NSF 2nd Year Report

CTRnet: Integrated Digital Library Support for Crisis, Tragedy, and Recovery

July 2011

1. Participants

1.1. Project Members

- Principal Investigator: Edward A. Fox
- Co-Principal Investigators: Donald J. Shoemaker, Steven D. Sheetz, Andrea L. Kavanaugh, and Naren Ramakrishnan
- GRAs: Venkat Srinivasan, Seungwon Yang
- Visiting Faculty: Riham Abdel Moneim
- Graduate Students: Tram Bethea, Bidisha Dewanjee, Yipan Deng, Seth Fox, Shamimul Hasan, Nádia P. Kozievitch, Lin Tzy Li, Min Li, Mohamed Magdy, Chao Peng, Chet Rosson, Travis F. Whalen
- Undergraduate Students: Bernadel Benoit, Jason Browning, Mario Calixte, Sherley Codio, Jennifer Francois, Jason Heim, Robert Leith, Fabrice Marcelin, Ashley Phelps, Jason Smith, Justin Tillar, Keith Wooldridge

1.2. Other Collaborators

- Hicham Elmongui, Alexandria University
- Apostol Natsev, IBM Watson Laboratory
- Lexing Xie, IBM Watson Laboratory

1.3. Other Organizations Involved as Collaborators

- Internet Archive (U.S.A.)
- IBM Watson Research Laboratory (U.S.A.)
- Monterrey Tech (Mexico)
- Arlington County (Virginia, U.S.A)

2. Activities and Findings

2.1. Research Activities and Findings

2.1.1. Building Collections for Crisis Events

Online News Collections
One of the important parts of this project is to collect and archive as much information as possible about various events that are related to crises, tragedies, and recovery (CTR). In order to do long-term archiving of information, we have worked with the Internet Archive (IA), a non-profit organization, whose goal is to archive the Internet. IA provides access to web crawlers that can be used to selectively crawl and archive webpages.

So far we have developed techniques for automatically identifying seeds (i.e., starting web pages), for performing the crawls. As soon as a CTR event is identified, using keywords specific to that event, we query online news sources, use the top results as seeds for IA crawlers, and launch a crawl. We have built several CTR collections using this approach. Another approach was to extract URLs from tweets that are relevant to CTR events, and then use those URLs as seeds. Our collections are summarized in Table 3 in Section 3.2.2.

The collections obtained from IA are in a compressed file format (.warc or .arc). These files cannot be added as-is to the digital library. We have developed (in some cases modified) parsers for extracting the constituent webpages inside these highly compressed files, so that individual HTML files and multimedia data that are embedded on Web sites can be easily accessed.

Tweet Collections
In disaster situations, it is well known that people use micro-blogging sites such as Twitter to reach their family and friends especially when their cell phones are not working due to high volume of traffic on the cell phone network. For this reason, tweet posts sometimes report CTR events faster than the mainstream news media. Those tweets often contain more detailed information, too, reported by the affected people on the site.

We have been archiving tweets (i.e., posts from Twitter.com) for both man-made and natural disaster events. Table 4 in Section 3.2.2 shows a list of tweet archives and the number of tweets in each archive. Collected tweets can be exported in various formats including .xsl, JSON, and HTML -- to be analyzed later using software tools.

2.1.2. Focused Crawling and Filtering

The collections that we have built using IA crawlers are incomplete. In other words, not all of the relevant content is located. One reason for this is that each set of seeds we have identified is limited in scope and size. Another reason is that the crawling process, which follows links from the seeds, and links from linked pages, etc., also is limited. Accordingly, we work to improve our identification of seeds. Further, to improve the crawling process we are developing improved focused crawlers, that can do a better job as they consider the seeds as starting points, and crawl the webpages that the seeds link to.

Another problem with our collections is that many of the pages crawled are not relevant. In other words, the precision of the process is low. Accordingly, we are working on techniques to classify which of the pages found are likely to be relevant.

In order to accomplish this, we are developing supervised machine learning techniques. We are using a modified version of the LibSVM classifier as provided by the data mining package WEKA. We first train a one class (LibSVM) classifier using the seed webpages as positive training examples. Once the classifier has been trained to give satisfactory performance on the training data, we will incorporate it into our focused crawler. Thus far we have been able to achieve a precision of 70% and a recall of 81% on 10-fold cross validation.

This classifier can be used in two ways. One way is inside a focused crawler. When a link is identified that can be used to further extend the crawl, it is helpful to determine if that link should be followed. A second way to use the classifier is for filtering the already crawled collections. We observed that many non-relevant pages tend to get picked up while crawling, so it is important to filter them out before adding the crawled data to our digital library. Once we have built our classifier, we will apply it to these collections, aiming to filter out non-relevant webpages.

