第 37 卷　第 4 期(819~828)　　中 国 地 震　　Vol. 37　No. 4
2021 年 12 月　　EARTHQUAKE RESEARCH IN CHINA　　Dec. 2021

文鑫涛,郑通彦,王钟浩,等,2021. 面向历史灾害地震的 Web 信息精确抽取与分析方法. 中国地震,**37**(4):819~828.

面向历史灾害地震的 Web 信息精确抽取与分析方法

文鑫涛[1)]　郑通彦[1)]　王钟浩[2)]　李华玥[1)]
李晨曦[2)]　吕文超[2)]

1)中国地震台网中心,北京　100045
2)防灾科技学院,河北三河　065201

摘要　以中国大陆地区灾害地震目录为基础,选取 2010—2019 年灾害地震的互联网信息,提出基于百度搜索引擎的信息获取技术,并以“时间、地名、震级”为关键词,设计一套 URL 生成规则。使用该技术进行百度检索,得到前 100 个站点的主体文字信息,建立地震信息基础语料库,形成灾害地震的网络灾情信息获取方法;通过采用已有的停用词词库剔除无用信息,对爬取到的信息进行初步清洗工作,进一步深入挖掘隐含信息,探索灾害关联关系,为震后互联网灾情信息快速获取建立基础。

关键词:　灾害地震　Web 信息抽取　灾情信息获取　数据分析
[文章编号] 1001-4683(2021)04-0819-10　[中图分类号] P315　[文献标识码] A

0　引言

基于地震灾害评估报告、地震应急基础数据等形成的地震宏观震情、灾情等真实调查信息是灾害评估和应急处置建议等研究的重要参考依据,但在深入分析灾情特点和提出处置建议时,结构化的数据形式仍存在诸多局限。

近年来,互联网信息的抽取技术发展迅速,其主要分为基于模板的抽取方法和与模板无关的全自动抽取方法(张儒清等,2017)。封化民等(2005)提出含有位置坐标树的 Web 页面分析和内容提取框架,在对网站信息的内容提取时,考虑了位置特征和空间关系,并通过对 120 个网站的 5000 个网页进行测试,结果表明该方法的准确率可达 93.78%。张儒清等(2017)实现与模板无关的全自动抽取算法,并开展基于模板的抽取算法的融合研究,结果显示,这种融合机制能有效提高抽取准确率,从而建立一个适用于任意网页、具有实用价值的信息抽取框架。张恺航等(2019)引入通配符节点话题权重的 Web 新闻抽取方法,降低 Web 新闻内容边缘噪音文本的错误识别率,提高抽取的新闻内容准确率。

目前 Web 提取与分析技术在其他相关领域得到了广泛应用,但在地震应急领域方面的

[收稿日期] 2021-07-09　[修定日期] 2021-11-01
[项目类别] 地震应急信息快速可视化技术研究(2018YFC1504506)资助
[作者简介] 文鑫涛,男,1988 年生,工程师,主要从事地震应急相关工作。E-mail:wenxintao@seis.ac.cn
郑通彦,通讯作者,女,1982 年生,高级工程师,主要从事地震应急相关工作。E-mail:zhengtongyan@seis.ac.cn

应用较少(庞晓克等,2019)。因此,本文以灾害地震目录为基础,截取 2010 年 1 月 1 日至 2019 年 12 月 31 日发生的总计 362 个灾害地震(图 1),以“发震时间、震中位置、震级”为关键词,爬取百度搜索前 100 位的网页正文文字内容,实现历史灾害地震互联网信息爬取、数据存储的全流程自动化,即通过读取历史灾害地震数据库、自动生成搜索关键词、依据关键词生成搜索引擎地址链接、获取该链接的前 10 页(100 条)URL 地址,爬取、解析网页正文文本内容。此外,按照既定的文本清洗策略和文件存储规则,对爬取到的数据进行预处理、数据清洗、分词和统计归类等工作。通过以上工作,一方面为震后灾情网络信息的快速获取探索可实现的方法,另一方面通过“历史灾害地震网络灾情信息数据库”的建立,为引入大数据分析方法提供一定的数据基础。

