

地质灾害气象风险预警实时检验客观工具 关键技术及应用

杨寅^{1,2} 包红军^{1,2*} 徐成鹏^{1,2}

(1 国家气象中心,北京 100081; 2 中国气象局-河海大学水文气象研究联合实验室,北京 100081)

摘要 中国气象局灾害管理系统中地质灾害关键词和灾害发生时间具有多种组织结构的信息特征,灾害信息分析依赖于人工,不利于地质灾害气象风险预警检验的实时性和客观化。本文结合灾情信息特征和中文自然语言处理软件包 Jieba 研发了地质灾害多元灾情信息智能分析技术,开发了地质灾害气象风险预警实时检验客观工具。检验工具集成了多元灾情信息智能分析,业务检验方法为主检验主客观预警,检验结果产品自动化输出等功能,可实现预警检验全部流程。2015—2019 年地质灾害气象风险预警检验应用表明,实时检验客观工具提取灾害信息准确率达到 97%,检验效率高,实时性和集成化优势明显。

关键词 地质灾害;气象风险预警;实时检验;多元灾情信息;智能分析;Jieba 自然语言;实时性,集成化

中图分类号: P429 **DOI:** 10.19517/j.1671-6345.20200130 **文献标识码:** A

引言

我国是全球地质灾害最严重国家之一,滑坡、崩塌、泥石流分布范围占国土面积 44.8%。根据自然资源部数据,2019 年全国共发生地质灾害 6181 起,造成 211 人死亡,13 人失踪,75 人受伤,直接经济损失 27.7 亿元。大量研究证明降水是引发地质灾害的主要因素^[1-3],全国县(市)地质灾害调查也指出降水是所有泥石流,90%滑坡和 81%崩塌的诱发因素^[4]。国土资源部和中国气象局 2003 年起联合开展气象因素(主要为降水)引发地质灾害的气象风险预警预报,对外正式发布国家级地质灾害气象风险预警,其中气象部门的业务承担单位为中央气象台^[5]。随着业务持续开展,中央气象台积累了丰富的预警经验,借鉴降水强度-时间曲线^[6]、降水驱动指标^[7]等研究成果,结合国内实际研发了基于临界雨量的第一代地质灾害气象预警客观模型^[8](简称客观模型 I)和融合临界雨量与地质灾害易发性信息量的第二代模型(简称客观模型 II)^[9]。此外,还有不少学者研究提出多种预警客观模型^[10-12],为地

质灾害气象风险预警提供了技术支持。

地质灾害气象风险预警效果和客观模型性能需要及时检验评估以反馈业务开展和改进客观模型。预警定量检验是一个对比实况灾情信息和预警产品,输出数值检验结果的过程,目前主要困难是灾情信息及时获取。地质灾害灾情信息的两个主要获取途径是遥感数据和灾情报告,如文广超等学者研究利用高分辨率的卫星或无人机遥感数据开展灾情信息提取^[13-15],但由于数据获取费用高、卫星重访周期长、无人机覆盖范围小、遥感地质解译科学发展慢等因素影响,这些技术多在小区域内试验,无法在面向全国的地质灾害气象风险预警业务中运用。实际业务中,灾情信息通常来源于中国气象局灾害管理系统中民政、国土等部门共享或气象信息员汇报的灾害报告,报告内容以中文自然语言为主,文本表达格式灵活,需要人工分析地质灾害发生时间、类型等信息^[16,17],在美国的山洪地质灾害预警效果评估工具中,灾情信息获取也依赖于人工分析^[18]。由于地质灾害气象风险预警检验需要获得灾情信息后才能开展,现有检验工具需要与人工分析结合使用,预

<http://www.qxkj.net.cn> 气象科技

国家重点研发计划(2018YFC1507505、2018YFC1507205 和 2018YFC1508102)、国家自然科学基金项目(41775111)、中国气象局创新发展专项(CXFZ2021Z033)共同资助

