NSF Year 1 Report
for
CTRnet: Integrated Digital Library Support for Crisis, Tragedy, and Recovery

Jul 7, 2010

1 Participants

1.1 People Directly Working on the Project

- **PI** : Edward A. Fox
- **Co-PIs** : Donald Shoemaker, Steven D. Sheetz, Andrea Kavanaugh, Naren Ramakrishnan
- **GRA** : Venkat Srinivasan
- **GRA (<160hrs)** : Seungwon Yang

1.2 Student Collaborators :

- **Graduate Students**: Tram Bethea, Seth Fox, Min Li, Chao Peng, Yipan Deng, Nadia Kozievich, Chet Rosson
- **Undergraduate Students**: Sherley Codio, Jennifer Francois, Mario Calixte, Bernadel Benoit, Fabrice Marcelin, Keith Wooldridge
- **External** : Bidisha Dewanjee

1.3 Other Organizations Involved as Collaborators

- Internet Archive (USA)
- Monterrey Tech (Mexico)
- IBM Research Lab (USA)

2 Activities and Findings

2.1 Research Activities and Findings

The Crisis, Tragedy and Recovery network, or CTRnet, is a human and digital library network for providing a range of services relating to different kinds of tragic events. Through this digital library, we will collect and archive different types of CTR related information, and apply advanced information analysis methods to this domain. It is hoped that services provided through CTRnet can help communities, as they heal and recover from tragic events.

We have taken several major steps towards our goal of building a digital library for CTR events.

Different strategies for collecting comprehensive information surrounding various CTR events have been explored, using school shooting events as a testbed. Several GBs worth of school shootings related data has been collected using the web crawling tools and methodologies we developed. Several different methods for removing non-relevant pages (noise) from the crawled data have been explored. A focused crawler is being developed with the aim of providing users the ability to build high quality collections for CTR events focused on their interests.

Use of social media for CTRnet related research is being explored. Software to integrate the popular social networking site Facebook with the CTRnet digital library has been prototyped, and is being developed further. Integration of the popular micro-blogging site Twitter with the CTRnet digital library is being explored.

The following sections provide details of our activities and findings to date.

2.1.1 CTRnet Database

Different kinds of crisis related data will be collected, analyzed and presented to users as part of this project. CTR related data can occur in various formats, like reports, tweets, pictures, webpages, videos, etc. We have developed a schema for a database that will be used to store this data. Development of this CTRnet database schema was part of the independent study project of a graduate student, Chet Rosson. The schema is descriptive enough to handle diverse CTR data types and formats, and in addition also comprehensive enough to describe the user communities involved (users making contributions to the CTRnet digital library, for example).

Collection Name	Date Crawled	URL
Alabama University Shooting	March, 2010	http://www.archive-it.org/public/collection.html?id=1829
Chile Earthquake	March, 2010	http://www.archive-it.org/public/collection.html?id=1851
Haiti Earthquake[1]	March, 2010	http://www.archive-it.org/public/collection.html?id=1784
Typhoon Ketsana	March, 2010	http://www.archive-it.org/public/collection.html?id=1621
International School Shootings	April, 2010	http://www.archive-it.org/public/collection.html?id=1909
Virginia Tech April16 Shootings Remembrance	April, 2010	http://www.archive-it.org/public/collection.html?id=1899
Northern Illinois University Shooting	April, 2009	http://www.archive-it.org/public/collection.html?id=970
April16 Archive	April, 2008	http://www.archive-it.org/public/collection.html?id=694

Table 1: CTR Related Collections at IA built by Virginia Tech (detailed listing available at http://www.archive-it.org/public/partner.html?id=156)

2.1.2 Web Crawling and Data Collection

Gathering and analyzing CTR related data is an integral part of this project. So far, we have focused on collecting data relating to school shootings. As part of a graduate class project, involving 2 graduate students, Tram Bethea, and Seth Fox, we made attempts at building high quality collections for 10 of the highest impact school shooting events in the United States. We explored several different methodologies for generating a list of relevant seeds, or starting points for performing web crawling, and compiled close to 21000 seeds for each of the 10 events. We have developed a semi-automatic method for seed generation, which was used to produce the seeds for school shooting events. Using these seeds as starting points, we crawled the web using an open source web crawler, and built our collections.

