NSF 3rd Year Report

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

July 2012

1. Participants

1.1. Project Members

- Principal Investigator: Edward A. Fox
- Co-Principal Investigators: Donald J. Shoemaker, Steven D. Sheetz, Andrea L. Kavanaugh, Naren Ramakrishnan
- GRAs: Venkat Srinivasan, Seungwon Yang
- Former Visiting Faculty, Now Collaborating from Egypt: Riham Abdel Moneim
- Graduate Students: Tram Bethea, Yinlin Chen, Amine Chigani, Kiran Chitturi, Bidisha Dewanjee, Yipan Deng, Noha ElSherbiny, Seth Fox, S. M. Shamimul Hasan, Nadia P. Kozievitch, Sunshin Lee, Lin Tzy Li, Min Li, Mohamed Magdy, Chao Peng, Chet Rosson, Travis F. Whalen
- Undergraduate Students: Bernadel Benoit, Jason Browning, Mario Calixte, Sherley Codio, Rachel Coston, Jennifer Francois, Jason Heim, Robert Leith, Fabrice Marcelin, Ashley Phelps, Jason Smith, Justin Tillar, Keith Wooldridge
- Other Senior Personnel: Paul Mather (Library), John Tedesco (Communications)

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 (CA, U.S.A.)
- IBM Watson Research Laboratory (NY, U.S.A.)
- Monterrey Tech (Mexico)
- Arlington County (VA, U.S.A)
- American University in Cairo, Egypt (Manal M. Samra)
- Texas A&M University (Paul Logasa Bogen II)
- State Autonomous University of Mexico, Toluca (Rodrigo Sandoval Almazan)
- High Institute of Management of Tunis (Hamida Skandrani)
- Université Laval (Sehl Mellouli)
- Lucid Imagination, Inc. (Paul Doscher, Charlotte Goldsberry)

2. Activities and Findings

2.1. Research and Education Activities

The Crisis, Tragedy and Recovery (CTR) network, or CTRnet, is a human and digital library network for providing a range of services relating to different kinds of tragic events, including broad collaborative studies related to Egypt, Tunisia, Mexico, and Arlington, Virginia. Through this digital library, we 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, initially using school shooting events as a testbed. Many GBs worth of 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 has proceeded well, and is being further automated, becoming a key part of our methodology.

Our web site includes selected results of studies, including a map of collections, and dynamic tag clouds from ongoing disasters.

We helped broaden interest in enhancing the use of information science and technology advances to aid with disasters through two activities. One was assisting with the April 24, 2012 workshop sponsored by NSF and CRA/CCC on "Computing for Disasters". The second was our 90 minute webinar "Emergency Informatics and Digital Libraries" on July 24, 2012, with 33 participants (http://www.ctrnet.net/webinar).

See additional details below.

2.2. Major Findings

2.2.1. CTRnet Database

Different kinds of crisis related data are 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).

Additional content collections follow other principles related to information retrieval or information science, including for tweets.

2.2.2. Web Crawling and Data Collection

Gathering and analyzing CTR related data is an integral part of this project. Initially, we 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 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 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 also leveraged the web crawling and archiving facility provided by our partner, Internet Archive (IA), to assist in their efforts to develop collections for many other CTR events. 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 web pages 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.

2.2.3. Filtering

Web crawlers typically also gather web pages that are unrelated to the topic of interest. We are developing 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 developed and tested Naive Baye's and SVM based algorithms for filtering some school shootings related data. We continue refining our methods to produce more robust filtering techniques.

2.2.4. 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 prototyped several pieces of this crawler (UI, modules for filtering, etc.). Research on focused crawlers is ongoing with the help of a number of graduate students who have volunteered their help.

2.2.5. 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.

2.2.6. Use of Social Media for CTRnet

We 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 build on CTRnet outcomes, pursuing research on social media in situations of crisis or social convergence (i.e., large crowds), in part through a project funded by the Center for Community Security and Resiliency (CCSR), a partnership between Virginia Tech, IBM, and Arlington County, Virginia. Over the period June - December 2010, we investigated methods to utilize social media sources to meet a variety of Arlington city, county, and community needs.

This research included 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 performed 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.

Targeted information sources included 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.

This work has been broadened to help with understanding social change and revolution in Egypt, Tunisia, Mexico, and other locations. Some of the work has broadened the CTR network to include a number of additional collaborators at institutions able to carry our surveys in locations of particular interest.

2.2.7. 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. A dissertation made clear it could be applied, with significant tailoring, to several domains. We are currently working on enhancing the Storytelling algorithm so as to have it work on the data from the CTR domain. See additional details below.

2.3. Training and Development

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

As part of a graduate course in Information Storage and Retrieval offered in Fall 2009, a class project was designed to help prototype data crawling and filtering modules for the CTRnet project. 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 during the period February-October 2010, helped coordinate the work with photographs, providing the EVA system for content-based image retrieval.

Additional graduate student activities were carried in the Fall 2010 offering of Information Storage and Retrieval, and in the Fall 2011 offering of Digital Libraries. Additional undergraduate student activities were carried out in the Spring offerings of Multimedia, Hypertext, and Information Access in Spring 2011 and 2012.

Further, a number of students and others new to this area participated in our 90 minute webinar “Emergency Informatics and Digital Libraries” on July 24, 2012, with 33 participants.

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.

2.4. Outreach Activities

We have disseminated findings from the project to wider audiences, and enlisted 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 global partners were enlisted to provide local support for various CTR events. Researchers in Monterrey Tech in Mexico were interested in building bi-lingual collections (English and Spanish) on swine flu. Our collaborators in Russia were interested in building CTR collections for future CTR events in Russia and the Balkans.

We launched our collaboration with IBM Research and Arlington County. Through this collaboration, we developed tools to help Arlington County make use of social media. 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.).

In 2011-2012 we extended and continued collaboration to include sites in Egypt (Alexandria University and the American University in Cairo), Mexico (State Autonomous University of Mexico, Toluca), and Tunisia (High Institute of Management of Tunis, collaborating with Universite Laval). We also launched a collaboration regarding software to aid those dealing with CTR-related collections, with the assistance of Lucid Imagination, Inc. Further, we ran a 90 minute webinar "Emergency Informatics and Digital Libraries" on July 24, 2012, with 33 participants.

2.5. Details on Research Activities and Findings

2.5.1. Introduction

The members of the CTRnet project, which has been funded for three years from the National Science Foundation (IIS-0916733), have been archiving online resources related to both natural and man-made disasters and emergencies since the summer of 2009. Through a close collaboration with the Internet Archive (IA), currently we are maintaining 43 webpage archives as shown in Section 3.2.2. We also have been building public tweet archives since February 25, 2011. Section 3.2.2 presents a total of 87 tweet archives; for each we give: keywords or hashtags used to collect relevant tweets, a brief description, the number of tweets, and the creation time.

The initial step to build a webpage archive is to prepare a list of seed URLs, which point to relevant resources, so that a crawler such as Heritrix can visit the site and collect pages. In the beginning, we relied on human labor to prepare these seeds, which was a good approach to acquire high quality seeds, but it required much time and effort. In the case of large scale disasters and emergency situations, this manual seed preparation approach would never keep up with the speed of newly generated online information. This naturally led us to devise a somewhat more automated approach. The main idea to automate seed preparation is to harvest URLs included in tweets; these are readily available in our disaster tweet archives. A benefit of using the URLs in tweets as seeds is the ability to archive information that people in the disaster-affected and nearby areas actually share when helping each other. To deal with non-relevant / noise resources that nevertheless are found in our archives, we studied a machine learning based filtering approach. Section 2.5.2.1 presents the details of our automation of archiving and filtering studies.

When searching and browsing the massive amount of disaster information we have collected, the ability to visualize the relevant resources, with their time and location, is critical for sense-making of events. Thus, we have been working on a new user interface (UI). The UI presents the location of the disaster events on a Google Map; the date and time of the events are shown using a timeline visualization. Searching for a query word results in a ranked list of resources with their metadata such as title, short description, date, and original URL. Details of the UI are explained in Section 2.5.2.2.

Another study included in the CTRnet project was to monitor tweets to identify where water main breaks have occurred. Tweets that contain keywords about water main breaks are continuously collected using an open source tool. After filtering, available location data such as latitude and longitude, as well as geo-parsed city/road names, are used in the visualization on a Google map. All of this software fits into our digital library framework, which includes collection of tweets, analysis, and visualization. For details, see Section 2.5.2.3.

2.5.2. Experiments and Prototypes

2.5.2.1. Archive Development and Filtering

In the event of emergencies and disasters, massive amounts of web resources are generated and shared. Due to the rapidly changing nature of those resources, it is important to start archiving them as soon as a disaster occurs. This led us to develop a prototype system for constructing archives with minimum human intervention using seed URLs extracted from tweet collections. In Section 2.5.2.1.1 we present the details of our prototype system. We applied it to five tweet collections that had been developed in advance, for evaluation. We also identified five categories of non-relevant files; below we give a discussion of findings from the evaluation.

When archiving resources broadly, it is inevitable to collect non-relevant resources along with relevant ones. To filter the non-relevant ones, we experimented with machine learning based techniques, using our collections for earthquake and flood events. Details are presented in Sections 2.5.2.1.2 and 2.5.2.1.3.

2.5.2.1.1. Automating archive development using open source tools

(Note: the content of this section is adapted from the JCDL'12 poster by Yang et al.)
The Crisis, Tragedy, and Recovery Network project (http://www.ctrnet.net) group has been archiving disaster-related webpages and tweets in collaboration with the Internet Archive [1]. The scope of the archive spans from the U.S. to international disaster events both in the natural (e.g., earthquakes, floods, hurricanes, volcanic eruptions) and man-made (e.g., shootings, transportation crashes, terrorism, political turmoil) disaster domains.

The usual steps to develop an archive of webpages are:

1. Prepare a list of seed URLs to use when crawling data.
2. Provide those seed URLs to the Heritrix crawler [2].
3. Fetch the webpages (e.g., HTML, image on the page) and store them as compressed files (i.e., in WARC format [15]).
4. Extract textual resources from the compressed files.
5. Index extracted textual resources for search/browse services.

The procedure above requires much human labor and time, especially for steps 1, 2, and 4. Also, it is difficult for a human archivist to keep up with the speed of newly generated webpages in the event of massive scale disasters such as the Japan Earthquake of 2011. Our approach to further automate the process is described in Section 2.5.2.1.1.

Related Studies

The LiWA project aims to provide a framework for building archives appearing identical to the online version [5]. Two application scenarios are supported. One is the Streaming Archive application, which preserves audio and video Web information. The other is the Social Web application that captures the dynamics of user interactions in the social web. Similar to LiWA, our goal is to develop archives of textual documents in an efficient manner using automation, which allows us to cover massive scale disasters.

