王琳，姜立新，杨天青，张维佳，2019．地震应急信息自动分类方法研究．震灾防御技术，**14**（4）：907—916．doi：10.11899/zzfy20190422

地震应急信息自动分类方法研究[1]

王　琳[1）]　姜立新[2）]　杨天青[2）]　张维佳[2）]

1）中国地震局地震预测研究所，北京 100036
2）中国地震台网中心，北京 100045

摘要　地震应急信息的高效处理为地震应急救援工作提供了重要支撑。本文根据地震应急信息分类的需求，构建了一种高效便捷的地震信息分类处理方法。以震前、震时、震后为时间主线，将地震应急信息分为震前基础背景信息、地震震情灾情信息及震后应急救援信息，并采用“关键词分类”的方法，在计算机语言的支持下，将多渠道汇集的应急信息进行自动分类，在一定程度上缩短了应急信息加工处理与服务的时间，能快速高效地为应急指挥提供信息服务。

关键词：地震　应急信息　自动分类　关键词分类

引言

管理者经常面临着与决策相关信息缺失和不相关信息泛滥的问题，往往会对管理者的决策造成负面影响（Detrick，2002）。此情况在地震灾害应对过程中尤为突出，信息缺失或冗余往往造成抗震救灾指挥决策的滞后，甚至导致救援力量和资源投放重点出现偏差。

近年来，中国地震局在应急救援领域先后开展了“九五首都圈防震减灾示范项目”“十五中国数字地震观测网络项目”和“国家地震社会服务工程”。应急触发、灾情研判、快速响应及辅助决策等应急科技产出的日益丰富为国家及各省抗震救灾指挥部实施地震应急救援提供有力的科学依据和技术支持。我国虽建成了较完整的应急指挥体系及相应的指挥技术系统，但在应急信息管理方面仍存在一些问题，具体表现为：①技术产出较丰富，直接有效利用率较低；②内容重复，存放分散；③尚未建立有效的灾情管理技术。

为此国内不少专家学者对地震应急基础信息及灾情信息的收集、整理与分类编码进行了大量研究。付继华等（2009）、聂高众等（2002）从建立数据库的角度分别讨论了地震应急数据的分类。《地震学专业分类表》（梁凯利等，2011）严格按照《中国图书馆分类法》的要求，结合地震科技资料分类的自身特点，对地震学专业进行了分类；白仙富等（2010）按照信息内容的本质属性，依据发生什么事件、产生什么影响、对产生的影响人们做出什么响应、针对响应有何成效的思路对地震应急现场信息进行分类；张翼等（2016）根据地

1 基金项目　地震科技星火项目（1840807302）；川滇地区人员伤亡动态研究和国家重点研发计划（2018YFC1504506）
［**收稿日期**］ 2018-12-11
［**作者简介**］ 王琳，女，生于 1994 年。中国地震局地震预测研究所 2016 级在读硕士。研究方向地震应急。E-mail：wanglin199@126.com
［**通讯作者**］ 姜立新，男，生于 1966 年。研究员。主要从事地震灾害与地震应急技术的研究。E-mail：jlx@seis.ac.cn

震应急信息产品管理、更新及共享的需要，针对地震应急信息产品属性、服务、时间、传递等特性，在借鉴地震应急基础理论研究及相关行业分类标准的基础上，研究地震应急信息产品的分类方法。

但对于多渠道的上传机制，加之震后大量的灾情及背景信息，使信息归类难度较大。面对紧迫的时效性压力和不同指挥决策部门对信息的不同需求，仅靠人工手动进行信息分类提取的方式难以达到令人满意的效果，因此建立条理更为清晰、标准更具实践应用意义、信息自动化程度更高的信息分类管理技术十分必要，以适应应急指挥决策部门对应急救援信息的快速获取要求。

