Summer 2012

Visualizing Digital Collections at Archive-It

Kalpesh Padia
Old Dominion University

VISUALIZING DIGITAL COLLECTIONS AT ARCHIVE-IT

by

Kalpesh Padia
B.E. June 2009, Visvesvaraya Technological University, Belgaum, INDIA

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2012

Approved by:

Michele C. Weigle (Director)

Michael L. Nelson (Member)

Ravi Mukkamala (Member)

ABSTRACT

VISUALIZING DIGITAL COLLECTIONS AT ARCHIVE-IT

Kalpesh Padia
Old Dominion University, 2012
Director: Dr. Michele C. Weigle

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. While this interface provides good searching capabilities, it is not efficient for browsing. In the absence of keywords, a user has to spend large amount of time trying to locate a web page of interest. In order to provide a better visual experience to the user, we have studied the underlying characteristics of Archive-It collections and implemented six different visualizations (treemap, time cloud, bubble chart, image plot, timeline and wordle), each highlighting one or more of the underlying characteristics of the collection. Archive-It supports grouping of web pages into categories, however, it does not enforce its usage. As a result there are many collections with missing or improper grouping. For such collections, we present a method of grouping web pages based on a set of pre-defined rules.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Michele Weigle, my advisor, for her guidance, constant encouragement and infinite patience, without which this may not have been possible. She knew just when to nudge and when to push me through, and for that I am very grateful. I would also like to thank my thesis committee, Dr. Michael Nelson and Dr. Ravi Mukkamala for their guidance and input.

I would like to thank our partners at Archive-It, Kristine Hanna, Lori Donovan and Kate Odell, for providing us the opportunity to work with them. I would also like to thank Alex Thurman, librarian and curator at Columbia University for evaluating our visualizations and providing us with an informal feedback.

I am grateful to Yasmin AlNoamany, friend, project partner and co-author, for all her invaluable contributions to this work. I would like to thank Karthik Navuluri, for pushing me to go for the extra mile and for making sure I am on track. I would also like to thank Priyanka Dharamshi for providing me moral support whenever I needed, and my roommates, Abhishek Biswas, Shreyas Ramesh, Arumugam Kamesh and Santosh Venkateswara, for all the fun we had in the last two years.

Lastly and most importantly, I would like to thank my parents and my sister for their unyielding support and for encouraging me to pursue higher education. To them I dedicate this thesis.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Archive-It[1] is a web archiving service that allows individuals and organizations, called partners, to create and archive collections of web pages. It provides a convenient web-based interface for its partners to create and manage collections. It extends a similar textual web interface to the general public for exploring the web collections.

The textual interface lists all web pages in the collection alphabetically and provides options for searching/filtering through this list based on a number of meta-data parameters like group, subject and creator. However, it still poses significant problems to the user for browsing through a collection. For instance a user may have to navigate through a series of pages before actually visiting a web page of interest, when the exact title or URI of the page is not known. This is because no means exist to enable the user to navigate through the collection based on content synopsis. Also, since Archive-It does not require that meta-data always be present, in many collections this information is missing. This adds to the difficulty in looking for a certain web page in the collection when missing meta-data makes filters unavailable.

In order to alleviate the problems in browsing through Archive-It collections, we have implemented multiple visualizations which present the collections to the user as different visual representations. We have identified the metrics around which every Archive-It collection is built and leverage those metrics to create efficient and meaningful visualizations.

These visualizations help the user to not only browse and explore the collection visually, but also provide the user insight about collection structure, collection statistics and allow the user to quickly grasp the gist of the collection. For those collections where there is no curator-defined grouping/categorization, we introduce a

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

rule-based categorization scheme which organizes web pages in a collection based on some preset rules. These pre-defined rules induce some structure into the collection while interacting with the visualizations and allowing the user to quickly identify a web page of interest or gain insight about the collection.

I.1 MOTIVATION

I.1.1 ARCHIVE-IT

The web has become an integral part of our lives, serving as a key component in our social and economic interactions. While the web is growing at a rapid rate, with at least 8.58 billion pages in the indexed web (as of March 2012) [1], the average lifespan of a web page was a mere 75 days in 2001 [2] and about 100 days in 2003 [3]. Though the average lifespan of a web page is slowly increasing, there is a need to archive the web to preserve the reflection of our social and cultural heritage. Founded in 1996, the Internet Archive[2] is dedicated towards archiving the web and it alone contains more than 5 petabytes of data. But, because of the massive size of the web, the Internet Archive cannot archive every page. There are many other organizations that perform archiving services. Several of these are associated with national libraries, such the Library of Congress Web Archives [4], the UK Web Archive [5], the Pandora project [6] at the National Library of Australia, the Greek Web project [7], Swedens Kulturarw3 [8], and Frances BnF [9]. Others, such as the California Digital Library [10] and Stanfords WebBase project [11], are associated with university libraries.

However, these archives cannot save everything. There is material on the web that smaller institutions or organizations would want to archive and is not covered by current archiving projects. Thus there is a need for services that allow institutions and individuals to create there own archives. Deployed in 2006, Archive-It is a subscription-based web archiving service by Internet Archive that allows institutions and individuals to build, manage, and archive their own collections. Subscribers, or partners, can harvest, catalog, and archive their collections, and then search and

[2] http://archive.org/

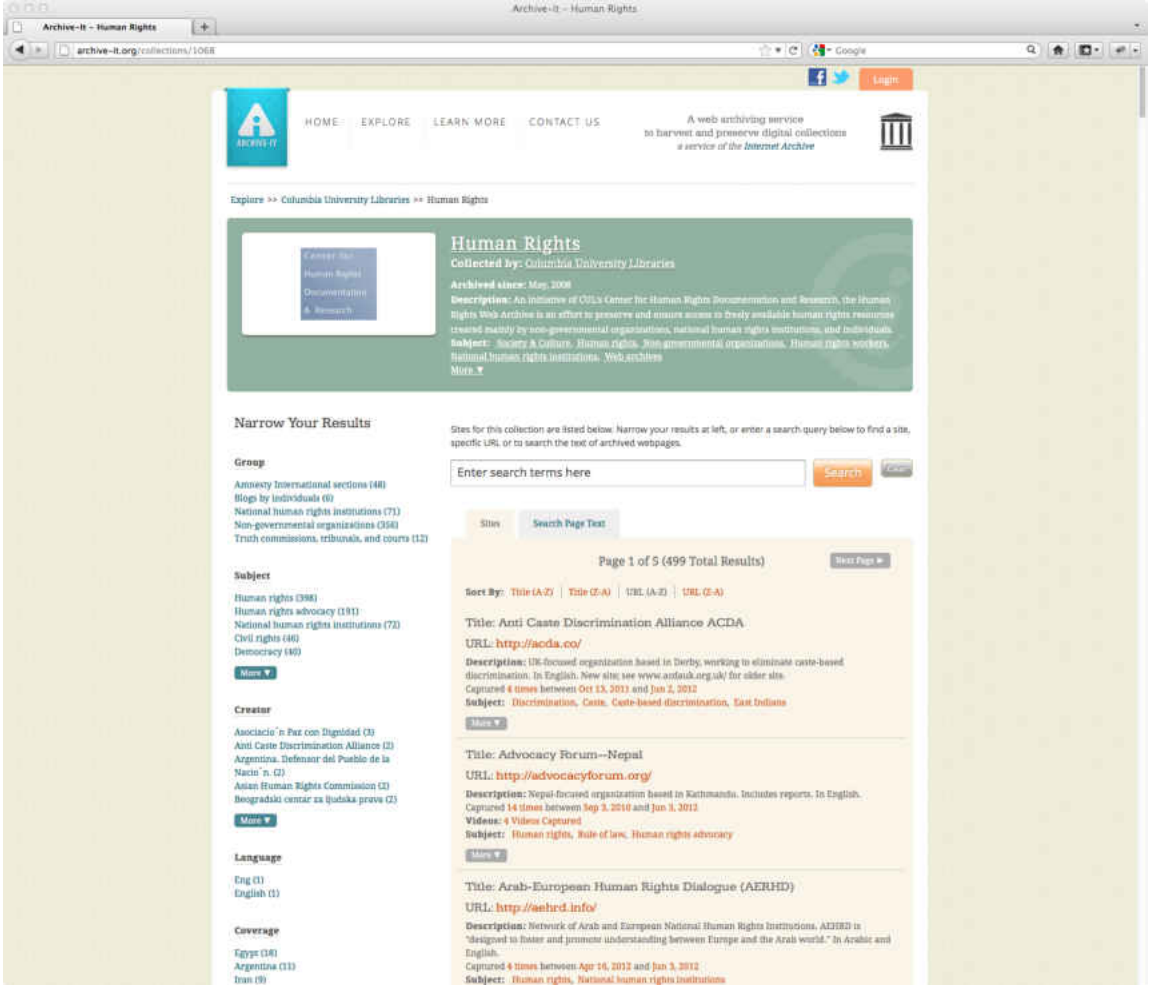
browse the collections when complete. Collections are hosted at the Internet Archive data center and are accessible to the public through the Archive-It website within 24 hours of being archived. There are more than 180 Archive-It partners in the US and 15 countries. Archive-It has captured over 2.7 billion URIs for over 1,700 collections [12] whose subject matter ranges from politics to news to current events to social media to simple institutional archives.

The main difference between Archive-It and other crawler-based archiving services like the Internet Archive is that all pages are user-contributed, resulting in focused collections of web pages. Curators contribute a set of seed URIs to a collection and specify the archiving frequency for each URI. This set of seed URIs identifies the theme of the collection. New URIs, representing content with similar theme, can be later added to expand and enrich the collection.

I.1.2 EXPLORING ARCHIVE-IT COLLECTIONS

In order to explore the collections, Archive-It provides a simple textual interface (Figure 1) with searching capabilities. A user is presented with a list of all URIs in the collection along with the number of times and dates over which each web page was archived. The user can search through the collection either using keywords or by entering a particular URI. Clicking on any URI in the list presents a table (Figure 2) listing dates when archived pages, or mementos [13], were captured. Clicking on any date displays the archived version of the web page in the traditional Internet Archive Wayback Machine [14].

Collection curators may also choose to organize the collection by assigning URIs into groups and subject, specifying coverage, and creating author tags. When a collection is organized in this manner, as with the Human Rights collection (Figure 1), the user can filter the list of URIs in the collection using the options provided under the heading “Narrow Your Results”. If the curator has not defined such groups or tags, then no filtering options are available, as with the Pakistan Floods collection (Figure 3).



A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Columbia University Libraries >> Human Rights

Center for Human Rights Documentation & Research

Human Rights

Collected by: Columbia University Libraries

Archived since: May, 2008

Description: An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.

Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives

More ▼

Narrow Your Results

Group

Amnesty International sections (48)
Blogs by individuals (6)
National human rights institutions (71)
Non-governmental organizations (356)
Truth commissions, tribunals, and courts (12)

Subject

Human rights (398)
Human rights advocacy (191)
National human rights institutions (72)
Civil rights (46)
Democracy (40)

Creator

Asociacio´n Paz con Dignidad (3)
Anti Caste Discrimination Alliance (2)
Argentina. Defensor del Pueblo de la Nacio´n. (2)
Asian Human Rights Commission (2)
Beogradski centar za ljudska prava (2)

Language

Eng (1)
English (1)

Coverage

Egypt (18)
Argentina (11)
Iran (9)

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.

Enter search terms here

Page 1 of 5 (499 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Anti Caste Discrimination Alliance ACDA
URL: http://acda.co/
Description: UK-focused organization based in Derby, working to eliminate caste-based discrimination. In English. New site; see www.acdauk.org.uk/ for older site.
Captured 4 times between Oct 13, 2011 and Jun 2, 2012
Subject: Discrimination, Caste, Caste-based discrimination, East Indians

Title: Advocacy Forum--Nepal
URL: http://advocacyforum.org/
Description: Nepal-focused organization based in Kathmandu. Includes reports. In English.
Captured 14 times between Sep 3, 2010 and Jun 3, 2012
Videos: 4 Videos Captured
Subject: Human rights, Rule of law, Human rights advocacy

Title: Arab-European Human Rights Dialogue (AERHD)
URL: http://aehrd.info/
Description: Network of Arab and European National Human Rights Institutions. AEHRD is "designed to foster and promote understanding between Europe and the Arab world." In Arabic and English.
Captured 4 times between Apr 16, 2012 and Jun 3, 2012
Subject: Human rights, National human rights institutions

Fig. 1. Human Rights collection in Archive-It's interface (from http://archive-it.org/collections/1068).

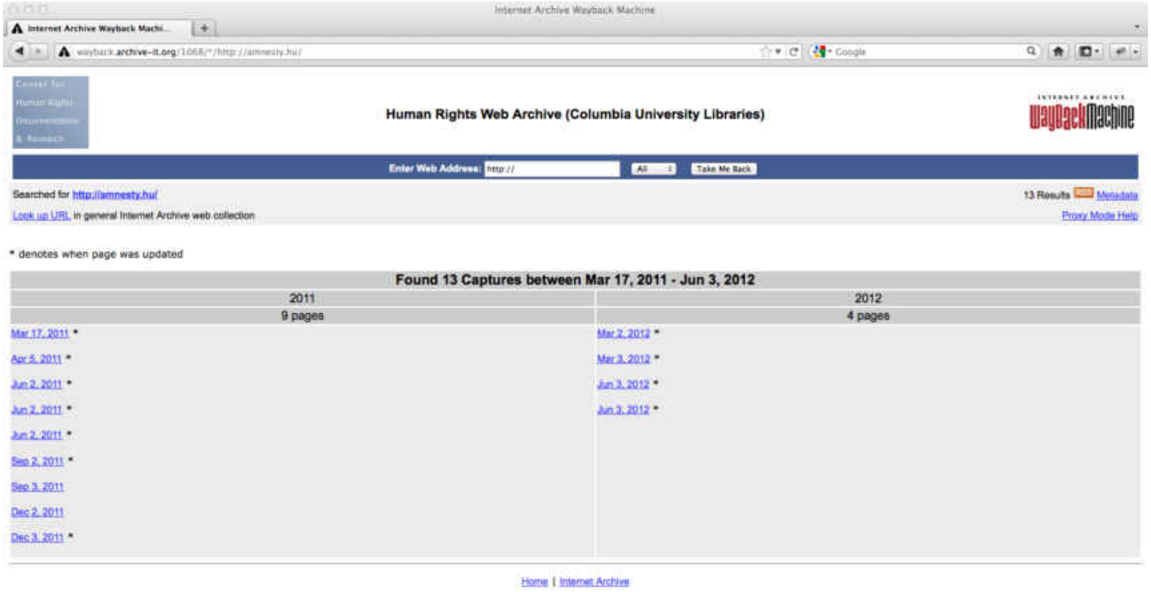
I.1.3 ARCHIVE-IT'S DRAWBACKS

While the current Archive-It interface is simple and easy to use, it has the following shortcomings:

1. Discovering individual web pages in the collection is difficult.

2. The lack of any visual representation of web pages in the collection, like screenshots, makes it difficult for the user to gain quick insight about the contents of the collection without actually visiting one or more web pages in the collection.

3. Archive-It does not mandate tagging and categorization of URIs in a collection


Internet Archive Wayback Machine
Internet Archive Wayback Machi...
wayback.archive-it.org/1068/*/http://amnesty.hu/
Google
Center for Human Rights Documentation & Research
INTERNET ARCHIVE WaybackMachine
Enter Web Address: http://
All
Take Me Back
Metadata
Proxy Mode Help
Home | Internet Archive

Human Rights Web Archive (Columbia University Libraries)

Searched for http://amnesty.hu/

Look up URL in general Internet Archive web collection

13 Results

* denotes when page was updated

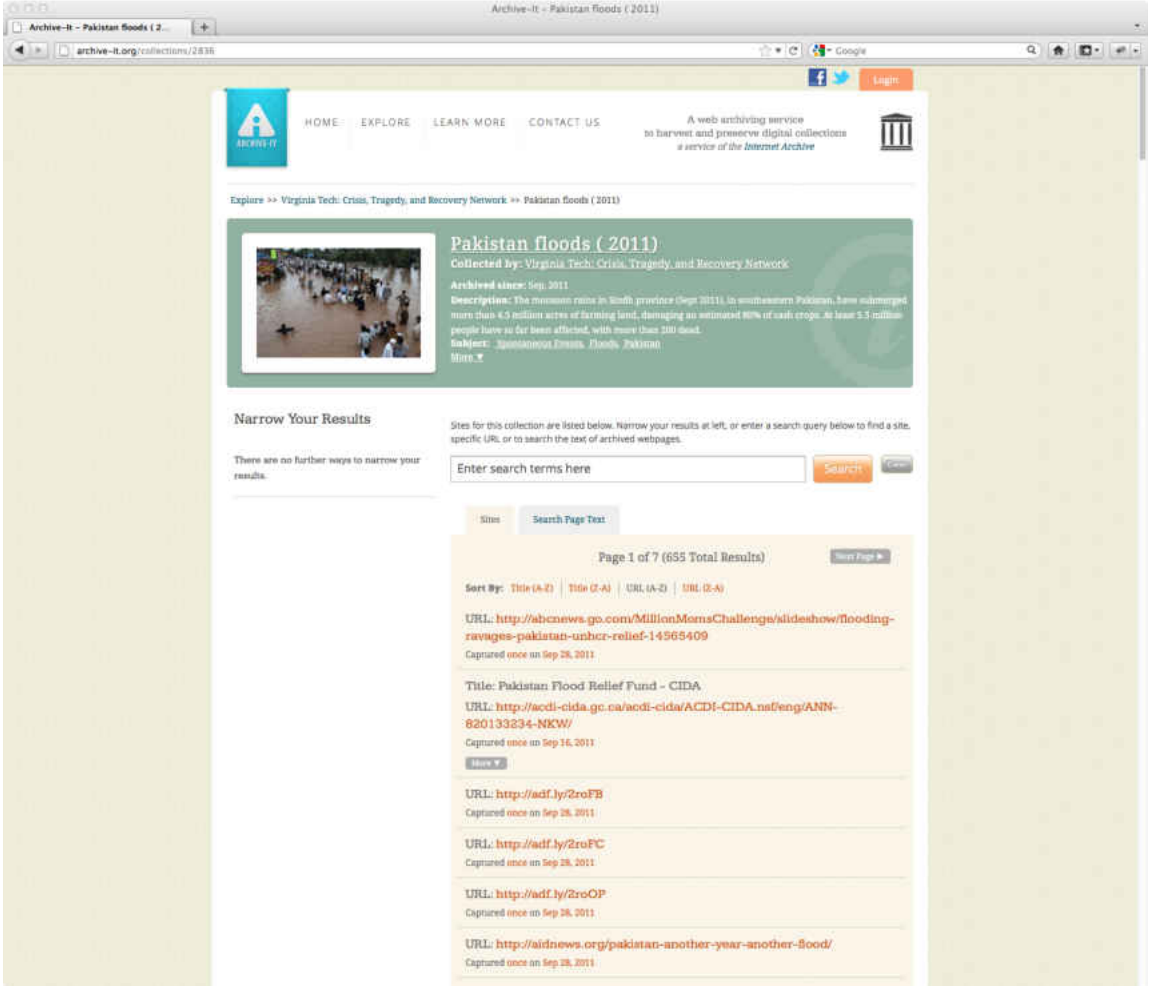
Found 13 Captures between Mar 17, 2011 - Jun 3, 2012	
2011	2012
9 pages	4 pages
Mar 17, 2011 *	Mar 2, 2012 *
Apr 5, 2011 *	Mar 3, 2012 *
Jun 2, 2011 *	Jun 3, 2012 *
Jun 2, 2011 *	Jun 3, 2012 *
Jun 2, 2011 *	
Sep 2, 2011 *	
Sep 3, 2011	
Dec 2, 2011	
Dec 3, 2011 *	

Fig. 2. Wayback Machine interface for a web page's mementos from the Human Rights collection (from http://wayback.archive-it.org/1068/*/http://amnesty.hu/).

and does not provide any means of automatically categorizing contents of a collection. The task of organizing a collection is thus left to collection curators. Lack of organization in a collection increases the effort required by a user to locate a particular web page in the collection.

4. When a curator does not categorize or tag URIs in the collection, the knowledge regarding the collection theme and structure stays with the curator. Users could leverage this knowledge to aid their understanding of a collection and to explore it more efficiently.

5. Each collection grows over time as new web pages are added into the collection and mementos are created. In the current interface, the size and timespan of the collection is not apparent until the user explores each web page in the collection.