Kandylas and Dasdan discuss that search engine researchers are becoming more interested in adding Twitter posts to their search results [6]. The main problem with tweets is assessing quality. The authors discuss the characteristics of the shortened URLs in the post, showing that most URLs fall into two categories: high quality or spam. We attempt to overcome the URL quality issues through our planned integration of a component for document filtering. Hughes and Palen show how the URLs in tweets play an important role in emergency events as a Web information broker and a means for distributing information resources on the Web [7]. Their study presents why it is important to use URLs found in tweets as seeds for crawls, and also shows the need for automation because there is a higher percentage of URLs in tweets during emergency events.

Prototype System

Figure 1 shows components of the prototype system and the data used as inputs and outputs.

- Input data comes from (1) Tweet Archive DB, built from [3].
- URL Extractor (A) periodically identifies URLs embedded in tweets from (1) and stores them into the Extracted URL DB (2).
- The Heritrix crawler (B) visits each URL in (2) and retrieves webpages.
- Crawled data are stored in Web ARChive (WARC) (3) format.
- Data Extractor (C) retrieves resource types such as HTML pages, PDFs, images, etc. (4) from multiple WARC files.

(A) and (C) have been implemented using Python. We used an open source version of Heritrix for (B). Further prototyping has integrated all the components in the Django framework [4] except for (A). The next step was to connect with the UI.

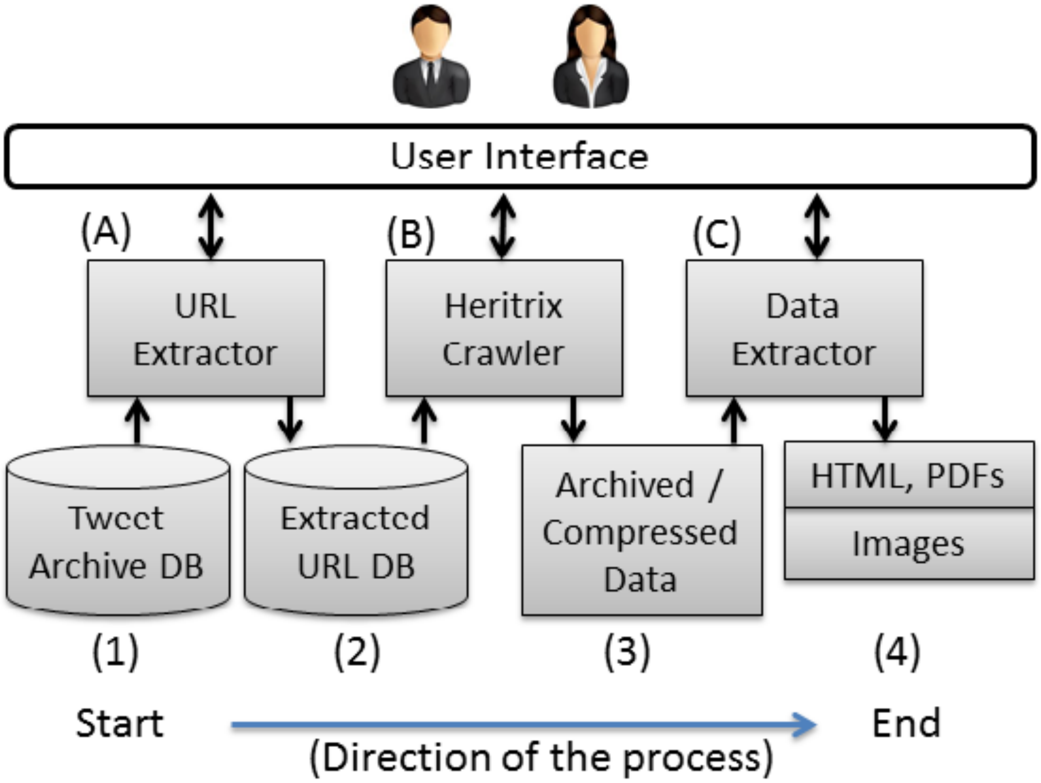
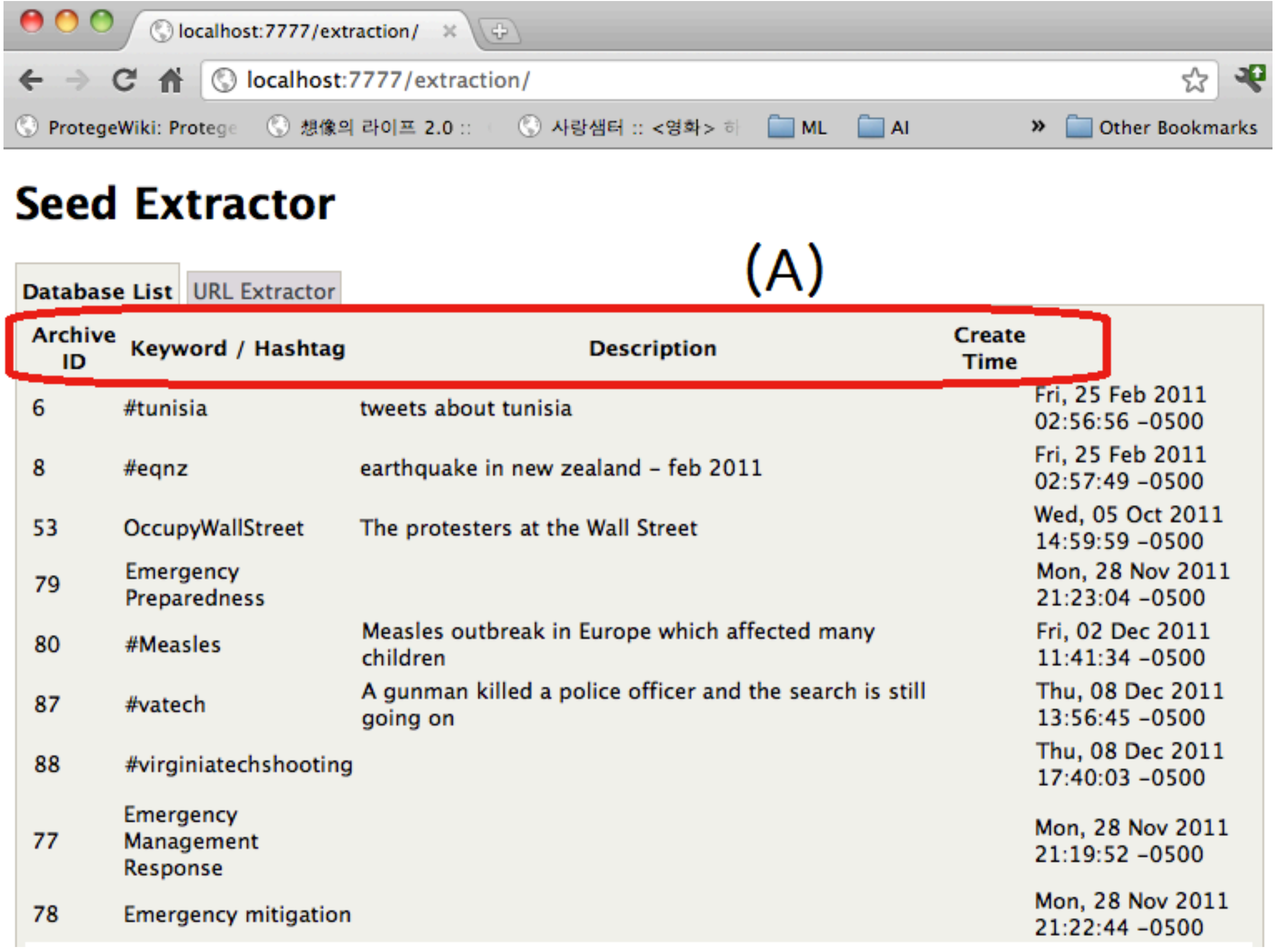


Figure 1. An overview of the prototype system. (A)-(C) are processing components. (1)-(4) are components for the produced data and its management.

UI Design and Using the System

Figure 2 presents a list of Extracted URL DB tables, each of which contains seed URLs from its corresponding Tweet Archive DB table. Users identify an archive ID of interest in the Database List tab. Then they start URL extraction at the URL Extractor tab.



Archive ID	Keyword / Hashtag	Description	Create Time
6	#tunisia	tweets about tunisia	Fri, 25 Feb 2011 02:56:56 -0500
8	#eqnz	earthquake in new zealand - feb 2011	Fri, 25 Feb 2011 02:57:49 -0500
53	OccupyWallStreet	The protesters at the Wall Street	Wed, 05 Oct 2011 14:59:59 -0500
79	Emergency Preparedness		Mon, 28 Nov 2011 21:23:04 -0500
80	#Measles	Measles outbreak in Europe which affected many children	Fri, 02 Dec 2011 11:41:34 -0500
87	#vatech	A gunman killed a police officer and the search is still going on	Thu, 08 Dec 2011 13:56:45 -0500
88	#virginiatechshooting		Thu, 08 Dec 2011 17:40:03 -0500
77	Emergency Management Response		Mon, 28 Nov 2011 21:19:52 -0500
78	Emergency mitigation		Mon, 28 Nov 2011 21:22:44 -0500

Figure 2. A list of Extracted URL DB tables with keywords, hashtags, descriptions, and the creation times of the original tweet archives.

Evaluation – Analysis for Relevance

Table 1 presents five tweet archives that we selected for this study. The number of HTML files extracted was less than the number of seed URLs for all the archives, since in some cases the website's robots.txt policy blocked the site from being crawled. The archive, *Cagayan de oro*, named after one of the typhoon-hit cities in Philippines, had a low precision of 31.2%. We used the keyword 'cagayan de oro' as a keyword to collect tweets. As a result, many tweets about travel, job advertisements, and other issues in the city of Cagayan De Oro also were captured, and their URLs were extracted for crawling. However, due to the use of 'violence' along with 'South Sudan', we had observed a high precision of 84.0% in the *Violence South Sudan* archive.

Table 1. Details of the tweet archives used in the analysis.

Keywords, Hashtags	Tweets	Seed URLs	HTML files	Precision (%)
#virginiatechshooting	456	149	43	48.8
#measles	761	174	155	54.4
Cagayan de oro	1,132	262	237	31.2
Violence South Sudan	1,118	434	431	84.0
Emergency preparedness	7,427	641	386	41.2

We also classified the non-relevant HTML pages into five categories: too short; empty-error pages; non-English pages; blocked pages; and off-topic. Detailed results were presented on our poster.

Conclusion and Future Plans

We presented how our prototype system integrated multiple components, and showed how these components work together to construct archives with minimal human intervention. Prototype capabilities were clarified as we applied it to five selected tweet archives. To ensure high quality in the archives, a de-duplication process, which is based on a content comparison instead of a URL strings comparison, is necessary. Another necessity is a sophisticated filtering component. Updating the UI and fine-tuning of the parameters in the Heritrix configuration file were considered important areas for ongoing work.

