Real-time Archiving of Spontaneous Events (Use-Case : Hurricane Sandy)

Kiran Chitturi,
Digital Library Research Laboratory,
Virginia Tech.
Background

• Crisis, Tragedy and Recovery Network (CTRnet) [1]
  – Dr. Edward A. Fox (Principal Investigator)
  – 3+ year project started in August 2009 with grant from NSF
  – Researching the problems of integrating content, community, and services related to crises, tragedies, and recovery (CTR)
  – Building a domain specific distributed digital library, integrating information from many sources, with ontology support

• Partnership with Internet Archive for archiving data
Background

• Data collected using Archive-it
  – 8.8 TB + (43 archive collections)

• Collections
  – School Shootings
  – Earthquakes
  – Hurricanes
  – Floods
  – Plane Crashes
Hurricane Sandy [2]

- Devastating tropical cyclone
  - Formed: October 22, 2012
  - Dissipated: October 31, 2012
  - Damage: At least $65.6 billion
  - Fatalities: 253 total
  - Spanning Seven Countries
Archiving Hurricane Sandy - Tweet Collection

• Tweet collection based on hash tags and keywords
  – Example: #HurricaneSandy, #Sandy, Sandy prepare, Sandy Help, Sandy Shelter, Hurricane Sandy ...

• Our Sample
  – Collecting tweets since October 25, 2012
  – 8 million tweets

• Seed Extraction through tweets

@...To friends and family up north, please take pre-caution and hope it all works out. This hurricane is shaping up to b... http://t.co/HiNSDbB4
Archiving Hurricane Sandy - Seeds from Tweets

- Seeds – tweets statistics (Oct.25 – Nov.11)
  - 338,505 distinct seeds from 49,083 domains
  - Some top domains and seeds

<table>
<thead>
<tr>
<th>Domain</th>
<th>Unique Seeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instagram.com</td>
<td>43098</td>
</tr>
<tr>
<td>Twitter.com</td>
<td>40863 (pictures)</td>
</tr>
<tr>
<td>Facebook.com</td>
<td>26176</td>
</tr>
<tr>
<td>youtube.com</td>
<td>12596</td>
</tr>
<tr>
<td>huffintonpost.com</td>
<td>4519</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>3305</td>
</tr>
</tbody>
</table>
Archiving Hurricane Sandy - Seed Collection and Crawling

- Collecting seeds from public
  - Crawl type based on user choice

- Seeds extracted from tweets
  - Crawl type: single page

- Google news real-time coverage
  - Crawl type: News/RSS
  - Crawl Frequency: daily twice

- Popular news, business and weather portals
  - Crawl type: News/RSS
  - Crawl Frequency: daily once
Archiving Hurricane Sandy - Seed Collection and Crawling (continued)

• Government, Emergency Organizations
  – Crawl type: whole website
  – Crawl Frequency: twice daily

• Facebook, twitter pages of local governments, and non-profit organizations
  – Crawl type: News/RSS
  – Crawl Frequency: daily once

• 83 Facebook pages about recovery of local communities (prepared by Steven Clift [3])
  – Crawl type: News/RSS
  – Crawl Frequency: weekly once
Archiving Hurricane Sandy – Weather maps from Wayback machine

- 10/27/2012 – 12.05 UTC
- 10/27/2012 – 19.43 UTC
- 10/28/2012 – 19.41 UTC
- 10/29/2012 – 07.42 UTC
- 10/29/2012 – 23.46 UTC
- 10/30/2012 – 12.45 UTC
- 10/30/2012 – 23.23 UTC
- 10/31/2012 – 11.22 UTC

Tropical Storm Warning
Hurricane Warning
Storm Warning
Red Flag Warning
Winter Storm Warning
Lakeshore Flood Advisory
Existing Problems: Real-time archiving

- Big collection sizes with frequent crawls
  - Distributed computing for crawling, filtering and analytics

- Seed collection after the event (recovery)
  - Crowd sourcing and identifying local sources

- Covering varied sources of information
  - Information Source graph (covering organizations and websites per geographic area)

- Disappearing webpages over time
  - Identifying and covering resources in time

- Spam
  - Classification and Filtering
Future Goals

• Use Lucid Works Big Data Software with the archive data
  – Integrated platform for solving information access problems with the tools for Search, Discovery, and Analysis of massive content sets

• Digital Library to provide access to the data
  – Features
    • Faceted Search
    • Image and Video extraction
    • Filtered, relevant documents per collection
Summary

• Real-time Archiving
  – Seed Collection
  – Big Data
  – Access

• Collaboration
  – Spotting Events
  – Access to data and software
  – IDEAL proposal
References and Resources

1. www.ctrnet.net
2. www.en.wikipedia.org/wiki/Hurricane_Sandy
Analyzing and Visualizing (Tweet) Archives

Presenter: Seungwon Yang

12/3/2012

Internet Archive Partners Meeting
Overview

• Description
  – Four Phase Model of Emergency Management

• Data Flow Diagram
  – Selection/cleaning \(\rightarrow\) classification \(\rightarrow\) visualization

• PhaseVis Components
  – What, When, Where, and Who

• A Sample Use Case
Tweet Analysis & Visualization Study

• Title:

*PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management Through the Lens of Social Media*

• Conference Submitted:
The 10th International ISCRAM Conference’13
(Information Systems for Crisis Response and Management)

• Prototype Link:
• [http://spare05.dlib.vt.edu/~ctrvis/phasevis/](http://spare05.dlib.vt.edu/~ctrvis/phasevis/)
## Four Phases of Emergency Management

