引用格式：仇培元,陆锋,张恒才,等.蕴含地理事件微博客消息的自动识别方法[J].地球信息科学学报,2016,18(7):886-893. [Chou P Y, Lu F, Zhang H C, *et al*. 2016. Automatic identification method of micro-blog messages containing geographical events. Journal of Geo-information Science, 18(7):886-893.] **DOI:**10.3724/SP.J.1047.2016.00886

蕴含地理事件微博客消息的自动识别方法

仇培元[1,2],陆　锋[1],张恒才[1*],余　丽[1,2]

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101；2. 中国科学院大学,北京 100101

Automatic Identification Method of Micro-blog Messages Containing Geographical Events

QIU Peiyuan[1,2], LU Feng[1], ZHANG Hengcai[1*] and YU Li[1,2]

1. State Key Lab of Resources and Environmental Information System, IGSNRR, CAS, Beijing 100101, China;

2. University of Chinese Academy of Sciences, Beijing 100101, China

Abstract: Micro-blogs usually contain abundant types of geographical event information, which could compensate for the shortcomings of traditional fixed point monitoring technologies and improve the quality of emergency response. Identify the micro-blog messages that containing the geographical event information is the prerequisite for fully utilizing this data source. The trigger-based and the supervised machine learning methods are commonly adopted to identify the event related texts. Comparatively, the supervised machine learning methods have better performance than the trigger-based ones for unrestricted texts. Unfortunately, the lack of large-scale tagged corpuses cause the supervised machine learning methods cannot be implemented to identify the geographical event related messages. In this paper, we propose an automatic method for recognizing micro-blogs that are related to geographical events based on the topic model and word vector. This method could achieve a satisfying identification result by increasing the corpus scale rapidly. Firstly, the topic model is capable to extract topics from documents. Thus, the web pages fetched by a search engine are grouped by the topics, and the corpus is obtained after combining the pages under the topics that are related to geographical events through judging their keywords of each topic. Secondly, the distributed representation word vector model is introduced to compensate the lack of context in the micro-blog, which is caused by its character count limit. These word vectors are integrated into the context semantic information from corpus training during the vector generation process. Thirdly, the correlation between the micro-blog message and the given geographical event is calculated and applied to determine whether this message contains the specified geographical event or not. In addition, some heuristic rules are used to correct the error correlations of very short messages. Experiments where the rainstorm is set as the targeting geographical event are conducted to validate the feasibility of this approach. The test conducted on Sina topic micro-blog shows that the F-1 of identification reaches 71.41% and is 10.79% higher than the traditional machine learning algorithm based on Support Vector Machine. Based on the premise that the precision loss is limited, the recall rate would rise with an increase in the corpus scale. The recognition precision could achieve 60% in a dataset containing five million micro-blog texts that simulating the actual data content and environment. These recognized event related micro-blogs could be used to extract detailed information elements in the future.

Key words: micro-blog; geographical event; event text identification; topic model; word vector

***Corresponding author:** ZHANG Hengcai, E-mail: zhanghc@lreis.ac.cn

摘要：微博客文本蕴含类型丰富的地理事件信息,能够弥补传统定点监测手段的不足,提高事件应急响应质量。然而,由于大

收稿日期：2015-09-07；**修回日期：**2015-11-03.

基金项目：国家“863”计划课题(2013AA120305)；国家自然科学基金项目(41401460)。

作者简介：仇培元(1986-),男,博士生,研究方向为互联网空间信息搜索。E-mail: qiupy@lreis.ac.cn

***通讯作者：**张恒才(1985-),男,博士后,研究方向为互联网空间信息搜索,轨迹数据管理与数据挖掘。
E-mail: zhanghc@lreis.ac.cn

规模标注语料的普遍匮乏，无法利用监督学习过程识别蕴含地理事件信息的微博客文本。为此，本文提出一种蕴含地理事件微博客消息的自动识别方法，通过快速获取的语料资源增强识别效果。该方法利用主题模型具有提取文档中主题集合的优势，通过主题过滤候选语料文本，实现地理事件语料的自动提取。同时，将分布式表达词向量模型引入事件相关性计算过程，借助词向量隐含的语义信息丰富微博客短文本的上下文内容，进一步增强事件消息的识别效果。通过以新浪微博为数据源开展的实验分析表明，本文提出的蕴含地理事件信息微博客消息识别方法，识别来自事件微博话题的消息文本的F-1值可达到71.41%，比经典的基于SVM模型的监督学习方法提高了10.79%。在模拟真实微博环境的500万微博客数据集上的识别准确率达到60%。

