IRDR Working Paper Series
Integrated Research on Disaster Risk

Extraction and Analysis of Earthquake Events Information based on Web Text

HAN Xuehua
WANG Juanle
YUAN Yuelei

IRDR
Integrated Research on Disaster Risk

doi: 10.24948/2019.06

About the Series

This Working Paper Series is a new publication of Integrated Research on Disaster Risk (IRDR), following the decision of the IRDR Scientific Committee in April 2019 to act to ‘Expand IRDR Network and Scientific Output’ (No. 5 of the IRDR Action Plan 2018-2020).

IRDR is an international scientific programme under co-sponsorship of the International Science Council (ISC) and United Nations Office for Disaster Risk Reduction (UNISDR) and with support from China Association for Science and Technology (CAST) and Chinese Academy of Sciences (CAS). Started in 2010, the Programme has been pioneering in the promoting international and interdisciplinary studies on DRR and has made its contributions through scientific publication and policy papers as well as dialogue toward shaping international agenda in the understanding disaster risks, bridging science and policy gaps and promoting knowledge for actions, all required in the Sendai Framework for Disaster Risk Reduction 2015-2030 (SFDRR) and its top priorities. Over time, the scientific agenda of IRDR has attracted many international renowned expertise and institutions. IRDR community is now, institutionally speaking, characterized by its strong Scientific Committee and six thematic working groups, thirteen IRDR national committees (IRDR NCs) and one regional committee (IRDR RC), sixteen international centres of excellence (IRDR ICoEs), a group of some one hundred fifty Young Scientists (IRDR YS) and a broad partnership with national, regional and international institutions working for SFDRR.

This Working Paper Series is thus specially made to facilitate the dissemination of the work of IRDR NCs, ICoEs, YS and institutions and individual experts that IRDR considers relevant to its mission and research agenda, and of important values for much broader range of audience working in DRR domains. As one will notice, all working papers in this series has anchored their relevance and contributions of their work toward SFDRR, IRDR, SDGs and Paris Agreement on climate change. It is the hope of the authors of the working papers and IRDR that this working paper series will not only bring new knowledge, experience and information toward disaster risk reduction, but also helped build better coherence of DRR with the mainstream agenda of UN today toward inclusive, resilient and sustainable human societies.

Team of IRDR-IPO

IRDR
Integrated Research on Disaster Risk

Extraction and Analysis of Earthquake Events Information based on Web Text

HAN Xuehua[1,2]
WANG Juanle[1,3*]
YUAN Yuelei[1,3]

[1] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[2] University of Chinese Academy of Sciences, Beijing 100049
[3] Disaster Risk Reduction Knowledge Service, International Knowledge Center for Engineering Sciences and Technology (IKCEST), Beijing 100088

*Corresponding Author: wangjl@igsnrr.ac.cn

Published in 2019 by Integrated Research on Disaster Risk (IRDR), an interdisciplinary body of the International Science Council (ISC).

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of IRDR or the ISC concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of IRDR or the ISC and do not commit the Organizations.

Cover photo: http://www.sohu.com/a/202171510_781810

Design: Fang LIAN

Edit: Qunli HAN, Lucy JONES

Printed by IRDR

Printed in Beijing, China

Citation of this publication: Han, X., Wang, J., & Yuan, Y. (2019). Extraction and Analysis of Earthquake Events Information based on Web Text. *IRDR Working Paper Series,* 17 pages. doi: 10.24948/2019.06

Content

Abstract of this Working Paper

In the era of big data, the extraction of disaster event information from huge quantities of network data is important in facilitating research on disaster prevention and reduction. The Sendai Framework for Disaster Risk Reduction 2015–2030 (SFDRR) emphasized the promotion of the collection, analysis, management, and use of relevant data and the enhancement of disaster monitoring, risk assessment, and service capacity to strengthen the utilization of big data, social media, and the Internet. Extracting and visualizing the spatiotemporal and attribute information of disaster events from web text is not only an exploration of the goal of the SFDRR but also an innovative application in disaster emergency response and disaster risk reduction. Based on the Disaster Risk Reduction Knowledge Service (http://drr.ikcest.org/) subordinate to the International Knowledge Center for Engineering Sciences and Technology (IKCEST), this study explores the automated spatiotemporal and attribute information extraction of earthquake events from web news reports and online official reports in China from 2015 to 2017. Combined with a web crawler, natural language processing, and geographic information science, a set of rules is created to automatically extract spatiotemporal and attribute information for earthquake events from web text, and geocoding is applied to map the earthquake events. The spatial and temporal distribution characteristics of earthquake events extracted from the web texts are then analyzed. This study compares the differences between web news reports and online official reports in terms of the quantity and spatiotemporal distribution of the earthquake events.

Keywords

Earthquake, web text, information extraction, spatiotemporal distribution, China

Indications of contributions to IRDR Science Plan and UN Agendas

IRDR Sub-objectives	1.1
SFDRR targets	
SDGs and/or Climate Goals	
S/T Roadmap actions	

1. How does this study contribute to IRDR research objectives?

Article 1.1 of IRDR research objectives point out that "identifying hazards and vulnerabilities leading to risks". This study automatically analyzed and extracted the spatiotemporal and attribute information of earthquake events from news reports and official reports, analyzes the distribution characteristics of earthquake events in China from 2015–2017 and compares the differences between the web news reports and online official reports. The results identify the hot spots of earthquake in China from 2015–2017.