<table>
<thead>
<tr>
<th>Phases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitigation</td>
<td>Activities providing a critical foundation in the effort to reduce or eliminate the risks to persons or property, and to lessen or avoid the impact of a disaster</td>
</tr>
<tr>
<td>Preparedness</td>
<td>Range of deliberate, critical tasks and activities necessary to build, sustain, and improve the operational capability to prevent, protect against, respond to, and recover from emergency and disasters.</td>
</tr>
<tr>
<td>Response</td>
<td>Immediate and ongoing actions following an emergency or disaster, which are designed to provide emergency assistance for casualties such as search/rescue, emergency shelter, medical care, mass feeding.</td>
</tr>
<tr>
<td>Recovery</td>
<td>The development, coordination, and execution of service and site restoration plans. The reconstruction of government operations and services.</td>
</tr>
</tbody>
</table>
### Four Phases of Emergency Management - Tweet Examples -

<table>
<thead>
<tr>
<th>Phase (Category)</th>
<th>Example Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Response</td>
<td>Red Cross provided 20,000+ overnight shelter stays and served 400,000+ meals/snacks to help those impacted by #Hurricane #Isaac. THANK YOU!!</td>
</tr>
<tr>
<td>2: Recovery</td>
<td>Hurricane #Isaac relief efforts in the Gulf continue. Pls help! Txt REDCROSS to 90999, to give $10.</td>
</tr>
<tr>
<td>3: Mitigation</td>
<td>FEMA mitigation to offer rebuilding tips in St. John Parish at Home Depot in LaPlace, La. <a href="http://t.co/8UFJFEz5">http://t.co/8UFJFEz5</a> #Isaac</td>
</tr>
<tr>
<td>4: Preparedness</td>
<td>RT @jeffiel: Seriously, everybody should have this #RedCross app on their phone. What to do in #emergencies <a href="http://t.co/MtrQcOo1">http://t.co/MtrQcOo1</a> #isaac</td>
</tr>
</tbody>
</table>
Visualizing the Four Phases of Emergency Management

Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013

T. Comes, F. Fiedrich, S. Fortier, J. Geldermann and L. Yang, eds.

Table 4. Number of tweets with disaster organization names, both in the target set and in our training set (within parentheses).

* Note: some tweets contain more than a single organization name; therefore, the sum of the numbers in column 2, 3, and 4 is greater than the total (column 5: tweets that have at least one organization name).

Figure 1. A data flow diagram: selection, classification, and visualization. Details are described in the following sections.

Since tweets often embed shortened links to webpages, those webpages might provide essential information required to understand the tweets themselves. Thus, we followed the first embedded link (if it exists) and extracted the title of the webpage. Then we concatenate the resource title with the tweet to form a richer tweet text with the intent of strengthening the distinguishability of the classifier. As a preprocessing step, we removed non-informative English stopwords, special characters, and URLs from tweets, and applied the Porter stemmer provided in the WEKA toolkit.

3.2. Classification

Each tweet in our training set was manually categorized as one of '1:Response', '2:Recovery', '3:Mitigation', '4:Preparedness', and '5:Other' based on the activities described in the content. Selected labeled training examples are presented in Table 5. We excluded 331 tweets categorized as '5:Other'. In total, 1,121 tweets were used as our training set.

<table>
<thead>
<tr>
<th>Phase (Category)</th>
<th>Example Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Response</td>
<td>Red Cross provided 20,000+ overnight shelter stays and served 400,000+ meals/snacks to help those impacted by #Hurricane #Isaac. THANK YOU!!</td>
</tr>
<tr>
<td>2: Recovery</td>
<td>Hurricane #Isaac relief efforts in the Gulf continue. Pls help! Txt REDCROSS to 90999, to give $10.</td>
</tr>
<tr>
<td>3: Mitigation</td>
<td>FEMA mitigation to offer rebuilding tips in St. John Parish at Home Depot in LaPlace, La. <a href="http://t.co/8UFJFEz5">http://t.co/8UFJFEz5</a> #Isaac</td>
</tr>
<tr>
<td>4: Preparedness</td>
<td>RT @jeffiel: Seriously, everybody should have this #RedCross app on their phone. What to do in #emergencies <a href="http://t.co/MtrQcOo1">http://t.co/MtrQcOo1</a> #isaac</td>
</tr>
</tbody>
</table>

Using our training set, we trained three classification algorithms -- multi-class SVM, multinomial Naïve Bayes,
What, When, Where, and Who in PhaseVis

(a) Phase View

WHAT

(b) Tweet View

02:45 09/07–07:08 09/10
1. Response: 88
2. Recovery: 151
3. Mitigation: 7
4. Preparedness: 22

(c) Social Network View

WHO

(d) Map View

WHERE

6 https://developers.google.com/maps/documentation/geocoding/

3.3.3. Component Details of PhaseVis

As briefly mentioned in Section 3.3.2, PhaseVis consists of four distinct views: the Phase View (a), Tweet View (b), Social Network View (c), and Map View (d).
A Sample Use Case (1/2)

8/24 - 8/26
Preparedness is dominant

(A)

9/7 - 9/9
Recovery is dominant

(B)
A Sample Use Case (2/2)

(A)

HumaneSociety

RedCross

(1)

(2)

(B)

RedCrossSELA

RedCross

(3)

Figure 4. Who and Where aspects of time intervals (A) (8/24 - 8/26) and (B) (9/7 - 9/9), visualized with a user mention graph and a user location map respectively.
Acknowledgments

• Members of the Digital Library Research Laboratory, Virginia Tech

• We thank NSF for supporting the CTRnet project (IIS-0916733), of which this work is part. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
Thank you!
Questions?