基于社交网络数据的交通突发事件识别方法*

刘　昭　何赏璐▲　刘英舜

（南京理工大学自动化学院 南京 210094）

摘　要：为了从社交网络数据中挖掘出交通突发事件，研究了基于机器学习的文本识别方法。通过关键词和地点定位，利用网页爬虫“Beautiful Soup”爬取到原始文本。采用正则匹配、重复度计算以及“0-1”标记预处理原始文本。基于预处理后文本特征，研究了基于特征权重的特征词选取方法；其中，特征权重的计算综合了词语的出现频率和含有该词语的文本所占比例，通过将二者归一化并加权合并，获得训练集突发事件文本中各个无重复词语的特征权重；依据此值选择确定特征词，并用于后续分类器的输入。测试对比了不同的分类器以及特征词选择方法，结果表明，所提特征词选取方法与XGBoost分类器结合，在交通突发事件识别上具有最好的综合表现，精确率为0.679 6，召回率为0.648 1，F1值为0.663 5，AUC值为0.759 4。

关键词：智能交通；社交网络数据；交通突发事件识别；文本分类；机器学习

中图分类号：U492.8　　**文献标识码：**A　　**doi:**10.3963/j.jssn.1674-4861.2021.02.007

A Method to Identify Traffic Incidents Based on Social Network Data

LIU Zhao　HE Shanglu▲　LIU Yingshun

（School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China）

Abstract: A text classification method based on machine learning is studied to identify traffic incidents by mining the data from the social networks. The original texts are crawled by web crawler “Beautiful Soup” based on the keywords and location. These texts are preprocessed using regular expression matching, duplicate removing, and “0-1” marking. According to the features of preprocessed texts, the paper proposes a method to select feature words based on feature weights. The feature weight is calculated by normalizing, weighting, and combining the word frequency and the ratio of the text containing that word. Accordingly, the feature weight of each unique word in the training set of the traffic incident text can be achieved, used as a criterion for selecting feature words, and as the inputs of classifiers. A test is conducted to compare different classifiers and methods to select feature words. The results show that the proposed method to select feature words combined with the XGBoost classifier has the optimal performance, with a precision rate of 0.679 6, a recall rate of 0.648 1, an F1 value of 0.663 5, and an AUC value of 0.759 4.

Keywords: intelligent transportation; social network data; traffic incident identification; text classification; machine learning

0　引　言

移动互联时代孕育出一批被数以亿计的用户所使用的社交网络平台，例如，QQ、微信、微博、脸书（Facebook）、推特（Twitter）等等。用户通过社交网络平台实时地分享所见所闻，其中，也蕴藏了与交通突发事件相关的信息。尽管传统的交通检测技术，如线圈、微波雷达、视频等，已在交通事件监测方面具有较多成熟的应用，而伴随高速公路路网的不断壮大，传统技术的检测范围局限等均催生了对交通事件监测技术的更新和创新的需求。用户在社交网络平台所共享的海量信息，为交通事件信息的挖掘

收稿日期：2020-06-18

* 江苏省自然科学基金项目（BK20180486）、中国博士后科学基金项目（2018M642257）、中央高校基本科研业务费专项资金（30920021140）资助

第一作者简介：刘　昭（1995—），硕士研究生.研究方向：交通信息工程及控制.Email:1486729439@qq.com

▲ 通信作者：何赏璐（1987—），博士，讲师.研究方向：智能网联交通系统.E-mail：slhemickey@126.com

提供了丰富的信息资源池，可发展作为获取交通信息的1种补充手段[1]。

近些年来，研究者们已尝试了从不同社交网络平台挖掘交通相关信息。例如，郑治豪等[2]使用网络爬虫，通过定位交通关键词，从微博中抓取到交通主题文本，从中提取了事故内容、时间、地点等关键信息。滕靖等[3]基于微博、微信和新闻客户端，提取了交通事件舆情特征，构建了交通事件的网络舆情分析系统。张恒才等[4]将微博信息进行路网匹配，采用模糊C聚类方法对微博信息进行分析，获取了所描述路段的畅通度水平。Gu等[5]聚焦于Twitter上的交通数据挖掘，通过Twitter rest API构建起了1个交通事件分类系统。D'Andrea等[6]研究了Twitter Streaming API，用以抓取实时的交通事故和拥堵等事件数据。

