Visualizing Digital Collections at Archive-It

Kalpesh Padia
kpadia@cs.odu.edu

Yasmin AlNoamany
yasmin@cs.odu.edu

Michele C. Weigle
mweigle@cs.odu.edu

Department of Computer Science
Old Dominion University
Norfolk, VA 23529

ABSTRACT

Archive-It, a subscription service from the Internet Archive, allows users to create, maintain and view digital collections of web resources. The current interface of Archive-It is largely text-based, supporting drill-down navigation using lists of URIs. To provide an overview of each collection and highlight the collection's underlying characteristics, we present four alternate visualizations (image plot with histogram, wordle, bubble chart and timeline). The sites in an Archive-It collection may be organized by the collection curator into groups for easier navigation. However, many collections do not have such groupings, making them difficult to explore. We introduce a heuristics-based categorization for such collections.

Categories and Subject Descriptors

H.3.5 [**On-line Information Services**]: Commercial services; H.3.7 [**Digital Libraries**]: Collection; H.5.2 [**User Interfaces**]: Graphical user interfaces

General Terms

Human Factors

Keywords

Digital collections, information visualization, bubble chart, timeline, image plot, wordle

1. INTRODUCTION

Archive-It[1] is a subscription service developed by the Internet Archive to allow institutions to harvest and preserve collections of digital content. For each collection, Archive-It provides a listing of all URIs in the collection along with the number of times and dates over which each site was archived. Archive-It also provides full-text search of archived sites, allowing users to quickly search for topics of interest within a collection or spanning multiple collections. However, the organization of a collection is dependent upon the collection's curator. If the curator has organized URIs into groups and assigned subjects, coverage, and creator tags, Archive-It provides filtering of the URIs on the left-side of the listing page (Figure 1, foreground, Human Rights collection[2]). If the curator has not defined such groups or tags, then there are no filtering options available (Figure 1, background, Pakistan Floods collection[3]). If there are no groups defined or if there are a large number of sites in a group, exploration of the collection can be cumbersome.

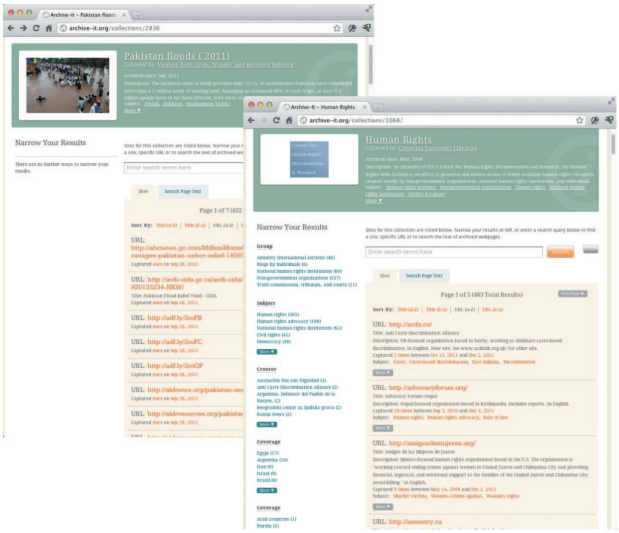

Figure 1: Archive-It's interface

We present an alternate interface for exploring Archive-It collections. This interface consists of multiple visualizations, namely image plot with histogram, wordle, bubble chart, and timeline. These visualizations help to provide an overview of each collection and highlight the collection's underlying characteristics, allowing the user to progressively gain insight into the collection. For those collections that lack a curator-defined grouping, we provide a heuristics-based categorization to make the new visualizations more meaningful. To test this interface, we picked several Archive-It collections that differ widely from each other. The collections used and their characteristics are listed in Table 1.

[1] `http://www.archive-it.org`

[2] http://archive-it.org/collections/1068

[3] http://archive-it.org/collections/2836

JCDL'12, June 10–14, 2012, Washington, DC, USA.