2. How does this study contribute to SFDRR targets?

Article 24(f) and 25(c) of SFDRR emphasized the promotion of the collection, analysis, management, and use of relevant data and the enhancement of disaster monitoring, risk assessment, and service capacity to strengthen the utilization of big data, social media, and the Internet. Combining geographic information systems (GIS) and network data mining technology, this study explored disaster information acquisition for disaster risk reduction from the Internet, and accomplished the automated spatiotemporal and attribute information extraction of earthquake events from web news reports and online official reports. This study is not only an exploration of the goal of the SFDRR but also an innovative application in disaster emergency response and disaster risk reduction

3. Main recommendations to DRR policy if not yet highlighted in the main texts

a) Strengthening the open access of web data resources. For example, the web data platform（e.g. social media platforms, news media websites） can set emergency data interface in the disaster period.

b) Improving the social media information processing and analysis ability of disaster risk reduction agencies. For example, more big data analysis training workshop for disaster risk reduction should be held.

Main Text

1. Introduction

In recent years, the Internet has become the foremost means of disseminating information and knowledge. By combing network big data, extracting and visualizing the spatiotemporal and attribute information of disaster events from web text is a growing area of disaster emergency response and disaster risk reduction. The Sendai Framework for Disaster Risk Reduction 2015–2030 (SFDRR) has a goal to "promote real-time access to reliable data, make use of space and in situ information, including geographic information systems (GIS), and use information and communications technology innovations to enhance measurement tools and the collection, analysis and dissemination of data." (United Nations 2015; Fan 2015). More specifically, this report emphasized the promotion of the collection, analysis, management, and use of relevant data, and the enhancement of disaster monitoring, risk assessment, and service capacity in order to strengthen the utilization of big data, social media, and the Internet. In comparison with traditional information, disaster information based on web text is dynamic, heterogeneous, and massive; has space-time characteristics; and accesses multiple sources (Lu et al. 2017; Li et al. 2015; Lyu et al. 2016). Owing to the sheer volume, high velocity, and varied structure of web texts, one significant challenge that arises in this context is how to deal with this 'unstructured data' to separate the 'wheat from the chaff': how to automatically pick disaster spatiotemporal and attribute information out of massive web texts.

Within the SFDRR and network big data background, extracting and analyzing disaster event information from web text, tracking dynamic change patterns and trends of such events over space and time, and building an application for the analysis and acquisition of Internet disaster information are not only explorations of the goal of the SFDRR mentioned above but also innovative applications in disaster emergency response and disaster risk reduction. With the development and popularization of the Internet, web news media with the characteristics of authority, objectivity, reality, and comprehensiveness has become important access to information for people and is also one of the important sources of disaster information. An increasing number of studies have extracted and analyzed of disaster events information from web text, most of which focus on innovative methods for extracting disaster information. Valero et al. (2010) described a system based on machine learning methods to improve the acquisition of disaster information from online news reports. Zhang et al. (2013) combined a rules model and a statistical model to extract and visualize spatiotemporal information on earthquake events in web news texts. Wang et al. (2015) applied ontology to extract spatiotemporal and semantic information on typhoons from web news reports. Liu et al. (2015) studied the extraction and visualization of spatial-temporal and attribute information on landslide disasters from web news reports by using a rules and statistical approach. Herford et al. (2015) presents an approach to enhance the identification of relevant messages from social media that relies upon the relations between georeferenced social media messages and geographic features of flood phenomena. Song et al. (2017) constructed a web information extraction algorithm that supported dynamic convergence, according to the characteristics of the timeliness of disaster information. Yang et al. (2013) proposed a method for spatial information extraction from earthquake event news, based on geographic names and semantic technology. Stewart et al. (2010) automatically extracted spatial and temporal references from web text, and represented the spatiotemporal characterizations of events in a dynamic mapping environment. Fan et al. (2018) created an extraction rule based on syntax analysis to identify earthquake events and extract information from web news text. Li et al. (2017) introduces a novel approach to mapping the flood in near-real time by leveraging twitter data in geospatial processes. Wang et al. (2016) analyzed the wildfire-related Twitter activities to gain insights into the usefulness of social media data in revealing situational awareness. Those studies seek to identify useful information from web text and establish solid foundation for disaster information extraction method. However, most of the studies ignored the influence of different web text on the extracted results. Understanding the characteristics of disaster information between varied web sources makes the extraction of disaster information more targeted.

Therefore, using the 2015–2017 earthquakes in China as the case, this study realizes the automatic extraction of spatiotemporal and attribute information of earthquake events from web news reports and online official reports, and explore the differences between the web news reports and online official reports in terms of the quantity and spatiotemporal distribution of earthquake events. The methods used for data acquisition and earthquake event information extraction are introduced in the following section. Finally, a summary is provided, as well as descriptions of the limitations of this study and future work that should be undertaken.

2. Earthquake Event Extraction and Analysis Process

Sina News is an online news media source with the largest user group in China. The China National Commission for Disaster Reduction (NCDR-China) is one of the leading institutions that provides support to the government in addressing disaster-related issues by focusing on the entire cycle of disaster management. Therefore, the Sina News website （https://www.sina.com.cn/） and NCDR-China website are selected as data sources.

The study period spans from January 1, 2015, to November 31, 2017. The experiment is divided into three main components, which are shown in Figure 1.