2.5.2.1.2. Filtering non-relevant resources using machine learning - earthquakes

One of the future plans from our previous automation for archive development study was to have a sophisticated filtering component. For this, we experimented with machine learning techniques in two natural disaster domains - earthquakes and floods.

Cleaning out the noise data from our earthquake collections required us to develop a training set. Our proposed filtering approach semi-automatically generates a training set and helps to achieve high quality archives.

Dataset:

CTRnet digital library contains a number of earthquake archives. Table 2 provides statistics about the part of our collection related to earthquakes.

Table 2: Earthquake Collections.

Archive Title	Number of webpages
Virginia Earthquake	9313
New Zealand Earthquake	603
Haitian Earthquake (one month)	1204824
Sikkim Earthquake	2252
Turkey Earthquake	382910
Chile Earthquake	119
Chilean Earthquake	150
Japan Earthquake (one month)	3187346
Haiti Anniversary	665
Total number of webpages	4788182

Training set Development:
Our proposed machine learning based filtering approach needs a good training set for classification. The following section explains our methodology.

Experiment Design:
Phase 1- Gold Standard: In this paper we used New Zealand and Virginia earthquake archive samples for our experiments. We randomly selected 140 webpages from the New Zealand collection and 100 webpages from the Virginia collection. After that, one human subject studied each of the webpages and provided a label. We used two types of labels: 1) relevant and 2) non-relevant. Thus we obtained a gold standard for our study.

Phase 2- Automatic Training Set Development: We wrote a Python script for automatic training set development. The purpose of this script is to extract the content of a webpage and automatically assign a label. The script implements the following steps:
1. If 'quake' or 'earthquake' is found in the title of the webpage, we assign 4 points.
2. The number of occurrences of the word 'quake' in all the strings between <p></p> tags is added to the score. (The script removes all the HTML tags before counting.)
3. If the total point >= 6, the script considers the HTML file as relevant to that particular earthquake. Otherwise, it is considered non-relevant.

Experimental Results:
For each of the webpages we compared the script-generated label against our gold standard and calculated the precision, recall, and F1-Score. Our experimental results are

available in the gold standard result column of Table 3. For each of the earthquakes we achieved very good precision, recall, and F1-Score, which indicates that our automatic training set development script performs very well on the labeling task.

We also applied the WEKA [13] machine learning toolkit to measure the performance of the training set. We applied “Filteredclassifier” in the training sets generated in Phase 2. In the Filteredclassifier algorithm we used “StringToWordVector” as a filter and the “Logistic” classification algorithm. We also use “term frequency (TF)”, “inverse document frequency (IDF)”, and “SnowballStemmer” for filtering; for the logistic algorithm our maximum iteration size was 80. All our WEKA experiments were performed using the 10-fold cross validation approach. Our results are available in Table 3 -- see the WEKA result columns. We note that the WEKA classification result and the gold standard comparison result are very close. So, we may conclude that our automatic training set development scheme performs very well. As future work we could apply our training set to classify archive webpages. Our proposed approach should help to increase the accuracy of an archive.

Table 3: Earthquake Experiment Results

Collection/ Evaluation	New Zealand		Virginia		Overall	
Metrics	Gold Standard Result	WEKA Result	Gold Standard Result	WEKA Result	Gold Standard Result	WEKA Result
Precision	1	0.99	0.93	0.917	0.97	0.913
Recall	0.91	0.99	0.86	0.918	0.89	0.92
F1-Score	0.95	0.99	0.894	0.918	0.93	0.916

2.5.2.1.3. Filtering non-relevant resources using machine learning - floods

Another disaster archive we used for a filtering experiment was in the flood domain. In order to choose the best algorithm, we used the same flood training set with every algorithm, and ten-fold cross validation (CV). In CV, the training set is split into two groups. One group is used to train the algorithm, while the other group is used to test the performance of the trained algorithm. After the algorithm is trained, WEKA attempts to classify the validation set. It then reports the percentage of instances that are correctly identified. Since ten-fold cross validation was used, this was done ten times. A total of five algorithms were chosen based on research for text-based classifiers. The results are shown in Table 4. Bayesian Logistic Regression performed the best on the training set. Now that we know what algorithm is best, we can begin processing our test sets for other collection domains.

Table 4: Comparison of Algorithms

	Naïve Bayes	Bayesian Logistic Regression	Logistic Regression	Classification Via Regression	Classification Via Clustering
Correctly Classified Instances (Number)	1650	1965	1771	1929	1569
Correctly Classified Instances (Percent)	78.24%	93.17%	83.97%	91.47%	74.40%

An important step to understand WEKA's classification is to tabulate the true positives, true negatives, false positives, and false negatives. The results of the tabulation are shown in Table 5. These results help to highlight the large number of negative documents that were originally bringing down the precision. To determine if precision was improved, precision, recall, and the F-1 score need to be recalculated for our newly classified data. The sampling error is negligible because each web page was manually inspected after running WEKA to verify the validity of the results. As illustrated in Table 6, both precision and the F-1 score have greatly improved. Recall, however, has decreased slightly due to an increase in false negatives.

Table 5. The Machine-Labeled Statistics (tf: true positive, fp: false positive, tn: true negative, and tn: false negative)

	tp	fp	tn	fn	Total
China Flood	152	5	8	219	384
Philippines Flood	36	3	4	193	236
Thai Flood	279	4	19	787	1089

Table 6. Experimental Results (Original: Ori, Experimental: Exp, Precision: pre, Recall: rec, F-1 score: f-1)

	Ori pre	Ori rec	Ori f-1	Exp pre	Exp rec	Exp f-1
China Flood	0.369	1	0.567	0.968	0.95	0.959
Thai Flood	0.153	1	0.265	0.923	0.90	0.911
Philippines Flood	0.256	1	0.408	0.936	0.986	0.965

This study shows that the Bayesian Logistic Regression algorithm delivers higher precision on the CTRnet flood archives than other traditional machine learning algorithms implemented in the WEKA tool suite. By removing the concern of training set quality through the use of human classification, machine learning algorithms were examined to determine which one provides the best precision. These findings can be used to complement research on the automated creation of training sets.

2.5.2.2. User Interface Design

Figure 3 illustrates the workflow related to our user interface activities.

Over the period of three years of support from NSF, we have archived 43 events in collaboration with the Internet Archive (IA) [1], in addition to some on our own. We also collected data from other CTR (Crisis, Tragedy and Recovery) events that IA and their partners have archived.

Table 7 shows the top three CTR (Crisis, Tragedy and Recovery) events according to the estimated total size of the archives. There are several factors that affect the total size of the archive, like number of seeds, depth of the crawl, and frequency of the crawl. The WARC files [15] from the first month of each collection were downloaded into Amazon cloud machines, and a size estimate was made based on the total number of WARC files.

As downloading all these archives required huge amount of storage, we have developed an initial framework as presented in Figure 3. It shows what each archive should go through. The final phase of the framework is the presentation of data in a User Interface that can be easily accessed. Once an archive is processed, we keep only the data that is needed.

The phases of the framework are

1) Extraction
2) Filtering
3) Indexing
4) Web-framework (or User-Interface)

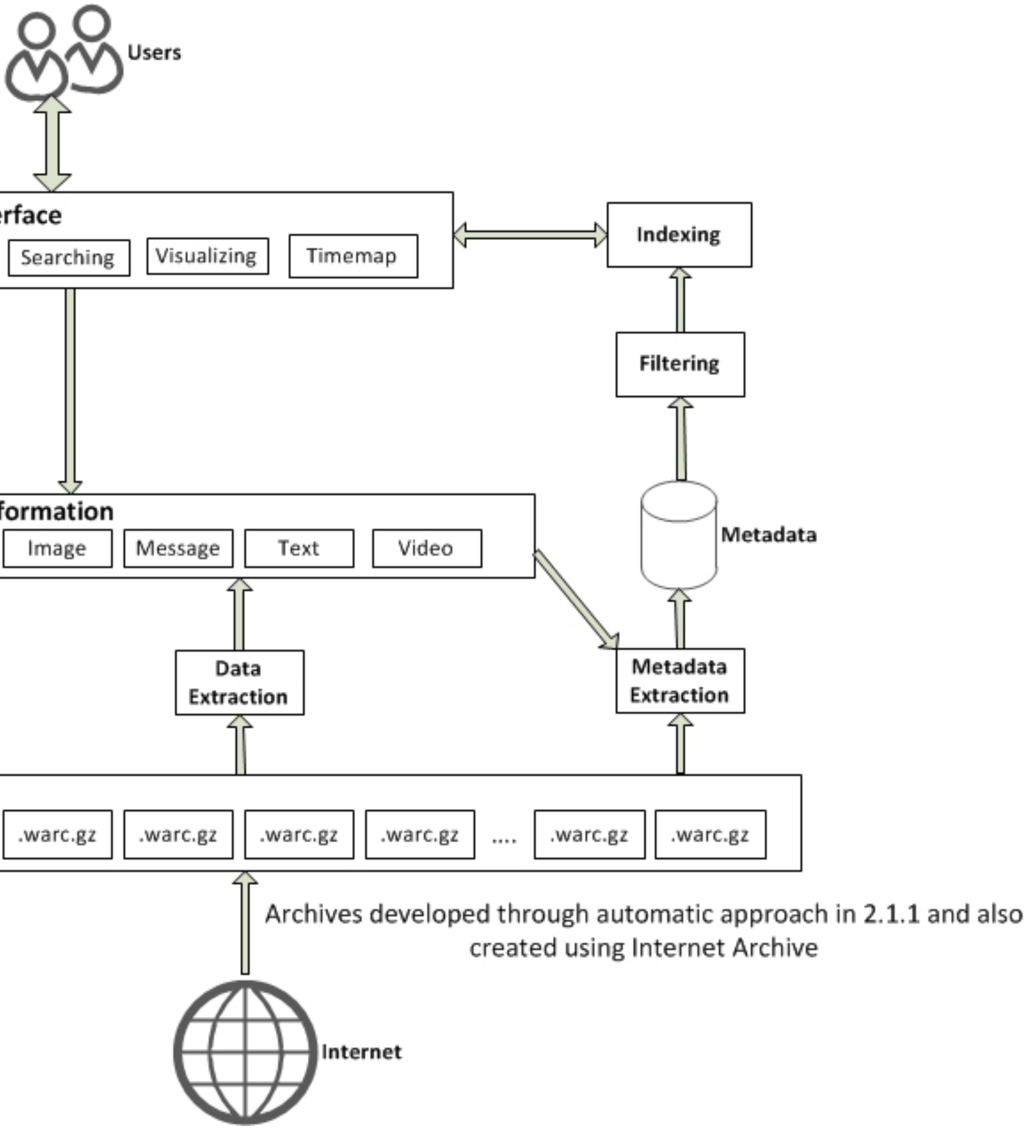


Figure 3. Data flow diagram of current work.