In addition to building collections locally for school shootings, we have also leveraged the web crawling and archiving facility provided by our partner, Internet Archive (IA), to assist in their efforts to develop collections for

[1] Collection was built by IA and their partner(s). Virginia Tech assisted in seeds collection.

many other CTR events (see Table 1). Several webmasters don't allow general web crawlers to crawl their websites, but make an exception for IA's crawlers. Crawling using IA's crawlers thus ensures that we are able to get a broader set of webpages that would otherwise be missed if we were to use our own crawlers. In connection with the subcontract to IA, we have obtained dumps of their CTR related collections to allow us to analyze them and provide additional services.

2.1.3 Filtering

Web crawlers typically also pick-up webpages that are unrelated to the topic of interest. We are attempting to develop Machine Learning based algorithms to filter out non-relevant pages from the crawled data, and ensure that the collections that are built are of high quality. As part of a graduate class project, involving 3 graduate students, Min Li, Chao Peng, and Yipan Deng, we have developed and tested Naive Baye's and SVM based algorithms for filtering some school shootings related data. We are currently refining our method to produce a more robust filtering method.

2.1.4 Focused Web Crawler

We are attempting to build a simple and intuitive focused web crawler that allows non-sophisticated users to build collections for CTR events of interest. As part of an undergraduate class project, and 2 undergraduate research projects, involving Mario Calixte, Fabrice Marcelin, and Bernadel Benoit, we have been able to prototype several pieces of this crawler (UI, modules for filtering, etc.) and are currently working on building a fully functional crawler.

2.1.5 Sociological Perspective on CTR Events

Besides taking a technology-centric approach to studying school shootings we have also looked at these events from a Sociological perspective. We worked on collecting comprehensive information about incidents of school shootings worldwide. So far, we have identified 60 reported school shooting incidents in North America and Europe, from 1996 to the present. The data we have collected reveal several categories of information about the shootings. The categories are year in which the incident occurred, city or town and state, or other location of the event, name of the school, name of the shooter(s), gender of shooter(s), number of deaths, number injured, circumstances of the incident (such as how the shootings occurred and related instances to the shooting), possible motives for the shooting (such as a response to bullying), and type of weapon used, such as rifle or handgun.

We developed protocols for an institutional review regarding a survey of Virginia Tech administrators, on preparedness for tragedies or disasters on campus. The purpose of the intended survey was to identify useful information on school shootings and natural disasters and to incorporate this information into the digital library to develop information resources which would help the administrators as information aids on crisis situations like school shootings and natural disasters. For this purpose a short questionnaire was developed. The questions included what kind of information they think is useful for responding to human tragedies and crisis situations on campus for planning and preparation, immediate response, and long-term investigations. However, this survey was discontinued because Virginia Tech's policy prohibits administrators from discussing any issue connected to the tragedy of April 16, 2007. Although this research was not been conducted on the Virginia Tech campus due to campus policy, we are exploring the possibility of using the same survey at other universities around the country.

2.1.6 Use of Social Media for CTRnet

We have built an application for integrating Facebook with the CTRnet website, as part of a project involving 3 undergraduate students, Jason Browning, Jason Heim, and Jacqueline Nguyen. This provides a good way for the general public to contribute information and resources to the CTRnet digital library directly via Facebook, without explicitly logging into the CTRnet website. The application is currently available only for test use within the VT domain; we are working on making it generally available through Facebook.

We also have been reviewing research on the use and impact of social media in emergency or crisis situations. In this review process, we have focused attention not only on the use of social network sites, such as, Facebook, but also other forms of user generated content or social interaction, such as, twitter, youtube, and emerging location-aware social applications, such as, foursquare. We are concerned not only with information sharing among the general public, but also among rescue personnel and government officials.

We are building on the CTRnet outcomes to date and pursuing research on social media in situations of crisis or social convergence (i.e., large crowds) through a recently funded planning proposal ($15K) from the Center for Community Security and Resiliency (CCSR), a partnership between Virginia Tech, IBM, and Arlington County, Virginia. Over a six-month period (June - December 2010) we will investigate methods to utilize social media sources to meet a variety of Arlington city, county, and community needs. This research includes leveraging and further developing a platform for collecting large amounts of public information relevant to Arlington County and its community, archiving collected social media data over a period of

time into a digital library, correlating multiple information sources, and studying analytic applications for city, county, and community services. With the help of CCSR we also plan to perform focused interviews with county officials, city planners, emergency responders, and public safety agencies in order to identify the most promising applications of social media analytics and mining for Arlington County services and operations. Target information sources include official county publicity portals, blogs, news, community forums, as well as relevant postings on social media sites such as twitter, Facebook, youtube, and flickr. Possible applications of such analyses include monitoring public opinions before and after large public events, monitoring planned or unplanned activities, identifying and categorizing important community issues over time and location, enhancing community recovery in response to crises or tragedies, and monitoring and tracking the development of long-running themes in civil life.

