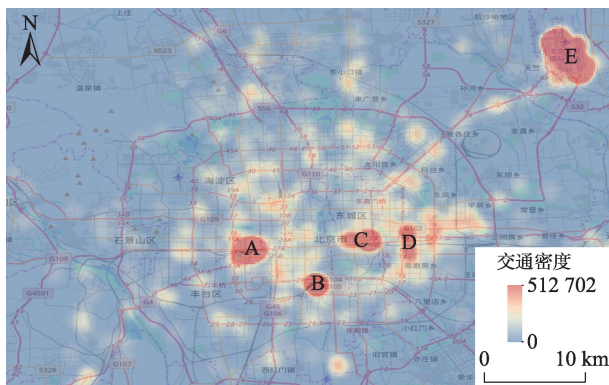


图8 2012年北京暴雨交通类灾害信息的空间聚类

Fig. 8 Cluster map of the traffic disaster information obtained from Beijing rainstorm in 2012

果显示了基于CNN的方法在监测这些地区发生的强降雨事件中的有效性。

5 结论与讨论

多年来,灾害信息获取与分析一直是政府及科研机构研究的重点。社交媒体能够让官方、受灾者共同成为灾害信息提供者和传播者,也因其广泛、实时、及时反馈等特点,成为灾害应急信息的重要来源。而随着技术的不断进步,学者们研究灾害信息获取的方法也在逐步优化和拓宽。

本文使用的基于卷积神经网络的多分类模型,尽管仅运用于社交媒体中暴雨灾害信息的分析,但它同样适用于短时间内在社交媒体上引起大范围轰动的其他类别的灾害事件。在本文中,首先使用了网络爬虫结合微博API的方式,获取了具有位置、时间信息的微博数据,这将有助于后续处理与分类。结合先验知识阅读微博文本,接着,将其标注为不同的暴雨灾情主题,甄别出其中的灾害应急相关的信息。之后,基于CNN,本文构建了一个适合微博短文本灾害提取的网络框架。在经过了控制过拟合、网格参数优化等优化操作后,模型在测试集上的准确度得到了提高,分类精度也达到90%以上。在新获取的2016年数据上进行验证的结果更加说明将模型运用于灾害信息分类具有一定的准确度。最后通过对数据进行可视化与统计分析,发现灾害信息与实际灾害发展阶段一致,说明了基于卷积神经网络的方法在监测北京暴雨灾害事件的有效性,能够有效帮助官方进行灾害决策。

但是,在社交媒体提取灾害信息方面,本研究还存在一些不足:①社交媒体数据本身可能存在部分垃圾信息,造成数据质量低下,从而影响分类精度。②社交媒体数据挖掘获取的信息仍然是片面的,它受到人口、气象、地形等多个因素的影响。后续可以结合这些因素进行综合分析。③未来将设计实时灾害地理信息系统实现对灾害信息的动态展示和监控。

参考文献(References):

- [1] 韩雪华,王卷乐,卜坤,等.基于Web文本的灾害事件信息获取进展[J].地球信息科学学报,2018,20(8):1037-1046.
- [Han X H, Wang J L, Bu K, et al. Progress on disaster events information acquisition from web text[J]. Journal of Geo-information Science, 2018,20(8):1037-1046.]

