DATA INTEGRATION TECHNOLOGY FOR MULTI-SOURCE GEO-SPATIAL INFORMATION FROM WEB FOR RAPID EMERGENCY RESPONSE

Wang Yong u1*, Liu Jiping1, Zhang Xiuhong2, Luo An1

1Chinese Academy of Surveying and Mapping; No. 28 Lianhuachi West Road, Haidian District, Beijing, 100830, China;
2Wuhan University; :No.129 Luoyu Road, Wuhan, 430079, China;
*Corresponding author; e-mail: wangyong@casm.ac.cn;

Abstract
The basic geo-spatial information for disaster reduction, including environment, transportation, construction, population and so on, plays a crucial role in emergency management. With the rapid development of Internet, web has become an important channel for the rapid information acquisition. However, most of the information exists in the Deep Web that is difficult to obtain, and different sources of information can not be directly integrated due to their heterogeneity and structural diversity. Therefore, in this paper we propose a multi-source geo-spatial information integration method for disaster reduction. By constructing web focused crawler, we can realize the on-demand aggregation and classification cleaning of multi-source geo-spatial information, and implement the aggregation and duplicate removal of heterogeneous geographic entities based on semantic similarity calculation. Experiments show that this method can support the rapid integration of multi-source geo-spatial information for disaster reduction in the case of sudden disasters and provide rich geo-spatial information support for emergency management.

Keywords: Web, focused crawler, data integration, geo-spatial information, disaster reduction

INTRODUCTION

In recent years, natural disasters happened frequently around the world and have caused enormous losses of life and property to human society. Adequate, accurate and timely disaster information is of great significance for disaster emergency management and risk reduction. The advances and access to new technologies such as semantic Web and Web 2.0 have helped progress information sharing efforts in the field (Harvard, 2011). During the disasters, major websites continue to provide up-to-date status reports about ongoing emergencies. At the same time, social networks such as Twitter, Facebook and micro-blog have become an important channel for finding real-time information in the disaster relief process (Kaplan and Haenlein, 2010). Internet has become the mainstream media for disaster information dissemination and collection, as well as the important channel for the rapid acquisition of disaster information.

The basic geo-spatial information for disaster reduction includes environment, transportation, construction, population, enterprise, institution, economy and so on, which plays a crucial role in the field of emergency management, disaster response and disaster recovery. The web contains a large amount of spatiotemporal information related to emergency management (Shi, 2011). How to obtain the spatial location, scope, semantics and Spatiotemporal evolution of disaster events in the massive data resources, and quickly analyze and judge the disaster situation, is an important part of scientific guidance for disaster relief command. In addition, the data obtained from the Internet are diverse and heterogeneous, mainly reflected in the differences in data sources, data formats (e.g., text, multi-media, GIS information) and content descriptions. How to integration these scattered and heterogeneous data and excavate valuable information become the bottleneck of the precise judgement of the disaster.

In recent years, related research and demonstration applications of information mining for disaster reduction from web have been gradually carried out. The focus of the applications is mainly in two aspects: emergency response and rescue, early warning and risk analysis. For example, the RESCUE program, funded by the National Science Foundation (NSF), searches and downloads relevant web information from various sites, and then obtains the information on the affected population, relief materials and disaster event progress through information extraction technology, and evaluates the disaster losses (Ma, 2007). The DRH project launched by Japanese scientist Hiroyuki Kameda proposed the idea of disaster reduction technology and knowledge sharing platform based on Web construction (Hiroyuki,2006).

In the face of the network disaster data with "multi-source, large volume, complex content and real-time", the traditional data acquisition and processing technology is difficult to meet the actual needs. Moreover, there is lack of an
integrated approach to support real-time integration of Multi-source Geo-spatial Information from Web in order to facilitate an Rapid Emergency Response. In view of the problem, in this paper we propose a multi-source geo-spatial information acquisition and integration method for disaster reduction. we can achieve the on-demand aggregation and classification cleaning of multi-source geo-spatial information for disaster reduction by developing a new web focused crawler system, and implement the aggregation and duplicate removal of heterogeneous geographic entities based on semantic similarity calculation.

TECHNICAL ARCHITECTURE

In this section, the key technologies of Geo-spatial information acquisition from web and multi-source information integration are described based on the existing technology (Yang, 2007) and our recent research results. Figure 1 illustrates the modules and interactions between them.

Information Collection and Access

This module is mainly to achieve the acquisition and storage of information from multiple online information sources (news, blogs, satellite images, GIS databases, etc.). It can provides the parallel and real-time visit to the websites and obtains initial pages through web crawler tools.