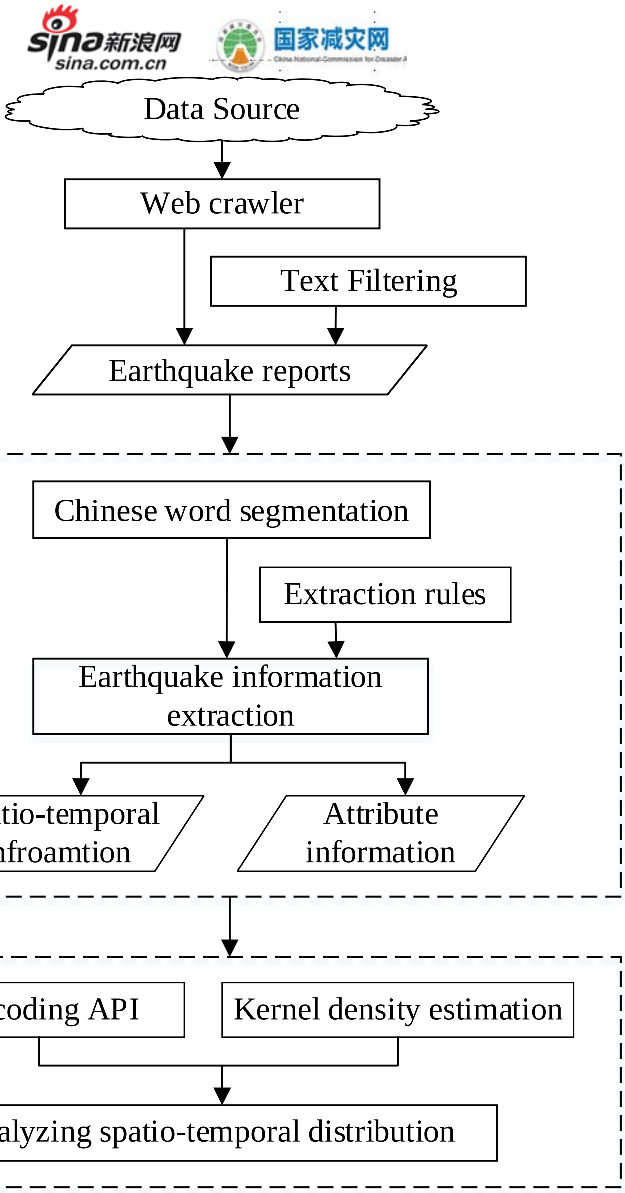
sina新浪网
sina.com.cn
国家减灾网
Data Source
Web crawler
Text Filtering
Earthquake reports
Chinese word segmentation
Extraction rules
Earthquake information extraction
Spatio-temporal infroamtion
Attribute information
Geocoding API
Kernel density estimation
Analyzing spatio-temporal distribution

Figure 1. The processes of the experiment

(1) Data acquisition. A web crawler is developed to collect text reports related to earthquakes from the above source websites.

(2) Earthquake event information extraction. Supported by Chinese Word Segmentation and Regular Expression, a set of extraction rules relevant to earthquake events is built to achieve the spatiotemporal and attribute information extraction of earthquake events.

(3) Analysis of spatial and temporal distribution characteristics of earthquake events. We analyze the spatial and temporal distribution characteristics of earthquake events in China that occurred from 2015 to 2017 by integrating the Baidu Map Geocoding API[1], GIS methods, and kernel density methods.

[1] http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding

2.1 Data Acquisition and Preprocessing

2.1.1 Data Collection

(1) Gathering news reports related to earthquake from the SINA.com. The website provides a search page for news reports via keywords of news content and title. The search pages return a search result list that meets the search criteria. Using a web crawler, earthquake news reports can be collected, including title, time, and text. We used all names of provinces in China, and "earthquake" ("地震" in Chinese) and "occur" ("发生" in Chinese) as the search keywords. We gathered 2,963 news reports relevant to earthquakes from January 1, 2015 to November 31, 2017 in total.

(2) Gathering official reports related to earthquakes from the NCDR-China website. The "latest disaster" ("最新灾情" in Chinese) column of the website is selected as seed pages for the crawler. Using the Beautiful Soup Library in Python[2], we parsed the pages and collected 1,062 official reports about natural disasters from January 1, 2015, to November 31, 2017. Of these, 219 were related to earthquakes.

2.1.2 Text Filtering

There are many repetitive and similar texts in the news reports from SINA.com According to the timing of a news release, related news reports usually emerge in large numbers within 2–3 days after a disaster. Therefore, we sort the news reports according to the release time. Then, the event location text is extracted from the news title and contrasted one by one. If the location and magnitude from two news report are same, they will be considered to be duplicated, and the latest news text is retained. Since they have less repetitive and similar texts, the official reports needn't to be filtered. Finally, we obtained 984 news reports and 219 official reports, stored in an SQLite[3] database.

2.2 Earthquake event information extraction

2.2.1 Earthquake event information rules

1) Temporal Extraction Rules

There are various temporal expressions in texts related to earthquakes, such as absolute time and relative time. For example, relative time refers to "tomorrow" ("明天" in Chinese), "the day after tomorrow" ("后天"in Chinese), "today" ("今天" in Chinese), etc. Absolute time refers to "January 1, 2015" ("2015 年 1 月 1 日" in Chinese), etc. According to the characteristics of temporal expressions of earthquake events, a temporal expression dictionary is constructed and shown in Table 1. Then, a set of temporal extraction rules is designed for each type of temporal expression. Examples of rules are listed in Table 2.