2.1.3. Mapping CTR Collections

Location or geographic information is one of the inherent characteristics of CTR events, therefore it makes sense to show the CTR collections positioned in a map. From there people can browse to documents/images about events that have occurred at a given place. For example, after a user has selected an event of interest on a map, the system can provide a link to a service where users can find documents (be it text, news, images, videos, etc.) about the event. Users also can follow links to other events connected with that place, at different dates.

We have prototyped this browsing method using the Google maps API. We are studying Drupal's module to integrate such maps with our current website. A proposal is to geo-parse and geocode our collection so we can analyze them under a location / geographic point of view. Geo-parsing is a process of recognizing references to locations inside documents and ignoring false references (e.g., place names that also are the name of an organization or person). Geocoding is a process to associate a document with some specific latitude and longitude, often based on locations recognized by geo-parsing. Once collections are geocoded, they can be placed on maps, analyzed, and searched by geographic location.

Our prototype CTR event map is presented in Section 3.2.4.

2.1.4. Twitter Use During an Emergency Event: the Case of the University of Texas, Austin Shooting

One topic studied in this project is the use of social media by local governments to respond to emergency events in towns and counties. Monitoring social media information for unusual behavior can help identify these events once we can characterize their patterns. In one example, we analyzed the campus shooting at the University of Texas, Austin, on September 28, 2010. In order to study the pattern of communication and the information communicated using social media on that day, we collected publicly available data from Twitter.

There were three phases in our procedure to prepare and analyze the dataset. In Phase I, we collected information about the followers of UTAustin. In Phase II, we crawled tweets that had a time stamp of Sept. 19, 2010 or later by using the follower information from Phase I. In Phase III, we analyzed the dataset using MySQL queries and the Natural Language Toolkit (NLTK). Specifically, NLTK's feature to find frequently collocated word pairs was helpful. The results were then visualized using multiple word clouds and a bar graph.

Analyzing the number of posts per day, we noted that for the day of the event (Sept. 28), there was a peak of over 15,000 posts, while for the other days the maximum number of posts was around 6,000 to 10,000. Considering the most common words in Twitter posts of Sept. 28, we found that words such as ``shooter'', ``gunman'', ``shooting'', ``utshooting'', ``suspect'', and ``university'' -- besides ``campus'', ``UT'', ``Austin'', and ``RT'' (which stands for retweets) -- were the most frequent words.

Comparing the distribution of number of posts over time for Sept. 28 to the day before and after it, the Twitter posts nearly doubled to 900 per hour at 8 AM as compared to other days when by this time it would be below 500 posts. The peak of Sept. 28 was at 9 AM with 2,623 Twitter posts, while on other days it would be around 600 by that time. Thus, we observed an increase of over 400%.

Users mostly tweeted about the shooting event for 7 hours after it happened. At 7 AM there was nothing related to it, but by 8 AM, around when the event happened, the words ``UT'', ``campus'', ``gunman'' and ``shooter'' were among the most used words found on tweet posts. The UT shooting dominated the Twitter posting of this community until 3 PM, when the most visible words were the shooter's name at UT Austin (Colton Tooley). From this study we observed that during crises people used Twitter to share and to comment on information about the event. We find that a spike in the number of tweets, and changes in the ideas in the tweets, signal that a significant event is occurring or has recently occurred.

2.1.5. Research Related to School Shootings

A number of events of interest relate to shootings in schools. To help researchers understand some aspects of this type of tragedy, we have worked to identify key related concepts, and to determine how those concepts are related. One partial goal is the development of a concept map, connecting personal and social experiences to school shootings. In particular, we have focused on the connection between immigration, school

bullying, and other forms of social ostracism and/or victimization in the school setting, with school violence, especially school shootings. In addition, we have developed a small bibliography of published research on the relationship between immigration, social isolation, school bullying and other victimization, and violent outcomes, such as school shootings. We also are developing a list of journals in sociology, psychology, criminology, and youth-related subjects, in an effort to broaden our literature search. We plan to connect this work with research on text analysis and text mining that was outlined in our NSF proposal, using a technique called story telling.

2.1.6. Visualization of the Social Media Content

As shown in Table 4, the team has been archiving tweets that are related to the natural and man-made disasters by using an open source tool, *yourTwapperKeeper* (http://your.twapperkeeper.com/). To visualize the current key terms from recently posted tweets, we have been building new word clouds, every 10 minutes, for the Japan earthquake disaster as well as the Libya revolution.