既有研究反映出微博、Twitter具有信息短、发布快、传播广等特点的社交网络平台是研究者们主要关注的信息挖掘对象[7]。据此，本文也将微博信息作为研究对象展开挖掘方法的研究。然而，海量的微博信息中包含了诸多干扰项，例如，某些主题与交通突发事件相似，但内容却毫不相关的干扰信息。微博文本内容在语义和形式上的复杂性，增加了文本信息挖掘的难度，也是笔者研究的1个重点问题。在既有的研究中，文本分类方法常被用于快速筛选出有效信息[8]。目前，常用的文本分类方法包括基于知识工程的人工分类法[9]、人工智能分类法等。其中，基于机器学习、深度学习等人工智能方法受到了广泛的关注，已在交通、金融、医疗等许多领域的文本信息处理中得到了应用[10-11]。在基于机器学习的方法当中，许多研究者将重点放在了特征词的选择上。宋呈祥等[12]提出1种改进卡方统计(chi-square statistic，CHI)的特征词选取方法，通过定义特征词频度分布相关性系数来提升不均衡数据集的分类指标。吴小晴等[13]提出1种改进TF-IDF的中文邮件识别算法，通过在传统的TF-IDF算法里面加入CHI和位置影响因子来改善一些重要词汇的权重。庄穆妮等[14]将LDA主题模型与Bert词向量融合，优化了主题向量的选取，在情感分类任务上，融合模型的分类性能优于单一的LDA模型。也有许多研究者关注于分类器的选择。曾奇[15]提出了基于相似度的K最近邻(K-Nearest Neighbor，KNN)算法，该算法较单一的KNN算法在微博短文本分类上的表现效果更好。柳本民等[16]以美国公路的追尾事故数据为样本，建立了基于支持向量机(support vector machine，SVM)的2车追尾事故与连环追尾事故二分类模型，结果显示，SVM模型能较好地区分2车追尾事故与连环追尾事故。李晓峰等[17]针对淘宝商品自动类目识别需求，使用了基于XGBoost的分类算法，该方法较SVM等传统的分类算法有更高的分类准确度。徐婷等[18]通过车载OBD设备获取了货车驾驶人车辆行驶数据，然后将k均值聚类分析后的结果作为分类指标来训练AdaBoost分类器，结果显示分类模型具有较高的准确率。由此看来，最佳的文本信息分类识别方法尚无明确的定论，需依据文本特征建立适合的分类模型，以实现更加准确的信息挖掘。

交通突发事件表现为多种形式，包括交通拥堵、交通事故、封路、施工等。其中，封路、施工等道路管制信息，交通运营管理部门通常会提前发布相关预告和预警信息，公众可提前做好出行规划，减少出行影响。但是，诸如交通拥堵、交通事故等不可预测和规划的信息，将影响公众出行的安全，往往是公众更为关心的问题，而公众也更愿意在微博上发布、讨论此类相关信息。考虑到交通事故相较于道路拥堵具有更大的危害性，笔者将研究的突发事件类型聚焦在交通事故，后续相关方法可向拥堵等其它突发事件进行推广。

综上，笔者以社交网络平台“微博”所发布的信息为研究对象，通过网络爬虫技术，采集到了与高速公路交通突发事件相关的文本。基于预处理后文本特征，研究了基于特征权重的特征词选取方法，该方法加强了特征词与交通突发事件文本之间的映射，从而提高这些少数文本的分类精度。考虑到特征词选择方法在不同的分类器上的分类效果的差异性，笔者也对分类器的选择进行了研究。研究的整个过程希望为交通突发事件信息的获取提供1种新的思路和方法。

1 样本集的挖掘流程

从微博文本中挖掘交通突发事件信息，本文分以下步骤开展，见图1。

步骤1。爬取相关微博文本及预处理。本文提出了基于位置搜索和关键词搜索的相关联文本爬取方法。据此，利用Python工具，网页爬虫“Beautiful Soup”从网页的HTML语言中抽取到与高速公路交通突发事件相关联的文本内容。在对文本预处理时，采用了正则匹配、重复度计算以及“0-1”标记，以提高样本集的质量，为后续处理做准备。