[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[3] https://www.sqlite.org/index.html

Table 1 Temporal expression dictionary

temporal type	Expressions	Sample
Absolute time	xx 年 x 月 x 日 x 时 x 分	“2018 年 5 月 21 日 21 时 27 分”
	X 月 x 日 x 时 x 分	“05 月 19 日 11 时 09 分”
	X 日 x 时 x 分	“1 日 21 时 10 分”
	Year/Month/Day/Hour/Minute	“At 21:27 on May 21, 2018.”
	Month/Day/Hour/Minute	“At 11:09 on May 19”
	Day/Hour/Minute	“At 21:10 on first”
Relative time	“今日/ today,” “下午/afternoon” ,”早上/ morning,” “今天/today”	“今日下午 1 时 55 分”， “早上 8 点 48 分 22 秒”， ”今天凌晨 2 时 39 分” “1:55 this afternoon,” “8:48:22 in the morning,” “2:39 this morning”

Table 2 temporal extraction rules

temporal type	Extraction Rules	Sample
Absolute time	\d{4}年\d{1,2}月\d{1,2}日\d{1,2}时\d{1,2}分 \d{1,2}月\d{1,2}日\d{1,2}时\d{1,2}分 \d{1,2}日\d{1,2}时\d{1,2}分	xx 年 x 月 x 日 x 时 x 分 X 月 x 日 x 时 x 分 X 日 x 时 x 分 Year/Month/Day/Hour/Minute Month/Day/Hour/Minute Day/Hour/Minute
Relative time	今日/t 下午/t 早上/t 今天/t	今日，下午，早上，今天 Today, afternoon, morning, today

2) Location trigger dictionary

Sorts of locations mentioned in the earthquake texts include: ① Latitude and longitude. Most earthquake texts contain latitude and longitude coordinates of the earthquake location. ② Toponym or address. A few of the earthquake texts describe locations with toponyms or addresses, e.g., “A 3.3 magnitude earthquake occurred in Qingchuan County, Guangyuan, Sichuan.” There are several Named Entity Recognition (NER) software packages that can be used to identify locations in text. The highly regarded NER software includes the NLPIR developed by the Chinese Academy of Science (http://ictclas.nlpir.org/), the Language Technology Platform developed by the Harbin Institute of Technology (https://www.ltp-cloud.com/), etc. However, there are a variety of locations in earthquake texts, many of which are not related to earthquakes. For example:

“***中国***网 3 月 30 日讯，据***中国***地震台网消息，***北京***时间 3 月 30 日 9 时 28 分，***四川省内江市威远县***（***北纬 29.52 度，东经 104.56 度***）发生 3.0 级地震，震源深度 11 千米。”

“*China.org.cn*, March 30, a magnitude 3.0 earthquake struck *Weiyuan County, Neijiang, Sichuan (29.52º N, 104.56º E*), at 9:28 a.m. March 30 (*Beijing* time), according to *China* Earthquake Networks Center.”

Therefore, in order to extract the correct spatial information of earthquake events, a location trigger dictionary is built, as shown in Table 3. Using the trigger word information, we can accurately detect place names or the latitude and longitude of earthquake events.

Table 3 Location trigger dictionary

Location type	Trigger words	Sample
Locations related to earthquake	“纬度/ Latitude,” “北纬/ north latitude” ,”经度/ Longitude,” “东经/ east longitude,” “发生/occur”	中新社 2 月 9 日电(记者董飞)，***中国***地震台网测定，9 日 19 时许，***河南省南阳市淅川县***发生 4.3 级地震。暂无人员伤亡报告。***河南省***地震局当晚通报称，根据相关预案规定，启动Ⅱ级应急响应。震中位于该县马蹬镇(***东经 111.60 度，北纬 32.80 度***)，震源深度 10 千米。 “***China*** news service, February 9 (reporter Dong Fei), according to **China** Earthquake Network Center, at 19:00 p.m. February 9, a 4.3 magnitude earthquake occurred in ***Xichuan County, Nanyang City, Henan province***. There are no reports of casualties for the time being. The **Henan** Seismological Bureau said on the same night that according to the relevant pre-arranged plans, the second level emergency response should be launched. The epicenter is located Madeng town(**111.60 degrees east longitude, 32.80 degrees north latitude**) in the county, with a focal depth of 10 km.”
Locations unrelated to earthquake	“地震局/ Seismological Bureau,” “地震台网/ Earthquake Networks Center,” “时间/time,” “台网/ Networks Center”	

3) Attribute trigger dictionary

Attribute information of earthquake events includes the epicenter, magnitude, casualties, housing losses, economic losses, and other information. In Chinese, there are some specific descriptions for disaster attribute information. By analyzing the descriptive characteristics of earthquake attribute information in texts, we designed an attribute trigger dictionary for earthquakes, as shown in Table 4.

Table 4 Attribute trigger dictionary

Attribute type	Trigger words
Magnitude	“级/ magnitude unit,” “地震/earthquake,” “震级/ magnitude,” “里氏/ Richter,” ms
Focal depth	“震源/focal depth,” “深度/depth,” “深/deep,” “公里/km,” “千米/km,” km
Casualties	“名/Casualty unit,” “死亡/death,” “遇难/death,” “伤者/Wounded,” “尸体/dead body,” “死者/dead,” “丧生 dead,” “人数/number”
Epicenter	“震中/ epicenter,” “所在地/place,” “位/located,” “中心/center”

2.2.2 Information extraction

Based on the trigger dictionary and rules, we can obtain the space-time and attribute information of earthquake events by rule matching, which is divided into five steps: (1) Text preprocessing. We use the NLPIR of the Chinese Academy of Science for keyword filtering, text segmentation, and information annotation. The process is in the following manner: First, we break down a block of text into sentences, and filter the sentences according to some keywords of earthquakes. Then, word segmentation and word annotation are conducted with the NLPIR implemented in Python. (2) Temporal information extraction. Using the temporal extraction rules, we extract the time of an earthquake event by the forward maximum matching method (FMM) (Wang, 2011). Longer rules have priority. (3) Spatial information extraction. Based on the location trigger dictionary and the annotated text, the place names related to earthquakes are extracted, and the adjacent place names are combined into a complete spatial expression. (4) Attribute information extraction. Based on the attribute trigger dictionary and regular expressions, the earthquake attribute information is extracted by the FMM. (5) Result filtering. The result is checked manually and some incorrect records are deleted. Finally, an earthquake event information database is built. The extracted metadata of earthquake events included the time, location, magnitude, latitude, and longitude. There are 854 earthquake events from news reports and 134 earthquake events from official reports, with a spatial scope of the entire territory of China.