Raw tweets that were collected in the last 10 minutes are extracted from our database, which contains archives of those two events. Those tweets are then preprocessed to remove stopwords (e.g., is, a, the, about, it, etc.) and less meaningful words (e.g., 'Japan', 'earthquake', 'Libya', 'revolution', etc. since we already know they are about the Japan earthquake and Libya events). By using a script developed using the Processing Language (http://www.processing.org/) and a WordCram library (http://wordcram.org/), the preprocessed tweets are visualized as word clouds. This methodology allows viewers to see updated information on the clouds every 10 minutes. See more about this in Section 3.3.4.

2.1.7. Egypt Study

In this study, entitled "The Use of Different Media Sources during Street Demonstrations in Egypt: March- February 2011," we investigated the use of various communications media including television, radio, face to face with family/friends, face to face with people in the street, cell phone, internet, and social networking sites during the uprisings in Egypt this year. In collaboration with Dr. Riham Moneim and Professor Hicham from Alexandria University in Egypt, we used an online survey in Arabic to survey university students.

A total of 256 subjects completed the survey. The diversity of information sources studied showed varying levels of perceived accuracy and reliability across sources. Data analyses are in progress to identify key findings.
Our intent is to supplement social media we have collected during the period February-March 2011 (e.g., especially twitter data) with the Egyptian citizens' perspective of the overall information sources available (i.e., television, radio, face-to-face, telephone, mobile phone, Facebook, Twitter, YouTube). We want to put the use of social media in perspective in terms of its use relative to other sources of information. Questions addressing the following issues were asked: Who was using social media? What sources

of information did you use during the street demonstrations? Were you able to get the information that you needed? Which sources did you trust to be most reliable?

In addition, we are currently conducting face to face interviews with Egyptians that are not associated with universities. These interviews are expected to be completed in August.

2.1.8. Arlington County Research

Citizens are increasingly using social media for interpersonal and group interaction and for communication with community leaders, including government. Twitter, Facebook, Flickr, YouTube, and other services managing user-generated content have made a staggering amount of information available. Government officials seek to leverage these resources to improve services and communication with citizens. Yet, the sheer volume of social data streams generates noise that must be filtered, and calls for analysis to identify issues of concern, as well as patterns or trends in information flow. Mining the diverse real-time feeds of social streams related to concerns or events can enable government to act more effectively on matters both routine (e.g., ongoing issues of public concern) and critical (e.g., major weather or traffic disruption, public safety, or rapid response in emergencies).

This project leverages technology to help governments manage information and facilitate citizen interaction in meaningful ways in order to achieve broader public participation than is possible through normal channels (e.g., public commenting at county board meetings). Deep analysis of social media streams also can aid in providing access to segments of the community that have not participated in traditional ways.

We investigated methods to utilize social media sources to meet a variety of city, county, and community needs. This research included: 1) leveraging and further developing a platform for collecting large amounts of public information relevant to cities, counties, or communities; 2) archiving collected social media data over a period of time into a digital library; 3) correlating multiple information sources; 4) studying analytic applications for diverse services; 5) extending technology to perform analytics on the image content of social media streams; 6) conducting focus groups with government employees and citizens to identify their issues related to using social media.

Two examples of results from this project include a content analysis of postings on the county Facebook page and the categories of social networking issues identified by participants in the focus groups.

Arlington County government has maintained a Facebook page since early 2010 (http://www.facebook.com/ArlingtonVA). The page had roughly 4500 fans at the end of September 2010 (by February 2011, there were close to 4900 fans). We analyzed a two-month period (August -September, 2010) of posts by the County and responses (comments) from the public by conducting a simple content analysis by topic. There were a total of 112 posts; the top 10 most frequent topics are shown in Figure 1.

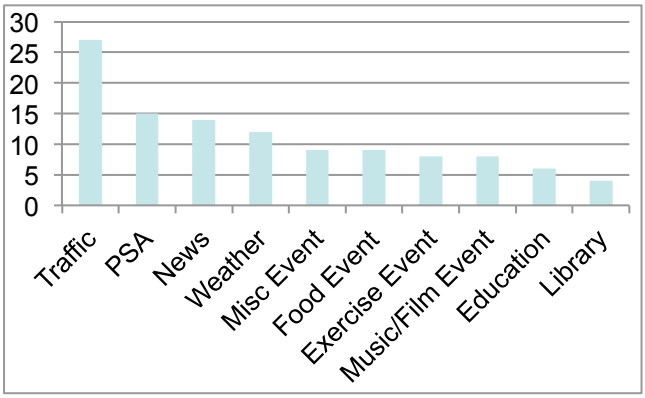

Figure 1. Top Ten Topics in Arlington County Facebook.

The most common posts by the County on the Facebook page were about traffic (e.g., conditions, closures, metro outages), followed by public service announcements (PSA). News (shout-outs, updates, and other County announcements) and weather related posts (National Weather Service and Arlington Weather Service advisories) were followed by various events (good, biking, walking, music or film) in terms of frequency of posts. There were only a few posts related to education (Arlington County School District) and library services (e.g., closures, speakers, special activities) during this two month period.

