突发事件Web信息挖掘方法及应用

蔡华利　潘守慧　杨跃翔◎著

Methods and Application
of Web Information Mining under Emergencies

中国质检出版社
中国标准出版社

突发事件Web信息挖掘方法及应用

Methods and Application
of Web Information Mining under Emergencies

蔡华利　潘守慧　杨跃翔　著

中国质检出版社
中国标准出版社

北　京

图书在版编目(CIP)数据

突发事件 Web 信息挖掘方法及应用/蔡华利,潘守慧,杨跃翔著.—北京:中国标准出版社,2014.11
ISBN 978-7-5066-7714-1

Ⅰ.①突… Ⅱ.①蔡… ②潘… ③杨… Ⅲ.①突发事件-互联网络-新闻报道-信息处理-研究 Ⅳ.①G210.7

中国版本图书馆 CIP 数据核字(2014)第 212871 号

中国质检出版社
中国标准出版社 出版发行
北京市朝阳区和平里西街甲 2 号(100029)
北京市西城区三里河北街 16 号(100045)
网址:www.spc.net.cn
总编室:(010)64275323 发行中心:(010)51780235
读者服务部:(010)68523946
北京京华虎彩印刷有限公司印刷
各地新华书店经销
*
开本 880×1230 1/32 印张 4.875 字数 118 千字
2014 年 11 月第一版 2014 年 11 月第一次印刷
*
定价 20.00 元

前　言

21世纪以来，各类突发事件频发，而且呈现接连不断的状态，如最近中国发生的“鲁甸地震”事件以及西非的“埃博拉病毒”事件对人身健康和财产都带来了巨大的损失，可以说人类社会面临前所未有的危机和灾难挑战。突发事件应急管理越来越受到世界各国的重视，建立应急管理体系，运用科学的理论、方法和工具，提高突发事件预警、响应、追踪处置和恢复能力在受到政府部门关注的同时，也成为学术界研究的重要科学问题。

近几年，网络技术的快速发展加快了信息的传播速度，丰富了信息传播的内容和观点。突发事件发生后，民众会通过多种渠道（如综合门户网站、网络论坛、点对点通信等）获取和扩散各种类别的Web信息，这些信息包含媒体对事件本身的描述、政府对事件进展的回应及相关应对措施，也包括广大民众发出的心理诉求甚至谣言等信息，信息量在短短的几天内会呈现指数增长，积累成海量信息甚至是这一事件主题的大数据。如何充分挖掘这些信息，以对政府的应急管理工作提供有价值的信息资源以及对民众提供真实、确切的信息是近年来学术界研究的热点，也是撰写本书的出发点。本书主要以突发事件发生后互联网上产生的Web信息为研究对象，以网络环境下的突发事件应急管理创新工作为研究目标，研究探讨几类突发事件Web信息的挖掘方法，并通过在实际事件中实践对所提方法进行改进和完善。

本书共分七章，中国标准化研究院蔡华利主要撰写第一章部分内容、第二章、第三章、第四章及第六章和全书的通稿及校对工作，国家农业信息化工程技术研究中心潘守慧主要撰写第五章、第七章，中国标准化研究院杨跃翔负责第一章部分内容的撰写工作。全书内容摘要如下：第一章介绍与本书相关的概念，并对相关研究方向的国内外研究现状给予综述，为本书的研究工作奠定了基础。第二章研究突发事件Web新闻的精确分类方法，该方法首先分析了突发事件Web新闻的分类，给出三层分类器的构造方法，第一层和第二层通过规则定制来完成，第三层通过统计学习来训练并实现，试验结果表明，所提方法的分类效果优于其他方法。第三章提出从Web新闻中提取突发事件发生时间的抽取方法，从突发事件Web新闻的时间构成、时间位置特征以及时间常用词三个方面分析突发事件Web新闻的表达特征，并提出突发事件Web新闻的时间抽取方法，试验证明，效果比较理想。同时还将抽取到的事件发生时间与Timeline工具结合，实现突发事件新闻按时间顺序的可视化展示。第四章研究并识别突发事件发生的地点信息，提出基于规则推理的地名实体识别方法，从辖区范围变化规律、所处位置分布规律、多地名实体同现的情形等方面分析突发事件地名实体在Web新闻中的表达特征，构造突发事件地名实体的多个识别规则，提出事件发生地点的识别方法。试验结果表明，识别精确率理想，并将抽取到的突发事件地名信息和地图工具进行结合，实现突发事件新闻按地点位置的可视化展示。第五章研究突发事件话题跟踪技术，能够将孤立、零散的网络信息串联起来并掌握事件发展态势，通过引入时间距离和考虑网页的标题、元数据和正文之间的内容相似度等要素，改进了现有的模型，试验证明方法理想。第六章利用Web挖掘

