CTRnet DL for Disaster Information Services

Seungwon Yang[1], Andrea Kavanaugh[1], Nádia P. Kozievitch[4], Lin Tzy Li[1,4,5], Venkat Srinivasan[1]
Steven D. Sheetz[2], Travis Whalen[3], Donald Shoemaker[3], Ricardo da S. Torres[4], Edward A. Fox[1]

[1]Computer Science,
[2]Accounting and Information Systems, [3]Sociology,
Virginia Tech, VA, USA 24060

[4]Institute of Computing,
University of Campinas,
Campinas, SP, Brazil 13083-852

[5]Telecommun. Res. & Dev. Center, CPqD Foundation,
Campinas, SP, Brazil, 13086-902

{seungwon, kavan, svenkat, sheetz, shoemake, tfw115, fox}@vt.edu,
{nadiapk, lintzyli, rtorres}@ic.unicamp.br

ABSTRACT

We describe our work in collecting, analyzing and visualizing online information (e.g., Web documents, images, tweets), which are to be maintained by the Crisis, Tragedy and Recovery Network (CTRnet) digital library. We have been collecting resources about disaster events, as well as campus and other major shooting events, in collaboration with the Internet Archive (IA). Social media data (e.g., tweets, Facebook data) also have been collected and analyzed. Analyzed results are visualized using graphs and tag clouds. Exploratory content-based image retrieval has been applied in one of our image collections. We explain our CTR ontology development methodology and collaboration with Arlington County, VA and IBM, in a Center for Community Security and Resilience funded project.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collections.*

General Terms: Management, Measurement, Documentation, Design, Human Factors.

Keywords: Digital libraries, natural disasters, man-made disasters, tweets, Crisis Informatics, ontology.

1. INTRODUCTION

The inception of the CTRnet DL began as an extension of our prior work, The 4/16 Digital Library, which collected data and provided services relating to the 4/16/2007 campus shooting at Virginia Tech [3]. A goal of CTRnet is to develop integrative approaches to collect, analyze, and visualize (present), under a DL environment, so that the DL system can archive resources and provide services to its stakeholders efficiently and effectively. Our proposed CTRnet system architecture includes multiple modules to collect, analyze and visualize. We have been exploring each of them using our developed modules (e.g., Facebook app, scripts for tweets, CBIR module, Brainstorming tool, etc.) as well as existing online software tools (e.g., Heritrix crawler, The Desktop Archivist, 140kit.com) for various types of content as shown in Table 1.

An overview of the work procedures in Table 1 is explained in the following sections. In addition, we present our current work on developing the CTR ontology development.

Table 1. Collect, analyze and visualize sequences, with content and technologies used.

	Collect	Analyze	Visualize
Content	Web sites, images	Image similarity	Organize images by similarity
	Tweets	Content, user profiles	Patterns, frequencies
	Facebook content	Usage of social media (SM)	SM use
	Focus group interviews/surveys	Usage of SM	SM use/needs
Technology	Crawler	CBIR algorithm	CBIR visualization interface
	Online tools, scripts, APIs	NLP toolkit, SQL	Graphics
	Facebook app	Spreadsheets	
	Brainstorming tool	Brainstorming tool	

2. COLLECTIONS

The CTRnet Web resource collections include content from natural disasters, shooting events, and their anniversary/ remembrance observations (as a way of coping with the tragic event). We collaborated with the Internet Archive (IA), a non-profit organization devoted to preserving Internet resources, so that those resources could be archived forever. The Heritrix crawler developed by IA has been used to gather materials online. Table 2 summarizes our developed collections as well as other related ones.

It is necessary to develop a list of seed URLs for Heritrix to crawl. To reduce manual effort/time for this task, we are devising methods to generate seeds in an automated fashion so that bigger collections can be built more quickly, supported by our ontology.

Microblogging (e.g., Twitter) data has been collected (Table 3). We used online tools such as The Archivist[1] and 140kit.com. To collect from specific Twitter IDs, we developed PHP scripts with database tables and a Twitter API.

JCDL'11, June 13-17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

[1] http://visitmix.com/labs/archivist-desktop/

Table 2. Natural disasters and shootings collections in IA[2].

Event type	# Collections
Natural Disasters (e.g., earthquakes, volcanic eruptions, floods, wildfires, tsunami)	12
School Shootings – USA	4
Remembrances (VT April 16, Haiti)	2
School Shootings – International	1
Tucson, AZ Shooting	1
Political Crises (e.g., Egypt, Tunisia)	5

Table 3. Tweets of disasters, shootings, and political crises.

