

Facilitating Situation Assessment through GIR with Multi-scale Open Source Web Documents

Brian M. Tomaszewski¹, Chi-Chun Pan², Prasenjit Mitra³, and Alan M. MacEachren¹

¹Department of Geography and GeoVISTA Center

²Department of Industrial and Manufacturing Engineering

³College of Information Science and Technology

The Pennsylvania State University

University Park, PA 16802

(1+) 814-865-4448

{bmt139, julianpan, pmitra, maceachren} @psu.edu

ABSTRACT

In this paper, we present our preliminary work on a Geographic Information Retrieval (GIR) system that utilizes loosely coupled web services and Google Earth™ (GE) to retrieve, extract, combine, and visualize situation information from multi-scale, open source web documents. Our intent with this work is to support situation assessment in the crisis management domain through tools that link and geographically contextualize information contained in text documents retrieved from multiple sources. In particular, our present work focuses on combining two data sources – The Federal Emergency Management Agency (FEMA) National Situation Updates and Google News™.

Categories and Subject Descriptors

H.3.3 [Information Storage And Retrieval]: Information Search and Retrieval – *Information filtering, Query formulation, Retrieval models, Search process, Selection process.*

General Terms

Algorithms, Design

Keywords

Open Source Information, Situation Assessment and Awareness, Crisis Management, Google Earth™, Web Services, Text Extraction

1. INTRODUCTION

Situation assessment is the process by which information on the state of the environment is acquired such that situation awareness, or the comprehension of the state of the environment within a given time/space extent, may be achieved [2]. For application domains such as crisis management, geospatial information that may be of potential use in situation assessment is contained in textual references within web-based, open source documents such

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Workshop on Geographic Information Retrieval (GIR'07), November 9, 2007, Lisboa, Portugal.

Copyright 2007 ACM 1-58113-000-0/00/0004...\$5.00.

as news stories and government-issued reports. The operational jurisdiction of the entity publishing the documents may create an implicit geographic scale, scope and extent to the information contained within the document. For example, a local news media outlet will more likely report on situations relevant to its locality versus a federal agency that reports on situations at state-level or regional scales and does not provide as much localized geographic detail for a situation.

Although individual documents from disparate sources may be limited in geographic extent or geographic detail, when combined, they can provide a more comprehensive picture of a situation across geographic scales and lead to improved situation assessment. It is the combination of these types of information sources that motivates our present work. We present our preliminary work on a Geographic Information Retrieval (GIR) system that utilizes loosely coupled web services to retrieve, extract, and combine situation and geographically contextualized information from multi-scale, open source web documents for situation assessment in the crisis management domain.

2. DATA SOURCES

Our present work focuses on combining two data sources – The Federal Emergency Management Agency (FEMA) National Situation Updates¹ (FEMA N.S.U.) and Google News™² (GN). FEMA N.S.U. provide a daily overview of various hazards, disaster, recovery and other information relevant to crisis management planning and operational activities within the United States (U.S.). The geographic scale of entities discussed in the FEMA N.S.U. reports tends to not go beyond the County level with entities at the U.S. state or regional scale being the most common. GN is a computer generated news aggregation service that uses the powerful search and retrieval engines of Google to gather news articles from around the world. The scale of geographic entities within stories found by GN varies dramatically.

3. PROCESS AND TOOLS

Conceptually, we support situation assessment using an approach based on the data sources described in section 2. FEMA N.S.U. are first used to give a state-level assessment of situations within the U.S. Based on the situations found, GN is then used to retrieve news stories about those situations in order to develop a more

¹ <http://www.fema.gov/emergency/reports/index.shtm>

² <http://news.google.com/>

localized perspective on the situation. We implement our approach to situation assessment by combining the functionality of two GIR tools that are loosely coupled through web services and utilize GE for geographic data presentation and visualization—FEMARepViz and the Context Discovery Application (CDA) [4].

3.1 FEMARepViz

The process begins with FEMARepViz. FEMARepViz is a visualization generation web service tailor-made for the FEMA N.S.U.. FEMARepViz uses FactXtractor, a named entity and entity relation extractor, to identify location names in FEMA N.S.U reports and the relationships between those locations. Each report is segmented into individual incidents and classified into 9 pre-defined categories using an *n*-gram language model. Note that each incident may contain multiple location names. We assign incidents and locations as (incident, location) pairs.

