白华，林勋国. 基于中文短文本分类的社交媒体灾害事件检测系统研究[J]. 灾害学，2016，31(2)：19－23. [Bai Hua and Lin Xunguo. Sina Weibo Disaster Information Detection Based on Chinese Short Text Classification[J]. Journal of Catastrophology，2016，31(2)：19－23.]

基于中文短文本分类的社交媒体灾害事件检测系统研究*

白 华[1]，林勋国[2]

（1. 哈尔滨工业大学 管理学院，黑龙江 哈尔滨 150001；2. 澳大利亚联邦科工院，澳大利亚 堪培拉 2601）

摘 要：随着移动互联业务的蓬勃发展，在灾害信息传播的过程中，不同类型的社交媒体在一个个突发性灾害事件中显示出了强大的力量。以微博为代表的在线社交媒体因在信息传播速度、传播内容、传播形式及传播效果等方面的优势，确立了其在灾害应急管理中特殊的传播价值。鉴于此，利用成熟的文本挖掘技术，面向中文新浪微博平台，开发了高效的灾害事件即时检测系统，从而能充分利用近于实时的灾害博文数据，使其更好地为灾害应急管理过程服务，有效提高灾害的应急管理能力。

关键词：社交媒体；新浪微博；灾害信息；灾害检测

中图分类号：X43 **文献标志码**：A **文章编号**：1000－811X(2016)02－0019－05

doi：10.3969/j.issn.1000－811X.2016.02.005

进入到 21 世纪以来，自然灾害在世界范围内造成了严重的经济损失和人员伤亡，毋庸置疑，日益恶化的自然环境及不断加快的城市化进程加重了这一趋势。遗憾的是，许多突发重大灾害难以及时预测，灾害的影响区域及人群也难以准确估计。由此，有效提高灾害的应急管理能力至关重要。信息的收集、处理和交流是突发灾害应急管理过程中的重大挑战。充足、准确、及时的灾害信息在防灾减灾过程中发挥着重要作用，可以有效降低灾害风险，减少灾害损失。

随着移动互联业务的蓬勃发展，以微博为代表的社交媒体应用已经成为人们生活中不可缺少的重要组成部分，微博等社交媒体平台也已成为灾害信息管理过程中的重要信息来源和沟通媒介。2004 年亚洲海啸事件中，许多第一手资料及统计来自社交网络，包括幸存者的经历、新闻信息的发布、救援努力、人道主义援助以及灾害情绪释放等等[1]；美国红十字会的调研[2]表明，当灾害事件发生后而 911(应急)电话无法接通时，20% 的美国民众通过移动应用收到灾害信息，76% 的美国民众通过社交媒体发布求助信息，40% 的美国民众在灾害事件中采用社交媒体与亲人取得联系；2013 年 10 月，受台风“菲特”的影响，浙江余姚遭遇了建国以来的最大降雨量，由于部分通讯及交通基础设置瘫痪，救援人员无法及时进入灾区，余姚市部分县镇成为了一座座“孤岛”，在无线网络仍能发挥作用的情况下，社交网络扮演了信息高架桥的角色。KCIS 观察发现[3]，灾害发生一周内，关于“余姚水灾”的微博搜索超过 30 万条，而且很多灾民利用评论及回复功能通过“@余姚发布”来发布受困求助信息。

近几年，美国、澳大利亚、日本等国家先后开展相关领域的研究，并取得较大的进展。他们相继开发了“Did You Feel it?”[4]，“Toretter”[5－6]，“Twicident”[7]，“Tweet4act”[8]，“CrisisTracker”[9]，“Ushahidi platform”[10]，“Twitter Earthquake Detector”[11]，“Emergency Situation Awareness”[12]，“EARS”[13]等面向互联网用户及社交媒体（Twitter）用户的灾害事件检测应用系统。

中国幅员辽阔，人口众多，自然灾害频繁发生。近些年，随着经济和科技的快速发展，越来越多的中国人拥有电脑和手机，并开始使用在线社交媒体。根据中国互联网信息中心的报告[15]，截至 2013 年末，中国微博用户达到了 2.81 亿，其中接近 70% 用户用手机的方式登陆微博账号。因此，中文灾害微博信息的研究势在必行，并且具有充分的数据资源。