Table 7. The top three collections that have the largest resource.

CTR Event	One Month Size (Gigabytes)	Months Archived	Estimated Total Size (Gigabytes)
Middle East Collections	80	16	3406
Japan Earthquake	572	14	3069
Haiti Earthquake	726	3	1415

2.5.2.2.1. Archive Extraction

The archives that we have created in collaboration with the Internet Archive [1] and also on our own servers are extracted by using a modified version of the Hanzo warc-tools [14]. The archives were extracted based on the MIME3 types (Content types) [16] present and initially were stored in the Amazon cloud.

2.5.2.2.2. Metadata Collection

The metadata collection is done in two phases; one is through archives while the other is through HTML files:

1) While the archives are extracted using the tools, we also collect metadata that comes with each MIME3 type, into a database. Table 8 shows the metadata associated with each WARC record. For more information see [14, 15].

Table 8. Metadata associated with WARC records.

Name	Definition
warc_content_length	The number of octets in the block
warc_subject_uri	The original URI whose collection gave rise to the information content in this record
warc_file	The path of the WARC file from which information is extracted
warc_uri_date	The timestamp that represents the time record creation began
Outfile	The path in which the information is saved
wayback_uri	The URI of the file in the Wayback Machine
warc_type	The type of WARC record
warc_content_type	The MIME type of information contained in the record's block
warc_id	Unique ID assigned to the WARC record

2) The HTML pages that are extracted are parsed and then more metadata is extracted. If it is a news article we are extracting the date when the article was written, the content of the article, as well as results from name-entity recognition.

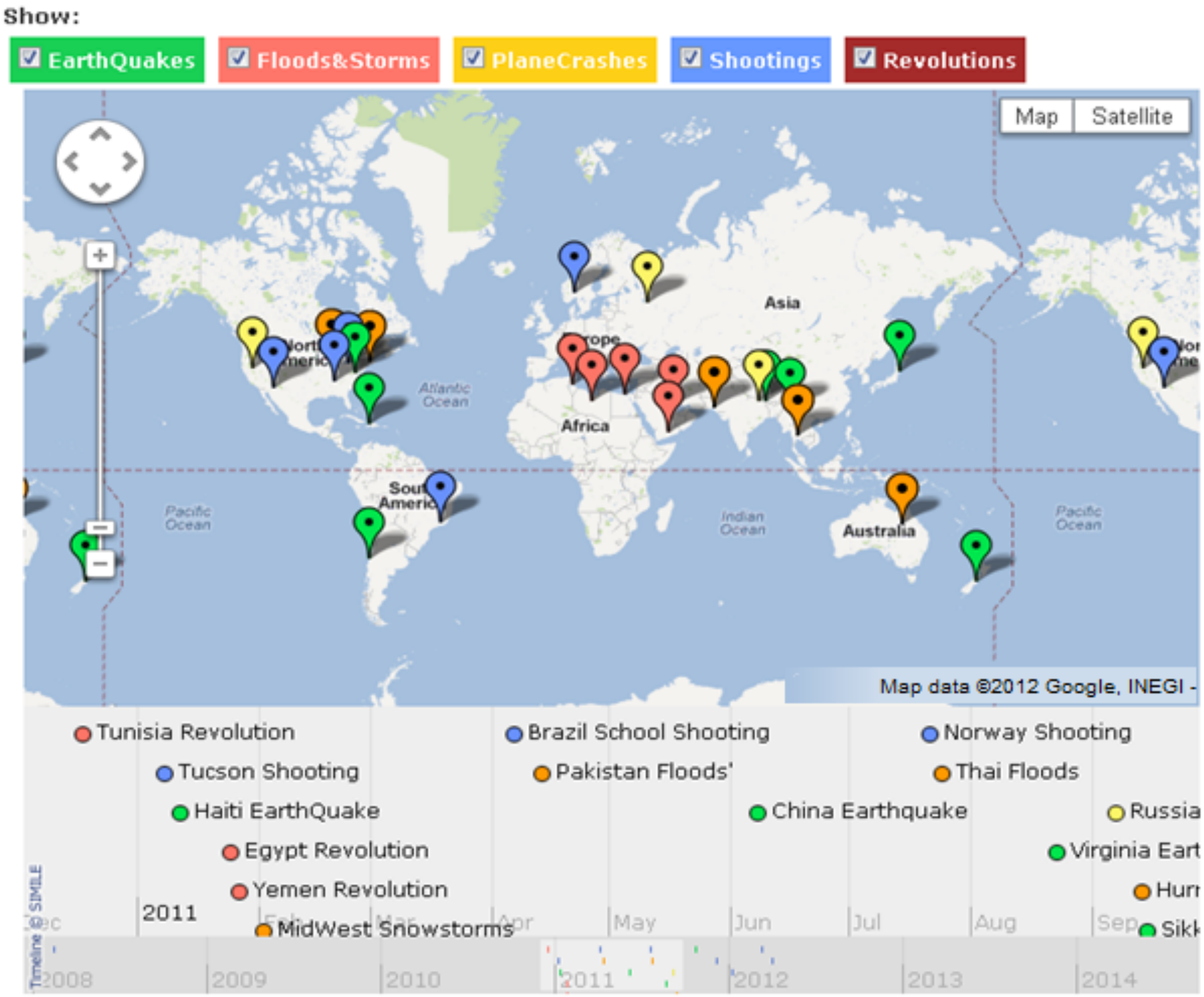
2.5.2.2.3. Filtering and Indexing

The records that are extracted go through a filtering process as discussed in Sections 2.5.2.1.2 and 2.5.2.1.3, using the metadata. The metadata that is collected and saved in the database goes through full text indexing using Apache SOLR [17].

2.5.2.2.4. User Interface: Timeline, Browsing, Profiles

The UI consists of several components: timeline, event browsing, and event profiles.

A timeline visualization of the events is developed using the Timemap [18], a Javascript library, as shown in Figure 4. It includes all of the events we have archived (web and tweet collections). They can be browsed using the timeline, which is created based on the dates events began.



Show:

☑ EarthQuakes ☑ Floods&Storms ☑ PlaneCrashes ☑ Shootings ☑ Revolutions

Figure 4: Timeline map in the UI.

All the events can be browsed through their categories or through the timeline by clicking on the respective event on the map. The selection of an event will give the related videos, images, and webpages. The webpages also can be browsed through facets.

Searching is supported through the Ajax-Solr interface [19] using full-text indexing. The content of the articles, titles, or event collections can be searched. For filtering the results, faceted search also is provided.

2.5.2.3.Monitoring for Water Main Breaks (Tweets, Info. Extraction)

(Note: The content of this section is adapted from the JCDL'2012 poster by Lee et al.) This section describes a prototype of a digital library for water main break identification and visualization. Many utilities rely on an emergency call to detect water main breaks, because breaks are difficult to predict. Collecting the information by call requires time consuming human effort. Furthermore, it is not archived and not shared with others. Collecting and archiving the information by tweets, news, and web resources helps users to identify relevant water main breaks efficiently. In developing this prototype, we extracted location information from text instead of using GPS data. We also describe the importance of tweet visualization by location, and how we show tweets on a map.

The detection and archiving of water main breaks can be useful to utility workers, as well as researchers who analyze and design the infrastructure of cities and towns. There is no automated technique for the collection or identification of water-related events. Current collections are maintained by manual entry of events, which is cumbersome and requires extensive work.

We prototyped a digital library for water main break identification and visualization, which automatically collects data about water-related events from tweets, news, and other web sources; identifies the event location and time; locates each events on a map; and provides a notification system.

In related work, researchers have used Twitter as a human sensor [12]; others have used Twitter to classify text messages and to deliver relevant information to the appropriate personnel during the recovery from the Haiti Earthquake [9]. Another method for extracting detailed location information from tweets is studied in [10], and a different approach to extracting location information using USGS data has been presented in [11].

2.5.2.3.1. Proposed Digital Library

Figure 5 shows an outline of the digital library for water main break identification and visualization. It involves a 3-step process: collecting data about water-related events, identifying key information about the event, and finally visualizing the information.

The digital library holds a collection of data about various water main breaks; this data is collected from tweets, news articles, and various web sources. Keywords such as "water main break" are used as search terms in yourTwapperKeeper [3], to capture water-related events. The data collected from these sources is cleaned up and stored in the digital library for the identification process.

The identification process involves extracting three main pieces of information about every water event: Where, When, and Who. Where has the event occurred? When did the event occur? Who reported it and any information related to it?

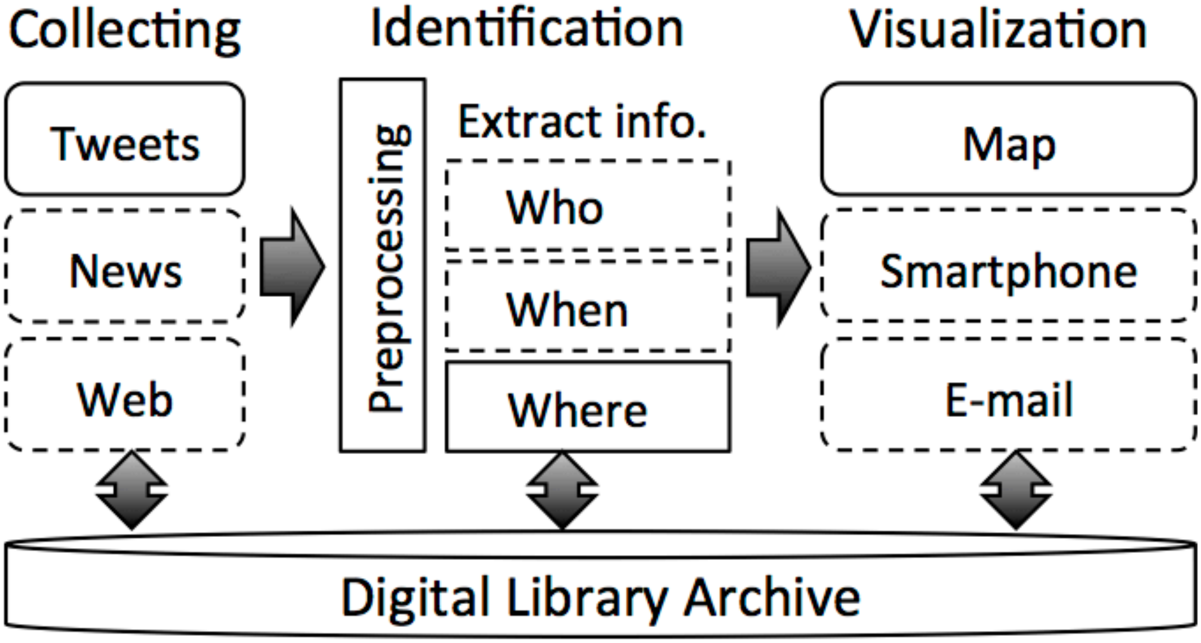


Figure 5. An outline of the digital library

In order to identify *where* the event occurred, a Named Entity Recognizer can be used to identify location information. Identifying *when* an event occurred can be determined by searching for words or number combinations that identify a date or time. In order to identify *who* reported it, machine learning tools can be used. After the identification of where, when, and who, the event is stored in the digital library for visualization.