Processed reports are stored in a repository and can be retrieved by a web interface. The output visualization of processed report data is a KML document that provides dynamic updates, interactive visualization, and a *query CDA* link. Users can click on the query CDA link to retrieve local news stories through GN that are relevant to the incident that was extracted by FEMARepViz.

3.2 The Context Discovery Application

The CDA performs an automated retrieval of news stories using Google News™ RSS feeds configured using situational and geographic scope information encoded by FEMARepViz. Once the news stories are retrieved, the CDA performs geocoding and visualization of geographic place names and possible relationships between places across user-defined geographic scales over time from within the stories retrieved. The CDA performs geocoding using a custom pattern matching algorithm and the Generic Architecture for Text Engineering (GATE) [1] program and attempts to construct “geophrases” that capture near-context word information around location words for disambiguation, an approach similar to that discussed in [3].

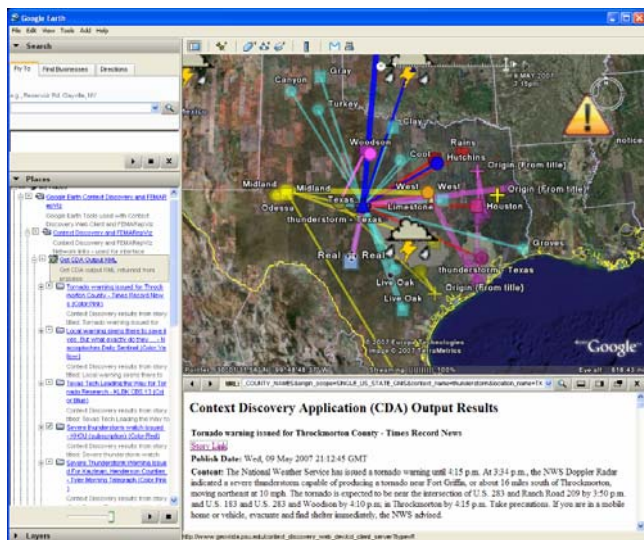


Figure 1: FEMARepViz and CDA output visualizations shown in Google Earth™. Visual representations of locations extracted from news stories about thunderstorms in Texas are shown in this figure.

3.3 The FEMARepViz/CDA Visualization

Figure 1 shows the results of FEMARepViz and CDA processes as rendered in Google Earth™. FEMARepViz organizes incidents found in the FEMA N.S.U. by the general category to which the incident is assigned and uses pictographic point symbols to represent the category of hazard or emergency found. In Figure 1, thunderstorms in Texas are being investigated.

The CDA creates a spider graph of the news stories found based on the type of incident and location of the incident as derived from FEMARepViz. As shown in Figure 1, each location found in the CDA news search is connected by a line to the origin point derived by the CDA for the news story. Line thickness indicates the number of times a place was referenced in the news story and line transparency indicates how old the story is relative to the time when the geovisualization was created. Point symbol size represents the geographic scale of the entity found (town, county etc.). Users can also access contents of the story retrieved.

4. FUTURE WORK

Future work will include the spatiotemporal indexing of situational information retrieved based on the category of incident. The goal will be to support development of geotemporal context for present situations, for example emergency planners who wish to review past reactions and responses to wildfires.

5. CONCLUSIONS

Muti-scale, open-source web documents provide an important source of text-based, geospatial situational information. By combining GIR tools, loosely coupled through web services, that can capture, geocode, and integrate diverse situational information contained in open source documents, processes of situation assessment in application domains such as crisis management can be better informed and capable of exploiting diverse information sources.

6. ACKNOWLEDGMENTS

The research reported here has been supported by the National Science Foundation under Grant EIA-0306845. This work is also supported by the National Visualization and Analytics Center, a U.S. Department of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL). PNNL is a U.S. Department of Energy Office of Science laboratory. The authors thank Google™ for permission to use the Figure 1 image.

7. REFERENCES

- [1] Cunningham, H. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36 (2). 223-254 (2002).
- [2] Endsley, M.R. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37 (1). 32-64 (1995).
- [3] Li, H., Srihari, R., Niu, C. and Li, W. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*. 39-44 (2003).
- [4] Tomaszewski, B., Mapping Open-Source Information to Support Crisis Management. in *First Annual DHS University Network Summit on Research and Education*, (Washington, D.C., 2007).