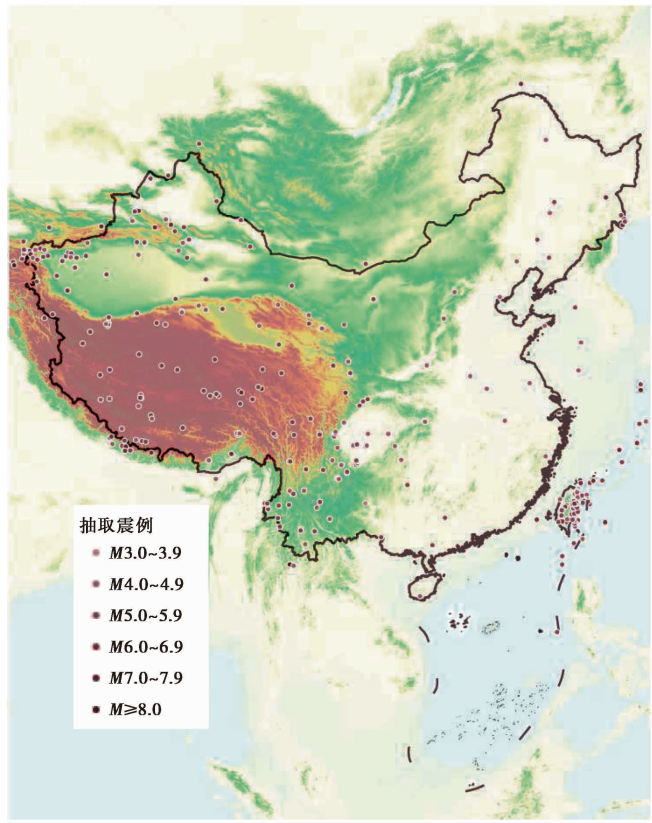

图 1 爬取震例的震级和空间分布

1 Web 信息精确抽取与分析框架

根据历史地震获得的关键词,从互联网中抽取相关信息,再对相关信息进行文本数据的预处理,包含文本信息去重和清洗过滤垃圾文本等,建立地震网站信息基础语料库,并在此基础上对地震信息进行统计与分析(张开敏,2014)。具体步骤如下:

(1)采用基于搜索引擎的信息获取技术,针对互联网地震信息的特点,设计地震信息搜索链接的 URL 生成规则,并生成地震信息 URL 链接列表。对地震信息 URL 链接列表中的

网页站点进行访问，针对网页结构一致性要求较高、算法复杂和实现效率较低的问题，采用并行的网页解析算法，对网站信息进行批量化解析，建立地震信息基础语料库。

（2）使用 Simhash 算法，结合 BP 神经网络对地震信息基础语料库中的重复文本信息和垃圾文本信息进行清洗，形成去重过滤后的地震信息基础语料库，为下一步的数据规范化处理做好准备。

（3）读取地震网站信息基础语料库，采用改进的 TF-IDF 算法对语料库中的地震信息文本进行训练、统计与分析，探索灾害信息间的关联，挖掘地震文本中包含的人员伤亡、地质条件、次生灾害等灾害发生后的急需信息，为震后互联网灾情信息快速获取建立基础（图 2）。

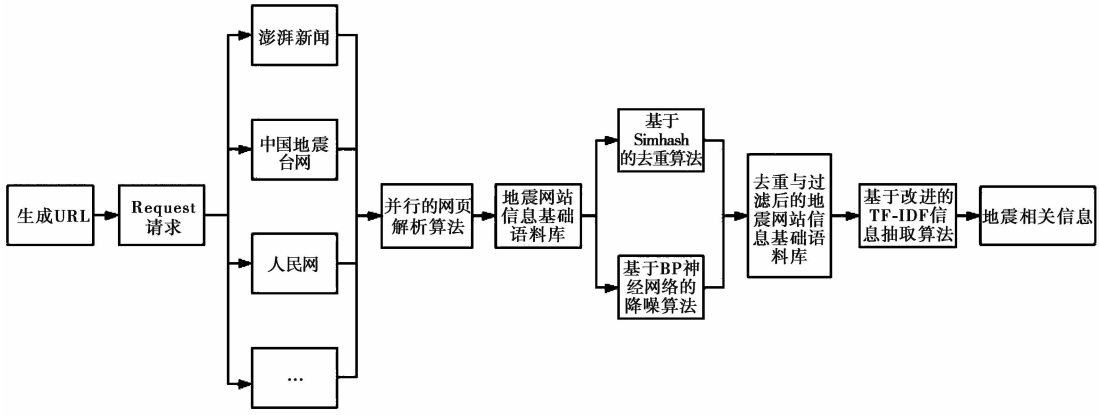

图 2　Web 信息精确抽取与分析框架

2　关键技术及算法

2.1　地震 Web 信息的精确抽取方法与数据获取方法

2.1.1　基于百度搜索引擎的 URL 数据获取

在互联网时代，搜索引擎是获取信息的最佳选择，也是获取地震 Web 信息的常用工具。百度、谷歌、必应（bing）为常见的搜索引擎，其中百度搜索引擎支持中文编码标准，基于字、词结合的信息处理技术符合地震三要素的中文提取方式。同时，百度搜索引擎的检索结果能够标识出网页的基本属性（如标题、网址、时间、大小、编码、摘要等），方便进行地震相关信息的提取。因此，本文采用基于百度搜索引擎的震后 Web 信息获取技术。

基于历史灾害地震信息数据库，获取 2010—2019 年的所有地震震例的三要素信息，构建地震关键词基础语料库，再结合百度搜索链接的 URL 生成规则和基于禁忌搜索算法的爬虫主题分析技术，进行地震专题 Web 页面 URL 数据获取。