关键词：微博客；地理事件；事件文本识别；主题模型；词向量

1 引言

近年来，随着智能终端和移动互联网的普及，位置服务应用不断增长，与空间位置密切相关的地理事件信息成为人们日常生活关注的焦点，如“南方雪灾”、“721暴雨”、台风登陆等。这类地理事件导致的城市内涝、路面积水、设施垮塌等现象极易影响周边人群正常生活。然而，上述现象发生的空间位置具有不确定性，难以及时被传统定点监测手段发现，使应急响应滞后。与此同时，社会化网络媒体的参与度高、双向交流、人人生产内容、公开共享、社区化和多媒体化等特征[1]，使其成为人们信息交流与分享的重要渠道。其中，微博客平台具有更强的开放性和时效性，成为事件信息快速传播的重要媒介，针对地理事件状态的描述信息也不断出现在微博客消息文本中。因此，抽取微博客消息蕴含的地理事件信息，能够进一步补充和完善地理事件影响的时空范围和实时状态，改善职能机构决策和公众信息服务的质量。

地理事件信息抽取属于自然语言处理的事件信息抽取任务，包括事件文本识别和事件属性提取。其中，事件文本识别是从候选文本集合中筛选出目标事件相关文本，是开展事件属性提取的基础。因此，识别出蕴含地理事件的微博客文本，可以提高后续地理事件属性提取的效率。事件信息文本的识别方法主要有基于触发词过滤的方法和基于监督学习的方法：前者依据是否含有事件触发词来判断文本是否为事件相关文本；后者基于给定的学习特征，通过标注语料训练机器学习模型，从而自动识别事件文本，常见模型有最大熵模型[2]、贝叶斯分类器[3]、KNN[4]、支持向量机[5-6]等。基于触发词过滤的方法没有考虑触发词与上下文之间的语义联系，易将含有触发词的无关文本识别为相关文本。因此，该方法主要用于处理事件相关性较强的文本[7]。对于内容自由度更高的开放文本，监督学习识别方法的效果更好，但监督学习方法需要足够的标注语料进行模型训练，大部分研究使用的语料主要来源于开放的测评语料[4,8]或人工标注语料[2,9]。然而，现阶段缺乏大规模开放的地理事件标注语料，人工标注语料需要的人力和时间成本较高，且不同的地理事件需要不同的语料资源。因此，当前蕴含地理事件信息文本识别，主要通过关键词搜索或匹配的方式实现，如从搜索结果、新闻报道或来源较为固定的微博客文本中获取自然灾害[10-11]、道路交通[12-13]和地理要素变化[14]等事件。监督学习方法则针对网络新闻报道，通过小规模人工标注语料实现[15]，无法适用于内容自由的开放微博客文本。

基于以上现状，本文提出一种蕴含地理事件微博客消息的自动识别方法，借助主题模型和分布式表达词向量模型提高语料处理的自动化程度，通过扩大语料规模弥补和改善非监督学习方法在识别精度方面的损失。

2 识别方法

蕴含地理事件微博客消息的识别是从微博客集合中识别出含有与目标类型一致的地理事件的消息文本。在信息检索领域，地理事件主要指发生于地表空间的各种自然和社会现象，由时间、空间位置和事件现象组成[16]，其中事件现象描述是判断事件类型的重要参考。因此，本文识别方法首先利用事件关键词采集地理事件信息候选语料，之后借助主题模型从候选语料中提取事件相关文本。微博客消息和普通网页文本均可通过关键词搜索方式得到。其中，微博搜索主要基于简单关键词匹配实现，由于微博客内容随意性较强，口语化程度较高，返回结果易掺杂地理事件无关信息，影响语料提取效果。而通用搜索引擎则会对搜索结果进行

优化,使搜索结果与关键词具有较强的相关性,利于事件语料的快速采集。此外,虽然微博客消息短文本与新闻报道等网络长文本之间存在差异[17],但对事件的描述仍会使用基本语法结构,如主谓、谓宾结构。因此,本文识别方法将普通网页文本作为语料来源,以词和词法为基本单元,在识别过程中将长文本语料应用于微博客短文本,从而完成对蕴含地理事件微博客消息的识别。识别方法主要流程为:(1)利用主题模型提取地理事件语料和事件核心词集合;(2)由地理事件语料构建词向量集合;(3)计算待识别微博客消息与地理事件的相关度,筛选出蕴含地理事件的微博客消息。具体流程如图1所示。

