Web Archiving and Digital Library Projects and Technologies

Edward A. Fox (fox@vt.edu) and Zhiwu Xie (zhiwuxie@vt.edu)

With slides and through collaboration with:

Co-PIs: Jefferson Bailey (IA), Andrea Kavanaugh, Donald Shoemaker, Steve Sheetz

GRAs: Prashant Chandrasekar, Liuqing Li, Ziqian Song

Former GRAs: Mohamad Farag (Alexandria), Sunshin Lee (Radford), Islam Harb (on leave), Seungwon Yang (LSU), . . .

Presentation for Linux/Unix Users Group @ VT (VTLUUG)
15 March 2018

Outline

• Context
• GETAR proposal
• IDEAL results – Sunshin Lee
• IDEAL results – Mohamed Farag
• Selected GETAR projects
• Welcoming collaboration
Context - 1

• Understanding the world by collecting, archiving, analyzing
• Providing access to information: digital libraries
• From theory: 5S (Societies, Scenarios, Spaces, Structures, Streams)
• To algorithms, applications, systems, collections, user studies

• Library related
• Any type of information: multimedia as well as text
• Applications: archaeology, autism, civil engineering, education, epidemiology, events, fingerprinting, fisheries, global change, hurricanes, national archiving (Qatar), neuroscience, news, physics, school shootings, sociology, trails, Web, . . .
Context - 2

- Digital Library Research Laboratory, 2030 Torgersen Hall
  - Director, Edward A. Fox, [http://fox.cs.vt.edu](http://fox.cs.vt.edu)

- University Libraries
  - Center for Digital Research & Scholarship [http://scholar.lib.vt.edu/staff/zxie/](http://scholar.lib.vt.edu/staff/zxie/)
    - Zhiwu Xie, Director, Digital Library Development


- VTechWorks sites for DLRL and related courses
  - [https://vtechworks.lib.vt.edu/handle/10919/18732](https://vtechworks.lib.vt.edu/handle/10919/18732), 47780, 19081, 18655, 50956
Part 1

• Extracts from the GETAR proposal (NSF IIS-1619028 and 1619371)
• Virginia Tech and Internet Archive, 2016 – 2020

• Global Event and Trend Archive Research
  • http://eventsarchive.org
Problems / Questions

• How can K-12 students, the general public, and interdisciplinary teams study and research the important global events and trends that relate to worldwide grand challenges?

• How can information systems support those needs in an integrated fashion, empowering users through interaction with content across the broad information life cycle?

• How can the growing collections of Internet archives be integrated with both the constantly changing current version of the WWW and stream-oriented communications like tweets?

• How can those involved in planning, policy making, innovation, economics, engineering, and the social sciences engage in focused as well as longitudinal studies (from the End of Millennium to the present: 1997-2020), in an interdisciplinary context, through those systems and collections?
Goals

• To aid interdisciplinary research and education, regarding important global events and trends, which can benefit from DL access to Internet content, starting in 1997, up through 2020.

• To develop next generation interactive and integrated information systems, connecting DLs and archives, connecting sources and documents, and connecting webpages and tweets (and other user submitted content).

• To advance the state-of-the-art in DL and NLP with regard to handling archives, analyzing documents, adding value to metadata and collections, and expanding the scope of interaction across the information life cycle.
Objectives - 1

• To develop methods, deploy them at the Internet Archive (IA), and
  • get high quality collections from IA’s archives (of WARC files),
  • aiming to find all relevant webpages
  • (including forms like Usenet posts that have similar functionality to tweets)
  • for important events we identify over the period 1997-2000.
• To devise interactive techniques resulting in rich models for events and trends,
  • that will lead to enhanced focused crawling, accurate classifiers, and
  • helpful information visualization.
• To devise interface development methods for DLs, that lead to
  • generic solutions where possible, but also facilitate tailoring interfaces
  • to the needs of particular disciplines.
• To aid stakeholders, through interactive interfaces, to engage in
  • development and curation of collections of tweets and webpages
  • about events and/or trends, with high quality, that will support the
  • needs in their discipline, as well as assist in interdisciplinary research studies.
Objectives - 2