Archive-It - Pakistan floods (2011)
Archive-It - Pakistan floods (2...
archive-it.org/collections/2836
Google
Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
Search
Clear
Sites
Search Page Text
Next Page ▶
More ▼

A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)

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 5.3 million people have so far been affected, with more than 200 dead.

Subject: Spontaneous Events, Floods, Pakistan

More ▼

Narrow Your Results

There are no further ways to narrow your results.

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.

Enter search terms here

Page 1 of 7 (655 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409
Captured once on Sep 28, 2011

Title: Pakistan Flood Relief Fund - CIDA
URL: http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/
Captured once on Sep 16, 2011

URL: http://adf.ly/2roFB
Captured once on Sep 28, 2011

URL: http://adf.ly/2roFC
Captured once on Sep 28, 2011

URL: http://adf.ly/2roOP
Captured once on Sep 28, 2011

URL: http://aidnews.org/pakistan-another-year-another-flood/
Captured once on Sep 28, 2011

Fig. 3. Pakistan Floods collection in Archive-It's interface (from http://archive-it.org/collections/2836).

I.2 GOALS

Our main goal is to develop a proof of concept visualization framework for exploring Archive-It collections. The framework should present a collection using multiple interactive visualizations which allow the user to progressively gain insight into the collection. The visualizations should provide an overview of each collection and highlight the collection's underlying characteristics like structure, size, timespan, topic themes and its evolution over time. For those collections that lack a curator-defined

TABLE 1
Collection metrics for Archive-It collections.

Time Span	Small	1 Day - 2 Weeks
	Medium	2 Weeks - 4 Months
	Large	> 4 Months
Groups	Small	1
	Medium	2 - 5
	Large	>5
URI Domains	Small	1 - 10
	Medium	11 - 20
	Large	>20
# of web pages	Small	1 - 10
	Medium	11 - 99
	Large	>99

grouping, we wish to provide a rule-based categorization scheme to make the visualizations more meaningful.

I.3 METHODOLOGY

To present Archive-It collections as interactive visualizations, it was necessary to identify the collection properties, or metrics, which define a collection. Understanding how collections vary along different metrics allows us to develop visualizations which efficiently represent the collection structure and its contents. After careful examination of various collections we identified the following collection metrics:

- Time span: the range of time period over which web pages have been archived
- Groups: the maximum number of categories in each collection
- URI domains: the number of domains (and sub-domains) which contribute to the collection
- Number of web pages: the number of web pages in the collection

The above metrics were further classified as Small, Medium or Large as specified in Table 1. We used this classification to identify collections with different ranges for

TABLE 2
Representative collections used for developing and testing the visualizations.

ID	Collection Name	Time Span	Groups	URI Domains	# of web pages
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	14 Years	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

collection metrics, and used them to develop and test the effectiveness of the various visualizations. These collections are listed in Table 2.

Once the collections were identified, we processed them and created equivalent JavaScript Object Notation (JSON) representations which could be easily used within our visualization framework. Using JavaScript[3] and jQuery[4], we developed six different visualizations for visualizing Archive-It collections. These visualizations are treemap, time cloud, bubble chart, image plot, timeline and wordle. Chapter IV discusses the design and implementation of these visualizations in greater detail.

I.4 CONTRIBUTION

In this work, we implement six different visualizations for exploring Archive-It collections. These are treemap, time cloud, bubble chart, image plot, timeline and

[3] https://developer.mozilla.org/en/JavaScript
[4] http://jquery.com/

wordle. Each of these visualizations leverages one or more collection metrics (Table 1) to create an effective visual representation of the collection, providing better insight into the collection.

The treemap and timeline visualizations use the temporal component of a collection to present collection growth and structural changes over time. We have implemented image plot to provide a novel way of visually browsing through the collection. We present the use of wordle to present a visual summary of collection. The time cloud visualization combines the temporal component of collections with a visual summary to provide an interactive visualization which allows users to explore the change in collection synopsis over time. We also present the use of bubble chart for insight into collection statistics.

Exploring a collection without a curator-defined grouping requires more effort in finding the desired content. We present the use of rule-based categorization for such collections and demonstrate the effectiveness of such categorization in visually exploring a collection.

We also published the results of this work [15] at the Joint Conference on Digital Libraries, 2012.

I.5 THESIS ORGANIZATION

The remainder of this thesis is structured as follows. Chapter II discusses related work done by other researchers in visualizing web archives and the various visualizations in general. In Chapter III we discuss the data preprocessing required to visualize the collections in our system. Chapter IV discusses the design and implementation of our visualization framework. We discuss a few case studies in Chapter V, our future work in Chapter VI and conclude in Chapter VII.

CHAPTER II

RELATED WORK

Similar to the way in which document collections are collections of digital documents forming a corpus, web archives are collections of web pages belonging to a common theme. Thus visualization techniques developed for a document corpus can be easily implemented for a web collection or archive. However, web pages in an archive are usually “captured”, or “archived”, over a period of time. Thus, web archives have both spatial (number of items in a collection) as well as temporal (timespan of archive) components. In this chapter we will first discuss the related work in visualizing document collections, followed by related work in visualizing web archives and other web page collections.

II.1 VISUALIZING DOCUMENT COLLECTIONS

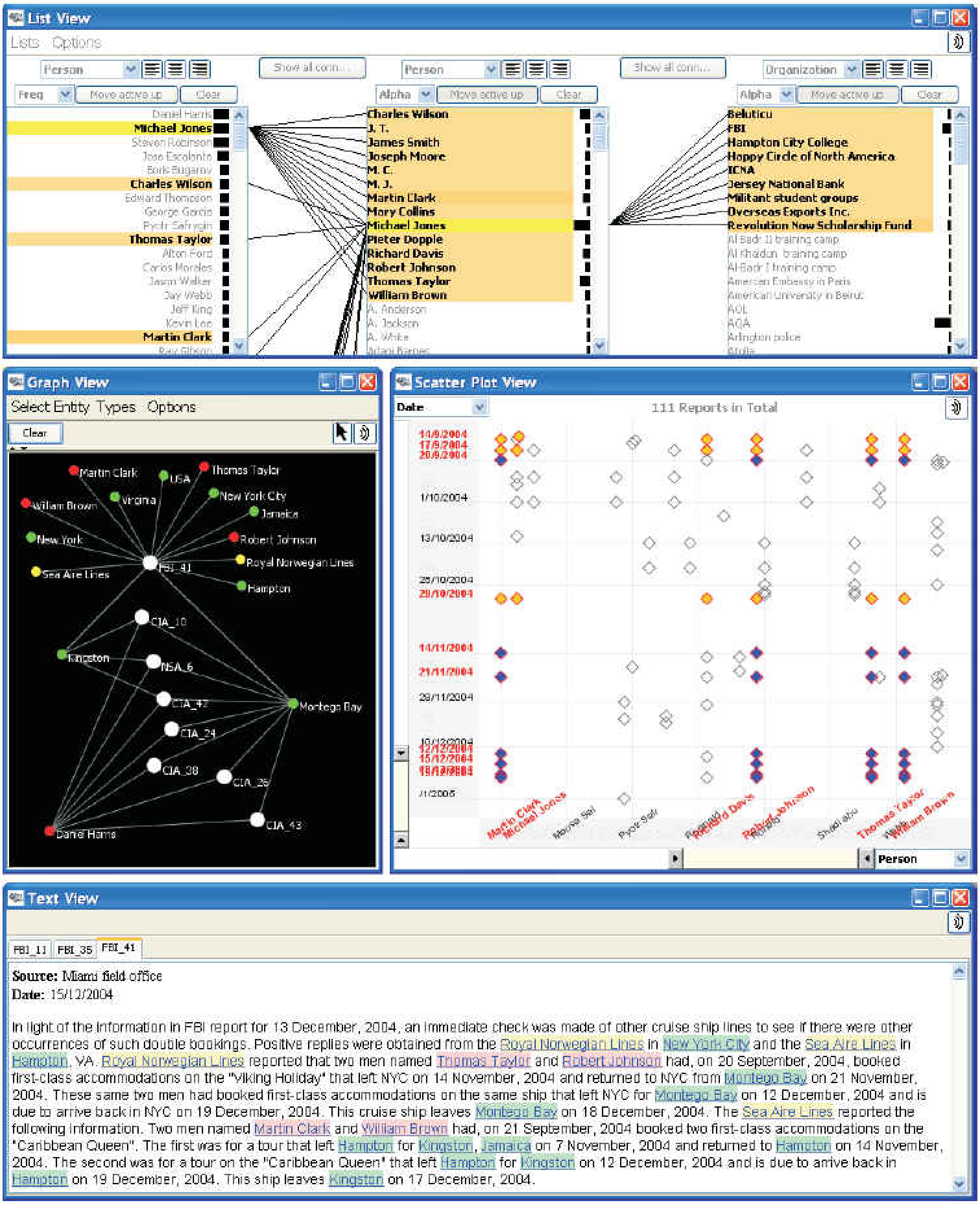
Considerable research has been dedicated towards developing visualizations for viewing and querying documents, and towards graphical querying and browsing of results. Ahlberg and Shneiderman [16] introduced various visualization techniques for visually browsing large datasets with an emphasis on the following:

- rapid filtering - rapid, incremental and reversible changes in query parameters. For example, dragging a slider to change query values.

- progressive refinement - altering parameters to get other results by reformulating the goals of search.

- visual scanning - identifying the results “visually” by creating visual representations.

Most visualization techniques for visualizing large documents are based on these concepts.



Source: Miami field office
Date: 15/12/2004

In light of the information in FBI report for 13 December, 2004, an immediate check was made of other cruise ship lines to see if there were other occurrences of such double bookings. Positive replies were obtained from the Royal Norwegian Lines in New York City and the Sea Aire Lines in Hampton, VA. Royal Norwegian Lines reported that two men named Thomas Taylor and Robert Johnson had, on 20 September, 2004, booked first-class accommodations on the "Viking Holiday" that left NYC on 14 November, 2004 and returned to NYC from Montego Bay on 21 November, 2004. These same two men had booked first-class accommodations on the same ship that left NYC for Montego Bay on 12 December, 2004 and is due to arrive back in NYC on 19 December, 2004. This cruise ship leaves Montego Bay on 18 December, 2004. The Sea Aire Lines reported the following information. Two men named Martin Clark and William Brown had, on 21 September, 2004 booked two first-class accommodations on the "Caribbean Queen". The first was for a tour that left Hampton for Kingston, Jamaica on 7 November, 2004 and returned to Hampton on 14 November, 2004. The second was for a tour on the "Caribbean Queen" that left Hampton for Kingston on 12 December, 2004 and is due to arrive back in Hampton on 19 December, 2004. This ship leaves Kingston on 17 December, 2004.

Fig. 4. Various visual interfaces in Jigsaw, a visualization system [17].

Jigsaw [17] is a system which provides a series of visual interfaces for investigative analysis across collections of text documents (Figure 4). The system provides a List View containing multiple reorderable lists of entities. Related entities are colored the same and linked to each other. It provides a Graph View displaying connections between entities and documents in a node-link diagram, allowing dynamic exploration of the documents by showing or hiding links and nodes. A Scatter Plot View highlights pair-wise relationships between any two entities. A Document View shows the original document and provides information about how many times it has been accessed. Additionally, various documents in a collection can be clustered together in this view. Jigsaw also provides a calendar view for adding temporal context to the documents.

Jigsaw however does not preserve hierarchies in a document collection. Also, Jigsaw supports only text documents and cannot be used to visualize multimedia documents such as web pages containing images and videos. Thus, it is not suited for visualizing web archive collections, such as those at Archive-It.

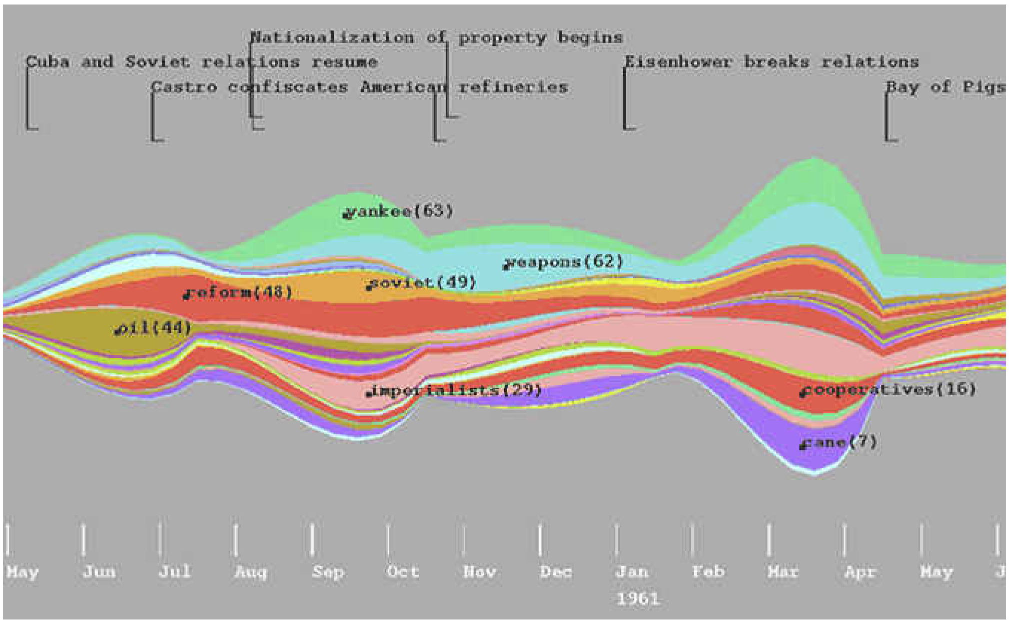


Fig. 5. Use of river metaphor in Themeriver to represent temporal changes [18].

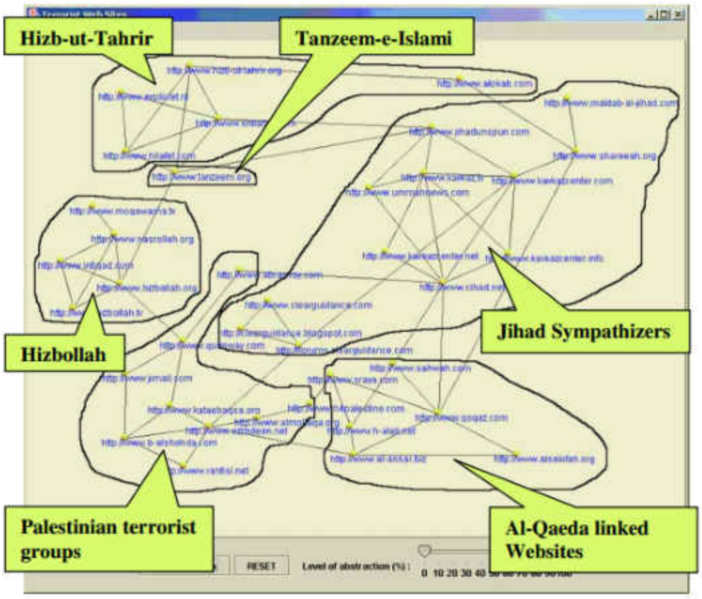


Fig. 6. Visualizing automatically generated hyperlinked communities from a Jihad network [19].

For temporal visualization of large document collections, ThemeRiver [18] provides contextual information through thematic changes within the documents over time (Figure 5) using a river metaphor. TIARA (Text Insight via Automated Responsive Analytics) [20] applies the ThemeRiver metaphor to visually summarize a text collection based on the topic content. It combines text analytics and interactive visualization to help users explore and analyze large collections of text documents. However, visualizing thematic changes alone in a web archive is not enough, and thus TIARA is a poor candidate for visualizing Archive-It collections.

Hearst et al. [21] experimented with the use of hierarchical meta-data and hyperlinked images as results for the purpose of browsing and searching through information on the web. A usability study conducted by the authors suggested that about 50 percent of users used images as a primary means of searching for and browsing information “all the time”. Their finding indicates that users are more inclined towards visual methods of querying and browsing rather than textual methods.

Reid et al. [19] consider collections of web pages on the Internet as document

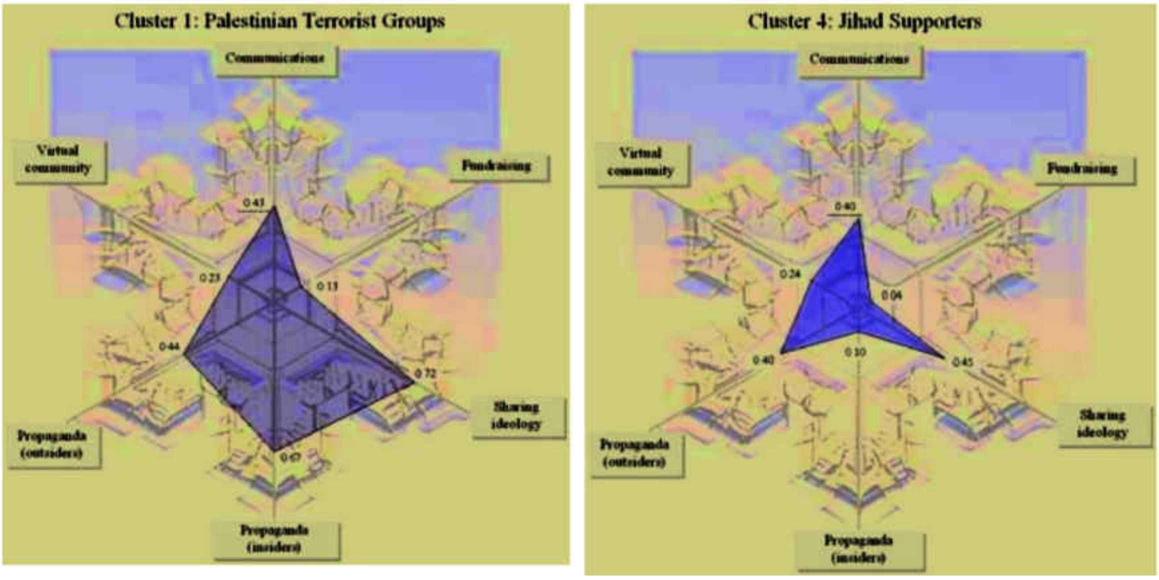


Fig. 7. Palestinian terrorist groups and Jihad supporters web communities visualized as snowflake diagrams [19].

collections. In their system, a group of related web pages forms a web community, much like a document corpus. They investigated the use of multiple visualizations (Figures 6 and 7) for studying and analyzing the content of websites related to Jihad terrorism. They have coined the term “Jihad terrorism Web Infrastructure” for such a collection of websites. Visualizing hyperlinked communities, where web pages from one community in the infrastructure link to those in another, facilitates the analysis of such networks. Most of the connections inside the network are easily revealed by visualizations, and hidden patterns also emerge allowing for refined content analysis. While these visualizations are helpful in analyzing a collection of web pages linking to each other, a web archive often contains individual web pages which do not link to each other. Thus, these visualizations cannot be directly applied to visualize Archive-It collections.