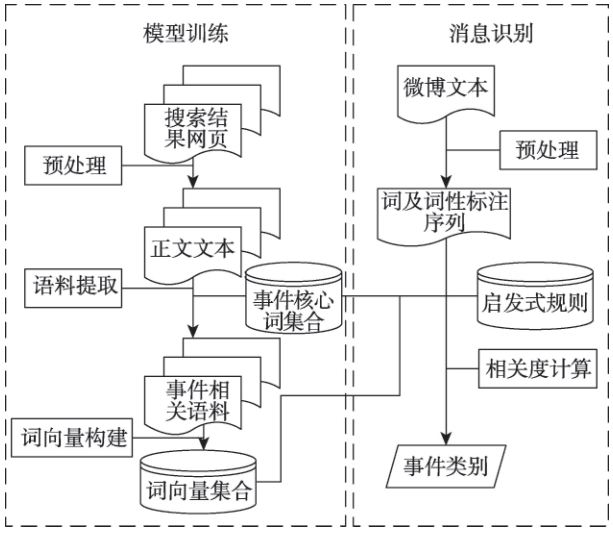

图1 识别方法流程

Fig.1 Flowchart of the identification method

2.1 语料提取

使用爬虫抓取关键词搜索结果指向的网页页面并解析页面正文,得到目标地理事件候选语料。该候选语料仍含有与地理事件关联较弱的信息类型,如应急指南、纪念报道等,需进一步提取事件相关文本。

由于网页文本集合中的信息类别未知,并且缺乏标注,监督方法难以应用,因此使用无监督的主题模型筛选地理事件相关文本。主题模型能够从文档集合中根据语义联系生成主题集合,并获取各文档的主题概率分布,以及各主题的词项概率分布。因此,利用主题模型可以将候选语料按主题划分,并根据各主题的高概率关键词集合筛选地理事件相关主题,从而得到地理事件文本语料。其中,浅层狄利克雷分布(Latent Dirichlet Allocation,LDA)是应用最广泛的主题模型之一,当前大多数主题模型研究均与之有关[18]。LDA是由Blei等提出[19],是在概率隐性语义索引(Probabilistic Latent Semantic Indexing,PLSI)基础上扩展的三层贝叶斯概率模型。模型假设文档中的每一个词都是由"一定概率选择了某个主题,并以一定概率从该主题中选择了某个词"的生成过程得到,且2个概率均服从Dirichlet分布。由于在不预先进行人工判读的情况下,难以确定网页文本集合包含的信息类别,因此研究选择层次LDA(Hierarchical LDA,HLDA)模型提取目标地理事件相关文本。HLDA由Blei等在LDA模型基础上改进,用于建立主题之间的树状层次关联,并能通过CRP(Chinese Restaurant Process)自动估计每一层的主题数量[20]。

具体步骤为:首先,对抓取的网页文本分词并去停用词;然后,利用HLDA提取候选语料中隐含的主题类别及各主题对应的关键词;最后,根据主题关键词,人工判读各主题是否与目标地理事件相关,保留相关主题下的文本作为目标地理事件语料,同时合并相关主题关键词作为目标地理事件核心词集合。图2以暴雨事件为例,展示了主题提取和相关性判读的结果。其中,左侧为主题提取结果,每行代表一个主题,由主题关键词组成;右侧为相关性判读结果,即根据主题关键词判断该主题是否与暴雨事件相关。

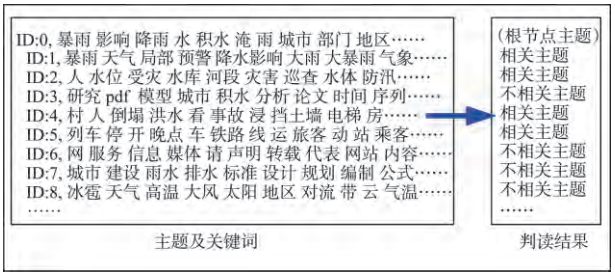

图2 暴雨事件候选语料的主题提取及相关性判读示意

Fig.2 An example of topic extraction and correlation interpretation from the candidate corpuses about rainstorm

2.2 词向量构建

提取的地理事件核心词是描述地理事件发生和状态的重要词汇,而其它非核心词也对地理事件具有指示作用。因此,在微博客消息识别过程中融入核心词与非核心词之间的语义联系,有助于提高消息识别的效果。

在自然语言的计算过程中,通常利用词向量表示词语。分布式表达模型(Distributed Representa-

tion)在词向量生成过程中,融入词汇在语料文本中的上下文语义,使其拥有比较词间语义相似性或相关性的能力。Bengio等在提出的神经网络语言模型(Neural Network Language Model, NNLM)中引入分布式表达词向量[21],成为后续相关研究的基础。其后,Mikolov等提出的CBOW和Skip-gram模型[22]则去除NNLM中的神经网络隐含层,通过损失一部分准确率以大幅提高模型训练效率。其中,Skip-gram模型输出结果在语义相似性计算上的效果较好,因此基于该模型由事件相关语料构建词向量集合。图3为词汇相关性计算结果示例,每列下方的列表是与首词相关性最高的10个词。

