Applied Geography 43 (2013) 147–158

Contents lists available at SciVerse ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Web data mining for automatic inventory of geohazards at national scale[☆]

Alessandro Battistini*, Samuele Segoni[1], Goffredo Manzo[1], Filippo Catani[2], Nicola Casagli[3]

Department of Earth Sciences, University of Firenze, Via La Pira n°4, 50121 Firenze, Italy

Keywords:
Geotagging
News
Geodatabase
Geohazards
Inventories
WebGIS

A B S T R A C T

In this study, we present a fully automated procedure to analyze online news using data mining techniques. It is then used to compile and continually update a geohazard database. The procedure is based on new technologies that publish news on the internet, i.e., the news is analyzed, georeferenced and attributed to a category of geohazards (the current categories are landslides, floods and earthquakes). A continuous flow of georeferenced events is established to populate and update the geodatabase automatically and in near-real time. We tested the procedure for 2 years at a national scale, creating a geodatabase containing more than 20,000 news items concerning geohazards that occurred in Italy. This procedure enables continuous feedback from events in the real world, such that information about geohazards can be fully exploited rapidly (compared to traditional techniques based on remote sensing, field surveys and historical inventories).

Introduction

Geohazards, including earthquakes, floods and landslides, are a common cause of casualties and economical losses (Petley, 2012; Wirtz, Kron, Löw, & Steuer, 2012); therefore, significant efforts are always ongoing to mitigate related risks. When studying geohazards, especially at a regional or a national scale, the availability of archives and geodatabases (Papagiannaki, Lagouvardos, & Kotroni, 2013; Wirz et al., 2012), which can provide information about past and recent events (including magnitude, timing and location), is of primary importance.

In particular, the availability of updated and complete inventories is of paramount importance for hazard and risk assessment (Hilker, Badoux, & Hegg, 2009; Van Westen, Castellanos, & Kuriakose, 2009) and for the creation of early warning models (Brázdil, Kundzewicz, & Benito, 2006; Martelloni, Segoni, Fanti, & Catani, 2012; Montesarchio, Ridolfi, Russo, & Napolitano, 2011; Rossi, Catani, Leoni, Segoni, & Tofani, 2013). Unfortunately, one of the primary limitations of the existing archives and databases (especially for landslides and floods) is their update rate and methodology. They are usually compiled manually using field surveys (Brunsden, 1985; Santangelo, Cardinali, Rossi, Mondini, & Guzzetti, 2010) and, sometimes, remote sensing (Lu, Casagli, Catani, & Tofani, 2012; McKean & Roering, 2003; Segoni et al., 2009; Soeters & Van Westen, 1996). Systems using automated or real-time updates are still uncommon and only used for a few types of geohazards.

An earthquake is the geohazard that can rely on the most effective and fast geolocalization and characterization methods. A worldwide network of sensors and processing stations exists that is able to record and localize the occurrence of major and medium events, even at the global scale in near-real time (Wald et al., 2008). In addition, several national agencies provide the same information at a national scale, including minor events, in real time. Global and national scale information is easily accessible on the Internet or via specific applications that also provide notification services (Satriano, Wu, Zollo, & Kanamori, 2011; Wald et al., 2008; Wetmiller, Adams, & Woodgold, 2007).

* Corresponding author. Tel.: +39 055 275 7548; fax: +39 055 275 6296.

E-mail addresses: alessandro.battistini@gmail.com (A. Battistini), samuele.segoni@unifi.it (S. Segoni), goffredo.manzo@gmail.com (G. Manzo), filippo.catani@unifi.it (F. Catani), nicola.casagli@unifi.it (N. Casagli).

[1] Tel.: +39 055 275 7551; fax: +39 055 275 6296.

[2] Tel.: +39 055 275 7559; fax: +39 055 275 6296.

[3] Tel.: +39 055 275 7523; fax: +39 055 275 6296.

0143-6228/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apgeog.2013.06.012

A flood is a geohazard with an overwhelming state of the art (e.g., hydraulic studies performed on the Arno river (Italy) date back at least to the 12th century, as reported by Morelli, Segoni, Manzo, Ermini, & Catani, 2012). In addition, the location of the primary elements at risk and sources of hazards are usually well-known and documented (Cummings, Todhunter, & Rundquist, 2012; Hilker et al., 2009; Morelli et al., 2012; Salvati, Bianchi, Rossi, & Guzzetti, 2010; Varis, Kummu, & Salmivaara, 2012). Despite that, the study of floods and hydraulic risk in general requires the use of long time series (Ahilan, O'Sullivan, & Bruen, 2012; Gräler et al., 2013; Ricci, Piacentini, Thual, Le Pape, & Jonville, 2011). Most countries rely on a number of gage stations able to monitor water levels and fluvial discharges with high accuracy. Many national or regional hydrological services have been tracking these data for decades, or even centuries, providing the data for scientific study (Alfieri et al., 2013; Böhm & Wetzel, 2006; Brázdil et al., 2006; Schmocker-Fackel and Naef, 2010; Silva, Portela, & Naghettini, 2012; Zhou, Persaud, & Wang, 2006). In many countries, it is possible to retrieve and use detailed descriptions of flood events (at the basin or even on a national scale) that occurred prior to the establishment of runoff measurements (Barriendos et al., 2003; Barriendos & Rodrigo, 2006; Schmocker-Fackel & Naef, 2010; Thorndycraft et al., 2005).

