From CTRnet to IDEAL (and Qatar, VT, SiteStory, UPS, ...)

NSF IA WIRE Workshop
Harvard -- June 16, 2014

Edward A. Fox, fox@vt.edu, http://fox.cs.vt.edu
Professor, Dept. of Computer Science, Virginia Tech
Director, Digital Library Research Laboratory
Director, Networked Digital Library of Theses and Dissertations

http://www.eventsarchive.org
• Related Funding:
  – 2009-2013: NSF IIS-0916733, Crisis, Tragedy, and Recovery network (CTRnet)
  – 2012-2015: Qatar NPRP 4-029-1-007, Establishing a Qatari Arabic-English Library Institute
• The Internet Archive (Kristine Hanna, co-PI):
  – Heritrix crawler and other tools and support
  – Hosting the crawls and resulting archives
• LucidWorks (software and support - open jobs, internships)
Acknowledgments - 2

• IDEAL: VT: PI: Fox, co-PIs: Andrea Kavanaugh, Steve Sheetz, Don Shoemaker; GRAs: Mohamed Magdy, Sunshin Lee; Egypt: Riham Mansour
• CTRnet: also Naren Ramakrishnan (co-PI); GRAs Seungwon Yang and Venkat Srinivasan
• DL-VT416: also Christopher North and Weiguo Fan
• Computing in Context: Villanova PI Robert Beck; Students: Xuan Zhang, Tarek Kanan: class to learn Computational Linguistics by 5-way better summarizing Web archive collections (extract words/sentences, find topics, use event templates)
• Qatar: Lead PI Fox, Co-PIs Mohammed Samaka (Qatar U.), Somaya Al-maadeed (QU), Krishna RoyChowdhury (Qatar National Library), C. Lee Giles (Penn State), Rick Furuta (Texas A&M); consultant John Impagliazzo (Hofstra), VT GRA Tarek Kanan
• Mellon: PI Zhiwu Xie, co-PI Fox, GRA Prashant Chandrasekar
• Other students: Kiran Chitturi, Rachel Coston, Ishita Ganotra, S.M.Shamimul Hasan, Christopher Jones, Rohan Kaul, Jun Kim, Lin Tzi Li, Ying Ni, Braeden Sebastian, and teams in CS4624, 5604, 6604
• Collaborators in: Egypt, Tunisia, Mexico, Philippines

• WE WELCOME OTHER COLLABORATORS!
Memento – Time Travel for the Web: Across-Archive Method for Linking the Current & Past Web


http://www.mementoweb.org/

RFC 7089 (Martin Klein)
Related Projects

• Mellon/Columbia: enhance SiteStory by devising a webserver that also archives; use the archive automatically when server is down; capture the VT Web and bring up UPS on multiple campus sites

• Qatar: at Qatar U., Qatar National Library
  – Build a digital library community (consulting center)
  – Build digital library infrastructure:
    • SiteSeer (CiteSeerX, ChemXseer, TableSeer, ...) with Arabic and CLIR support
    • Heritrix, Wayback Machine, Solr, ...
Web Archives

- 13 TB of IA Collections, e.g., 2013: Boko Haram attack, Boston Marathon blast, Global Emergency Overview, Texas fertilizer plant explosion

<table>
<thead>
<tr>
<th>Category</th>
<th>No. of Archives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accidents (plane crash, building collapse, ferry)</td>
<td>11</td>
</tr>
<tr>
<td>Bombings</td>
<td>4</td>
</tr>
<tr>
<td>Earthquakes (Japan)</td>
<td>12</td>
</tr>
<tr>
<td>Fires</td>
<td>2</td>
</tr>
<tr>
<td>Floods</td>
<td>4</td>
</tr>
<tr>
<td>Hurricanes (Sandy), Tsunami, Cyclones, Typhoons</td>
<td>8</td>
</tr>
<tr>
<td>Shootings</td>
<td>17</td>
</tr>
</tbody>
</table>
Tweet Collections

- > 120 Event-specific and general collections
- Total of 600 million tweets, from streaming APIs using hashtags and keywords

<table>
<thead>
<tr>
<th>Category</th>
<th>No. of collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident (transportation)</td>
<td>33</td>
</tr>
<tr>
<td>Bombing</td>
<td>8</td>
</tr>
<tr>
<td>Community</td>
<td>10</td>
</tr>
<tr>
<td>Earthquake</td>
<td>18</td>
</tr>
<tr>
<td>Fire</td>
<td>6</td>
</tr>
<tr>
<td>Flood</td>
<td>11</td>
</tr>
<tr>
<td>General (including health)</td>
<td>67</td>
</tr>
<tr>
<td>Hurricane, Tsunami</td>
<td>39</td>
</tr>
<tr>
<td>Political (Middle East, Iran)</td>
<td>40</td>
</tr>
<tr>
<td>Shooting</td>
<td>29</td>
</tr>
</tbody>
</table>
CTRnet