- To **aid stakeholders**, through interactive interfaces (with NLP), to *deal with errors, spam*,
  - variations in doc. length / structure / focus, multiple languages / sublanguages, and their
  - varying needs for analyzing and representing/desccribing/summarizing interesting content.
- To **aid stakeholders**, through interactive interfaces, to *analyze, visualize, and access*
  - those collections (with maps, timelines, social networks, and faceted browse, search, and
  - exploration of content), in ways appropriate for their needs and disciplines, and to
  - integrate the interaction with the collection development and curation activities.
- To **integrate DL and archive methods** so collections can both be preserved
  - for the long term, and easily accessed through highly interactive interfaces.
- To **aid stakeholders** to advance *research and education concerning important global events*
  - and trends, including climate change, development, disasters, energy, globalization,
  - innovation, policies, resilience, social movements, and violence.
GETAR Architecture

- Sources
- WWW
- Twitter
- Collections
  - Source Identification
  - Model Building
  - Collection Development
  - Utilization
  - Curation
- Users
  - Analysis
  - Info Extraction
  - Organizing
    - Correcting/Revising
  - Trends
  - NLP
  - Classifying
- Events
  - Searching
  - Visualizing
  - Recommending
  - Selecting
  - Browsing
- Services
  - Data/Info/Knowledge
- Linked Data Archives
- Phases
### Selected Events from 54 Identified at End of Millennium (EoM, 1997-2000)

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>El Nino warms ocean wildlife (thru '99)</td>
</tr>
<tr>
<td></td>
<td>Toyota Prius makes a debut in Japan</td>
</tr>
<tr>
<td></td>
<td>Kyoto Global Warming Conf/Protocol</td>
</tr>
<tr>
<td></td>
<td>'bird flu'; Hong Kong kills 1M chickens</td>
</tr>
<tr>
<td></td>
<td>Hong Kong (transfer of sovereignty)</td>
</tr>
<tr>
<td></td>
<td>Massacres in Algeria</td>
</tr>
<tr>
<td></td>
<td>Iraq expelled US weapons inspectors</td>
</tr>
<tr>
<td>1998</td>
<td>Microsoft releases Windows 98</td>
</tr>
<tr>
<td></td>
<td>Smoking ban: CA restaurants &amp; public</td>
</tr>
<tr>
<td></td>
<td>Al Qaeda bombs US embassies</td>
</tr>
<tr>
<td></td>
<td>Ethiopian-Eritrean War: &gt;10K dead,</td>
</tr>
<tr>
<td></td>
<td>Hurricane Mitch in Honduras</td>
</tr>
<tr>
<td></td>
<td>Congo/Africa's World War: kills &gt;2.5M</td>
</tr>
<tr>
<td></td>
<td>Google, Inc. is founded</td>
</tr>
<tr>
<td>1999</td>
<td>UN announces the 6 billionth baby born</td>
</tr>
<tr>
<td></td>
<td>Napster (music download) debuts</td>
</tr>
<tr>
<td></td>
<td>Kosovo War: NATO air strikes</td>
</tr>
<tr>
<td></td>
<td>Turkey: Richter scale 7.4 earthquake</td>
</tr>
<tr>
<td></td>
<td>Columbine High School shooting</td>
</tr>
<tr>
<td></td>
<td>Two viruses afflict computers worldwide</td>
</tr>
<tr>
<td></td>
<td>Kargil War between India and Pakistan</td>
</tr>
<tr>
<td>2000</td>
<td>Y2K (Year 2000 problem)</td>
</tr>
<tr>
<td></td>
<td>Toyota released Prius worldwide</td>
</tr>
<tr>
<td></td>
<td>Cyclone Eline Mozambique</td>
</tr>
<tr>
<td></td>
<td>Al-Qaeda attack on USS Cole in Yemen</td>
</tr>
<tr>
<td></td>
<td>Mad cow disease alarms Europe</td>
</tr>
<tr>
<td></td>
<td>W. Nile Virus: Israel, France, Jordan, US</td>
</tr>
<tr>
<td></td>
<td>US Bush v. Gore election; no FL recount</td>
</tr>
</tbody>
</table>
### PhaseViz

#### Phase View

- **WHAT**
- **WHEN**

#### Map View

- **WHERE**

#### Social Network View

- **WHO**

---

### Tweets in Selected Region (Classes: 1-Response, 2-Recovery, 3-Mitigation, 4-Preparedness)

<table>
<thead>
<tr>
<th>ID</th>
<th>is_R</th>
<th>Text</th>
<th>Class</th>
<th>Date</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>33836</td>
<td>0</td>
<td>More #RedCross volunteers prepared to drive straight into #Isaac path #Fox59<a href="http://trib.al/dQAORO">http://trib.al/dQAORO</a></td>
<td>4</td>
<td>Wed, 29 Aug</td>
<td>21346209617</td>
</tr>
<tr>
<td>33829</td>
<td>1</td>
<td>RT @CraigatFEMA: Hurricane #Isaac, dangerous storm surge, heavy rainfall to be follow by additional floodinghttp</td>
<td>4</td>
<td>Wed, 29 Aug</td>
<td>21346209620</td>
</tr>
</tbody>
</table>
# IDEAL event webpage collections:
## CS4984 (Computational Linguistics)