林子雨等（2010）根据关系数据库的关键词查询问题研究背景，阐述解决该问题的基于模式图和数据图的优缺点、困难和挑战，提出利用排序函数解决关键词查询时匹配结果可能很多的情况，最终反馈给用户一个最相关信息。张晓民（2017）设计了基于关键词数据库信息检索方法及时态检索算法，主要采用时间修剪策略，同时提出时态边权重的计算方法，实现了基于关键词的关系数据库时态检索原型系统。通过借鉴关键词在信息检索中的应用，本文将关键词分类法应用于地震应急信息管理中。

1 地震应急信息分类方法

信息分类方法主要包括线分类法、面分类法、混合分类法（耿庆斋等，2014）。现有与地震信息分类有关的标准与研究多采用线分类法，其特点是层次较清晰，易于理解；缺点是结构可塑性较差，一旦分类深度和每层级类目容量固定后，修改层级和插入新类将受限（刘若梅等，2004）。面分类法将选定的分类对象若干属性或特征视为若干个“面”，每个“面”中又可分成彼此独立的若干个类目，对于解决同种类型要素在不同应用中分类的矛盾具有优势。

参考不同分类方法（杨天青等，2016；和锐等，2011），考虑自动分类结果的时效性与实用性，本文采用线与面相结合的混合分类法，以信息服务的高效便捷为目的，按照应急信息自身的特征属性、地震发生时间线产生的直接与间接损失信息（即震前、震时与震后所造成的破坏与损失信息），针对产生的影响采取相应的应急救援信息，将地震应急信息分为震前基础背景信息、地震震情灾情信息、震后应急救援信息，如表 1 所示。

表 1 地震应急信息分类定义

Table 1 Definition of classification of seismic emergency information

信息类别	定义	内容概述	意义
震前基础背景信息	震区客观存在的各类自然灾害、基础社会信息	震区人文、构造，救援力量背景、基础设施设备，自然灾害等	对地震应急行动起宏观的参考价值
地震震情灾情信息	地震自身具有的动态性、可变性的各类自然属性特征与其间接造成的破坏或损失信息	地震发生的时间、地点、震级、余震监测信息、强震动监测信息、震情发展趋势判断、成因分析及地震烈度分布、房屋破坏与人口伤亡信息、次生灾害信息等	判断地震灾害的规模和成灾方式，直接控制地震应急救援行动的总体规模和方式
震后应急救援信息	为应对地震事件、处理地震灾害而采取的应急救援信息	应急救援决策信息、救援调度信息、指挥部工作状态、现场工作状态、具有相关参考性的历史救援案例信息等	掌握救援行动的进展，动态调整救援方案

2　地震应急信息自动分类方法的研究

（1）通过实地调研河北省、山西省、内蒙古自治区、四川省的基本人文地理环境信息概况，本文选择收集四川省 4 次地震应急资料的主要原因为：1）对同一省份的地震应急资料进行文档分词处理时，可直接忽略地名类固定性且不具实际区分意义的属性词，且同一省份文本文档之间的语义描述差异性相对较小；2）相对于地震易发的其他 3 个省来说，四川省地势地形地貌相对较复杂，建筑物水库大坝等公共基础设施种类结构相对复杂，且抗震救灾技术较成熟，从而使得到的信息更丰富和全面；3）四川省已建成一套独立的信息上传与协同管理体系，有助于提高资料分析和研究的准确性。

（2）应急信息资料分析统计

共收集 2013 年 4 月 20 日芦山 7.0 级地震、2014 年 11 月 22 日康定 6.3 级地震、2017 年 8 月 8 日九寨沟 7 级地震、2017 年 9 月 30 日广元青川 5.4 级地震资料，由于收集到的数据较零散，且震级较小的数据资料较少，所以本文将 4 次地震中相同类别的信息统计在同一文件夹下，如表 2 所示。

表 2　信息文档分类统计

Table2　Classification statistics of information documents

类别	文档数
震前基础背景信息	12
地震震情灾情信息	405
震后应急救援信息	138
总文档量	555

（3）应急信息分类关键词的选取

中文分词(Chinese Word Segmentation）指将一个汉字序列切分成一个一个单独的词，作为文本挖掘的基础，对输入的一段中文进行中文分词，可达到自动识别语句含义的效果（赵小华，2010）。