2.2.3 Geocoding

In this study, for earthquake events without latitude and longitude coordinates, we use the Baidu Map Geocoding API to geocode the location information. First, according to the code format of the Geocoding API, we standardize the extracted location and generate a valid URL. Then, we send an HTTP request to the Geocoding API for returned JSON data. By parsing the “latitude” and “longitude” parameters from the JSON data, the geographic coordinates are assigned to each extracted location.

3. Results

According to China’s national preparatory plan for earthquake emergencies (2012 Revised Edition), earthquakes can be divided into four types: general earthquake ($4.0 \leq Ms < 5.0$), larger earthquake ($5.0 \leq Ms < 6.0$), major earthquake ($6.0 \leq Ms < 7.0$), and massive earthquake ($Ms \geq 7.0$).

This earthquake classification standard was applied in our study. The levels of earthquake events from the news reports and official reports are compared in Figure 2. For earthquake events with $Ms \leq 4$, there were 594 earthquake events from news reports and 2 earthquake events from official reports. For general earthquakes ($4.0 \leq Ms < 5.0$), there were 190 earthquake events from news reports and 104 earthquake events from official reports. For larger earthquakes ($5.0 \leq Ms < 6.0$), there were 51 earthquake events from news reports and 20 earthquake events from official reports. For major earthquakes ($6.0 \leq Ms < 7.0$), there were 17 earthquake events from news reports and 7 earthquake events from official reports. The number of massive earthquakes from both types of reports was only 1.

We find that the news reports are more comprehensive and cover all levels of magnitude. However, the official reports are mainly focused on earthquakes with *Ms* ⩾ 4 or earthquakes that may cause damage, suggesting that the official reports focus on disasters that cause damage to society or the public. The total number of earthquake events from official reports is much lower than that from news reports, but there is little difference between the number of earthquake events with *Ms* ⩾ 4.

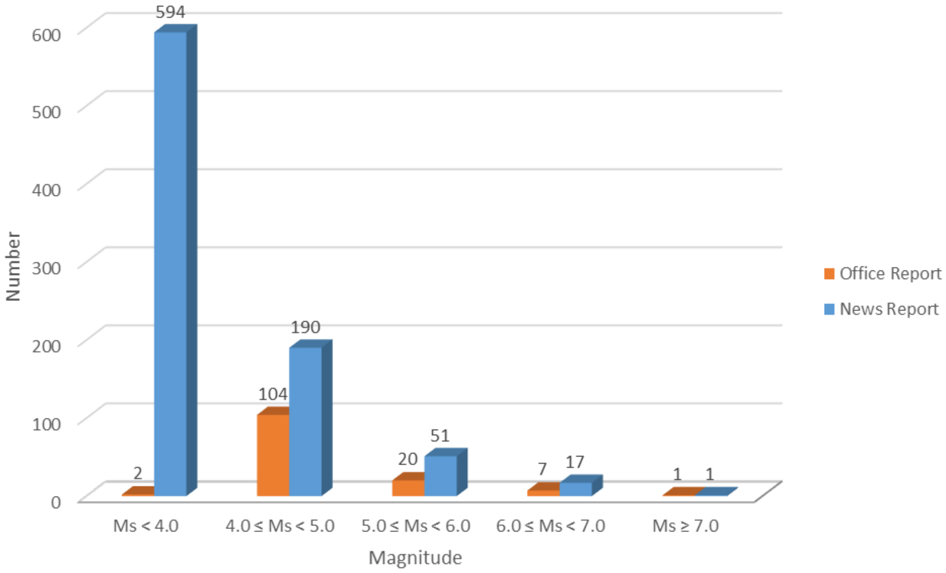

Figure 2. Levels of earthquake events from news reports and official reports from January 1, 2015, to November 31, 2017

Both datasets (the earthquake events from news reports and earthquake events from official reports) were counted for each province in China. As shown in Table 5 and Figure 3a, earthquake events from news reports are widely distributed across 30 provinces in China except for Shanghai, Hainan, Hong Kong, and Macao. The province with the highest number of earthquake events from news reports is Sichuan province, accounting for 17%. The number of earthquakes in Xinjiang, Xizang, Taiwan, Yunnan, and Qinghai decreased successively, accounting for more than 5%. As shown in Figure 3c, the provinces with higher proportions (≥5%) of earthquake events from news reports are located in China's western region and Taiwan region. North China is the second, and the South is the lowest. The earthquake events are reduced from west to east in spatial distribution.

The distribution of earthquake events from official reports in 18 provinces is listed in Table 6. Xinjiang Autonomous Region had the highest number of earthquake events, accounting for 27.61%. The number of earthquakes in Xizang, Yunnan, Sichuan, and Qinghai decreased successively, accounting for more than 5%. As shown in Figure 3d, the China western region had the higher proportion (≥5%) of earthquake events. Compared with the news reports, the earthquake events from official reports are mainly concentrated in southwest and northwest China.

