计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150746
53(3):611-620,2016

面向网络舆情数据的异常行为识别

郝亚洲　郑庆华　陈艳平　闫彩霞
(陕西省天地网技术重点实验室(西安交通大学)　西安　710049)
(西安交通大学计算机科学与技术系　西安　710049)
(hyzxjtu@qq.com)

Recognition of Abnormal Behavior Based on Data of Public Opinion on the Web

Hao Yazhou, Zheng Qinghua, Chen Yanping, and Yan Caixia
(*SPKLSTN Laboratory* (*Xi'an Jiaotong University*), *Xi'an* 710049)
(*Department of Computer Science and Technology*, *Xi'an Jiaotong University*, *Xi'an* 710049)

Abstract　With the increasing popularity of the social network, public awareness and participation to hot topics has been much improved, mobile terminal equipment and fast Internet access make the spread of public opinion quickly. Public opinion on the Web has freedom, interactivity, diversity, deviation and burstiness as characteristics, has become an important factor that affects social stability. Therefore, how to timely detect, control and guide the development of public opinion is of great significance to the social stability. This article focuses on the behaviors that spread on the Web and contain "destruction", "dangerous" and "loss" involves public security or judicial justice, and the behaviors is defined as abnormal behavior. We define the types of abnormal behavior that this article focuses on are aggression, injury, death, and arrests, four categories. From the point of view of information extraction, our method recognizes the abnormal behavior by identifying sentences that contain the abnormal behavior and constructs co-occurrence network of abnormal behavior, with provide the visualization analysis approach of public opinion on the Web.

Key words　public opinion; event extraction; recognition of abnormal behavior; co-occurrence network; data mining

摘　要　社交网络的日益普及和移动设备快捷的网络接入,使得网络舆情的传播十分迅捷,民众对热点话题的关注度和参与度得到很大的提升.网络舆情具有自由性、交互性、多元性、偏差性、突发性等特点,能够左右民众的情感和判断,能推动和改变事件的发展和走向,容易被反对分子利用,已经成为影响社会稳定的重要因素.因此,及时检测、控制并引导舆情的发展具有十分重要的意义.研究关注网络中传播的蕴含有"破坏"、"危险"、"损失"等涉及公共安全或涉及司法公正的行为.根据课题的需要,定义4种关注的异常行为类型:攻击行为、受伤行为、死亡行为、拘捕行为.从数据挖掘和信息抽取的角度研究识别异常行为的方法,首先通过分类器和触发词从海量的数据中过滤出包含异常行为的句子,然后抽取异常行为句中包含的命名实体,最后利用抽取的实体构建异常行为共现网络,为分析人员提供可视化的网络舆情分析方法.

关键词　网络舆情;事件抽取;异常行为识别;共现网;数据挖掘

中图法分类号　TP391

收稿日期:2015-10-29;**修回日期**:2015-11-18

基金项目:国家自然科学基金项目(91118005,91218301,91418205);国家"八六三"高技术研究发展计划基金项目(2012AA011003)
This work was supported by the National Natural Science Foundation of China (91118005,91218301,91418205) and the National High Technology Research and Development Program of China (863 Program) (2012AA011003).

通信作者:郑庆华(qhzheng@mail.xjtu.edu.cn)

近年来,我国网民规模一直呈现十分迅速的增长趋势,在全球互联网中占据越来越重要的位置.随之而来的是网络数据的急速增加,互联网已经成为我国最大的社交平台和信息集散地.据中国互联网信息中心(China Internet Network Information Center, CNNIC)于2015年1月发布的第35次中国互联网发展状况调查统计报告显示,我国网民规模和互联网普及率较上年有明显提升.伴随着我国网络规模的不断扩大,网络媒体被越来越多的民众认可和使用,成为一种新的信息传播方式,并且渐渐地超越了传统媒体,为信息的发布、传递和获取带来了更方便和快捷的全新概念.

与传统的信息传播媒体相比较而言,网络媒体上的信息交流具有门槛低、规模大、传播迅速、参与群体庞大、实时性强等特点,再加上BBS论坛、微博、博客、新闻跟帖等社交媒体的虚拟隐蔽性等特点,导致广大网民积极地通过网络参与热点话题的讨论与传播,这些热点话题大多属于网络新闻.在参与热点话题的过程中,网民会对社会热点表达自己的观点,如果这种观点引起了较大范围网民的关注和共鸣,并通过网上讨论、跟帖、转帖等逐步形成一种网络舆论倾向,就形成了网络舆情.

网络舆情中最引人关注的是现实中发生的,民众热切关注和议论并蕴含有“破坏”、“危害”、“损失”等涉及公共安全或涉及司法公正的行为.本文将这些行为定义为异常行为.异常行为通常关系到广大民众的切身利益,影响到社会的安定和谐.而且部分网民通过网络进行个人情绪的发泄,从而发表一些过激片面的言论.部分反动或恐怖组织也会利用网络发布虚假反动的信息,这类信息通常就属于异常行为.由于信息传播的不对称性,大多数网民是无法辨别这些信息真伪的,很多人会盲目相信并传播这些虚假言论,从而影响政府的形象和社会的安定.因此,及时地发现网络中的异常行为并辨别其真实性至关重要.对于真实发生的异常行为,要调查核实并维护社会公平正义,对于虚假宣传的异常行为,需要及时停止其在网络上的传播和扩散,以免产生更大的危害.