2. RELATED WORK

Considerable research has been dedicated towards developing visualizations for viewing and querying documents, and towards graphical querying and browsing of results [1, 2]. Jatowt et al. [3] describe a browser for viewing past web resources with changes animated on a timeline. Popular web archives such as Archive-It, Internet Archive[4], California Digital Library[5], and Library of Congress[6] provide a good user interface for searching through the collections, however they still use list view as their main interface which makes it difficult for users to gain insight about the whole collection. One exception, the UK Web Archive[7], provides a 3D wall visualization for selected collections, allowing interaction through zooming.

3. DESIGN

Our approach for an alternate interface for viewing Archive-It collections consists of visualizations and heuristics-based categorization of sites.

3.1 Visualizations

We have created four different visualizations for collections in Archive-It viz. image plot with histogram, wordle, bubble chart, and timeline, presented in an integrated interface allowing users to switch between visualizations via the filters tab (Figure 2, black band at bottom).

3.1.1 *Image plot with histogram*

The *image plot with histogram* view (Figure 2) is an implementation of an inverted stacked bar chart to represent all sites in a collection in a graphical manner. The chart is divided based on the collection's defined groups. This representation allows the user to explore the collection by viewing a screenshot of a recent version of each archived site. Each screenshot is linked to Archive-It's list of archived versions of that site. The inverted representation allows the user to see both larger and smaller groups side by side. Since it is likely that not all sites will be viewable in a single window, a resizable histogram in the bottom right corner shows the number of sites in each group, so that the user gains an overview of the distribution of sites over the groups.

3.1.2 *Wordle*

Hovering over any image in the image plot reveals a *wordle* [5] (Figure 2 overlay) summarizing the content discussed on the site. For multimedia content in the collection, the wordle summarizes the comments (if present). This wordle representation aids the understanding of a site in the collection by supplementing the visual representation provided by the image plot. Analyzing various wordles allows the user to quickly grasp the key ideas of the collection.

3.1.3 *Bubble chart*

The *bubble chart* visualization (Figure 3) provides a quick summary of the collection by displaying each group in the collection as a bubble, where the size of the bubble represents the number of sites in each group. The number of total sites, archived copies, and duration over which the collection has been constructed is also visible to the user below the bubble chart. Each bubble links to Archive-It's default list of sites in that group, allowing the user to quickly filter through the collection by group.

Figure 2: Image plot with histogram visualization for the Human Rights collection, showing the wordle for the highlighted site.

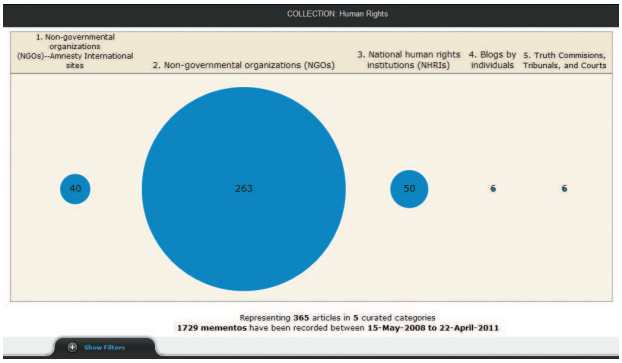

Figure 3: Bubble chart visualization for the Human Rights collection.

3.1.4 *Timeline*

The previous visualizations provide a quick statistical summary (bubble chart) and content summary (image plot with histogram, wordle) of the collection. However, curators may be interested in discovering how the collection developed over time to correlate events in history with the organization of the collection. We provide a *timeline* visualization (Figure 4) for visualizing the development of the collection over time. In this visualization, each site is represented as a single horizontal line, the length of which denotes the duration over which its archived copies have been captured.

[4] http://www.archive.org
[5] http://www.cdlib.org
[6] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html
[7] http://www.webarchive.org.uk/ukwa

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

Table 1: Example collections used for developing and testing the visualizations.