暴雨	水	道路	水位	交通
席卷	浸	桥梁	上涨	瘫痪
大暴雨	一头	场站	警戒	中断
趋于	教室	环路	江水	几近
时隔	泡	塌方	大运河	拥堵
沉重	涨	中断	限	改作
阵风	台阶	受损	突破	管制
雷暴	漫	交	干流	路段
雷电	脚踝	国道	流量	轨道
来临	海景	民房	水库	环路
入夏	汹涌	辅	支流	客流

图3　基于词向量的相关词计算结果示例

Fig.3　An example of related words computation based on the word vector

2.3 事件消息识别

利用构建的事件核心词集合及词向量集合,识别微博客消息文本是否蕴含给定类型的地理事件信息。

2.3.1 相关度计算

首先,对微博客消息文本作分词和词性标注预处理后,依次计算文本各词与各核心词之间的相关度,选择相关度最大的核心词作为该词的最相关核心词。然后,将所有最相关核心词的平均值作为微博客消息文本与目标地理事件的相关度。

若词 w_i 与 w_j 的词向量分别为 $w_i=\left[vec_i^1,vec_i^2,\cdots,vec_i^k,\cdots,vec_i^n\right]$, $w_j=\left[vec_j^1,vec_j^2,\cdots,vec_j^k,\cdots,vec_j^n\right]$,则基于夹角余弦的词间相关度 $rel_{\text{word}}(w_i,w_j)$ 可通过式(1)计算。

$$rel_{\text{word}}\left(w_i,w_j\right)=\frac{\sum_{k=1}^{n}vec_i^k\cdot vec_j^k}{\sqrt{\left(\sum_{k=1}^{n}\left(vec_i^k\right)^2\right)\cdot\left(\sum_{k=1}^{n}\left(vec_j^k\right)^2\right)}} \quad (1)$$

若微博客消息文本为 $text=\{w_1,w_2,\cdots,w_k,\cdots,w_n\}$,核心词集合为 $dic=(keyw_1,keyw_2,\cdots,keyw_g,\cdots,keyw_m)$,则微博客消息文本与目标地理事件 *topic* 的相关度 $rel_{\text{event}}(text,topic)$ 计算公式如式(2)所示。

$$rel_{\text{event}}(text,topic)=\frac{\sum_{k=1}^{n}\max_{1\le g\le m}(rel_{\text{word}}(w_k,keyw_g))}{n} \quad (2)$$

2.3.2 启发式规则约束

(1)词法约束

微博客消息描述随意性较强,易出现字数过少的文本,由于缺失上下文信息,文本内容的意义不明,如“严重关切”、“表示愤慨”等。同时,若文本参与相关度计算的词语较少,出现的核心词会导致相关度偏高,造成事件类型识别错误,因此需要对文本进行约束。

虽然因字数限制,微博客消息文本与普通长文本在语言描述上存在差异,但在表达内容的过程中仍会符合基本的句法规则(如“主谓宾”、“谓宾”等),因此可以利用这些语言规则对识别结果加以约束。由于目前中文句法分析在实际应用中表现不佳[23],研究用词法规则代替句法规则,如“名词–动词–名词”、“动词–名词”等词性模式。

词法规则基于网络开放事件信息标注语料(http://www.datatang.com/data/44588)统计得到。该语料包含3000篇新闻报道文本,标注了事件发生的时间、地点、人物、内容、过程等信息。选取语料中标注为事件内容(what)和过程(how)的文本,经分词和词性标注处理,统计不同词性模式出现的次数,提取出现次数超过10次的模式,并剔除同时缺少主语和宾语的模式(即未包括名词的模式),最终得到的词性模式如表1所示。

表1　词性模式

Tab.1　Some instances of speech patterns

模式	出现次数
v n	327
n v	170
n n	72
m q n	19
n m q	18
n d v	16
a n	16
v m n	15
n a	11
v b n	10
m n p v	10
v u n	10