Jatowt et al. [22] propose an interactive visualization system to explore the evolution of web pages and summarize their content over time. They use a temporal tag cloud as a structure for visualizing prevailing and active terms appearing on web

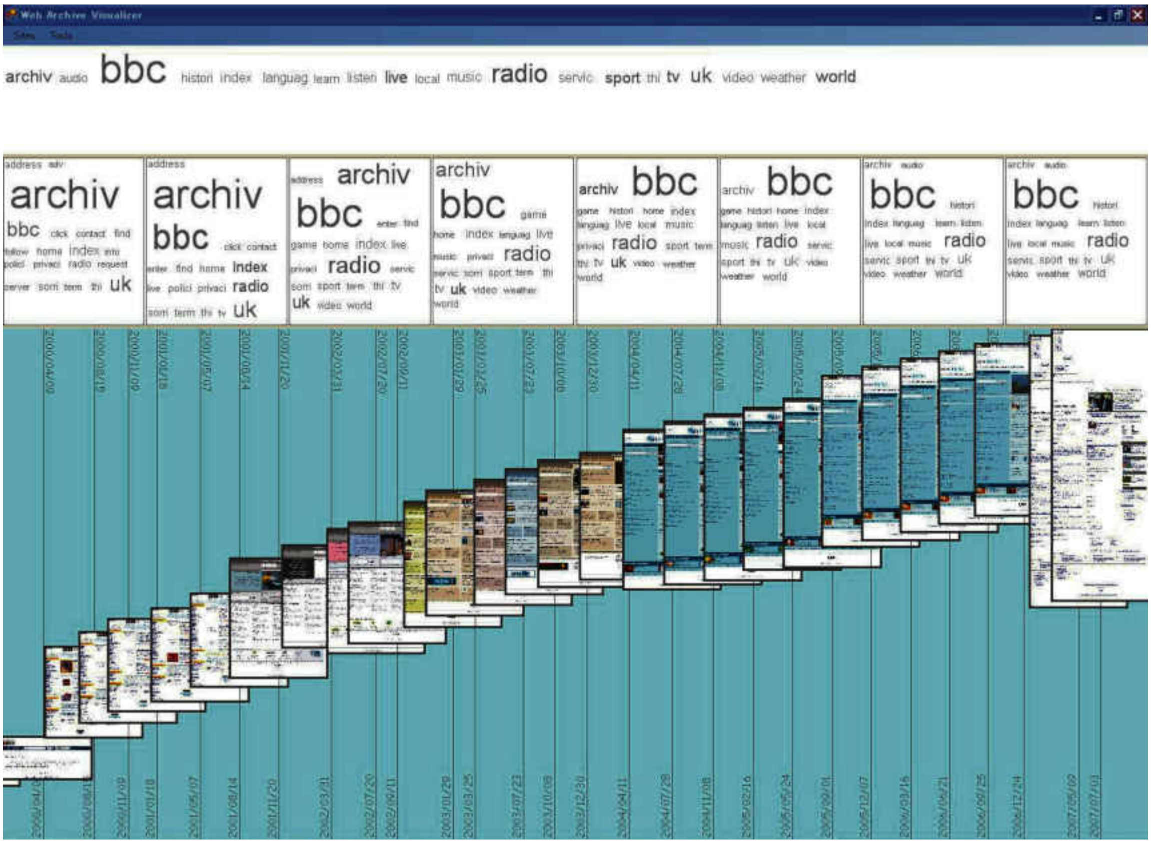

Web Archive Visualizer
Sites Tools
archiv audio bbc histori index languag learn listen live local music radio servic sport thi tv uk video weather world

Fig. 8. The history view of BBC homepage in Page History Explorer [22].

pages. Figure 8 displays the history view of the BBC Homepage[1] in the Page History Explorer. This visualization displays clouds of top 20 terms over the specified time period in the top frame. Additionally tag clouds consisting of up to 20 terms, for smaller time periods are shown below the top frame. Each snapshot on the timeline represents the view of a web page as it existed during that period. This visualization system can be used to visualize how a single web page in a collection changes over time by representing its various mementos over the timeline. However, it can not be used to visualize an entire collection with multiple web pages and thus is not suited for directly visualizing Archive-It collections.

Pivot [23], an experimental application for exploring large data sets with smooth visual interactions, was released by Microsoft Live Labs in 2009. Pivot allows users

[1] http://www.bbc.co.uk/

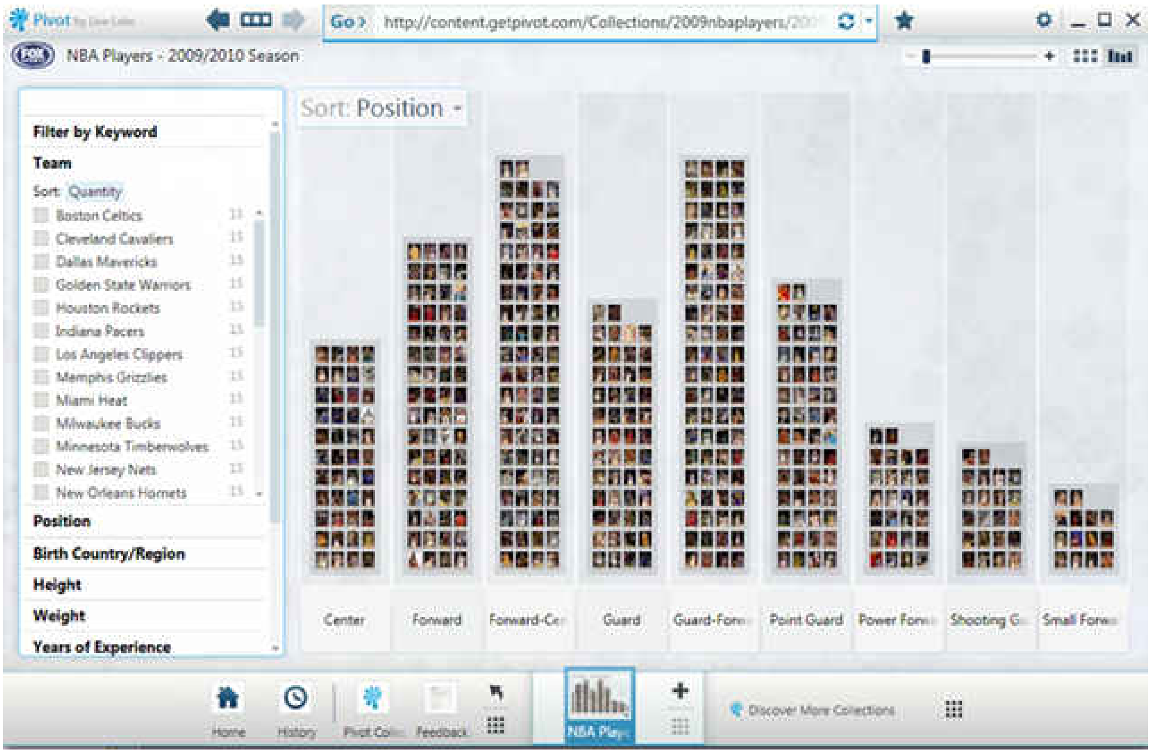
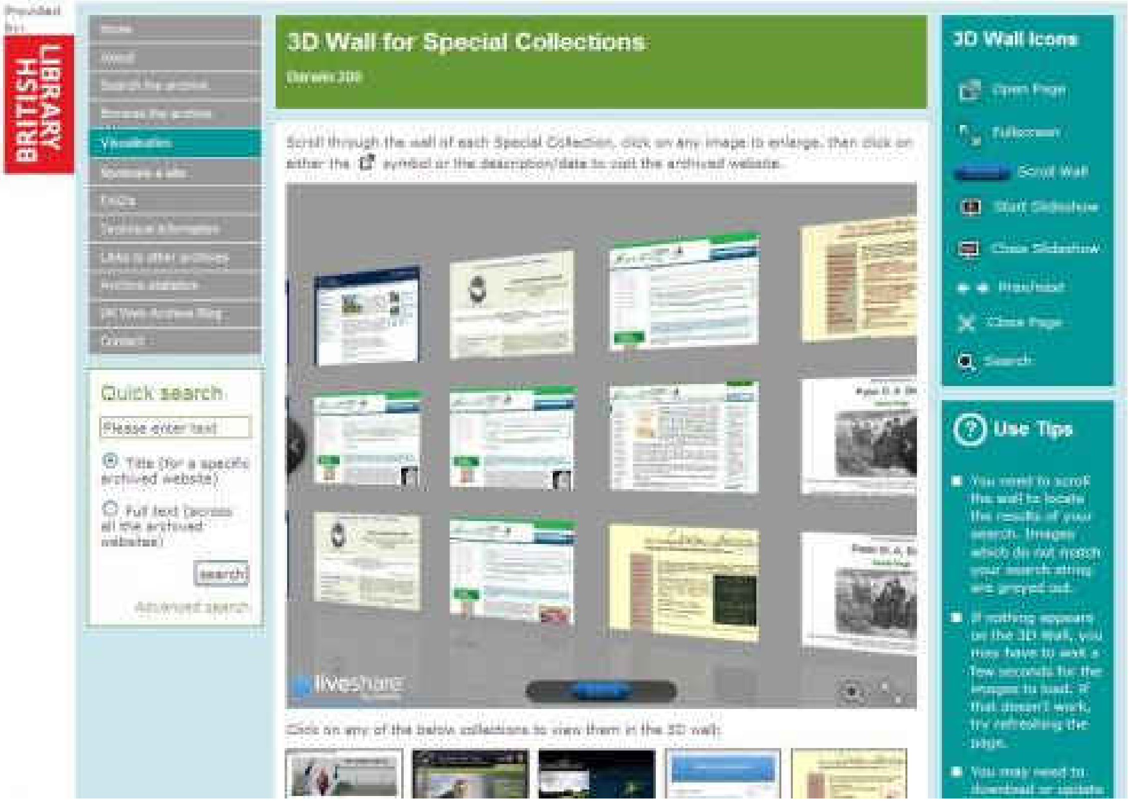


Fig. 9. Microsoft Pivot showing positions of NBA players from 2009/2010 season (from http://www.michaelmcclary.net/image.axd?picture=image_12.png).

to first visualize and then sort, organize and categorize data dynamically, uncovering trends and patterns in a visual format. Pivot can load any form of data and represent it as a deck of cards, with similar cards stacked together. Such representation is similar to a stacked bar chart which uses images to represent each data item in the stack and is called an image plot. Figure 9 shows positions of NBA players from 2009/2010 season visualized as an image plot in Pivot. The toolbar on the left allows users to add additional filters to modify the query used to populate the image plot at runtime. The ability of Pivot to be able to handle multiple data types makes it very convenient for the user to visualize any kind of data as an image plot and interact with it. The image plot representation provided by Pivot can be used to represent the various web pages in an Archive-It collection. However the tool lacks content summarization and temporal tracking features out of the box, and would thus be less suited for visualizing Archive-It collections efficiently.

II.2 VISUALIZING WEB ARCHIVES



● Title (for a specific archived website)

○ Full text (across all the archived websites)

Scroll through the wall of each Special Collection, click on any image to enlarge, then click on either the symbol or the description/date to visit the archived website.

Click on any of the below collections to view them in the 3D wall:

- You need to scroll the wall to locate the results of your search. Images which do not match your search string are greyed out.
- If nothing appears on the 3D Wall, you may have to wait a few seconds for the images to load. If that doesn't work, try refreshing the page.
- You may need to download or update

Fig. 10. 3-D wall visualization for collections in U.K. Web Archive (from http://www.webarchive.org.uk/ukwa).

This section introduces related work in visualizing web archives. Since each visualization system targets a specific web archive, it cannot be directly used to visualize Archive-It collections.

Many web archives such as Archive-It, California Digital Library[2], Library of Congress[3], and Pandora - Australia's Web Archive[4] provide a textual interface for interacting with the archived collections. Though such interface facilitates easier searching, it is difficult for users to gain insight about the whole collection using the textual interface. Several attempts have been made recently in order to provide

[2] http://www.cdlib.org

[3] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[4] http://pandora.nla.gov.au/

an alternative visual interface for other web archives. The UK Web Archive[5] provides a 3D wall visualization (Figure 10) for selected collections, allowing interaction through zooming. However exploring a large collection using this visualization can be cumbersome due to repeated scrolling and potential loss of context when user navigates from one page to another.

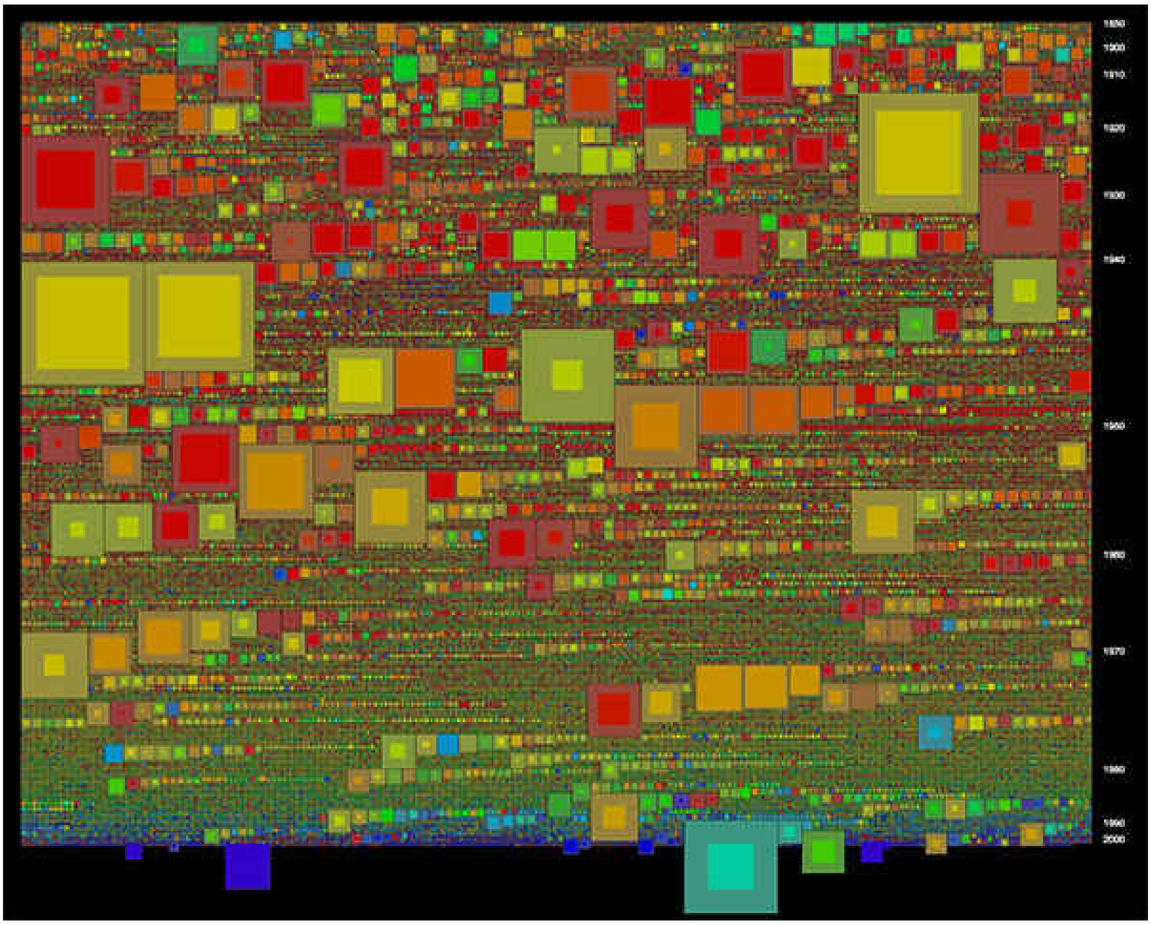
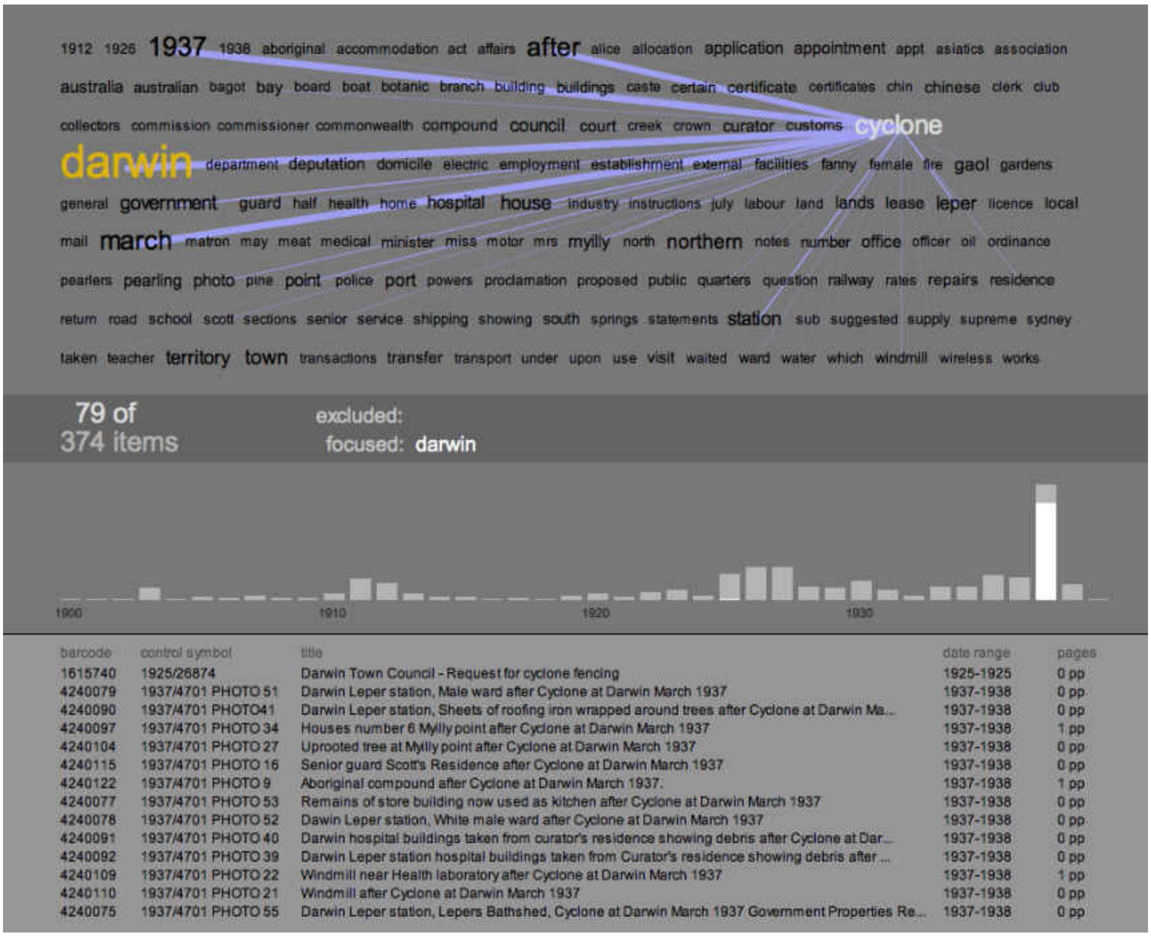


Fig. 11. Series Browser showing treemap of all collections by an agency in the National Archives of Australia [24].

Whitelaw [24] has discussed multiple visualizations for exploring collections from the National Archives of Australia[6]. Figure 11 shows the "Series Browser" visualization which represents the collection in form of a treemap. Figure 12 shows the "A1 Explorer" interface which displays a tag cloud of top terms in a collection. A

[5] http://www.webarchive.org.uk/ukwa
[6] http://www.naa.gov.au/


1912 1926 1937 1938 aboriginal accommodation act affairs after alice allocation application appointment appt asiatics association
australia australian bagot bay board boat botanic branch building buildings caste certain certificate certificates chin chinese clerk club
collectors commission commissioner commonwealth compound council court creek crown curator customs cyclone
darwin department deputation domicile electric employment establishment external facilities fanny female fire gaol gardens
general government guard half health home hospital house industry instructions july labour land lands lease leper licence local
mail march matron may meat medical minister miss motor mrs myilly north northern notes number office officer oil ordinance
pearlers pearling photo pine point police port powers proclamation proposed public quarters question railway rates repairs residence
return road school scott sections senior service shipping showing south springs statements station sub suggested supply supreme sydney
taken teacher territory town transactions transfer transport under upon use visit waited ward water which windmill wireless works
1900
1910
1920
1930

79 of
374 items

excluded:
focused: darwin

barcode	control symbol	title	date range	pages
1615740	1925/26874	Darwin Town Council - Request for cyclone fencing	1925-1925	0 pp
4240079	1937/4701 PHOTO 51	Darwin Leper station, Male ward after Cyclone at Darwin March 1937	1937-1938	0 pp
4240090	1937/4701 PHOTO41	Darwin Leper station, Sheets of roofing iron wrapped around trees after Cyclone at Darwin Ma...	1937-1938	0 pp
4240097	1937/4701 PHOTO 34	Houses number 6 Myilly point after Cyclone at Darwin March 1937	1937-1938	1 pp
4240104	1937/4701 PHOTO 27	Uprooted tree at Myilly point after Cyclone at Darwin March 1937	1937-1938	0 pp
4240115	1937/4701 PHOTO 16	Senior guard Scott's Residence after Cyclone at Darwin March 1937	1937-1938	0 pp
4240122	1937/4701 PHOTO 9	Aboriginal compound after Cyclone at Darwin March 1937.	1937-1938	1 pp
4240077	1937/4701 PHOTO 53	Remains of store building now used as kitchen after Cyclone at Darwin March 1937	1937-1938	0 pp
4240078	1937/4701 PHOTO 52	Dawin Leper station, White male ward after Cyclone at Darwin March 1937	1937-1938	0 pp
4240091	1937/4701 PHOTO 40	Darwin hospital buildings taken from curator's residence showing debris after Cyclone at Dar...	1937-1938	0 pp
4240092	1937/4701 PHOTO 39	Darwin Leper station hospital buildings taken from Curator's residence showing debris after ...	1937-1938	0 pp
4240109	1937/4701 PHOTO 22	Windmill near Health laboratory after Cyclone at Darwin March 1937	1937-1938	1 pp
4240110	1937/4701 PHOTO 21	Windmill after Cyclone at Darwin March 1937	1937-1938	0 pp
4240075	1937/4701 PHOTO 55	Darwin Leper station, Lepers Bathshed, Cyclone at Darwin March 1937 Government Properties Re...	1937-1938	0 pp

Fig. 12. A1 Explorer combining tag cloud and histogram to reveal patterns in the dataset [24].

histogram below the tag cloud shows the number of items with start dates in each year that are related to the selected keywords in the tag cloud. Users can also select and request a visual record of any of the museum items in the collection. These visualizations provide a sense of large-scale collection structure and reveal patterns and relationships in the collection.