There were a total of 824 public comments to the County posts during this two-month period. Half of the comments pertained to about a fifth (19%) of the County posts (the top 21 posts by the County). Figure 2 shows the distribution of the bulk of the comments on the same top 10 County posts seen in Figure 1.

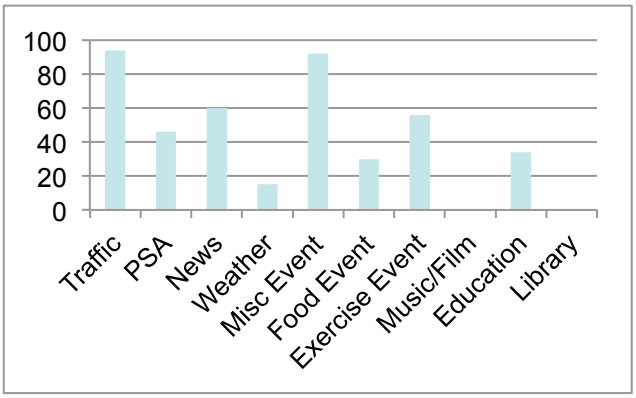

Figure 2. Public Comments by Arlington Facebook Topic.

The comments are predominantly related to traffic and miscellaneous events (that is, events that do not fall into the other 'event' categories shown, such as food, exercise, music and film). Exercise events (bikes, walks) and News announcements generated the next most frequent number of comments from the public. Almost all the comments were highly consistent with the social media policy of the County (e.g., no profanity or off topic comments) and were overwhelmingly positive in tone, including many "Likes" hits.

We also conducted three focus group sessions consisting of two steps, beginning with the participants engaged in electronic brainstorming to generate a substantive number of ideas quickly, followed by their identifying categories that grouped the ideas by similarity.

Using group support software, the participants anonymously generate and enter ideas, beliefs, issues, or concepts, in the form of short sentences or phrases that they feel are important to the situation. We developed and used a set of framing questions (shown in Table 1 to cue participants to begin entering ideas. Participants' responses to these questions are visible on each participant's computer screen as the responses are generated, allowing ideas generated by one person to be expanded by others or to cue others to generate related ideas. As a single group interview participants aided by a facilitator discuss, create and name the meaning of response items or categories in order to organize their ideas by similarity.

Table 1. Framing questions used in previous focus groups.

- What are the missions and objectives of your organization?
- What are you trying to accomplish using social media?
  - Do you feel you are currently accomplishing this goal effectively with social media? (if yes, why?)
  - If not, what do you need [to know? – to do? --in order] to use social media more effectively?
- What concerns do you have about using social media?
- What difficulties do you have about using social media?
- What information would you like to have about how your organization uses social media?
- What information would you like to have about how social media is being used in your community?
- Is there anything else you would like to know about social media that would be helpful?

The categories generated by the focus groups are merged based on the ideas that participants group together by similarity into the taxonomy started in the previous groups. Participants in our three focus group sessions identified shows the 23 categories of ideas (shown in Table 2). The categories consist of factors related to the organization and factors related to the information exchanged between the organization and the community.

Table 2. Categories of issues related to using social media in county government.

- ❖ Communication Issues
  - ➢ Community Outreach (emergency, crime/traffic alerts, 24/7 level of service, recruitment)
  - ➢ Feedback (from community to organization, social trends, locale, fast spreading ideas)
  - ➢ Population Reached (misses traditional/older population or can't afford technology)

- One Way vs. Two Way (pushing out vs. creating dialogue, effort/costs different)
- Tone (Government presents just the facts, not stories, not press release, listen then education)

- Cost Issues
  - Resource Issues (SM adds to previously full time job, other duties, limit 24/7 expectation)
  - ROI/cost to value (how to measure value, who are we reaching, enough received messages)

- Information Issues
  - Quality of Content (accuracy, facts of situation, unvetted information, misinformation)
  - Quantity of messages (how to be heard, from 1 to 10 to 1000s, overwhelming, loss of control)
  - Personal Level (information overload, ability to write complete thoughts, nuances of FTF lost)

- Legal Issues
  - Data Maintenance (FOIA data maintenance and related costs)
  - Owing Vs Using Someone Else (official outlet versus imposter, use in investigations.
  - Public Record/FOIA (are SM public record, tools needed to save, outdated polices)

- Policy Issues
  - Management Buy-In (unknown expectations, under valued, set culture)
  - Control Issues (how much to control, what we can control, telling how/what to think/do)
  - HR Components (job descriptions, evaluation, expertise, dialogue, positive and negative)
  - SM Communications Policy (what not to do/say, right people to make SOP, moving target)
  - Professional Level (privacy concerns, devices owned by county, investigative purposes)