2.1.7 CTRnet Data Analysis

We have attempted to leverage our earlier research on Storytelling in order to find connections between different CTR events, as well as to verify hypotheses for a specific event. A specific hypothesis surrounding a school shooting event for example could be, whether or not bullying played a role in an individual deciding to commit violence against the school community. Storytelling can reveal potentially interesting connections between different events, and also reveal a logical chain of reasoning drawn from the data at hand that leads to proving (or disproving) a specific hypothesis.

We have attempted to apply the Storytelling algorithm to some school shootings data that we had crawled. The results of our analysis revealed that the algorithm in its current form may not be particularly suited for CTR related data. We are currently working on enhancing the Storytelling algorithm so as to have it work on the data from the CTR domain.

2.2 Training and Development Activities

The CTRnet project has afforded learning opportunities to several graduate and undergraduate students.

As part of a graduate course in Information Storage and Retrieval offered in Fall 2009, a class project to help prototype data crawling and filtering modules for CTRnet project was designed. Five graduate students who worked in this project had the opportunity to apply the principles learned in the class on a cutting edge research project.

As part of an undergraduate Multimedia, Hypertext, and Information Access class offered in Spring 2010, 1 undergraduate student worked on applying content based image retrieval modules on natural disaster photographs. Three other undergraduate students working for independent

study, and another for undergraduate research credits in Spring 2010, focused on content-based image retrieval for the photographs, and on developing focused crawlers for crawling CTR data. A Ph.D. student from UNICAMP in Brazil, visiting our Digital Library Research Laboratory for February-October 2010, helped coordinate the work with photographs, providing the EVA system for content-based image retrieval.

Personnel directly working on the project (faculty and students) had several opportunities to take part in online training sessions offered by the Internet Archive. These sessions helped impart knowledge about general web crawling and archiving principles, which are directly useful in the context of the CTRnet project.

2.3 Outreach activities

We have made several attempts to disseminate findings from the project to the wider audience, and to enlist collaborations with various organizations.

We presented a poster at Virginia Tech's International Outreach NOW conference in Sept. 2009 outlining our plans for the CTRnet project and its relevance to the global audience. Our poster won the award for the best poster in one of the several award categories.

We submitted a presentation to Internet Archive's partners conference in Nov. 2009 where we summarized our experiences with using IA's crawling tools for crawling CTR data, and made several suggestions for improving IA's crawling tool.

Several new global partners have been enlisted to provide local support for various CTR events. Researchers in Monterrey Tech in Mexico are interested in building bi-lingual collections (English and Spanish) on swine flu. Our collaborators in Russia are interested in building CTR collections for future CTR events in Russia and the Balkans.

We have recently entered into a collaboration with IBM Research and Arlington County. Through this collaboration, we are developing tools to help Arlington County make use of social media for various purposes. The tools developed as part of this initiative are generic, and will subsequently be extended to suit other situations and at a bigger scale (national, worldwide etc.).

3 Publications and Products

3.1 Publications

- Edward A. Fox, Donald J. Shoemaker, and Steven D. Sheetz. "Initiatives to Assemble a Record". Presented at the conference, Aftermath

Dynamics and Management: Through the Lens of the Virginia Tech Incident, July 22, 2009.

- Venkat Srinivasan, Bidisha Dewanjee, Edward A. Fox, Donald J. Shoemaker, Steven D. Sheetz, Andrea Kavanaugh, and Naren Ramakrishnan. “CTRnet: A Distributed Digital Library for Rescue and Recovery”. Poster presented at International Outreach NOW conference at Virginia Tech, Blacksburg, Virginia, Sep 2009.

- Uma Murthy, Edward Fox, Naren Ramakrishnan, Andrea Kavanaugh, Steven Sheetz, Donald Shoemaker, and Venkat Srinivasan. “Building an Ontology for Crisis, Tragedy, and Recovery Network”. Poster presented at the 8th Networked Knowledge Organization Systems and Services Workshop, Corfu, Greece, October 1, 2009.

