International Journal of Geo-Information

MDPI

Article

Earthquake Information Extraction and Comparison from Different Sources Based on Web Text

Xuehua Han [1,2,3] and Juanle Wang [1,3,4,*]

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; hanxh@lreis.ac.cn
2 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
3 The International Knowledge Centre for Engineering Sciences and Technology (IKCEST) under the Auspices of UNESCO, Beijing 100088, China
4 Jiangsu Centre for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: wangjl@igsnrr.ac.cn; Tel.: +86-010-6488-8016

Received: 12 April 2019; Accepted: 26 May 2019; Published: 30 May 2019

Abstract: Web text, using natural language to describe a disaster event, contains a considerable amount of disaster information. Automatic extraction from web text of this disaster information (e.g., time, location, casualties, and disaster losses) is an important supplement to conventional disaster monitoring data. This study extracted and compared the characteristics of earthquake disaster information from web news media reports (news reports) and online disaster reduction agency reports (professional reports). Using earthquakes in China from 2015 to 2017 as a case study, a series of rules were created for extracting earthquake event information, including temporal extraction rules, a location trigger dictionary, and an attribute trigger dictionary. The differences in characteristics of news reports and professional reports were investigated in terms of their quantity and spatiotemporal distribution through statistical analysis, geocoding, and kernel density estimation. The information extracted from each set of reports was also compared with authoritative data. The results indicated that news reports are more extensive and have richer information. In contrast, professional reports are less repetitive as well as more accurate and standardized, mainly focusing on earthquakes with $Ms \geq 4$ and/or earthquakes that may cause damage. These characteristics of disaster information from different web texts sources can be used to improve the efficiency and analysis of disaster information extraction. In addition, the rule-based approach proposed herein was found to be an accurate and viable way to extract earthquake information from web texts. The approach provided the technical basics and background information to support further research seeking human-centric disaster information, which cannot be acquired using traditional instrument monitoring methods, from web text.

Keywords: earthquake disaster; web text; information extraction; spatiotemporal distribution; China

1. Introduction

In recent years, the Internet has become the foremost means for disseminating information and knowledge and has enabled the government, professional organizations, news media, and even the affected population to quickly publicize an overwhelming amount of disaster information. For example, within 48 h after the earthquake in Jiuzhaigou, Sichuan, China, a huge volume of web text related to the earthquake emerged, including 85,000 news reports, 40,000 Application (APP) news articles, 70,000 WeChat articles, and 1.79 million microblogs [1]. Being multi-source, dynamic and heterogeneous, web text is a useful source of data to improve emergency response and strengthen disaster information

acquisition. Numerous studies have explored the extraction and analysis of disaster information from web text. However, very few studies have focused on the influence of different web text (e.g., news articles, official reports, and microblogs) on the extracted disaster information. The structure of sentences, the information concerns, and the reporting perspectives vary among different Web texts sources [2]. How to make effective use of multiple web texts to assist disaster management, based on their diverse features, remains an unresolved question. Therefore, the characteristics of disaster information from varied web texts need to be further explored.

Most existing studies focused on innovative methods for extracting disaster information from one type of web text. Valero et al. described a system based on machine learning methods to improve the acquisition of disaster information from online news reports [3]. Zhang combined a rules model and a statistical model to extract and visualize spatiotemporal information on earthquake events in web news texts [4]. Wang and Stewart applied ontology to extract spatiotemporal and semantic information on typhoons from web news reports [5]. Liu 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 [6]. Herford et al. sought to enhance the identification of relevant messages from social media using an approach that relies upon the relations between georeferenced social media messages and the geographic features of flood phenomena [7]. Song et al. constructed a web information extraction algorithm that supported dynamic convergence, according to the characteristics and timeliness of disaster information [8]. Yang et al. proposed a method for spatial information extraction from earthquake event news, based on geographic names and semantic technology [9]. Stewart and Wang automatically extracted spatial and temporal references from web texts and represented the spatiotemporal characterizations of events in a dynamic mapping environment [10]. Fan created an extraction rule based on syntax analysis to identify earthquake events and extract information from web news reports [11]. Li et al. introduce a novel approach to mapping floods in near-real time by leveraging Twitter data into geospatial processes [12]. Wang et al. analyzed the wildfire-related Twitter activities to gain insights into the usefulness of social media data in revealing situational awareness [13]. Shin et al. presented an information extraction method for government reports based on patterns and dictionaries [2]. Shengxiang et al. proposed an algorithm based on statistical and rule-based methods for extracting topic time from Web news reports [14]. Most of these studies focused on methods of identifying useful information from news reports, official reports, or social media while ignoring the characteristics of different web texts sources. We need information on the relative merits of the different web texts sources so future studies can make an informed decision on what source(s) to use and also utilize web texts resources more efficiently.

Based on previous works, this study explored the difference between web news media reports (herein referred to as 'news reports') and online disaster reduction agency reports (herein 'professional reports') on disaster information. Due to the integrity of the monitoring system and high openness of authoritative earthquake data, earthquakes occurring in China between 1 January 2015, and 31 December 2017, were used as the case study. A set of rules was created to automate the extraction of spatiotemporal information of earthquake events from both news reports and professional reports. Then the differences in the spatiotemporal distribution of earthquake events, between the news reports and professional reports, were investigated. An authoritative earthquake data set acquired from the China Earthquake Networks Centre (CENC) was utilized to evaluate the comparative results.

2. Materials and Methods

2.1. Data and Pre-Processing

According to China's national preparatory plan for earthquake emergencies (2012 revised edition), earthquakes should be classified 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 system was used in this study.

Data includes news reports data, professional report data and authoritative monitoring earthquake data. They are introduced below, followed by preprocessing.