本文利用信息抽取的技术识别异常行为.信息抽取(information extraction, IE)是把文本里包含的信息进行结构化处理.抽取出的是结构化或半结构化的信息,将抽取结果存储到数据库中,方便人们进行相关的查询和处理,从而在很大程度上提高人们的工作效率.本文根据事件抽取的相关概念,定义异常行为的识别对象和方法.事件抽取本质上是信息抽取领域一个很重要并且应用十分广泛的研究方向,在信息检索等诸多领域都有着广泛的应用.在事件抽取领域主要有2个权威的研究机构:信息理解研讨会议(Message Understanding Conference, MUC)[1]和自动内容抽取(automatic content extraction, ACE)会议[2].

MUC会议是ACE会议的前身,每2年举办一次,只从1987年维持召开到1998年,总共举行了7届,但即使如此,它也为事件抽取的任务目标与相关理论的制定做出了相当大的贡献.在MUC会议停止召开后的2000年开始至今,美国NIST组织举办了ACE自动内容抽取会议,该会议召开后,由于其与信息理解研讨MUC会议研究内容和研究领域的相似性,人们就认为该会议是MUC会议的扩展和延伸,被越来越多的专业人士所认可和关注,该会议所制定的抽取标准和目标也就自然而然成为了该领域比较权威的标准.

ACE将事件抽取的任务进行了更明确的规定,将其定义为事件的检测与识别(event detection and recognition, VDR),即我们进行事件抽取的目标是从大量的文本数据中识别出所关注的某些特定类型的事件描述句,并对这些事件描述句进行相关信息的确定和抽取,例如事件的类型和子类型、事件的元素等.

现阶段进行事件抽取有2种方法被广泛应用,即模式匹配和机器学习,这2种方法各有利弊,针对不同领域的事件抽取任务,可采用对应的抽取方法.代表性的工作有1995年Riloff和Shoen[3]提出的基于无标注语料的自动事件抽取方法、2001年Yangarber[4]提出的基于种子模式的自举信息抽取模型学习系统ExDisco系统,这个系统以人工构造的质量较高的种子模板为基础,多次迭代增量式地学习新的模板.2002年Chieu和Ng[5]在事件元素抽取问题上首次引入了最大熵分类器.2005年姜吉发[6]提出了基于领域无关概念知识库的事件抽取模式学习方法GenPAM,它的优点是完全无指导,且对于标注语料基本没需求.需要人工参与的部分只是要给出事件抽取的事件类型、事件元素及其所属角色,最后人工对学习到的模式进行评价.如此,自动学习事件抽取模板,大幅度地减少了需要人工参与的工作量.2006年Ahn[7]将MegaM和TiMBL这2种机器学习方法进行结合并在ACE语料库验证证明优于单一算法的性能.2007年于江德等人[8]使用隐Markov模型(HMM)完成事件元素的抽取.2009年Chen和Ji[9]打破了事件抽取中分类的思想,

从序列标注的角度来识别事件.同年付剑锋等人[10]提出了基于依存分析的事件识别.2010 年 Llorens 等人[11]使用 CRF 模型进行 TimeML 事件抽取中的语义角色标注,提升了系统的性能.同年许红磊等人[12]提出自动识别事件类别的中文事件抽取技术,取得较好的效果.由于基于机器学习的事件抽取方法客观高效的优点,目前国内外大多采用机器学习的方法进行事件抽取,而本文只需要识别事件类别,因此也采用该方法进行事件类型识别.

本文提出异常行为识别,基于此构建异常行为共现网的方法,为蕴含在网络中的异常行为提供可视化的分析方法.

1 异常行为识别研究综述

1.1 异常行为识别相关概念

定义 1. 事件.由触发词和描述时间结构的元素组成,表示一个动作的发生或状态的变化.往往由动词驱动,也可以由能表示动作的名词等其他词性的词来触发,它包括参与该动作行为的主要成份(如人物、地点、时间等).

定义 2. 触发词.触发词是最能表现事件发生的词语,通常是一个动词或者能够代表动作发生的名词.

定义 3. 事件描述句.事件描述句是文本中描述事件信息的句子或片段,通常包含了一个触发词.

定义 4. 异常行为.现实中发生,民众热切关注和议论并蕴含有"破坏"、"危害"、"损失"等涉及公共安全或涉及司法公正的行为.

1.2 异常行为识别研究目标

进行异常行为识别首先要确定我们需要关注的行为类型.本文中采用的是 ACE 中定义的事件类型,包括 8 个大类和 33 个子类,如表 1 所示:

Tabel 1 Type of Event Defined in ACE

表 1 ACE 评测中定义的事件类型

Type	Subtype
Life	Born, Marry, Divorce, Injure, Die
Conflict	Demonstate, Attack
Transaction	Transfer-Money, Transfer-Ownership
Business	Start-Org, End-Org, Declare-Bank
Personnel	Nominate, Elect, End-Position, Start-Position
Justice	Sue, Arrest-Jail, Pardon, Sentence, Extradite, Execure, Convict, Release-Parole, Appeal, Fine, Charge-Indict, Trial-Hearing
Contact	Meet, Phone-Write
Movement	Transport

根据定义 4,我们关注的是现实中发生,民众热切关注和议论并蕴含有"破坏"、"危害"、"损失"等涉及公共安全或涉及司法公正的行为.这些行为类型是 ACE 中定义的全部事件类型的子集.根据研究的需要和项目的需求,我们定义本文关注的异常行为类型是攻击行为、受伤行为、死亡行为、拘捕行为 4 类,分别对应 ACE 事件类型中的 Attack, Injure, Die, Arrest-Jail.