以 2017 年四川九寨沟 7.0 级地震为例，依据历史灾害地震数据库数据“2017 年 8 月 8 日四川九寨沟 7.0 级地震”，对其字符进行中文编码，结合百度搜索链接的 URL 生成规则，进行 URL 的拼接，生成一级 URL；进一步采用 Requests 库，基于一级 URL 地址对浏览器进行请求，获得一级 URL 对应的所有二级 URL 地址；最后使用 BeautifulSoup 库对二级 URL 地址中的 HTML 源文件进行解析，得到震后 Web 信息的二级 URL 元数据。表 1 为根据地震关键词信息所获取的不同地震的二级 URL 数量。

表 1 二级 URL 元数据爬取数量

时间(年-月-日)	震中	震级	URL 元数据量
2017-08-08	四川九寨沟	7.0	99
2015-11-14	东海海域	7.2	108
2016-06-26	吉尔吉斯斯坦	6.7	102
2012-08-12	新疆于田	6.2	109
2015-04-15	内蒙古阿左旗	5.8	102
2017-05-11	新疆塔什库尔干	5.5	99
2013-08-28	云南德钦	5.1	97
2018-08-03	青海治多	5.1	128
2014-05-30	云南盈江	5.1	98
2017-01-28	四川筠连	4.9	120

2.1.2 网页站点信息的精确抽取方法

基于二级 URL 元数据进行网页正文信息的抽取是地震 Web 信息抽取的核心工作,对这些正文信息进行深入分析,可获得更有价值的地震灾害深层信息。

常用的正文抽取方法包括:采用 DOM 树结构抽取、应用机器学习中的聚类分析抽取以及隐马尔可夫模型等方法,这些方法或存在对网页结构一致性要求较高的问题,或存在算法复杂、实现效率较低的问题,对地震网页信息的多源异构和震后信息快速获取要求时效性较高的情况均不适用。因此,本文采用了一种并行的异构网页解析算法,该算法基于数据并行的方案,通过将输入数据划分成多个部分,对其进行并行处理,再合并各个部分的结果以得到最终结果。

该算法的基本思想是将 HTML 文档划分成多个片段,每个片段包含 1 个或多个 HTML 单元,随后使用传统的串行解析算法同时解析多个片段,最后将片段的解析结果归并得到此文档最终的解析结果。以 2017 年四川九寨沟 7.0 级地震为例,对获取到的 1 个二级 URL 内容进行解析,算法的运行过程分为以下 3 个步骤:

(1)分段:逐字符地扫描 HTML 并找出所有“<”字符。将 HTML 分割成 N 个片段 F_1、F_2、…、F_N,其中每个 F_N 均以“<”起始。由于地震文本信息中的地震三要素、地形地貌、人口信息与伤亡人数等关键信息大概率分布在不同的 HTML 标签内,且内容相对独立,因此这些内容被分割到了不同片段中,可以用于下一步的并行解析。

(2)并行解析:并行解析分段步骤所得到的片段 F_1、F_2…F_N,并将解析出的 HTML 单元放入全局集合 R 中(初始化为空)。对 1 个片段 F_K 解析时,首先创建并初始化 1 个空白的有限状态机 FSM_K,然后从 F_K 的起始位置 SF_K 开始,使用 FSM_K 解析 HTML 文本。每解析出 1 个 HTML 单元 U,将 U 加入 R 中,如果 U 引用了外部内容(例如引用其他地震网站的数据信息),则根据引用的 URL 下载此内容。

(3)归并:为了保证最终结果的正确性,归并是 1 个串行的过程。按照在 HTML 文档中出现的顺序对 R 中的 HTML 单元进行排序,随后依次合并 HTML 单元形成 DOM 树,并使用 BeautifulSoup 提取出所有的文本内容,包括地震三要素、地形地貌、人口信息、伤亡人数等有

用信息和一些无用信息（垃圾信息）。

基于上述并行异构网页解析算法，可对二级 URL 元数据所指向的 HTML 内容进行正文信息的快速精确抽取，并建立地震信息基础语料库（图 3），并对 2010—2019 年地震 web 正文信息进行入库操作，为后续数据清洗和分析操作提供数据支持。