步骤2。自然语言处理。将样本集分为训练集和测试集。其中，训练集用于对后续文本分类器的

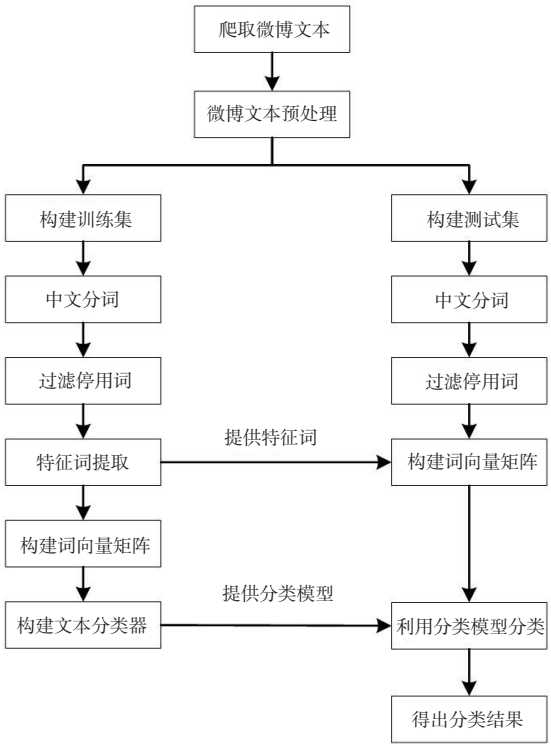


图1　研究流程图

Fig. 1　Flow for the study

标定和训练，而测试集用于验证标定后分类器对交通突发事件识别的效果。对于训练集，自然语言处理过程包括中文分词、过滤停用词、特征权重计算、特征词选取。而对于测试集，特征权重计算、特征词选取已完成，故这2个步骤可省略。

步骤3。文本分类器的构建、对比与测试。本文分别构建了基于SVM，KNN，AdaBoost和XGBoost算法的分类器，并将训练集的词向量输入到各个分类器中进行训练，调节4种分类器参数以获得最优性能，最后，利用测试集对比分析4个分类器识别交通突发事件信息的效果。

2　样本集的获取

2.1　微博文本爬取

许多学者在研究社交网络交通类文本分类的时候，研究对象常聚焦于全国的交通状况，优势是信息来源广泛，更容易获取到样本集[19]。但是驾驶人或交通管理者往往关注于自身所处区域的交通情况，而其他区域的交通信息会对驾驶人或交通管理者的判断产生干扰。在挖掘文本内容时，笔者通过限定搜索范围，将所要研究的区域精确定位到江苏省上，实现了交通突发事件在位置方位上的更精细化研究。采用“高速&事故”组合词对突发事件进行爬取，这个组合词既能表征研究的范围是“高速公路”，又能表征研究的对象是“交通事故”，相较于其它的词组“事故”“高速”等，减少了多次重复爬取的过程，具有更强的针对性。通过这种方式，笔者从微博平台中爬取了2019年全年和2020年6月—12月的相关微博文本信息（2020年上半年因疫情影响，多数高速公路处于管制状态，因此，不作为爬取的时间段）。考虑到样本集的不均衡特性，没有划分验证集，将2019年的样本作为训练集，2020年的样本作为测试集。

在微博信息爬取过程中，笔者使用了网页爬虫的“Beautiful Soup”。“Beautiful Soup”是1种可以从网页HTML语言中快速提取文本内容的Python库，而Python中的Requests库可以方便地爬取到网页的HTML语言。具体使用时，通过“Beautiful Soup”遍历标签名、属性等提取文本信息。

2.2　样本集预处理

爬取到相关信息后，在构建样本集之前，需对原始数据进行清洗，以提升数据质量，具体流程见图2。

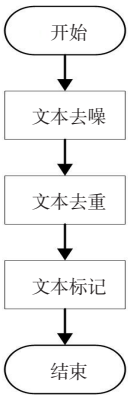


图2　样本集预处理流程图

Fig. 2　Flow of preprocessing the sample set

样本集的预处理主要包括：①文本去噪。包括文本格式上的噪声（“\n”“\t”“空格”等）和字符噪声（一些难以通过停用词库过滤的特殊字符等），采用正则匹配的方法，通过“re.compile(u"[^a-zA-Z0-9\u4E00-\u9F-A5]")”命令来去噪；②文本去重，即去掉重复度高的文本。以其中2个文本为例说明：首先统计2个文本中重复出现的汉字总数，然后分别计算重复数占2个文本总字数的百分比，即重复度，如果文本重复度达到80%，则删除字数少的文本。③文本标记。本文采用“0-1”方式来标记训练集样本，其中，“1”代表文本内容与交通突发事件相关（以下简称相关文本），“0”代表文本内容与交通突发事件无关（以下简称非相关文本）。非相关文本中主要包括3类：①与实际交通事故发生没有关联的文本；②官方事故文本；③非实时性文本（一般来说，这些文