Table 5. Statistics at province level of earthquake events from news reports from 2015 to 2017

Province	$Ms < 4.0$		$4.0 \leq Ms < 5.0$		$5.0 \leq Ms < 6.0$		$6.0 \leq Ms < 7.0$		$Ms \geq 7.0$		Total	
	Number	proportion /%	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)
Anhui	4	0.67	1	0.53	0	0.00	0	0.00	0	0	5	0.59
Beijing	5	0.84	0	0.00	0	0.00	0	0.00	0	0	5	0.59
Fujian	4	0.67	0	0.00	0	0.00	0	0.00	0	0	4	0.47
Gansu	18	3.03	4	2.11	0	0.00	0	0.00	0	0	22	2.58
Guangdong	14	2.36	1	0.53	0	0.00	0	0.00	0	0	15	1.76
Guangxi	3	0.51	6	3.16	1	1.96	0	0.00	0	0	10	1.17
Guizhou	6	1.01	0	0.00	1	1.96	0	0.00	0	0	7	0.82
Hebei	26	4.38	3	1.58	1	1.96	0	0.00	0	0	30	3.52
Henan	3	0.51	1	0.53	0	0.00	0	0.00	0	0	4	0.47
Heilongjiang	5	0.84	0	0.00	0	0.00	0	0.00	0	0	5	0.59
Hubei	2	0.34	3	1.58	0	0.00	0	0.00	0	0	5	0.59
Hunan	4	0.67	0	0.00	1	1.96	0	0.00	0	0	5	0.59
Jilin	6	1.01	4	2.11	0	0.00	0	0.00	0	0	10	1.17
Jiangsu	7	1.18	1	0.53	0	0.00	0	0.00	0	0	8	0.94
Jiangxi	5	0.84	0	0.00	0	0.00	0	0.00	0	0	5	0.59
Liaoning	9	1.52	3	1.58	1	1.96	0	0.00	0	0	13	1.52
Neimenggu	21	3.54	1	0.53	1	1.96	0	0.00	0	0	23	2.70
Ningxia	8	1.35	3	1.58	0	0.00	0	0.00	0	0	11	1.29
Qinghai	43	7.24	10	5.26	0	0.00	2	11.76	0	0	55	6.45
Shandong	9	1.52	3	1.58	0	0.00	0	0.00	0	0	12	1.41

Shanxi	11	1.85	3	1.58	0	0.00	0	0.00	0	0	14	1.64
Shanxi	9	1.52	0	0.00	3	5.88	0	0.00	0	0	12	1.41
Sichuan	115	19.36	25	13.16	4	7.84	0	0.00	1	100	145	17.00
Taiwan	17	2.86	45	23.68	19	37.25	6	35.29	0	0	87	10.20
Tianjin	3	0.51	0	0.00	0	0.00	0	0.00	0	0	3	0.35
Xizang	66	11.11	32	16.84	9	17.65	2	11.76	0	0	109	12.78
Xinjiang	106	17.85	21	11.05	6	11.76	7	41.18	0	0	140	16.41
Yunnan	55	9.26	17	8.95	4	7.84	0	0.00	0	0	76	8.91
Zhejiang	1	0.17	1	0.53	0	0.00	0	0.00	0	0	2	0.23
Chongqing	9	1.52	2	1.05	0	0.00	0	0.00	0	0	11	1.29
Total	594	100.00	190	100.00	51	100.00	17	100.00	1	100	853	100.00

Table 6. Statistics at province level of earthquake events from official reports from 2015 to 2017

Province	*Ms* < 4.0		4.0 ≤ *Ms* < 5.0		5.0 ≤ *Ms* < 6.0		6.0 ≤ *Ms* < 7.0		*Ms* ≥ 7.0		Total	
	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)
Anhui	0	0.00	2	1.92	0	0.00	0	0.00	0	0.00	2	1.49
Gansu	1	50.00	3	2.88	0	0.00	0	0.00	0	0.00	4	2.99
Guangxi	0	0.00	1	0.96	1	5.00	0	0.00	0	0.00	2	1.49
Hebei	0	0.00	1	0.96	0	0.00	0	0.00	0	0.00	1	0.75
Heilongjiang	0	0.00	0	0.00	0	0.00	1	14.29	0	0.00	1	0.75
Hubei	0	0.00	2	1.92	0	0.00	0	0.00	0	0.00	2	1.49
Jilin	0	0.00	2	1.92	0	0.00	0	0.00	0	0.00	2	1.49
Liaoning	0	0.00	2	1.92	0	0.00	0	0.00	0	0.00	2	1.49

Ningxia	0	0.00	1	0.96	0	0.00	0	0.00	0	0.00	1	0.75
Qinghai	0	0.00	9	8.65	1	5.00	1	14.29	0	0.00	11	8.21
Shandong	0	0.00	1	0.96	0	0.00	0	0.00	0	0.00	1	0.75
Shanxi	0	0.00	3	2.88	0	0.00	0	0.00	0	0.00	3	2.24
Sichuan	0	0.00	13	12.50	2	10.00	0	0.00	1	100.00	16	11.94
Xizang	1	50.00	19	18.27	6	30.00	2	28.57	0	0.00	28	20.90
Xinjiang	0	0.00	27	25.96	7	35.00	3	42.86	0	0.00	37	27.61
Yunnan	0	0.00	15	14.42	2	10.00	0	0.00	0	0.00	17	12.69
Zhejiang	0	0.00	1	0.96	0	0.00	0	0.00	0	0.00	1	0.75
Chongqing	0	0.00	2	1.92	1	5.00	0	0.00	0	0.00	3	2.24
Total	2	100	104	100.00	20	100.00	7	100.00	1	100.00	134	100.00

To further investigate the geographical distribution of earthquake events from web text in China, we performed a kernel density estimation using ArcGIS aiming to identify the hot spots of the earthquakes. Three parameters were used in the kernel density analysis: kernel search radius (bandwidth) for calculating the density, cell size for the output raster data and population field to weight some features more heavily than others. By repeated experiments, a kernel search radius of 500 km was used to avoid creating a map that was too smooth or too ambiguous to interpret. And a cell size of 6 km was used to show sufficient detail. The magnitude was used to equal the Population field value. The areas with high densities of earthquake events are represented in the darkest hue of red.