label	text	cutword	cutwordnum	Earthquake_name	time	url
地震信息	昨日21时19分在四川阿坝州九寨沟县（北纬33.20度，	昨日 21 时 19 分在 四川 阿坝州 九	2800	九寨沟地震	2017年8月9日	http://blog.sina.com.cn/s/blog_48d84e6f0102wu7u.html
地震信息	原标题：泪崩！九寨沟地震前后对比照，上天收回了	原 标题 ： 泪崩 ！ 九寨沟 地震 前	2872	九寨沟地震	2017年8月19日	https://www.sohu.com/a/165774442_661771
地震信息	四川省阿坝州九寨沟7.0级地震8日夜发生后，中新社记	四川省 阿坝州 九寨沟 7.0 级 地震 8	813	九寨沟地震	2017年8月9日	http://news.sina.com.cn/c/nd/2017-08-09/doc-ifyitamv
地震信息	用微信扫描二维码分享至好友和朋友圈扫描到手机&tin	用 微信 扫描 二维码 分享 至 好友 和	989	九寨沟地震	2017年8月9日	https://news.china.com/focus/jiuzhaigoudz/news/13000
地震信息	最新消息-九寨沟地震与三峡工程有关？工程院院士回	最新消息 - 九寨沟 地震 与 三峡工程	190	九寨沟地震	2017年8月9日	http://www.jxnews.com.cn/zt/system/2017/08/09/016331
地震信息	四川九寨沟7.0级地震已造成25人死亡 1人未确定身份	四川 九寨沟 7.0 级 地震 已 造成 25	7405	九寨沟地震	2017年8月13日	http://news.cyol.com/node_61044.htm
地震信息	首页新闻博客人才会议基金论文视频小柯机器人帮助He	首页 新闻 博客 人才 会议 基金 论文	884	九寨沟地震	2017年8月11日	http://blog.sciencenet.cn/home.php?mod=space&uid=279
地震信息	原标题：关于九寨沟地震，你最好弄清楚这10件事！1	原 标题 ： 关于 九寨沟 地震 ， 你	2670	九寨沟地震	2017年8月11日	https://www.sohu.com/a/163804053_246933
地震信息	2017九寨沟地震受伤死亡人数汇总 九寨沟7.0级地震伤	2017 九寨沟 地震 受伤 死亡 人数 汇	1094	九寨沟地震	2017年8月8日	http://www.mnw.cn/news/shehui/1810723.html
地震信息	划重点：截至8月13日20时，地震造成25人遇难(其中24	划 重点 ： 截至 8 月 13 日 20 时 ，	1098	九寨沟地震	2017年8月14日	https://news.qq.com/a/20170808/059190.htm
地震信息	中新社成都8月11日电（贺劭 清杨珺)2008年汶川地震、	中新社 成都 8 月 11 日电 （ 贺劭	1171	九寨沟地震	2017年8月11日	http://www.chinanews.com/gn/2017/08-11/8302315.shtml
地震信息	该楼层疑似违规已被系统折叠 隐藏此楼查看此楼当然，	该 楼层 疑似 违规 已 被 系统 折叠	122	九寨沟地震	2017年8月10日	http://tieba.baidu.com/p/5266159019
地震信息	四川省阿坝州九寨沟7.0级地震8日夜发生后，中新社记	四川省 阿坝州 九寨沟 7.0 级 地震 8	813	九寨沟地震	2017年8月9日	http://news.sina.com.cn/c/nd/2017-08-09/doc-ifyitamv
地震信息	手机用户编辑数字“1”或“2”发送短信至“1069999	手机用户 编辑 数字 “ 1 ” 或 “	243	九寨沟地震	2017年8月14日	https://news.china.com/focus/jiuzhaigoudz/
地震信息	原标题：九寨沟地震发生后，这几张图片传疯了，看完	原 标题 ： 九寨沟 地震 发生 后 ，	451	九寨沟地震	2017年8月10日	https://www.sohu.com/a/163582780_394179
地震信息	网站简介|产品服务|招聘信息|广告服务|联系我们|版	网站 简介 | 产品 服务 | 招聘 信息	110	九寨沟地震	2017年8月9日	http://news.hsw.cn/xwzt08/xw/jzgdz/
地震信息	作者：刘忠俊 来源：中国新闻网 发布时间：2017/8/1	作者 ： 刘忠俊 来源 ： 中国 新闻	1495	九寨沟地震	2017年8月10日	http://news.hexun.com/2017-08-10/190394339.html
地震信息	四川九寨沟发生7.0级地震 关于它的成因都在这里了百	四川 九寨沟 发生 7.0 级 地震 关	1935	九寨沟地震	2017年8月9日	https://baijiahao.baidu.com/s?id=1575222930988414&wf
地震信息	四川九寨沟地震日记，地震无情人有情，关于地震中各	四川 九寨沟 地震 日记 ， 地震 无情	1621	九寨沟地震	2017年8月15日	http://www.qbaobei.com/jiaoyu/943043.html
地震信息	请点击“”选择在浏览器打开全网最火短视频九寨沟地	请 点击 “ ” 选择 在 浏览器 打开	33	九寨沟地震	2017年8月10日	http://t.cn/R9Rj2Xe
地震信息	浙江在线8月9日讯（浙江在线编辑 洪慧敏）正值旅游	浙江 在线 8 月 9 日讯 （ 浙江 在线	2303	九寨沟地震	2017年8月9日	http://china.zjol.com.cn/ktx/201708/t20170809_476925
地震信息	姹躲窚鏀伴潹 鎴戝湪楦樹腑 鎴戝爜鐲ュ垚闅戒笂澶у	姹 躲 窚 鏀 伴 潹 鎴 戝 湪 楦 樹	105	九寨沟地震	2017年8月23日	https://m.iask.sina.com.cn/comments/bp-598c2d99e4b01
地震信息	2008年5月12日14时28分，相信很多国人都没有忘记这	2008 年 5 月 12 日 14 时 28 分 ，	855	九寨沟地震	2017年8月15日	http://yuqing.people.com.cn/n1/2017/0815/c209043-294
地震信息	康康侃球分享最独特的体坛资讯，不仅仅关注足球2017	康康侃球 分享 最 独特 的 体坛 资讯	172	九寨沟地震	2017年8月10日	https://zhidao.baidu.com/question/437558343608264484
地震信息	【字体：大 中 小】 语音播报8月8日21时19分，四川	【 字体 ： 大 中 小 】 语	899	九寨沟地震	2017年8月10日	https://www.douban.com/note/632442290/
地震信息	实探合肥高新区乐富强悦湖合肥二手房价格跌了！这些	实探 合肥 高新区 乐 富强 悦湖 合肥	1847	九寨沟地震		http://365jia.cn/news/2017-08-10/8CD7427AF21F9A29_al
地震信息	实探合肥高新区乐富强悦湖合肥二手房价格跌了！这些	实探 合肥 高新区 乐 富强 悦湖 合肥	1847	九寨沟地震	2017年8月9日	https://www.douban.com/note/632442290/
地震信息	假如你遇到九寨沟地震怎么办 自救手册值得你拥有 -	假如 你 遇到 九寨沟 地震 怎么办	3557	九寨沟地震	2017年8月9日	https://www.d1xz.net/wenhua/minzu/art217627.aspx
地震信息	您的浏览器Javascript被禁用，需开启后体验完整功能	您 的 浏览器 Javascript 被 禁用 ，	19	九寨沟地震	2017年8月10日	https://wenku.baidu.com/view/0e4839d5988fcc22bcd126f
地震信息			2	九寨沟地震	2017年8月8日	https://www.iqiyi.com/v_19rr8hchhk.html
地震信息	扫描二维码下载新浪视频AppCopyright ? 1996-2017	扫描 二维码 下载 新浪 视频 AppCo	37	九寨沟地震	2017年8月11日	http://video.sina.com.cn/p/news/c/doc/2017-08-08/225