方法尝试研究并测度了突发事件主题的破坏性，对突发事件主题、主题破坏性、破坏特征的维度进行定义，并构建破坏词表，分别研究单条 Web 文档和突发事件主题的破坏指数测度方法。试验证明，所提方法和事件自身表现的破坏程度基本吻合。第七章提出了质量安全领域突发事件个性化信息服务的用户兴趣挖掘方法，通过分析用户浏览过程中的多种行为以挖掘用户关注的兴趣点，基于遗忘曲线原理给出了用户兴趣模型的动态更新方法。经试验验证，本方法能有效地挖掘用户兴趣。

本书以质检行业公益性项目《“双打”中日用消费品质量状况动态监测及分析研究》、国家自然科学基金项目《基于 Web 的农产品安全信息传播与干预模型研究》、国家社会科学基金重点项目《我国质量安全评价与网络预警方法研究》、国家科技支撑计划项目《产品质量安全风险监测及信息分析技术研究》的研究成果为基础，结合突发事件的具体特征，形成了一些创新性的研究成果。本书适用于从事应急管理工作的政府、行业和企业人员，为高效分析突发事件 Web 信息、掌握事件状态提供帮助；同时，对公共管理、管理信息系统、计算机等相关专业的教师、学生开展教学和研究工作也有较大参考作用。

由于作者水平有限，不当之处敬请读者批评指正。

蔡华利

2014 年 8 月

目　录

第一章　绪论

第二章　突发事件 Web 新闻多层次自动分类方法

第三章　突发事件 Web 新闻中时间信息分析及抽取

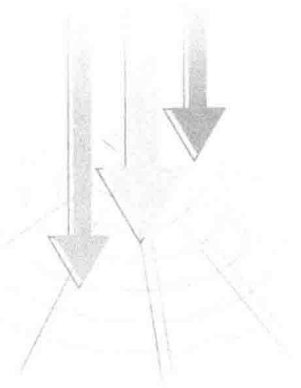
第四章　基于规则推理的突发事件发生地点识别

第五章　突发事件 Web 新闻话题跟踪

第六章　基于 Web 挖掘的突发事件破坏指数测度

第七章　产品质量安全突发事件用户个性化信息服务兴趣建模

第一章　绪论



第一节　研究背景和意义

一、研究背景

进入21世纪以来，各类突发事件频发，人类社会面临前所未有的危机和灾难挑战。突发事件应急管理越来越受到世界各国的重视，建立应急管理体系，运用科学的理论、方法和工具，提高突发事件预警、响应、追踪处置和恢复能力在受到政府部门关注的同时，也成为学术界研究的重要科学问题。

20世纪60年代伊始，突发事件应急管理开展了初期的探索性研究，20世纪90年代后期转向对各类公共危机和企业危机事件的应对方法和管理框架的研究，至今已进入体系化研究阶段，研究对象涉及自然灾害、事故灾难、公共卫生和社会安全等各类突发事件，研究主题涵盖突发事件的信息处理与演化规律建模、应急决策理论、紧急状态下个体和群体的心理反应与行为规律等多项内容，研究视角呈现出管理学、社会学、信息学、传播学和心理学等多学科特点。