作者简介:杨寅,男,1986 年生,硕士,工程师,主要从事水文气象预报业务和科研工作,Email:yy6194819@126.com

收稿日期:2020 年 4 月 10 日;定稿日期:2020 年 9 月 7 日

* 通信作者, Email:baohongjun@cma.gov.cn

警检验集成化程度不高,尚未完全实现客观化。此外,由于人工效率不高,同时全国范围气象灾害报告数量多,地质灾害信息无法实时获取,导致工具也缺乏实时检验能力。

随着计算机科学发展,近年来自然语言处理技术有很大发展,出现了 Jieba、SnowNLP 等多种自然语言处理软件,实现了很多成功的中文文本数据处理案例^[19-22],这为利用计算机程序替代人工分析,实现地质灾害灾情信息分析客观化,快速收集提取灾害发生时间、地点、损失等多元灾情信息提供了技术基础。

在统计中国气象局灾害管理系统 2015—2019 年全国地质灾害灾情信息特征的基础上,基于中文自然语言处理软件研发了地质灾害多元灾情信息智能分析技术,建立了具备实时性和集成化优势的地质灾害气象风险预警实时检验客观工具,能为预警业务开展和客观模型改进提供信息反馈。

1 灾情报告信息特征分析

在国家级地质灾害气象风险预警业务中,中央气象台收集灾情信息的主要渠道是中国气象局灾害管理系统,系统中灾情报告由各基层气象局以县级行政区为基本单位,通过民政和自然资源部门信息共享、气象信息员汇报等途径获取数据后上传,包含灾害所属县行政区编码、灾害类型、灾害开始时间与结束时间、灾情概况、雨情概况、人员和财产损失等要素,其中灾情与雨情概况为一段或若干段中文文本,灾害类型为内容和种类确定的词语,其余要素均为格式固定的数字。

理想情况下,根据灾害时间、地点、类型、损失要素可以全面描述地质灾害事件。分析 2015—2019 年全国地质灾害灾情报告发现,灾害发生县的行政区编码要素项准确可靠,但存在着人员伤亡和财产损失要素项数据缺失,灾害开始和结束时间跨度超过国家级地质灾害气象风险预警时效,以及地质灾害事件与台风、强对流等明显降水过程及其诱发的更为严重的渍涝、溃坝、中小河流洪水等次生灾害同时发生时,由于灾害类型要素为单一值,业务人员优先选择更为严重的灾害填报的情况。灾情报告中的灾害时间、类型、损失要素可能无法提供多方位的有效信息,但灾情和雨情概况的中文文本往往包含着描述地质灾害类型、时间、地点、人员伤亡和财产损

失的内容,灾情信息更丰富全面。

灾情和雨情概况由全国各基层气象局上报,对地质灾害事件及影响的文字描述格式往往不统一,具有很强的主观灵活性,收集多元灾情信息时需要人工分析,工作效率低,同时为保证全面收集地质灾害灾情,需要对所有气象灾害报告开展分析,工作量大。图 1 是中国气象局灾害管理系统 2015—2019 年气象灾害和地质灾害报告总数对比,由图可见,气象灾害报告数量在 4600~6000 条之间浮动,年平均约 5400 县次,灾害总数大,其中地质灾害整体约占气象灾害数量的 12%,年平均约 670 县次。以国家级地质灾害气象风险预警年度检验为例,人工分析全国气象灾害报告提取地质灾害灾情信息的平均花费时间超过 3 个工作日,这无法根据检验需求实时获得灾情信息分析结果并开展地质灾害气象风险预警效果定量检验。