(1) News Reports. News reports related to earthquakes were gathered from SINA.com, which is an online news media source with the largest user group in China. 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. Using the names of all the provinces in China, and "earthquake" and "occur" as the search keywords, a total of 2963 news reports, relevant to earthquakes from 1 January 2015, to 31 November 2017, were gathered.

(2) Professional reports. We gathered professional reports related to earthquakes from the website of China National Commission for Disaster Reduction (NCDR-China), which 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. The "latest disaster" column of the website was selected as the seed page for the crawler. Using the beautiful soup library in Python, we parsed the pages and collected 1062 professional reports about natural disasters from 1 January 2015, to 31 November 2017. Of these, 219 were related to earthquakes.

(3) Authoritative earthquake data. Through the website of China Earthquake Networks Centre (CENC), which is one of the most important hubs of China's earthquake disaster reduction network and the basis of information for the international community, we acquired an earthquake data set that included information about the location, time and magnitude of each earthquake between 1 January 2015 and 31 November 2017.

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 sorted the news reports according to their release time. Then, the event location and magnitude were extracted from the news title and contrasted one by one. If the location and magnitude from two news report are the same, they were considered to be duplicates, and the most recent news text was retained. Since the professional reports have less repetition and similar texts, they did not need to be filtered. In the end, we compiled 984 news reports and 163 professional reports, which were stored in an SQLite database (https://www.sqlite.org/index.html).

2.2. Named Entity Recognition

Named entity recognition (NER) addresses the problem of the identification (detection) and classification of predefined types of named entities [15,16], such as organizations, persons, place names, temporal expressions, and numerical and currency expressions. There are several highly regarded NER software packages, including the NLPIR-ICTCLAS Chinese lexical analysis system developed by the Chinese Academy of Science (http://ictclas.nlpir.org/) and the Language Technology Platform developed by the Harbin Institute of Technology (https://www.ltp-cloud.com/). Rule-based and statistical learning-based methods are commonly used for information extraction. In this study, as the language in news reports and professional reports tends to be standardized, we use a rule-based method for information extraction of earthquake events. Supported by the NLPIR-ICTCLAS Chinese lexical analysis system, a set of extraction rules relevant to earthquake events was built to achieve our goals.

2.2.1. Earthquake Information Extraction Rules

(1). Rules for extracting temporal information

There are various temporal expressions in texts related to earthquakes, including relative time and absolute time. For example, relative time refers to concepts such as "tomorrow", "the day after tomorrow", and "today". Absolute time, in contrast, refers to specific dates, like "1 January 2015". According to the characteristics of temporal expressions of earthquake events, a temporal expression dictionary was constructed and is shown in Table 1. Then, a set of temporal extraction rules was

designed for each type of temporal expression using Regular Expression. Regular expressions contain both special and ordinary characters to provide matching operations. "\d" matches any decimal digit. "{m}" specifies that exactly m copies of the previous characters should be matched. For example, "\d{4}" will match exactly four decimal digit. "d{1,2}" will match exactly one or two decimal digits. Examples of temporal rules are listed in Table 2.

Table 1. Temporal expression dictionary.

Temporal Type	Expressions	Examples
Absolute time	Year/Month/Day/Hour/Minute Month/Day/Hour/Minute Day/Hour/Minute	"At 21:27 on 21 May 2018" "At 11:09 on May 19" "At 21:10 on first"
Relative time	"today," "afternoon" "morning," "today"	"1:55 this afternoon," "8:48:22 in the morning," "2:39 this morning"

Table 2. Temporal extraction rules.

Temporal Type	Extraction Rules	Examples
Absolute time	\d{4}year\d{1,2}month\d{1,2}day\d{1,2}hour\d{1,2}minute \d{1,2}month\d{1,2}day\d{1,2}hour\d{1,2}minute \d{1,2}day\d{1,2}hour\d{1,2}minute	Year/Month/Day/Hour/Minute Month/Day/Hour/Minute Day/Hour/Minute
Relative time	Today|afternoon|morning|today	Today, afternoon, morning, today

(2). Location trigger dictionary

The types 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.' The NLPIR-ICTCLAS Chinese lexical analysis system can identify most locations in text, but there are a variety of locations in earthquake texts, many of which are not related to the reported earthquake. For example:

"China.org.cn, March 30, a magnitude 3.0 earthquake struck Weiyuan County, Neijiang, Sichuan (29.52° N, 104.56° 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 was 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.

(3). Attribute trigger dictionary

Earthquake attribute information may include the epicenter, magnitude, casualties, housing losses, and economic losses. In Chinese, there are some specific descriptors for disaster attributes. By analyzing the descriptive characteristics of earthquake attribute information in texts, we designed an attribute trigger dictionary for earthquakes, as shown in Table 4.

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. This is divided into five steps: (1) Text preprocessing. We used the NLPIR for keyword filtering, text segmentation, and information annotation. First, we broke down a block of text into sentences and filtered the sentences according to some earthquakes keyword. Then, word segmentation and word annotation were conducted with the NLPIR implemented in Python. (2) Temporal information extraction. Using the temporal extraction rules, we extracted the time information of earthquake events by the regular expression operations ('re' module) in Python. (3) Spatial information extraction. Based on the location trigger dictionary and the annotated text,

the place names related to earthquakes were extracted, and the adjacent place names were combined into a complete spatial expression. (4) Attribute information extraction. Based on the attribute trigger dictionary and regular expressions, the earthquake attribute information was extracted using the regular expression operations ('re' module) in Python. (5) Result filtering. The result was checked manually, and some incorrect records are deleted.

Table 3. Location trigger dictionary.