- Technology Issues
  - Security (network exposed to world)
  - Technology and Equipment (cost of technology and maintenance, cost savings, training)
  - What SM Outlets (knowing audience/expertise, users expect transparency, so many outlets)

- Training and Education
  - Education (tools to manage, learning from each other, train constituents where

to go)
- Training (best practices for dividing duties, case studies, understanding management's concerns)
- Other (educate nonusers, establish boundaries)

2.1.9. External Advisory Board Meeting

There was an external advisory board meeting in May 12, 2011. Two board members, Kristine Hannah (Internet Archive) and Susan Metros (University of Southern California), participated in the meeting this time. The topics discussed include:

- Upgrading the CTR ontology
- Social media archiving in IA
- Managing the overflow of information
- Comments regarding the project homepage
- Development regarding the CTR toolkit
- Engaging with national agencies that are related to emergency response
- Needs of sociological and media theories

We will have meetings with the other members of the external advisory board, too.

2.1.10. Local Advisory Board Meeting

As part of our planned meetings with members of the local advisory board, on June 21, 2011, the project team met with Erv Blythe, who is Vice President for Information Technology and Chief Information Officer at Virginia Tech. A number of topics were discussed. They include:

- Geo-spatial information access during the crisis situations
- Identity management resources
- Crawling – increasing the accuracy of the crawl
- Threat assessment and mitigation tools
- Similarities and differences with other projects on campus
- Disaster mitigation plans for regions

2.2. Training and Development Activities

In the fall of 2010, in CS5604, Information Storage and Retrieval, and in the spring of 2011, in CS4624, Multimedia, Hypertext, and Information Access, students were briefed about the CTRnet project and invited to attend project meetings.

Students listed as participants in this project were all involved in meetings and related training and development activities.

For example, a Virginia Tech graduate student in Egypt, Noha Elsherbiny, enrolled in our VT-MENA program, worked with PI Fox on digital library security, motivated by the needs of CTRnet, and considering our requirements as specified by the project team, including Professor Shoemaker. This led to a book chapter and to an MS thesis.

2.3. Outreach Activities

We have multiple activities including the poster and paper presentations. In March 25, the project team presented a poster in the department open house to publicize our efforts regarding the project. Visitors from companies and on-campus institutions such as Virginia Bioinformatics Institute (VBI) showed interest in our project.

In addition to this, three posters have been presented, at the Joint Conference on Digital Libraries (JCDL 2011), Digital Government Conference (dg.o 2011) and the International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011), respectively. The team also conducted a tutorial session for using open source tools to collect and visualize tweets, and a paper presentation, at dg.o 2011 conference.

3. Publications and Products

3.1. Publications During Year 2

The project has led to a variety of publications.

Notable is the content available at our WWW site: http://www.ctrnet.net/

This includes a page about publications, which has a number of links to copies of the works: http://www.ctrnet.net/publications

- Fox, E. Introduction to (Teaching / Learning about) Digital Libraries. Refereed 1/2 day tutorial for Research and Advanced Technology for Digital Libraries, Proc. 14th European Conference, ECDL2010, Glasgow, Sept. 6-10

- Kavanaugh, A., Fox, E. A., Sheetz, S., Yang, S., Li, L. T., Whalen, T., Shoemaker, D., Nastev, A., Xie, L. Social Media for Cities, Counties and Communities. Final Grant Report to VT CCSR. March 11, 2011, Virginia Tech, Dept. of Computer Science Technical Report TR-11-09, Blacksburg, VA. http://eprints.cs.vt.edu/archive/00001148/

- Kavanaugh, A., Yang, S., Sheetz, S. D., Li, L. T., Fox, E. A. (2011). Microblogging in Crisis Situations: Mass Protests in Iran, Tunisia, and Egypt. Refereed paper at workshop held in conjunction with the ACM Conference on Human Factors in Computing Systems (CHI'11). May 7-12, 2011, Vancouver, Canada. http://www.ctrnet.net/sites/default/files/Twitter%20Use%20in%20Iran%20Tunisia%20Egypt.Kavanaugh.Final__0.pdf

- S. Sheetz, E. Fox, A. Fitzgerald, S. Palmer, D. Shoemaker, A. Kavanaugh. "Why students use social networking sites after crisis situations". Poster presented at Information Systems for Crisis Response and Management (ISCRAM) conference,

May 8-11, 2011, Lisbon, Portugal. http://www.ctrnet.net/sites/default/files/ISCRAM2011WhyFacebookPoster.pdf