Scharnhorst et al. [25] discuss the visualization of a hierarchical digital archive consisting of documents related to research data. The archive exposes collection meta-data (subject or classification, depositors, date and access rights), and their visualization system EASY visualizes the archive contents along these dimensions. Since access rights for a document are hierarchical in nature, they can be efficiently

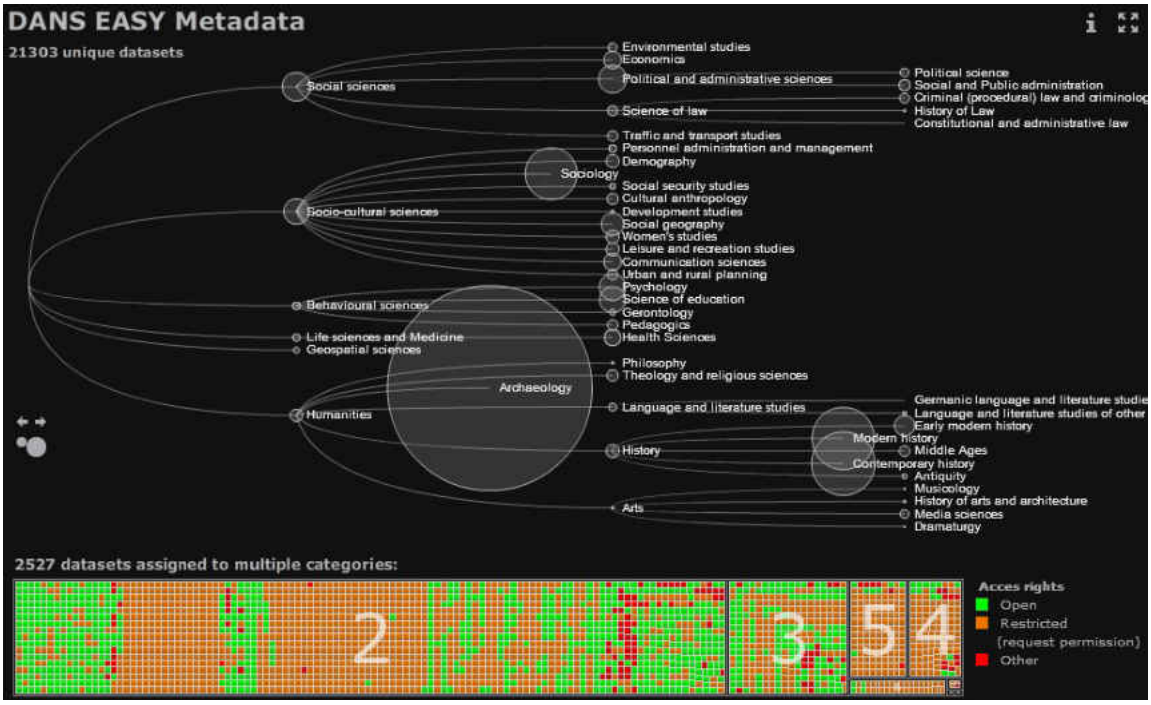


Fig. 13. Exploring meta-data in EASY using spacetrees and treemaps [25].

depicted using treemaps. Figure 13 shows the use of treemaps and space trees to visually explore access rights in an archive. Figure 14 shows the use of bubble charts to visualize subjects' classification where the size of each bubble is relative to number of web pages classified under that tag and the nesting of bubbles depicts the classification hierarchy. Alternatively, the authors also provide a spacetree (Figure 15) for visualizing the classification hierarchy.

Similar to their digital archive, Archive-It also structures all collections in a hierarchy (Figure 16) and associates meta-data information with each collection. Like EASY, treemaps can be implemented for Archive-It collections to represent collection hierarchy. Also, bubble charts can be implemented for providing a statistical view of the collection.

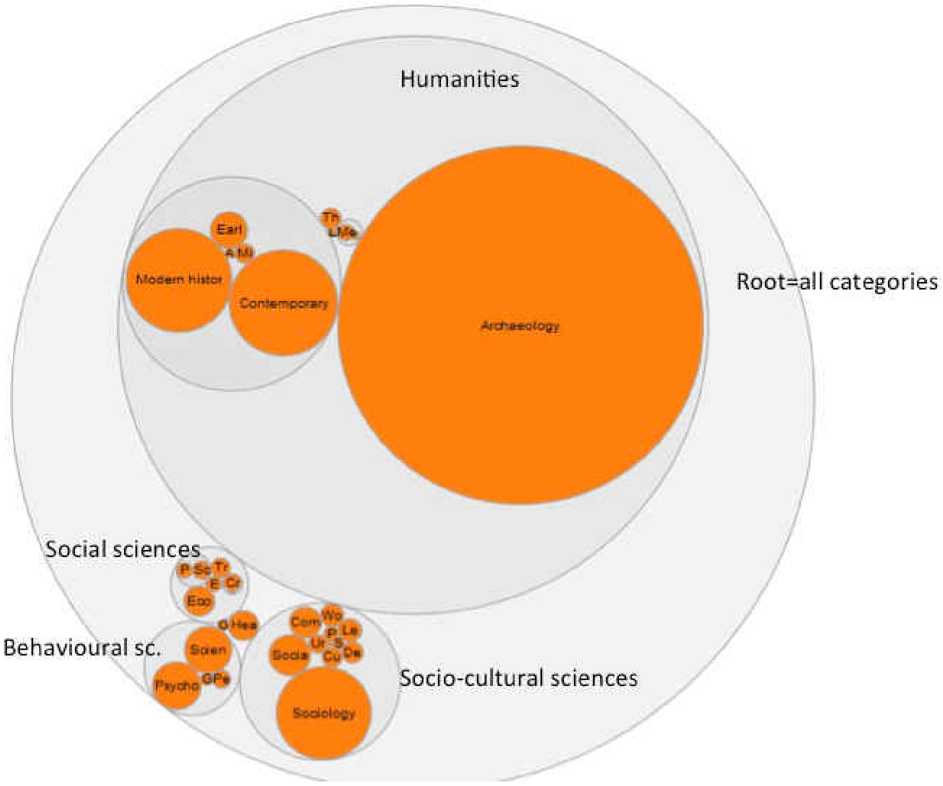


Fig. 14. Exploring subjects' classification in EASY using bubble chart [25].

II.3 VISUALIZATIONS

In order to visualize the collections at Archive-It, we selected several visualizations: treemap, tag cloud, wordle, bubble chart, histogram and timeline, and modified them to fit our needs. Each visualization represents either spatial, temporal or both components in a manner such that the overall structure and organization of the collection is effectively conveyed to the user. We will now provide a general discussion of each type of visualization here. In Chapter IV we will discuss their implementation for visualizing Archive-It collections.

II.3.1 TREEMAP

Treemaps [26] enable viewing of large sets of hierarchical data by organizing it as a set of nested rectangles. Each level of the tree is represented using a rectangle which is further tiled with smaller rectangles to represent the next level in the hierarchy. The

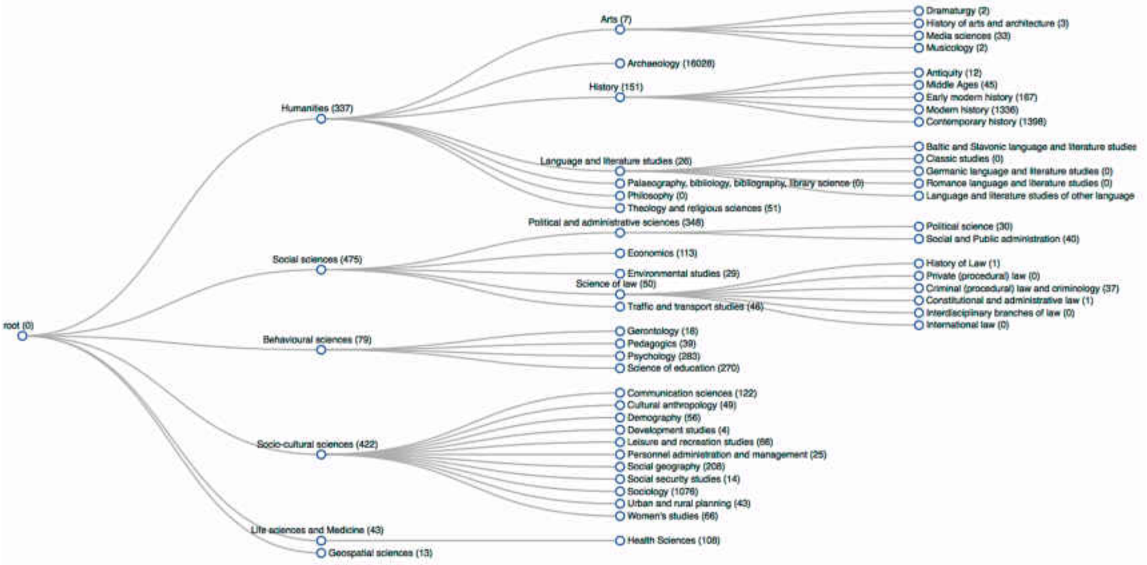

Arts (7)
Dramaturgy (2)
History of arts and architecture (3)
Media sciences (33)
Musicology (2)
Archaeology (16028)
History (151)
Antiquity (12)
Middle Ages (45)
Early modern history (167)
Modern history (1336)
Contemporary history (1398)
Humanities (337)
Baltic and Slavonic language and literature studies
Classic studies (0)
Germanic language and literature studies (0)
Romance language and literature studies (0)
Language and literature studies of other language
Language and literature studies (26)
Palaeography, bibiology, bibliography, library science (0)
Philosophy (0)
Theology and religious sciences (51)
Political and administrative sciences (348)
Political science (30)
Social and Public administration (40)
Social sciences (475)
Economics (113)
Environmental studies (29)
Science of law (50)
Traffic and transport studies (46)
History of Law (1)
Private (procedural) law (0)
Criminal (procedural) law and criminology (37)
Constitutional and administrative law (1)
Interdisciplinary branches of law (0)
International law (0)
root (0)
Behavioural sciences (79)
Gerontology (18)
Pedagogics (39)
Psychology (283)
Science of education (270)
Communication sciences (122)
Cultural anthropology (49)
Demography (56)
Development studies (4)
Leisure and recreation studies (66)
Socio-cultural sciences (422)
Personnel administration and management (25)
Social geography (208)
Social security studies (14)
Sociology (1076)
Urban and rural planning (43)
Women's studies (66)
Life sciences and Medicine (43)
Health Sciences (108)
Geospatial sciences (13)

Fig. 15. Exploring subjects' classification in EASY using spacetree [25].

area of the rectangle is directly proportional to the amount of data represented by the node. Treemaps make efficient use of space and can be used to display huge amounts of data and hierarchical relationships together simultaneously and efficiently.

Figure 17 shows an example of a treemap which displays the daily status of the iTunes 100 most popular songs, grouped by genre (rock, pop, hip-hop, etc.). The highest ranked songs are larger, and color-coding shows whether a song has moved up or down in the past days.

II.3.2 TAG CLOUD AND WORDLE

A tag cloud (also term cloud or word cloud) is a special visualization in which an overview-style visualization of the text can be obtained by emphasizing the importance of the words. The frequency of appearance of a word in a document corpus determines its importance, which is represented by varying the size and the color. Tag clouds are commonly used to visualize user-generated tags for social tagging in Web 2.0 services such as Flickr[7] (Figure 18) and Delicious[8].

[7] http://flickr.com/

[8] http://delicious.com/

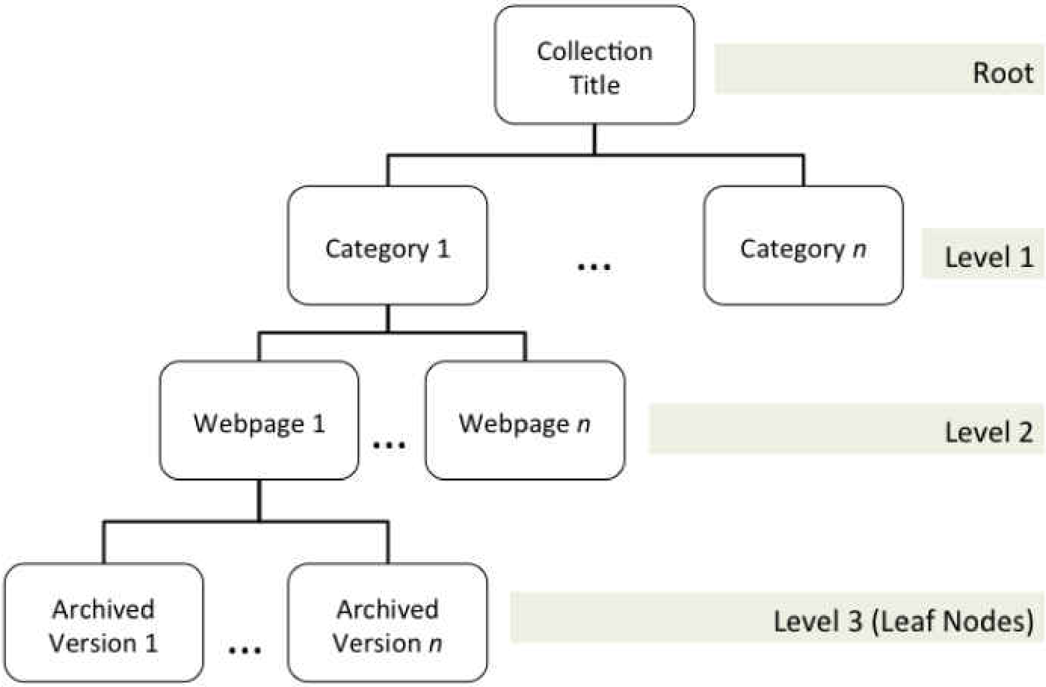
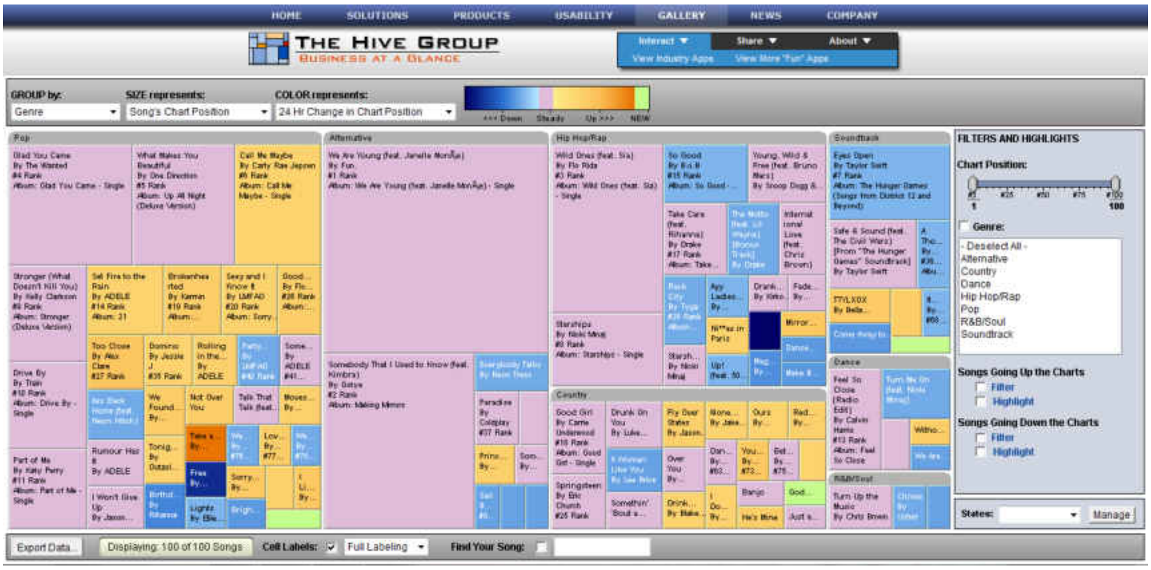


Fig. 16. Tree based layout of Archive-It collections.

Various research efforts that attempt to understand the effectiveness and utility of tag clouds in describing contents of web pages or blogs have been conducted [27, 28]. A systematic analysis of what people say when they write about tag clouds on the web was piloted by Hearst et al. [28]. The authors studied the reception by new users and found that there was disagreement about the emotional or aesthetic appeal of tag clouds. However, tag clouds continue to be used heavily on the web for creating visual summaries.

Wordles (Figure 19) were first created in 2008 by Jonathan Feinberg as a part of a social bookmarking application Dogear [29]. Wordle is close to tag cloud in encoding word frequency information via font size. However, wordle differs from tag cloud in many ways. While wordle allows the words to be in different orientations, tag clouds provide only horizontal orientation without any strict alignment. Furthermore, by packing words tightly, a wordle can make better use of available screen space and is very useful in situations where space is a constraint. Conversely, tag clouds waste an



GROUP by: Genre

SIZE represents: Song's Chart Position

COLOR represents: 24 Hr Change in Chart Position

Cell Labels: ☑ Full Labeling

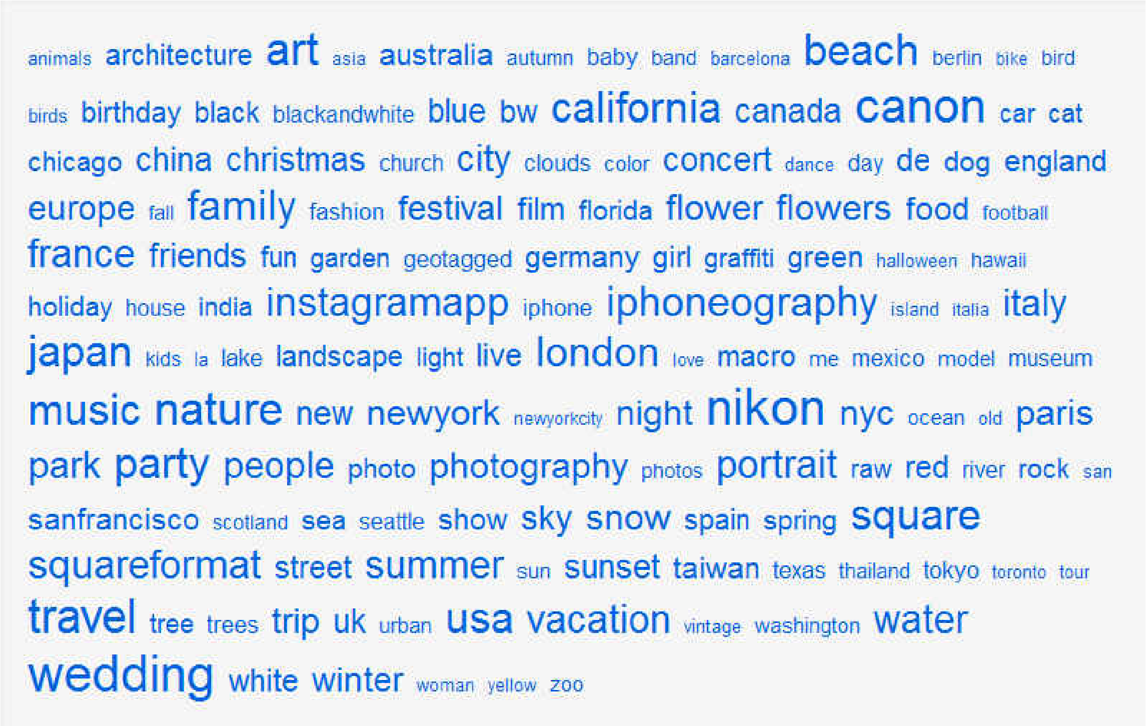
Fig. 17. Treemap by Hive Group displaying iTunes Top 100 for March 23, 2012, grouped by genre, sized by chart position, and colored by 24-hour change in chart position (from http://www.hivegroup.com/gallery/itunes.html).

enormous amount of whitespace around small words, because the size of the line is specified based on the largest word in the line.