TF 词频(Term Frequency）指某一个给定的词语在该文件中出现的次数。IDF 反文档频率(Inverse Document Frequency）的主要思想是：如果包含词条的文档越少，IDF 越大，则说明词条具有很好的类别区分能力。TF-IDF 是一种用于信息搜索和信息挖掘的常用加权技术，在搜索、文献分类和其他相关领域中的应用较为广泛（施聪莺等，2009）。

本文在对文本信息进行分析处理时，根据建立的分类标准，对收集到的信息进行分类，应用 TF-IDF 技术，在 Excel 表里对各类文本信息进行分词和词频统计。此种方法的局限是处理的文档只能是文本文档（.txt）格式。按名词和动词的词性，统计 IDF 和词频数排名前 20 的词，如图 1-3 所示。

由图 4 可知，地震、级地震、地震局、水库 4 个词语的出现总频数超过 1000，其中地震出现频数高达 2439。各类别信息里的频数具体为：震区背景信息 119 次、震区震情灾情信息 1105 次、灾区应急救援信息 914 次，占各类别信息前 20 频数的比例分别为 9%、15%、13%，在总文档里所占比例为 16%，平均出现频率占 12.3%。

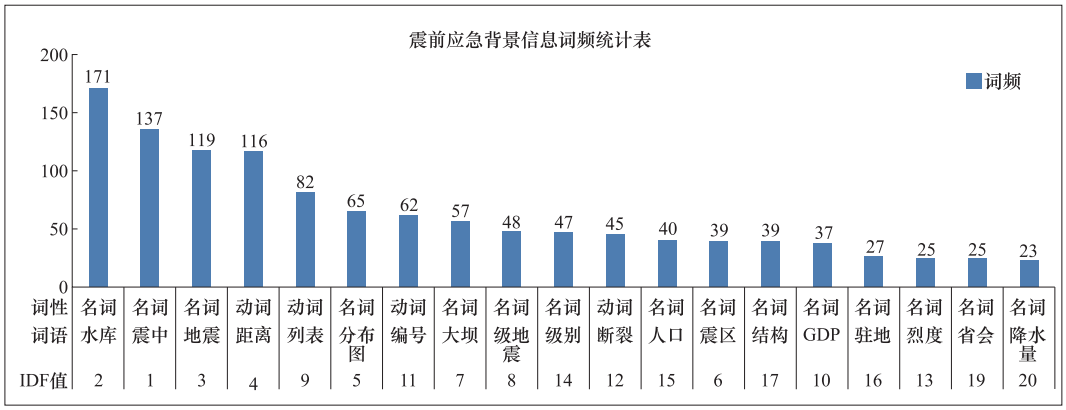

图 1 震前应急背景信息词频统计

Fig. 1 Frequency statistics of emergency background information before earthquake