Creating complete and updated databases is more difficult for landslides (Galli, Ardizzone, Cardinali, Guzzetti, & Reichenbach, 2008; Roth, 1983; Santangelo et al., 2010). For this geohazard, efforts are required for the development of models and their application, and efforts are also required for the collection of complete data sets to be used in the calibration and validation process. Therefore, the detection and mapping of landslides can often be the primary goal of a single research study (Tofani, Segoni, Agostini, Catani, & Casagli, 2013, and references therein). However, several global landslide databases are currently operative (GSC, 2011; Kirschbaum et al., 2009; Petley, 2012; USGS, 2011). Although they are useful tools for estimating the hazard and its impact on society, they are characterized by a significant degree of incompleteness because they primarily contain major events with catastrophic effects. At a national scale, several existing archives and geodatabases may have more detail and thus a higher degree of completeness and homogeneity (Dikau, Cavallin, & Jager, 1996; Hilker et al., 2009; Salvati et al., 2010; Trigila, Iadanza, & Spizzichino, 2010; Van Den Eeckhaut & Hervás, 2012). These instruments, although very useful for susceptibility assessments (Van Den Eeckhaut & Hervás, 2012), present some drawbacks, which prevent wider applications in landslide studies, i.e., they are not continually updated and seldom provide systematic information about the landslide time of occurrence (thus they cannot be useful for the temporal calibration/validation of forecasting models). The collection of recent landslide data can be a very demanding task, even at detailed scales, because it is usually accomplished using field surveys (Brunsden, 1985; Santangelo et al., 2010), remote sensing techniques (Lu et al., 2012; McKean & Roering, 2003; Segoni et al., 2009; Soeters & Van Westen, 1996) or manual retrieval of information from newspapers and technical reports (Brunetti et al., 2010; Hilker et al., 2009; Lagomarsino, Segoni, Fanti, & Catani, 2013; Rosi, Segoni, Catani, & Casagli, 2012), thus demanding a significant amount of time and human resources.

The primary goal of this study is to address this need with a data mining technique that automatically retrieves geohazard events from the Internet and organizes them in a continually self-updating geodatabase. The proposed approach is based on the concept that whenever a geohazard event produces a noticeable effect, the news is reported on the Internet. Therefore, Internet news can be considered a continuous stream of information, and appropriate data mining techniques can be set up to separate non-relevant information from pertinent items. Once the latter are identified by an automated procedure, each single event can be analyzed and stored in a geodatabase, together with related information (including a reference location and the time of occurrence). The proposed data mining procedure retrieves news in really simple syndication (RSS) format and analyzes it to identify the geohazard typology and its time of occurrence. In addition, the comparison with a geodatabase of toponyms is used to assess the spatial location of the events. The data mining procedure uses algorithms that are specifically calibrated for a single geohazard typology.

We tested the proposed data mining technique at the national scale, using Italy (approximately 300,000 km^2) as the test site and landslide as the primary target geohazard. After a test period of 2 years (November 2009–November 2011), a national scale geodatabase of approximately 5500 dated and georeferenced landslides was created automatically. These data can be used to update existing geodatabases or to calibrate or validate landslide forecasting models. This study focused on landslides because they represent the geohazard for which the creation of a model is most needed; our results could thus be more immediately and directly exploited. We also include the results of preliminary applications to earthquakes and floods to show that the proposed data mining technique can be applied (with appropriate specific adjustments) to almost any type of geohazard and to all news categories.

Materials and methods

Most of the news distributed via telematic channels is currently structured in RSS format (RSS Advisor Board, 2009) or in its analog, Atom (Nottingham & Sayre, 2005). RSS and Atom are XML-based formats that establish the distribution of documents via the web. They categorize the news and ensure that appropriate software can utilize news from different sources. This software, called a feed aggregator or feed reader, distributes news in a standardized format and in a unique environment. Actually, a feed is a summary of web news content along with links to their full versions. By subscribing to a particular web feed, it is possible to receive constantly updated summaries of the news collected from the site and, filtering them, to obtain only topics of interest (e.g., news about landslides and floods).

Unfortunately, the information structured in the RSS or Atom format does not have a native geographical location; the news itself is not associated with any structured geo-location. The RSS format was expanded in 2007, in its georeferenced version, adding a suitable geotag, GeoRSS (Geographically encoded objects for RSS feeds). According to Reed (2006), GeoRSS is a simple method that attaches a location, or "geotag", to an RSS feed. GeoRSS standardizes the way in which "where" is encoded with enough simplicity and descriptive power to satisfy the majority of needs and to describe the location of the Web content. With the geotag, a news item gains a reference to a specific location (area or geographical point) defined in the WGS84 coordinate system. News in this format is easily collected to be used in our database. Unfortunately, the GeoRSS format is not established as a standard, so networks rarely provide such information. In particular, in Italy and in the Italian language (in which this work was performed), it is not possible to find a news bulletin able to distribute news in the GeoRSS format.

The only way to obtain geolocalized news is to use an RSS-to-GeoRSS conversion service. This service, using data-mining procedures, compares the words in the news item to a database of locations to search for terms that may have a geographical location. The primary application (GeoNames, http://www.geonames.org) currently available online uses a limited set of locations (a few thousand for the entire country of Italy), resulting in poor localization accuracy.