Wordle has been greatly appreciated by responders of the surveys conducted by Viégas et al. [30]. The survey showed that the wordle layout is very compelling through the people's reaction. The huge popularity of wordle showed the prominence of aesthetics and expressiveness in attracting people to use visualizations. However, the efficacy of wordle and tag clouds is disputed, often blamed for providing a pretty picture but poor insight. For our work, we chose to use tag clouds and wordles for the ease of representing visual summary of web pages.

II.3.3 BUBBLE CHART

A bubble chart is a chart where each data point in a graph is represented in the form of a bubble. Depending on the type of entities being depicted, each bubble in a bubble chart can depict up to three distinct parameters along which they can be compared. It is not necessary for all parameters being depicted to be of numeric type. Figure 20 shows a bubble chart of the Chicago Deposit Market generated



animals architecture art asia australia autumn baby band barcelona beach berlin bike bird
birds birthday black blackandwhite blue bw california canada canon car cat
chicago china christmas church city clouds color concert dance day de dog england
europe fall family fashion festival film florida flower flowers food football
france friends fun garden geotagged germany girl graffiti green halloween hawaii
holiday house india instagramapp iphone iphoneography island italia italy
japan kids la lake landscape light live london love macro me mexico model museum
music nature new newyork newyorkcity night nikon nyc ocean old paris
park party people photo photography photos portrait raw red river rock san
sanfrancisco scotland sea seattle show sky snow spain spring square
squareformat street summer sun sunset taiwan texas thailand tokyo toronto tour
travel tree trees trip uk urban usa vacation vintage washington water
wedding white winter woman yellow zoo

Fig. 18. All-time most popular tags from Flickr as of March 23, 2012 (from http://www.flickr.com/photos/tags/).

by the FDIC where the market's concentration of deposits is compared against the concentration of companies in a region using a bubble. At the same time, the area of each bubble represents the volume of insured deposits. Bubble charts are thus able to convey more than one insight about data to the user in a very simple fashion.

II.3.4 HISTOGRAMS AND BAR CHART

A histogram is traditionally constructed for data in the form of key-value pairs where each key is listed on one of the axes (usually X-axis) and the value is plotted on the other axis (usually Y-axis). Placement of keys along an axis creates intervals termed "bins" on the axis. The value of each bin (represented on the other axis) is same as that for corresponding key and is some statistical parameter (total, max, min, average, etc.). A rectangle in each bin has height equal to the value for that bin and width equal to that of the bin. Figure 21 shows a histogram of travel time

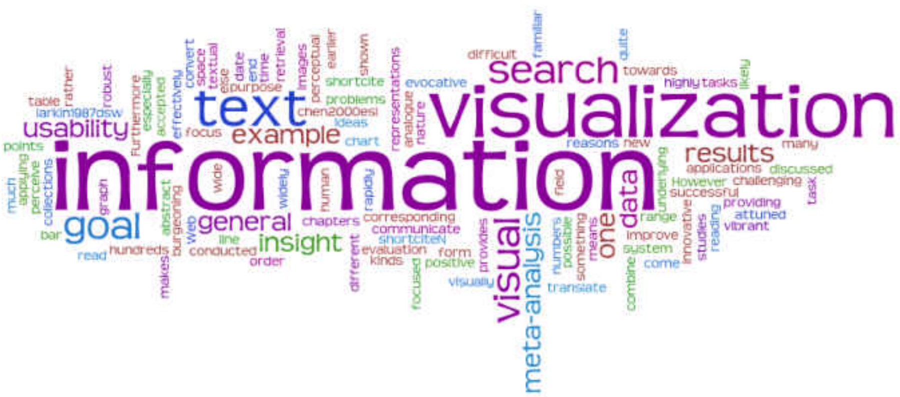


Fig. 19. Wordle created using text from the book Search User Interfaces (from http://searchuserinterfaces.com/book/sui_ch11_text_analysis_visualization.html).

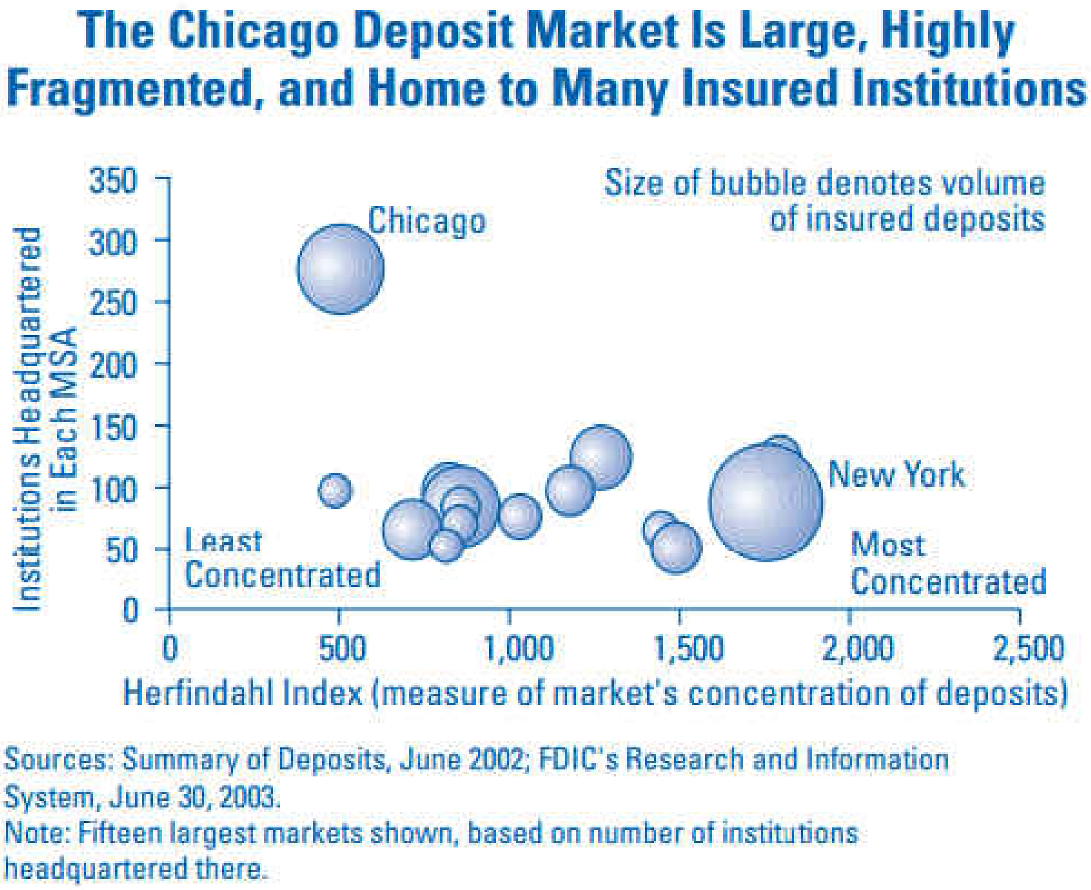
in minutes as per the US 2000 Census.

A bar chart, or bar graph, is a special histogram where the value for each key to be plotted is a discrete value or a number representing the key. Figure 22 shows a bar graph of worldwide incarceration rates of prisoners per 100,000 of the population of the country, from selected countries as per the 2006 statistics on World Prison Population List.

Stacked bar charts provide an alternate way of representing multidimensional data by stacking one bar over another. Stacked bar charts also help in visualizing relationship of parts to the whole.

II.3.5 TIMELINE

A timeline graphically depicts events that have occurred over a period of time, presented in chronological order. It is typically a graphic design showing a long bar labeled with dates alongside itself and (usually) events labeled on points where they would have happened. There is no specific domain of data to which timelines are restricted, and they can be used to represent any kind of data movement over time. Furthermore, timelines can use any time scale, depending on the subject and the data, though most use a linear scale, where a unit of distance is equal to a set



Sources: Summary of Deposits, June 2002; FDIC's Research and Information System, June 30, 2003.
Note: Fifteen largest markets shown, based on number of institutions headquartered there.

Fig. 20. Bubble chart of the Chicago Deposit Market [31].

amount of time. Figure 23 depicts a timeline of events in Benjamin Frankin's life. In this figure the timeline is represented on the X-axis, with the events represented on either side of the timeline along the Y-axis.

II.4 TF AND TF-IDF

In our visualizations we use the TF and TF-IDF metrics to generate word clouds and wordles. Here we give an introduction to these metrics.

Term frequency (TF) refers to how often a word appears within a document. Term frequency in a document is typically evaluated as the frequency of a term in a given document normalized by the frequency of the most prevalent term in the document. The assumption is that the more a term appears in a document, the

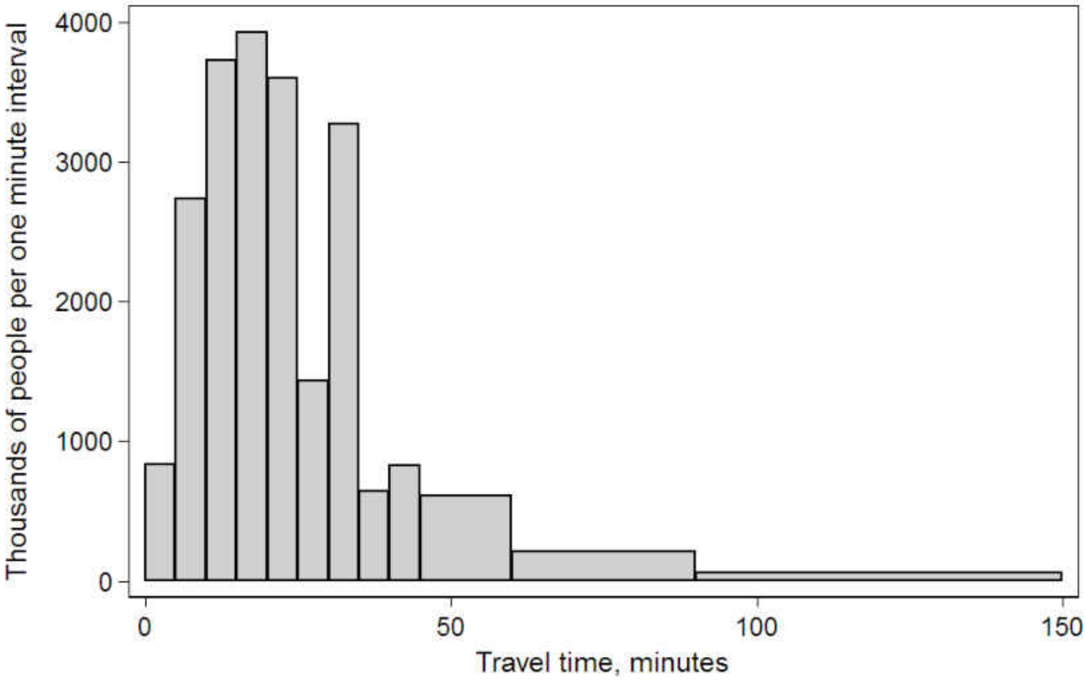


Fig. 21. Histogram of travel time, US 2000 Census (from http://en.wikipedia.org/wiki/Histogram).

greater the connection between the term and the document. If a one-term query is issued against two documents, where one document contains two instances of the term and the second document contains 10 instances of it, then according to term frequency, if this term is the most frequent in both, then the second document would be more relevant.

Thus TF faces two issues. First, if a term appears with high frequency in most documents within a collection, then they will likely all be returned, which would reduce the precision of the result set. Second, it fails to determine the importance of one term over another.

To solve the above two issues, Salton and Buckley introduced the concept of Inverse Document Frequency (IDF) [32]. The IDF of a term is evaluated at a collection level by taking the log of the total number of documents (N) over the number of documents in which the term appeared (n), such that a term's IDF varies inversely with

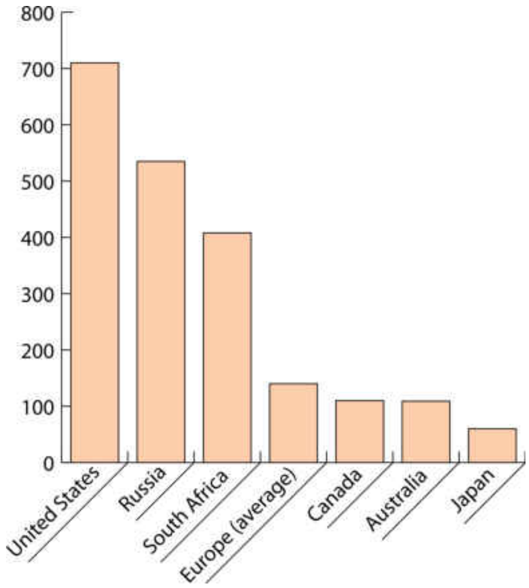


Fig. 22. Bar graph of worldwide incarceration rates (from http://www.nicic.org/Library/020631).

N/n. Thus,

$$IDF(t, N) = log(\frac{|N|}{|n|}).$$

When TF and IDF are combined together, both the frequency of a term in a document along with its relative uniqueness in the collection can be determined [33]. To rank documents with the greatest number of matches to a given term, both the query and documents are used, taking into account the uniqueness of terms in the scope of the collection, as well as each terms' frequency in a given document. Since IDF factors in the significance of the term in determination of ranking, multiple instances of a very unique term (to the collection) in a document would increase the document's rank in the collection. Thus, using TF-IDF, a document containing a very unique term would be ranked higher than another document which contains several terms of lesser significance.

II.5 SUMMARY

This chapter discussed some of the existing work in visualizing document collections and web archives. The chapter also discussed the visualization techniques

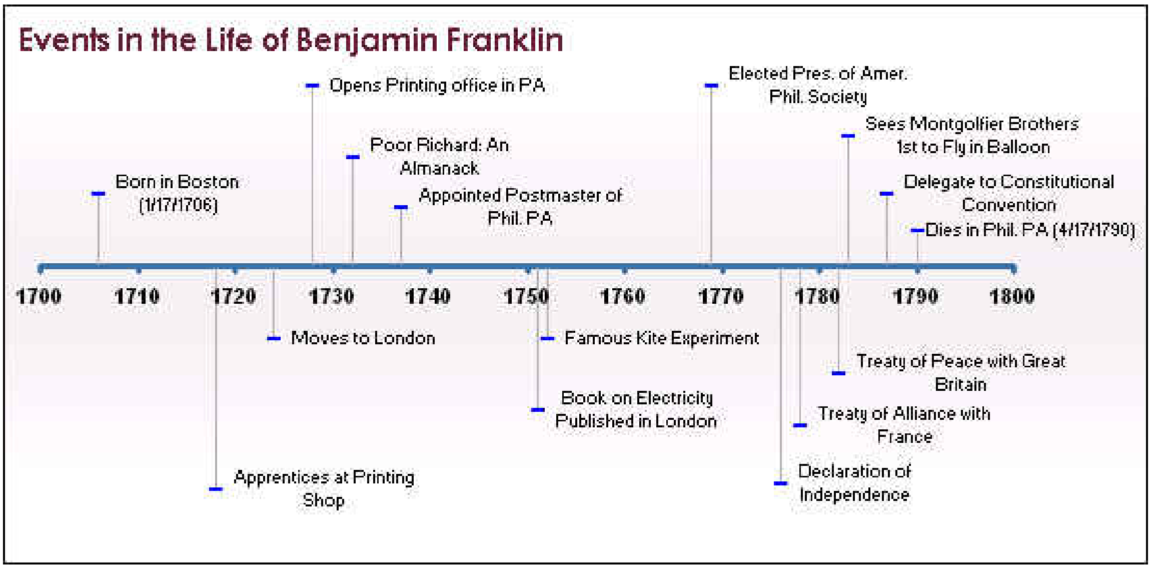


Fig. 23. Timeline of events in the life of Benjamin Franklin (from http://www.vertex42.com/ExcelArticles/Images/timeline/Timeline-for-Benjamin-Franklin.gif).

implemented by other web archives or digital collections which can be applied to Archive-It collections.

CHAPTER III

DATA PREPROCESSING

Archive-It stores all of its digital content at Internet Archive's data centers [12], where it is stored in the WARC (Web ARChive) file format [34], a revision of the Internet Archive's ARC file format that has traditionally been used to store "web crawls" as sequences of content blocks harvested from the World Wide Web. These WARC files are created with the help of Heritrix [35, 36], an open source web crawler developed by Internet Archive specially designed for web archiving.

As Heritrix crawls the web, it stores the web resources in a WARC file. A WARC file stores multiple archived resources in a single file in order to avoid managing a large number of small files. The size of a WARC file ranges between 100 to 600 MB. The file consists of a sequence of URL records, each with a header containing meta-data about how the resource was requested followed by the HTTP header and the response. Each collection at Archive-It is stored as multiple WARC files. Each WARC file exposes meta-data about the files contained within, into separate meta-data files in the WAT (Web Archive Transformation) file format [37]. This meta-data is used to populate the textual interface for viewing collections at Archive-It. When a user requests a certain archived version of a web page, it is then served via Internet Archive's WayBack Machine.

In order to create visualizations for Archive-It collections, we needed both content and meta-data for each web page in the collection as well as its screenshot. The required meta-data could be obtained from WAT files, however, the screenshots would have to be captured separately. Also, since Heritrix lacks client side support, it cannot process JavaScript and other dynamic content embedded in a web page like Flash videos. Thus, even though WARC files may contain references to dynamic content in a web page, they fail to archive them correctly. Due to these limitations, we decided to harvest the archived data locally. The harvested data was further processed before being fed to the visualization framework. The methods described in the following sections were applied to the collections at Archive-It to convert them into a format suitable for visualizing within our framework. These methods can be easily applied to any other digital archive and its collections visualized in our framework.

III.1 WEB SCRAPING

The first step in visualizing collections was to replicate collection data locally. Harvesting the content directly from WARC and WAT files requires use of WARC extractor tools like arcreader, included in Heritrix, to extract the content of a WARC file. Also, in order to capture screenshots of each web page in the collection, we need to load each page in a web browser. We realized that a less efficient way of capturing the collection in the form of web scraping would eliminate the need of working with bulky WARC files and WARC extractors, in order to extract and parse web page content. Also, screenshots of web pages could be easily created during the web scraping process. Thus in order to avoid obtaining WARC files, using a WARC extractor and to facilitate easier creation of screenshots, we decided to web scrape Archive-It collections. We are aware that this technique is cumbersome and prone to failure when the web page layout changes, however its ease of implementation enables us to quickly gather collection data for building a proof of concept visualization framework.

Archive-It structures its collections in a hierarchy as shown in Figure 16 (in Chapter II). The presence of categories in the hierarchy is optional and often varies from collection to collection. In order to perform web scraping of collections, it is thus necessary to identify how this hierarchy exists in a certain collection presented to the user. Once this hierarchy is identified, a set of DOM (Document Object Model) [38] navigation rules can be written that can parse any collection available in the archive.

After scraping the collections, we store the following information regarding each collection in our database:

- Collection ID - generated by Archive-It

- Collection Title - assigned by collection curator
- Category - group in the collection created by curator
- Web page Title - title of a web page in a category or collection as assigned by curator
- Web page URI - URI of a web page in collection
- Archive URI - URI of an archived copy of a web page
- Archive Date - date when web page was archived at Archive-It.
- Content - HTML content of the archived copy of a web page
- Snapshot - an image of the archived copy of a web page, generated during scraping

We formulated the DOM navigation and scraping rules for Archive-It collections as an algorithm (Algorithm 1) and implemented it in Ruby[1]. To create snapshots of each archived copy of the web page, we accessed each archived copy in Mechanize[2], a headless browser and captured snapshots using PyWebShot[3], a Python[4] based image capture tool.