Each point on the line represents an archived copy of the site. Hovering over a point displays a list of archives of other sites captured on that same day. A curator can easily see the growth of a collection over time by looking at site density and analyzing the addition (or removal) of sites from the collection.

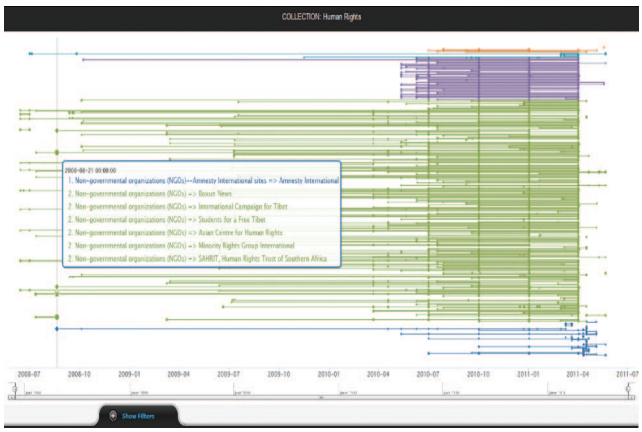

Figure 4: Timeline visualization for the Human Rights collection.

3.2 Heuristics-based categorization

Several Archive-It collections do not have the sites of the collection organized into groups (*e.g.*, Pakistan Floods collection), making it difficult for the user to explore the collection. Thus, we provide an option of exploring the collection using a heuristics-based categorization. The usefulness of URIs for webpage classification has been demonstrated before [4]. We implement a similar approach. Our rules for heuristics-based categorization are based on the hostname component of URI. While most rules were run on all collections, certain rules were applied only to specific collections. Examples of our categorization rules are below:

- If the hostname contains facebook, twitter, or wiki, group the site under "Social Media".
- If the hostname indicates a news site, such as bbc, cnn, nytimes, group the site under "News Web Sites".
- If the hostname contains blog or wordpress, group the site under "Blogs".
- If the hostname contains youtube or dailymotion, group the site under "Videos".

We also developed rules based on the top level domain name (TLD) of a website. For example, the sites in Collection 11, South Dakota Government, were grouped based on the TLD. All sites with a TLD ".gov" were grouped as government web sites and those with the TLD ".edu" were grouped as education web sites. We also developed specific rules for particular collections. For example, Collection 194, NC State Government, has specific rules for the "North Carolina Web Sites" group. If we were unable to group a site using any predefined (either generic or specific) rule, it was grouped under "Others". Such heuristics-based categorization is helpful in organizing the collection, and also in helping users understand which sources and what media types contribute the most to a collection.

4. CASE STUDIES

From the list of collections in Table 1, we highlight two collections to demonstrate how our interface can expose hidden characteristics and provide meaningful insight about the collection to both users and curators. One collection is well-curated and one did not have any curator-defined groups.

4.1 Collection: Human Rights

The Human Rights collection (ID:1068) is a large collection built over a period of 3 years and is an example of a well-curated collection. Due to its large size (365 sites, 1729 archived copies, 5 groups), it can be overwhelming for the user to understand the common topics discussed in the collection, identify the most active sites (and groups), and observe the collection's growth over time.

Figures 2-4 show the various visualizations for this collection (image plot, bubble chart, and timeline, respectively). By looking at the image representation of sites in the collection (Figure 2), the user can get visual feedback and easily locate a site of interest. Upon hovering over any site, the associated wordle effectively summarizes the content, allowing the user to quickly understand the basic content without going to the site. This summary saves time during preliminary exploration. Also the histogram in the bottom right corner quickly summarizes the size of each group, informing the user of the largest (and perhaps the most important) group in the collection.

Since this collection spans a long time, it is interesting for a curator to see how the collection evolved over time using the timeline view (Figure 4). There are a few insights that can be gained through this visualization:

- The "Non-governmental organizations (NGOs)" group

has been the largest since the beginning of the collection.