近几年，网络技术的发展加快了人与人之间的沟通速度，

突发事件发生后，民众会通过多种渠道（新闻发布网站、网络论坛、网络日志、点对点通信）发布和扩散各种类别的 Web 信息，这些信息包含对事件本身的描述、广大民众真实的心理诉求、谣言信息等，有些信息如果不进行人为干预，可能会导致一系列新的社会稳定、和谐发展及国家安全的问题[1]；除了民众发布和扩散的各类 Web 信息之外，各类新闻媒体、政府机构在事件的进展过程中也会实时发布相关的新闻。利用 Web 挖掘技术能够从某种程度上识别民众的心理诉求、发现事件的发展态势，且能够在事件发展的过程中在适当时机做出预警决策，并从某种程度上减缓事件向更恶劣的方向发展。

Web 挖掘技术包括内容挖掘、日志挖掘及结构挖掘。经过多年的发展，Web 挖掘在电子商务、知识管理以及数字图书馆等领域已经取得了很多有意义的研究成果。其中，部分研究成果证明，利用已有的 Web 挖掘技术，可以较高效率地实现对突发事件 Web 信息进行分类、聚类，可以实现对关键信息的提取以及其他关键领域 Web 知识的获取。对于实时监测事件的发展、民众的心理情绪、事件的变化趋势和发展规律，进而为政策干预、预警决策提供了科学依据。基于 Web 挖掘技术的应急管理模式能够促使应急管理工作向着数据驱动模式的方向发展，很大程度上提高了应对突发事件的整体能力，所以基于 Web 挖掘技术的应急管理将是未来几年应急管理及 Web 挖掘领域的研究热点问题。

二、研究意义

突发事件预警干预是应急管理的关键环节之一。预警干预包括预警分析和干预两项任务。预警分析是对突发事件各阶段的相关信息进行监测、识别、诊断与评价并及时报警的管理活动；干预是在预警分析的基础上，对不良趋势进行纠正、预防与控制的管理活动。

信息资源是突发事件预警分析的重要依据。在网络化信息时代，除了制度化渠道之外，非制度性渠道中的 Web 信息资源汇集了各类事件新闻的报道和舆论评价，是突发事件的重要信息平台，对事件信息管理与控制具有广泛作用[2]。然而，在突发事件的 Web 信息融合与分析方面，我国仍处于以手工方式为主，以智能计算为辅的低水平状态，先进的 Web 信息分析技术在我国应急管理体系中尚未得到充分应用。

目前，Web 文本信息抽取、主题检测与跟踪由于其广泛的应用前景受到世界各国学者的关注，取得了有意义的研究成果。在这些领域相关理论方法的指导下，研究突发事件的 Web 信息分类、信息抽取、主题跟踪、破坏性测度以及相关的预警方法，对应急管理部门发现“未然态”危机因素，跟踪和评估危机态势，进而实现事件预警具有重要意义[3]。为此，需要解决如下问题：①如何对突发事件 Web 新闻进行多层次分类，以快速定位、查找所需信息；②如何从海量的 Web 信息中抽取事件的发生时间和发生地点，自动、快速地确认事件的基本信息和发展状态；③如何能有效地把零散、孤立的信息汇集并组织起来，从而可以帮助人们全面地掌握某个事件的整体发展态势以及事件相互之间的关联性；④如何测度突发事件的破坏性，为政府决策与预警提供直接依据；⑤如何构建基于 Web 挖掘成果的预警策略，为群众提供事件发展态势、为政府提供决策支持等；⑥如何及时、准确地向用户推荐其关注的突发事件信息等。

本书的研究成果将为突发事件的 Web 信息管理和预警决策提供理论基础、技术支撑和应用范例，进而推动我国应急管理工作向数据驱动模式方向发展，从而提高应急预警、准备、响应和恢复过程的科学性。因此，本书的研究工作具有较大的理论意义和广泛的应用价值。

第二节　相关概念

一、突发事件

“突发”一词含有突如其来、出乎预料、令人猝不及防的意思；“事件”一词，按照《辞海》的解释，是指历史上或社会上发生的大事情。目前，“突发事件”还没有公认的界定。本书采纳《国家突发公共事件总体应急预案》[4]对突发事件的定义——造成或者可能造成重大人员伤亡、财产损失、生态环境破坏和严重社会危害，危及公共安全的紧急事件。根据突发事件的发生过程、性质和机理，突发事件主要分为以下四类：