注:a代表形容词;b为区别词;d为副词;m为数词;n为名词;p为介词;q为量词;u为助词;v为动词

(2)完整性约束

地理事件信息包含空间要素、时间要素和事件要素。上述工作主要集中在对文本是否包含地理事件要素的判断,此外还需识别文本是否包含空间要素和时间要素。对于空间要素识别,通过判断文本中是否包含地理命名实体实现。对于时间要素识别,由于微博客消息本身带有发表时间元数据,对每条消息必然能够提取出时间信息,因此不将文本是否含有时间要素作为完整性约束的内容。需要注意的是,本研究目的是识别蕴含地理事件信息的微博客消息,选择正确的地理实体作为事件发生位置及抽文本描述中的时间是后续事件信息抽取的内容。

2.3.3 分类阈值

微博客消息文本与目标地理事件的相关度计算结果为区间[0,1]的值,在实际识别过程中需要设定一个分类阈值,以最终判断该微博客消息与目标地理事件相关或不相关。为此,利用上述相关度计算方法,统计候选语料中的各文本与目标地理事件的相关度,将相关度平均值作为识别该目标地理事件的分类阈值。

3 实验分析

3.1 实验环境

以暴雨事件作为目标地理事件,验证识别方法效果。事件训练语料来源于百度搜索结果,分别利用“北京暴雨”、“广东暴雨”、“上海暴雨”、“成都暴雨”等24组关键词采集相关网页,经去重、正文提取后,得到10 041篇网页文本作为候选语料。测试数据来源于:(1)标注微博数据集。利用爬虫抓取“#北京暴雨#”、“#广东暴雨#”、“#成都暴雨#”、“#南京暴雨#”、“#上海暴雨#”、“#天津暴雨#”、“#重庆暴雨#”7个暴雨相关微博话题下的微博客消息,去除话题标签后,人工判读并标注暴雨事件相关消息。各随机选取500条相关消息和不相关消息组成实验数据集;(2)500万微博数据集。北京理工大学张华平博士开放的新浪微博数据集,包含4 993 581条微博消息。提取算法基于Java语言实现,其中,分词算法调用NLPIR 2015工具包(http://ictclas.nlpir.org/),HLDA算法调用Mallet工具包(http://mallet.cs.umass.edu/;https://github.com/chyikwei/topicModels),skip-gram词向量生成算法调用Google word2vec工具包(https://code.google.com/p/word2vec/)。

实验采用准确率(P)、召回率(R)和F-值3个指标对方法性能进行评价。3个指标的计算如式(3)–(5)所示。

$$P=\frac{\text{正确识别的相关消息数量}}{\text{识别的相关消息数量}} \tag{3}$$

$$R=\frac{\text{正确识别的相关消息数量}}{\text{应识别的相关消息数量}} \tag{4}$$

$$F\text{-值}=\frac{\left(\beta^2+1\right)\times P\times R}{\left(\beta^2\cdot P\right)+R} \tag{5}$$

F-值基于准确率和召回率对识别方法效果作综合评价。其中,β用于调节准确率和召回率的比重,一般取1,即准确率和召回率重要性相同(式(6))。

$$F\text{-}1\text{值}=\frac{2\times P\times R}{P+R} \tag{6}$$

由于实际应用过程中优先考虑消息的可靠性,即识别结果的准确性,因此需同时考察$\beta=0.5$时的F-值(式(7))。

$$F\text{-}0.5\text{值}=\frac{1.25\times P\times R}{0.25\times P+R} \tag{7}$$

3.2 实验结果

3.2.1 分类阈值计算

基于暴雨事件训练语料和标注微博数据验证提出的分类阈值计算方法效果。由暴雨事件训练语料计算得到的分类阈值为0.505,识别结果为P=69.71%,R=73.20%,F-1值=71.41%,F-0.5值=70.38%。依次计算分类阈值为[0.1,0.8]的识别结果,如图4所示,图中数字为各分类阈值对应的F-0.5值。

由图4可看出,F-0.5值和F-1值最好结果分别为71%和74%左右,即利用提出方法计算的分类阈值可使识别结果的F-0.5值接近最佳,F-1值稍差。因此,该阈值虽无法使识别结果在准确率和召回率