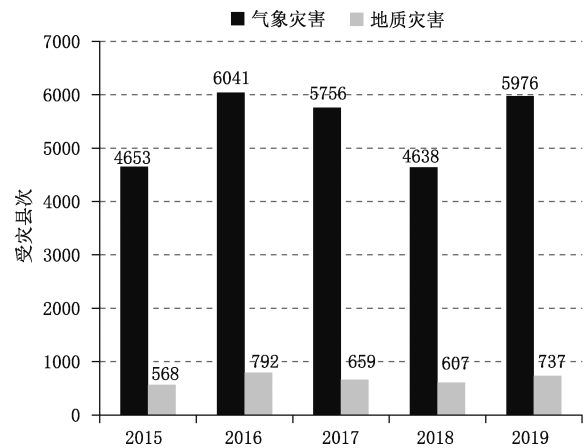


图 1 2015—2019 年中国气象局灾害管理系统中气象灾害和地质灾害数量对比

分析地质灾害报告中灾情和雨情概况的中文文本,发现语句格式虽然不统一,但对灾害事件描述总体可分为 2 种类型,第 1 种具有“具体时间+灾害事件”的信息特征,如某月某日某时某地发生滑坡,第 2 种具有“时间段+灾害事件”的特征,其中时间段的文字描述通常以上午、下午等词语,或者使用到、至、破折线、破浪线等文字或符号将两个时间连接而成,如某日某时至(或到)某日某时,或某日某时~某日下午明显降雨,某地发生泥石流的描述性语句。对于人员伤亡和财产损失情况,绝大多数灾情概况具有“死亡 X 人”、“X 人死亡”、“X 人重伤”、“受伤 X 人”等固定文本格式的信息特征。如果利用计算机

替代人工分析灾情报告信息特征,能更客观高效地处理收集地质灾害的灾情信息。

2 多元灾情信息智能分析技术

2.1 中文语句分析工具选择

地质灾害多元灾情信息智能分析技术是指对中文文本内容进行自然语言处理,利用灾情和雨情概况中灾害时间和损失内容的信息特征,智能分析提取地质灾害发生时间、人员伤亡和财产损失等多元灾情信息。目前常用的中文分词工具有 Jieba、SnowNLP、PKUSeg、Thulac、HanLP、pyhanlp 等,其中 Jieba 以其开源、词典可自定义等优点而使用广泛,可参考借鉴成功案例数量多,研究选用 Jieba 处理灾情报告。

Jieba 软件包利用中文词库确定汉字之间的关联概率,将概率大的汉字组成词组,具有精确模式、全模式、搜索引擎模式和 Paddle 模式选项,适用于不同用途。本研究目的是提取文本中地质灾害多元灾情信息,属于文本分析,因此选用精确模式。

2.2 地质灾害类型关键词分析

灾情数据分析发现,由于灾情报告由基层气象局上报,各地业务人员的惯用词语和语句组织方式多样,灾情概况中存在着多种描述地质灾害的词语。图 2 是利用 Jieba 分词软件包对 2015—2019 年地质灾害灾情报告开展词频分析得到的灾害类型关键词分布。由图可见,对以降水为主要诱发因素的崩塌、滑坡、泥石流 3 种地质灾害,灾情报告对同一类型灾害存在着意义相近的不同名称,如塌方、坍塌、垮塌均表示崩塌,溜方的语义等同于滑坡,为完整提取多元灾情信息,研究将以上描述地质灾害的关键词加入 Jieba 分词软件包自定义词典。

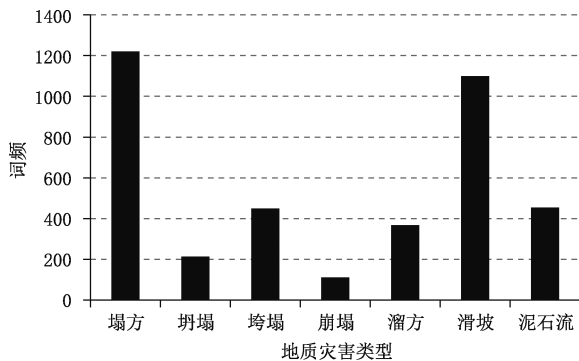


图 2 地质灾害类型关键词词频分布

此外,根据灾情报告文本内容的信息特征分析结果,研究也将上午、下午、8 时、8 点、截至、至、到、连接符“—”和“~”等参与构成灾害事件信息特征的字和词语加入 Jieba 自定义词典,以提高软件包正确分割语句的能力。

2.3 多元灾情信息智能分析

根据地质灾害类型关键词和灾情报告信息特征分析结果,结合 Jieba 中文分词软件包,研究尝试对中国气象局灾害管理系统的灾情报告进行计算机客观分析,收集地质灾害的多元灾情信息。