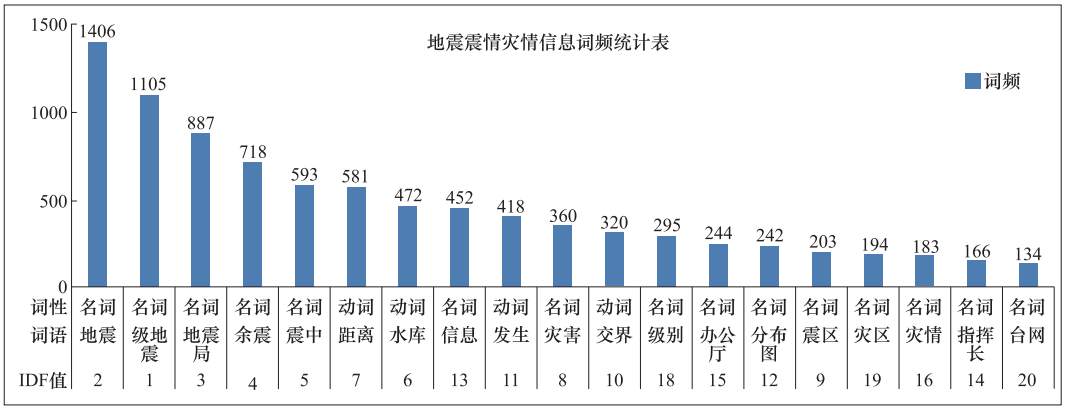

图 2 地震震情灾情信息统计

Fig. 2 Statistical table of disaster information in earthquake area

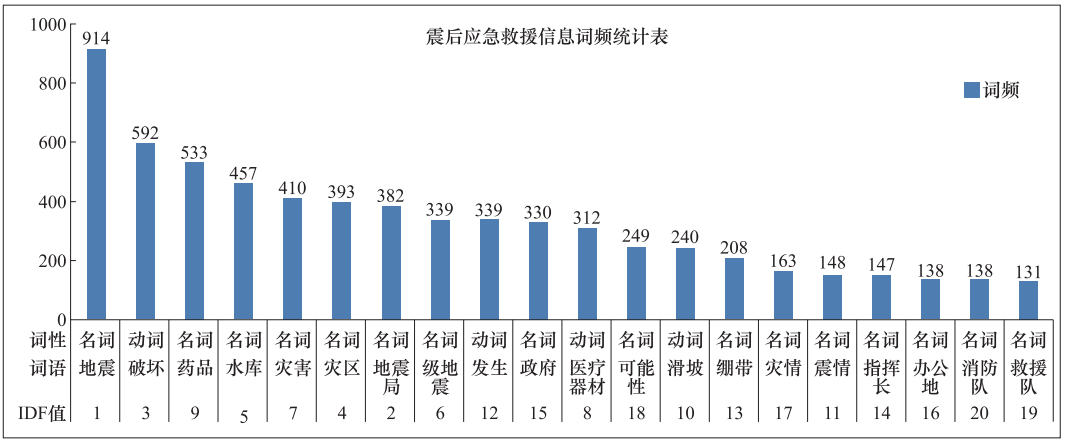

图 3 震后应急救援信息词频统计

Fig. 3 Frequency statistics of emergency rescue information after earthquake

对未分类的所有初始文本进行统计，结果如表 6 所示。

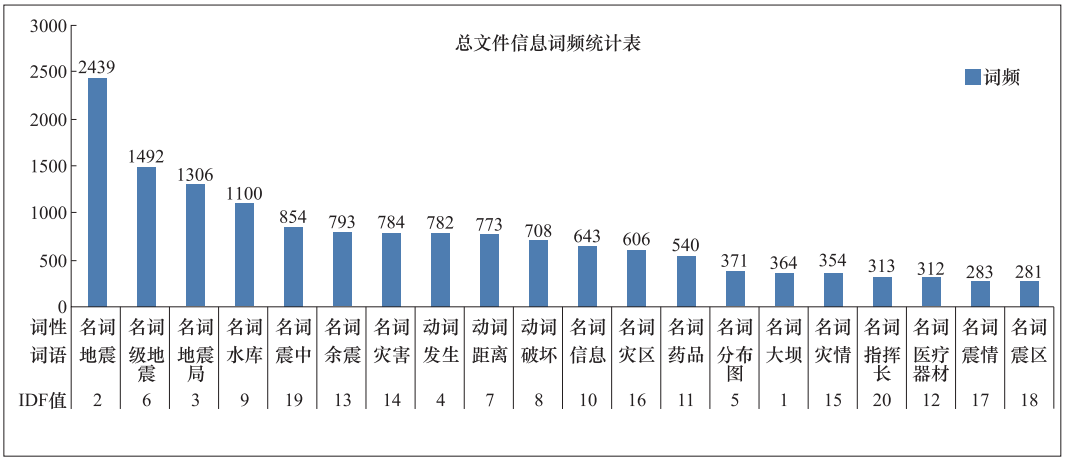

图 4　总文档信息词频统计

Fig. 4　Total Document Information frequency table

频数为 700—1000 的词语共 6 个，分别为震中 854 次、余震 793 次、灾害 784 次、发生 782 次、距离 773 次、破坏 708 次，占所有词频的比例为 4.7%—5.7%，其中发生和灾害 2 个词语的频数相差 2，在进行词语筛选时，任选其一即可。