The visualization and dissemination of the information about a water-related event is an important feature of the digital library. The events are categorized by location and are visualized on Google Maps using Google Fusion Table and geocoding.

In order to visualize relevant tweets within a specific time period, the “Time Window” concept can be applied. For example, “Most recent”, “A week”, “One month”, “One year”, or “5 years” time windows can be created to cover all recent and historical data.

The digital library also can be used to provide a notification system. Users can get information about water-related events that have occurred near them and may affect them. Notifications can be sent by email or directly to a smartphone.

Figure 5 shows the proposed DL; solid lines indicate parts completed in the prototype. In the collecting phase our prototype uses tweet messages, in the identification phase the location information is identified to answer “where”, and in the visualization phase the events are plotted on maps.

2.5.2.3.2. Prototype

Collecting: To collect data from Twitter.com, *yourTwapperKeeper* is used, which is an open source tool for collecting tweets [3]. The program is installed on Red Hat Fedora 15

Linux, and a LAMP (Linux, Apache, MySQL and PHP) server was installed to provide a user interface to manage the program.

The tool collects tweets using a *Search API* and a *Stream API*, which are provided by Twitter.com. The *Search API* is used in the archiving process where the tweets that match certain keywords are retrieved and stored in the database. The archive is updated every 5 minutes, where the new relevant tweets are added using the *Stream API,* and the database is updated.

We have been collecting 17 different water-related events using keywords such as water main break, water pipe leak, sewage spill, etc. For our prototype, we selected the most appropriate dataset (keyword: water main break), which has less noise.

Table 9 shows that at most 1.17 percent of tweet messages have GPS location information (longitude, latitude); this is a very small percentage, justifying the need to use another method for location extraction along with the GPS data. All datasets were collected between 10/24/2011 and 1/30/2012.

Table 9. A sample dataset.

Dataset Keyword	Total tweets	# of tweets which have GPS information (percentage)
water main break	13,382	156 (1.17 %)
water pipe leak	967	1 (0.10 %)

Identification: In order to automatically extract location information from tweets, a Named Entity Recognizer (NER) is applied. It can identify people, organizations, and locations from a text. The Stanford NER [8] is a widely used implementation.

The twitter data needs to be cleaned before the location or other detailed information can be extracted from it. Removing special characters, such as '#' and '@', and removing URLs, helps the SNER to find location information more accurately.

SNER extracts location information from each tweet message, and then it returns a set of geonames as location information. The selected dataset has 3333 tweets. Table 10 shows that we get 1,473 relevant tweets by extracting location information from text. In contrast, we only get 36 tweets using GPS data.

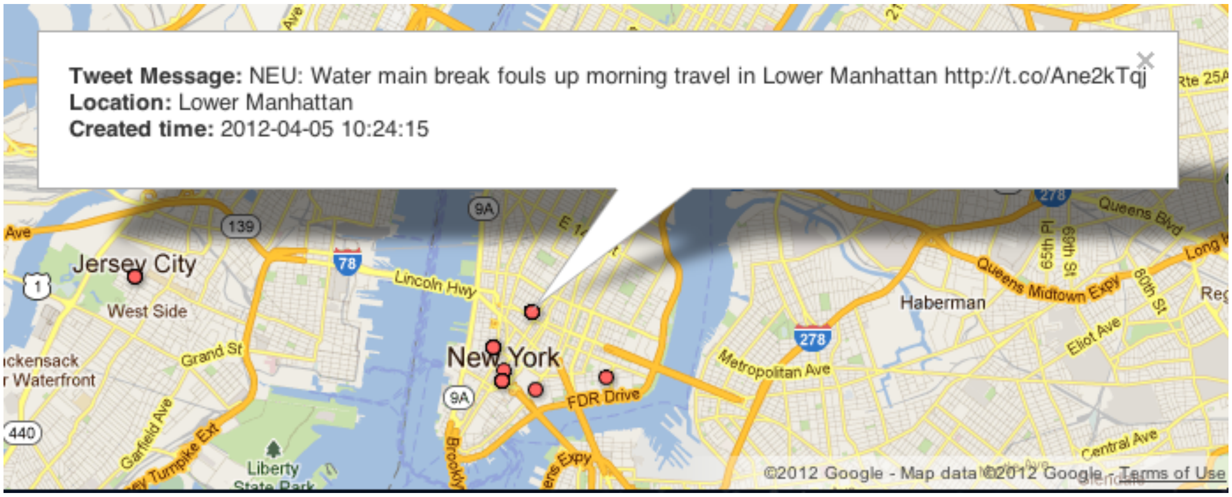
Visualization: The categorization of tweets by location may facilitate the search for relevant information. For example, a Blacksburg utility worker usually monitors water main break events that occur within the town of Blacksburg.

Table 10. Comparison of # of tweets: GPS data vs. location data extracted from text.

Location information type	# of tweets
GPS data (longitude, latitude)	36 (1.08 %)
Location information extracted from text	1,473 (44.19 %)

To visualize tweets on a map, locations (longitude, latitude) are required. The Google Fusion Table, which enables gathering, visualizing, and sharing data online, provides a geocoding function to visualize tweets according to Google Maps locations.

Figure 6 shows an example of the visualized tweets on a map of the New York area, USA. On the Google Maps, each dot represents a tweet event. When a dot is clicked, a pop-up displays a tweet message, location, and created time.



Tweet Message: NEU: Water main break fouls up morning travel in Lower Manhattan http://t.co/Ane2kTqj
Location: Lower Manhattan
Created time: 2012-04-05 10:24:15

Figure 6. An example of the visualized tweets on Google Maps

2.5.3. Conclusion and Future Work

The studies introduced above show the promise of the main components of our disaster/emergency information preservation and research digital library. There are a number of lines of future research. First is to develop an efficient way to extract URLs from tweets when building the archives. Each URL extraction requires an HTTP request, which, given the huge volume of tweets generated, takes a great deal of time, including to expand the shortened form into its original URL form for de-duplication. Parallelism might be a solution for this.

We conducted machine learning based approaches to filter out non-relevant resources. In addition, we have been building a sophisticated rule-based classifier, which examines specific parts of the text and labels it as relevant or non-relevant. This classifier can be integrated seamlessly in our WARC extraction and metadata extraction steps, considering that the machine learning based approach requires much human labor for training set development.

For the water main break identification study, it appears promising to build a system to send messages to people near the affected areas so that they could be aware of potential water shortages or traffic jams.

2.5.4. References Supporting Discussion in 2.5

[1] The Internet Archive. URL http://www.archive.org accessed Jan. 21, 2012.
[2] The open source Heritrix crawler from the Internet Archive at https://webarchive.jira.com/wiki/display/Heritrix/Heritrix accessed Jan. 21, 2012.
[3] yourTwapperKeeper, an open source tweet collection tool at https://github.com/jobrieniii/yourTwapperKeeper accessed Jan. 21, 2012.
[4] Django Web Framework at https://www.djangoproject.com/ accessed Jan. 21, 2012.
[5] T. Risse, J. Masanes, A. A. Benczur, and M. Spaniol, "Turning pure Web Page Storages into Living Web Archives," in *Proc. Cultural Heritage on line. Empowering users: an active role for user communities*" Florence, Italy, pp. 151-155, Dec. 15-16, 2009.
[6] V. Kandylas and A. Dasdan, "The utility of tweeted URLs for web search," in *Proceedings of the 19th international conference on World Wide Web,* Raleigh, North Carolina, USA, 2010, 1127-1128.
[7] A. L. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergency events," International Journal of Emergency Management, vol. 6, pp. 248-260, 2009.
[8] The Stanford NLP group, Stanford NER, accessed Jan 2012 http://nlp.stanford.edu/software/CRF-NER.shtml.
[9] C. Caragea, M. McNeese, A. Jaiswal, G. Traylor, H. Kim, P. Mitra, D. Wu, A. Tapia, C. Giles, J. Jansen, and J. Yen. Classifying text messages for the Haiti earthquake. In *8th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pages 1-10, Lisbon, Portugal, May 8-11, 2011.
[10] W. J. Corvey, S. Vieweg, T. Rood, and M. Palmer. Twitter in mass emergency: what NLP techniques can contribute. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 23–24, Stroudsburg, PA, USA, 2010.
[11] S. M. Paradesi. Geotagging tweets using their content. In *R. C. Murray and P. M. McCarthy, editors, FLAIRS Conference*. pages 355-356, AAAI Press, 2011.
[12] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web* (WWW '10). ACM, New York, NY, USA, 851-860. DOI=10.1145/1772690.1772777.
[13] WEKA Machine Learning Toolkit. http://www.cs.waikato.ac.nz/ml/weka/ accessed April 27, 2012.
[14] Hanzo warc-tools URL http://code.hanzoarchives.com/warc-tools/overview accessed May 4, 2012.
[15] WARC file format URL http://archive-access.sourceforge.net/warc/warc_file_format-0.16.html
accessed May 4, 2012.

[16] MIME types URL http://en.wikipedia.org/wiki/Internet_media_type accessed May 4, 2012.
[17] Apache Solr URL http://lucene.apache.org/ accessed May 4, 2012.
[18] Time map URL http://code.google.com/p/timemap/ accessed May 4, 2012.
[19] Ajax-Solr https://github.com/evolvingweb/ajax-solr accessed May 4, 2012.