（1）自然灾害，主要包括水旱灾害、气象灾害、地震灾害、地质灾害、海洋灾害，生物灾害和森林草原火灾等。

（2）事故灾难，主要包括工矿商贸等企业的各类安全事故、交通运输事故、公共设施和设备事故、环境污染和生态破坏事件等。

（3）公共卫生事件，主要包括传染病疫情、群体性不明原因疾病、食品安全和职业危害、动物疫情以及其他严重影响公众健康和生命安全的事件。

（4）社会安全事件，主要包括恐怖袭击事件、经济安全事件和涉外突发事件等。

二、Web 信息

在宏观层面上，Web 信息可以理解为通过计算机网络获取的各种信息资源的总和，而本书关注的 Web 信息仅包括 Web 新闻、博客和 BBS 上发布的各类与某主题相关的信息以及在互联网其他渠道中发布的各种与某主题相关的评论信息等。在这些 Web 信息中，Web 新闻是本书重点关注的信息类型。

Web 新闻，指通过因特网发布、传播的新闻，其途径包括万维网网站、新闻组、邮件列表。国务院新闻办公室、工信部联合发布的《互联网站从事登载新闻业务管理暂行规定》[5]规定了 Web 新闻发布机构应具有的资质、条件等。

三、突发事件 Web 信息

突发事件 Web 信息是突发事件发生后，互联网上针对该事件报道及转载的各种新闻、专家发表的评论、政府的应对举措、网民在博客中撰写的博文以及其他与突发事件有关的 Web 文档等。

随着 Web 技术的发展以及互联网用户的增加，突发事件 Web 信息在事件的诞生、发展过程中以惊人的速度传递，Web 信息的数量也呈指数增加。如自鲁甸地震发生以来，相关的 Web 信息以难以置信的速度增加，截至 2014 年 8 月 8 日，百度搜索引擎收录了 60 余万条 Web 信息。

四、应急管理

2003 年以前，关于应急管理的研究主要集中在灾害管理方面。2003 年抗击“非典”的过程中暴露了我国政府管理特别是应急管理工作中的若干薄弱环节，同时推动了我国应急管理理论与实践的发展。

应急管理作为一门新兴的学科，受到国内外相关专家空前的关注，比较有代表性的定义有：

(1) 美国联邦应急管理署(FEMA)[6]将应急管理界定为面对突发事件时准备、缓解、反应和恢复的过程，并且认为这个过程是动态的，每一个子过程互相关联。

(2) Hoetmer[7]将应急管理定义为一门应用科学、技术、计划和管理等多方面的知识来处理或管理可以造成民众伤亡、财产损失或者严重影响社会正常生活秩序的突发事件，以避免或

减少这些突发事件所造成严重损失的学科。

(3) 国内应急管理领域专家计雷等人[8]将应急管理定义为在应对突发事件的过程中,为了降低突发事件的危害,达到优化决策的目的,基于对突发事件的原因、过程及后果进行分析,有效地集成社会各方面资源,对突发事件进行有效预警、控制和处理的过程。

本书认为应急管理是指政府及其他公共管理机构在突发事件的事前预防、事发应对、事中处置和善后处置过程中,通过建立必要的应对机制,采取一系列必要的措施和手段,保障公众生命财产安全、促进社会和谐健康发展的系列活动。

五、Web 挖掘

Web 挖掘是利用数据挖掘技术从 Web 文档和服务中自动发现和获取信息,对 Web 上的有用信息进行分析[9,10],是一个发掘特定用户感兴趣的、有用的、以前不知道的信息或知识的过程。它可以将 Web 文档进行分类,抽取主题,分析用户浏览站点的行为特点,以帮助用户获取、归纳信息,改进站点结构,为用户提供个性化服务等。处理的信息包括 Web 文本、Web 图片、Web 视频、Web 日志等各种媒体信息。