频数为 300—700 的词语共 8 个，其中 400 以上的有 3 个，分别为信息 643 次、灾区 606 次、药品 540 次；其余 5 个为分布图、大坝、灾情、指挥长、医疗器材，频数为 300—400。8 个词语从分类属性来看，主要属于应急救援信息，占总文档词语的比例为 2%—4%。

整体来看，出现频率越高的词语，在分类过程中起到的作用越低，即作为关键词的代表性越不强，本文最终选取的各类别信息关键词是在各类信息词语统计里频率不高且在其他类别信息里频率较低或没有的词语。根据频数统计规律可知，本文关键词的取舍主要按以下规则：①对 4 个频数数据按词语词频占所有 20 个词语词频的比例，将频率域划分为 2%以下、2%—4%、4%—6%、6%—8%、8%五个区间；②按各类信息的定义，每个区间选取一个词（选取与本类信息最相关的词语）作为 3 类信息的基础关键词。如第一区间地震局、第二区间水库、第三区间破坏、第四区间灾情、第五区间震情，这个组合归至震情灾情信息类；③每个区间选取 2—4 个固有关键词，与基础关键词重合的排除，低频率区间的词语多选，重复词语与高频词语尽量不选，最终每类信息选出 15 个关键词，如表 7 所示。

表 3　关键词选取结果

Table 3　Keyword selection results

震前基础背景信息	地震灾情震情信息	震后应急救援信息
地震	级地震	地震局
水库	水库	水库
距离	截止	药品
大坝	交界	政府
震中	震情	绷带

续表

震前基础背景信息	地震灾情震情信息	震后应急救援信息
人口	震中	搜救
结构	余震	消防队
烈度	灾害	救援队
省会	灾情	医疗器材
分布图	大坝	办公地
降水量	经度	发生
断裂	灾区	可能性
GDP	防震减灾	食品
房屋	发生	震情
震源	烈度	灾区

3 地震应急信息的自动分类

百度、谷歌等搜索引擎成功显示出关键词检索的方式已被广大用户所接受（张晓民，2017）。本文为解决应急信息的自动分类，采用 “关键词分类法”，根据分类标准，对原始文本进行结构化处理，通过中文分词、词频筛选与统计实现信息关键词的提取，此阶段中的中文分词将一串连续汉字序列按动词、名词的规范重新组合成词语序列。词频统计与筛选即对分词结果进行统计，去除一些无效词后，生成关键词词库，用匹配词库的方法实现信息的自主分类，具体过程如下：①收集震后国家中心、各研究所、各省（自治区）地震局上传至应急信息共享平台、评比 FTP 站点、台网中心台网部 FTP 站点的震后产出成果，建立相对完整的产出目录；按照之前建立的地震应急信息分类标准，对收集到的条目进行梳理归类。②对所有文档按词性进行词频统计，将无效词语去除后，对每个大类建立相应的关键词词库。由于高频词语的重合度较高，因此在建立关键词词库时，需综合考虑词频和词语含义，首选该分类独有且出现频率较高的词语。③以提取的特征词作为自动分类程序中的词库，进行自动分类处理，在计算机语言的基础上，实现信息的自动分类。要求程序在震后启动，自动完成当前地震产生在各不同平台上的信息分类，并将产出成果保存至本地服务器。根据已建立的分类类别和各应急指挥部门需求，可进一步实现对产出成果的重命名（非必要）和重新分发。分类流程如图 5 所示。