III.2 CALCULATION OF TF AND TF-IDF

One of the goals of developing this visualization framework is to allow users to gain insight about the evolution of a collection over time. A user can use the time cloud visualization (tag clouds presented on a timeline) to see how the content of the collection has changed over time. Similarly, the wordle visualization presents a quick summary of a web page in the collection. Both tag cloud and wordle use TF and TF-IDF of words appearing in the collection to create visual representations of top terms appearing in a document. For the purpose of creating these visualizations, we

[1] http://www.ruby-lang.org/en/

[2] http://mechanize.rubyforge.org/

[3] https://github.com/coderholic/PyWebShot

[4] http://www.python.org/

Algorithm 1 DOM navigation and scraping for Archive-It collections.

```
Require: $coll\_id \leftarrow$ id of collection
Require: $coll\_name \leftarrow$ name of collection
Require: $collections\_page \leftarrow$ link to collection's page on Archive-It
  $page \leftarrow$ OPEN($collections\_page$)
  for all $candidate$ such that $page\_element$ = "$/html/body/div/ul/li/b[1]/a$" do
    if $candidate$.href = NULL then
      $categories$.add($candidate$.HTML)
    else
      $webpages$.add(href ⇒ $candidate$.href, title ⇒ $candidate$.HTML, category ⇒
      "UNCATEGORIZED")
    end if
  end for
  for all $category$ in $categories$ do
    for all $webpage$ such that $page\_element$ =
        "$/html/body/div[2]/ul/li[category$.index$]/ul/li/h5/b/a$" do
      $webpages$.add(href ⇒ $webpage$.href, title ⇒ $webpage$.HTML, category ⇒
      $category$.HTML)
    end for
  end for
  for all $webpage$ in $webpages$ do
    $art\_page \leftarrow$ OPEN($webpage$.href)
    for all $archive$ in $art\_page\_element$ = "$/html/body/table[3]/tr[4]/td/a''$ do
      $archives$.add(href ⇒ $archive$.href, date ⇒ $archive$.HTML, webpage ⇒
      $webpage$)
    end for
  end for
  for all $archive$ in $archives$ do
    $arch\_uri \leftarrow$ OPEN($archive$.href)
    $content \leftarrow arch\_uri$.HTML
    $snapshot \leftarrow$ CAPTURE_SNAPSHOT($arch\_uri$)
    $archive$.snapshot $\leftarrow snapshot$
    $art\_name \leftarrow archive$.webpage.title
    $art\_uri \leftarrow archive$.webpage.href
    $category \leftarrow archive$.webpage.category
    SAVE_TO_DB($coll\_id$,$coll\_name$,$category$,$art_name$,$art_uri$,$arch\_uri$,$content$,
    $snapshot$)
  end for
```

Algorithm 2 Term Frequency (TF) calculation for a web page.

```
Require: $text$ ← text of a web page
Require: $stop\_words$ ← list of stop words (Appendix A)
  for each $word$ in $text$ do
    if $word$ present in $stop\_words$ then
      $text$.remove($word$)
    end if
  end for
  for all $word$ in $text$ do
    stem($word$)
  end for
  $TF$ ← NEW(Hash)
  for all $word$ in $text$ do
    $TF[word]$ ← $TF[word]$+1
  end for
  $TF$ ← SORT_DESC($TF$)
```

Algorithm 3 Term Frequency-Inverse Document Frequency (TF-IDF) calculation for a web page.

```
Require: $TF$ ← list of TF for each relevant term in the text of a web page
Require: $total\_TF$ ← total of TF for each relevant term in the text of a web page
Require: $total\_indexed\_pages$ ← number of indexed pages in Yahoo search engine
  [1]
  for each $term$ in $TF$ do
    $DF$ ← frequency of occurrence of $term$ in Yahoo search engine
    $normalized\_TF$ ← $TF[word]/total\_TF$
    $IDF$ ← log($total\_indexed\_pages/DF$)
    $TF\_IDF$ ← $normalized\_TF * IDF$
  end for
  $TF\_IDF$ ← SORT_DESC($TF\_IDF$)
```

calculated TF and TF-IDF scores for each archived copy of every web page in the database and stored them along with other data for the archive.

Stop words are frequently appearing words in text that do not add any important significance to the text as a whole. Stop words include articles, conjunctions, prepositions, pronouns and interjections. They may also include words belonging to other parts of speech if those words are commonly used. Removing stop words helps us to increase the weight of relevant terms in a document relative to the overall weight of all terms in the document. In order to compute TF and TF-IDF scores for a web page, we first removed all stop words (Appendix A) appearing in the text of the web page. These stop words were sourced from Onix Text Retrieval Toolkit [39], a fast full text indexing engine. We next stemmed the remaining words using Porter Stemmer [40] for the purpose of normalizing the text of the web page and then calculated the TF scores for each word in the text (Algorithm 2). Algorithm 3 illustrates how we calculated the TF-IDF scores for each web page using the computed TF score for each word in the text of the web page and the same word's frequency of occurrence in all web pages in the Yahoo search engine [41, 42].

In the time cloud visualization, first we retrieve the top 10 TF and TF-IDF terms for all web pages in a collection for the specified date range and aggregate them together. We then choose the top 20 terms from this aggregation and display them to the user. To create a wordle for a web page, we aggregate all terms from all archived versions of the web page and then generate a wordle. This allows us to present a more detailed visual summary covering all memes discussed on a web page over time.

III.3 RULE-BASED CATEGORIZATION

Several Archive-It collections do not have the web pages in the collection organized into groups (for example, the Pakistan Floods collection), making it difficult for the user to explore the collection. Thus, we provide an option of exploring the collection using a rule-based categorization. Our approach for rule-based web page classification is similar to the one discussed by Kan and Thi [43]. Our rules for rule-based categorization use the hostname component of URI. The following rules were run on all collections classifying the web pages into broad categories:

- If the hostname contains the words “Facebook”, “twitter” or “wiki”, put the URI in the Social Media category.
- If the hostname contains the words “bbc”, “cnn”, “nytimes”, “msnbc”, “huffingtonpost”, “foxnews”, “reuters” or “abcnews”, put the URI in the News Web Sites category.
- If the hostname contains the words “blog” or “wordpress”, put the URI in the Blogs category.
- If the hostname contains the words “YouTube” or “dailymotion” words, put the URI in the Videos category.

At the same time, certain rules were applied only to specific collections based on the top level domain name (TLD) of a website to achieve a finer granularity in category assignment. For example, the web pages in Collection 11 (South Dakota Government) were grouped based on the TLD. All web pages with a TLD “.gov” were grouped as government web pages and those with the TLD “.edu” were grouped as education web pages.

We also developed specific rules for particular collections. For example, Collection 2836 (Pakistan Floods collection) has specific rules for the “Pakistan News Sites” and “Relief Websites” category. These rules are mentioned below:

- If the top level domain is .co.pk or .pk, put the URI in the Pakistan News Sites category.
- If the hostname contains the words “relief” or “aid”, put the URI in the Relief Websites category.

If we were unable to group a web page using any predefined (either generic or specific) rule, it was grouped under “Others”. Such rule-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.

III.4 SUMMARY

This chapter discussed the data acquisition and preprocessing techniques that were used to copy the collections listed in Table 2 into our database. In order to avoid using WARC extractors, the inability of WARC or WAT files to archive client side interactive content like Flash videos, and our need to create screenshots, we decided to screen scrape our collections. Once stored in our database, we calculated TF and TF-IDF of all terms in a web page in a collection and performed rule-based categorization of web pages in all collections. This processing of data allowed us to easily create JSON representations for visualization.

CHAPTER IV

VISUALIZING ARCHIVE-IT COLLECTIONS

The structure of all Archive-It collections is governed by the metrics listed in Table 1 (in Chapter I). To meaningfully capture the essence of the collection and present it to the user in a visual form, we need to use visualizations which are effective at depicting both temporal and spatial information. The visualizations discussed in Chapter II are most suited for this purpose. We have implemented these visualizations using JavaScript, jQuery, JavaScript InfoVis Toolkit (JIT)[1] and Highcharts JS[2] to present the collections inside a web browser, much like the current Archive-It interface.

IV.1 TREEMAP

Archive-It collections are organized in a hierarchy, as shown in Figure 16 (in Chapter II). This makes them a natural fit for visualization as a treemap. The collection title can be seen as the root of the tree, with curator-created categories as its children, in level 1. The individual web pages comprise level 2, as children of each category. For each web page, there are one or more archived versions at level 3.

To create our treemap visualization we used JIT, a JavaScript-based visualization toolkit. Separate JSON objects representing each level in the collection hierarchy were created and fed into the toolkit to create a nested treemap. Unlike usual treemaps, instead of showing sub-rectangles representing a node's children, we show a screenshot of a web page along with the title of the category or web page and the number of children the node has. The user can left-click on a node to navigate to the next level, or right-click on it to navigate to a higher level. At the category level (Figure 24) and web page levels (Figure 25) (levels 1 and 2), the size of each

[1] http://thejit.org/

[2] http://www.highcharts.com/

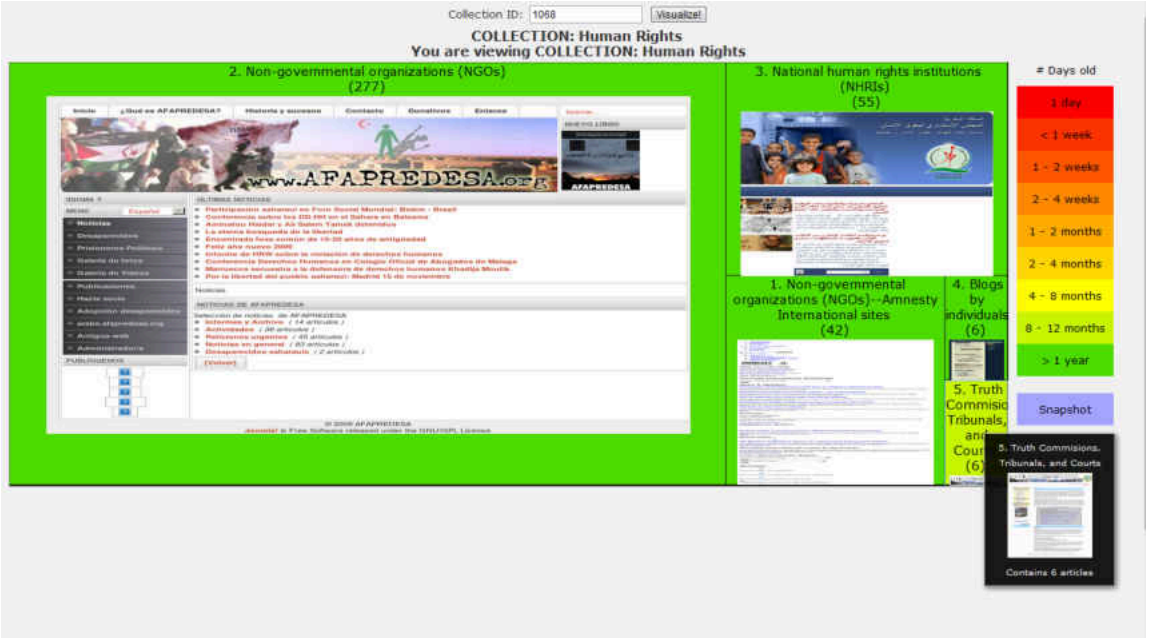

Collection ID: 1068
Visualize!
COLLECTION: Human Rights
You are viewing COLLECTION: Human Rights
2. Non-governmental organizations (NGOs) (277)
3. National human rights institutions (NHRIs) (55)
1. Non-governmental organizations (NGOs)--Amnesty International sites (42)
4. Blogs by individuals (6)
5. Truth Commissio Tribunals, and Cour (6)
# Days old
1 day
< 1 week
1 - 2 weeks
2 - 4 weeks
1 - 2 months
2 - 4 months
4 - 8 months
8 - 12 months
> 1 year
Snapshot
5. Truth Commisions, Tribunals, and Courts
Contains 6 articles

Fig. 24. Category level (level 1) visualization of the Human Rights collection. The number of web pages in each category is represented by area of the rectangle, and the timespan of the archived versions of the web pages is depicted by color.

rectangle represents the number of child nodes and the color represents the timespan over which archives have been collected for web pages represented by that rectangle. Together the size of the rectangle and color of the node quickly provide an insight to the user about the relevance of a web page to the collection.

At level 1 (Figure 24), the screenshot visible in each rectangle gives the user an example of the type of web pages present in each category. This screenshot is also visible when the user hovers over an individual node, which is especially useful for smaller categories where the screenshot cannot fit inside the rectangle.

At level 2 (Figure 25), all web pages in a category are revealed to the user. Here, the size of each rectangle represents the number of archived versions of the web page, and the color represents the timespan of the archives for the web page. The screenshot displayed in the rectangle and on the mouse-over is the latest archived version of the web page.

At level 3 (Figure 26), the user can see all archived versions of a selected web

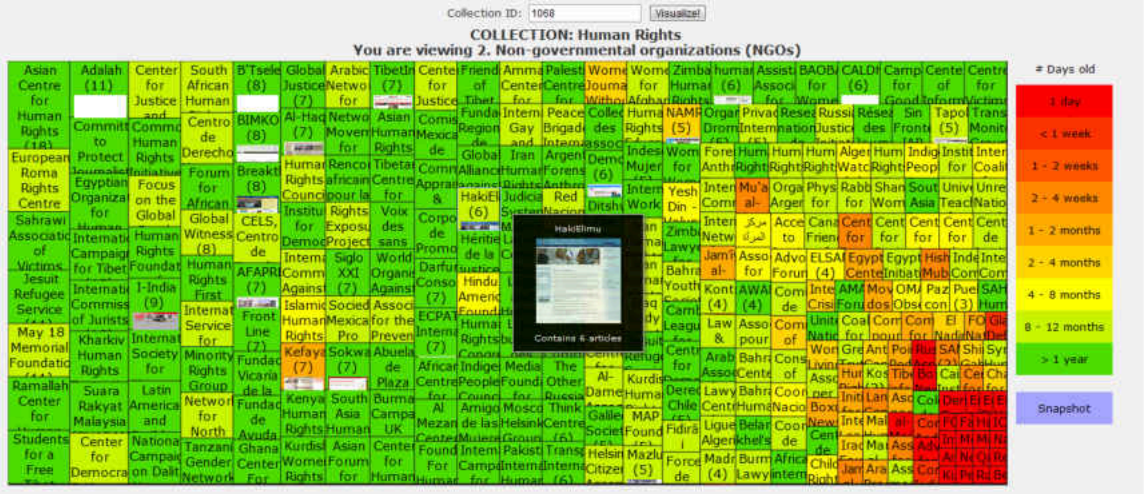


Fig. 25. Web page level (level 2) visualization of the NGO category of the Human Rights collection. All web pages within the NGO category are represented.

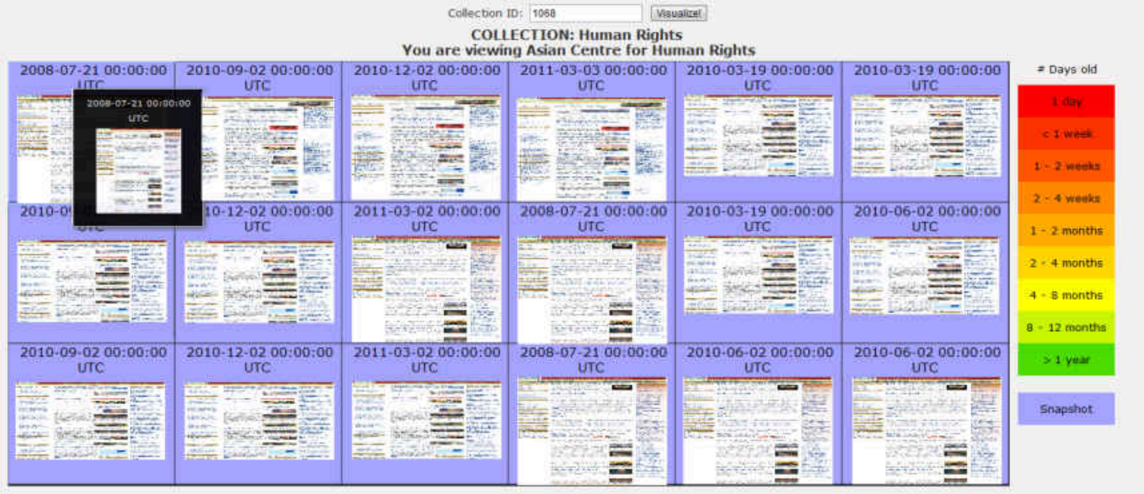


Fig. 26. Archived version level (level 3) visualization of a web page in the NGO category of the Human Rights collection. All nodes are of equal size and have the same color.

page. At this level, all rectangles are the same size and have the same color. The screenshots here can help the user determine how the web page has changed over time. Clicking on any of the archived versions will take the user to the web page as captured by Archive-It on the specified date.

IV.2 TIME CLOUD

For the purpose of visualizing changes in document collections, we decided to add the dimension of time to the frequency of terms in documents and create a time cloud visualization which depicts the movement of tag clouds over time, thus allowing the user to visualize how the theme of the collection changes over time.

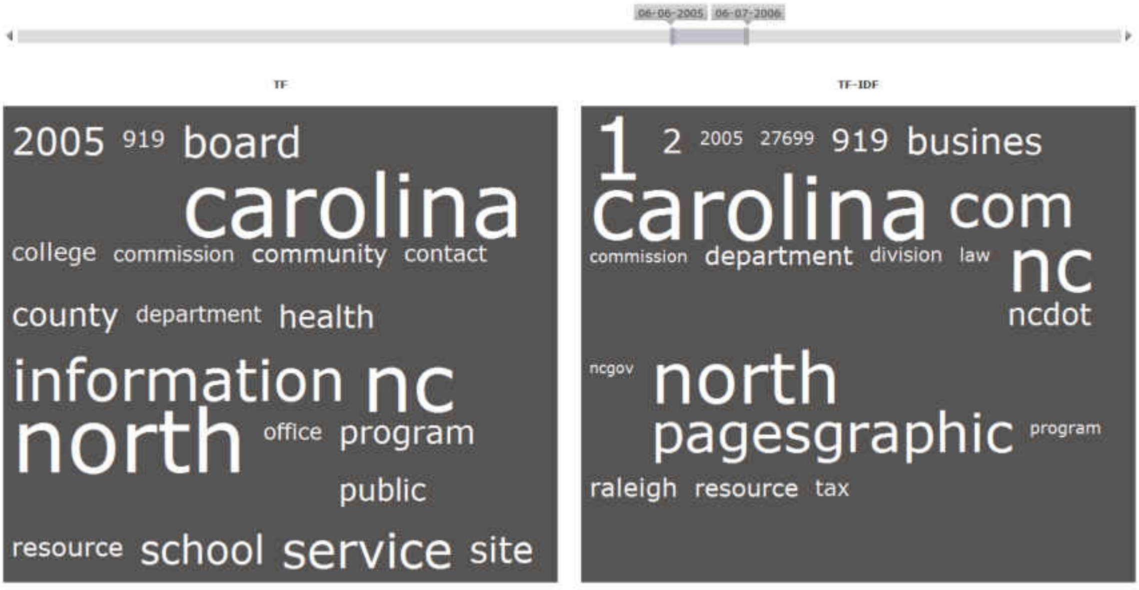
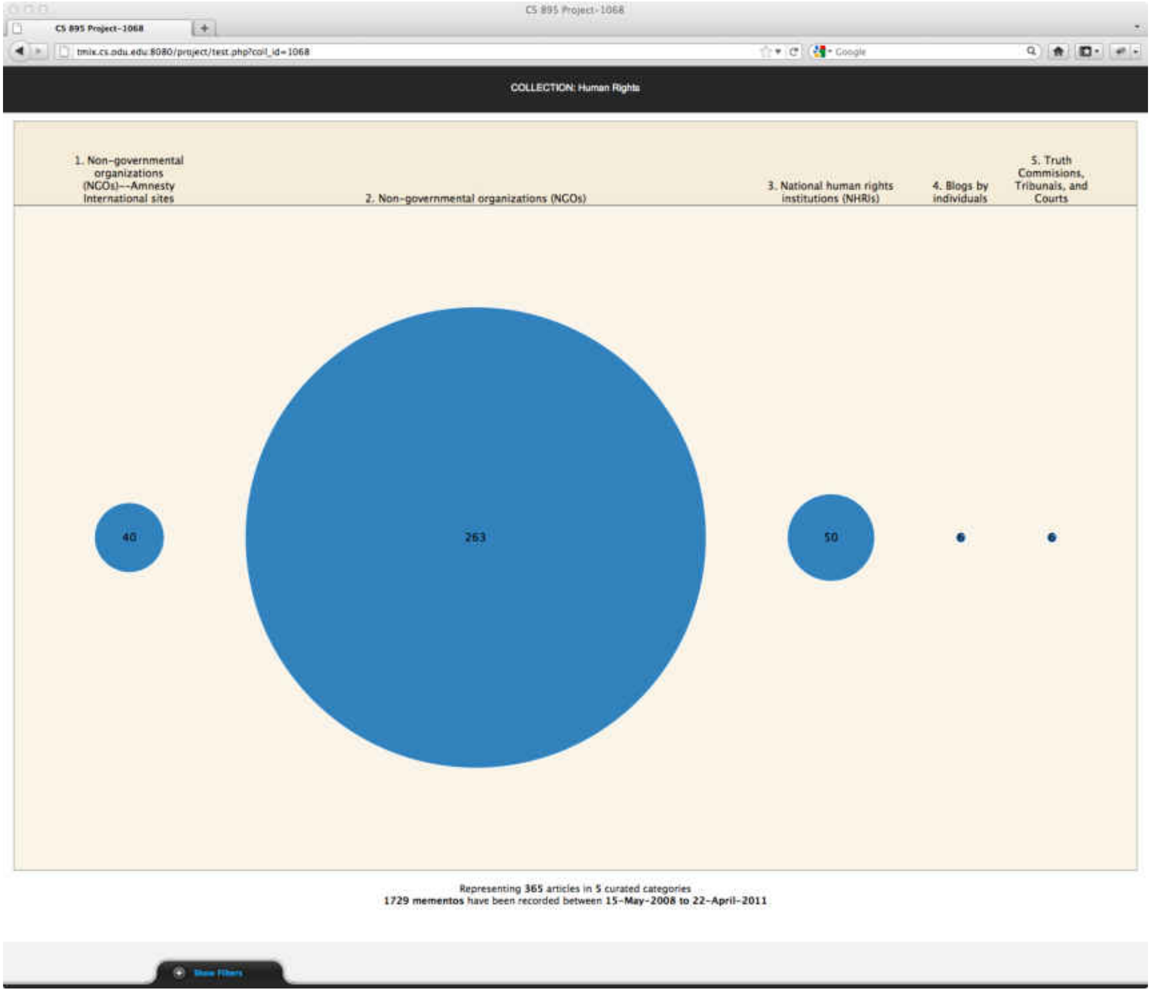

06-06-2005
06-07-2006
TF
TF-IDF
2005 919 board carolina college commission community contact county department health information nc north office program public resource school service site
1 2 2005 27699 919 busines carolina com commission department division law nc ncdot ncgov north pagesgraphic program raleigh resource tax

Fig. 27. Time clouds for the North Carolina State Government Web Site collection. The tag cloud on left is based on TF score from the collection. The tag cloud on the right is based on the TF-IDF score from the collection.