Figure 4a provides a glimpse of the spatial distribution, showing that the high-density areas of earthquake events from news reports are located in Beijing-Tianjin-Hebei region, Sichuan province, Xizang and Qinghai border regions, Xinjiang province, and Taiwan province. As shown in Figure 4b, the high-density areas of earthquake events from official reports are located in China’s western region, such as Sichuan and Yunnan adjacent region, Xizang and Qinghai adjacent region, and Xinjiang province. In addition, the density of earthquake events tends to decrease away from the center.

In comparing Figure 4a and b, the spatial distribution of earthquake events from news reports is wider in China, showing five high-density centers not only in China’s western region but also in the north and Taiwan province. However, the earthquake events from official reports are mainly distributed in China’s western region, showing three high-density centers. One possible explanation is that most of the earthquakes that occurred in northern China were *Ms* ≤ 4 earthquakes, and most of the earthquakes on Taiwan occurred near the sea area without causing damage from 2015–2017. Therefore, there are fewer earthquake events in the north of China and on Taiwan in official reports. To a certain extent, the news reports are more extensive and rich with information, while the official reports are less repetitive and more accurate. Combining the news reports with the official reports makes the analysis of spatial distribution of earthquakes more comprehensively.

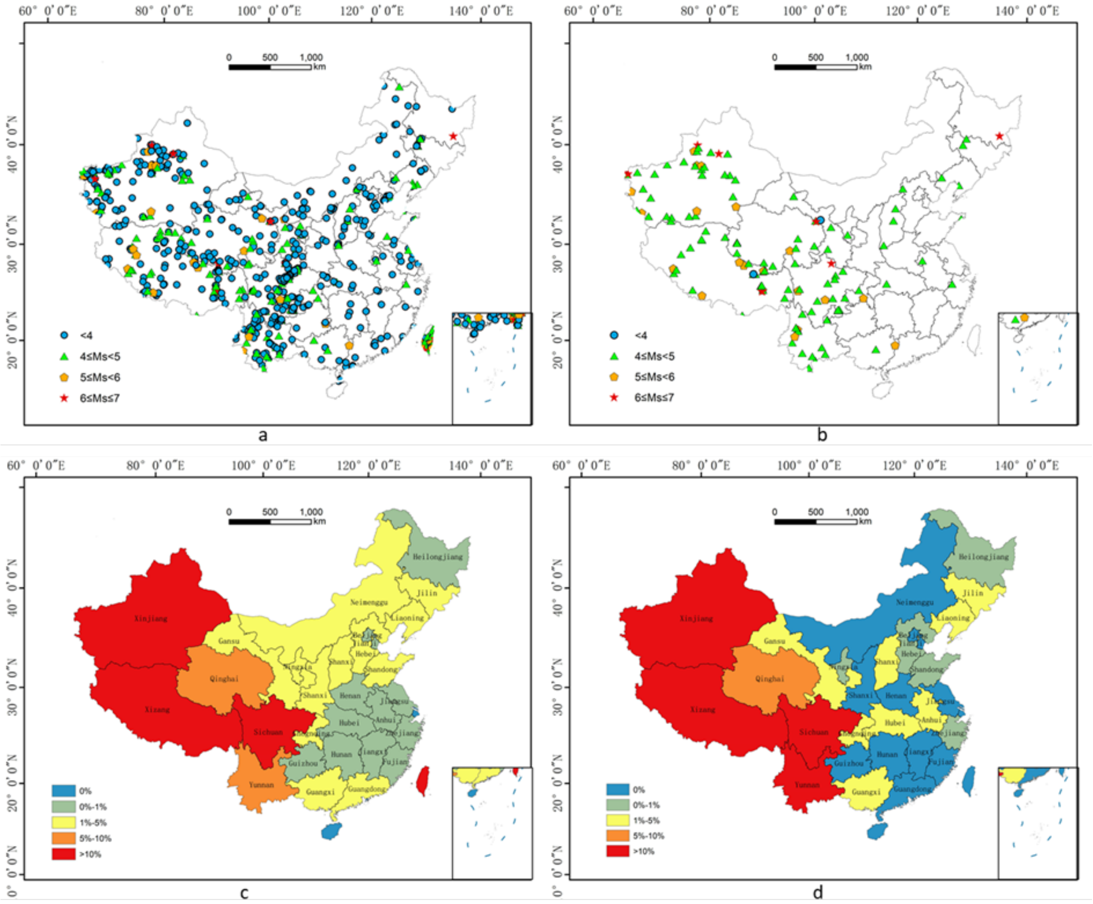

Figure 3. (a) Distribution of earthquake events from news reports from 2015 to 2017, (b) distribution of earthquake events from official reports from 2015 to 2017, (c) distribution at province level of earthquake events from news reports from 2015 to 2017, (d) distribution at province level of earthquake events from official reports from 2015 to 2017