然而，当前国内对利用社交媒体来进行灾害

* 收稿日期：2015－09－16 修回日期：2015－11－07
基金项目：国家自然科学基金资助项目(71372091)，国家留学基金委公派联合培养博士生资助项目(201306120166)
作者简介：白华(1985－)，女，辽宁沈阳人，博士研究生，主要研究方向为灾害信息学. E-mail：baihua1727@163.com

信息管理的趋势尚未足够重视，相关研究成果较少。Qu 等[16]和 Zhou 等[17]均对青海玉树地震后的相关微博进行了研究，前者主要分析了灾害相关微博的内容、趋势和扩散路径，后者则从救援角度采用贝叶斯算法将灾害博文分类，但是，二者均未涉及灾害事件的检测方法研究。

基于此，本文利用自然语言处理及文本挖掘技术，面向中文微博平台，开发高效的灾害事件检测方法，从而充分利用中文灾害微博数据，使其更好地为灾害应急管理过程服务，有效提高灾害的应急管理能力。

1 中文短文本分类

1.1 分类方法

经过观察可以发现，灾害爆发后，微博平台往往会在在短时间内产生大量的相关博文，存在严重的信息冗余现象。因此，我们需要对相关博文进行文本分类，从而为后续救援提供及时、准确的灾区信息。微博文本一般比较短，且在表达方式上非常口语化，经常包含大量的表情符号、标点符号及网络用语等，这一特点为文本分类领域的研究提出了很大的挑战。本研究过程中主要讨论了四种常见的文本分类算法：支持向量机(Support Vector Machine)、朴素贝叶斯(Na? ve Bayes)、K 近邻(K Nearest Neighbor)及随机森林(Random forests)。这四种方法均在传统文本分类领域中取得了很好的分类效果，但在面对不同特征属性的样本时表现各异，各有优劣[18-20]。

1.2 训练集

为了训练事件分类器，需要收集历史微博数据作为训练集。根据新浪微博 API 的调用方法，我们在新浪活跃用户中随机选择 50 000 用户作为采集目标，采集其最新的 1 000 条微博信息。由于很多用户历史微博信息数量尚未达到 1 000 条，最终，历史数据集中包含了近 2 600 万条微博信息。在此基础上，我们采用关键词(如“地震”等)过滤的方式获取灾害微博数据集。然后，在这个数据集中进行人工筛选，抽取与灾害事件相对应的即时灾害信息作为 Positive 数据集(标签“+”)，同时随机抽取等量的不相关信息(此类信息也包含灾害关键词但不为即时信息)作为 Negative 数据集(标签“-”)。筛选后的训练集如表 1 所示。经过人工筛选及标注，地震相关微博文本训练集合计包括了 934 条信息(其中 467 条含即时相关信息，另一半含非即时信息或非相关信息)。

1.3 特征选择

在对训练集数据进行观察后，可以发现即时地震信息往往较短，包含问号或感叹号，文字中经常提到“晃”、“摇”等描写地震感觉词语。为了更好地进行特征提取，不遗漏重要特征，在分类器构造过程中，我们采用了 10-fold 交叉方法对所有特征组合进行了测试。因此，针对每一个分类算法，我们进行了 $2^8-1=255$ 次试验，分别获取了各个分类算法中任一特征组合的准确率(Accuracy)、F1 值、精确率(Precision)及召回率(Recall)．根据测试结果，最终为四个分类算法选取最优特征组合如表 2 所示。

1.4 训练集最优规模测试

经过上一节所示的特征提取过程后可以发现，支持向量机分类器表现最为优异，F1 值达到了到 0.890。但是，由于我们是预设的训练集，因此尚不确定训练集规模变化对各个分类器表现的影响，也不确定更大规模的训练集是否可以取得更好的分类精度。由此，需要进行最优训练集规模测试，以确定不同大小的训练集规模对测试结果的影响。

从图 1 中可以看出，地震数据集的最优训练规模大致为 600 条信息，且随着地震数据量的增加，准确率、召回率及 F1 值之间的差异逐步缩小，这说明扩大训练集规模对提高分类器精度是无意义的。

表 1 训练集示例

信息内容	类标签
有地震千万要是晚上啊	-
地震！抖的好凶哦!!	+
那些牛蛙叫一晚上了，吵死了 ~~~~~~~~~~~~~~~~~~~~是不是又要地震了。。。[怒][怒]	-
地震！地震！地震！	+
地震，震的不是地，应该是人的灵魂…祈祷	-