- [2] 范春波,张鹏,汪洋,等.大力推进灾情管理科技应用[J].中国减灾,2017(11):22-25. [Fan C B, Zhang P, Wang Y, et al. Vigorously promote the technology application of disaster management[J]. Disaster Reduction in China, 2017(11):22-25.]
- [3] 贾平,张云霞,刘克俭.提升综合应急装备水平增强应急决策信息保障能力—灾害现场信息获取技术研究与应用示范获得国家重点研发计划支持[J].中国减灾,2016(21):30-31. [Jia P, Zhang Y X, Liu K J, et al. Improve the level of comprehensive emergency equipment, enhance the ability of emergency decision-making information support - research and application demonstration of disaster site information acquisition technology has won the support of national key research and development plans [J]. Disaster Reduction in China, 2016(21):30-31.]
- [4] 苏亚丽,郭旭东,雷莉萍,等.基于多源卫星遥感的暴雨灾情时空动态信息的提取[J].地球信息科学学报,2018,20(7):1004-1013. [Su Y L, Guo X D, Lei L P, et al. Spatio-temporal dynamic information extraction of heavy rain disaster impacts on the growth of crop using multi-satellites observing data[J]. Journal of Geo-information Science, 2018,20(7):1004-1013.]
- [5] 陈梓,高涛,罗年学,等.反映自然灾害时空分布的社交媒体有效性探讨[J].测绘科学,2017,42(8):44-48. [Chen Z, Gao T, Luo N X, et al. Empirical discussion on relation between realistic disasters and social media data[J]. Science of Surveying and Mapping, 2017,42(8):44-48.]
- [6] Zhang N, Chen H, Chen J, et al. Social media meets big urban data: A case study of urban waterlogging analysis [J]. Computational intelligence and neuroscience, 2016, 2016:No.3264587.
- [7] Wang Z, Ye X. Social media analytics for natural disaster management[J]. International Journal of Geographical Information Science, 2018,32(1):49-72.
- [8] 王森,肖渝,黄群英,等.基于社交大数据挖掘的城市灾害分析——纽约市桑迪飓风的案例[J].国际城市规划, 2018,33(4):84-92. [Wang S, Xiao Y, Huang Q Y, et al. Research on urban disaster analysis based on the big data mining of social media: Case study of hurricane sandy in New York[J]. Urban Planning International, 2018,33(4): 84-92.]
- [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] Crooks A, Croitoru A, Stefanidis A, et al. # Earthquake: Twitter as a distributed sensor system[J]. Transactions in GIS, 2013,17(1):124-147.
- [11] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-time event detection by social sensors[C]. International Conference on World Wide Web. ACM, 2010:851-860.
- [12] Wang Y D, Wang T, Ye X, et al. Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm[J]. Sustainability, 2016,8(1):25.
- [13] Zhu R, Lin D, Jendryke M, et al. Geo-Tagged social media data-based analytical approach for perceiving impacts of social events[J]. ISPRS International Journal of Geo-Information, 2019,8(1):15.
- [14] Mouzannar H, Rizk Y, Awad M. Damage identification in social media posts using multimodal deep learning[C]. Proceedings of the 15th ISCRAM Conference, 2018:529-543.
- [15] Zhang Y, Liu Z, Zhou W. Event recognition based on deep learning in Chinese texts[J]. PloS one, 2016,11(8): e0160147.
- [16] Li J, He Z, Plaza J, et al. Social Media: New perspectives to improve remote sensing for emergency response[J]. Proceedings of the IEEE, 2017,105(10):1900-1912.
- [17] Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning?[J]. Journal of Machine Learning Research, 2010,11(3):625-660.
- [18] Kim Y. Convolutional neural networks for sentence classification[J]. Empirical Methods in Natural Language Processing, 2014:1746-1751.
- [19] Mikolov T, Chen K, Corrado G S, et al. Efficient estimation of word representations in vector space[J]. ArXiv Preprint, arX-iv:1301.3781,2013.
- [20] Lai S, Liu K, He S, et al. How to generate a good word embedding[J]. IEEE Intelligent Systems, 2016,31(6):5-14.
- [21] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems, 2012: 1097-1105.
- [22] Zhang Y, Wallace B C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. International Joint Conference on Natural Language Processing, 2017:253-263.
- [23] 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):1549-1568.
- [24] 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. [Sun J S, He N, Wang G R, et al. Preliminary analysis on synoptic configuration evolution and mechanism of a torrential rain occurring in Beijing on 21 July 2012[J]. Torrential Rain and Disasters, 2012,31(3):218-225.]