Crisis, Tragedy, and Recovery
Network Digital Library (CTRnet) + Web Archiving in Qatar and VT

Edward A. Fox, Seungwon Yang, & CTRnet Team
Department of Computer Science, Virginia Tech

Workshop at WADL’13, July 25-26, 2013
Outline

- Introduction
  - Project goal
  - Members & collaborators
- Main Archiving Tasks
- Sub-Projects
- Dissemination Efforts
- IDEAL Project
- Qatar
- VT
- Acknowledgments
- Collaboration
## CTRnet Project Goal

Developing integrative approaches:
- Collect, analyze, and visualize disaster information with a DL

<table>
<thead>
<tr>
<th>Content</th>
<th>Collect</th>
<th>Analyze</th>
<th>Visualize</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web sites, images</td>
<td></td>
<td>Image similarity</td>
<td>Organize images by similarity</td>
</tr>
<tr>
<td>Tweets</td>
<td></td>
<td>Content, user profiles</td>
<td>Patterns, frequencies</td>
</tr>
<tr>
<td>Facebook content</td>
<td></td>
<td>Usage of social media (SM)</td>
<td>SM use</td>
</tr>
<tr>
<td>Focus group interviews/surveys</td>
<td></td>
<td>Usage of SM</td>
<td>SM use/needs</td>
</tr>
<tr>
<td>Technology</td>
<td>Crawler</td>
<td>CBIR algorithm</td>
<td>CBIR visualization interface</td>
</tr>
<tr>
<td>Online tools, scripts, APIs</td>
<td></td>
<td>NLP toolkit, SQL</td>
<td>Graphics</td>
</tr>
<tr>
<td>Facebook app</td>
<td></td>
<td>Spreadsheets</td>
<td></td>
</tr>
<tr>
<td>Brainstorming tool</td>
<td></td>
<td>Brainstorming tool</td>
<td></td>
</tr>
</tbody>
</table>
Members & Collaborators

- Project members from multi-disciplinary areas
  - Computer Science (HCI, Information Retrieval)
  - Accounting and Information Systems
  - Sociology

- Collaboration with the Internet Archive (IA)
  - Developed web archives
    - Heritrix crawler
    - Crawled data hosted by Wayback Machine in IA
    - Raw data downloaded and locally analyzed
  - Attended Archive-It Partners Meeting
    - Introduced the CTRnet team’s crawling approach using tweets
Outline

- Introduction
- Main Archiving Tasks
  - Disaster webpage archives
  - Disaster tweet archives
- Sub-Projects
- Dissemination Efforts
- IDEAL Project
- Qatar
- VT
- Acknowledgment
- Collaboration
Disaster Webpage Archives

- Webpages, PDFs, and multimedia content crawled from the Web
  - 45 archives and growing (8.8 TB+)
- Active archives:

<table>
<thead>
<tr>
<th>Boston marathon blast 2013</th>
<th>Global Emergency Overview 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boko Haram Attack 2013</td>
<td>Hurricane Sandy 2012</td>
</tr>
<tr>
<td>Center for Research on the Epidemiology of Disasters (CRED) 2012</td>
<td>Japan Earthquake 2011</td>
</tr>
</tbody>
</table>
Disaster Tweet Archives

- More than 120 tweet archives and growing
  - Use Twitter Streaming API
  - Hashtags and keyword-based archiving

<table>
<thead>
<tr>
<th>Natural</th>
<th>floods, earthquakes, wildfires, tsunami, hurricanes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man-made</td>
<td>shooting, transportation accidents, plane crash</td>
</tr>
<tr>
<td>Political</td>
<td>Middle East protests, Iran elections</td>
</tr>
<tr>
<td>Health</td>
<td>diabetes, obesity, cancer, mental illness</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Main Archiving Tasks
- Sub-Projects
  - Social media use during political crisis
  - Topic tagging of webpages
  - Visualizing emergency phases in tweets
  - Water main break visualization
  - Focused crawling
  - LucidWorks tool for big data processing
- Dissemination Efforts
- IDEAL Project
- Qatar
- VT
- Acknowledgment
- Collaboration
Total 514,782 tweets
Opinion Leadership in Egypt Uprising 2011

- 514,782 tweets (one week around Mubarak’s resignation)
- Total 79,000 unique users
  - Presumably posting from Egypt → 4,710
  - Individuals excluding organizations → 3,675

Opinion leaders

- 500-27,000 followers in top 10% (365) individuals
- Bios: blogger/activist, writer/reporter, lawyer/executive director, social media consultant,… → ‘elite’ type actors
Topic Tagging of Webpages: Xpantrac

(a) EXPANSION

Xpantrac

Input Text

Query units

Corpus

Term-doc matrix

(b) EXTRACTION

Search Engine API

Web

Term 1

Term 2

Term 3

Term n

Doc 1 Doc 2 Doc 3 ... Doc m Sum

Term 1 3 1 0 ... 4 12

Term 2 1 2 4 ... 1 8

Term 3 4 0 0 ... 3 9

... ... ... ... ... ...