Web 挖掘包括 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。

(1) Web 内容挖掘,指对 Web 页面内容及后台交易数据库进行挖掘,从 Web 文档内容及其描述中的内容信息中获取有用知识的过程。

(2) Web 结构挖掘,指对 Web 的组织结构和链接关系进行挖掘,从人为的链接结构中获取有用知识的过程。

(3) Web 使用挖掘,主要通过分析用户访问 Web 的记录了解用户的兴趣和习惯,对用户行为进行预测,以便于提供个性化的产品信息和服务。挖掘的数据是用户与 Web 交互过程中

留下的用户访问过程的数据[11]。

本书研究的对象是突发事件发生后网络上传播的各类报道的内容、标题以及时间信息等，这些信息一般会存在于 Web 页面的内容之中，所以本书涉及的研究方法属于 Web 内容挖掘的研究范畴。

第三节 国内外研究现状

突发事件的频繁发生，使得国家和民众蒙受难以估计的经济和精神损失。突发事件的应急管理研究也逐渐被提上日程。随着网络多媒体技术的发展，Web 信息已经成为与突发事件应急管理不可分离的组成部分。在应急管理以及 Web 挖掘等领域，目前已经有较多的研究成果。

本节针对与本书相关的研究方向及其研究成果进行综述。

一、Web 信息话题检测与跟踪

话题检测与跟踪（Topic Detection and Tracking，TDT）是与突发事件 Web 信息分析密切相关的一个热点研究领域。TDT 研究始于 1996 年，是对新闻媒体信息进行未知话题识别和已知话题跟踪的信息处理方法。TDT 研究涉及三个基本概念：①话题（Topic），由一个核心事件或活动，加上与其直接相关的其他事件和活动构成[12]；②事件（Event），发生在特定时间、地点，有一定参与者或涉及者，并可能伴随某些结果的一个事件[13]；③报道（Story），指与某个话题紧密相关，包含事件陈述句子的新闻片段。

2004 年美国国家标准与技术研究院（NIST）将 TDT 研究分为五项子任务：新闻报道切分、话题跟踪、话题检测、首次报道检测和关联检测。其中，话题检测和话题跟踪是 TDT 研究的核心，以下重点综述与此相关的国内外研究进展。

（一）话题检测

话题检测包括以下子任务：

（1）话题/报道的模型化

常用建模方法包括向量空间模型[14,15]和语言模型[16,17]，但均存在特征空间的数据稀疏性问题。Ponte 等[18]基于特征上下文对向量空间模型进行了扩展，该研究选择权重较大的特征作为扩展对象，不仅有助于解决数据稀疏问题，同时可以削弱特征的歧义性。而解决话题/报道建模数据稀疏性和歧义性问题的研究成果较少。

（2）话题与报道、报道与报道间的相似度计算

该领域的研究主要借助自然语言处理和统计方法，抽取报道文本的命名实体（Named Entities，NE）[19-21]，实现基于内容的相似度计算。其中，跨语言命名实体识别成为多语言新闻话题检测的研究重点[22-26]。此外，将时间特征引入检测模型，构造衰减函数，改进基于内容的相似度计算，可以提高话题检测性能。

（3）基于聚类的话题检测算法

话题检测旨在将陆续到来的新闻报道聚类到不同的话题簇，相当于无指导增量聚类过程[27]。对聚类算法进行选择与融合，以适应新闻数据流特征是话题检测研究的一个重要内容。除选用 K 均值、凝聚法等传统聚类算法外，数据流聚类挖掘逐渐成为话题检测的主流技术[28-30]。此外，随着层次话题检测（Hierarchical Topic Detection，HTD）概念的提出，寻找适于 HTD 任务的聚类机制成为 TDT 研究的另一热点。HTD 的研究输出是一个非循环有向图（Directed Acyclic Graph，DAG），它表达了话题内部子话题、子事件间的层次关联，相比传统话题检测，更能体现事件信息的本质。已提出的 HTD 代表方案包括：基于种子报道构建初始 DAG 体系，再以增量方式插入新