<table>
<thead>
<tr>
<th>Category</th>
<th>Collection</th>
<th>Event Year(s)</th>
<th>Location</th>
<th># of Webpages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>Ebola</td>
<td>2014</td>
<td>World</td>
<td>15,000</td>
</tr>
<tr>
<td>Earthquake</td>
<td>Virginia earthquake</td>
<td>2011</td>
<td>Virginia, USA</td>
<td>8,765</td>
</tr>
<tr>
<td>Fire</td>
<td>Brazil club fire</td>
<td>2013</td>
<td>Brazil</td>
<td>690,281</td>
</tr>
<tr>
<td>Flood</td>
<td>Pakistan flood</td>
<td>2011</td>
<td>Pakistan</td>
<td>20,416</td>
</tr>
<tr>
<td>Hurricane</td>
<td>Hurricane Sandy</td>
<td>2012</td>
<td>East Coast, USA</td>
<td>75,929</td>
</tr>
<tr>
<td>Shooting</td>
<td>Tucson shooting</td>
<td>2011</td>
<td>Tucson AZ, USA</td>
<td>37,829</td>
</tr>
<tr>
<td>Community</td>
<td>Blacksburg events</td>
<td>2011-2012</td>
<td>Blacksburg VA, USA</td>
<td>16,024</td>
</tr>
</tbody>
</table>
Archives, the GETAR DL, and WWW, used to extend IA extraction results.
GETAR DL architecture

**Producers/Sources**
- Social Media
- Tweets (IDEAL, NDSSL, VTS, EMBERS)
- Usenet/Newsgroup
- Webpages (WARC)
- Internet Archive
- Heritrix
- Focused Crawler
- VT WebArchive

**Preservation Planning**

**Ingest**
- DB import (Sqoop2)
- WARC (Behemoth)
- Web Crawling (Nutch)
- ETL (Pig)

**Data Management**

**Archival Storage**
- NoSQL Database
- Taxonomy
- Topics
- Metadata
- Ontology
- Raw data
- Named Entities
- Storage (HDFS)

**Access**
- Machine Learning
- Apache Spark
- NLP (NLTK)
- PhaseVis
- NER (SNER)
- Summarization
- Database Querying
- Maps
- Python
- Trend Analysis
- Indexing (Solr)
- Faceted Search

