Agenda

- Introduction
- Motivation
- Related Work
- Collection Retrieval and Processing
- Visualizations
- Case Studies
- Future Work
- Conclusion
INTRODUCTION AND MOTIVATION
Digital Archives

http://www.loc.gov/index.html
http://digitalcollections.library.yale.edu/
Archive-It Collection Hierarchy

- **Root**
  - Collection Title

- **Level 1**
  - Category 1
  - Category n

- **Level 2**
  - Web page 1
  - Web page n

- **Level 3 (Leaf Nodes)**
  - Archived Version 1
  - Archived Version n
Exploring Archive-It Collections

http://archive-it.org/collections/1068
Exploring Archive-It Collections

http://archive-it.org/collections/1068
Exploring Archive-It Collections

http://archive-it.org/collections/1068
Exploring Archive-It Collections

http://wayback.archive-it.org/1068/*http://acda.co/

<table>
<thead>
<tr>
<th>Found 4 Captures between Oct 13, 2011 - Jun 2, 2012</th>
</tr>
</thead>
</table>
| Oct 13, 2011  
Dec 2, 2011 | 2 pages |
| Mar 2, 2012  
Jun 2, 2012 | 2 pages |

http://wayback.archive-it.org/1068/*http://acda.co/
Exploring Archive-It Collections

http://archive-it.org/collections/1068
Exploring Archive-It Collections

http://archive-it.org/collections/2836
Drawbacks

- No visual feedback
- Discovering individual pages is difficult
- Optional metadata and categorization
- Collection structure known only to curator
Contribution

Interactive visualizations
- Treemap
- Time cloud
- Bubble chart
- Image plot
- Wordle
- Timeline

Temporal exploration of collections
Uncover collection structure
RELATED WORK
Microsoft Pivot

http://www.microsoft.com/silverlight/pivotviewer/
3D Wall

http://www.webarchive.org.uk/ukwa/wall/Blogs
Johnson and Shneiderman, "Space-Filling Approach to the Visualization of Hierarchical Information Structures" in proceedings of the 2nd conference on Visualization '91
Series Browser

Wordle

DATA RETRIEVAL AND PROCESSING
### 11 Collections, 2K+ Web pages, 70K+ Mementos

<table>
<thead>
<tr>
<th>ID</th>
<th>Collection Name</th>
<th>Time Span</th>
<th>Groups</th>
<th>URI Domains</th>
<th># of Webpages</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>South Dakota Government</td>
<td>3 Days</td>
<td>1</td>
<td>50</td>
<td>88</td>
</tr>
<tr>
<td>12</td>
<td>State Minnesota Sites</td>
<td>3 Weeks</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>Ari Salomon Archive</td>
<td>1 Day</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>194</td>
<td>NC State Government Web Site Archive</td>
<td>14 Years</td>
<td>1</td>
<td>451</td>
<td>609</td>
</tr>
<tr>
<td>499</td>
<td>Archive Montana: Preserving State Agency Websites</td>
<td>15 Years</td>
<td>36</td>
<td>132</td>
<td>144</td>
</tr>
<tr>
<td>667</td>
<td>The New York Greens</td>
<td>1 Day</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>677</td>
<td>Actors Equity Association</td>
<td>1 Day</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1068</td>
<td>Human Rights</td>
<td>3 Years</td>
<td>5</td>
<td>341</td>
<td>365</td>
</tr>
<tr>
<td>1621</td>
<td>Chile Earthquake</td>
<td>1 Day</td>
<td>1</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>2323</td>
<td>Jasmine Revolution - Tunisia 2011</td>
<td>5 Months</td>
<td>4</td>
<td>147</td>
<td>223</td>
</tr>
<tr>
<td>2836</td>
<td>Pakistan Floods (2011)</td>
<td>20 Days</td>
<td>1</td>
<td>253</td>
<td>623</td>
</tr>
</tbody>
</table>
Data Retrieval & Processing

Retrieval:
- Screen scrape
- Copy collection hierarchy
- Store page content

Processing:
- Calculate TF and TF-IDF
- Generate screenshots
- Generate wordles
- Rule-based categorization
- Construct JSON
No categorization
Improper Categorization

http://www.archive-it.org/collections/2323
Rule based categorization

Narrow Your Results

There are no further ways to narrow your results.

News Web pages

Blogs

Social Media

Videos
Special URI and TLD based categorization

Pakistani news web pages

http://www.archive-it.org/collections/2836
VISUALIZATIONS
Time Cloud
### Bubble Chart, Image Plot & Timeline

#### COLLECTION: Human Rights

| 1. Non-governmental organizations (NGOs) – Amnesty International sites |
|---|---|---|---|---|---|
| 2. Non-governmental organizations (NGOs) |
| 3. National human rights institutions (NHRIs) |
| 4. Blogs by individuals |
| 5. Truth Commissions, Tribunals, and Courts |

- 40
- 263
- 50
- 6
- 6

---

Representing 365 articles in 5 curated categories

1729 mementos have been recorded between 15-May-2008 to 22-April-2011
1. Collection Building and Growth
2. Re-Categorization
(Pakistan Flood: no categorization)
2. Re-Categorization
(Pakistan Flood: after categorization)
3. Collection Synopsis

Pakistan floods (2011)

Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network

Archived since: Sep, 2011

Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead.