Term n 2 7 1 ... 0 17

Topics

Term n

Term 1

Term 3
Visualizing Emergency Phases in Tweets (ISCRAM 2013) (1/2)

Four phases of emergency management model
Visualizing Emergency Phases in Tweets (2/2)

http://spare05.dlib.vt.edu/~ctrvis/phasevis/index_may.html
## Water Main Break Visualization

<table>
<thead>
<tr>
<th>Dataset Keyword</th>
<th>Total tweets</th>
<th># of tweets which have GPS information (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>water main break</td>
<td>13,382</td>
<td>156 (1.17 %)</td>
</tr>
<tr>
<td>water pipe leak</td>
<td>967</td>
<td>1 (0.10 %)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location information type</th>
<th># of tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPS data (longitude, latitude)</td>
<td>36 (1.08 %)</td>
</tr>
<tr>
<td>Location information extracted from text</td>
<td>1,473 (44.19 %)</td>
</tr>
</tbody>
</table>

**Tweets collected with keywords**

**Selected tweets with location information (lat/long, geonames)**

**Event locations displayed with details**

*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
Focused Crawling

- **IA collections**
  - Identify a CTR event, list keywords
  - Query online news sources, identify URLs in tweets
  - Use URLs as initial seeds for crawling; IA provides access

- **Modified version of the LibSVM classifier**
  - Reduced noise
  - 3000 documents about school shootings

- **Next-generation focused crawler**
  - Combines evidence signals for relevance estimation (using Bayesian networks)
  - Solves Tunneling problem using AI approaches (Reinforcement Learning)
LucidWorks Big Data Tool

- Powerful tool with components:
  - Hadoop – for distributed computing
  - Lucene & Solr – for indexing, searching
  - Hbase – distributed database for Hadoop
  - Mahout – distributed machine learning
  - Oozie – workflow
  - Kafka: high throughput distributed messaging
  - Zookeeper: maintaining distributed coordination
  - Pig: high-level platform for creating MapReduce programs
- Packaged as a virtual appliance in Ubuntu for easy installation
- Processing of WARC files downloaded from IA
Outline

- Introduction
- Main Archiving Tasks
- Sub-Projects
- Dissemination Efforts
  - Conferences
  - Journal papers
  - Meetings attended
- IDEAL Project
- Qatar
- VT
- Acknowledgment
- Collaboration
Dissemination Efforts

- Conferences, Workshops
  - JCDL, ISCRAM, Digital Government, CHI, WADL

- Meetings Attended
  - NSF workshop: Crisis Informatics 2012, 2011
  - Archive-It Partners Meeting
    - 2012 (Annapolis, MD), 2011 (Lexington, KY)

- Publications
  - Please see [http://www.ctrnet.net/publications](http://www.ctrnet.net/publications)
Outline

- Introduction
- Main Archiving Tasks
- Sub-Projects
- Dissemination Efforts
- IDEAL Project
  - Extension of CTRnet
  - Scope broadened beyond crisis events (e.g., community)
  - NSF funding pending
- Qatar
- VT
- Acknowledgment
- Collaboration
Integrated Digital Event Archive and Library (IDEAL) Project
http://www.eventsarchive.org/

- Extension of CTRnet with broadened scope:
  - Event detection
  - Event data archiving & processing
    - Multimedia (images, videos) shared in social media
- Digital government research
  - Community issue detection
  - Public opinion mining, mood perception, information flow
- Technologies:
  - Focused crawling, analysis/visualization services, integration of archive and DL capabilities
Outline

- Introduction
- Main Archiving Tasks
- Sub-Projects
- Dissemination Efforts
- IDEAL Project
- Qatar
- VT
- Acknowledgment
- Collaboration
Project Objectives/Aims

A. Research and prototype digital library systems and infrastructure for Qatar, focusing initially on Qatari information related to government and scholarly activities.

Leverage the crawling engine from Penn State’s SeerSuite software infrastructure, and extend it beyond its current focus on English to support Arabic-English collections, and to cover a broad range of scholarly disciplines, and all types of government information.

… (with collaboration of National Library)
Project Objectives/Aims (cont’d)

B. Research and build the digital library community in Qatar, supporting digital library use, services, collection development, tailored systems, and advancing toward a Knowledge Society.

Study scholarly activities, and engage in community building in Qatar, so DLs can be tailored to specific domains and to the unique needs of Qatar. Through workshops, a consulting center at the proposed Institute, and collaborative efforts with libraries and museums in Qatar, we will identify particular needs and uses, and tailor collections, systems, and services, to lead toward the Qatari Knowledge Society.
Half of campus web servers use the central CMS
Many other web servers cover varied content
Coverage by Internet Archive is OK, but for parts of the overall campus Web, crawling is infrequent

Discussions with IT, Library, University Relations, about
  Heretrix
  Memento support
  SiteStory
Outline

- Introduction
- Main Archiving Tasks
- Sub-Projects
- Dissemination Efforts
- IDEAL Project
- Qatar
- VT
- Acknowledgment
- Collaboration
Acknowledgment

- NSF for funding:
  - Grant: CTRnet IIS-0916733
  - Proposal: IDEAL IIS-1319578, Integrated Digital Event Archive and Library

- The Internet Archive:
  - Heritrix crawler
  - hosting the crawls and resulting archives
Collaboration

- We invite anyone to collaborate with us!

- Contact:
  - Edward A. Fox  <fox@vt.edu>
Thank you!

Questions/Comments?