**Consumers/Users**
- Researcher
- Practitioner
- Affected
- Interdisciplinary Team
- K-12 Students
- General Public

**Administration**
<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
<th>Investigators</th>
<th>Courses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core Research</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis, access</td>
<td>Databases, HCI, information visualization, machine learning, ontologies, statistics</td>
<td>Fox, Franck, Huang, North, Sheetz</td>
<td>BIT 4524, 4544, 4614; CMDA/CS/ STAT3654; CS5764</td>
</tr>
<tr>
<td>Library, information, data</td>
<td>Archives, big data, curation, data management, decision support, exploring, knowledge engineering, searching</td>
<td>Fox, French, Nicholls, Speer, Thomas, Zobel</td>
<td>CS4624, 5604, 6604; FOR3604; GRAD5134</td>
</tr>
<tr>
<td>NLP</td>
<td>Arabic, document analysis, errors, information extraction, summarization, topic identification</td>
<td>Eubank, Fox, Rozovskaya</td>
<td>CS4624, 4984, 5984, 6804</td>
</tr>
<tr>
<td><strong>Applied Research Across Disciplines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geospatial</td>
<td>Car crashes, data structures, GIS, maps, queries, traffic, tweets, weather and crops</td>
<td>Baird, Lu, Sforza</td>
<td>GEOG1115, 1116</td>
</tr>
<tr>
<td>Simulation</td>
<td>NDSSL: epidemiology, diffusion in networks, planning response</td>
<td>Eubank, Lewis, Swarup</td>
<td>GBCB5874, 7994</td>
</tr>
<tr>
<td>Climate change</td>
<td>Adaptation, biodiversity, conservation, ecology, ecosystems, effects on plants &amp; animals, environment, sea-level rise</td>
<td>Bukvic, Jelesko, Kalkstein, Quinn</td>
<td>GEOG2994, 4974, 4994; PPWS4994</td>
</tr>
<tr>
<td>Economics</td>
<td>Development, families, game theory, Middle East, smart cities, social networks</td>
<td>Ball, Korkmaz, Salehi-Isfahani</td>
<td>ECON3004</td>
</tr>
<tr>
<td>Area</td>
<td>Description</td>
<td>Investigators</td>
<td>Courses</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Energy</td>
<td>Green engineering, nuclear policies</td>
<td>Avey, McGinnis</td>
<td>ENGR3124</td>
</tr>
<tr>
<td>History</td>
<td>Globalization, Soviet history</td>
<td>Ewing</td>
<td>HIST1214, 1215, 2124, 3394, 3554</td>
</tr>
<tr>
<td>Innovation</td>
<td>CIE: entrepreneurship, impact of resources, industry collaboration, social</td>
<td>Junkunc</td>
<td>MGT3064, 3074, 4064, 4094</td>
</tr>
<tr>
<td></td>
<td>and technology based ventures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resilience</td>
<td>Concentrations, dependency, disasters, evacuations, planning, policy,</td>
<td>Bohland, Bukvic, Lawrence,</td>
<td>CEE5620, 5660; GRAD5134</td>
</tr>
<tr>
<td></td>
<td>relocation, supply chains, urban and regional, vulnerability</td>
<td>Murray-Tuite, Zobel</td>
<td></td>
</tr>
<tr>
<td>Sociology</td>
<td>Crises, global issues, social inequality, social movements, social</td>
<td>Baird, Kavanaugh, Shoemaker,</td>
<td>SOC2034, 2044, 3304, 3504, 3854, 4354, 4424, 4444, 4764, 5424</td>
</tr>
<tr>
<td></td>
<td>participation, violence, social networks, communication behavior and</td>
<td>Wimberley</td>
<td></td>
</tr>
<tr>
<td></td>
<td>effects (incl. in Maasai society)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political Science</td>
<td>National security, world politics, nuclear policies</td>
<td>Avey, Nicholls</td>
<td>PSCI1004, 1024, 1034, 2034, 2054, 2064, 3114, 3514-6, 3524, 3544, 3564, 3515-6, 3624, 3634, 3684-5, 3794, 4734, 5254, 5264, 5284, 5384, 5424, 5444, 5464, 5474, 5514, 5524, 5584, 5624, 5634, 5644</td>
</tr>
</tbody>
</table>
Part 2

• Slides from Dr. Sunshin Lee
• IDEAL: NSF IIS-1319578 – 2013-2017

• Integrated Digital Event and Archive Library
Collecting Webpages

• Started 2007
• Used Internet Archive (IA)
  • 66 collections
  • ~11TB
• Shooting, earthquake, bombing, hurricane
• Problem: very low precision

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Last Crawl</th>
<th>Data (all time)</th>
<th>Docs (all time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texas fertilizer plant explosion</td>
<td>#98765: Feb 2, 2014</td>
<td>1.5 TB</td>
<td>7,398,544</td>
</tr>
<tr>
<td>Hurricane Sandy (October 2012)</td>
<td>#130896: Oct 9, 2014</td>
<td>921.5 GB</td>
<td>14,085,550</td>
</tr>
<tr>
<td>Global Food Crisis</td>
<td>#57243: Oct 24, 2012</td>
<td>716.6 GB</td>
<td>6,151,325</td>
</tr>
<tr>
<td>Boston Marathon Bombing: Twitter</td>
<td>#71632: May 24, 2013</td>
<td>545 GB</td>
<td>7,103,524</td>
</tr>
<tr>
<td>Guatemala Earthquake</td>
<td>#59578: Dec 2, 2012</td>
<td>521.9 GB</td>
<td>2,719,020</td>
</tr>
<tr>
<td>April 16 Archive</td>
<td>#5098: Apr 28, 2008</td>
<td>287.9 GB</td>
<td>4,742,265</td>
</tr>
<tr>
<td>CTRnet - Emergency Preparedness</td>
<td>#131089: Oct 11, 2014</td>
<td>176.3 GB</td>
<td>2,333,228</td>
</tr>
<tr>
<td>Indonesian Volcanic Eruption, Tsunami</td>
<td>#23669: Nov 2, 2010</td>
<td>147.8 GB</td>
<td>2,528,739</td>
</tr>
<tr>
<td>Brazil NightClub Fire</td>
<td>#63372: Feb 2, 2013</td>
<td>94.9 GB</td>
<td>2,118,616</td>
</tr>
<tr>
<td>Virginia Tech Shootings (December)</td>
<td>#41480: Jan 3, 2012</td>
<td>57.9 GB</td>
<td>1,505,721</td>
</tr>
<tr>
<td>Northern Illinois University Shootings</td>
<td>#4916: Apr 16, 2008</td>
<td>45.5 GB</td>
<td>631,082</td>
</tr>
</tbody>
</table>
Collecting tweets