- Kavanaugh, A., Fox, E., Sheetz, S., Yang, S., Li, L.T., Whalen, T., Shoemaker, D., Natsev, A., Xie, L. Social Media Use by Government: From the routine to the critical. Paper at ACM 2011 Digital Government Research Conference (dg.o 2011), June 12-15, 2011 (College Park, MD). New York: ACM Press. http://www.ctrnet.net/sites/default/files/dgo.2011.Paper%20Final.pdf

- Li, L.T., Yang, S., Kavanaugh, A., Fox, E. A., Sheetz, S. D. and Shoemaker, D. Twitter User During an Emergency Event: the Case of UT Austin Shooting. Poster at ACM 2011 Digital Government Research Conference (dg.o 2011), June 12-15, 2011 (College Park, MD). New York: ACM Press. http://www.ctrnet.net/sites/default/files/dgo2011-cameraready2.pdf

- Seungwon Yang and Andrea Kavanaugh. Half-Day Tutorial: Collecting, Analyzing and Visualizing Tweets using Open Source Tools. ACM 2011 Digital Government Research Conference (dg.o 2011), June 12-15, 2011 (College Park, MD). New York: ACM Press. http://www.ctrnet.net/sites/default/files/Twitter%20tutorial%20submitted.pdf

- Seungwon Yang, Andrea Kavanaugh, Nádia P. Kozievitch, Lin Tzy Li, Venkat Srinivasan, Steven Sheetz, Travis Whalen, Donald Shoemaker, Ricardo da S. Torres and Edward A. Fox. CTRnet DL for Disaster Information Services. Poster at ACM/IEEE Joint Conference on Digital Libraries (JCDL'11). June 13-17, 2011. Ottawa, Ontario, Canada. http://www.ctrnet.net/sites/default/files/jcdl212p-yang-submitted.pdf

- E. Fox. Introduction to Digital Libraries. Refereed 1/2 day tutorial for ACM/IEEE Joint Conf. on Digital Libraries, JCDL 2011, Ottawa, June 13-17

- Kavanaugh, Dr. Andrea and Yang, Mr. Seungwon and Sheetz, Dr. Steven and Li, Ms. Lin Tzy and Fox, Dr. Edward (2011) Between a Rock and a Cell Phone: Social Media Use during Mass Protests in Iran, Tunisia and Egypt. Technical Report TR-11-10, Computer Science, Virginia Tech. http://eprints.cs.vt.edu/archive/00001149/
- Seungwon Yang and Andrea Kavanaugh. (2011). Collecting, Analyzing and Visualizing Tweets using Open Source Tools. Technical Report TR-11-13, Computer Science, Virginia Tech.

3.2. Products

3.2.1. Refined CTRnet Homepage

We refined the project homepage at http://www.ctrnet.net/ (see Figure 3). The list of updates is:

- Added two word clouds, one for the Japan earthquake and the other for the Libyan revolution (http://www.ctrnet.net/node/2924).
- Added a list of publication information (http://www.ctrnet.net/publications).
- Added links to our newly developed archives in IA (http://www.ctrnet.net/manmade and http://www.ctrnet.net/natural).
- Changed the user interface to simplify the login process
- Added a map of disaster events