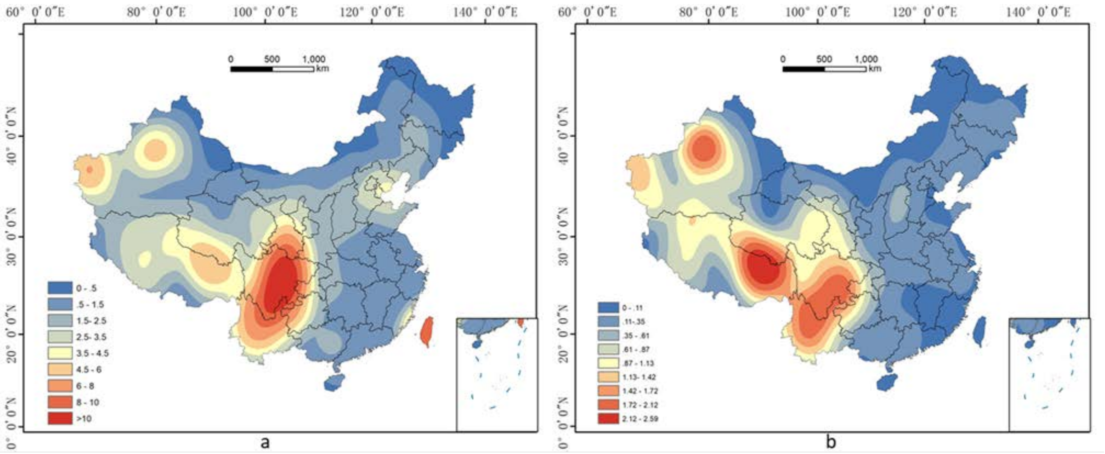

Figure 4. (a) Kernel density analysis for earthquake events in news reports, (b) kernel density analysis for earthquake events in official reports

4. Conclusion

(1) This study summarized the description characteristics of earthquake information in news reports and official reports, and built temporal extraction rules, a location trigger dictionary, and an attribute trigger dictionary for earthquake information extraction. By combining the dictionary and rules, we automatically analyzed and extracted the spatiotemporal and attribute information of earthquake events from news reports and official reports, and formed structured earthquake data. This study provides an effective solution for the acquisition of the spatiotemporal and attribute information of earthquake events in web texts, and provides a new data source for earthquake disaster research.

(2) This study analyzes the distribution characteristics of earthquake events in China from 2015–2017. By analyzing the occurrence frequency of earthquake events at the province level and performing a kernel density estimation, the conclusions are as follows: From 2015–2017, earthquakes occurred mainly in China's western region, northern region, and Taiwan region. The spatial distribution is reduced from west to east. Most earthquakes ($Ms \geq 4.0$) were distributed in China's western region, mainly in Sichuan, Xinjiang, Xizang autonomous region, Taiwan, Yunnan, and Qinghai.

(3) In comparing news reports with official reports, the earthquake coverage in news reports is more comprehensive and covers all levels of magnitude, while the earthquake coverage in official reports is mainly focused on earthquakes with $Ms \geq 4$ or earthquakes that may cause damage. The result of a kernel density estimation showed some differences that earthquake events from news reports focused on five high-density centers in China's western region and in China's north and Taiwan region, but the earthquake events from official reports were mainly distributed in China's western region with only three high-density centers. To some extent, the news reports are more extensive and richer with information, while the official reports are less repetitive and more accurate. By combining the news reports with the official reports, the features of

earthquake events or other disaster events can be discovered more comprehensively in China.

Acknowledgements: This work was supported by the International Knowledge Center for Engineering Sciences and Technology (IKCEST) under the Auspices of UNESCO.

5. References

Fan H, LI H, Du W. Web Based Extraction of Spatiotemporal Information of Earthquake Event by Semantic Technology [J]. Engineering Journal of Wuhan University, 2018(2):183-188. (in Chinses).

Fan Y D. The Future Challenge of Disaster Risk Management in China: Interpretation "Sendai Framework for Disaster Risk Reduction 2015-2030" [J]. Disaster Reduction in China, 2015(7):18-21.

Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. International Journal of Geographical Information Science, 29(4), 667-689.

Li Y, Feng L, Zhang H. Extracting Geographic Information from Web Texts: Status and Development [J]. Journal of Geo-Information Science, 2015, 17(2):127-134.

Liu, S. Y. (2015). Extracting Landslide Disaster Information from Web Pages. Southwest Jiaotong University. (in Chinses).

Li, Z., Wang, C., Emrich, C. T., & Guo, D. (2017). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. Cartography & Geographic Information Science, 1-14.

Lu F, Yu L, Qiu P Y. On Geographic Knowledge Graph [J]. Journal of Geo-Information Science, 2017, 19(6):723-734.

Lyu X F, Chen S Y. Review of Natural Disaster Network Public Opinion Information Analysis and Management [J]. Geography and Geo-Information Science, 2016, 32(4):49-56.

Stewart, K., & Wang, W. (2010). Representing dynamic phenomena based on spatiotemporal information extracted from web documents. International Conference on Geographic Information Science.

Song, J. G., Wang, Z. X., Li, Q. Y., Ma, S. L., & Lv, J. H. (2017). Internet information process oriented to the earthquake response. Journal of Beijing University of Aeronautics and Astronautics, 43(6), 1155-1164. (in Chinses).

Téllez Valero, A., Manuel, M. Y. G., & Villaseñor Pineda, L. (2009). Using Machine Learning for Extracting Information from Natural Disaster News Reports. Computación Y Sistemas, 13(1), 33-44.

United Nations. Sendai Framework for Disaster Risk Reduction 2015-2030 [J]. United Nations Office for Disaster Risk Reduction (UNISDR), 2015.

Wang R. An Improved Forward Maximum Matching Algorithm for Chinese Word Segmentation [J]. Computer Applications & Software, 2011.

Wang, W., & Stewart, K. (2015). Spatiotemporal and semantic information extraction from Web news reports about natural hazards. Computers Environment & Urban Systems, 50, 30-40.

Wang, Z., Ye, X., & Tsou, M. H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. Natural Hazards, 83(1), 523-540.

Yang, J., Hong, F., & Huaiyuan, L. I. (2013). Spatial Information Extraction of Web Seismic Event Based on Geographic Names Semantic Technology. Journal of Geomatics, 38(6), 10-13. (in Chinses).

Zhang, C. J. (2013). Interpretation of Event Spatio-temporal and Attribute Information in Chinese Text. Nanjing Normal University. (in Chinses).