Figure 27 shows the time cloud visualization for the North Carolina State Government Web Site collection. A slider on the top allows the user to select a date range for visualization. Beneath the slider, a tag cloud created using TF is shown on the left, while a tag cloud created using TF-IDF is shown on the right. Both clouds show the top 20 terms in the selected date range, sorted in alphabetical order. While TF tells the user how often a term appears in the collection, TF-IDF evaluates a term's importance in the collection. Comparing term frequency with TF-IDF also provides further insight to the user about which terms are most popular during a timespan and which are most common over a larger time period.

IV.3 BUBBLE CHART


CS 895 Project-1068
CS 895 Project-1068
tmix.cs.odu.edu:8080/project/test.php?coll_id=1068
Google
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 Commisions, Tribunals, and Courts
40
263
50
Show Filters

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

Fig. 28. Bubble chart visualization for the Human Rights collection.

We intend the bubble chart visualization (Figure 28) to provide a quick summary of the collection. For this purpose, the bubbles are placed side by side. The placement of a bubble on the axis marks which category it depicts in our visualization, while the area of the bubble represents the number of web pages in each category. The number of total web pages, 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 web pages in that group, allowing the user to quickly filter through the collection by group.

IV.4 IMAGE PLOT

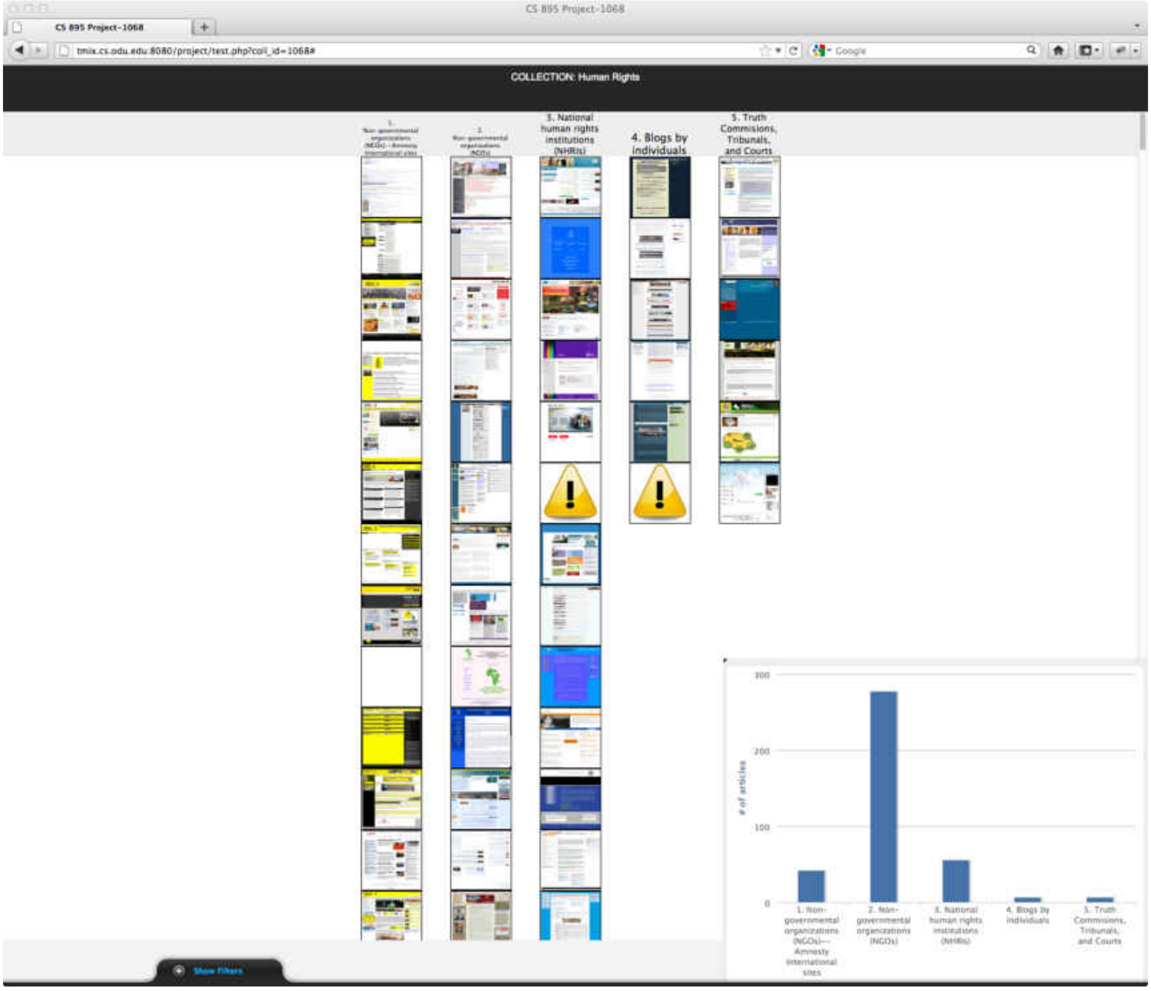

CS 895 Project-1068
CS 895 Project-1068
tmix.cs.odu.edu:8080/project/test.php?coll_id=1068#
Google
COLLECTION: Human Rights
1. Non-governmental organizations (NGOs)—Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
Show Filters
300
200
100
0
# of articles
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 Commisions, Tribunals, and Courts

Fig. 29. Image plot visualization for the Human Rights collection.

The image plot (Figure 29) is an implementation of an inverted stacked bar chart to represent all web pages 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 the latest capture of each archived web page. Each screenshot is linked to Archive-It's list of archived versions of that web page.

The inverted representation allows the user to see both larger and smaller groups side by side. Since it is likely that not all web pages will be viewable in a single window, a resizable histogram in the bottom right corner shows the number of web pages in each group, so that the user gains an overview of the distribution of web pages over the groups.

IV.5 TIMELINE

Often collection curators are interested in discovering how the collection evolved over time to correlate events in history with the organization of the collection. A recent study of web archive users found that providing a timeline interface may help users to better understand the temporal nature of web archives [44]. We provide a timeline visualization (Figure 30) for visualizing the development of the collection over time. We created this visualization with the help of Highcharts JS, a JavaScript-based charting engine. Here, each web page is represented as a single horizontal line, the length of which denotes the duration over which its archived copies have been captured. A line's color depicts the category to which the web page belongs to. Thus all web pages belonging to a category are depicted by the same color. Each point on the line represents an archived copy of the web page. Hovering over a point displays a list of archives of other web pages captured on that same day (Figure 31). A slider at the bottom of the timeline allows the user to easily change the resolution of the timeline from months to days to hours and vice versa to observe the collection at varying granularity level (Figure 32). A curator can thus easily see the growth of a collection over time by looking at web page density and analyzing the addition (or removal) of web pages from the collection. Interesting patterns about the structure and evolution of collection are observed when the collection is visualized on a timeline in this manner. While short-lived web pages stand out, a web page with multiple archived copies clearly signifies its importance to the collection.

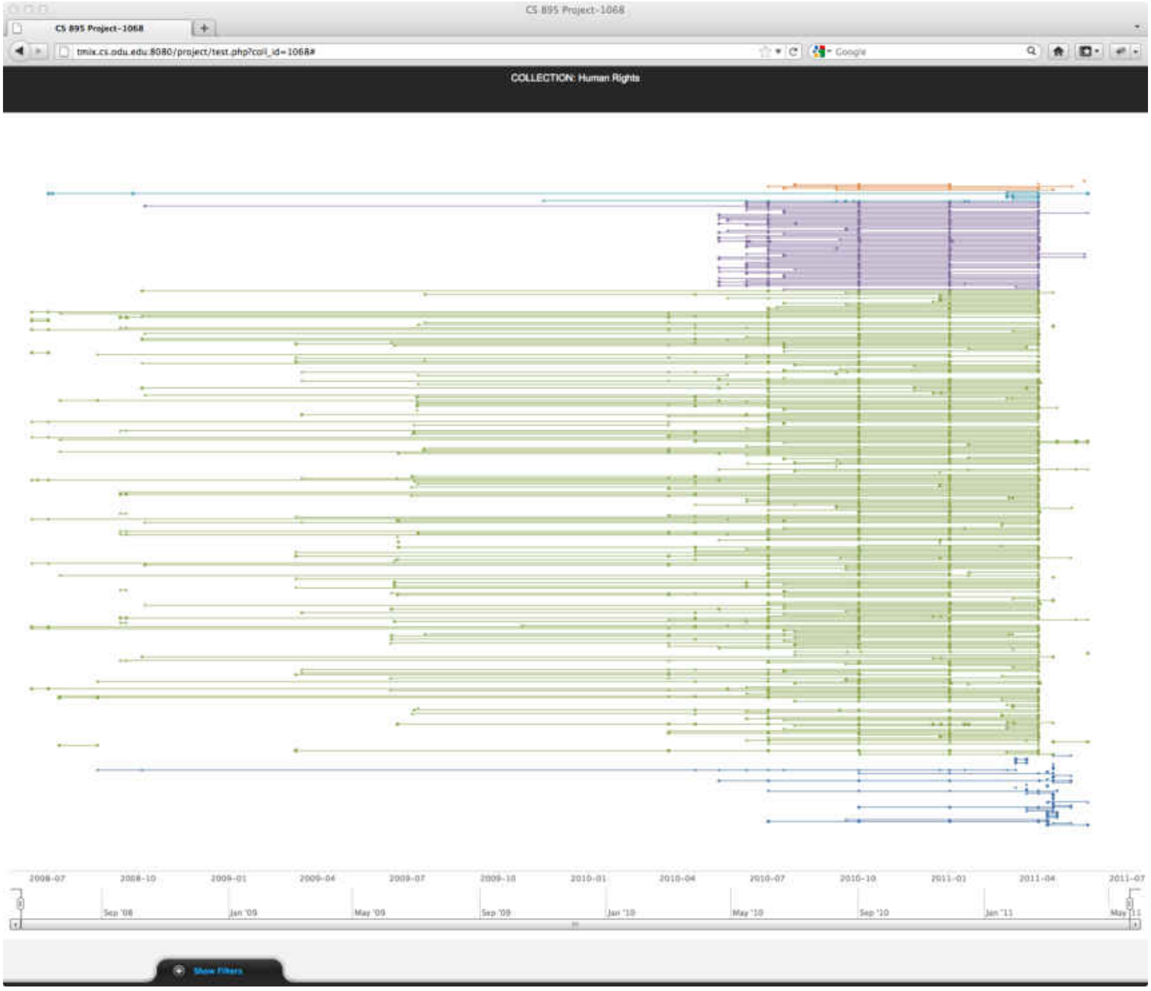
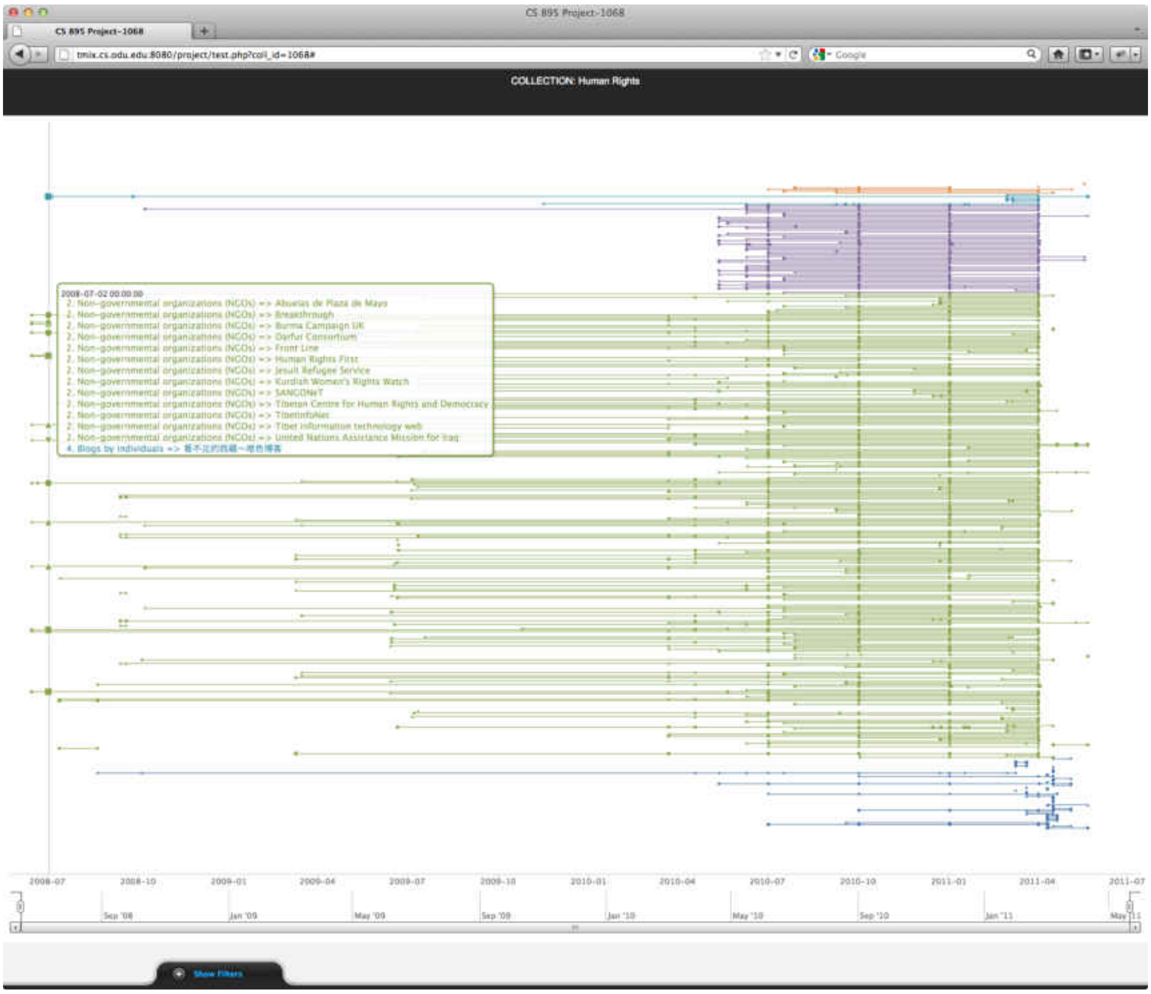


Fig. 30. Timeline visualization for the Human Rights collection.

IV.6 WORDLE

We use wordle for summarizing content (text) of Archive-It collections. For multimedia content in the collection, the wordle summarizes the comments (if present). To create a wordle for any web page, we provided the top 20 terms for the web page based on their TF score as an input to pyTagCloud[3], a Python library based on Feinberg's algorithm [29] . The library provided us with a wordle created from the input terms as a JPEG image. Figure 33 shows a wordle for a web page in Human Rights collection. We have integrated the wordle visualization with image plot in

[3] https://github.com/atizo/PyTagCloud



2008-07-02 00:00:00
2. Non-governmental organizations (NGOs) => Abuelas de Plaza de Mayo
2. Non-governmental organizations (NGOs) => Breakthrough
2. Non-governmental organizations (NGOs) => Burma Campaign UK
2. Non-governmental organizations (NGOs) => Darfur Consortium
2. Non-governmental organizations (NGOs) => Front Line
2. Non-governmental organizations (NGOs) => Human Rights First
2. Non-governmental organizations (NGOs) => Jesuit Refugee Service
2. Non-governmental organizations (NGOs) => Kurdish Women's Rights Watch
2. Non-governmental organizations (NGOs) => SANGONeT
2. Non-governmental organizations (NGOs) => Tibetan Centre for Human Rights and Democracy
2. Non-governmental organizations (NGOs) => TibetInfoNet
2. Non-governmental organizations (NGOs) => Tibet information technology web
2. Non-governmental organizations (NGOs) => United Nations Assistance Mission for Iraq
4. Blogs by individuals => 看不见的西藏—唯色博客

Fig. 31. Timeline visualization for the Human Rights collection showing archived copies on a specific date.

our interface. Hovering over any image in the image plot reveals a wordle (Figure 34 overlay) summarizing the content discussed on the web page. This wordle representation aids the understanding of a web page 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.

IV.7 SUMMARY

This chapter discussed the various visualizations we implemented as a part of this work. The treemap visualization allows the user to visualize the hierarchical

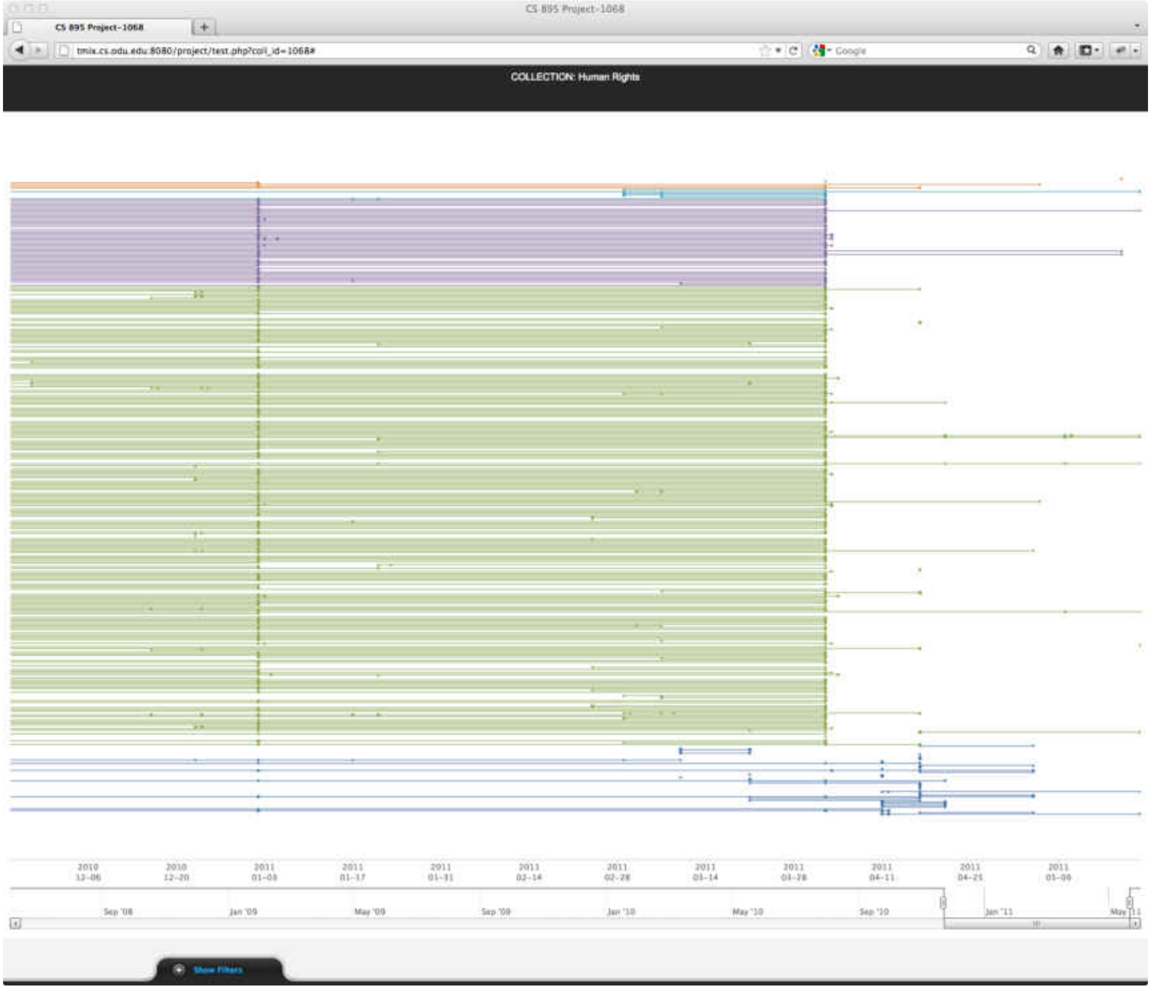


Fig. 32. Tuning granularity levels on timeline visualization.

structure of the collection in a very efficient manner. While wordle provides a visual summary of the collection, time cloud enables the user to visualize the top terms in the collection over a temporal component. Timeline is another visualization that leverages the temporal component of collections to depict the overall collection growth and structure. The image plot represents all web pages as an inverted bar chart, and allows visual exploration of web pages using their screenshots. Finally, the bubble chart gives a bird's eye view of the collection statistics viz. number of web pages, number of mementos, timespan and number of categories.