Location Type	Trigger Words	Examples
Locations related to earthquake	"latitude," "north latitude" "longitude," "east longitude," "occur"	China news service, February 9 2018, (reporter Dong Fei), according to China Earthquake Network Center, at 19:00 p.m. on February 9 2018, 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 in 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	"Seismological Bureau," "Earthquake Networks Center," "time," "Networks Center"	

Table 4. Attribute trigger dictionary.

Attribute Type	Trigger Words
Magnitude	"magnitude unit," "earthquake," "magnitude," "Richter,"
Focal depth	"focal depth," "depth," "deep," "km," "km,"
Casualties	"Casualty unit," "death," "death," "Wounded," "dead body," "dead," "dead," "number"
Epicenter	"epicenter," "place," "located," "center"

2.3. Geocoding

Geocoding is the computational process of transforming an address description into a spatial coordinate. It includes two steps: address standardization and address matching. At present, there are several commonly used geocoding software packages available, including the ArcGIS Geocoding module, Yahoo's Geocoding application programming interface (API), Google Geocoder service, and Baidu Map Geocoding API.

In this study, for earthquake events without latitude and longitude coordinates, the Baidu Map Geocoding API was used to geocode the location information. First, we standardized the extracted location according to the code format of the Geocoding API and generated a valid uniform resource locator (URL). We then sent a hypertext transfer protocol (HTTP) request to the Geocoding API for the returned JavaScript object notation (JSON) data. By parsing the "latitude" and "longitude" parameters from the JSON data, the geographic coordinates were assigned to each extracted location.

2.4. Kernel Density Estimation

Kernel density estimation is generally used to detect the intensity of events by generating a smooth surface using a quadratic kernel function [17]. Let $(s_1, \ldots, s_i, \ldots, s_n)$ be a series of event samples

distributed with an unknown density, $\hat{\lambda}(s)$, in the study area. We then can estimate the shape of this function, $\hat{\lambda}(s)$. Its kernel density estimator is:

$$\hat{\lambda}(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k(\frac{s - s_i}{\tau}) \quad (1)$$

Here, k is the kernel function, τ is a smoothing parameter called the bandwidth, i.e., the search radius within which to calculate density, and $s - s_i$ is the distance between s and s_i.

This study performed a kernel density estimation using ArcGIS aiming to identify the hot spots of earthquakes. Three parameters were used in the kernel density estimation: kernel search radius (bandwidth) for calculating the density, cell size for the output raster data, and population field to denote population values for each feature. 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. A cell size of 6 km was used to show sufficient detail. The magnitude was used as the population field value.

2.5. Evaluation Approach

In information extraction, the most commonly used evaluation approach is to compute precision, recall and F-measure metrics over a set of evaluation data. Precision is the fraction of the correct extraction results among the number of extraction results, and recall is the fraction of the correct extraction results among the numbers of results that should have been extracted. The F1-measure is used as a weighted harmonic mean of precision and recall. Higher values of the F1-measure indicate that the extraction method is more effective [18]. Precision (P), recall (R), and F1-measure (F1) are defined as follows:

$$P = \frac{\text{the number of correctly extracted results}}{\text{the number of extracted results}} \times 100\% \quad (2)$$

$$R = \frac{\text{the number of correctly extracted results}}{\text{the number of results expected to be extracted}} \times 100\% \quad (3)$$

$$F1 = \frac{\mathrm{P} \times \mathrm{R} \times 2}{\mathrm{P} + \mathrm{R}} \times 100\% \quad (4)$$

In this study, to compute the accuracy (precision, recall and F1-measure) of the extraction results, the two automatically processed extraction results, by rules as above, were manually evaluated in order to acquire the numbers of correctly extracted earthquake events, incorrectly extracted earthquake events, and missing earthquake events. Then, the extraction accuracies of news reports and processional reports, respectively, were calculated.

3. Results

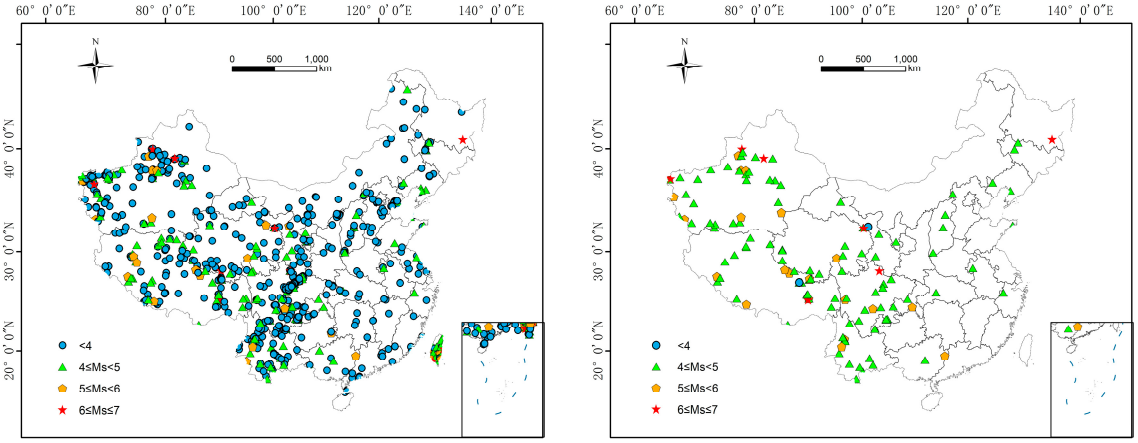
3.1. Results Description

The extracted information about earthquake events included their time, location, magnitude, latitude, and longitude. We processed 850 earthquake events from news reports and 133 earthquake events from professional reports for the time period from 1 January 2015, to 31 November 2017. Table 5 shows the classification statistics of the extracted earthquakes and the authoritative earthquake data from CENC. CENC reported 1149 earthquakes with $Ms < 4$, of these 594 were documented in news reports, compared to only two in professional reports. For general earthquakes ($4.0 \leq Ms < 5.0$), there were 189 earthquake events found in news reports, 104 earthquake events in professional reports, and 308 earthquakes from CENC. For larger earthquakes ($5.0 \leq Ms < 6.0$), there were 51 earthquake events in news reports, 20 earthquake events in professional reports and 65 reported by CENC. For major earthquakes ($6.0 \leq Ms < 7.0$), there were 15 earthquake events in news reports, six earthquake events in professional reports and 15 earthquakes from CENC. There was only one massive earthquake, which was identified in both types of reports and by CENC.