图 3 是地质灾害多元灾情信息智能分析处理流程:①按照中文文本的行文习惯,以句号和回车符为标志,使用 PYTHON 计算机语言将灾情和雨情概况分割为若干个完整语句,之后利用 jieba 将语句拆分为词语组合。②对比拆分后的词语是否含有滑坡、塌方、泥石流等地质灾害类型关键词。③关键词对比匹配条件下,开展地质灾害信息特征智能分析,对于地质灾害发生时间,考虑到地质灾害属于突发性现象,首先搜索灾情概况是否包含灾害发生具体时间,其次分析灾情概况中灾害发生时间段,最后基于明显降水诱发地质灾害时间滞后性一般不明显的规律,提取雨情概况中最明显降水时段作为灾害发生时间;对于灾害损失,按照人员伤亡和财产损失的文本信息特征搜索灾情概况;考虑到目前国家级地质灾害气象风险预警时效为 24 h,即预警发布日 20:00 至次日 20:00,检验工具采用 20:00 为截断时间划分灾害日期,对于灾害发生时段包含 20:00 的

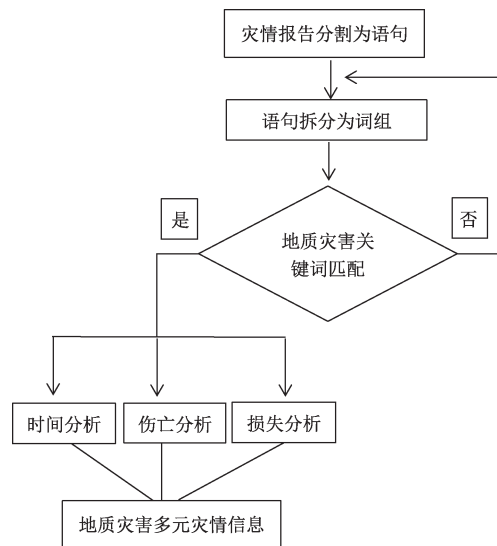


图 3 地质灾害多元灾情信息智能分析流程图

情况,以灾害发生时间与截断时间,以及截断时间与灾害结束时间二者中时段较长者为灾害发生时段,如果灾害开始和结束时间跨度超过 24 h,大于目前地质灾害气象风险预警的时效长度,采用文献[18]的方法质量控制。④遍历分析灾情和雨情概况要素项的完整语句,选择最精确的时间点或持续最短的时间段为灾害发生时间。

3 实时检验客观工具架构

地质灾害气象风险预警实时检验客观工具是一个可独立运行的平台,包含多元灾情信息智能分析,预警定量检验和检验产品自动化输出 3 个功能模块,能够完成适用于业务的实况信息获取、实况和预报对比、检验结果输出的预警检验全部流程,具有高度客观化和集成化的特点。考虑到业务有检验固定时段预警效果的实际需求,检验工具能够自动开展多尺度(日,周,月等)的预警评估,此外平台还具有交互功能,允许用户针对具体地质灾害过程自定义输入开始和结束时间进行预警检验。

3.1 多元灾情信息智能分析模块

本模块基于地质灾害多元灾情信息智能分析技术,利用 Python 和 JAVA 语言开发应用程序,根据用户指定的检验时间段从中国气象局灾害管理系统数据库查询、下载灾情报告,开展智能分析,提取地质灾害发生时间、地点、人员和财产损失等多元灾情信息。