本中会出现表征过去时间状态的特征)。表1展示了样本标记示例。

表1 样本“0-1”标记示例

Tab. 1 Cases of “0-1” labeled samples

标记	文本样本
0	高速公路的设计,忌讳一直都笔直向前,哪怕是大平原上,危险容易造成事故
1	@江苏交警警察同志,请问,在图1沪宁高速路段发生交通事故,一些车辆(图2~7)可以“借用”应急车道通行吗? 如果可以,下次,我们都这样做!

样本集预处理后,得到2019年的样本,共计1 350条,其中标记为“1”的有345条;2020年的样本共计726条,其中标记为“1”的有216条。图3~4为2019年全年和2020年下半年的相关文本分布情况。

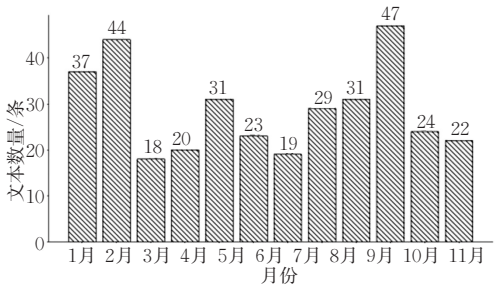


图3 2019年交通突发事件文本数量

Fig. 3 Number of texts on traffic emergencies in 2019

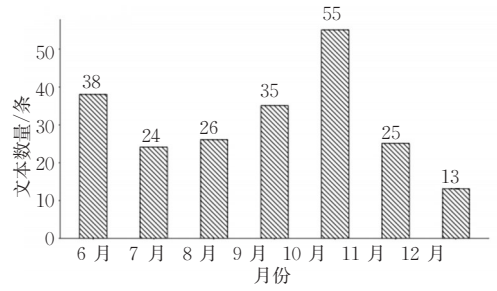


图4 2020年交通突发事件文本数量

Fig. 4 Number of texts on traffic emergencies in 2020

从2019年和2020年样本可看出,1月、2月和10月是交通突发事件文本数量较高的时期,可能是受春节假期和国庆假期高峰出行的影响,高速公路上的交通事故数量增加,加大了人们舆论的力度。

3 样本集的自然语言处理

本文构建的自然语言处理方法主要包括分词和过滤停用词、基于特征权重的特征词选取。通过分词和过滤停用词,提高待处理文本的质量;在此技术上,提出1种基于特征权重的特征词选取方法,通过加强特征词与相关文本间的映射关系,提高相关文本分类的准确性。

3.1 分词和过滤停用词

笔者对训练集中的样本进行了分词和过滤停用词。在分词的过程中,对比了Jieba分词工具和LTP分词工具。Jieba是Python中的中文分词库,因其对中文文本良好的分词功能,广泛地用于中文文本分词;LTP是哈尔滨工业大学开源的1套中文语言处理系统,因其在文本分词中较好的处理性能,被广泛应用于各大比赛中。根据后续的测试效果,Jieba分词速度远快于LTP分词速度,但LTP分词性能要优于Jieba,故本文最终选择了LTP分词工具。经过分词后,训练集中仍存在着大量的无意义词汇,被称为停用词,例如“的”“然后”“呀”等。笔者选用哈工大中文停用词库来对停用词进行过滤。以某条相关文本举例,经分词和停用词过滤后的形式见表2。

表2 分词和过滤停用词的结果示例

Tab. 2 A case for segmenting words and filtering stop words

原文本	@江苏交警警察同志,请问,在图1沪宁高速路段发生交通事故,一些车辆(图2~7)可以“借用”应急车道通行吗? 如果可以,下次,我们都这样做!
分词和过滤停用词处理后的文本	江苏/交警/警察/同志/请问/图/1/沪宁/高速/路段/发生/交通/事故/车辆/图/27/借用/应急/车道/通行/下次/都/做/