Table 5. Classification of earthquakes.

Levels of Earthquake Events	News Report	Professional Report	CENC
$Ms < 4$	594	2	1449
$4.0 \leq Ms < 5.0$	189	104	308
$5.0 \leq Ms < 6.0$	51	20	65
$6.0 \leq Ms < 7.0$	15	6	15
$Ms \geq 7.0$	1	1	1
Total	850	133	1838

Figure 1 shows the spatial distribution of earthquake events from both news reports and professional reports. The earthquakes identified from news reports cover almost all of China, while the earthquakes identified from professional reports were much sparser.



(**a**) (**b**)

Figure 1. (**a**) Spatial distribution of earthquake events from news reports, (**b**) spatial distribution of earthquake events from professional reports.

Table 6 shows the accuracy of earthquake information extraction. For news reports related to earthquakes, the precision, recall and F1-measure values were 88.4%, 86.4%, and 87.4% respectively. For professional reports related to earthquakes, the precision, recall and F1-measure values were 96.4%, 81.6%, and 88.4%.

Table 6. Evaluation results of earthquake extraction.

	News Reports	Professional Reports
Precision (P)	88.4%	96.4%
Recall (R)	86.4%	81.6%
F1	87.4%	88.4%

3.2. Statistical Analysis

Figure 2 shows a comparison of earthquake events from news reports and professional reports. The number of extracted earthquakes from news reports was about six times higher than that from professional reports. More specifically, except in massive earthquakes, the differences in the quantity of earthquakes from news reports and professional reports varied with earthquake types; the difference was the greatest in earthquakes with $Ms < 4$. Both datasets (the earthquake events from news reports and earthquake events from professional reports) were counted for each province in China. As shown in Figure 3a, the earthquake events extracted from news reports were widely distributed across 30

provinces and were absent only in Shanghai, Hainan, Hong Kong, and Macao. Sichuan province had the highest number of earthquake events from news reports (17.06% of the total events extracted from news reports). Xinjiang, Xizang, Taiwan, Yunnan, and Qinghai each accounted for more than 5%. The provinces with higher proportions (≥5%) of earthquake events from news reports are located either in China's western region or in Taiwan. North China had the next highest number of events while South China had the lowest. The number of earthquake events generally reduced from west to east.

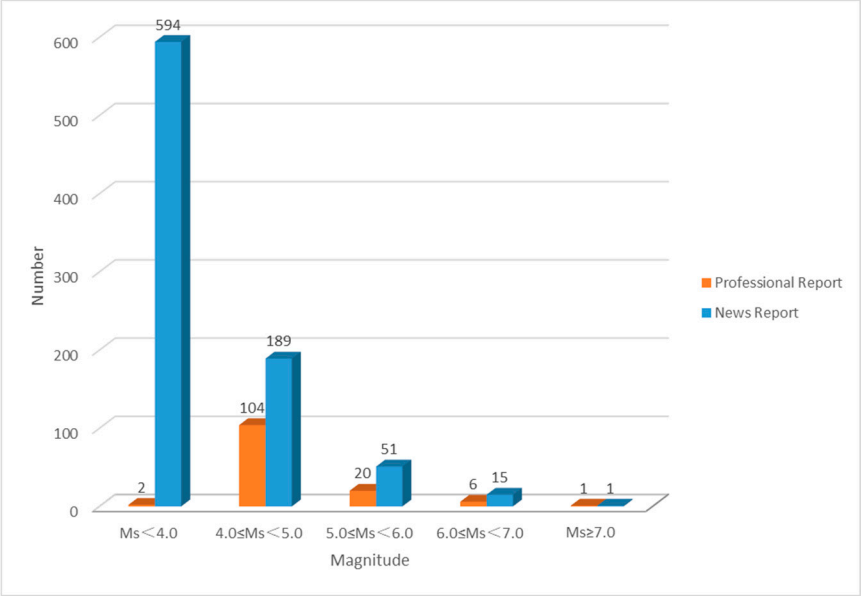


Figure 2. Number of earthquake events, by magnitude, from news reports and professional reports from 1 January 2015, to 31 November 2017.

