Filtering the crowd: EDIT A New Methodology for the Treatment of Non-authoritative Data in the Reconstruction of Risk and Damage Scenarios

Author: Valentina Cerutti | Supervisors: Eng. S.Traverso, Dr. K. Stock, Prof. M. Jackson, Prof. L. Ferraris

University of Genoa – Master Degree in Environmental and Energy Engineering – A.Y. 2012-2013

Nottingham Geospatial Institute

Introduction

EDIT - Extracting Disaster Information from Text - is a methodology to threat textual information extracted from web, originated by non-expert users and referring to the occurrence of a disaster event.

The project is born from a collaboration between CIMA Foundation (Italy) and the Nottingham Geospatial Institute (NGI) (United Kingdom).

The main goal is the creation of disaster scenarios using information extracted from non-authoritative data, i.e. non-GIS data like texts, photos, social media messages.

The project represents a new contribute to the vision of scenarios and it is focused on the possibility of mosaicking real-time information from heterogeneous sources.

Components and information harvested are the ones most relevant in emergency management, during preparedness, response or post event phases, depending on the type of risk and expected damage.

EDIT

EDIT methodology focuses on the possibility of developing a multi-perspective scenario of damage or potential risk using data from heterogeneous sources, collected during a disaster. It designs a framework of application, rules, specifications on data, and create the representation of the footprint of an event, by integrating unstructured knowledge into operative tools. In this way, EDIT can be seen as a new approach to emergency response applications.

EDIT is structured in many components and uses an operational tool based on Java for analyzing texts through Natural Language Processing (NLP) techniques, semantics and ontologies.

The methodology allows to make a semantic analysis starting from information contained in the web and categorizes results into a PostgreSQL scenario database containing significant disaster-related information (location, time and impacts). Accessing it, users can immediately query and visualize information on maps, for example using a QGIS environment. Data are enriched with the attribute of a semantic reliability index, introduced as a further filter for information.

Main steps

• Critical analysis of non-authoritative data and web search

Effective emergency management requires access to tools that can process data and quickly produce maps to analyze impacted areas and effects: it is vital to read information in a common language and in a harmonized environment. When official data are not available, non-authoritative data and Big Data represent a valid and cheap alternative to obtain useful information for crisis management and build scenarios. Non-authoritative data can be produced by non-expert users, are not homogenous, are not validated and can be misused but they can be integrated with traditional sources, allowing the creation of high definition datasets with the collaboration of a huge number of users, and are timely and cost-effective.

The starting point of the methodology is a web research of online news and tweets using key words.

• Creation of a domain ontology about disaster and its consequences

The ontology allows focusing on the definition of a formal model for data processing and reasoning through a shared understanding of a domain of interest, which can be used to solve the problem of semantic heterogeneous descriptions of the same topic. A disaster ontology is defined.

• Creation of a concept map

In order to create the logical base of the procedure and to better analyze and evaluate concepts and relationships, the ontology has been graphically represented into a concept map (Fig.I). Four main characteristics have been identified to describe the disaster event: natural hazard, location, time and impact.

▪ Database creation

While the concept map represents the graphical illustration of the ontology, the database represents the formalization of the ontology. An entity-relationship data model has been chosen. The open source object-relational database management system PostgreSQL is chosen to realize the EDIT database for disaster representation: eighteen tables are created in order to contain all the relevant information; six of them are code lookup tables that contain glossary and thesaurus. The empty 'scenario tables-structure' can be partially or completely populated during an event. The meaningful elements are saved into the formal database structure that can be queried to obtain analysis results, which can be used for mapping, risk assessment, prediction, disaster real time response, damage evaluation, possible mitigation.

• Natural language processing and semantic analysis

Space-temporal information and impact related information can be extracted from text only after the application of standards and techniques that permit to rationalize and validate contents published online by non-expert users.

In order to extract relevant information from textual documents a Java code has been developed. It synthetizes the process of tokenization and bypass the POS tagging by comparing words contained in text with a list of keywords (the ones defined in the code lookup tables of EDIT database). The spatial information and the geo-localization has been realized by comparing sentences with data from OpenStreetMap (OSM) or other repositories of spatial data.

The semantic analysis is a fundamental step in the process of extraction relevant information from non-authoritative data: it is necessary to move from the natural language to a formal language in order to populate the scenario database.