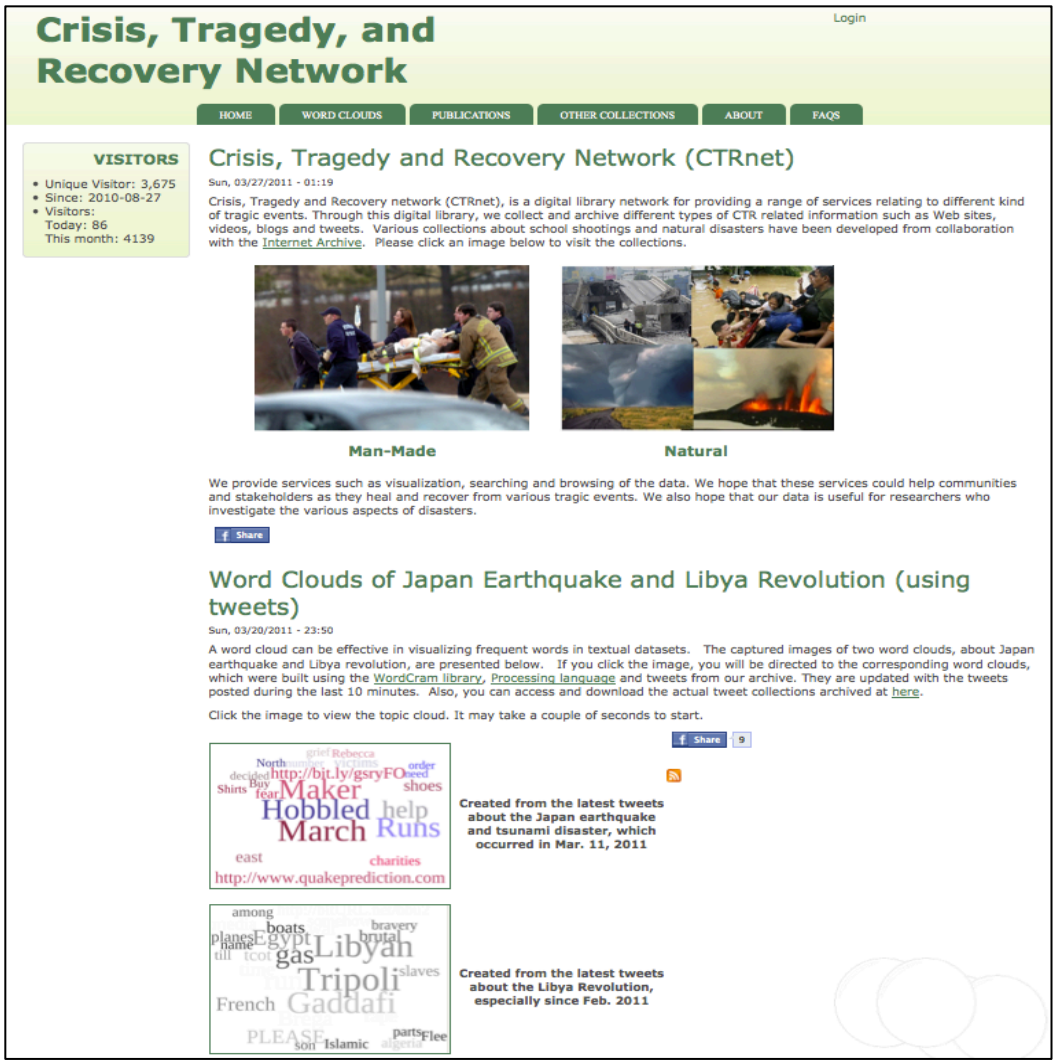

Figure 3. CTRnet homepage.

3.2.2. CTR Collections

As mentioned in Section 2.1.1, we have added more collections using the Heritrix crawler and our seeds. Table 3 shows our 17 collections currently hosted by IA. The list of collections is increasing as new events occur.

Table 3. CTR collections created at the Internet Archive.

Disasters	**Event Type**
Natural	Chilean Earthquake
	New Zealand Earthquake
	Indonesian Volcanic Eruption, Tsunami, Earthquake
	Japan Earthquake
	Typhoon Ketsana
	U.S. Midwest Snowstorm
	U.S. South Eastern Storm
	Cyclone Yasi
	Virginia Tech Global Disasters Collection
Man-Made	April 16 Archive
	Brazilian School Shooting
	International School Shooting
	Northern Illinois University Shooting
	Tucson Shooting
	Youngstown Shooting
	Haiti Earthquake Anniversary
	VT April 16 Shooting Remembrance

Tweets are continuously archived into our tweet archives (Table 4). Since the Middle Eastern revolutions and Japan earthquake/nuclear radiation disaster are on-going events, the number of tweets related to those is increasing.

Table 4. Tweet archives for CTR events.

Disasters	**Tweet Archive Description**	**Count**
Man-Made	Egypt Revolution (#egypt)	2,549,380
	Egypt Revolution (#jan25)	1,438,010
	Libya Revolution	2,948,142
	Tunisia Revolution (#tunisia)	192,043
	Tunisia Revolution (#sidibouzid)	111,393
	Bahrain Revolution	3,091,375
	Yemen Revolution	537,407
	Syria Revolution	1,000,168
	Swaziland Protest	873
	Uganda Protest	4,154
	Brazilian School Shooting (English tweets)	3,574
	Brazilian School Shooting (Brazilian tweets)	12,021
	Virginia Tech Shooting Anniversary	479
	San Jose State University Shooting	1926
Natural	Spain Earthquake	11,416
	Joplin, Missouri Tornado	61,081
	New Zealand Earthquake	122,776
	Japan Earthquake	1,302,698

	Florida Wildfire	1,938
	U.S. Storm in Southern States	7,001
	U.S. Tornado	819,700

3.2.3. Word Clouds

As explained in Section 2.1.6, we have built two word clouds that show approximately top 20 frequent words extracted from the tweets during the last 10 minutes. Figure 4 was developed based on tweets from the Japan earthquake event. Figure 5 was about the Libya revolution. These clouds are updated every 10 minutes to reflect the current issues.