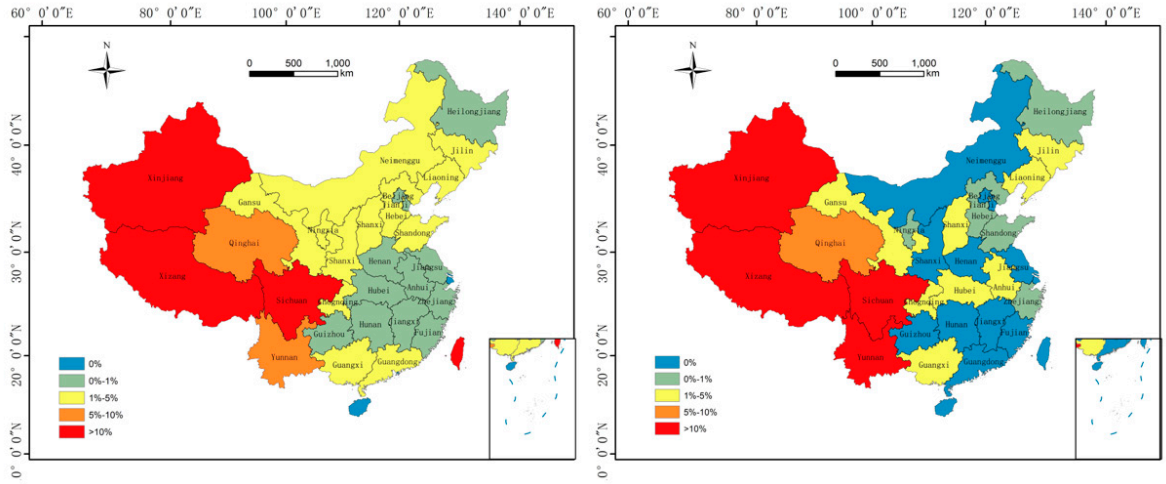


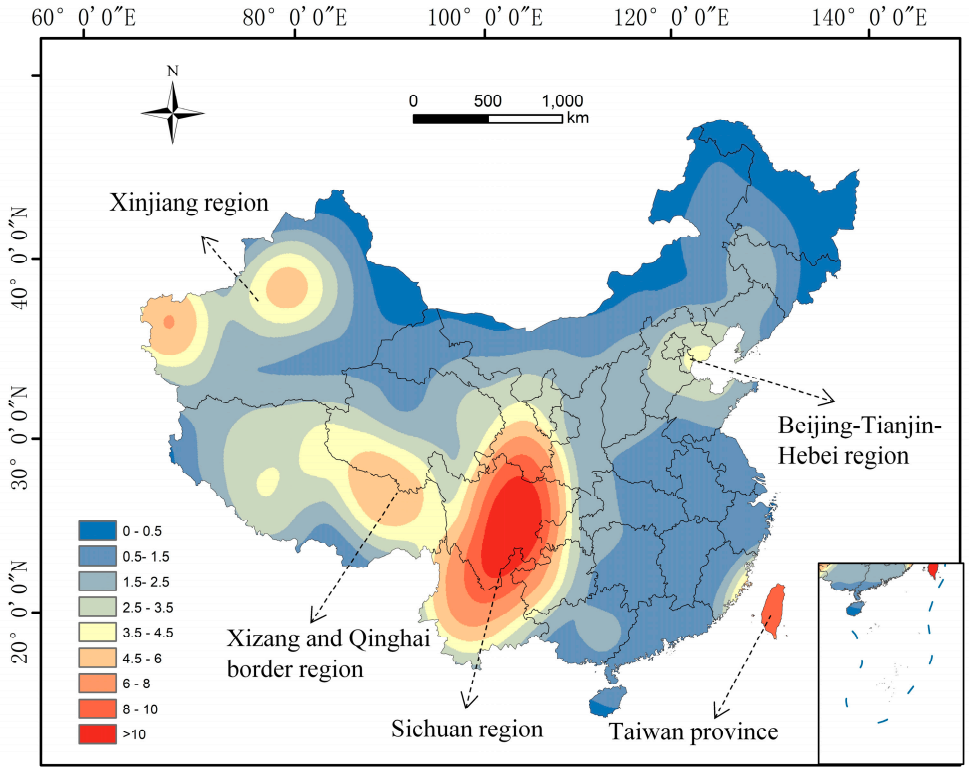
Figure 3. (**a**) Distribution of earthquake events at the province level from news reports from 2015 to 2017, (**b**) distribution of earthquake events at province level from professional reports from 2015 to 2017.

The distribution of earthquake events from professional reports across 18 provinces is shown in Figure 3b. Xinjiang had the highest number of earthquake events, accounting for 27.82% of the total events extracted. Xizang, Yunnan, Sichuan, and Qinghai each accounted for more than 5%. The China western region had the highest proportion (≥5%) of earthquake events. Compared with the news reports, the earthquake events from professional reports were mainly concentrated in southwest and northwest China.

3.3. Kernel Density Estimation

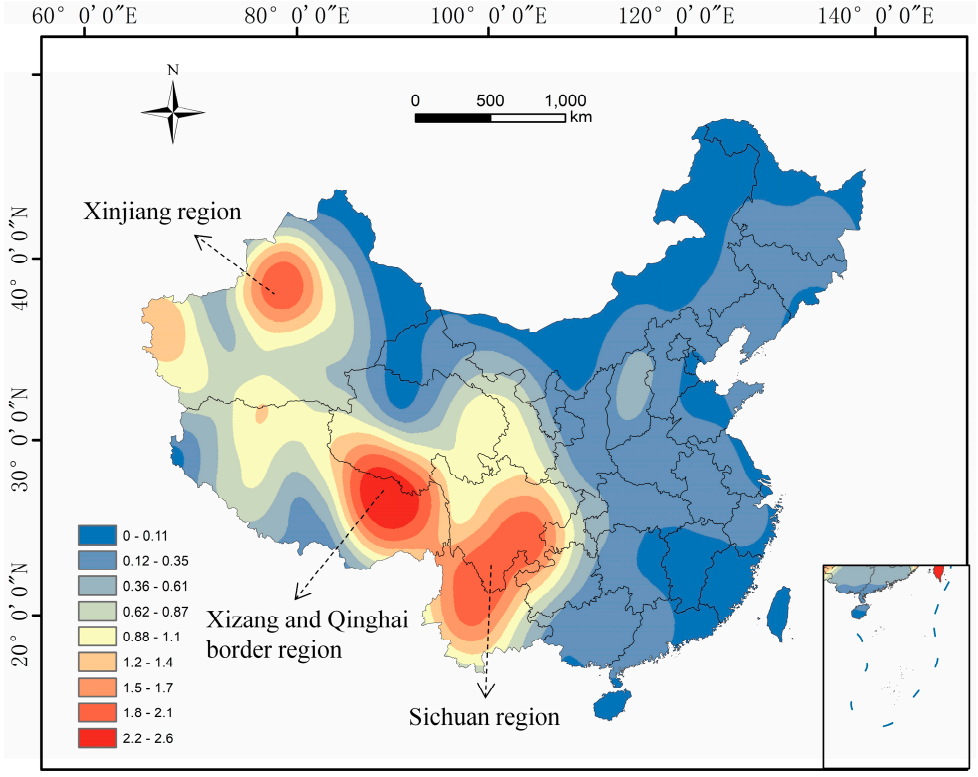
Figure 4a shows the spatial distribution of the kernel density estimation, indicating that the high-density areas of earthquake events extracted from news reports were located in the Beijing-Tianjin-Hebei region, Sichuan province, the Xizang and Qinghai border region, Xinjiang province, and Taiwan province. As shown in Figure 4b, the high-density areas of earthquake events extracted from professional reports were located in China's western regions, showing three high-density centers (Sichuan and Yunnan adjacent region; Xizang and Qinghai adjacent region; and Xinjiang province). In Figure 4c, the earthquake data obtained from CENC is mapped by kernel density estimation and shows a similar spatial pattern of high-density earthquake events as both news reports and professional reports. The density of earthquakes from news reports and professional reports is lower than that of CENC in the southwest area of Xinjiang province.