Subject: Spontaneous Events, Floods, Pakistan

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.
3. Collection Synopsis
3. Collection Synopsis
3. Collection Synopsis

![Diagram showing collection synopsis of news websites, others, blogs, and social media.](image-url)
3. Collection Synopsis
4. Theme Tracking
4. Theme Tracking

1997 - 1998
Primary: General Information
Secondary: Transportation
4. Theme Tracking

2002 - 2003
Primary: General Information
Secondary: Election
4. Theme Tracking

2007 2008 carolina
center community contact
county department
home information
link interaction
nc north office
program public resource
school service site

2008 - 2009
Primary: Education, Health
Secondary: Government, Business

27699 919 business
department dhh division governor
health
nc ncmail north office program
raleigh resource
Informal User Evaluation

Alex Thurman, Columbia University Libraries

Feedback on

- ease of browsing and obtaining information
- user-friendliness of the interface
- whether they prefer textual or graphical interface
- most effective visualization
- effectiveness of the rule-based categorization in exploring archives
Feedback

Effective visualizations:

- Treemap – color coding useful for identifying newer additions
- Image plot – screenshots with mouse-over wordles allow for good navigation
- Timeline – useful for visualizing development of groups in collection

Suggestions

- Broader timescale for treemaps
- Include stop words from other languages
FUTURE WORK AND
CONCLUSION
Future Work

- N-Gram wordles
- Term expansion
- Krovetz stemmer (dictionary based stemmer)
- Integration with Archive-It
- Detailed user evaluation
- Implementation for other archives
Conclusion

*Identified metrics for collections*
Conclusion

- Identified metrics for collections

Visualizations

- Treemap
Conclusion

- Identified metrics for collections
- Visualizations
  - Treemap
  - Time cloud
Conclusion

- Identified metrics for collections
- Visualizations
  - Treemap
  - Time cloud
  - Bubble chart

![Diagram showing bubble chart with categories and numbers]

- Representing 265 articles in 9 curated categories
- 1729 mentions have been recorded between 15 May 2008 to 22 April 2011
Conclusion

- Identified metrics for collections
- Visualizations
  - Treemap
  - Time cloud
  - Bubble chart
  - *Image plot*
Conclusion

- Identified metrics for collections
- Visualizations
  - Treemap
  - Time cloud
  - Bubble chart
  - Image plot
  - Wordle
Conclusion

- Identified metrics for collections
- Visualizations
  - Treemap
  - Time cloud
  - Bubble chart
  - Image plot
  - Wordle
  - Timeline
Conclusion

- Identified metrics for collections
- Visualizations
  - Treemap
  - Time cloud
  - Bubble chart
  - Image plot
  - Wordle
  - Timeline
- Rule – based categorization
## Time Span

### Human Rights Web Archive (Columbia University Libraries)

<table>
<thead>
<tr>
<th>Time span</th>
<th>Small</th>
<th>1 Day - 2 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>2 Weeks - 4 Months</td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>&gt; 4 Months</td>
<td></td>
</tr>
</tbody>
</table>

### Found 10 Captures between May 14, 2009 - Mar 2, 2012

<table>
<thead>
<tr>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 page</td>
<td>4 pages</td>
<td>4 pages</td>
<td>1 page</td>
</tr>
<tr>
<td>Jun 2, 2010 *</td>
<td>Sep 2, 2010 *</td>
<td>Dec 2, 2010 *</td>
<td>Dec 2, 2011 *</td>
</tr>
</tbody>
</table>

---

http://wayback.archive-it.org/1068/*http://amigosdemujeres.org/

7/20/2012

MS Thesis - August 2012
Groups

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2 - 5</td>
<td>&gt; 5</td>
</tr>
</tbody>
</table>

http://www.archive-it.org/collections/1068
URI Domains

<table>
<thead>
<tr>
<th>URI Domains</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>1 - 10</td>
</tr>
<tr>
<td>Medium</td>
<td>11 - 20</td>
</tr>
<tr>
<td>Large</td>
<td>&gt; 20</td>
</tr>
</tbody>
</table>

http://www.archive-it.org/collections/2836
Number of Web Pages

<table>
<thead>
<tr>
<th># of Web Pages</th>
<th>Small</th>
<th>1 - 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Medium</td>
<td>11 - 99</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>&gt; 99</td>
</tr>
</tbody>
</table>

[http://www.archive-it.org/collections/2836]
The river

Wei et.al. in SIGKDD, 2010
Bubble Chart

<table>
<thead>
<tr>
<th>Collection: Human Rights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Non-governmental organizations (NGOs)--Amnesty International sites</td>
</tr>
<tr>
<td>2. Non-governmental organizations (NGOs)</td>
</tr>
<tr>
<td>3. National human rights institutions (NHRIs)</td>
</tr>
<tr>
<td>4. Blogs by individuals</td>
</tr>
<tr>
<td>5. Truth Commissions, Tribunals, and Courts</td>
</tr>
</tbody>
</table>

Representing 365 articles in 5 curated categories
1729 mementos have been recorded between 15-May-2008 to 22-April-2011
Image Plot with Wordle

Timeline