3.2 预警定量检验模块

本模块功能是对国家级地质灾害气象风险预警主观和客观产品开展定量检验。中国气象局 2016 年 10 月发布《暴雨诱发地质灾害气象风险预警业务规范》(以下简称《规范》),规定了地质灾害气象风险预警产品业务检验方法。检验工具采用《规范》确定的业务方法检验预警产品,并考虑到气象行业通常使用 TS 评分开展预报检验的实际情况,提供了地质灾害气象风险预警 TS 评分。《规范》规定,地质灾害气象风险预警产品的检验对象分为单点和县级区域两种,其中单点指出现地质灾害的地点,县级区域指发生地质灾害的县行政区,当某县行政区范围内出现一起及以上地质灾害时,认为该县发生地质灾害;否则认为未发生地质灾害。检验指标包括预报命中率 TSR、空报率 FAR 和漏报率 PO,计算公式如下:

$$TSR = \frac{NA + ND}{NA + NB + NC + ND} \times 100\% \quad (1)$$

$$PO = \frac{NC}{NA + NC} \times 100\% \quad (2)$$

$$FAR = \frac{NB}{NA + NB} \times 100\% \quad (3)$$

其中,NA 为预警服务产品发布正确次数,NB 为空报次数,NC 为漏报次数、ND 为无预警服务产品预报正确次数。

TS 评分计算公式如下所示,其中 NA,NB,NC 参数意义与业务检验方法相同。

$$TS = \frac{NA}{NA + NB + NC} \quad (4)$$

受中国气象局灾害管理系统中地质灾害发生地点只能精确到县行政区的约束,重点研发了以县行政区为检验对象的国家级地质灾害气象风险预警定量检验功能,使用县行政区内国家基本气象站代表所属的县,针对目前地质灾害气象风险预警业务主观产品和预警模型客观产品的落区、格点和站点 3 种格式设计了不同检验策略,对于落区预报,对比灾害发生县是否在落区范围内以确定预警命中、空报和漏报;对于格点预报,则通过判断一定半径范围内的格点是否覆盖灾害发生县;对于站点预报,对比灾害发生县与预警县是否重合。

3.3 检验结果产品输出模块

本模块功能是对地质灾害气象风险预警定量检验数字结果进行可视化输出,提高检验结果被理解和解读的效果。研究利用 GRADS 软件绘制预警产品落区和地质灾害发生县的相对位置,展示预警命中、空报和漏报的灾害数量,以及业务检验方法 TSR、PO 和 FAR 指标及 TS 评分定量检验结果。

4 实时检验客观工具应用

地质灾害气象风险预警实时检验客观工具已经应用于中央气象台地质灾害气象风险预警业务,2015—2019 年国家级地质灾害气象风险预警检验表明,该工具能够实时完成多元灾情信息智能分析、以业务检验方法为主检验主观和客观预警,检验产品自动化输出的全部流程,具有实时检验的能力,如完成年度地质灾害多元灾情信息分析和预警定量检验不超过 0.5 h,而依赖于人工分析灾情的同类检验工具耗时超过 3 个工作日,效率有明显提升,同时,灾情信息分析正确率达到 97%,能为预警效果

分析与客观模型改进提供信息参考。

展示预警检验工具应用个例,2016年9月28日,第17号台风“鲇鱼”强降水在浙江、福建两省多个县诱发地质灾害,自然资源部和中国气象局于9月27日20:00联合发布地质灾害气象风险橙色预警。图4是预警定量检验结果图形产品,由图可见,地质灾害气象风险预警命中8个县,漏报1个县,空报43个,无预警也无灾害县2465个,业务检验方法TSR、PO和FAR分别为98%、11%和84%,TS评分为0.15。

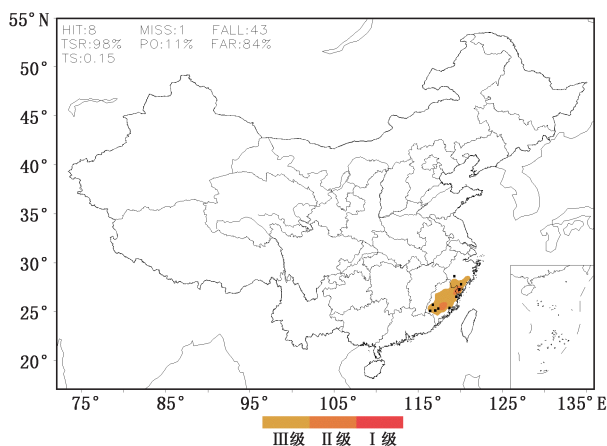


图4 地质灾害气象风险预警定量检验图形产品
(黑色方框表示地质灾害实测发生地,填色为地质灾害气象风险预测落区及等级)