To further examine this difference in the spatial distributions, we statistically and spatially analyzed earthquakes from news reports in the Beijing-Tianjin-Hebei region and Taiwan province. As shown in Table 7, 38 earthquake events occurred in Beijing-Tianjin-Hebei, of which 35 events had a magnitude of <4. There were 87 reported earthquakes in Taiwan, most of which had $Ms \geq 4$. Figure 5 shows the spatial distribution of earthquake events in the Beijing-Tianjin-Hebei region and Taiwan province. The earthquake events in the Beijing-Tianjin-Hebei region are mainly distributed in Tangshan city. In Taiwan, most earthquakes are distributed in the eastern sea area.

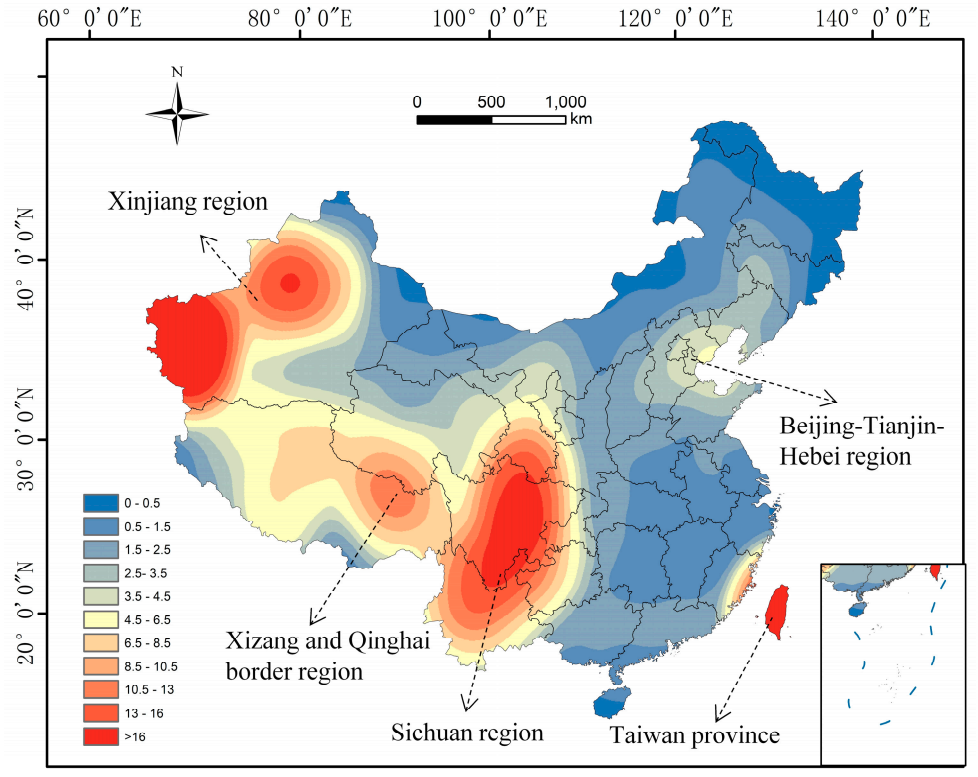


(**a**)

Figure 4. *Cont.*



(**b**)

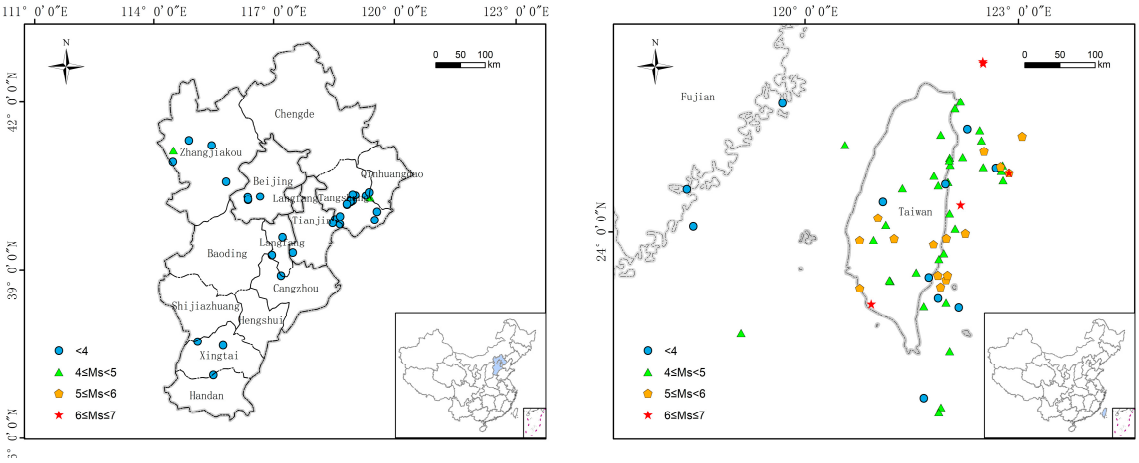


(**c**)