- Steven D. Sheetz, Edward A. Fox, Andrew Fitzgerald, Sean Palmer, Donald J. Shoemaker, Andrea Kavanaugh, and Naren Ramakrishnan, “Why Students Use Social Networking Sites After Crisis Situations”. Journal of Computer Mediated Communication, Submitted 2010.

3.2 Invited Presentations

- Edward A. Fox. “CTRnet: A Crisis, Tragedy, and Recovery Network”. Virginia College of Osteopathic Medicine Research Day, Oct. 16, 2009

- Edward A. Fox. “CTRnet (Crisis, Tragedy, and Recovery Network): A global human network and distributed digital library”. University of Iowa, Oct. 29, 2009

- Venkat Srinivasan, Edward A. Fox, Donald J. Shoemaker, Steven D. Sheetz, Andrea Kavanaugh, and Naren Ramakrishnan. "Building Digital Collections for Crises and Tragedies". Presented at Internet Archive’s partners conference, Washington DC, Nov 2009.

- Kristine Hanna, Edward A. Fox, Jamaica Jones, and Padmini Srinivasan. “Capturing Crisis: A Digital Library to Study Tragedy and Recovery from Around the World”. Panel at CNI Fall Meeting, Washington, DC, Dec. 9, 2008

- Edward A. Fox. “CTR”, As part of panel for session "Crisis, Tragedy, and Recovery Network (CTRnet)" at the Coalition for Networked Information Fall Meeting, Dec. 15-16, 2009, Washington, D.C.

3.3 Products

3.3.1 CTRnet Homepage

The project webpage at http://www.ctrnet.net is currently being used as an outreach tool describing the goals and vision of the project, and for inviting CTR related contributions from the general public. We are currently working on adding several features to the website in order to make the CTR collections and results of our analysis widely available.

3.3.2 CTR Collections

Digital collections for 10 high impact school shooting events have been built and are ready to be filtered and made available to researchers via the CTRnet website. Besides building collections for school shootings locally, we have also built several collections for other types CTR events (typhoons, earthquakes, etc.) using IA's facilities. These collections (also listed in Table 1) are accessible at http://www.archive-it.org/public/partner.html?id=156.

3.3.3 Software

An automatic seed generator has been developed that collates a user defined number of seeds for a CTR topic of interest from various web search engines.

Several different web crawlers have been developed to crawl Web 2.0 websites like blogs, forums, video sharing websites, etc. for a user specified CTR event.

Several other pieces of software (focused crawler, twitter crawler, etc.) are under development.

3.3.4 Social Networking Application

A Facebook application has been developed that will be used as an outreach tool. When complete, it will allow users to upload resources into the CTRnet digital library from within Facebook itself.

4 Contributions

4.1 Contributions to the principal disciplines of the project

One of the major goals of the CTRnet project is to develop tools and infrastructure for researchers to study and learn from different crises and tragedies. We have taken several steps toward this goal already, with the development of various CTR digital archiving and analysis tools.

As part of this initiative, we have been examining various social media used during crises (e.g., twitter, Facebook, blogs, flickr, youtube) and have developed a Facebook application allowing users to archive their digital content related to crises. Scholars note that it is the sharing of information in a disaster that is especially crucial, and different types of information and communication technology (ICT) facilitate sharing to a greater or lesser degree based on a variety of circumstances. We also have been further reviewing the research on the use and impact of social media in emergency or crisis situations.

4.2 Contributions to other aspects of public welfare

Our research underway into the use of social media in crisis situations should contribute to public safety and security. Specifically, it should help us understand how to optimize communication and information sharing among the public, and among first responders as well as longer term recovery agents and groups, including: rescue crews, police and fire, community leaders, voluntary associations, and government officials in charge of public communications. These analyses are intended to help cities and communities, such as Arlington County, know how and where to reach citizens in the event of a crisis or social convergence condition, as well as to monitor and make sense of the diversity of voices and information that enriches the quality of life in that community. The planning period and pilot study should advance technologies and systems in social media analysis, and inform day-to-day civil society.

4.3 Contributions to the development of human resources

Several graduate and undergraduate students working on the CTRnet project have 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 several IA training sessions, thereby enhancing their knowledge and skills with regard to building digital collections.