一个典型的事件抽取通常会识别事件的 6 个要素,我们将其作为一个六元组,即〈*trigger*, *type*, *subject*, *object*, *time*, *place*〉,其中,*trigger* 表示引发事件发生的触发词,每个触发词一般触发一种异常行为,从一定程度上决定了行为的类型;*subject* 表示事件的行为主体,是动作活动的发起、状态的主体,是有生命的人;*object* 表示事件涉及的行为客体,是行为中的动作活动所涉及或者影响到的人或事物,是行为的被动承受者;*time* 表示事件发生的时间;*place* 表示事件发生的地点.六元组中的 *trigger* 和 *type* 不能为空,其他项为可选项,可以为空.以下面的新闻报道为例:

"2013 年 6 月 26 日凌晨 5 时 50 分许,新疆吐鲁番地区鄯善县鲁克沁镇发生暴力恐怖袭击案件,多名暴徒先后袭击鲁克沁镇派出所.这是在新疆发生的民族分裂分子有预谋、有组织策划制造的公然扰乱社会秩序、制造恐怖气氛的恶性案件."

这段话中就包含一个典型的事件,其中,触发词"袭击"触发了一个攻击行为,再进一步分析句子中有关的主体、客体、时间、地点信息,可以得到完整的六元组〈"袭击","攻击行为","多名暴徒","鲁克沁镇派出所","2013 年 6 月 26 日凌晨 5 时 50 分许","新疆吐鲁番地区鄯善县鲁克沁镇"〉.

由于传统的事件识别正确率低,根据 ACE 评价标准,目前相关研究的性能在 30%左右.其原因首先在于需要抽取触发词,行为主体、客体、时间、地点等事件要素,抽取性能较差;其次在开放的大数据环境下,数据的异质性、噪音、碎片化等特点,更加影响抽取性能.而本文提出的基于句子分类的异常行为识别方法先将带异常行为的句子识别出来,再抽取异常行为句中的触发词、实体和实体的共现关系,相比 ACE 定义的事件抽取性能更好.同时识别出的句子加入人工干预,可以辅助舆情分析人员,提高效率.

本文的研究目标是从实际爬取的大量网络舆情文档集中,识别出攻击、受伤、死亡、拘捕 4 类异常行为,并存储在数据库中,便于人们进行查询,及时了

解网络舆情热点信息,并且可以帮助政府更好地分析具有某种行为倾向的人和地点等信息,对决策做出一定的支持.

本文进行异常行为识别的具体目标有 3 点:

1) 识别出异常行为句. 即为该新闻片段的第 1 个句子.

2) 判断异常行为类型. 该异常行为属于攻击行为.

3) 构建异常行为共现网. 将异常行为句中的实体及其共现关系表示在异常行为共现网中并进行相关分析. 如新疆吐鲁番地区鄯善县鲁克沁镇、多名暴徒、鲁克沁镇派出所这 3 个实体出现在一个异常行为句中,它们都属于共现网中的节点,并且两两有共现关系.

1.3 研究框架和技术路线

本文的研究框架如图 1 所示:

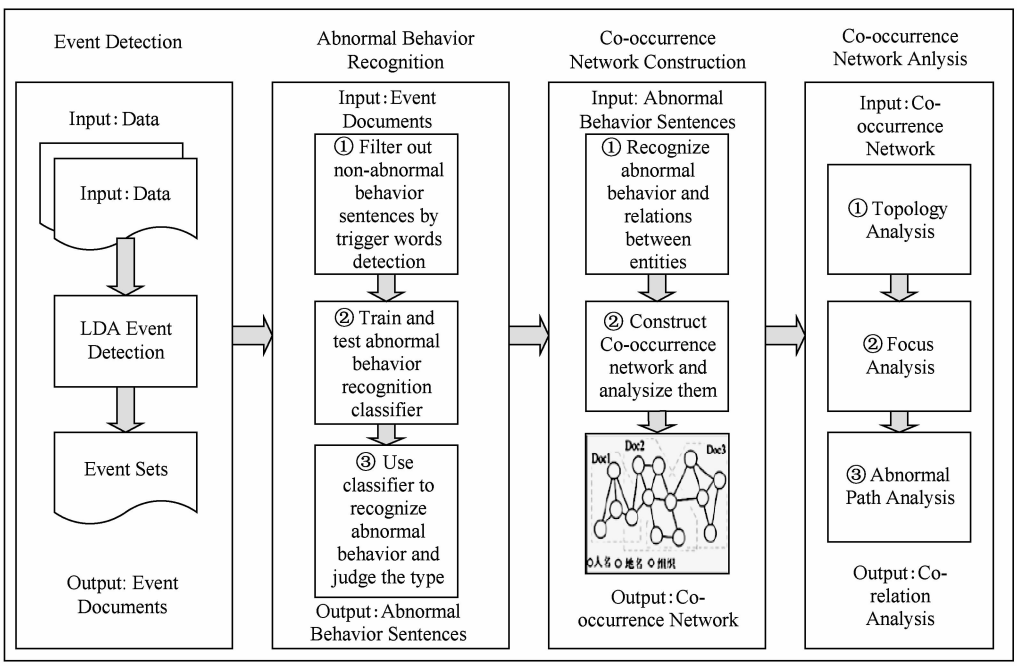

Fig. 1 General diagram of abnormal behavior recognition.