图 3　地震信息基础语料库

2.2　地震信息清洗

在精确抽取后形成的地震网站信息基础语料库中，会产生许多地震重复文本信息和垃圾文本信息，而地震网站信息基础语料库是后期地震分析的基石，其准确性会影响后期研究结果的可靠性。基于重复文本去重和过滤地震无关文本（垃圾文本）的原则，需要对地震网站信息语料库进行信息清洗处理。

在去除冗余数据方面，首先对地震网站信息基础语料库进行数据预处理，过滤特殊字符和大量重复字符，如“地震”、“九寨沟”、“震级”、“@：”、“【】”等，并删除多余空白。其次调用 Python 的 jieba 库函数，完成文本数据的分词处理，并结合现有的“中文停用词表”、“哈工大停用词表”、“百度停用词表”、“四川大学机器智能实验室停用词库”等 4 个词库共 3843 个停用词，去除文本中的空格、符号、乱码及中、英文常用的停用词。最后使用 Simhash 算法去除重复数据，Simhash 算法是当前公认去除重复数据的算法，该算法是一种局部敏感哈希算法，优势在于处理速度快、结果准确度高。

Simhash 算法的具体实现为，对于 2 个给定的变量 x、y，hash 函数 H 总是满足

$$\Pr_{H \in F}(H(x) = H(y)) = Sim(x,y) \tag{1}$$

其中，Pr 表示 $H(x)=H(y)$ 的可能性，$Sim(x,y) \in [0,1]$ 为相似度函数。

该算法可以将高维的地震文本数据进行概率性的降维，并映射成为位数少且固定的指纹，随后再对指纹进行相似度计算、比较，其结果可以反映地震文本数据间的相似度。本文使用海明距离算法进行地震文本的相似度计算（张彦波，2018），假设有 2 个等长（长度为 K）的地震数据信息：A =（九，寨，沟，地，震）、B =（阿，左，旗，地，震），二者的海明距离为