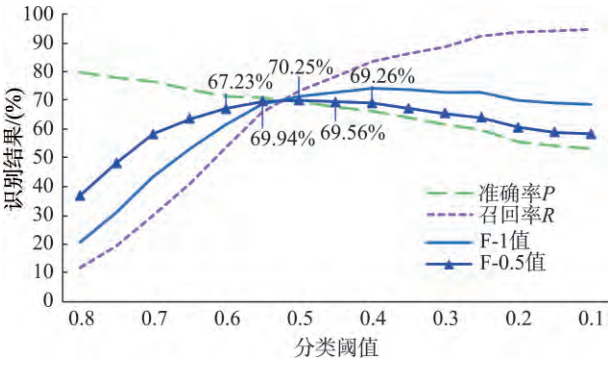

图4 不同分类阈值识别结果

Fig.4 Identification results under different thresholds

同等重要的情况下达到最佳，但能够优先保证识别事件消息的可靠性，表明该计算方法的有效性。

3.2.2 识别效果比较

利用暴雨事件训练语料和标注微博数据，比较提出方法与现有监督学习方法的识别效果。对比监督学习方法选用基于支持向量机（Support Vector Machine，SVM）的识别方法[24]，并参考文献[8]、[24]−[26]的工作，选取的识别特征包括：微博消息中词个数、各词词频、名词个数、停用词个数、事件词个数和数词个数。实验过程中，将测试数据随机分成5组，4组数据作为SVM模型的训练数据，剩余1组作为测试数据，交叉验证后的平均值作为最终识别结果，如表2所示。

表2　蕴含暴雨事件消息识别结果

Tab.2　Performance of the identification approach for micro-blogs containing rainstorm events

抽取方法	准确率/(%)	召回率/(%)	F-1值/(%)	F-0.5值/(%)
本文方法	69.71	73.20	71.41	70.38
SVM方法	68.48	54.88	60.62	65.00

实验结果显示，本文方法的召回率明显优于SVM方法，致使F-1值的提升也高于F-0.5值。结果表明，本文方法利用语料自动处理降低了事件语料获取成本，通过扩大语料规模改善了非监督方法的识别效果。分析识别错误原因，主要为识别过程缺乏对事件时态的判断，即将已结束事件识别为实时发生事件。可通过增加时效性判断改善识别效果。

此外，为验证语料规模对识别效果的影响，由候选语料中分别随机选取2000、4000、6000、8000篇网页文本重复上述识别过程，识别结果如图5所示。

由实验结果可看出，随着候选语料数量的增加，蕴含地理事件微博客消息识别的准确率下降约5%，召回率上升近20%，因此F-1值比F-0.5值有明显提升，表明在准确率没有明显降低的情况下，语料规模的增加显著提高了识别方法的召回率。

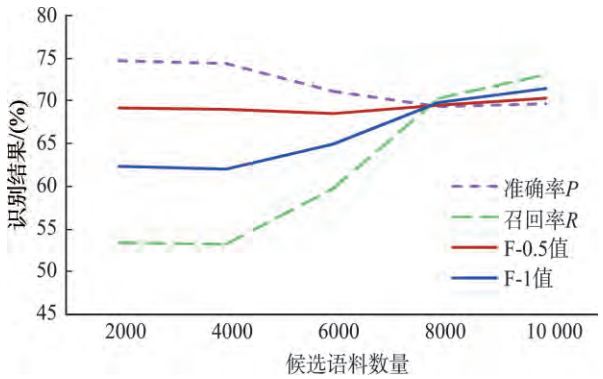

图5　不同规模候选语料识别结果

Fig.5　Identification results using different scales of candidate corpuses

3.2.3 开放环境实验

利用500万微博数据模拟真实环境下的微博数据流，验证本文方法在文本内容开放时的识别效果。识别出暴雨事件相关微博消息为2906条，人工随机判读其中的500条消息，识别正确的消息数量为307条，识别准确率为61.40%。若计算识别结果的召回率和F-值，需要人工判读整个数据集以获取实际包含的事件消息数量，工作量巨大，因此未在该实验中计算此类指标。

识别错误的主要原因除已提到的时态问题外，还在开放语料中易将与目标地理事件近似的无关事件识别为目标事件，如将寒潮、冰雹等气象事件识别为暴雨事件。这类事件易与目标地理事件同时出现在语料中，导致相关性计算结果较高。后续工作可尝试将单一事件识别扩展为多事件识别，通过比较事件差异提高对各类型地理事件的识别效果。

3.3 讨论

（1）本文方法仍需一定人工参与，主要在语料提取阶段，根据关键词判断各主题是否与目标地理事件相关。但主题数量远少于需标注的语料数量，如实验中暴雨事件候选语料提取的主题数量为51，文本数量为10 041，随着语料资源的增加，二者之间的差异将更加显著。因此，本文方法可以明显地减少人工成本，实现语料快速更新，满足新类型地理事件消息识别的需求。

（2）实验结果显示，本文方法在开放微博数据集的识别表现低于话题微博数据。主要原因在于，话题本身易使事件相关消息聚集，无关类型消息（个人状态、评论、广告等）较少，有利于识别效果的提升。而开放数据中的微博消息更接近用户真实使用环境，消息类型及内容更为复杂，目标地理事件信息所占比例低，易导致识别错误。参考命名实体识别[27]、事件识别[28]和关系提取[29]等研究工作，其开放环境下的实验准确率约为60%~70%，即提高方法在开放数据环境下的表现是自然语言处理未来研究的重点。