表 2 最优特征组合

分类器	最优特征组合结果
Support Vector Machine	Char count, link count, question mark count, exclamation mark count and unigrams
K Nearest Neighbor	Exclamation mark count and unigrams
Naïve Bayes	Links count, word count, chars count, question mark count and exclamation mark count
Random forests	All features

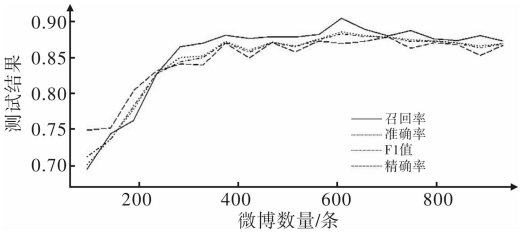

图 1　最优规模测试结果

2　系统框架及其可视化

2.1　系统框架

本文主要解决的问题是检测灾害事件发生后的即时微博相关信息，从而为后续的救援过程提供帮助。因此基于网络舆情计算的基本流程，本系统的基本框架设计如图 2 所示。

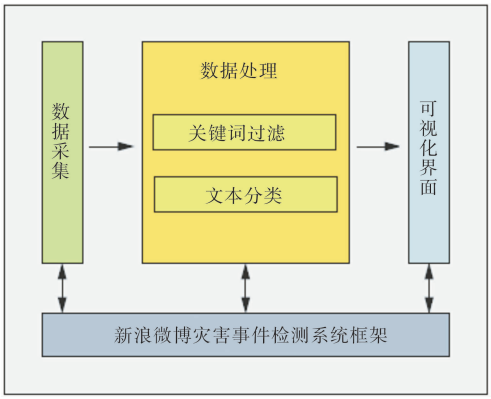

图 2　新浪微博灾害事件检测系统框架结构

2.2　数据采集

新浪微博 API 为开发者提供了不同目标的数据调用接口，本文介绍的新浪微博灾害事件检测系统中主要调用“statuses/public_ timeline”接口，从而获取最新的公共微博。根据新浪微博数据开放平台介绍，这一接口单页可以返回最多不超过 200 条信息(博文)。由此，基于新浪微博 API 对用户请求的限制(每小时不超过 150 次，即每 24s 可以发送一次请求)，本系统系统近于实时的数据流量大致为每小时近 30 000 条或每天大约 72 万条(如图 3 所示)。采用这一方法方法获取的公共微博是随机的，没有指定用户，因此可以排除统计偏差。

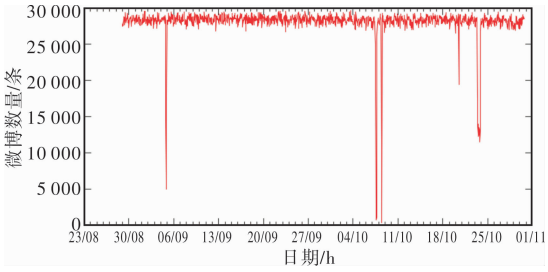

图 3　数据采集量示意图

2.3　数据处理

数据处理模块主要包括两个步骤，一是实现数据的实时过滤，二是对过滤后的数据进行文本分类。由于系统采集的数据包括大量信息，其中仅含有部分相关信息，噪声量巨大。为了简化系统的计算过程，实现即时检测目的，本研究采用关键词过滤的方法实现对大量数据的实时过滤。在对历史数据进行文本分析的基础上，选取灾害密切相关的关键词作为过滤词。经过测试，系统当前采用的过滤词如表 3 所示。

表 3　系统预设关键词列表

灾害类型	词语
地震	地震
洪水	洪水、水灾、淹、大水、涨水、涨大水、涝、积水
台风	台风
暴雨	风暴、暴风雨、暴雨、大雨
火灾	大火，火灾，失火

基于短文本分类实验结果，系统当前采用支持向量机作为文本分类算法，对过滤后的相关信息进行分类。除上文所述地震即时信息分类器外，我们使用相同的方法，在历史数据的基础上，面向火灾、暴雨、台风、洪水分别进行特征选择和训练集规模测试，为各个灾种构建了即时灾害信息分类器。

2.4　可视化

为了更加直观地呈现微博信息灾害检测过程，我们面向新浪微博平台开发了灾害事件检测系统界面（SWIM，https：//swim. csiro. au/swim/index. html)。如图 4 所示，这一界面主要由四部分组成。