3. Publications and Products

3.1.Publications During Year 3

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 many works: http://www.ctrnet.net/publications

For 2011-2012 we published the following:

Book:

- Edward A. Fox, Marcos Andre Goncalves, and Rao Shen. Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach. Morgan & Claypool Publishers, San Francisco, 2012, ISBN paperback 9781608459100, ebook 9781608459117, DOI 10.2200/S00407ED1V01Y201203CRM004, 180 pages

Journal:

- Kavanaugh, A.L., Fox, E.A., Sheetz, S.D., Yang, S., Li, L.T., Whalen, T., Shoemaker, D. J., Natsev, P., Xie L. Social Media Use by Government: From the Routine to the Critical. Government Information Quarterly (GIQ), Elsevier, in press

Refereed Conference Papers:

- Kavanaugh, A.L., Sheetz, S.D., Hassan, R., Yang, S., Elmongui, H.G., Fox, E.A., Magdy, M., Shoemaker, D. (2012). Between a Rock and a Cell Phone: Communication and Information Use During the Egyptian Uprising. In Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012). Apr. 22-25. Vancouver, Canada, 10 pages

Posters:

- Sunshin Lee, Noha Elsherbiny, Edward A. Fox. A Digital Library for Water Main Break Identification and Visualization. Poster in Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012), Washington D.C., June 10-14, 2012, 335-336, http://dx.doi.org/10.1145/2232817.2232878

- Seungwon Yang, Kiran Chitturi, Gregory Wilson, Mohamed Magdy, and Edward A. Fox. A Study of Automation from Seed URL Generation to Focused Web Archive Development: The CTRnet Context. Poster in Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012), Washington D.C., June 10-14, 2012, 341-342, http://dx.doi.org/10.1145/2232817.2232881

3.2. Products

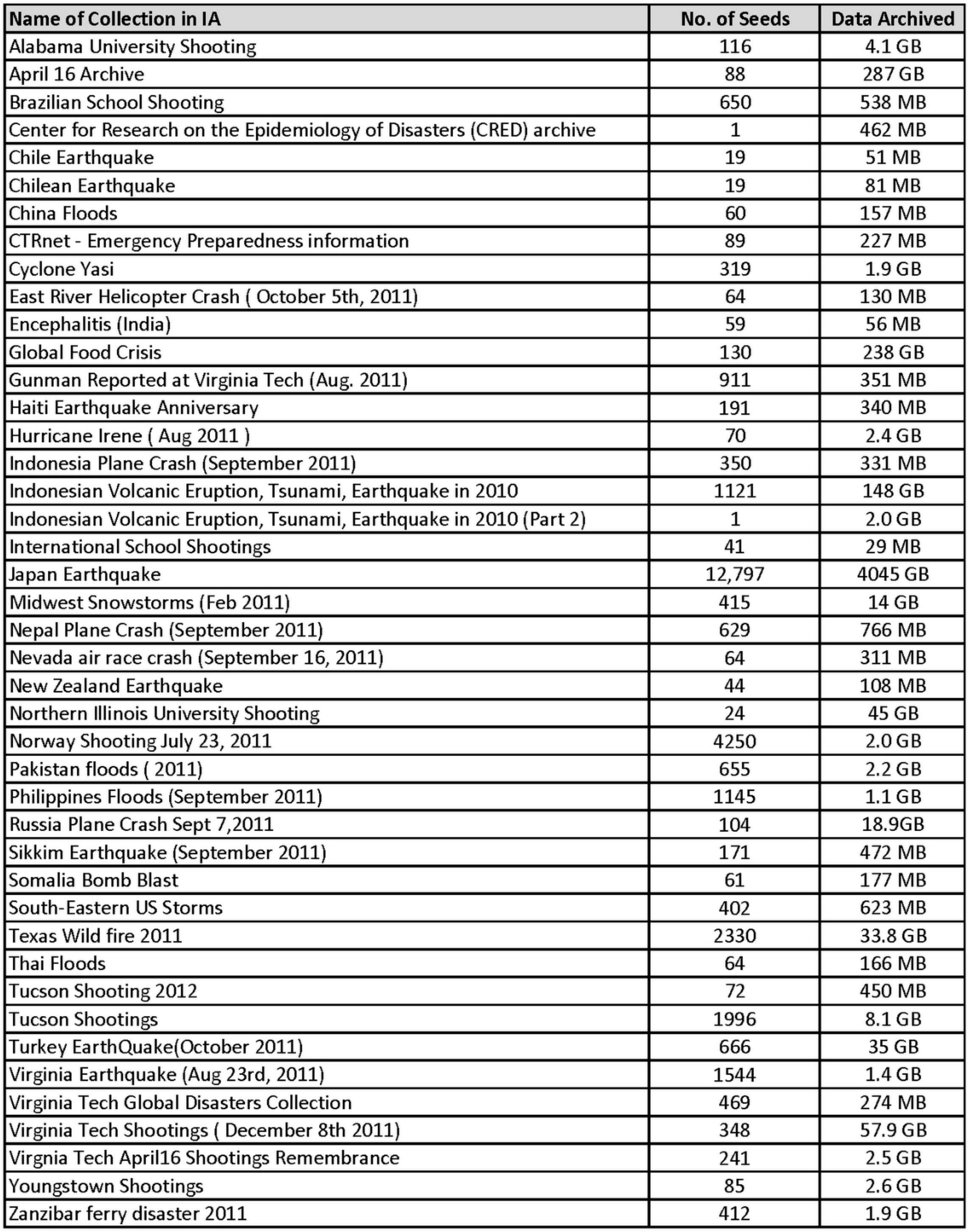
3.2.1. Refined CTRnet Homepage

We refined the project homepage at http://www.ctrnet.net/ (see the online version).

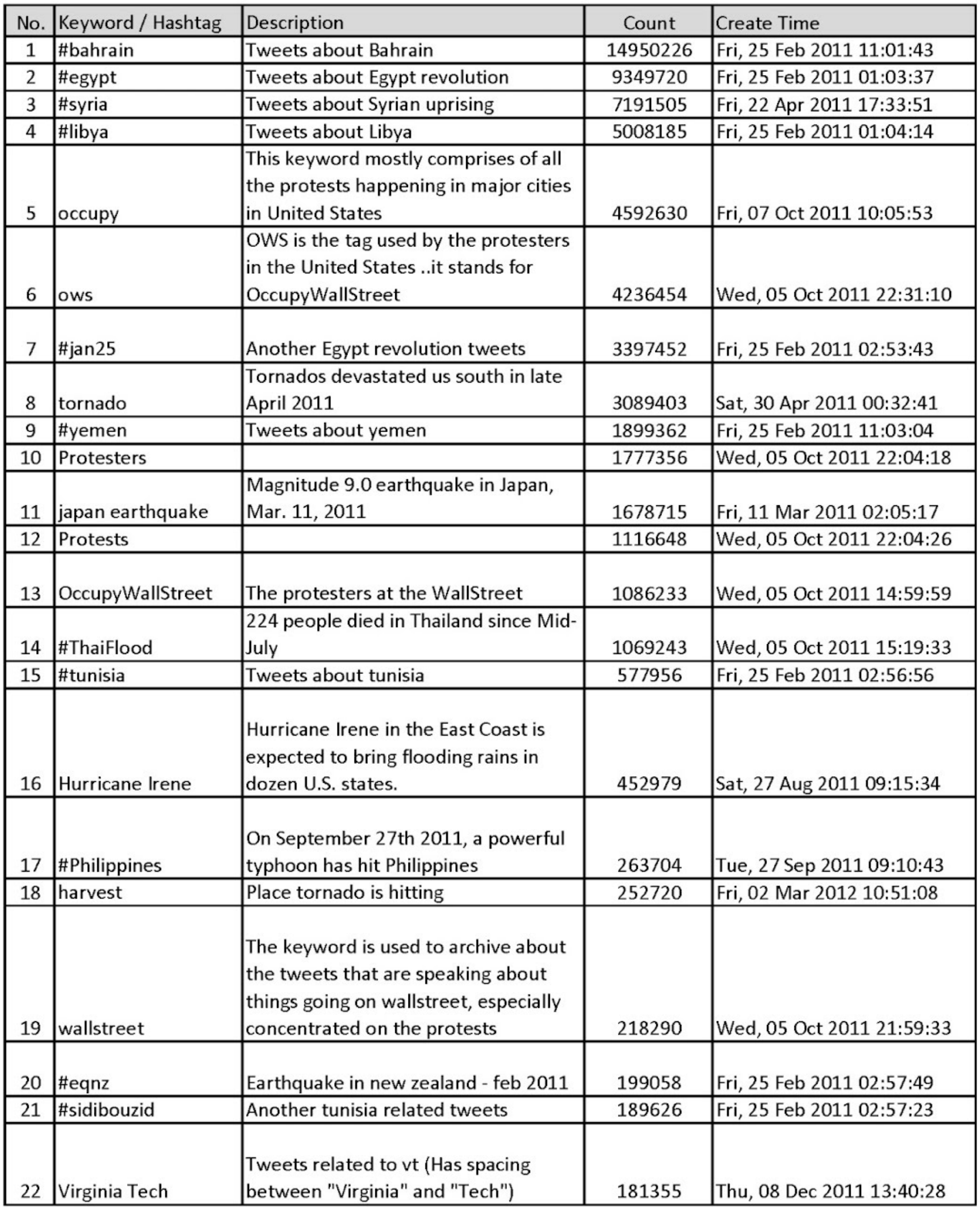
3.2.2. CTR Collections

Table 11 shows Web pages in part a; part b shows tweets in 4.5 pages.

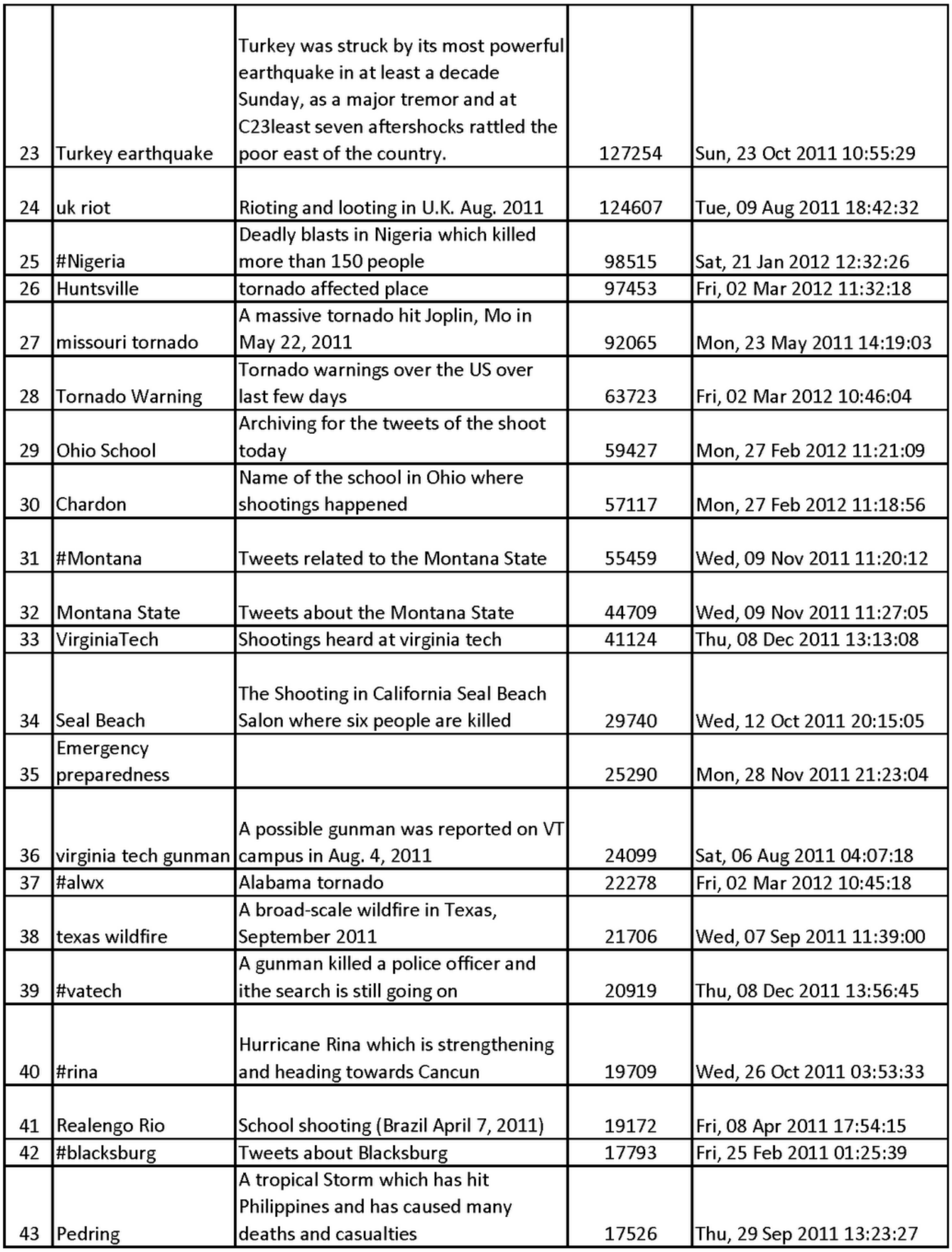
Table 11. a) Web Pages at Internet Archive, b) Tweets