$$\text{Distance} = \sum_{K=1}^{K} S_K \tag{2}$$

其中，S_K 表示为

$$S_K = \begin{cases} 0, & a_K = b_K \\ 1, & a_K \neq b_K \end{cases} \tag{3}$$

式中，a_K 和 b_K 分别代表 A、B 语句中的字符。因此，文档 A 和文档 B 的海明距离为 3。

对于垃圾文本过滤，本文采用 Word2vec 方法中的 CBOW 模型构造词嵌入文本特征矩阵，将各词向量相加作为文本的向量，进行文本特征提取（图 4）。

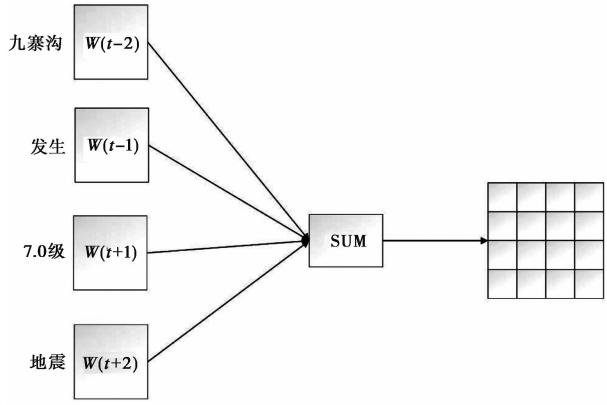

图 4 CBOW 词嵌入模型

采用 BP 神经网络结合地震文本特征矩阵，构建地震信息垃圾文本过滤模型，首先使用 BP 神经网络分类器处理地震文本信息，再对文本进行分类。BP 神经网络主体由输入层、隐藏层和输出层组成，各层之间采用权值为 W 的连接线连接，结合 sigmoid 激活函数，分类并检测出垃圾文本。

BP 神经网络分类器的基本原理为，输入数据 X_i 通过隐藏层作用于输出层，经过非线性变换输出的值包括输入向量 X 和期望输出值 t、输出值 Y 与期望输出值 t 之间的偏差，通过调整各层的权重值 W_{ij}（输入层—隐藏层）、W_{jk}（隐藏层—输出层）以及阈值，经过反复学习训练确定误差最小的权值和阈值。

隐藏层输出模型表示为

$$O_j = f\left(\sum W_{ij} \times X_i - \theta_j\right) \tag{4}$$

输出层输出模型表示为

$$Y_j = f\left(\sum T_{jk} \times O_j - \theta_k\right) \tag{5}$$

其中，f 表示非线性作用函数；θ 表示神经单元阈值。

激活函数本文用到（0,1）内连续取值的 sigmoid 函数，即

$$f(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

爬取网页数量最多的前 10 个震例 URL 元数据与经过去重过滤后的 URL 数量的对比，如

表 2 所示，去重过滤后的地震网站信息基础语料库为下一步的数据规范化处理奠定了基础。

表 2　　URL 数量对比

时间（年-月-日）	震中	震级	URL 元数据量	去重过滤后的 URL 数量
2017-08-08	四川九寨沟	7.0	99	70
2015-11-14	东海海域	7.2	108	70
2016-06-26	吉尔吉斯斯坦	6.7	102	60
2012-08-12	新疆于田	6.2	109	70
2015-04-15	内蒙古阿左旗	5.8	102	60
2017-05-11	新疆塔什库尔干	5.5	99	60
2013-08-28	云南德钦	5.1	97	65
2018-08-03	青海治多	5.1	128	64
2014-05-30	云南盈江	5.1	98	60
2017-01-28	四川筠连	4.9	120	80

2.3　地震信息抽取

通过 Simhash 算法已经删除掉大量重复信息，但网页中通常还存在一些与“人员伤亡、地质条件、次生灾害”等无关的干扰信息。地震文本内的信息通常为非结构化的数据，因此本文采用了基于词频与逆文档频率的关键字提取算法（Term Frequency-Inverse Document Frequency，TF-IDF）对干扰信息进行清除。结合信息论中的信息熵和相对熵 2 个概念，同时引入词语位置权重因子，采用基于 TF-IDF 的改进算法对地震文本信息中的地形地貌、人口信息与伤亡人数进行抽取。

使用 TF-IDF 算法对地震文本信息进行计算，首先需计算地震文本信息中每个单词的出现频率。假设所有地震文本信息中共有 words 个单词，其中某个单词的出现次数为 counter，那么该单词的词频 TF 的计算公式为

$$\mathrm{TF} = \frac{\mathrm{counter}}{\mathrm{words}} \tag{7}$$

其次，需要计算地震文本信息的逆文档频率（IDF）。假设共有 docs 条地震文本信息，其中包含某个词的地震文本信息条数为 have_docs，那么该单词的逆文档频率 IDF 的计算公式为

$$\mathrm{IDF} = \lg\left(\frac{\mathrm{docs}}{\mathrm{have\_docs} + 1}\right) \tag{8}$$