3.2 基于特征权重的特征词选取方法

传统的特征词选择方法,比如TF-IDF、卡方检验、LDA主题模型等,均同时关注样本集的不同类别的特征,但当样本集类别数目不均衡时,这些方法处理效果往往欠佳。训练集上,“1”和“0”样本数量比例为1:3左右,样本有较强的非均衡性。因此,本文建立了1种基于特征权重的特征词选取方法,通过将注意力全部聚焦在相关文本,以建立相关文本和特征词之间的精确映射,而不考虑非相关文本特征词的影响。该方法的计算步骤如下。

步骤1。从训练集共计345条相关文本(已过滤停用词)中提取到词语集合,共9 312,去除重复词语后,共2 311。

步骤2。通过式(1)~(2)计算词语出现频率的最大值和最小值;通过式(3)~(4)计算含有某个词语的文本数占总文本比例的最大值和最小值。

$$N_{\max}=\max\left\{\frac{n_1}{n},\frac{n_2}{n},\cdots,\frac{n_i}{n}\right\} \tag{1}$$

$$N_{\min}=\min\left\{\frac{n_1}{n},\frac{n_2}{n},\cdots,\frac{n_i}{n}\right\} \tag{2}$$

$$D_{\max}=\max\left\{\frac{d_1}{d},\frac{d_2}{d},\cdots,\frac{d_i}{d}\right\} \tag{3}$$

$$D_{\min}=\min\left\{\frac{d_1}{d},\frac{d_2}{d},\cdots,\frac{d_i}{d}\right\} \tag{4}$$

式中：n_i 为第 i 个词语在词语集出现的总次数；n 为词语集总词数；$N_{\max}$ 为词频序列的最大值；$N_{\min}$ 为词频序列的最小值；d_i 为含有第 i 个词语的文本数量；d 为文本总数；$D_{\max}$ 为文本比例序列的最大值；$D_{\min}$ 为文本比例序列的最小值。

通过式(5)计算每个词语的特征权重。

$$W_i=w\frac{\frac{n_i}{n}-N_{\min}}{N_{\max}-N_{\min}}+(1-w)\frac{\frac{d_i}{d}-D_{\min}}{D_{\max}-D_{\min}} \tag{5}$$

式中：W_i 为第 i 个词语的特征权重；w 为权重因子。

步骤3。将 W_i 序列降序排序，根据 W_i 依次选取特征词。后续通过选取不同的权重因子和特征词数量来测试分类器分类性能。

步骤4。构建词向量矩阵。以某1条文本举例：如果特征集合中的特征词出现在了文本中，则将相对应特征词的位置上赋值该特征词的特征权重。最终构建的训练集的词向量矩阵为[1 350，特征词数量]，测试集的词向量矩阵为[726，特征词数量]。

4　样本集的分类器构建

4.1　分类器构建

本文分别建立了基于KNN，SVM，AdaBoost，XGBoost的文本分类器。KNN算法无数据输入假定，k 的大小会影响分类精度，但 k 值选取没有1个确定的标准；SVM算法可以解决非线性的分类任务，但对参数和核函数的选择较敏感；AdaBoost和XGBoost属于集成学习，均基于Boosting算法，其中AdaBoost可以将不同的分类算法作为弱分类器，很好的利用了弱分类器进行级联运算，但是数据不平衡会导致分类精度下降；XGBoost可以采取并行优化策略，它加入了正则项，可以降低过拟合，XGBoost也提供了调节样本不均衡特性的参数，但XGBoost需要调试的参数比较多。综合来看，在分类器选择上，并没有1个确定的标准。本文利用训练集构成的词向量矩阵对各个分类器进行训练，以获得最优的分类性能。

4.2　分类器性能评估

本文选取了精确率（P）、召回率（R），$F1$ 值（$F1$-Score）和 AUC 值4项常用指标作为分类器评估标准。式(6)～(8)展示了前3项指标的计算公式。

$$P=\frac{TP}{TP+FP} \tag{6}$$

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$F1\text{-Score}=\frac{2P\times R}{P+R} \tag{8}$$

式中：TP 为标记为“1”且被分类为“1”的文本数量；FN 为标记为“1”但被分类为“0”的文本数量；TN 为标记为“0”且被分类为“0”的文本数量；FP 为标记为‘0’但被分类为‘1’的文本数量。