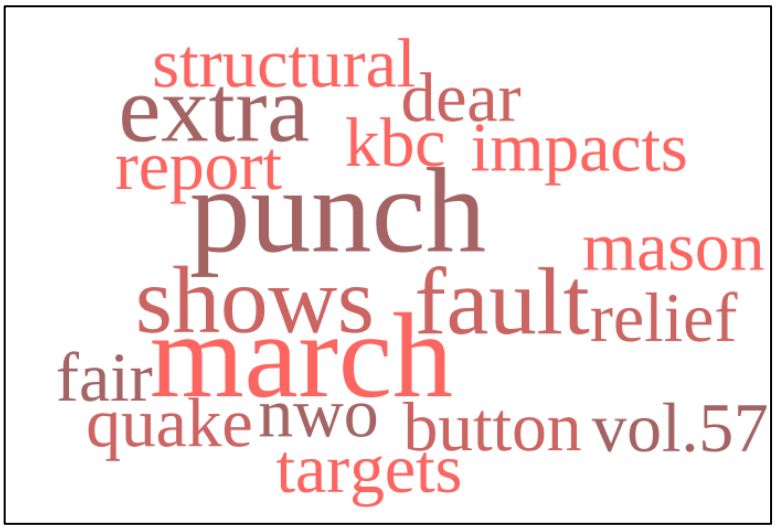

Figure 4. Word cloud developed from the Japan earthquake disaster tweets.

Figure 5. Word cloud developed from the Libya revolution tweets.

3.2.4. CTR Event Map

The prototype map of the disaster events was developed using Google Earth Outreach (http://earth.google.com/outreach/index.html), especially with the Spreadsheet Mapper 2.0 (see Figure 6). It allows entering the data for each event such as event name, location in latitude and longitude, related links, balloon type, etc. into a Google Spreadsheet, which can be published as a Web page and shared among multiple people. Then, this published information is used to place balloons on a map. When a balloon is clicked, it opens up a small window with more information (see Figure 7).

Figure 6. A prototype map of disaster events.

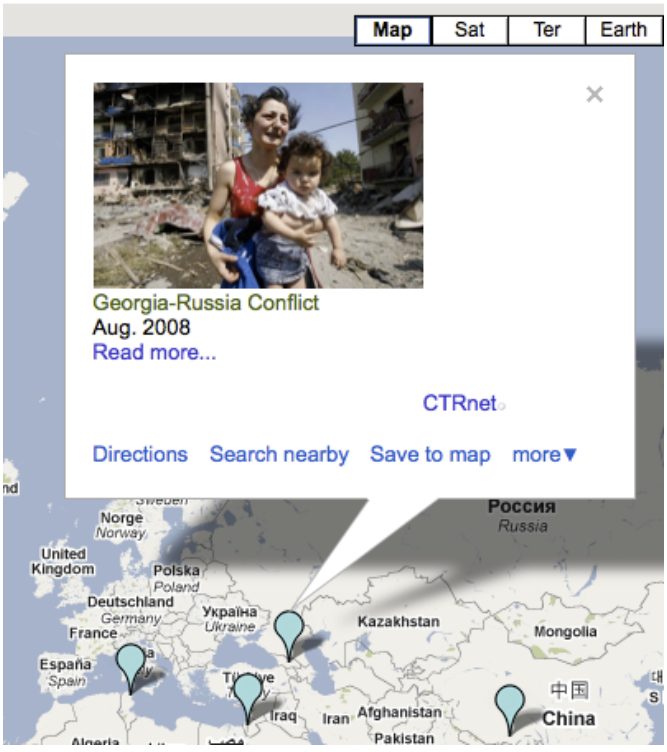

Figure 7. A pop-up window of a balloon to show more details.

4. Contributions

4.1. Contributions to the Principal Disciplines of the Project

The collected information is useful only when they are visible and their related services are provided to the visitors of the CTRnet. As an effort to make our tweet archives visible, we have connected the databases from a tweet collection tool with a Processing visualization script using the Python programming language. This approach allows viewers to see the dynamic and up-to-date visualization of important words from CTR-related tweets. The methodologies for tweet collection, preprocessing, and visualization will be presented in a tutorial session of the Digital Government conference on June 11, 2011.

4.2. Contributions to Other Aspect of Public Welfare

In the fall of 2010, we worked with personnel in Arlington county and the IBM Watson laboratory to study the various needs of social media information collection and analysis. This research helped us to understand real life needs such as methodologies to monitor and summarize social media information from certain regions and certain topics. Currently, two approaches, topic identification and sentiment analysis, are being conducted by the team. These technologies might be used to address those needs.

4.3. Contributions to the Development of Human Resources

More graduate students joined the CTRnet project since Fall 2010. They have been actively participating in the project, and gained valuable experience and skills in working in the emerging critical research area of crisis informatics. Personnel working on the project directly (PIs, co-PIs, and GRAs) have had opportunities to attend IA training sessions and Webinars about newly developed social media collection and analysis tools. It enhances their knowledge and skills with regard to building digital collections from Web pages as well as social media information.