以九寨沟 7.0 级地震产出为例：

报告及图件总数如表 8 所示。分类文件夹包括震前背景信息文件夹、震区灾情震情信息文件夹、震后应急救援信息文件夹和其他文件夹。

建立的分类词库较简单，结果与表 3 的关键词库高度匹配。震前背景信息特征词包括构造、交通、居民点、GDP、人口等，地震震情灾情信息特征词包括截止、余震、热力图、震动、态势、数据、精密、水准、伤亡、灾害、中央电视台、设防、展开、遇难等，震后应急救援信息特征词包括救援、救援队、搜救等。

分类标准建成后，以提取的关键词作为自动分类程序中的词库，进行自动分类，流程如图 6 所示。分类过程中各环节为：①将所有格式文档转为.txt 格式文件，并输出至原始文件

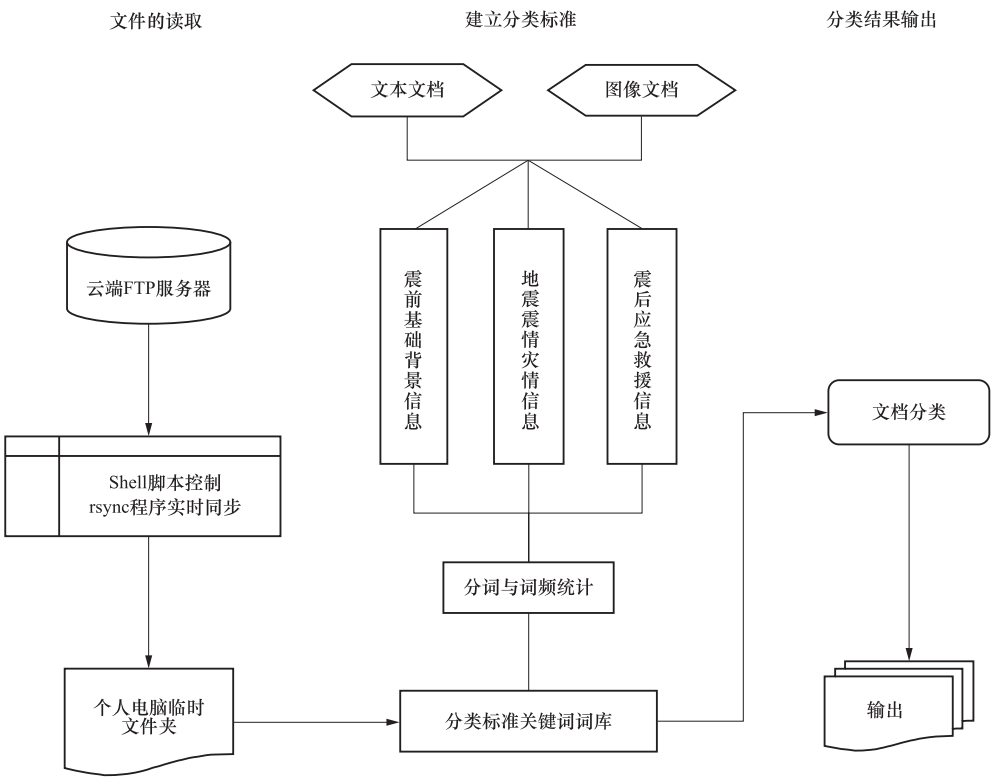

图 5　分类流程

Fig. 5　Classification flowchart

表 4　报告及图件总数

Table 4　Total number of reports and artworks

信息文件类别	文档	图件	文件总量
震前基础背景信息	21	0	21
地震震情灾情信息	119	13	132
震后应急救援信息	7	0	7

夹；②搭建主程序运行环境（Python2.7 环境、jieba 程序库）；③运行 shell 主程序，调用 Python 子程序模块，将原始文件夹下的所有文件进行分类处理。模块 1（cut）：获得文件对文件进行分词，并将其存至临时文件夹；模块 2（count）：对原文件进行词频统计，并对统计结果进行排序；模块 3（order）：分词词频统计排序前 15 的词进行排序；模块 4（set）：根据各类关键词筛选结果，得到关键词库；模块 5（classify）：将初始文档进行结构化处理后得到的前 15 词频作为该文档的关键词，将其与关键词库进行对比，通过文档关键词在所划分的 5 个频率域区间的关键词库匹配率决定文档的归属类别，将文档划分至匹配率最高的类别。判断该关键词属于哪个分类，按照文件归属，把文件归类至该目录下。某个文件可能属于多个类别，如果没有对应的目录，则把文件拷贝至其他文件夹。