精确率越高，代表分类器识别出的所有相关文本中，真实值也是“1”的比例越高；而召回率越高，表示了能有更多真实值是“1”的文本被识别出来，而遗漏的越少。$F1$ 值作为综合指标，可以平衡精确率和召回率的影响，较为全面地评价1个分类器。AUC 值定义为ROC曲线下方的面积。AUC 值适合评价样本不平衡中的分类器性能。AUC 越大，表示分类器性能越好。

4.3　分类器分类性能测试

本文通过训练集构建出的分类器来测试测试集上的文本分类性能。测试集上，将其他3种传统特征词选择方法作了对比。表3～6展示了不同组合下的相关文本的分类指标结果。

表3～6的对比结果证实了所提基于特征权重的特征词选取方法的有效性，相较于其它方法，所提特征词选取方法与XGBoost结合的方法综合表现最

表3　精确率对比

Tab. 3　Comparison of precision rates

	KNN	SVM	AdaBoost	XGBoost
新方法	0.583 9	0.647 4	0.720 5	0.679 6
TF-IDF	0.528 2	0.730 2	0.640 4	0.594 2
CHI	0.655 2	0.646 7	0.655 4	0.625 8
LDA	0.397 0	0.527 3	0.402 9	0.455 1

表 4 召回率对比

Tab. 4 Comparison of recall rates

	KNN	SVM	AdaBoost	XGBoost
新方法	0.370 4	0.518 5	0.537 0	0.648 1
TF-IDF	0.347 2	0.425 9	0.601 9	0.569 4
CHI	0.263 9	0.449 1	0.537 0	0.472 2
LDA	0.365 7	0.268 5	0.384 3	0.375 0

表 5 *F*1 值对比

Tab. 5 Comparison of *F*1 values

	KNN	SVM	AdaBoost	XGBoost
新方法	0.453 3	0.575 8	0.615 4	0.663 5
TF-IDF	0.419 0	0.538 0	0.620 5	0.581 6
CHI	0.376 2	0.530 1	0.590 3	0.538 3
LDA	0.380 7	0.355 8	0.393 4	0.411 2

表 6 *AUC* 值对比

Tab. 6 Comparison of *AUC* values

	KNN	SVM	AdaBoost	XGBoost
新方法	0.629 3	0.699 5	0.724 4	0.759 4
TF-IDF	0.607 9	0.679 6	0.729 4	0.702 3
CHI	0.602 5	0.672 6	0.708 7	0.676 3
LDA	0.565 2	0.583 3	0.571 5	0.592 4

优，具有最高的召回率值 0.6481，最高的 *F*1 值 0.663 5，以及最高的 *AUC* 值 0.759 4。表 7 为在最优组合下的参数设置情况（未列出的参数采用默认值）。

表 7 参数设置

Tab. 7 Parameter settings

特征词选择	*w*	0.5
	特征词数量	150
XGBoost	*max_depth*	5
	booster	gbtree
	objective	binary:logistic
	scale_pos_weight	3
	min_child_weight	1
	learning_rate	0.1
	n_estimators	100

笔者也探讨了不同特征词数量对分类指标的影响，结果见图 5。测试表明当特征词数量为 150 时，

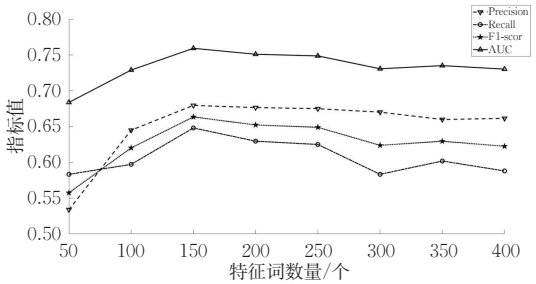


图 5 不同特征词数量下的指标对比

Fig. 5 Comparison of indices under different eigenvalues

分类器取得最优性能。

图 6 展示了由 150 个特征词构建出的词云图。其中，字号越大，代表了特征词的特征权重越大。

图 6 词云图

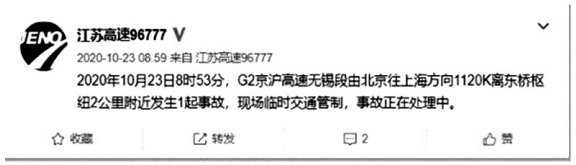
Fig. 6 Word cloud

5 案例分析

笔者随机选取了 2020 年 10 月的相关文本来进一步分析，共计约 55 条相关文本。与官方发布信息进行对比，比对成功共计 16 起。图 7 为其中 1 个匹配成功的案例。其中，图 7(a)是识别出的交通突发事件文本，图 7(b)是与之相应的官方报道。某用户于 08:00 发布的 1 条微博信息显示在 G2 京沪高速上发生交通事故，官方微博于 08:59 才发布了该事件的相关报道。可以看出，普通用户在相对早的时间发布了事故信息，而官方微博发布时间相对晚，可能是官方需要确认交通事故信息后才对外发布。然而，仍有 39 条交通突发事件信息未匹配到相关的官方信息，潜