- Over 1375 collections for multiple projects
  - https://docs.google.com/spreadsheets/d/13wUfD-BI49Wkloq8ZezfqTuuwQV0PKIIS9umm0RoM80/edit?ts=59b98e32#gid=0

<table>
<thead>
<tr>
<th>Project</th>
<th>Collection name</th>
<th>Total # of tweet</th>
<th>Started at</th>
<th>Collection tool</th>
<th>Analysis service</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDEAL</td>
<td>Archive DB</td>
<td>1,657,541,394</td>
<td>2012</td>
<td>yTK(^1)</td>
<td>Analysis using Hadoop</td>
</tr>
<tr>
<td>IDEAL</td>
<td>1% sampling</td>
<td>Maintenance</td>
<td>2015</td>
<td>DMI-TCAT(^2)</td>
<td>Analysis</td>
</tr>
<tr>
<td>IDEAL</td>
<td>User following</td>
<td>10,407,631</td>
<td>2015</td>
<td>DMI-TCAT(^2)</td>
<td>Analysis</td>
</tr>
<tr>
<td>IDEAL</td>
<td>Keyword tracking</td>
<td>20,984,747</td>
<td>2015</td>
<td>DMI-TCAT(^2)</td>
<td>Analysis</td>
</tr>
<tr>
<td>GETAR</td>
<td>Collection</td>
<td>127,148,171</td>
<td>2015</td>
<td>yTK(^1)</td>
<td>Analysis using Hadoop</td>
</tr>
<tr>
<td>GETAR</td>
<td>Collection</td>
<td>230,995,656</td>
<td>2016.9</td>
<td>SFM(^3)</td>
<td>Analysis</td>
</tr>
<tr>
<td>NIH</td>
<td>Keyword tracking</td>
<td>622,692</td>
<td>2015</td>
<td>DMI-TCAT(^2)</td>
<td>Analysis</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>2,047,700,291</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Collection Example: GETAR Prototype

• Research key global challenges, e.g., climate change (as well as opportunities), innovation, and resilience

• Initial Collection Effort
  • Started 10/8/2015
  • 315 collections
  • 127,148,395 tweets (as of 3/15/2018)
  • Including global warming,
    • Internet of things,
    • population,
    • and the environment
Collection Example: GETAR Prototype
Archiving and Analyzing using Bigdata Hadoop cluster

• Using Desktop PCs
  • # of Nodes: 20 + 1 (Solr)
  • CPU: Intel i5 Haswell Quad core 3.3Ghz * 20, + Xeon 8C
  • RAM: 704 GB (20 * 32 + 64)
  • HDD: 149 TB (20 * 7 + 9)
  • Backup: 32TB, 8.3TB NAS

• Servers
  • Tweet collecting
  • Web crawling
  • Geocoding
  • Search (Solr)
IDEAL System Architecture

* IDEAL project: Integrated Digital Event Archiving and Library
* Highlighted (as grey) are related to this research
External Tools

• Spark and Mahout:
  • Classification, clustering
  • Topic analysis (LDA), Frequent Pattern Mining
• Solr/Lucene and (Geo)Blacklight: Search/(Faceted) Browse
• Natural Language Processing and Named Entity Recognition
  • NLTK (Python)
  • SNER (Stanford NER)
• . . .
Analysis and Visualization Example
What causes Water Main Breaks?

AccuWeather.com

Heat Causes 700 Water Main Breaks Daily in Houston

Days of hot weather and aging water pipes have resulted in daily water main breaks throughout Texas.

Houston's Mayor Annise Parker told MSNBC that the city is experiencing over 700 breaks a day along 7,000 miles of pipes. The city normally averages about 200 breaks a day in the summer.

Arlington is averaging about six breaks a day, Fort Worth is dealing with about nine or 10 a day, and Ennis has gone from about one a week to about one a day, according to CBSlocal.com.

MassLive.com

Palmer water main break may be result of earthquake, says town official

Published: Tuesday, August 23, 2011, 9:07 PM Updated: Tuesday, August 23, 2011, 9:32 PM

By Patrick Johnson, The Republican

CAUSES OF WATER MAIN BREAKS
- CORROSION WEAKENS PIPES
- EXTREME WEATHER CONDITIONS
- SHIFTING OF GROUND

Pipe repair image courtesy of Photos.com
What causes Water Main Break?