Conventional search engines (General search, Meta-search Engines, etc.) can search for texts, pictures, videos and other information, but can not deal with rich geographic semantics and spatial relationships. Therefore, developing a special geographic information Web retrieval system is an effective way to retrieve a large amount of geographic information on Web. In order to achieve the spatial data crawling of vertical web and deep web, this paper designs a multi-strategy parallel method for multi-source spatial data acquisition. The detailed technical process is shown in Figure 2.
The prototype system (NetCrawler), based on the .NET 4 framework, combines asynchronous I/O model and multi-thread parallel strategy to improve the efficiency of Web spatial data downloading and parsing. After getting the required Web data, the spatial data parsing engine based on template mapping is invoked to realize the parallel and efficient parsing of multi-source heterogeneous spatial data in Web. Then, the generated standardized XML document is submitted to the remote data storage service through a data submitted thread. The obtained result document is organized into a standard GeoJSON object, stored in the MongoDB spatial database, and the index is set up to complete the whole data acquisition and management process. The system mainly includes four parts: multi-source spatial data crawling module, heterogeneous spatial data parsing engine, error logging module and system setup module.

**Information Extraction**

The purpose of information extraction is to automatically extract structured information from unstructured or semi-structured documents. There are abundant disaster related geospatial information in the internet. Taking a large amount of Web texts as the data source, extracting structured disaster related information involving disaster events, time, space location and scope, direct loss and secondary impact, and cause of disaster automatically is of great significance for disaster research and risk reduction. In this section, we will introduce our research on information extraction from four aspects of location extraction, time extraction, event extraction and Map images semantic classification. The detailed extraction process is shown in Figure 3.
**Location Extraction**

The purpose of location extraction is to identify the place name address or geographic entity, such as administrative division, organization, gate address, and so on, from the web text data. In view of the structural characteristics of web texts for disaster events, this paper proposes a location extraction method based on Prefix suffix feature words and rules. The detailed steps are as follows:

1. Remove the meaningless characters and special symbols in the web text, and build the prefix and suffix feature lexicon of the address.
2. Combining the multi-layer address expression model and the address reference library, the Chinese word segmentation, semantic analysis and the pre suffix identification technology of the address are used to realize the automatic identification and extraction of the address for the web text information.
3. Combined with a variety of open source geocoding API services, the spatial location statistics inference is used to realize the precise location and expression of disaster oriented events.

**Time Extraction**

In natural language, the description of time information is usually expressed by words, phrases, and syntax, so the recognition process needs to refer to the corresponding dictionary and grammar rules (Zhang, 2014). This section combines the trigger words with the rule model to realize the time information analysis in Chinese text. First, by analyzing and summarizing the description characteristics of temporal information in Chinese texts, the temporal lexicon dictionary and the time information description library are constructed. And then, the time information normalization expression and semantic inference algorithm are designed to realize temporal information analysis. This process can be regarded as the three level information extraction and filtering process of temporal semantic role labeling, phrase recognition and syntactic pattern matching.

**Event Extraction**

Event extraction not only probes the types of events and identifies the basic elements of events (time, place, role and behavior), but also pays more attention to the temporal and spatial evolution of events from occurrence to extinction. Therefore, geographical events extraction includes event type identification, event element identification and spatiotemporal change tracking. Event type are usually judged by triggers. In this study, we use the weighted LDA topic model to extract and classify disaster events, establish the disaster event topic feature knowledge base, and then, use the SVM text classification algorithm to realize the type identification of disaster events. Event element identification can be achieved by constructing the knowledge framework for disaster events (Fan, 2018). Spatiotemporal change tracking is mainly to analyze the trend and characteristics of temporal and spatial distribution of disaster events. The specific process is to detect the events of the same topic, integrate the information of multiple events, and then build the change trend maps of time and space according to the time continuation or space transfer of the event (Jannik, 2010).

**Map images semantic classification**

The map images contain abundant geo-spatial information, such as location information, contour lines, buildings, and water. Extracting these information from the map images can support a variety of applications and research fields. Due to a large number of images on the web are not map images, fast and effective identification of the map images, is an important step in the mining of geo-spatial information in map images. In this paper, convolution neural network is introduced into the recognition of map images to realize automatic semantic indexing of web map images. We use the geo-spatial information vertical search engine to capture the images on the Internet, and obtains the training samples by combining self learning and manual annotation. Then, the training samples are used to train the convolution neural network to obtain the deep fuzzy features of the map images. After building the classification model of the map, the captured image is indexed and the result is returned to the picture library. Experiments show that the method can quickly and effectively identify the map images on the Internet.
Multi-source Information Integration