表1是检验工具分别对主观和客观共3种预警产品的定量检验结果,可以看到,在业务检验方法3个指标中,预报员制作的国家级地质灾害气象风险预警主观产品具有最高的命中率TSR,以及最低的漏报率PO和空报率FAR,预警效果最优。客观模型I的PO与主观产品相同,但FAR更高,表现出模型以较高空报为代价降低漏报的特征,客观模型II的FAR与主观产品接近,但PO偏大,说明模型存在增加漏报以控制空报的策略,预警业务中需要注意不同模型的策略差异。主观预警产品的TS评分最高,说明了预报员对客观模型的订正作用。

表1 3种地质灾害气象风险预警产品定量检验结果对比

预警产品类别	业务检验方法/%			TS评分
	TSR	PO	FAR	
国家级地质灾害气象风险预警	98	11	84	0.15
客观模型I	97	11	89	0.10
客观模型II	98	44	85	0.06

5 结论和讨论

中国气象局灾害管理系统中地质灾害灾情报告内容具有多种组织结构,灾害信息获取依赖于人工分析,导致现有地质灾害气象风险预警检验工具实时性和客观化能力不足。在灾情报告分析中发现其具有滑坡、溜方、塌方、垮塌等多种描述地质灾害的关键词,以及灾害内容按照“具体时间+灾害事件”或“时间段+灾害事件”组织的信息特征,本研究中结合了灾情信息特征与中文自然语言处理软件包jieba研发了地质灾害多元灾情信息智能分析技术,建立了地质灾害气象风险预警实时检验客观工具。检验工具是一个由多元灾情信息智能分析,以业务检验方法为主检验主观和客观预警,检验结果产品自动化输出等功能模块组成的可交互平台,能够集成完成预警检验全部流程。

2015—2019年国家级地质灾害气象风险预警业务检验表明,实时检验客观工具提取灾情准确率达到97%,完成多元灾情信息分析和预警检验的效率比依赖于人工分析灾情的同类检验工具有明显提升,具有实时性和集成化的优势,能更好地为预警业务开展和客观模型改进提供信息反馈。

检验工具分析多元灾情信息正确率尚未达到100%的原因是灾情报告存在一定数量的错字或不常见文字表述,如“1日08时—20时”中,08时与20时的连接符号按照中文习惯为破折号,而灾情报告中实际是汉字“一”,或者日字被错误拼写为曰。未来将分析更多灾情报告,增强检验工具容错能力,提高灾情信息分析准确率。此外,现行地质灾害业务检验方法存在TSR指标受ND参数影响大和检验指标数量多的缺点,后期将研究更适用于地质灾害气象风险预警的检验方法。

参考文献

- [1] 谭万沛,王成华,姚令侃,等.暴雨泥石流滑坡的区域预测与预报[M].成都:四川科技出版社,1994,115:205-274.
- [2] Au S W. Rain-induced slope instability in Hong Kong [J]. Engineering Geology, 1988, 51(1):1-36.
- [3] Glade T, Crozier M, Smith P. Applying probability determination to refine landslide-triggering rainfall thresholds using an empirical Antecedent Daily Rainfall Model [J]. Pure and Applied Geophysics, 2000, 157(2): 1059-1079.
- [4] 李媛,孟晖,董颖,等.中国地质灾害类型及其特征——基于全