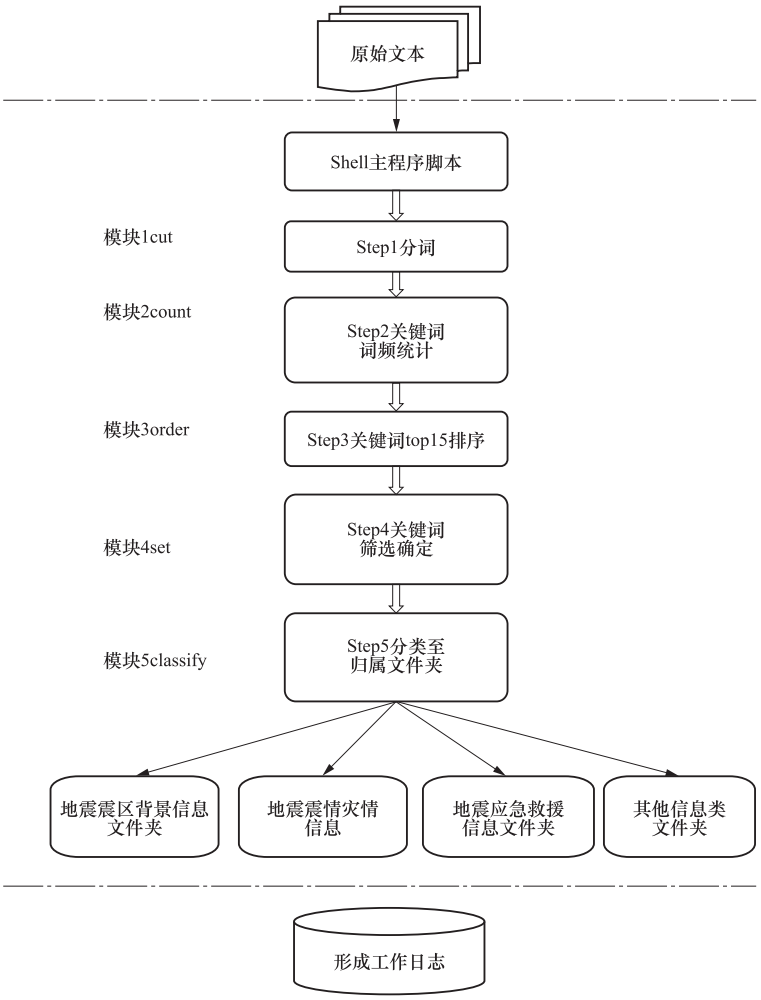

图 6 自动分类流程

Fig. 6 Flowchart of automatic classification

4 总结和应用探讨

目前我国地震应急信息是通过各省、市已建立的信息汇总渠道直接上传至相关服务平台，供指挥部及相关领导专家参阅，但大地震发生后面对的是大量灾情震情救援及背景信息，仅靠上述传输和提取方式不能达到令人满意的程度。本文的研究成果可实现多渠道应急信息的自动分类，辅助地震应急指挥控制与决策等。

（1）参考以往学者在地震应急信息分类与编号方面的研究，考虑分类信息的服务实用性，根据地震事件发生的时间轴，将地震应急信息分为震前应急背景信息、地震应急震情灾情信息和震后应急救援信息。