=> Earthquakes (USGS)

Mar. 1 – Apr. 5, 2012
Who is involved in a WMB?

- Fix water pipe
  - Water utility
  - City/town utility
- Traffic
  - Police
- Affected
  - Citizen
- Who else?

Lakewood, NJ, June. 2014

West Philadelphia, PA, June. 2015
Datasets for Geo-location Research

- 5 types of road side small disasters
  - Includes specific location info.
  - 3 small and 2 large collections
  - Collected 2/1/2013 to 6/30/2014
    - 17 months, start and end days varying

<table>
<thead>
<tr>
<th>Size</th>
<th>Dataset</th>
<th>Number of tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td></td>
<td>6,039,888</td>
</tr>
<tr>
<td>Mid-size</td>
<td>Water main break</td>
<td>155,657</td>
</tr>
<tr>
<td></td>
<td>Sinkhole</td>
<td>231,579</td>
</tr>
<tr>
<td></td>
<td>Pothole</td>
<td>324,849</td>
</tr>
<tr>
<td>Big-size</td>
<td>Car crash</td>
<td>2,510,317</td>
</tr>
<tr>
<td></td>
<td>Car accident</td>
<td>2,817,486</td>
</tr>
</tbody>
</table>
Features, combinations of features

Figure 11. Percentages of tweets that have particular features

Figure 12. Percentages of tweets with features or union of features

Up to 91.1%

LIWs may help to geo-locate tweets
State Level Distribution

• Pothole dataset
  • (a) Geo-coordinates: 18K
  • (b) Geonames (SNER): 30K
  • (c) Geonames (SNER)
    + LIWs (Hashtags): 45K

• National trends

Figure 20. State level distributions of unambiguously geo-coded tweets with
(a) geo-coordinates, (b) geoname (SNER), and (c) geoname (SNER) and hashtags in pothole collection
Part 3

- Slides from Dr. Mohamed Magdy Farag

• Integrated Digital Event and Archive Library
## Archive-It Collection Quality

### Table 4. Classification Results

<table>
<thead>
<tr>
<th>Collection</th>
<th>Rel. (%)</th>
<th>Non-rel. (%)</th>
<th># HTML Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama University Shooting</td>
<td>1.4</td>
<td>98.6</td>
<td>6470</td>
</tr>
<tr>
<td>Brazilian School Shooting</td>
<td>8.8</td>
<td>91.2</td>
<td>1120</td>
</tr>
<tr>
<td>Connecticut School Shooting</td>
<td>17.5</td>
<td>82.5</td>
<td>3238</td>
</tr>
<tr>
<td>Northern Illinois University Shooting</td>
<td>26.7</td>
<td>73.3</td>
<td>15385</td>
</tr>
<tr>
<td>Norway Shooting</td>
<td>13.5</td>
<td>86.5</td>
<td>7419</td>
</tr>
<tr>
<td>Youngstown Shooting</td>
<td>40.0</td>
<td>60.0</td>
<td>3427</td>
</tr>
</tbody>
</table>
Representative Event Types

<table>
<thead>
<tr>
<th>Event Types</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shooting</td>
<td>Oregon, California, Orlando</td>
</tr>
<tr>
<td>Earthquake</td>
<td>Ecuador, Japan</td>
</tr>
<tr>
<td>Flood</td>
<td>Texas Floods</td>
</tr>
<tr>
<td>Fire</td>
<td>California wild fire</td>
</tr>
<tr>
<td>Bombing</td>
<td>Boston bombing</td>
</tr>
<tr>
<td>Plane Crash</td>
<td>Egyptair, germanwings</td>
</tr>
<tr>
<td>Building Collapse</td>
<td>East Harlem</td>
</tr>
<tr>
<td>Protests/Riots</td>
<td>Egyptian revolution</td>
</tr>
<tr>
<td>Political Issue/Conflict</td>
<td>Brexit, Turkey coup, Greece Bailout referendum</td>
</tr>
<tr>
<td>Hurricane</td>
<td>Joaquin, Sandy, Katrina</td>
</tr>
<tr>
<td>Terror Attack</td>
<td>Paris, Brussels, Nice</td>
</tr>
<tr>
<td>Train Derailment</td>
<td>Amtrak188</td>
</tr>
<tr>
<td>Scandal</td>
<td>Panama Papers, Sepp Blatter</td>
</tr>
<tr>
<td>Community</td>
<td>Lovewins</td>
</tr>
</tbody>
</table>
Data Flow