（3）基于主题模型，提出方法能够将候选语料划分为事件相关语料和事件无关语料。然而，将分割后的语料作为训练数据直接应用于监督分类模型仍存在一定不足。首先为训练数据中的正反例

不平衡,相较于事件相关信息,事件无关信息的类型和内容更加丰富,本文方法分割后的事件无关语料不能完全覆盖事件无关信息,加之正反例比例对分类效果的影响难以界定[30],需要对训练语料进行分析和优化工作,以提高识别效果。其次,提出方法获取的地理事件相关语料来源于网页长文本,与微博客消息短文本之间存在差异,如应用于监督分类模型需在分类特征构建过程中充分考虑这种差异。因此,本文方法没有使用监督分类模型,而是通过蕴含了地理事件语义的相关度对微博客消息文本进行识别,以降低上述问题对识别结果造成的影响。

(4)地理事件候选语料来源于网页,需借助正文提取算法去除网页源码中的xml标签、导航栏文字等内容,以获得正文部分文本。由于网页本身的复杂性,本文算法无法保证对所有网页的正文都能正确解析,若将无关文字识别为正文则会影响语料质量。然而,由于网页的非正文部分本具有相似文字特征,主题提取算法能够将该部分文本提取至相同主题,可经人工判读后剔除,避免对词向量构建过程产生不良影响。因此,本文方法能够降低错误网页对识别结果的影响。

4 结论

本文提出了一种蕴含地理事件微博客消息的自动识别方法,利用主题模型生成文档主题集合的特性,实现地理事件语料的快速提取,降低语料资源的获取难度。同时,利用分布式表达词向量模型隐含的语义信息优化相关度计算过程。因此,提出方法能够通过扩大事件语料规模改善非监督方法的识别效果,实现对地理事件微博客消息的准确识别。以新浪微博为数据源开展的实验分析表明,本文所提出的消息识别方法,识别来自事件微博话题的消息文本的F-1值达到71.41%,比经典的基于SVM模型的监督学习方法提高了10.79%。在模拟真实微博环境的500万微博客数据集上的识别准确率达到60%。

下一步工作中,开展多类型地理事件消息的识别研究,尝试利用不同类型地理事件的特征差异增强识别效果。同时,通过增加时效性判断、优化事件核心词集和相关性计算方法提高地理事件信息相关度计算的准确性。

参考文献(References):

[1] 王明会,丁焰,白良.社会化媒体发展现状及其趋势分析[J].信息通信技术,2011(5):5-10. [Wang M, Ding Y, Bai L. Social media development status and trend analysis[J]. Information and Communications Technologies, 2011,5:5-10.]

[2] Li R, Tao X, Tang L, *et al*. Using maximum entropy model for Chinese text categorization[A]. In: Yu J X, Lin X, Lu H, *et al*(eds.). Advanced Web Technologies and Applications[C]. Springer-Verlag, 2004:578-587.

[3] Sankaranarayanan J, Samet H, Teitler B E, *et al*. Twitter Stand: News in tweets[C]. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2009:1-10.

[4] Jiang S, Pang G, Wu M, *et al*. An improved K-nearest-neighbor algorithm for text categorization[J]. Expert Systems with Applications, 2012,39(1):1503-1509.

[5] Kumar M A, Gopal M. A comparison study on multiple binary-class SVM methods for unilabel text categorization [J]. Pattern Recognition Letters, 2010,31(11):1437-1444.

[6] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]. Proceedings of the 19th International Conference on World Wide Web (WWW'10), ACM, 2010:851-860.

[7] 丁效,宋凡,秦兵,等.音乐领域典型事件抽取方法研究[J].中文信息学报,2011,25(2):15-20. [Ding X, Song F, Qin B, *et al*. Research on typical event extraction method in the field of music[J]. Journal of Chinese Information Processing, 2011,25(2):15-20.]

[8] Miwa M, Sætre R, Kim J-D, *et al*. Event extraction with complex event classification using rich features[J]. Journal of Bioinformatics and Computational Biology, 2010,8(1):131-146.

[9] Zhao L, Chen F, Dai J, *et al*. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling[J]. PLoS ONE, 2014,9(10):e110206.

[10] Wang W, Stewart K. Spatiotemporal and semantic information extraction from Web news reports about natural hazards[J]. Computers, Environment and Urban Systems, 2015,50:30-40.