Topics	Status
UT Austin campus shooting	Completed
34 civic organizations in Arlington County, VA	Completed
Tucson, AZ Shooting	Completed
Politicians tweets following Tucson shooting	Completed
Cyclone Yasi in Australia	Completed
Protests in the Middle East	In Process
Japan earthquake, tsunami, nuclear radiation	In Process

3. ANALYSIS EXAMPLES

The Natural Language Tool Kit[3] has been used to find word collocations and frequencies of tweets. We will apply its entity extraction in our further analysis of tweets. SQL queries were useful for finding the tweeting patterns per day (and per hour) during and after events [4].

Content-based image retrieval (CBIR) in digital libraries is important for retrieving multimedia information. A CBIR module takes a CTR image as input and attempts to find similar images from the image collection. An independent study [1] explored CBIR concepts along with the Eva [2] tool. The dataset included a list of 111 pictures, representing different areas affected by the earthquake in Haiti in January 2010.

4. VISUALIZATIONS

We applied CBIR technology to organize and visualize images based on their feature similarities (Figure 1). Tag clouds were useful in showing multiple popular terms extracted from tweet content. Figure 2 shows a tag cloud of member profiles from ArlingtonUW (i.e., Arlington Unwired). It is useful in showing a snapshot of popular words from tweets.

5. CTR ONTOLOGY

One capability common to many ontology development efforts is to describe data from diverse sources. Thus, we began our ontology development process by identifying several existing databases currently tracking disasters and derived the "ontology in situ" of their database (Figure 3).

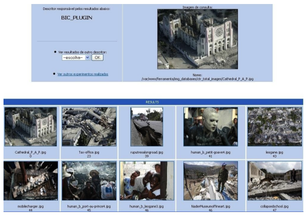

Figure 1. Image ranking by the BIC descriptor.

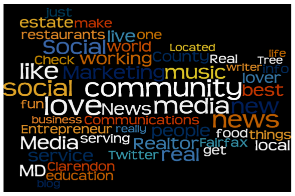
Just
estate make
restaurants live one
Social world
Located
check working County Real Tree life
like Marketing music writer info
lover
social community best
fun love News media new
business Communications news
Entrepreneur really people food things
Media serving Realtor Fairfax local
service Twitter real get
MD Clarendon education
blog

Figure 2. Tag cloud of tweets from an organization in Arlington, VA.

The resulting ontology consists of 185 elements and has the potential to support data sharing/aggregation across the databases considered. We are expanding this ontology to include other aspects of disasters such as recovery, preparedness, and mitigation.

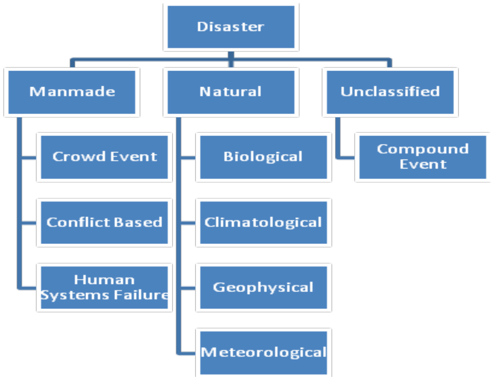

Figure 3. Top level concepts in the CTR ontology.

6. ACKNOWLEDGEMENTS

We thank NSF for support (IIS-0916733) and the Internet Archive, Arlington County, VA and IBM for their collaboration. We also thank our sponsors CAPES (BEX 1385/10-0), FAPESP, FAEPEX, and CNPq.

7. REFERENCES

[1] N. P. Kozievitch, et al. *Exploring CBIR concepts in the CTRnet Project*. Technical Report IC-10-32, Institute of Computing, University of Campinas, November 2010.

[2] O. A. Penatti and R. da Silva Torres. Eva: an evaluation tool for comparing descriptors in content-based image retrieval tasks. In *Proceedings of MIR '10*, pages 413–416, ACM, New York, NY, USA, 2010.

[3] E. A. Fox, et al. A digital library for recovery, research, and learning from April 16, 2007 at Virginia Tech. *Traumatology*, 14(1): 64–84, Mar. 2008.

[4] L. T. Li, et al. Twitter Use During an Emergency Event: the Case of UT Austin Shooting. In *Proceedings of the 12th Annual International Conference on Digital Government Research (dg.o 2011)*. Digital Government Society of North America. (to appear)

[2] http://archive-it.org/public/topic.html?topic=spontaneousEvents

[3] http://www.nltk.org/