图 1 异常行为识别总体框图

图 1 研究框架共分为 4 个部分:事件识别、异常行为句识别、异常行为共现网构建和共现网络分析. 其中,异常行为句识别和异常行为共现网构建这 2 部分是整个系统的核心部分,事件识别是预处理阶段,共现网络分析属于扩展部分.

识别过程可以概括为 3 个步骤:

1) 预处理. 采用 LDA 模型对网络舆情文档集进行文档事件识别,对识别出的每个文档事件分别进行后续操作. 该阶段可以识别出多个文档事件,如钓鱼岛事件、占中事件等,为后续的操作提供输入. 通过对文档事件单独进行操作,每次处理的数据量更小且更有针对性.

2) 异常行为识别阶段. 首先根据触发词表,用触发词检测的方法初步过滤掉非异常行为句,得到候选异常行为句的集合. 然后用 ACE 的标准数据集训练 SVM 异常行为识别分类器,选取句子的全词特征[13]作为特征向量. 最后用训练好的分类器对候选异常行为句进行异常行为识别,并判断行为类型.

3) 构建异常行为共现网. 利用中国科学院分词工具进行命名实体识别,将出现在同一个异常行为句中的实体定义为有共现关系. 用 igraph 构建出包含关键实体及其共现关系的异常行为共现网,为异常行为提供可视化的分析方法.

2 异常行为识别流程

2.1 触发词检测

1) 问题分析

在开放的网络环境中爬取的网络舆情数据具有数量大和异质性等特点. 其中包含大量的无用和干扰数据,如果对这些数据全都进行处理,不仅浪费时间,也会影响系统的处理结果和性能. 触发词检测可以去除噪音,过滤掉大量的无用数据,提高系统的效率.

异常行为是由具体的行为发生或状态改变所引发的,描述句通常包含一个触发词.异常行为触发词可以直接引起异常行为的发生,是决定行为类型的重要特征.对于不含异常行为触发词的句子,我们认为该句子不含异常行为,直接将其过滤掉.因此,我们可以根据句子中触发词的有无进行初步过滤.下面针对每种异常行为类型分别给出了一个含有触发词的句子.

① 死亡(Die).警方一直紧随其后,最终顺利击毙逃犯.触发词:击毙.

② 攻击(Attack).当天在加沙地带和约旦河西岸地区仍有零星的冲突发生.触发词:冲突.

③ 拘捕(Arrest-Jail).电焊工王呈泰等12名犯罪嫌疑人已被检查机关批准逮捕.触发词:逮捕.

④ 受伤(Injure).巴基斯坦方面说:最近在平泊尔地区,有很多士兵被打伤.触发词:打伤.

2) 实验数据集

本文采用的实验数据分为2个部分:①ACE2005中文语料库的682篇新闻报道;②人工标注网络爬虫爬取的真实网络舆情数据318篇(条).其中ACE2005语料是由ACE评测会议发布,其中的中文语料分布如表2所示:

Table 2 2005ACE System Training Corpus Statistics for Release LDC2005E18

表2 ACE2005中文语料

Source	Training Epoch	Approximate Size
Broadcast News	10/00－12/00	120 000 Words
Newswire	10/00－12/00	120 000 Words
Weblog	11/04－02/05	60 000 Words

Indication:1) Chinese Resources (1.5characters=1word);
2)Descriptions like mm/nn means month/year.

由2部分数据构成的实验数据总体分布如表3所示:

Table 3 Distribution of Data

表3 实验数据分布表

Statistical Data	ACE2005	Labeled Data	Each Kind
News	682	60(Sohu)	742
Forums	0	100(Tianya)	100
Microblog	0	158(Sina)	158
Total Number			1000

Indication:Units of the numbers in the table is sheet or post.

3) 问题解决

基于词的触发词检测的首要任务是建立初始的触发词表,实验数据所包含的1 000篇中文文档进行统计,33个子类别的触发词共计976个.进一步对这976个触发词筛选,选出其中的4个子类别"Attack","Injure","Die","Arrest-Jail"的触发词来构建初始的触发词表,所构建的触发词表一共包含338个触发词,具体内容如表4所示:

Table 4 Trigger Words

表4 触发词表

Trigger Words	Event Type
咬	Attack
行凶	Attack
昏死	Injure
不治	Die
砍杀	Attack
受伤	Injure,Attack
战火连连	Attack
拘捕	Arrest-Jail
自焚	Injure,Attack
谋杀	Die
⋮	⋮

根据触发词表,采用基于关键词匹配的方法,对于S中的每个句子进行检测,过滤掉不含异常行为触发词的句子,得到候选异常行为句的集合S'.