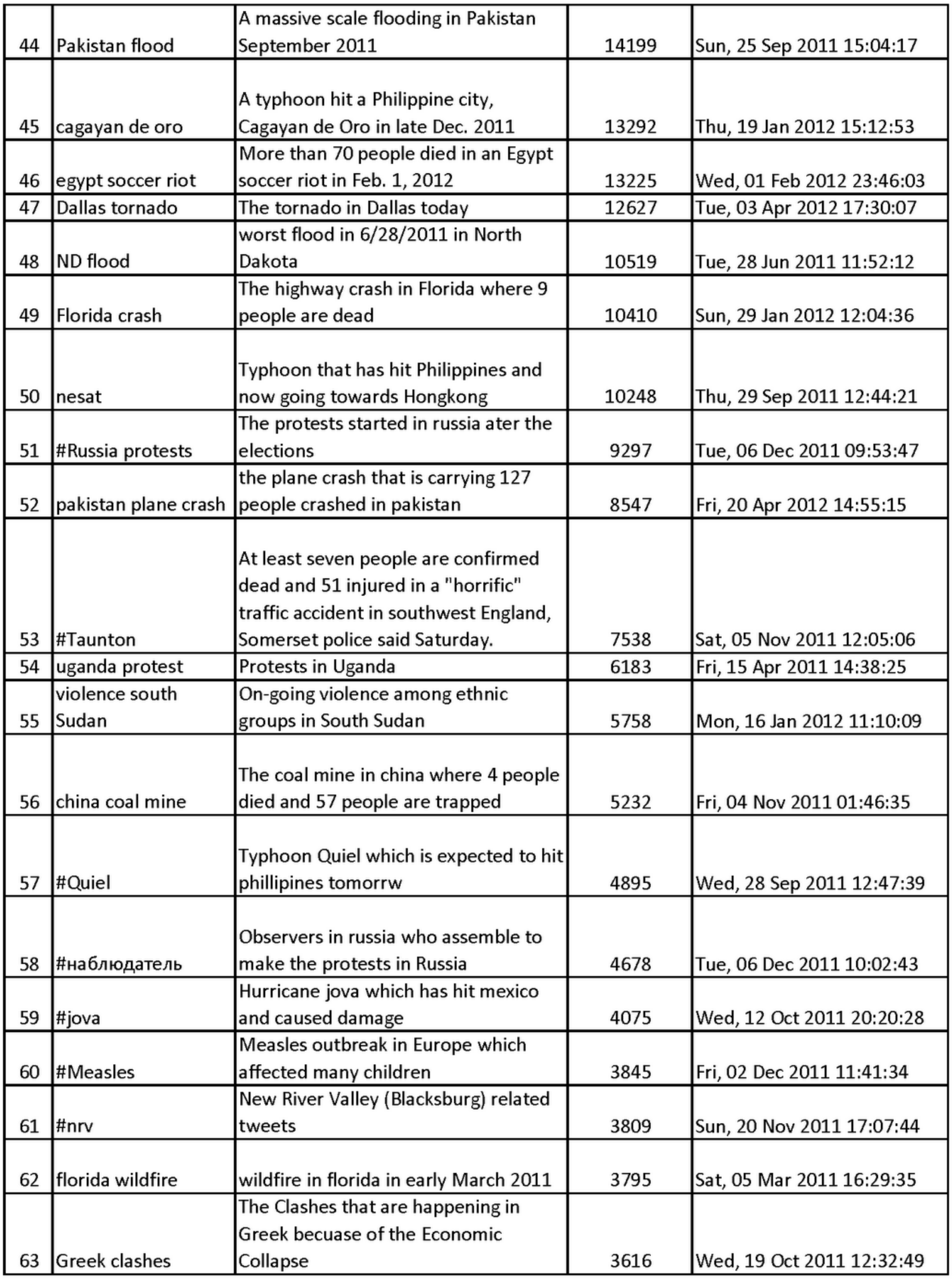
Name of Collection in IA	No. of Seeds	Data Archived
Alabama University Shooting	116	4.1 GB
April 16 Archive	88	287 GB
Brazilian School Shooting	650	538 MB
Center for Research on the Epidemiology of Disasters (CRED) archive	1	462 MB
Chile Earthquake	19	51 MB
Chilean Earthquake	19	81 MB
China Floods	60	157 MB
CTRnet - Emergency Preparedness information	89	227 MB
Cyclone Yasi	319	1.9 GB
East River Helicopter Crash (October 5th, 2011)	64	130 MB
Encephalitis (India)	59	56 MB
Global Food Crisis	130	238 GB
Gunman Reported at Virginia Tech (Aug. 2011)	911	351 MB
Haiti Earthquake Anniversary	191	340 MB
Hurricane Irene (Aug 2011)	70	2.4 GB
Indonesia Plane Crash (September 2011)	350	331 MB
Indonesian Volcanic Eruption, Tsunami, Earthquake in 2010	1121	148 GB
Indonesian Volcanic Eruption, Tsunami, Earthquake in 2010 (Part 2)	1	2.0 GB
International School Shootings	41	29 MB
Japan Earthquake	12,797	4045 GB
Midwest Snowstorms (Feb 2011)	415	14 GB
Nepal Plane Crash (September 2011)	629	766 MB
Nevada air race crash (September 16, 2011)	64	311 MB
New Zealand Earthquake	44	108 MB
Northern Illinois University Shooting	24	45 GB
Norway Shooting July 23, 2011	4250	2.0 GB
Pakistan floods (2011)	655	2.2 GB
Philippines Floods (September 2011)	1145	1.1 GB
Russia Plane Crash Sept 7,2011	104	18.9GB
Sikkim Earthquake (September 2011)	171	472 MB
Somalia Bomb Blast	61	177 MB
South-Eastern US Storms	402	623 MB
Texas Wild fire 2011	2330	33.8 GB
Thai Floods	64	166 MB
Tucson Shooting 2012	72	450 MB
Tucson Shootings	1996	8.1 GB
Turkey EarthQuake(October 2011)	666	35 GB
Virginia Earthquake (Aug 23rd, 2011)	1544	1.4 GB
Virginia Tech Global Disasters Collection	469	274 MB
Virginia Tech Shootings (December 8th 2011)	348	57.9 GB
Virgnia Tech April16 Shootings Remembrance	241	2.5 GB
Youngstown Shootings	85	2.6 GB
Zanzibar ferry disaster 2011	412	1.9 GB



No.	Keyword / Hashtag	Description	Count	Create Time
1	#bahrain	Tweets about Bahrain	14950226	Fri, 25 Feb 2011 11:01:43
2	#egypt	Tweets about Egypt revolution	9349720	Fri, 25 Feb 2011 01:03:37
3	#syria	Tweets about Syrian uprising	7191505	Fri, 22 Apr 2011 17:33:51
4	#libya	Tweets about Libya	5008185	Fri, 25 Feb 2011 01:04:14
5	occupy	This keyword mostly comprises of all the protests happening in major cities in United States	4592630	Fri, 07 Oct 2011 10:05:53
6	ows	OWS is the tag used by the protesters in the United States ..it stands for OccupyWallStreet	4236454	Wed, 05 Oct 2011 22:31:10
7	#jan25	Another Egypt revolution tweets	3397452	Fri, 25 Feb 2011 02:53:43
8	tornado	Tornados devastated us south in late April 2011	3089403	Sat, 30 Apr 2011 00:32:41
9	#yemen	Tweets about yemen	1899362	Fri, 25 Feb 2011 11:03:04
10	Protesters		1777356	Wed, 05 Oct 2011 22:04:18
11	japan earthquake	Magnitude 9.0 earthquake in Japan, Mar. 11, 2011	1678715	Fri, 11 Mar 2011 02:05:17
12	Protests		1116648	Wed, 05 Oct 2011 22:04:26
13	OccupyWallStreet	The protesters at the WallStreet	1086233	Wed, 05 Oct 2011 14:59:59
14	#ThaiFlood	224 people died in Thailand since Mid-July	1069243	Wed, 05 Oct 2011 15:19:33
15	#tunisia	Tweets about tunisia	577956	Fri, 25 Feb 2011 02:56:56
16	Hurricane Irene	Hurricane Irene in the East Coast is expected to bring flooding rains in dozen U.S. states.	452979	Sat, 27 Aug 2011 09:15:34
17	#Philippines	On September 27th 2011, a powerful typhoon has hit Philippines	263704	Tue, 27 Sep 2011 09:10:43
18	harvest	Place tornado is hitting	252720	Fri, 02 Mar 2012 10:51:08
19	wallstreet	The keyword is used to archive about the tweets that are speaking about things going on wallstreet, especially concentrated on the protests	218290	Wed, 05 Oct 2011 21:59:33
20	#eqnz	Earthquake in new zealand - feb 2011	199058	Fri, 25 Feb 2011 02:57:49
21	#sidibouzid	Another tunisia related tweets	189626	Fri, 25 Feb 2011 02:57:23
22	Virginia Tech	Tweets related to vt (Has spacing between "Virginia" and "Tech")	181355	Thu, 08 Dec 2011 13:40:28