Collect, analyze, and visualize disaster information with a DL
Social Media Use in Political Crisis
(1/2)(2/7 - 2/14, 2011)

- (2/7) Google executive released
- (2/8) Ghonim, freed activist, energizes protests
- (2/9) Widespread labor strikes
- (2/10) Mubarak refuses to step down
- (2/11) Mubarak resigns

Total 514,782 tweets
Social Media Use in Political Crisis (2/2)

• 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
Four phases of emergency management model
Visualizing Emergency Phases in Tweets (2/2)
Input: text file

Build query

- Every 5 words, 1 word overlap

Send query to search API

- Web search (Seungwon)

- Wikipedia, our collection(s): CS4624 Spring 2014: Sloane Neidig, Samantha Johnson, David Cabrera, Erika Hoffman

Find topics in retrieved documents

- Frequency of words

Select most frequent as “topics”

Output: topics
Topic Tagging of Webpages: Xpantrac - 2
Seungwon Yang (GMU postdoc now)
Water Main Break Visualization

Sunshin Lee

<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
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
Event Ontology

• Event model
  – Organizations/entities participating in the event

• What
  – Topics of the Event

• Where
  – Event location

• When
  – Event time frame (and later times of interest, e.g., anniversaries)
IDEAL Proposal Architecture

Preservation Planning

Data Management

Archival Storage

Access

Administration
IDEAL System Architecture
Sunshin Lee (built low-cost cluster)
IDEAL Data Architecture
Sunshin Lee
Event Focused Crawler
Mohamed Magdy

Focus of research
Baseline vs. Event Focused Crawler

Mohamed Magdy

Harvest ratio: relevant crawled webpages vs. cumulative set of crawled webpages
Extracted News Events on a Time Line
CS6604 Spring 2014: Tianyu Geng, Wei Huang, Ji Wang, Xuan Zhang

02/28
03/01
03/08
03/12
03/16
03/20
03/23
03/26
04/12
04/16

03/01
03/09
03/14
03/20
03/26

ukraine, yanukovich, crisis, minister, sign, russian
russia, bank, sanctions, ukraine, crisis, crimea
ukraine, tensions, data, rise, shares, china, stocks
ukraine, russia, talks, aid, crisis, sanctions, deal
gas, ukraine, russian, russia, europe, talks, energy

crimea, ukraine, russia, minister, referendum, vote
ukraine, crimea, crisis, putin, russia, minister
ukraine, house, imf, u.s, bill, white, aid
ukraine, aid, support, government, talks, house, russian
crimea, ukraine, russia, troops, border

History:
3/7 referendum annulled
3/14: UN draft resolution
News-Tweet Architecture

CS6604 Spring 2014: Tianyu Geng, Wei Huang, Ji Wang, Xuan Zhang
Welcome to IDEAL Spreadsheet

Here is how to use our tool:

1. Click the large green "v" button at the top of your screen.

2. Fill in the window with all your requests and click the "Add" button.

3. In the 'Operations' column type in your desired operation (example: "termCount").

4. In the 'Visualizations' column type in your desired visualization (example: "wordCloud").

<table>
<thead>
<tr>
<th>Collection</th>
<th>Query</th>
<th>Operation</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Illinois University Shooting</td>
<td></td>
<td>Word Count</td>
<td>WordCloud</td>
</tr>
<tr>
<td>Virginia Tech Shooting</td>
<td></td>
<td>Word Count</td>
<td>WordCloud</td>
</tr>
</tbody>
</table>

Add Row
Recommended Collection-Level Metadata
CS6604 Spring 2014: Michael Shuffett

• Dublin Core
  – Title, Description

• PROV-O
  – Starting Point Classes
  – Collection process, organization, hadMember, atLocation

• ISO 3166-2 for locations

• W3/XMLSchema#dateTime

• PLUS: TweetID tool for tweet collections
  – Extracts tweet and collection level metadata
  – Compares / combines tweet collections
Thank you!

Questions/Comments?

fox@vt.edu, http://fox.cs.vt.edu
540-231-5113
Office: 2160G Torgersen Hall
Campus Mail: 114 McBryde Hall, M/C 0106, Dept. of CS, Virginia Tech, Blacksburg, VA 24061