Figure 4. (**a**) Kernel density estimation for earthquake events in news reports, (**b**) kernel density estimation for earthquake events in professional reports, (**c**) kernel density estimation for earthquakes obtained from China Earthquake Networks Centre (CENC).

Table 7. Classification statistics of earthquakes from news reports in the Beijing-Tianjin-Hebei region and in Taiwan province.

Levels of Earthquake	Beijing-Tianjin-Hebei	Taiwan
$Ms < 4$	35	10
$4.0 \leq Ms < 5.0$	3	46
$5.0 \leq Ms < 6.0$	0	23
$6.0 \leq Ms < 7.0$	0	8
$Ms \geq 7.0$	0	0
Total	38	87



(**a**) (**b**)

Figure 5. (**a**) Spatial distribution of news reported earthquake events in the Beijing-Tianjin-Hebei region, (**b**) spatial distribution of news reported earthquake events in Taiwan.

In addition, we statistically analyzed and compared earthquakes from CENC, professional reports, and news reports in Xinjiang province (Table 8). From CENC and news reports, most earthquakes that occurred in Xinjiang province had a magnitude of <4. However, all of the earthquake events from professional reports were of a magnitude above 4.

Table 8. Classification statistics of earthquakes in Xinjiang province.

Levels of Earthquake	CENC	News Reports	Professional Reports
$Ms < 4.0$	564	105	0
$4.0 \leq Ms < 5.0$	67	22	27
$5.0 \leq Ms < 6.0$	10	8	7
$6.0 \leq Ms < 7.0$	4	3	3
$Ms \geq 7.0$	0	0	0
Total	645	138	37

4. Discussion

A rule-based approach was used to automatically extract spatiotemporal and attribute information about earthquakes from web news reports and professional reports. The extracted earthquake information was accurate and the process was viable. The evaluation results (Table 6) achieved by our method compared well with other evaluation outcomes, for example, the precision (83%), recall (82%) and F-measure (82.5%) values obtained using ontologies with natural language processing (NLP) and geographic information retrieval (GIR) techniques [4], the average evaluation results (precision 69%, recall 85%, and F-measure 76%) obtained using machine learning methods for extracting disaster information from online news reports [2]. Through kernel density estimation, the spatial distribution

of the earthquake events from the reports shows similar spatiotemporal patterns to the authoritative earthquake data from CENC, which demonstrates the validity of the data extracted from reports. Furthermore, the precision of the extracted results from professional reports was better than that from news reports. One possible explanation is that the disaster reports issued by the disaster reduction agency are more standardized, less repetitive and do not contain irrelevant information, which improves the accuracy of information extraction.

The comparison results show that news reports are more comprehensive and cover all levels of earthquake magnitude, while professional reports focus on earthquakes with $Ms \geq 4$ or earthquakes that may cause damage. The quantity of extracted earthquakes from news reports was far more than that from professional reports, especially in earthquakes with $Ms < 4$. Indeed, 98% of professional reports were about earthquakes with $Ms \geq 4$. The main differences in the spatial distribution of earthquakes were evident in the Beijing-Tianjin-Hebei region and Taiwan province. Few of the earthquakes in those two places were covered in professional reports. Further analysis suggested the possible explanation that most of the earthquakes in the Beijing-Tianjin-Hebei region had a magnitude of <4 while earthquakes in Taiwan mainly occurred near the sea without causing damage. A similar explanation may apply in Xinjiang province where most earthquakes events were $Ms < 4$, contributing to the low coverage in professional reports. These findings reinforce the idea that professional reports are mainly focused on earthquakes with $Ms \geq 4$, and/or on earthquakes that may cause damage. As might be intuitively anticipated, news reports have greater coverage of disaster information, while professional reports tend to focus on disasters impacts.

Nevertheless, this study has some limitations. First, due to the rapid spread of the mass information via the Internet, some of the information contained in news reports was likely to come from official reports and was not independent of the latter, which may influence the comparison results. Second, there was a limited amount of quantifiable reports of earthquakes in China because there were only two data source websites used in this study, Sina.com and the website of NCDR-China. More mainstream media websites and other internet information sources could provide richer disaster-related data. Third, the information gathered from web text in this study was relatively basic disaster information and could be assessed for accuracy against the same parameters of monitoring data. As the frequency of natural hazard increases, disaster management has greater human-centric information requirements (e.g., disaster losses, casualties, human activities, and disaster-related public opinion) to facilitate better decision-making for reduction of human and property loss. Fourth, the method used in the study was mainly based on current technologies and can be improved in several aspects. For example, the extracted rules in the study were manually built, which was time-consuming and gave poor portability. In other disaster scenarios, we may have to rebuild the extracted rules. This situation can be improved by combining statistical models and machine learning in future studies. Despite these limitations, this study has contributed to the existing research on the extraction and analysis of disaster information from web text by revealing the distinct characteristics of disaster information from different web texts. This awareness can improve the efficiency of extraction and analysis. Furthermore, the information extraction rules constructed in the study can serve as a reliable basis for further research about earthquake information extraction from web text, such as disaster loss, secondary disasters, public sentiment, and so on.