2.2 异常行为识别分类器

1) 构建分类器的原因

虽然"触发词"被定义为"最能表现事件发生的词语",但并不意味着"触发词的出现一定代表了事件的发生".例如"谋杀"这一触发词,在句子"根据加州法律,不管有意或无意杀害火车上的人,光是这一点就足以构成谋杀罪"中,这只是谋杀罪的一种构成方式,实际上并没有谋杀行为的发生.本文关注的是例如"他被控于1989年同其他几名成员一道将一名试图脱离这个组织的21岁的成员谋杀"句子中"谋杀"所触发的"Die"事件.因此,只通过句子有无触发词来判断异常行为的发生是不一定正确的.

为了验证触发词的出现是否代表着事件的发生,本文对触发词表中的每个触发词统计其在ACE语料中出现的总次数以及触发事件的次数,统计结果如表5所示:

Table 5 Frequency of Triggers and Events

表 5 触发词出现次数与触发事件次数统计

Trigger Words	Frequency of Occurrence	Frequency of Triggered Event
咬	4	3
去世	4	4
事件	108	1
击	176	10
交战	2	1
谋杀	10	1
侵略	18	6
⋮	⋮	⋮

对表 5 中的出现总次数和触发事件次数的对应关系分析,得到表 6:

Table 6 Ratio of Triggers and Events

表 6 触发词出现与触发事件的比例关系 %

Occur and Trigger Event	Not Always Trigger Event
69	31

因此需要选择合适的分类器筛选出真正地代表了 4 类事件发生的触发词所在的异常行为描述句,本文选择支持向量机(support vector machine, SVM)分类器,这是由于 SVM 能够将非线性的问题转化为高维空间的线性问题,从很大程度上降低了问题的难度,并且依据结构风险最小化的原则和核函数的思想,在解决有限样本的非线性以及高维模型识别问题中表现出优于其他模型的性能. 而本文就选用了高维特征,因此 SVM 能更好地用于解决本文的分类问题,下面对 SVM 的具体原理以及本文如何使用 SVM 进行详细介绍.

2) SVM 概述

SVM 是对线性分类器的一种最佳设计准则,1965 年由 Vapnik 和 Cortes 在统计学习理论基础上提出之后就被人们广泛应用,SVM 的主要思想概括为 2 点:

① SVM 本质上只分析和处理线性可分的情况,对于线性不可分的样本,它会通过非线性映射算法将低维空间的样本映射到高维空间的方法使得线性不可分的样本变得可分,从而就可以对这些高维空间的样本进行线性处理,降低处理的复杂度.

② SVM 为使分类的风险最小化,在特征空间中构建分割平面的时候,会构建使得学习器得到全局最优化的分割平面,并且在分类时的全局期望风险以某个概率满足一定上界,从而可以达到较好的分类效果. 二元分类问题的最优分割平面满足:

$$\boldsymbol{w}\cdot\boldsymbol{x}+\boldsymbol{b}=0, \tag{1}$$

其中,$\boldsymbol{w}\cdot\boldsymbol{x}$ 为多维向量,表示向量与向量的内积. 最优平面要求:如果训练样本被平面正确切分,并且距离平面越近的训练样本与平面的间距越大. 最小化的约束条件为所有的数据点到最优平面的距离大于 1,并且保证训练样本被正确切分. 同时,引入非松弛变量来解决部分样本不能被正确地分类的情况,因此平面最优解问题可以被表示为

$$\begin{cases}\min\left(\dfrac{1}{2}|w|^2+c\sum\limits_{i=1}^{n}\varepsilon_i\right),\\ y_i(w\cdot x_i+b)\geqslant 1-\varepsilon_i.\end{cases} \tag{2}$$

其中,$\varepsilon_i\geqslant 0$, $i=1,2,\cdots,n$.

目前,SVM 的开源工具有很多,其中使用最多的是台湾大学的林智仁教授等人开发的一个用于支持向量机分类的开源库 LibLinear,它也可以用来解决多类分类问题. LibLinear 由于程序小、运用灵活、输入参数少、易于扩展等优点成为目前国内应用最多的 SVM 库. 目前有 C++, Python, Java, R, Matlab 等多种语言的接口,可以方便地在 Windows 或 Unix 平台下使用. 另外,Windows 平台下还有可视化操作工具 SVM-toy.

3) 特征选择

本文训练分类器所用的实验数据共有 1 000 篇标注过的新闻文档. 要训练分类器,首先要选择合适的分类特征,由于分类器处理的是候选异常行为句这样的短文本,信息量较少,为了充分利用句子的信息,我们选用全词特征,逐字扫描句子,若匹配到在词典中出现的词,就把该词放入特征向量中,这样,句子中所有潜在的词都被放入特征向量中,向量的维度就是词典中词的个数. 全词特征解决了传统的分词导致的词语边界错误问题,最大限度地利用句子中的信息. 要想取得较好的效果,词典的选择就十分重要,我们将 2 个词典合并作为本文的词典. 第 1 个词典是 The Lexicon Common Words in Contemporary Chinese. 第 2 个词典由 ICTCLAS 分词工具对实际的舆情文档集分词得到,加入这个词典以提高性能.