Event

Keyword/Hashtag

Collect Tweets

Tweet Collection

Extract URLs

Shortened URLs

Expand

Original URLs

Archive/Organize/Analyze

Index

SOLR

Fetch

Webpages

Archive

WARC

Access

Search

Browse

Wayback
Baseline Focused Crawler

Focus of research
Event Focused Crawler
Event Modeling and Representation

• We define an event as
  • something (e.g., a disaster),
  • which happened in a certain place, and
  • at a certain time.
• Event E is a tuple <T, L, D>.
• T = topic of event, L = location, D = date
• i.e.: what, where, and when.
Building Event Model

Seed Webpages

Build Event Model

Extract Topic Vector
- Word Tokenization
- Frequency Analysis

Extract Locations Vector
- Sentence Tokenization
- Named Entity Recognition

Extract Publishing Date
- Regular expression
- HTML parsing

Event Model

Topic: [(t_1, f_1), (t_2, f_2), ..., (t_n, f_n)]
Loc: [(l_1, f_1), (l_2, f_2), ..., (l_m, f_m)]
Date: mm/dd/yyyy
## California Shooting Model

<table>
<thead>
<tr>
<th>Topic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>shoot</td>
<td>0.93</td>
</tr>
<tr>
<td>san</td>
<td>0.513</td>
</tr>
<tr>
<td>bernardino</td>
<td>0.465</td>
</tr>
<tr>
<td>said</td>
<td>0.357</td>
</tr>
<tr>
<td>wa</td>
<td>0.323</td>
</tr>
<tr>
<td>2015</td>
<td>0.321</td>
</tr>
<tr>
<td>peopl</td>
<td>0.31</td>
</tr>
<tr>
<td>california</td>
<td>0.305</td>
</tr>
<tr>
<td>polic</td>
<td>0.258</td>
</tr>
<tr>
<td>suspect</td>
<td>0.177</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>San Bernardino</td>
<td>1</td>
</tr>
<tr>
<td>California</td>
<td>0.51</td>
</tr>
<tr>
<td>Calif.</td>
<td>0.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-12-02</td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>Type</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>California Shooting</td>
<td>Shooting</td>
</tr>
<tr>
<td>Brussels Attack</td>
<td>Terrorist Attack</td>
</tr>
<tr>
<td>Oregon Shooting</td>
<td>Shooting</td>
</tr>
<tr>
<td>Egyptair Plane Crash</td>
<td>Plane Crash</td>
</tr>
<tr>
<td>Panama Papers Leak</td>
<td>Document Leak</td>
</tr>
<tr>
<td>Orlando Shooting</td>
<td>Shooting</td>
</tr>
<tr>
<td>Paris Attack</td>
<td>Terrorist Attack</td>
</tr>
<tr>
<td>Ecuador Earthquake</td>
<td>Earthquake</td>
</tr>
</tbody>
</table>
Large Scale: California Shooting
Large Scale: Brussels Attack

![Graph showing the comparison between Baseline Focused Crawler and Event Focused Crawler performance in terms of the percentage of relevant webpages over the total number of crawled webpages. The graph indicates that the Event Focused Crawler outperforms the Baseline Focused Crawler as the number of crawled webpages increases.]
Large Scale: Egyptair Plane Crash

![Graph showing Egyptair Plane Crash Baseline vs. Event Focused Crawler Performance](image)
Twider: A Hybrid Model for Role-related User Classification on Twitter

Presenter: Liuqing Li
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018
Framework
Discussion

• Features
Presenter: Liuqing Li
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018
Preliminary Results

- Data Collections
  - #55 – Quantico shooting 2013/03/22
  - #89 – Santa Monica shooting 2013/06/07
  - #145 – Nevada school shooting 2013/10/21
  - #186 – Shooting California 2014/05/25
  - #435 – Ottawa Shooting 2014/10/22
  - #443 – Marysville shooting 2014/10/24
**Preliminary Results**

### Tweets with Different Number of Short URLs

<table>
<thead>
<tr>
<th>ID</th>
<th># of URLs</th>
<th># of Unique URLs</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>#55</td>
<td>36,704</td>
<td>1,377</td>
<td>3.8%</td>
</tr>
<tr>
<td>#89</td>
<td>32,089</td>
<td>3,092</td>
<td>9.6%</td>
</tr>
<tr>
<td>#145</td>
<td>21,497</td>
<td>2,602</td>
<td>12.1%</td>
</tr>
<tr>
<td>#186</td>
<td>665,895</td>
<td>62,096</td>
<td>9.3%</td>
</tr>
<tr>
<td>#435</td>
<td>186,600</td>
<td>15,686</td>
<td>8.4%</td>
</tr>
<tr>
<td>#443</td>
<td>59,554</td>
<td>6,338</td>
<td>10.6%</td>
</tr>
</tbody>
</table>
Preliminary Results