5. Conclusions

This study explored the extraction from web text of spatiotemporal and attribute information about earthquakes, compared the quantity and spatial distribution of earthquakes in news reports and professional reports, and tested the viability of a rule-based extraction method. A set of temporal extraction rules, a location trigger dictionary, and an attribute trigger dictionary were created for extracting earthquake event information from web text. The results indicate that news reports are more comprehensive and extensive, while professional reports are more standardized, mainly focusing on earthquakes with $Ms \geq 4$ and/or earthquakes that may cause damage. The separate characteristic of

news reports and professional reports can provide helpful information for selecting data sources for disaster information extraction. In addition, the rule-based approach proposed here is an accurate and viable way to acquire earthquake information from web text.

Future work should strengthen the analysis of disaster information that cannot be provided by disaster monitoring data from more mainstream media websites and other internet information sources. Additionally, with the growth of volunteered geographic information (VGI), the use of social media for disaster management has been explored in various papers [19,20]. There is an urgent need to strengthen the acquisition and analysis of earthquake disaster information from social media following this study.

Author Contributions: Xuehua Han drafted the manuscript and was responsible for data preparation, data processing and analysis. Juanle Wang was responsible for the research design, result analysis, and reviewed the manuscript.

Funding: This research was funded by the Strategic Priority Research Program (Class A) of the Chinese Academy of Sciences (Grant No. XDA19040501), the National Natural Science Foundation of China, grant number 41421001; the Construction Project of the China Knowledge Center for Engineering Sciences and Technology (Grant No. CKCEST-2018-2-8), and the 13th Five-year Information Plan of the Chinese Academy of Sciences (Grant No. XXH13505-07).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. People's Daily Online. The Big Data Analysis of the Public Opinion 72 Hours after the Jiuzhaigou Earthquake. Available online: http://yuqing.people.com.cn/n1/2017/0815/c209043-29471816.html (accessed on 15 August 2017). (In Chinese).
2. Shin, S.; Hong, S.; Song, S. An Efficient Damage Information Extraction from Government Disaster Report. *J. Int. Comput. Serv.* **2017**, *18*, 55–63.
3. Téllez, V.A.; Manuel, M.Y.G.; Villaseñor, P.L. Using Machine Learning for Extracting Information from Natural Disaster News Reports. *Comput. Y Sist.* **2009**, *13*, 33–44.
4. Zhang, C.J. *Interpretation of Event Spatio-temporal and Attribute Information in Chinese Text*; Nanjing Normal University: Nanjing, China, 2013. (In Chinese)
5. Wang, W.; Stewart, K. Spatiotemporal and semantic information extraction from Web news reports about natural hazards. *Comput. Environ. Urban Syst.* **2015**, *50*, 30–40. [CrossRef]
6. Liu, S.Y. *Extracting Landslide Disaster Information from Web Pages*; Southwest Jiaotong University: Chengdu, China, 2015. (In Chinese)
7. Herfort, B.; Brenning, A.; Zipf, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 667–689.
8. Song, J.G.; Wang, Z.X.; Li, Q.Y.; Ma, S.L.; Lv, J.H. Internet information process oriented to the earthquake response. *J. Beijing Univ. Aeronaut. Astronaut.* **2017**, *43*, 1155–1164. (In Chinese)
9. Yang, J.; Hong, F.; Huaiyuan, L.I. Spatial Information Extraction of Web Seismic Event Based on Geographic Names Semantic Technology. *J. Geomat.* **2013**, *38*, 10–13. (In Chinese)
10. Stewart, K.; Wang, W. Representing dynamic phenomena based on spatiotemporal information extracted from web documents. In Proceedings of the Sixth International Conference on Geographic Information Science, Zurich, Switzerland, 14–17 September 2010. Extended Abstracts.
11. Fan, H.; Guo, D.; Li, H. Extraction of spatio-temporal information of earthquake event based on semantic technology. In *MIPPR 2015: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications*; International Society for Optics and Photonics: Bellingham, WA, USA, 2015; Volume 9815, p. 981509.
12. Li, Z.; Wang, C.; Emrich, C.T.; Guo, D. A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods. *Cartogr. Geogr. Inf. Sci.* **2018**, *45*, 97–110. [CrossRef]
13. Wang, Z.; Ye, X.; Tsou, M.H. Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Nat. Hazard.* **2016**, *83*, 523–540. [CrossRef]
14. Shengxiang, W.; Chaoliang, W.; Weixin, Y. Topic time extraction algorithm of Web pages based on hierarchical tree. *J. Comput. Appl.* **2017**, *37* (Suppl. S1), 270–272. (In Chinese)

15. Chinchor, N. MUC7 Named Entity Task Definition. In Proceedings of the Seventh Message Understanding Conference, Fairfax, VA, USA, 29 April–1 May 1998.
16. Zheng, J.; Fu, L.; Ma, X.; Fox, P. SEM+: Tool for discovering concept mapping in Earth science related domain. *Earth Sci. Inform.* **2015**, *8*, 95–102. [CrossRef]
17. Okabe, A.; Satoh, T.; Sugihara, K. A kernel density estimation method for networks, its computational method and a GIS-based tool. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 7–32. [CrossRef]
18. Piskorski, J.; Yangarber, R. Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 23–49.
19. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *Geojournal* **2007**, *69*, 211–221. [CrossRef]
20. Li, H.J.; Liang, H.B. Natural Disaster Forecasting System Based Association Rules. *Comput. Syst. Appl.* **2017**, *26*, 50–55. (In Chinese)