- Short green lines on the left of the visualization indicate that a few sites were initially added to the collection but quickly removed.
- A handful of sites have remained in the collection since the beginning, and regular archived copies have been made. This is indicated by long lines with frequent data points. This hints at the high importance of these sites in the collection and perhaps are ones the user should explore first.
- A change in the composition of the collection is evident from the abrupt ending of many long lines on the right-hand side of timeline (at 2011-04 time marker) with only a few remaining until the end.

4.2 Collection: Pakistan Floods (2011)

The Pakistan Floods (2011) collection is a good representative of collections with a large number of sites but no curator-defined grouping. The collection contains 655 archived copies of 623 sites collected over a period of 20 days. The absence of any grouping makes it difficult for the user to selectively explore the collection.

After running our heuristics-based categorization, the collection was organized into 8 groups: Blogs (35 sites), News Websites (259 sites), Pakistan News Sites (47 sites), Relief Websites (5 sites), Social Media (105 sites), Television (3 sites), Videos (18 sites) and Others (151 sites). This categorization is shown in the bubble chart (Figure 5). Since this collection is concerned with a natural disaster, we added specific rules to categorize news websites from Pakistan and websites discussing relief operations into their own groups. A large number of websites were categorized as "Others" because the URIs did not meet any specified rule. After this categorization, the composition of the collection can be presented more clearly to the user. Also by looking at the statistics below the bubble chart, the user can infer almost immediately that this collection is still in its early stages because for most of the sites there is only 1 archived copy.

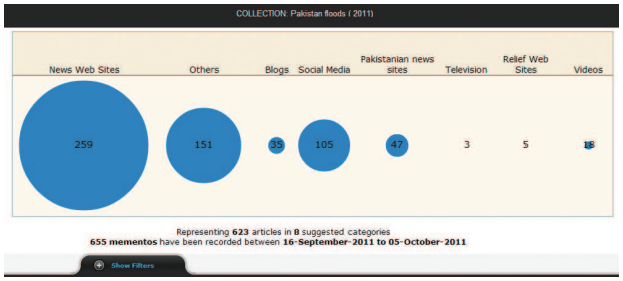

Figure 5: Bubble chart visualization for the Pakistan Floods (2011) collection.

As before, the timeline visualization is effective in visualizing the growth of the collection over time. Figure 6 shows that only one site has contributed consistently to the collection and a large number of sites were added on a single day. By looking at the visualization, one can infer that this collection was perhaps meant to be an archive of various news items related to the disaster and not as a repository of events taking place over time (as with the Human Rights collection).

Figure 6: Timeline visualization for the Pakistan Floods (2011) collection.

5. CONCLUSION AND FUTURE WORK

We have developed a novel visual interface for exploring digital collections at Archive-It. The various visualizations that constitute this interface provide insight into the collections. We have found the timeline visualization to be particularly useful in understanding the structure of a collection. The bubble chart provides a quick look at the collection statistics, while the image plot allows for visual exploration of the collection. Also, the wordles enable understanding of the collection by providing a summary of the selected site.

These visualizations are helpful in exploring the digital collections and offer insight into the collections over the textual interface. Based on feedback from Archive-It partners and collection curators, we plan to further develop these visualization techniques, and integrate them with the existing Archive-It interface. Also, the heuristics-based categorization is still in a naive stage and we plan to further develop it to include a broad range of collection types.

6. REFERENCES

[1] C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *Proceedings of ACM SIGCHI*, CHI '94, pages 313–317, 1994.

[2] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K. Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):42–49, 2002.

[3] A. Jatowt, Y. Kawai, S. Nakamura, Y. Kidawara, and K. Tanaka. A browser for browsing the past web. In *Proceedings of WWW*, pages 877–878, 2006.

[4] M.-Y. Kan and H. O. N. Thi. Fast webpage classification using URL features. In *Proceedings of ACM CIKM*, pages 325–326, 2005.

[5] F. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, 2009.