在的原因可能包括:①部分信息并不含有关于事件的具体地点定位;②在同1个时段内,部分信息定位到的地点无法与官方信息中提供的地点匹配。

(a) 识别结果



(b) 官方报道

图7　交通事故识别与官方报道对比

Fig. 7　Comparison between traffic accident identification and official reports

综上所述,所提出的交通突发事件识别方法可从社交网络平台微博文本中有效地挖掘出交通突发事件信息。由于微博信息发布的实时性特点,从微博中挖掘出的信息,可为实时的交通突发事件监测提供1种新的信息获取方式。然而,由于用户发布的信息在可靠性和严谨性等方面有所欠缺,因此,基于社交网络数据的交通突发事件识别目前仅能作为1种辅助方法。在获取识别结果后,仍需其他监测方法对结果进行核实。

6　结束语

为了实现从社交网络平台"微博"中挖掘出高速公路交通突发事件信息,提出了基于机器学习的文本识别方法,具体包括文本信息的爬取、预处理、分词和过滤停用词、特征权重计算、特征词选取及分类器构建。通过构建不同的分类器进行测试,并与其他3种传统特征词选择方法对比,论证了基于特征权重的特征词选取方法的有效性。该方法在XGBoost分类器上具有最高的分类性能。在案例分析时,笔者发现,通过所提出的挖掘方法识别出的高速公路交通突发事件,发布时间上会早于官方发布,在一定程度上可辅助高速公路交通运营和管理的相关单位和部门实现交通事件的监测。在未来的研究中,研究拟从以下方面继续深化:①提高文本识别的分类器性能;②考虑融合图像数据来做交通突发事件的识别,例如某些用户倾向于发布图片来描述交通突发事件,而非文本说明。

参考文献

References

[1] QIAO F X, YU L. Social media applications to publish dynamic transportation information on campus[C]. International Conference of Chinese Transportation Professionals, Nanjing, China: ICCTP, 2011.

[2] 郑治豪,吴文兵,陈　鑫,等.基于社交媒体数据的交通感知分析系统[J].自动化学报,2018,44(4):656-666.
ZHENG Zhihao, WU Wenbing, CHEN Xin, et al.A traffic sensing and analyzing systemusing social media data[J]. Acta Automatica Sinica, 2018, 44(4): 656-666. (in Chinese)

[3] 滕　靖,刘韶杰,龚　越,等.交通事件网络舆情分析方法[J].交通信息与安全,2019,37(6):139-148.
TENG Jing, LIU Shaojie, GONG Yue, et al. An analysis method of online public opinions on traffic incidents[J]. Journal of Transport Information and Safety, 2019, 37(6): 139-148. (in Chinese)

[4] 张恒才,陆　锋,陈　洁.微博客蕴含交通信息的提取[J].中国图象图形学报,2013,18(1):123-129.
ZHANG Hengcai, LU Feng, CHEN Jie.Extracting traffic information from massive microblog messages[J]. Journal of Image and Graphics, 2013, 18(1): 123-129. (in Chinese)

[5] GU Y M, QIAN Z, CHEN F. From twitterto detector: realtime traffic incident detectionusing social media data[J]. Transportation Research Part C: Emerging Technologies, 2016(67): 321-342.

[6] D'ANDREA E, DUCANGE P, LAZZERINI B, et al. Real-time detection of traffic from twitter stream analysis[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 2269-2283.