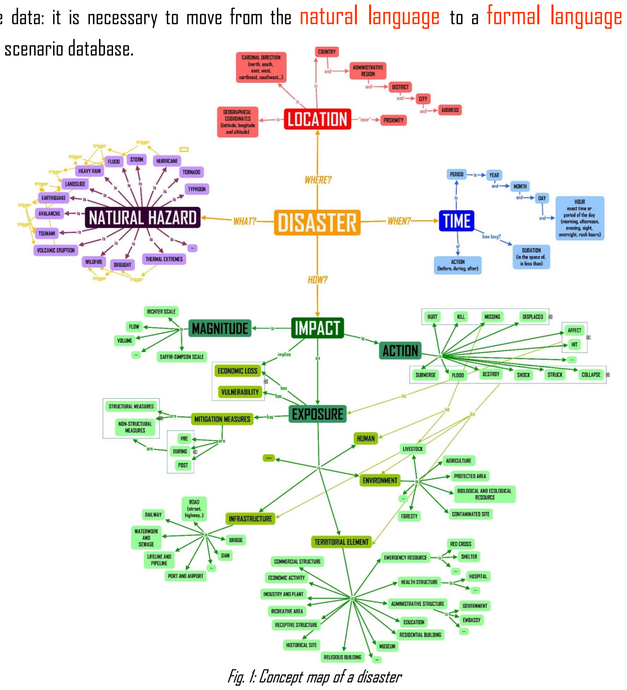

Fig. I: Concept map of a disaster

• Twitter

Online conversational text, exemplified by microblogs, chat, and text messages, is a challenge for natural language processing because it contains many nonstandard lexical items, unintentional errors, dialectal variation, etc. Applying EDIT Java code to the content of the tweet message it is possible to extract relevant news; in addition, selection of relevant words can be put in evidence considering ash tags. For researches on Twitter the integration with the free and open source Artificial Intelligence for Disaster Response (AIDR) tool (developed by Qatar Computing Research Institute) which leverages machine learning to automatically identify informative content on Twitter during disasters has been considered.

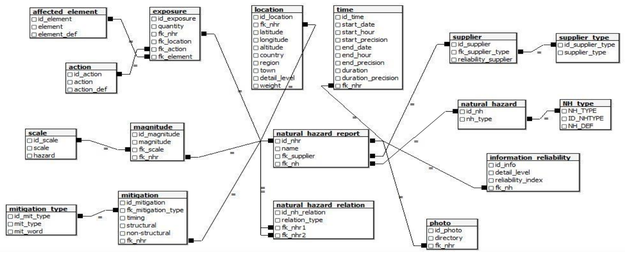

Fig.2: EDIT database for disaster representation

• Semantic reliability index definition

In order to use these information into a scenario, it is necessary the creation of a filter, both on the location – with the definition of a geographic location accuracy index - and on the content – with definition of a reliability index; after the filtering process, reliable information can be added for the construction of the scenario.

Some factors affect the quality of the information extraction, like the level of detail of the information, the accuracy and precision of the information, the correctness and completeness of a sentence, etc.

• Visualization of scenario maps

Representation of the scenario-tables of the database through disaster scenario maps in the QGIS environment reporting relevant features.

Application to two case studies

The first test zone is an example of historical research on a past event, the Christchurch Earthquake (New Zealand) in 2011. The use of semantic analysis on online news to extract relevant information to store in EDIT database has been performed using the English language.

The second case study is the flood of Secchia River (Italy) in 2014: near-real-time data reports from online news websites, tweets and photos have been investigated. The semantic analysis and the population of EDIT database has been realized using the manual and the automatic procedure, a translated version of database tables suitable for the Italian language.

The application to two different languages was intended to proof the flexibility of the model and has produced interesting results.

Innovative contributions and conclusions

The proposed topic is new and quite unstudied. The project represents the starting point of a new way of approaching disaster risk management and it is proposed as a research project with the hope to be implemented in the future on a large scale.

The innovative contribution of the project are:

- the use and validation of non-authoritative data;
- the conceptualization of a disaster event into a concept map;
- the integration of semantic and rule-based reasoning;
- the creation of a spatial database to handle and select the most relevant event-related information;
- the application of a filter of semantic reliability on data;
- the representation of these information into scenarios;
- The use of open source tools.

EDIT is a prototype and can be seen as a starting point for further developments; some aspects to be exploited are: an improvement of automatic research and storing processes, the widening to studies on crowd behavior, the definition of a location accuracy index based on geospatial criteria.