- 国市县地质灾害调查成果分析[J]. 中国地质灾害与防治学报, 2004, 15(2): 90-97.
- [5] 杨寅, 林建, 包红军. 2018 年地质灾害重点区域气象预警效果检验[J]. 中国地质灾害与防治学报, 2019, 30(6): 135-140.
- [6] Caine N. The rainfall intensity duration control of shallow landslides and debris flows [J]. *Geografiska Annaler*, 1980, 62: 23-27.
- [7] 詹钱登, 李明喜. 土石流发生降雨警戒模式[J]. 中华水土保持学报, 2004, 35(3): 275-285.
- [8] 薛建军, 徐晶, 牛若芸, 等. 区域性地质灾害气象预报方法研究[J]. 气象, 2007, 10(31): 24-27.
- [9] 徐晶, 张国平, 张芳华, 等. 基于 Logistic 回归的区域地质灾害综合气象预警模型[J]. 气象, 2007, 12(33): 3-8.
- [10] 李忠燕, 田其博, 章国材, 等. 铜仁地区滑坡临界雨量研究[J]. 气象科技, 2016, 44(4): 680-685.
- [11] 齐大鹏, 汪超, 韩小令, 等. 基于加密降水资料的贵州地质灾害概率预报模型[J]. 气象科技, 2016, 44(5): 788-792.
- [12] 范江琳, 郭晓军, 青泉. 基于降雨分级的泥石流降雨 I-D 预报模型[J]. 气象科技, 2020, 48(2): 242-247.
- [13] 文广超, 张哲玮, 肖学军, 等. 基于遥感数据的灾后滑坡信息快速提取方法[J]. 中国地质灾害与防治学报, 2020, 31(2): 80-86.
- [14] 张海涛. 基于高分影像的滑坡提取关键技术研究[D]. 武汉: 中国地质大学(武汉), 2017.
- [15] 刘俊龙. 基于无人机影像的精细化地形建模在滑坡识别中的应用[D]. 绵阳: 西南科技大学, 2019.
- [16] 刘艳辉, 温铭生, 苏永超, 等. 台风暴雨型地质灾害时空特征及预警效果分析[J]. 水文地质工程地质, 2016, (43)5: 119-126.
- [17] 杨寅, 包红军, 彭涛. 台风“鲇鱼”强降水引发的地质灾害气象风险预警检验与分析[J]. 暴雨灾害, 2018, 37(6): 221-228.
- [18] Jonathan G, Jessica M, Yang Hong, et al. Evaluation of tools used for monitoring and forecasting flash floods in the United States [J]. *Journal of Hydrological*, 2010, 27: 158-173.
- [19] 胡侠, 林晔, 王灿, 等. 自动文本摘要技术综述[J]. 情报杂志, 2010, 29(8): 144-147.
- [20] 吴思竹, 钱庆, 胡铁军, 等. 词干提取方法及工具的对比分析研究[J]. 图书情报工作, 2012, 56(15): 109-115.
- [21] 明拓思宇, 陈鸿昶. 文本摘要研究进展与趋势[J]. 网络与信息安全学报, 2018, 31(6): 5-14.
- [22] 吕佳高, 梁奎阳, 蔡伟. 基于文献计量和众包技术的前沿科技关键词挖掘[J]. 计算机科学, 2019, 46(3): 275-282.

Development of a Real-Time Verification Objective Tool for Meteorological Risk Warning to Geological Hazards

YANG Yin^{1,2} BAO Hongjun^{1,2} XU Chengpeng^{1,2}

(1 National Meteorological Center, Beijing 100081;

2 CMA-HHU Joint Laboratory for Hydrometeorological Studies, Beijing 100081)

Abstract: The hazard information was collected manually for keywords and the occurrence time of geological hazards in the hazard management system of the Chinese Meteorological Administration, which have multiple organization structures in information characteristics and are thus adverse to the real-time property and objectivity of verification for meteorological risk warning to geological hazard. In this study, an objective technique of intelligent analysis for multiple information of geological hazards by computer and a real-time verification objective tool for meteorological risk warning to geological hazard are developed based on the characteristic analysis of hazard reports and a natural language processing software named Jieba. This objective tool can accomplish all the processes of verification by integrating functions of hazard information analysis, verification to objective and subjective meteorological risk warning to geological hazards, and output verification result products automatically. Verification for meteorological risk warning to geological hazards from 2015 to 2019 shows that the objective tool has an accuracy rate of 97% in gathering hazard information, higher efficiency in analyzing hazards and verifying meteorological risk warning, which has obvious advantages of real time and integration compared with similar tools.

Keywords: geological hazard; meteorological risk warning; real-time verification; multi hazard information; intelligent analysis; Jieba natural language