[7] 徐　翔,刘　悦.全球社交网络中用户"社会互动位置—信息位置"同质效应研究——基于Twitter信息传播的数据挖掘和实证分析[J].华东理工大学学报(社会科学版),2019,34(5):92-102.
XU Xiang, LIU Yue. Research on the homogenouseffect of"social interaction locationinformation location"of the users in the global social networks: Data mining and empirical analysis based on twitter informationdissemination[J].Journal of East ChinaUniversity of Science and Technology (Social Science Edition), 2019, 34(5): 92-102. (in Chinese)

[8] 叶颖婕.基于关联规则的交通事故风险因素挖掘及预测模型构建[D]. 北京:北京工业大学,2018.
YE Yingjie. Research on mining algorithm and prediction mod-

el of traffic accident risk factors based on news data[D]. Beijing:Beijing University of Technology,2018. (in Chinese)

[9] 胡泽文,王效岳,白如江.国内外文本分类研究计量分析与综述[J].图书情报工作,2011,55(6):78-81+142.
HU Zewen,WANG Xiaoyue,BAI Rujiang. Quantitative Analysis and review of text classification research at home and abroad [J]. Library And Information Service,2011,55(6):78-81+142. (in Chinese)

[10] SALAS A, GEORGAKIS P, PETALAS Y.Incident detection using data from social media[C]. 20th International IEEE Conference on Intelligent Transportation Systems, yokohama, Janpan:IEEE,2017.

[11] SAKAKI T, MATSUO Y, YANAGIHARAT, et al. Realtime event extraction for drivinginformation from social sensors[C]. International IEEE Conference Cyber Technology in Automation, Control, and Intelligent Systems, Bangkok, Thailand: IEEE,2012.

[12] 宋呈祥,陈秀宏,牛 强.文本分类中基于CHI改进的特征选择方法[J].微电子学与计算机,2018,35(9):74-78.
SONG Chengxiang,CHEN Xiuhong,NIU Qiang. Improved feature selection methodbased on chi for text categorization[J]. Microelectronics & Computer,2018,35(9):74-78. (in Chinese)

[13] 吴小晴,万国金,李程文,等.一种改进TF-IDF的中文邮件识别算法研究[J].现代电子技术,2020,43(12):83-86.
WU Xiaoqing,WAN Guojin,LI Chengwen,et al. Research on improved TF-IDF Chinese mail recognition algorithm[J]. Modern Electronics Technique,2020,43(12):83-86. (in Chinese)

[14] 庄穆妮,李 勇,谭 旭,等.基于BERT-LDA模型的新冠肺炎疫情网络舆情演化仿真[J].系统仿真学报,2021,33(1):24-36.
ZHUANG Muni,LI Yong,TAN Xu,et al. Evolutionary simulation of online public opinion based on the BERT-LDA model under COVID-19[J]. Journal of System Simulation, 2021, 33(1):24-36.(inChinese)

[15] 曾 奇.面向微博的短文本分类算法研究[D].成都:电子科技大学,2019.
ZENG Qi. Research on short text classification algorithms for Microblog[D]. Chengdu: University of Electronic Scienceand Technology of China,2019. (in Chinese)

[16] 柳本民,闫 寒.基于SVM事故分类的连环追尾事故影响因素分析[J].交通信息与安全,2020,38(1):43-51.
LIU Benmin, YAN Han. An analysis of influencing factors of multi-vehicle rear-end accidentsbased on accident classification of SVM[J]. Journal of Transport Information and Safety, 2020,38(1):43-51. (in Chinese)

[17] 李晓峰,马 静,李 驰,等.基于XGBoost模型的电商商品品名识别算法研究[J].数据分析与知识发现,2019,3(7):34-41.
LI Xiaofeng, MA Jing, LI Chi, et al. Identifying commodity names based on XGBoost model[J]. Data Analysis and Knowledge Discovery,2019,3(7):34-41. (in Chinese)

[18] 徐 婷,张 香,张亚坤,等.基于AdaBoost算法的货车驾驶人安全倾向性分类[J].安全与环境学报,2019,19(4):1273-1281.
XU Ting,ZHANG Xiang,ZHANG Yakun,et al. Truck driver safety tendency classification based on the AdaBoost algorithm [J]. Journal of Safety and Environment, 2019, 19(4): 1273-1281. (in Chinese)

[19] 尹何举,咎红英,陈俊怡,等.交通事故的自动判案研究[J].中文信息学报,2019,33(3):136-144.
YI Heju, ZAN Hongying, CHEN Junyi, et al. Study on automatic judgment of traffic accidents[J]. Journal of Chinese Information Processing,2019,33(3):136-144. (in Chinese)