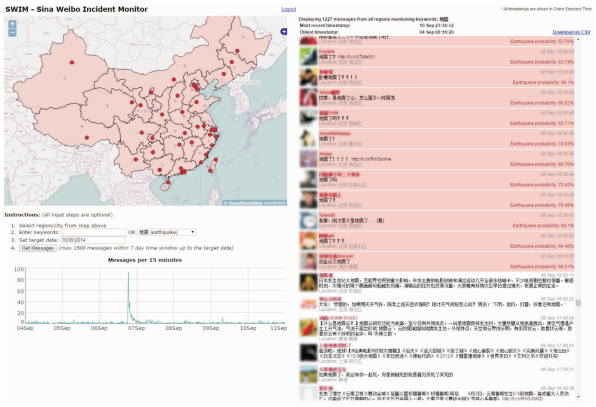

图 4　SWIM 系统界面

(1)中国行政区划图。地图起用于 OpenStreetMap 的界面，可以放大、缩小或移动，应用者可以根据地图选择目标省域，自定义检测地区。如果选择了一个省或市，系统的关键词搜寻将集中在来源于这个地区的博文里进行，若应用者未使用这一功能，则系统默认在全网范围内进行灾害爆发检测。

(2)自定义功能区。这一区域位于地图下方，应用者可以自定义检测时间段及检测关键词。SWIM 系统提供了“系统预设关键词”及“用户自定义关键词”两种关键词过滤方法，提高了系统的灵

活性，同时有效地扩大了系统的应用范围。此外，用户还可以自定义搜索时间段。

(3)关键词频率(Keyword Counts /15 min)曲线。这一曲线直观地呈现了包含灾害关键字的微博信息数量变化过程，显示了关键词的数目和时间的关系图，如果关键词相关灾害事件爆发，很可能会产生峰波，增加了确认事件发生的准确度。

(4)相关微博示例区。系统界面右侧显示了在自定义地区、时间段、关键词的情况下，系统自动采集的原始微博信息示例(当前，系统设定最多可显示 1 500 条信息)。显示的每条原始博文还包括微博用户名、头像和注册地区。如果用户手机的 GPS 开启，则博文后端显示信息发送时用户的具体位置。示例区的功能，提供了人工复查博文内容的可能性，也扩展了本系统的应用范围。界面上，如果微博信息被标注为红色(地震)，意味着博文经上节所述分类器分类为即时灾害信息，且在右下角显示了检测系统运算出来的成功概率作为参考。

3 结论与讨论

现代社会，在自然灾害发生后，灾情信息的传播过程高度依赖于互联网社交媒体平台。因此，社交媒体的灾害信息管理能力对于整体应急救援响应行动的开展至关重要。本文面向新浪微博平台，探索高效的中文灾害微博信息分类算法，借鉴澳大利亚科学院研发的英文推特灾害实时预警系统(ESA)的经验，开发了新浪微博灾害事件检测系统(SWIM)，成功实现了基于社交媒体平台的地震等灾害事件检测。

中国曾被称为“灾荒之国”，洪涝、干旱、台风、风暴潮、地震、森林草原大火等自然灾害种类繁多，发生频率高，分布地域广。这一现状为当前的灾害事件检测系统提出了更高的挑战。首先，当前的 SWIM 系统只实现了地震等既定灾害的爆发检测，未来将探索更多种灾害的综合分类器，以实现其他灾种及突发事件的实时检测；其次，对 SWIM 系统应继续完善，开发后续的灾害预警模块；第三，面向灾害救援响应过程，拟探索中文短文本聚类方法，根据灾害救援需求，实现合理的话题聚类与分析，从而更好地利用社交媒体平台的实时信息为救援减灾过程服务。

参考文献：

[1] Dorothy E Leidner, Gary Pan and Shan L Pan. The role of IT in crisis response: Lessons from the SARS and Asian tsunami disasters[J]. Strateg. Inf. Syst., 2009, 18(2): 80 - 99.

[2] American Red Cross. More Americans using mobile apps in emergencies [EB/OL]. (2012 - 08 - 31) [2013 - 04 - 10]. http://www.redcross.org/news/pressrelease/More - Americans - Using - Mobile - Apps - in - Emergencies.

[3] 马化展，常媛媛，陈泽然. 水灾 7 天：余姚的红与黑[EB/OL]. (2013 - 10 - 14)[2013 - 10 - 16]. http://www.kcis.cn/4409

[4] USGS. Did you feel it? [EB/OL]. (2005 - 03 - 21)[2012 - 09 - 26]. http://earthquake.usgs.gov/earthquakes/dyfi/.