最后,抽取每个句子的全词特征向量作为 SVM 分类器的输入,训练分类器.

2.3 异常行为类型识别

第 2.1 节、第 2.2 节介绍了事件识别和触发词检测,得到了候选异常行为句的集合. 因此,现在只需要调用异常行为识别分类器对所有候选异常行为

句进行分类. 设 $C=\{-1,1,2,3,4\}$ 为行为类别的集合,-1 代表非异常行为,1,2,3,4 分别代表 Attack,Injure,Die,Arrest-Jail 这 4 种异常行为类型. 我们的目标是识别每个句子对应的行为类别.

2.4 实验结果

1) 触发词检测:The Peoples Daily 的所有文档中共检测出 579 113 个候选异常行为句.

2) 异常行为识别分类器:一般情况(分类器的预测值大于 0.5 时,输出为正例)下的性能如表 7 所示:

Table 7 Performance of Classifier

表 7 分类器性能

Event Type	Precision	Recall	F1
Attack	0.901	0.821	0.859
Injure	0.806	0.703	0.751
Die	0.932	0.747	0.829
Arrest-Jail	0.798	0.604	0.686

改进的方向是:在开放的大数据条件下,异常行为的数量十分庞大,因此我们更加注重识别的准确率而不是召回率. 可以通过提高分类器的预测值来使准确率提高、召回率降低,符合我们的要求.

3) 异常行为识别:在所有的候选异常行为句中,共检测出 41 830 条异常行为.

3 异常行为共现网

异常行为共现网本质上来说是一种异质网络,网络中的每个点是一个实体(人名、地名、组织名),将这些实体作为共现网中的节点,对于共同出现在一个异常行为句中的实体,认为它们有共现关系,在共现网中给出连线. 这样,将一个事件中的所有实体及它们的关系都表现在共现网中就构成了异常行为共现网.

命名实体识别之后我们已经得到了异常行为句中的实体及其关系,可以根据共现次数筛选出其中的关键实体,将这些实体作为节点,实体关系作为边,我们就得到了异常行为共现网.

我们还可以对共现网进行相关分析,如网络拓扑分析,焦点分析和异常路径分析. 具体来说,我们可以从节点的度,节点所经过的最短路径的条数以及节点与其他节点之间的路径长度等角度来进行分析,从而挖掘共现网中中心性高的节点或比较关键的节点,并且衡量 2 个节点之间关系的密切程度.

以香港占中事件为例,数据库中共有 757 条记录,记录中包含异常行为句 2587 个、实体数 245 个、共现关系 249 个,我们分别采用 2 种策略构建共现网,根据共现次数多的方法构建的网络如图 2 所示:

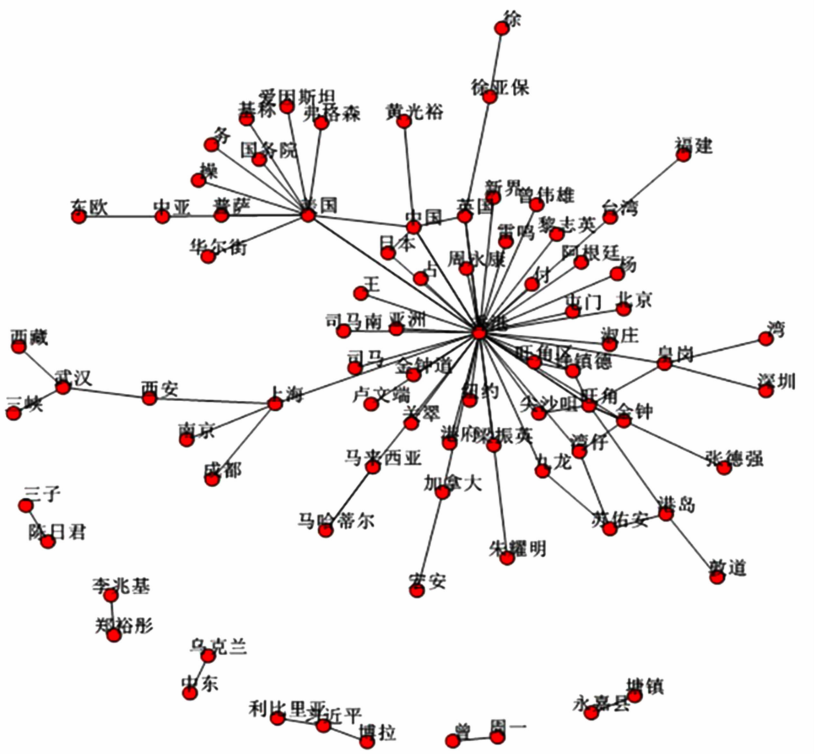

Fig. 2 Abnormal behavior co-occurrence network 1.

图 2 异常行为共现网 1

从图 2 可以看出,很多实体都与“香港”这个实体共现组成实体对,“香港”处于网络比较中心的位置,该网络就显示出了与“香港”这个中心节点共现次数最多的节点,结合实际,网络中表现的就是香港占中事件中,“香港”是中心节点以及与它共现次数最多的其他在该事件中比较关键的节点,从而可以帮助舆情分析人员及时发现热点事件中的关键实体(人名、地名、组织名),并及时采取相应举措控制和引导舆情向正常的方向发展.