（2）为实现地震应急信息的自动分类，研究采用 “关键词分类法”，以实现地震应急信息的自动分类，提高信息处理的目标性、针对性和有效性。

（3）通过分析，本文对应急信息进行分类、分词、词频统计，由前 15 位关键词信息统计

结果可知，各不同类别应急信息关键词之间存在较大差异，可见与传统信息直接上传法相比，“关键词分类法”能使信息条理性更强，分析处理时更方便直接。

（4）在大数据的背景下，相比于传统的信息分类方法，实现地震应急信息的自动分类，将大大提高信息利用率，并推动地震应急救援相关技术走向智能成熟化、自动服务化。

但对于有效应用关键词分类法实现应急信息的自动分类、降低某个文件可能属于多个类别的交叉情况，仍存在以下问题：

（1）如何建立关键词之间的语义关系和逻辑关联关系，处理并不断丰富分类关系树，还需对信息自身与信息相互之间更深层次的关联关系进行探讨，如时态上或语义上。

（2）对于关键词重复和冗余问题，目前只有少数研究提出了初步解决方案，还需结合信息自身的属性、信息之间的差异及用户对信息的需求，由相关函数（如排序函数）探索建立一个权衡的标准。

参考文献：

白仙富，李永强，陈建华等，2010．地震应急现场信息分类初步研究．地震研究，**33**（1）：111—118，120．

付继华，王建军，刘晓晢等，2009．灾情数据自动获取的地震灾情信息系统．数据采集与处理，**24**（S1）：310—314．

耿庆斋，王冠华，张伟兵等，2014．面向信息共享的科研单位信息分类编码体系研究．水利技术监督，**22**（4）：21—24．

和锐，冯义钧，张翼，2011．地震信息分类与编码研究．中国地震，**27**（3）：327—334．

梁凯利，吕金霞，王丽威等，2011．专业分类表的编制修订初探——以《中国图书馆分类法·地震学专业分类表》为例．国家图书馆学刊，**20**（2）：31—33，79．

聂高众，陈建英，李志强等，2002．地震应急基础数据库建设．地震，**22**（3）：105—112．

林子雨，杨冬青，王腾蛟等，2010．基于关系数据库的关键词查询．软件学报， **21**（10）：2454—2476．

刘若梅，蒋景瞳，2004．地理信息的分类原则与方法研究——以基础地理信息数据分类为例．全国地图学与GIS 学术会议论文集．

施聪莺，徐朝军，杨晓江，2009．TFIDF 算法研究综述．计算机应用，**29**（S1）：167—170，180．

杨天青，姜立新，董曼等，2016．基于共享模式的地震灾情集成发布平台设计与实现．震灾防御技术，**11**（2）：375—383．

张晓民，2017．基于关键词的关系数据库时态信息检索方法研究．大连：大连海事大学．

张翼，唐姝娅，王悦等，2016．地震应急信息产品分类编码研究．震灾防御技术，**11**（1）：132—143．

赵小华，2010．KNN 文本分类中特征词权重算法的研究．太原：太原理工大学．

Detrick G.，2002.Russell L.Ackoff.Academy of Management Learning and Education, **1**（1）：56—63.

Research on the Method of Automatic Classification in Earthquake Emergency Information

Wang Lin[1)], Jiang Lixin[2)], Yang Tianqing[2)] and Zhang Weijia[2)]

1）Institute of Earthquake Science, China Earthquake Administration,Beijing 100036,China
2）China Earthquake Networks Center, Beijing 100045,China

Abstract The efficient processing of seismic emergency information provides important support for earthquake emergency rescue. This paper, based on the demand of seismic emergency information classification, constructs an efficient and convenient method for classifying and processing seismic emergency information. According to the timeline before, during and after the earthquake, the seismic emergency information is classified into basic background information before the earthquake, information during the earthquake and information about the earthquake rescue. With the method of 'Key-word Classification' adopted, the emergency information collected from multiple ways is classified automatically with the support of computer language, which shortens the time needed for information processing and improves the efficiency and convenience of information service for emergency command.

Key words: Earthquake; Emergency information; Automatic classification; Keywords classification