[11] Murthy D, Longwell S A. Twitter and disasters[J]. Information, Communication & Society, 2013,16(6):837-855.

[12] 张恒才,陆锋,仇培元.基于D-S证据理论的微博客蕴含交通信息提取方法[J].中文信息学报,2015,29(2):170-178. [Zhang H, Lu F, Qiu P. Extracting traffic information from micro-blog based on D-S evidence theory[J]. Journal of Chinese Information Processing, 2015,29(2):170-178.]

[13] 仇培元,张恒才,陆锋.互联网文本蕴含道路交通信息抽取的模式匹配方法[J].地球信息科学学报,2015,17(4):416-422. [Qiu P, Zhang H, Lu F. A pattern matching method for extracting road traffic information from internet texts[J]. Journal of Geo-Information Science, 2015,17(4):416-422.]

[14] 王曙,吉雷静,张雪英,等.面向网页文本的地理要素变化检测[J].地球信息科学学报,2013,15(5):625-634. [Wang S, Ji J, Zhang, X, *et al*. Change detection of geographic features based on web pages[J]. Journal of Geo-Information Science, 2013,15(5):625-634.]

[15] 张春菊.中文文本中事件时空与属性信息解析方法研究[D].南京:南京师范大学,2013. [Zhang C. Interpretation of event spatio-temporal and attribute information in Chinese text[D]. Nanjing: Nanjing Normal University, 2013.]

[16] 刘纪平,栗斌,石丽红,等.一种本体驱动的地理空间事件相关信息自动检索方法[J].测绘学报,2011,40(4):502-508. [Liu J, Li B, Shi L, *et al*. An automated retrieval method of geo-spatial event information based on ontology[J]. Acta Geodaetica et Cartographica Sinica, 2011,40(4):502-508.]

[17] 张剑峰,夏云庆,姚建民.微博文本处理研究综述[J].中文信息学报,2012,26(4):21-27,42. [Zhang J, Xia Y, Yao J. A review towards microtext processing[J]. Journal of Chinese Information Processing, 2012,26(4):21-27,42.]

[18] 徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,34(8):1423-1436. [Xu G, Wang H. The development of topic models in natural language processing [J]. Chinese Journal of Computers, 2011,34(8):1423-1436.]

[19] Blei D M, Ng A Y, Jordan M I. Latent dirichl *et al.* location[J]. Journal of Machine Learning Research, 2003,3:993-1022.

[20] Blei D M, Griffiths T L, Jordan M I, *et al*. Hierarchical topic models and the nested Chinese restaurant process [A]. In: Advances in Neural Information Processing Systems[M]. Cambridge, MA: MIT Press, 2004.

[21] Bengio Y, Ducharme R, Vincent P, *et al*. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003,3:1137-1155.

[22] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space[C]. Proceedings of Workshop at International Conference on Learning Representations, 2013:1-12.

[23] 刘挺,马金山.汉语自动句法分析的理论与方法[J].当代语言学,2009,11(2):100-112,189. [Liu T, Ma, J. Theories and methods of Chinese automatic syntactic parsing: A critical survey[J]. Contemporary Linguistics, 2009,11(2):100-112,189.]

[24] Naughton M, Stokes N, Carthy J. Sentence-level event classification in unstructured texts[J]. Information Retrieval, 2009,13(2):132-156.

[25] 许旭阳,李弼程,张先飞,等.基于事件实例驱动的新闻文本事件抽取[J].计算机科学,2011,38(8):232-235. [Xu X Y, Li B C, Zhang X F, *et al*. News text event extraction driven by event sample[J]. Computer Science, 2011,38(8):232-235.]

[26] 许红磊,陈锦秀,周昌乐,等.自动识别事件类别的中文事件抽取技术研究[J].心智与计算,2010,4(1):34-44. [Xu H, Chen J, Zhou C, *et al*. Research on event type identification for Chinese event extraction[J]. Mind and Computation, 2010,4(1):34-44.]

[27] Li C, Weng J, He Q, *et al*. TwiNER: Named entity recognition in targeted twitter stream[C]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012:721-730.

[28] Zhou D, Chen L, He Y. An unsupervised framework of exploring events on twitter: filtering, extraction and categorization[C]. Proceedings of the 29 AAAI Conference on Artificial Intelligence, 2015:2468-2474.

[29] Wu F, Weld D S. Open information extraction using Wikipedia[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010:118-127.

[30] Weiss G M, Provost F. Learning when training data are costly: The effect of class distribution on tree induction [J]. Journal of Artificial Intelligence Research, 2002,19:315-354.