Percentage of Webpages (Code 200) per Year

Percentage of Retrieved Webpages per Year
School Violence - GETAR

Presenter: Jason Callahan & Dr. Shoemaker
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018
Themes to Evaluate/Refine

- Victim (gender, race, age, student vs non-student)
- Suspect (gender, race, age, student vs non-student)
- Type of weapon used (firearm, blade, etc.)
- Geographical location/region (population size)
- Suspect killed vs. suspect survives
*Themes of tweets/URL
*Locations of tweets (potential geotags)
*Media response tweets (news vs. non-news sources)
*Emotional response tones/themes (measured by volume/frequency)
*Clustering of terms/related incidents (hashtags of events/suspects/victims consolidated)
*Word clouds
*Pie/bar charts to illustrate the refined themes
*Time sequence tracking of refined themes
*Maps of twitter data if geotags are available
Technology on Trail Study

Presenter: Abigail Bartolome
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018
Trail Study

“Each one is different; each has a soul” - Triple Crown veteran, Karen Berger, on which trail is her favorite.

<table>
<thead>
<tr>
<th>Appalachian Trail Topics</th>
<th>Pacific Crest Trail Topics</th>
<th>Continental Divide Trail Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>#indigenous, #flats, #caryards, day, hike, mcafee, traila</td>
<td>California, #pc2017, 2, story, tips, resupply, #pcs</td>
</tr>
<tr>
<td>Topic 2</td>
<td>va, catawba, suit, halfway, #backpacking, just, co4444x2400</td>
<td>@pcassociation, lisa, today, #lifef, win, did, great</td>
</tr>
<tr>
<td>Topic 3</td>
<td>days, trip, long, mountain, complete, miles, week</td>
<td>mount, adams, goat, rocks, @hgranoe, washington, view</td>
</tr>
<tr>
<td>Topic 4</td>
<td>#travel, #bestseller, black, 1, awel, books, 2</td>
<td>#shesh, booth, gear, come, free, #pdc2017, @donnie</td>
</tr>
<tr>
<td>Topic 5</td>
<td>new, going, woman, eightyold, side, sisters, twin</td>
<td>wild, lost, #travel, #bestseller, @organ, #7, #9</td>
</tr>
<tr>
<td>Topic 6</td>
<td>hiker, #pdc2017, @thelake, #update, #trail, #hikers</td>
<td>taking, #backpacking, carries, job, #big, better, 4, day</td>
</tr>
<tr>
<td>Topic 7</td>
<td>#pdc2017, @thelake, #trail, gear, limit, things, #photography</td>
<td>#pedays, new, instagram, year, weeks, go, #breeze</td>
</tr>
</tbody>
</table>

From January-May, topic analysis reflected Appalachian Trail valued experiences, while Pacific Crest Trail focused more on the logistics of planning a hike. Are these part of hiking culture? Was this influenced by geographical (and schedule) differences?

Trail Cultures:
• Avid Hiking
• Conservation practices

What can we learn about these cultures? What can we learn about their language?

Surprising Trail Countercultures:
• Nude Hiking
• Actively denying conservation practices

Why do these hikers behave in this way? What are their motivations? Who is attracted to these countercultures and can they be infiltrated?
GETAR Collection + GeoBlacklight

Presenter: Ziqian Song
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018
Homepage

Presenter: Shou Niu
Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA, 24061
February 20, 2018
TweetBank

• A web portal to explore GETAR Twitter collection.
• Developed in Fall17
• Functions:
  • Searching
  • Tweet viewing
  • Social network
  • User information
  • Time-line
  • Keywords
  • Geo-locations

http://mule.dlib.vt.edu/cs5604f17_fe/TweetBank/src/
Planned Activities – Welcoming Involvement

- Collaboration with Internet Archive to aid research community
- Aid some 30 local stakeholders
- Variety of interfaces across information life cycle
- Collect, Add Value, Archive, Analyze, Search/Browse, Visualize

- Displays outside 2030 Torgersen Hall (DLRL)

- Many volunteers: CS4624, CS5604, CS6604, Theses, Independent Studies, and others at all levels
Summary

• Context
• GETAR proposal
• IDEAL results – Sunshin Lee
• IDEAL results – Mohamed Farag
• Selected GETAR projects
• Welcoming collaboration