On the one hand, production of data from different sources is relatively independent, and the information standards and data processing requirements used during different production are different. On the other hand, different data acquisition approaches and times are also different (Alamdar F, 2016). Therefore, the data from multi-sources may differ in position, attribute, etc., which results in inconvenience in the application of these data. Therefore, data from different sources need to be fused ahead of application. This study aims for emergency response, multi-source information are integrated at the following levels:

Location Integration

On Internet, it’s unknown or heterogeneous in spatial reference, and data is published in encrypted form, which lead an inconsistency in location for data from different sources. The traditional location integration methods like spatial reference transformation, affine transformation processing, correction table based approach, open API based approach are difficult to effectively achieve location integration of multi-source geo-spatial data on the Internet environment. Therefore, this study proposes a consistency processing method based on geographic grid and quadratic multinomial transformation. Consistency processing is implemented locally for each unit in geographic grid, through which global approximation revise of nonlinear excursion can be received. Moreover, we present a method for grid automatic construction and control point generation based on iterative detection. Automatic generation of geographic grid unit can come true under given error threshold. Experiment shows that this method can better match position of multi-source geo-spatial data. The details of location integration of multi-source geo-spatial data are shown in Figure 5.

Figure 4. The process of Map images semantic classification
**Attribute Integration**

The production of data from different data sources is independent, and the classification, gradation and description of elements are different. There are great differences in the attribute expression of the same geo-spatial target in different information sources. For example, many words represent the same concept, or different structures denote the same (or similar) information, or the same concept expresses different meanings in different information sources. In addition, because of the lack of unified POI classification standards at the present stage, the classification information name and the contained semantic information are not exactly the same, which is inconvenience to match and compare the POI data. Attribute integration can be generalized as two aspects: attribute information fusion and category mapping.

**Attribute information fusion:** Fusion of attribute information from different sources of the same geo-spatial target can avoid shortage in information and unavailability of a single data source. To achieve the fusion of multi-source data attribute information, similarity measure for describing information should be built first. Here, the similarity of describing information is measured mainly based on vector space model (VSM) and word co-occurrence model. The text is represented as a space vector by VSM. Different words composing the text are given a certain weight based on word co-occurrence model (Wang, 2011), cosine of angle between two vectors reflects the similarity between two descriptions.

**Category mapping:** Different classification system and category description method lead to inconsistency of multi-source POI’s categorical attribute. To solve this problems, we proposes a mapping method based on Formal Concept Analysis. By using chinese text segmentation and synonym merging method, the semantic factors are extracted and optimized, and the background matrix for multi-source POI classification can be formed. Then the improved Chein algorithm is used to realize the automatic generation of the concept lattice of the POI classification system. The experiment shows that conversion and mapping of heterologous POI classified information can be effectively implemented by Formal Concept Analysis method.

**EXPERIMENTAL DEMONSTRATION AND VISUALIZATION**

Based on .Net framework, we have developed a visual platform for web Geo-spatial Information acquisition and integration. We have designed a new visual crawling engine to achieve efficient access to geo-spatial information from web (Figure 6). Figure 7 shows the crawling task management tracking portal, which is mainly used for crawling task creation and crawling status tracking.

As shown in Figure 8(a), we provide the location consistency processing tool for multi-source geo-spatial information from Web, which supports the conversion of multiple coordinate systems, and accepts various types of input data such as shp, dbf, xls, txt. Taking the two bus stations of "Taiping Road East" and "Jinjiacunqiao North" in Beijing as an example, the position offset before correction is more than 500 meters. When the spacing of grid processing units is set to 30 points and 15 points, the maximum position offset is reduced to 30 meters and 18 meters respectively (Figure 8(b)). In addition, we have also developed a consistency processing tool for categorical information of multi-source POI, which supports mapping and conversion of heterologous POI classified information and POI information extraction across web sites (Figure 9(a)). For instance, based on the classified information conversion and mapping
system, a total of 25438 agricultural trade markets around China were extracted from the resources database of 90 million POI, which were provided to relevant departments for the analysis of influenza transmission routes (Figure 9(b)).
CONCLUSION

After disaster, accurate, timely and reliable disaster information and available disaster relief resources information play a key role in rescue decision-making. How to use network geographic information to acquire relevant technologies and transform massive and heterogeneous information into decision making action information is particularly important. This paper has described current technology and our work on multi-source geo-spatial information acquisition and integration for disaster reduction. Experiments show that our method can be effectively applied to disaster emergency management and enhance disaster information service level. In the future research, we will face greater challenges in the field of information extraction, multi-source data Integration and application.

REFERENCES