23	Turkey earthquake	Turkey was struck by its most powerful earthquake in at least a decade Sunday, as a major tremor and at C23least seven aftershocks rattled the poor east of the country.	127254	Sun, 23 Oct 2011 10:55:29
24	uk riot	Rioting and looting in U.K. Aug. 2011	124607	Tue, 09 Aug 2011 18:42:32
25	#Nigeria	Deadly blasts in Nigeria which killed more than 150 people	98515	Sat, 21 Jan 2012 12:32:26
26	Huntsville	tornado affected place	97453	Fri, 02 Mar 2012 11:32:18
27	missouri tornado	A massive tornado hit Joplin, Mo in May 22, 2011	92065	Mon, 23 May 2011 14:19:03
28	Tornado Warning	Tornado warnings over the US over last few days	63723	Fri, 02 Mar 2012 10:46:04
29	Ohio School	Archiving for the tweets of the shoot today	59427	Mon, 27 Feb 2012 11:21:09
30	Chardon	Name of the school in Ohio where shootings happened	57117	Mon, 27 Feb 2012 11:18:56
31	#Montana	Tweets related to the Montana State	55459	Wed, 09 Nov 2011 11:20:12
32	Montana State	Tweets about the Montana State	44709	Wed, 09 Nov 2011 11:27:05
33	VirginiaTech	Shootings heard at virginia tech	41124	Thu, 08 Dec 2011 13:13:08
34	Seal Beach	The Shooting in California Seal Beach Salon where six people are killed	29740	Wed, 12 Oct 2011 20:15:05
35	Emergency preparedness		25290	Mon, 28 Nov 2011 21:23:04
36	virginia tech gunman	A possible gunman was reported on VT campus in Aug. 4, 2011	24099	Sat, 06 Aug 2011 04:07:18
37	#alwx	Alabama tornado	22278	Fri, 02 Mar 2012 10:45:18
38	texas wildfire	A broad-scale wildfire in Texas, September 2011	21706	Wed, 07 Sep 2011 11:39:00
39	#vatech	A gunman killed a police officer and ithe search is still going on	20919	Thu, 08 Dec 2011 13:56:45
40	#rina	Hurricane Rina which is strengthening and heading towards Cancun	19709	Wed, 26 Oct 2011 03:53:33
41	Realengo Rio	School shooting (Brazil April 7, 2011)	19172	Fri, 08 Apr 2011 17:54:15
42	#blacksburg	Tweets about Blacksburg	17793	Fri, 25 Feb 2011 01:25:39
43	Pedring	A tropical Storm which has hit Philippines and has caused many deaths and casualties	17526	Thu, 29 Sep 2011 13:23:27



44	Pakistan flood	A massive scale flooding in Pakistan September 2011	14199	Sun, 25 Sep 2011 15:04:17
45	cagayan de oro	A typhoon hit a Philippine city, Cagayan de Oro in late Dec. 2011	13292	Thu, 19 Jan 2012 15:12:53
46	egypt soccer riot	More than 70 people died in an Egypt soccer riot in Feb. 1, 2012	13225	Wed, 01 Feb 2012 23:46:03
47	Dallas tornado	The tornado in Dallas today	12627	Tue, 03 Apr 2012 17:30:07
48	ND flood	worst flood in 6/28/2011 in North Dakota	10519	Tue, 28 Jun 2011 11:52:12
49	Florida crash	The highway crash in Florida where 9 people are dead	10410	Sun, 29 Jan 2012 12:04:36
50	nesat	Typhoon that has hit Philippines and now going towards Hongkong	10248	Thu, 29 Sep 2011 12:44:21
51	#Russia protests	The protests started in russia ater the elections	9297	Tue, 06 Dec 2011 09:53:47
52	pakistan plane crash	the plane crash that is carrying 127 people crashed in pakistan	8547	Fri, 20 Apr 2012 14:55:15
53	#Taunton	At least seven people are confirmed dead and 51 injured in a "horrific" traffic accident in southwest England, Somerset police said Saturday.	7538	Sat, 05 Nov 2011 12:05:06
54	uganda protest	Protests in Uganda	6183	Fri, 15 Apr 2011 14:38:25
55	violence south Sudan	On-going violence among ethnic groups in South Sudan	5758	Mon, 16 Jan 2012 11:10:09
56	china coal mine	The coal mine in china where 4 people died and 57 people are trapped	5232	Fri, 04 Nov 2011 01:46:35
57	#Quiel	Typhoon Quiel which is expected to hit phillipines tomorrw	4895	Wed, 28 Sep 2011 12:47:39
58	#наблюдатель	Observers in russia who assemble to make the protests in Russia	4678	Tue, 06 Dec 2011 10:02:43
59	#jova	Hurricane jova which has hit mexico and caused damage	4075	Wed, 12 Oct 2011 20:20:28
60	#Measles	Measles outbreak in Europe which affected many children	3845	Fri, 02 Dec 2011 11:41:34
61	#nrv	New River Valley (Blacksburg) related tweets	3809	Sun, 20 Nov 2011 17:07:44
62	florida wildfire	wildfire in florida in early March 2011	3795	Sat, 05 Mar 2011 16:29:35
63	Greek clashes	The Clashes that are happening in Greek becuase of the Economic Collapse	3616	Wed, 19 Oct 2011 12:32:49

64	oikos university shooting	A school shooting occurred at the Oikos University Shooting in April 2, 2012	3445	Mon, 02 Apr 2012 22:16:22
65	mexican helicopter crash	The helicopter crash in mexico where mexicos interior minister died along with the other 8 people	3303	Fri, 11 Nov 2011 15:55:33
66	tornado norman	tornado in oklahoma norman	3299	Fri, 13 Apr 2012 20:57:43
67	sikkim earthquake	On september 18th 2011 an earthquake centered within the Kanchenjunga Conservation Area, near the border of Nepal and the Indian state of Sikkim has occured where 111 people died and major casualties reported	2818	Wed, 28 Sep 2011 08:36:27
68	#swva	SouthWestVirginia (VTS)	2762	Wed, 23 Nov 2011 17:39:15
69	ohio school shooting	A campus shooting occurred at Chardon high school in Ohio in Feb. 2012	2757	Mon, 12 Mar 2012 16:35:59
70	ometepec mexico earthquake	Magnitude 7.4 earthquake hit Mexico in March, 20, 2012	1877	Wed, 21 Mar 2012 10:43:49
71	swaziland protest	Ongoing protest and crackdown in Swaziland	1765	Fri, 15 Apr 2011 14:35:17
72	Papua New Guinea ferry	The ferry which sank in the Papua New Guinea	1689	Thu, 02 Feb 2012 20:11:56
73	emergency management response		1374	Mon, 28 Nov 2011 21:19:52
74	russia plane crash	The plane crash in Siberia, Russia where 31 people died	1298	Mon, 02 Apr 2012 10:00:55
75	#Oakland shooting	The shooting at Okis University where 6 people died	984	Mon, 02 Apr 2012 16:58:47
76	#EastRiver	The private helicopter that crashed in to East River, killing one passenger and injuring three	916	Wed, 05 Oct 2011 14:32:13
77	#Oklahoma quake	The 5.6 magnitude earthquake in Oklahoma	869	Sun, 06 Nov 2011 11:51:32
78	florida fires	The wild fires in florida	743	Tue, 10 Apr 2012 21:03:55
79	emergency management plan	Tweet archive to collect emergency management plans	599	Mon, 28 Nov 2011 21:16:51
80	cyclone irina	A tropical cyclone Irina hit Madagascar killing more than 70 people in March 2012	513	Wed, 07 Mar 2012 14:29:10
81	#virginiatechshooting		467	Thu, 08 Dec 2011 17:40:03

82	Iran earthquake	Magnitude 5.5 earthquake occurred in Iran May 3, 2012	397	Thu, 03 May 2012 12:06:00
83	emergency management recovery		359	Mon, 28 Nov 2011 21:19:23
84	tucson shooting anniversary	Anniversary of the Tucson shooting	251	Wed, 11 Jan 2012 14:35:37
85	Emergency mitigation		206	Mon, 28 Nov 2011 21:22:44
86	pkfloods2011	Floods of Pakistan 2011	192	Thu, 29 Sep 2011 13:26:43
87	giffords resign	Gabrielle Giffords resigns late Jan. 2012	156	Thu, 26 Jan 2012 14:40:13

4. Contributions

4.1.Contributions within Discipline

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 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 reviewing research on the use and impact of social media in emergency or crisis situations.

Our collected information is useful only when visible, and when 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. Our contribution in this area of applied computing has been developing methodologies for tweet collection, preprocessing, and visualization; these were presented first in a tutorial session of the Digital Government conference on June 11, 2011. At JCDL 2012, a more highly automated version was presented, along with related software to show the location of water main breaks. Other contributions of our work include in connection with geo-parsing, information retrieval, supervised machine learning (especially classification), and archiving.

4.2. Contributions to Other Disciplines

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 findings of our study should advance technologies and systems in social media analysis, and inform day-to-day civil society.

Thus, 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 studied by the team; these technologies might be used to address those needs. Our contribution to sociology and political science, in this area of our work, has been the identification of these requirements and of two possible approaches to address the requirements.

Over 2011-2012, we also collected useful data regarding CTR events, including online and survey information related to Egypt, Tunisia, and Mexico, among other locations. This is helpful in understanding social change, revolutions, technology adoption, communications, and a number of social sciences. The water main break effort related to civil engineering.

4.3.Contributions to Human Resource Development

A growing body of students has been joining the CTRnet project. 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, GRAs, as well as other graduate and undergraduate students) have had opportunities to attend IA training sessions and Webinars about newly developed social media collection and analysis tools. Those enhance their knowledge and skills with regard to building digital collections from webpages as well as social media information. Further, in 2011-2012, seven additional partners collaborated, allowing faculty, students, and professionals to develop additional skills.

4.4. Contributions to Resources for Research and Education

Our project has developed content to aid in research and education. The collections at the Internet Archive, and our local collections of tweets, represent unique archives, at appropriate scale, regarding Web and Twitter information about recent crises and tragedies. Additional collections include from studies of Egypt, Tunisia, Mexico, etc., and from our collaboration with IBM and Arlington County, Virginia.

4.5. Contributions Beyond Science and Engineering

Our project is contributing improved methods of collecting and archiving, which are important in all walks of life, in the Internet Age. Further, we are making available and accessible a broad range of information about crises, tragedies, and disasters, including both man-made and natural events, along with efforts made toward recovery. These can help with human life and welfare, with planning and policy making, with humanitarian and compassionate activities, and, we hope, may in some cases alleviate future suffering. The importance of these contributions is clearly articulated in the report "Computing for Disasters" (http://cra.org/ccc/docs/init/computingfordisasters.pdf), edited by Robin Murphy and Trevor Darrell, reporting on the April 24, 2012 workshop sponsored by NSF and CRA/CCC, for which PI Fox served on the Steering Committee and Co-PI Kavanaugh aided as a workshop attendee.