根据度数最大的方法构建的网络如图 3 所示:

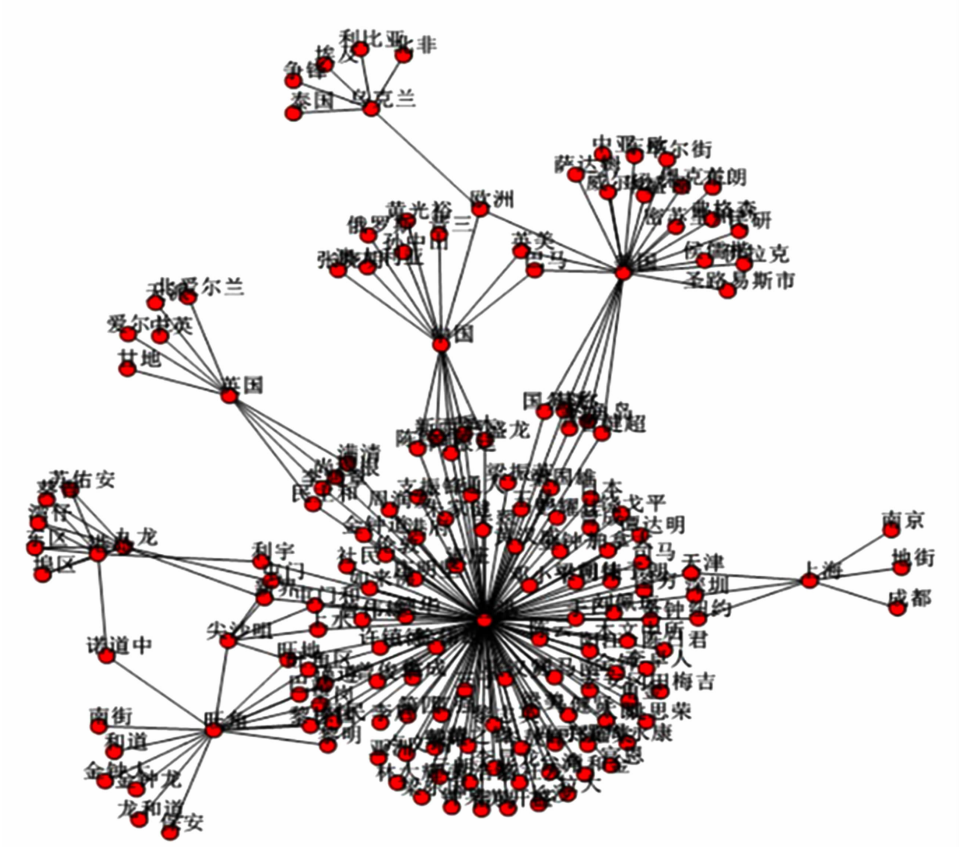

Fig. 3 Abnormal behavior co-occurrence network 2.

图 3 异常行为共现网 2

从图 3 可以看出,在香港占中事件中,度数最大的一些节点分别是“香港”、“美国”、“中国”、“英国”、“九龙”、“旺角”等,这些都是该事件中比较重要的一些地点,需要重点分析它们之间的联系,网络中还有与这些关键节点共现过的其他节点,可以进行辅助分析,通过分析可以掌握关键节点之间的隐含联系,为舆情分析和决策做出一定的支持.

对于 2 种共现网络,我们都可以对其进行异常路径分析和焦点分析,对于网络中的任意 2 个实体节点,我们都可以找出它们之间的所有路径和最短路径,从而分析该实体对与路径上的实体之间的关系以及是如何通过这些实体进行连接的. 例如在图 4 中,“周永康”和“朱耀明”这 2 个支持占中的人名之间的最短路径上就有“香港”和“梁振英”. 拥有较高度数的节点,与其他较多节点之间有最短路径的节点,其他节点对之间的最短路径通过次数较多的节点等具有较高的“中心性”,把中心性高的节点作为网络中的焦点进行重点分析,也可以根据网络的动态变化来动态跟踪关键实体. 这些分析都是现实可行的,可以挖掘出某些热点事件中隐含的信息,对决策做出一定的支持,有很重要的理论和现实意义.

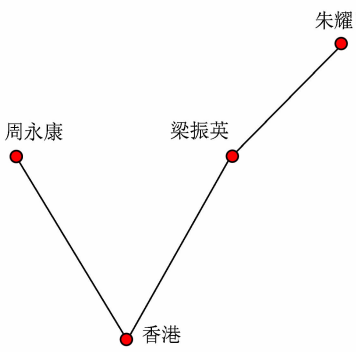

Fig. 4 Analysis of abnormal path.

图 4 异常路径分析

4 总结与展望

4.1 工作总结

本文主要探索了对网络舆情文本进行事件抽取的研究,采用事件抽取的主流会议 ACE 中对事件和事件抽取子任务的定义,结合本文所依托的课题背景和网络舆情分析的需求,对网络舆情分析中关注的 4 类异常行为进行抽取.