最后，计算出 TF-IDF 的整体词频值，计算公式为

$$\mathrm{TF - IDF} = \mathrm{TF} \times \mathrm{IDF} \tag{9}$$

根据上述公式可以发现，如果一个词越常见，那么分母就会越大，逆文档频率就越小，对 TF-IDF 频率的影响则越小。

根据历史灾害地震，选取爬取到的 337 例地震文本信息中的 9731 个字作为 TF-IDF 算法的训练样本，对地震文本形信息中的地形地貌、人口信息与伤亡人数进行抽取。

3 结果分析

针对 2010—2019 年 360 个历史灾害地震，共搜索得到 29565 个页面信息，其中有 337 个地震可以爬取出有效的文本数据，可通过 Web 信息自动抽取方式抽取出文字内容的网页有 9731 个（图 5），经文本数据清洗处理后，实际共解析得到 2480567 字，有效震例和文本数据的总量较为可观。

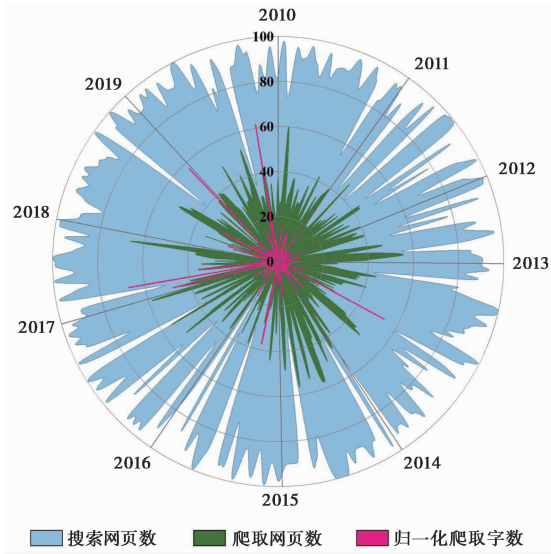

图 5 2010—2019 年灾害地震网页搜索和文字内容提取情况

当以 100 个页面为限时，平均每个震例通过百度搜索得到的网页数为 82.125 个，可以解析出 7360.7 个有效文字，为进一步文本分析和数据挖掘工作提供了丰富的数据基础。

以“发震时间、震中位置、震级”为关键词进行百度搜索，当“震中位置”为空值时，搜索引擎会采用模糊搜索策略（模糊搜索策略指当搜索关键词不存在时，扩大搜索范围），搜索内容的精确性存在一定偏差，因此不作为进一步分析的样本。

通过计算爬取网页个数、爬取文字字数与地震发生时间、震中位置和震级之间相关系数（表 3），可以看出通过互联网爬取的地震信息与目标地震的发震时间和震级相关性较低，但结合爬取的网页信息散点图（图 6）分析，爬取网页的信息量与震中所在省份紧密相关，因此可以通过该手段较稳定地获取地震相关信息。

表 3 **地震参数与爬取效果的相关系数**

参数	爬取网页个数	爬取文字数
地震年份	−0.077	0.114
震级	−0.200	0.024
震中位置	−0.055	0.003

注：负号表示负相关。

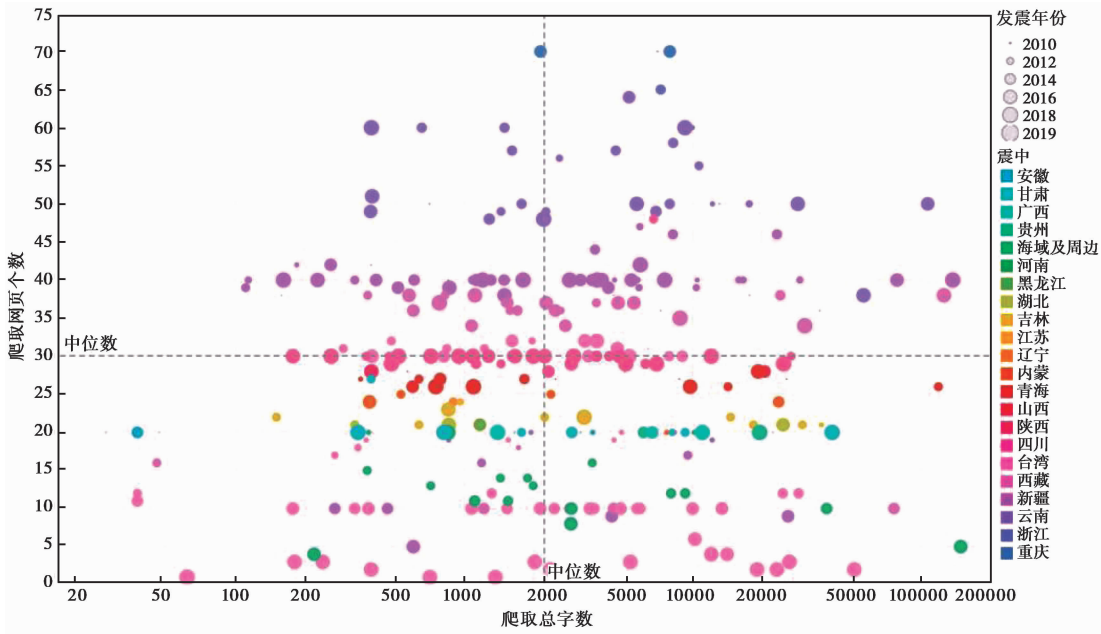

图 6　爬取网页信息字数散点图

同时，基于信息论的改进 TF-IDF 信息抽取算法对于地震文本信息的抽取有可靠的精度，能够成功对地震文本信息中的地质地貌、人口信息与人员伤亡信息进行提取（表 4）。