[5] Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. Earthquake shakes twitter users: real - time event detection by social sensors [C]//The 19th International Conference on World Wide Web, WWW'10. New York, ACM 2010: 851 - 860.

[6] Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. Tweet analysis for real - time event detection and earthquake reporting system development [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 919 - 931.

[7] Fabian Abel, Claudia Hauff, Geert - Jan Houben, et al. Twitcident: fighting fire with information from social web streams[C]// The 21st International Conference Companion on World Wide Web, WWW' 12Companion. New York: ACM, 2012: 305 - 308.

[8] Soudip Roy Chowdhury, Muhammad Imran, Muhammad Rizwan Asghar, et al. Tweet4act: Using incident - specific profiles for classifying crisis - related messages[C]// The 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM). Kmstiansand: ISCRAM, 2013.

[9] Jakob Rogstadius, Maja Vukovic, Claudio Teixeira, et al. Crisistracker: Crowdsourced social media curation for disaster awareness [J]. IBM Journal of Research and Development, 2013, 57(5): 411 - 413.

[10] Omidyar Network. Ushahidi: The African Software Platform Helping Victims in Global Emergencies[EB/OL]. (2013 - 1 - 22) [2013 - 7 - 08]. http://www.ushahidi.com/.

[11] Paul S Earle, Daniel C Bowden and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world[J]. Annals of GeoPhysics, 2012, 54(6): 708 - 715.

[12] Mark A Cameron, Robert Power, Bella Robinson, et al. Emergency situation awareness from Twitter for crisis management[C]// The 21st International Conference Companion on World Wide Web, WWW ' 12 Companion. New York: ACM, 2012: 695 - 698.

[13] Bella Robinson, Robert Power and Mark Cameron. An evidence based earthquake detector using twitter[C]//The Workshop on Language Processing and Crisis Information. Nagoya: LPCI, 2013: 1 - 9.

[14] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, et al. EARS (earthquake alert and report system): a real time decision support system for earthquake crisis management[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'14, New York: ACM, 2014: 1749 - 1758.

[15] 中国互联网络信息中心. 第 33 次中国互联网络发展状况统计报告[EB/OL]. (2014 - 01 - 16)[2014 - 01 - 20]. http://www.199it.com/archives/187745.html.

[16] Yan Qu, Chen Huang, Pengyi Zhang, et al. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake[C]//The ACM 2011 Conference on Computer Supported Cooperative Work, Hangzhou, ACM 2011: 25 - 34.

[17] Yanquan Zhou, Lili Yang, Bartel Van de Walle, et al. Classification of microblogs for support emergency responses: Case study Yushu earthquake in China[C]//The 46th Hawaii International Conference on System Sciences. Hawaii: IEEE, 2013: 1553 - 1562.

[18] Burbidge R, Trotter M, Buxton B, et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis [J]. Computers and Chemistry. 2001, 26 (5): 5 - 14.

[19] Beyer K, Goldstein J, Ramakrishnan R, et al. When is "nearest neighbor" meaningful? [C]//Database Theory - ICDT'99. Israel: IEEE, 1999: 217 - 235.

[20] Breiman L. Random forests[J]. Machine learning, 2001, 45 (1): 5 - 32.

Sina Weibo Disaster Information Detection Based on Chinese Short Text Classification

Bai Hua[1] and Lin Xunguo[2]
(1. *School of Management*, *Harbin Institute of Technology University*, *Harbin* 150001, *China*;
2. *CSIRO Digital Productivity Flagship*, *G. P. O. Box* 664, *Australia Canberra*, *ACT* 2601)

Abstract: Weibo, a popular Chinese social media service, has received much attention recently. More and more people use Weibo as an information tool, especially when the disaster happens. We present a work to develop a disasters detector based on Sina Weibo messages. This system captures public messages from Sina Weibo platform at first, and then processes messages filter and text classification to determine if messages correspond to people experiencing a disaster. We also offer an interface for users to view the processed messages. Our long term aim is to develop a general alert stem for various disaster event types in China, and it would be very useful for the disaster rescue.

