Crisis, Tragedy and Recovery Network (CTRnet)

http://www.ctrnet.net

Slides by Kiran Chitturi, Edward A. Fox, and the CTRnet team
CTRnet

- 3+ year project started in August 2009 with grant from NSF
- PI and Co-PIs
  - Edward A. Fox\(^1\)
  - Andrea L. Kavanaugh\(^1\)
  - Steven D. Sheetz\(^2\)
  - Donald J. Shoemaker\(^3\)
  - Naren Ramakrishnan\(^1\)

\(^1\)Dept. of Computer Science
\(^2\)Dept. of Accounting and Information Systems
\(^3\)Dept. of Sociology
CTRnet

• Research the problem of integrating content, community, and services related to crises, tragedies, and recovery (CTR)

• Goals:
  – Integrating heterogeneous information in a specific domain, making it accessible, and preserving it for long-term reuse;
  – Extending the scope of digital libraries so they are closely but flexibly coupled with a wide variety of services to support diverse emerging communities; and
  – Supporting information exploration with advanced methods (Stepping Stones and Pathways (SSP), PathRank, and Storytelling) that facilitate searching, browsing, and discovery

• Building a domain specific distributed digital library, integrating information from many sources, with ontology support
System Stakeholders
CTR Toolkit Services
Collections archived so far

- 9.5 TB (45 archive collections)
- 87 tweet archives (16 million tweets)
- Types of Collections
  - Middle East Revolutions
  - Earthquakes
  - Floods
  - School Shootings
  - Plane Crashes
- Click [here](#) for more details about some of our collections
## Collections Sample (Earthquakes)

<table>
<thead>
<tr>
<th>Archive Title</th>
<th>Number of webpages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virginia Earthquake</td>
<td>9313</td>
</tr>
<tr>
<td>New Zealand Earthquake</td>
<td>603</td>
</tr>
<tr>
<td>Haitian Earthquake (one month)</td>
<td>1204824</td>
</tr>
<tr>
<td>Sikkim Earthquake</td>
<td>2252</td>
</tr>
<tr>
<td>Turkey Earthquake</td>
<td>382910</td>
</tr>
<tr>
<td>Chile Earthquake</td>
<td>119</td>
</tr>
<tr>
<td>Chilean Earthquake</td>
<td>150</td>
</tr>
<tr>
<td>Japan Earthquake (one month)</td>
<td>3187346</td>
</tr>
<tr>
<td>Haiti Anniversary</td>
<td>665</td>
</tr>
<tr>
<td>Total number of webpages</td>
<td>4788182</td>
</tr>
</tbody>
</table>
Types of Data

• Metadata
  – Stored as WARC files

• Multimedia Raw Data
  – HTML pages
  – PDF files
  – Images
  – Videos
Goals for next few months

• Integrating a cleaned-up version of existing metadata and data, with a variety of services, operating through the digital library

• Automating procedures to: archive automatically (by using tweets, web crawling, filtering), index the content, and make it accessible through the digital library

• Extracting additional metadata out of data archived

• Recognizing names and entities to enhance access

• Switching to Hadoop based processing for crawling, indexing, and other steps
Above shows components of the prototype system and related data flows:

1. Input data comes from (1) Tweet Archive DB, built from yourTwapperKeeper.
2. URL Extractor (A) periodically identifies URLs embedded in tweets from (1) and stores them into Extracted URL DB (2).
3. Crawled data are stored in Web Archive (WARC) (3) format.
4. Data Extractor (C) retrieves resource types such as HTML pages, PDFs, images etc. (4) from multiple WARC files.

Methods are developed to crawl from web as well as tweets.
Workflow (from warc files)

Takes WARC files in (3) Archived Data as input and

- Extracts webpages from WARC files using Hanzo and WARC tools
- Extracts metadata from archives and HTML pages into database using CTRnet Ontology
- Filters the data using Waikato Environment for Knowledge Analysis (WEKA)*
- Indexes the metadata from the database using Apache SOLR
- Makes the archive content accessible through Digital Library

*Filtering won’t be there if the data that is archived is already relevant (based on the URLs that are crawled, with new approaches). This filtering is only for data already archived.
References and Contact Info.


- See http://www.ctrnet.net

- Contact information for additional information:
  - Professor Edward A. Fox, http://fox.cs.vt.edu, fox@vt.edu