表 4　　地震文本信息抽取结果（以 2017 年九寨沟 7.0 级地震为例）

地震	地震文本信息	伤亡人数	地形地貌	人口信息	次生灾害
九寨沟地震	九寨沟地震发生于 2017 年 8 月 8 日 21 时 19 分 46 秒。四川省北部阿坝州九寨沟县发生 7.0 级地震，震中位于九寨沟核心景区西部 5km 处比芒村，震中 5.1km 范围平均海拔约 3827m，属高山深谷碳酸盐堰塞地貌。震中东距九寨沟县城永乐镇 39km、南距松潘县 66km、东北距舟曲县 83km、东南距文县 85km、西北距若尔盖县 90km，东偏北距陇南市 105km，南距成都市 285km。截至 2017 年 8 月 13 日 20 时，地震造成 25 人死亡（其中 24 名遇难者身份已确认），525 人受伤，6 人失联，176492 人（含游客）受灾，73671 间房屋不同程度受损（其中倒塌 76 间）。据移动人口大数据分析，震中 20km 范围内人口数约 2.1 万，50km 范围内约 6.3 万，100km 范围内约 30 万。震区属于高山峡谷区，地震引发的次生地质灾害较为严重	25 人死亡，525 人受伤	平均海拔约 3827m，属高山深谷碳酸盐堰塞地貌	震中 20km 范围内人口数约 2.1 万，50km 范围内约 6.3 万，100km 范围内约 30 万	地震引发的次生地质灾害较为严重，导致人员伤亡和部分道路交通中断

4　结论

本文提出了一种基于百度搜索引擎、以地震信息作为关键词的地震信息获取与并行的

网页解析算法,采用 Simhash 算法去重、word2vec 结合 BP 神经网络过滤垃圾文本和 TF-IDF 算法对获取到的地震信息文本进行处理与分析。实验表明,该算法能够准确且有效地提取互联网上存在的“人员伤亡、地质条件、次生灾害”等地震文本信息,并进行统计分析,生成地震文本信息的关键词以及历史灾害地震网络灾情信息数据库。但该方法在数据抽取的速度、地震应急灾情网络的形成速度等方面仍有欠缺,对此还需要进行深入的研究。

参考文献

封化民,刘飚,刘艳敏,等,2005. 含有位置坐标树的 Web 页面分析和内容提取框架. 清华大学学报(自然科学版),**45**(增刊 I):1767~1771.

庞晓克,聂高众,张昕,等,2019. 基于手机位置数据的地震灾情指标选择. 中国地震,**35**(1):144~157.

张恺航,徐克付,张闯,2019. 基于通配符节点话题权重的 Web 新闻抽取方法. 计算机工程,**45**(4):275~280.

张开敏,2014. 一种并行的网页解析算法. 小型微型计算机系统,**35**(2):193~198.

张儒清,郭岩,刘悦,等,2017. 任意网页的主题信息抽取研究. 中文信息学报,**31**(5):127~137.

张彦波,2018. 基于改进 Levenshtein 距离算法的轮廓匹配技术及实现,硕士学位论文,郑州:河南大学.

A Method of Accurate Extraction and Analysis of Web Data on Historical Disaster Earthquakes

Wen Xintao[1)] Zheng Tongyan[1)] Wang Zhonghao[2)] Li Huayue[1)] Li Chenxi[2)] Lü Wenchao[2)]

1) China Earthquake Networks Center, Beijing 100045, China

2) Institute of Disaster Prevention, Sanhe 065201, Hebei, China

Abstract Taking the earthquake catalogs from the Internet information of earthquakes between 2010 to 2019 in mainland China as an example, we propose an information acquisition technique based on Baidu search engine, and generate a set of URL generation rules with “time, place name, magnitude” as keywords. The first 100 sites retrieved by Baidu by using this technique are used to build a basic corpus of earthquake information and to form a method for acquiring Internet disaster information on earthquakes. The existing deactivation thesaurus is used to eliminate useless information, and then to conduct preliminary cleaning of the crawled information. The further digging into the implied information is performed in order to explore disaster correlations, and to establish a basis for rapid acquisition of Internet disaster information after earthquakes.

Keywords: **Disaster earthquake; Web information extraction; Disaster information acquisition; Data analysis**