Key words: social media; Sina Weibo; disaster information; disaster detection

《灾害学》杂志征稿简则

《灾害学》杂志是由陕西省地震局主办的把灾害问题作为一门科学在我国最早创办(1986 年)的核心学术期刊。主要刊载的内容有：对各种灾害(自然灾害和人文灾害)进行综合系统探讨研究；通过对各种灾害事件的分析讨论，总结经验，吸取教训；广泛交流灾害科学的学术思想，研究方法，研究成果；报导国内外关于灾害问题的研究动态和防灾抗灾对策；揭示和探索各种灾害发生演化的客观规律。

来稿要求和注意事项：

(1) 篇首页下注请附作者简介，含姓名(出生年月 -)，性别，民族，籍贯，职称(务)及主要研究方向。同时提供方便联系的电话、E - mail 等。并附该文受何种基金项目(编号)资助。

(2) 文章应包括：标题、作者、作者单位、摘要、关健词、引言、正文、结语、参考文献和英文摘要。字号不小于 5 号字。

(3) 来稿要求论点明确，论据可靠，文字精炼，数据准确。学术论文、综述文章 8000 ~ 12000 字左右(包括图表)，其中插图以不超过 6 幅为宜；科技报导、简介等一般不超过 3000 字，其中插图以不超过 3 幅为宜。

(4) 中文摘要不少于 250 字，应含文章主旨、所用方法和主要结论。关健词 3 ~ 6 个。另需附英文摘要，并与中文摘要相对应。

(5) 图表要求

① 图、表须有名称和编号，其内容要与正文中的编号和解释一致；图幅不超过 A4。图件需保证图内线条和字符清晰。表格用三线表。

② 有坐标系的插图，纵横坐标上均要有适宜的刻度、对应的数据，并注明其所代表的物理量和单位。文字中涉及的主要地名在图件中需要标注，便于读者对照理解。

③ 涉及国界的图件须以正式出版的地图为底图绘制，并注出底图的出处及线段比例尺。

(6) 文稿中的计量单位一律采用中华人民共和国国家标准《量和单位》中颁布的法定计量单位，变量的符号用斜体，常量及单位的符号用正体。

(7) 参考文献应列全，最少不少于 8 条。凡在正文中引用的公开发表的著作、论文，必须在参考文献中列出；参考文献的序号按文中出现的先后顺序，以阿拉伯数字标注，用方括号括起置于正文引用处右上角。参考文献中请保留 3 位作者姓名，从第 4 位作者起，用“等”来代替。

中英文参考文献均按下列格式排列：

专著：[序号]著者姓名. 书名[M]. 出版地：出版者，出版年：起止页码.

文集：[序号]文献著者姓名. 析出文献名[C]//文集编著者. 文集名. 出版地：出版者，出版年：起止页码.

期刊论文：[序号]著者姓名. 论文题名[J]. 刊名，出版年，卷号(期号)：起止页码.

学位论文：[序号]作者. 论文题名[D]. 授予单位地址：授予单位，授予时间：起止页码.

网络下载文献：[序号]作者. 论文题名[EB/OL]. (文献上传时间)[下载时间]. 下载网址.

其它文献类型均参考上述格式标明各项。

(8) 需提供电子文稿。稿件请自留底稿，自投稿之日起，3 个月内未接到刊用通知者可改投他刊。

(9) 凡经本刊录用的文章，需与办刊签订“投稿协议书”，把文章的专有许可权和独家代理权授予本刊。本刊对文章具有以下专有使用权：汇编权、发行权、印刷权和电子版的复制权、信息网络传播权以及代理许可国内外文献检索系统或数据库收录权。不同意者，请另投他刊。

(10) 来稿需保证为原创作品，与灾害研究密切相关，有自己的思想和一定的创新，无一稿两投，没有抄袭，并且不涉及保密及其他与知识产权有关的侵权问题。文责自负。

(11) 来稿一经审查通过，本刊将按所占版面多少通知作者交纳论文发表费和审稿费，一经发表，亦将按采稿通知的约定付给作者稿酬，并赠送本期期刊 3 本。

(12) 作者可以参考《灾害学》已经发表的文章格式，可以到《灾害学》网站 http：//www. zaihaixue. com 免费下载最新的《灾害学》登载文章的电子版作为格式参考。

(13) 作者可登录《灾害学》网站或者直接把稿件发到编辑部信箱的形式投稿，请注明详细联系地址、邮编和电话。

(14)《灾害学》编辑部没有授权任何代理机构代收、代写、代发稿件，请作者朋友小心上当受骗。

《灾害学》编辑部

地址：陕西省西安市碑林区边家村水文巷 4 号《灾害学》编辑部
邮编：710068　　电话(传真)：(029)88465341
电子信箱：zhx@ eqsn. gov. cn　zhx02988465341@ 163. com
网站地址：http：//www. zaihaixue. com