本文通过对真实数据的实验验证了该原型系统的有效性和可行性.论文的主要工作可以总结如下:

1) 根据 ACE 中对事件的定义并结合本文的研究需求,明确本文中事件的定义.结合网络舆情的研究现状和本文的研究需求,确定本文的抽取目标.

2) 对网络舆情数据进行事件识别和触发词检测,过滤干扰数据.

3) 使用合适的特征来训练异常行为识别分类器,进行行为类别识别.

4) 构建异常行为共现网,为舆情分析提供可视化的研究方法.

5) 开发基于本文工作的原型系统,并在真实数据上进行验证.

4.2 未来展望

面向网络舆情数据的异常行为识别是一个非常有意义的方向,可以从 3 个方面对本文的工作进行扩展和改进:

1) 完善未知触发词识别.本文目前采用的触发词表是固定大小的,包含了大部分的触发词,必定也会有一些触发词的遗漏,这样有一些异常行为句就在触发词检测时被错误过滤掉,影响了整体的性能.因此我们在下一步的工作中需要完善触发词表,可以采用基于词语构词结构和语义相似度的方法来识别未知触发词,并将其加入原有触发词表中,或者在本文方法的基础上使用基于监督的方法来动态扩展触发词表.

2) 完善对分类特征的选择.本文采用的分类特征是全词特征,更加注重的是句子在词法方面的信息,而要对异常行为进行分类,仅仅有词法信息是不够的,因此我们下一步需要在特征中加入更多的语法和语义信息,比如可以选择句子中触发词左右的 n 个词及其词性作为分类特征,完善异常行为识别分类器的分类效果.

3) 完善对异常行为共现网的分析.本文构建了异常行为共现网,但并没有进行共现网的详细分析.因此未来需要完善网络的分析,更加明确地体现出异常行为共现网的应用价值.

参 考 文 献

[1] Grishman R. Message Understanding Conf (MUC) [EB/OL]. Philadelphia, PA: University of Pennsylvania1. (2002-07-01) [2013-07-21]. http://en.wikipedia.org-/wiki/Message_Understanding_Conference

[2] Garofolo J. Automatic Content Extraction (ACE) [EB/OL]. Philadelphia, PA: University of Pennsylvania1. (2005-07-01) [2013-07-21]. http://www.itl.nist.gov/iad/mig/-tests/ace/2005

[3] Riloff E, Shoen J. Automatically acquiring conceptual answer patterns without an annotated corpus [C] //Proc of the 3rd Workshop on Very Large Corpora. San Francisco: Morgan Kaufmann, 1995: 148-161

[4] Yangarber R. Scenario customization for information extraction [D]. New York: New York University, 2001

[5] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text [C] //Proc of the 18th National Conf on Artificial Intelligence. Edmonton, Alberta: American Association for Artificial Intelligence, 2002: 786-791

[6] Jiang Jifa. A method to do Chinese event ie from a multiple sentences' event narration [J]. Computer Engineering, 2005, 31(2): 27-29 (in Chinese)
(姜吉发.一种跨语句汉语事件信息抽取方法[J].计算机工程, 2005, 31(2): 27-29)

[7] Ahn D. The stages of event extraction [C] //Proc of the Workshop on Annotations and Reasoning about Time and Events. Stroudsburg, PA: Association for Computational Linguistics, 2006: 1-8

[8] Yu Jiangde, Xiao Xinfeng, Fan Xiaozhong. Event information extraction from Chinese text based on hidden Markov models [J]. Microelectronics & Computer, 2007, 24(10): 92-94 (in Chinese)
(于江德,肖新峰,樊孝忠.基于隐马尔可夫模型的中文文本事件信息抽取[J].微电子学与计算机, 2007, 24(10): 92-94)

[9] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction [C] //Proc of Human Language Technologies: The 2009 Annual Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2009: 209-212

[10] Fu Jianfeng, Liu Zongtian, Fu Xuefeng, et al. Dependency parsing based event recognition [J]. Computer Science, 2009, 36(11): 217-219 (in Chinese)
(付剑锋,刘宗田,付雪峰,等.基于依存分析的事件识别[J].计算机科学, 2009, 36(11): 217-219)

[11] Llorens H, Saquete E, Navarro-Colorado B. TimeML events recognition and classification learning CRF models with semantic roles [C] //Proc of the 23rd Int Conf on

Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 725-733

[12] Xu Honglei, Chen Jinxiu, Zhou Changle, et al. Research on event type identification for Chinese event extraction [J]. Mind and Computation, 2010, 4(1): 34-44 (in Chinese)
(许红磊, 陈锦绣, 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)

[13] Chen Yanping, Zheng Qinghua, Zhang Wei. Omni-word feature and soft constraint for Chinese relation extraction [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2014: 572-581

Hao Yazhou, born in 1989. PhD candidate in Xi'an Jiaotong University. Student member of China Computer Federation. His research interests include data mining, natural language processing and social media mining.

Zheng Qinghua, born in 1969. Professor and PhD supervisor. His main research interests include multi-media e-learning, computer network security, intelligent e-learning theory and algorithm.

Chen Yanping, born in 1980. PhD candidate in Xi'an Jiaotong University. His research interests include natural language processing, information extraction and data mining.

Yan Caixia, born in 1992. Master candidate in Xi'an Jiaotong University. Her research interests include information extraction and data mining.