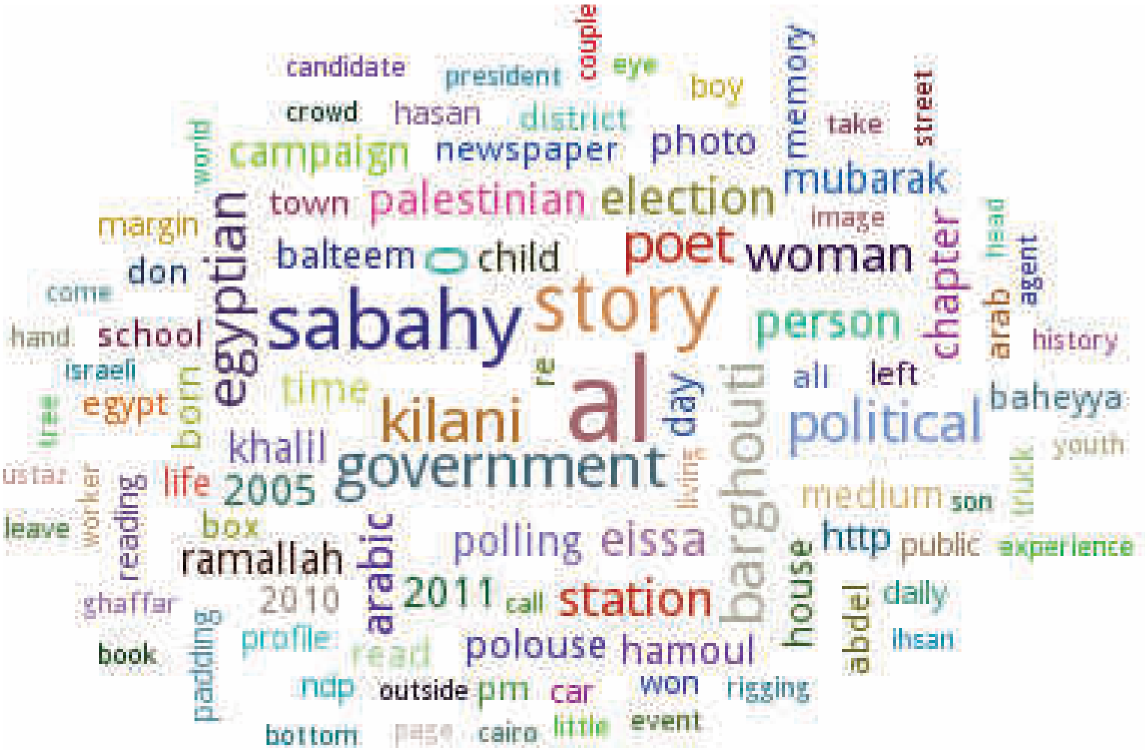


Fig. 33. Wordle visualization for a web page in Human Rights collection.

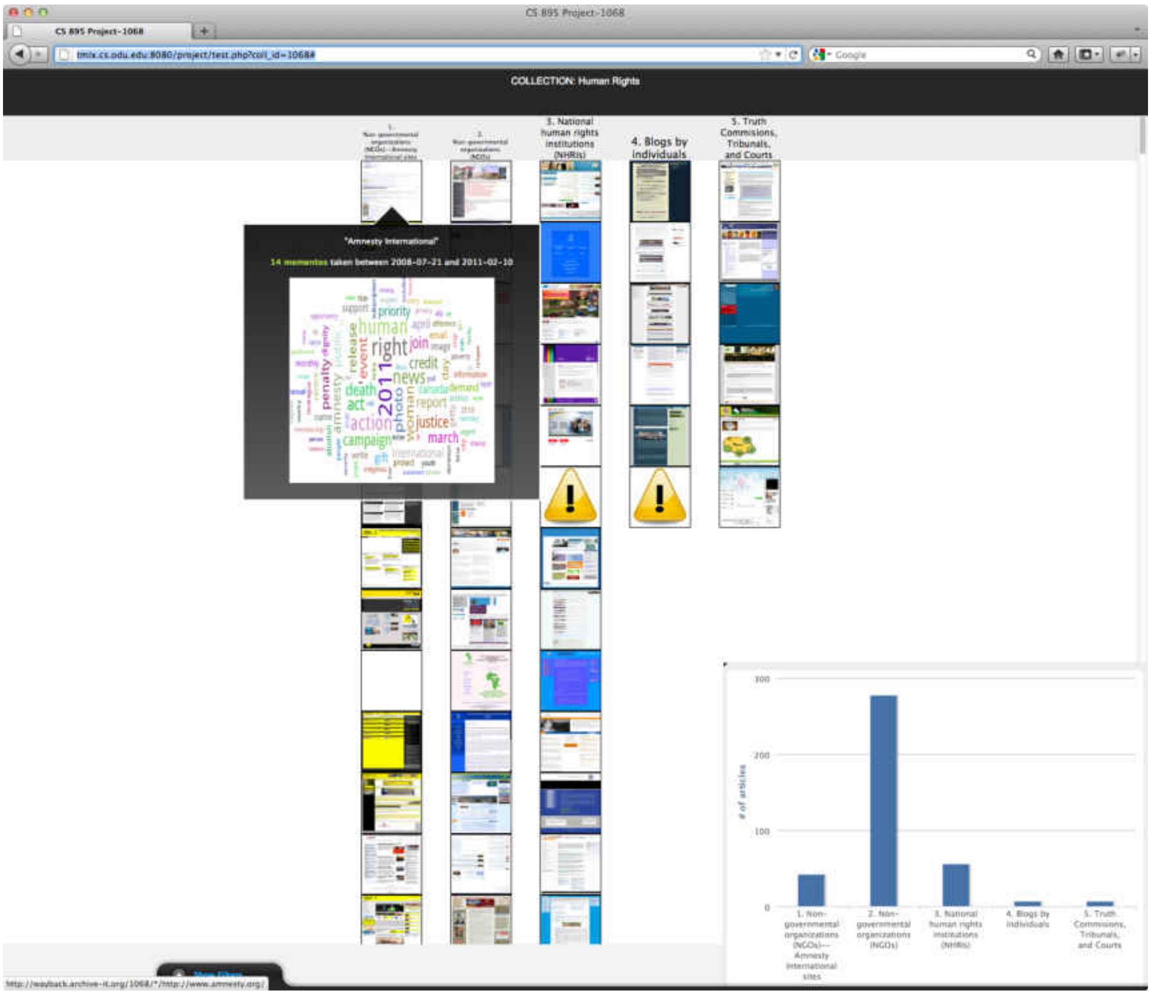

CS 895 Project-1068
tmix.cs.odu.edu:8080/project/test.php?coll_id=1068#
Google
COLLECTION: Human Rights
1. Non-governmental organizations (NGOs)—Amnesty International sites
2. Non-governmental organizations (NGOs)
3. National human rights institutions (NHRIs)
4. Blogs by individuals
5. Truth Commisions, Tribunals, and Courts
"Amnesty International"
14 mementos taken between 2008-07-21 and 2011-02-10
human
right
2011
news
penalty
amnesty
release
event
death
action
campaign
justice
report
credit
photo
woman
march
join
priority
support
day
act
international
# of articles
300
200
100
0
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 Commisions, Tribunals, and Courts
http://wayback.archive-it.org/1068/*/http://www.amnesty.org/

Fig. 34. Image plot visualization for the Human Rights collection, showing the wordle for the highlighted web page.

CHAPTER V

CASE STUDIES

We will now discuss various case studies to illustrate how each of the visualization techniques aids in better understanding archived collections.

V.1 COLLECTION BUILDING (AND GROWTH)

Most often an archived collection grows over time due to the addition of new web pages to the collection. Also, new mementos of already archived web pages cause the collection to grow over time. If a collection continues to add new web pages or capture mementos at a steady rate, the number of captures during a certain time interval continues to be monotonically increasing. For example, when a memento is captured every month for each web page in a collection with $\boldsymbol{x}$ web pages, the collection grows at a steady rate of $\boldsymbol{x}$ captures every month. The collection is thus said to have a positive rate of growth.

A collection experiences negative rate of growth if during a certain time interval, the number of captures (or additions) are less than the previous time interval. For example, if a collection captures $\boldsymbol{x}$ mementos every month, but during a certain month captured only $\boldsymbol{y} < \boldsymbol{x}$ mementos, then it experiences a negative rate of growth in that time interval, though overall the growth rate might still be positive.

Visualizing the rate of growth of a collection is important in understanding how the structure of the collection changes over time. While a positive rate of growth depicts an actively growing collection, a negative growth rate hints towards possible structural changes in the collection such as removal of web pages from the archiving schedule because their content no longer contributes to the collection theme or a slow down in the rate of memento capture due to slower updates of web pages. It can also be used to identify those collections which have stopped growing and are not maintained anymore, thus identifying closed archives.

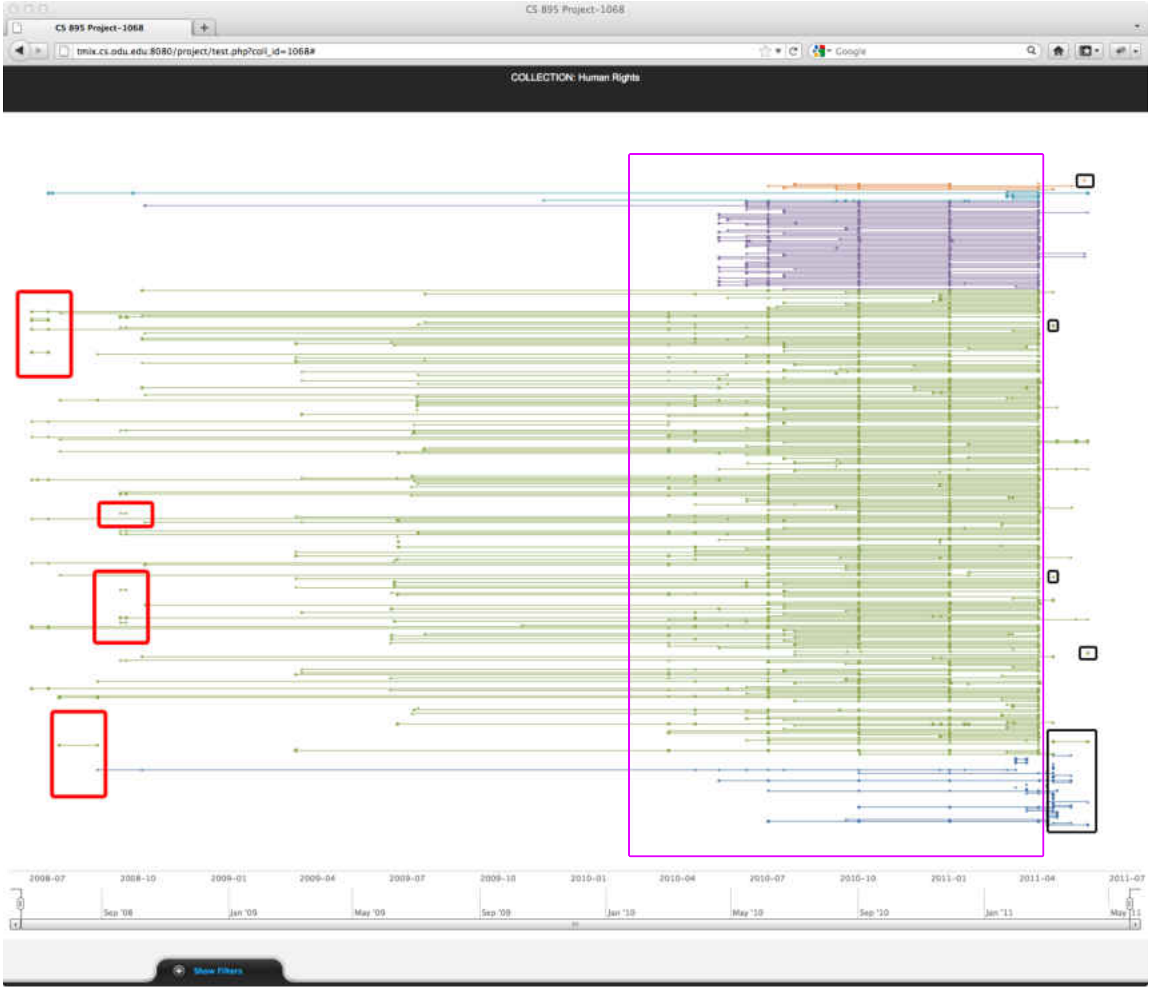
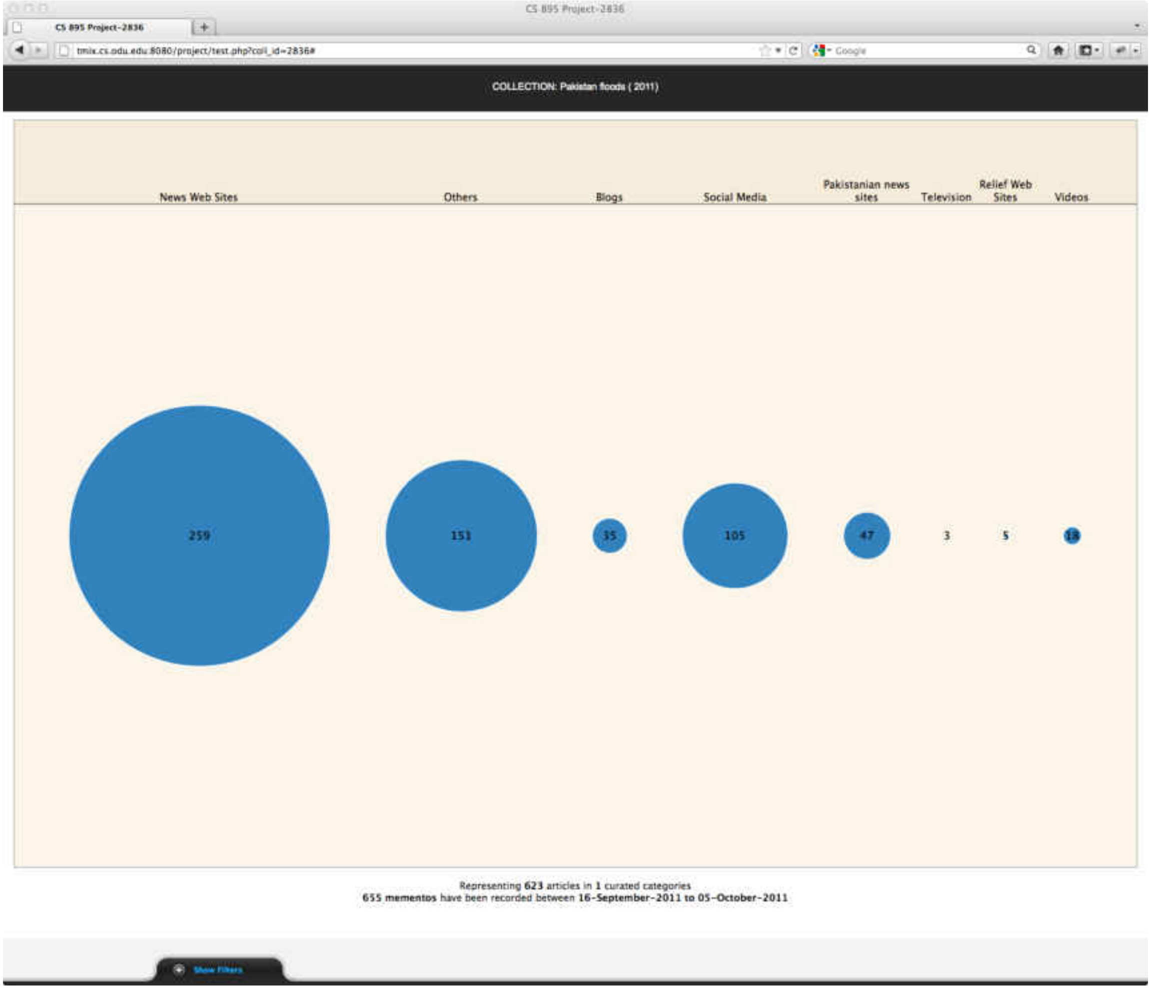


Fig. 35. Annotated timeline visualization for Human Rights collection. Lines within red annotations depict web pages that were added to the collection but quickly removed from the archiving schedule. Lines within purple annotation depict new additions to the collection and also introduction of a quarterly crawl cycle from July 2010 onwards. Points/lines within black annotations depict recently added web pages to the collection.

The timeline visualization discussed in Chapter IV allows users to visualize the growth of the collection and also inspect when web pages have been added (or removed) from the collection. Figure 35 shows the timeline visualization for the entire duration over which the Human Rights collection was archived. The changes in the collection are easily noticeable in this visualization. At the extreme left of the visualization (Figure 35, red annotations), the user can identify the web pages that were added to the collection, but quickly removed from the archiving schedule. Loosely packed lines in the center of the timeline between September 2008 to May 2010 signify a more or less steady growth of the collection, with new web pages being occasionally added to the collection and mementos for existing web pages being captured at regular intervals. Densely packed lines between May 2010 to April 2011 (Figure 35, purple annotation) represent a steep increase in the size of the collection. Not only were a lot of new web pages added to the collection, but also mementos were captured at a faster rate during this period. The timeline visualization also uncovers the introduction of a more regular, quarterly crawl rate for web pages in the collection. It can be seen that irrespective of the previous crawl rate, the new crawl rate for most web pages was set to once every 3 months. The abrupt end in the lines representing these web pages at April 2011 was perhaps because we obtained the collection in June 2011 while the next crawl for these web pages was scheduled for July 2011. Also it can be noticed that few new web pages were added to the collection after April 2011 (Figure 35, black annotation).

While Heritrix can make such information available to a collection curator in the form of statistical data, such analysis of the collection structure is not possible using the text based interface of Archive-It available to general users. The timeline visualization can thus aid curators in making informed decisions about addition or removal of web pages from collection and to fine tune their archiving schedule. It also allows general users to gain insight about how the collection structure changes over time along with information about addition and removal of web pages to/from the collection.

V.2 RE-CATEGORIZATION



COLLECTION: Pakistan floods (2011)

Representing 623 articles in 1 curated categories
655 mementos have been recorded between 16-September-2011 to 05-October-2011

Fig. 36. Bubble chart visualization for the Pakistan Floods (2011) collection.

Archive-It allows curators to organize the web pages in a collection under various groups or categories. Often, however, this grouping is missing from the collection. The lack of such grouping makes it difficult for the user to quickly filter through the collection to identify web pages of interest. We have implemented a rule-based categorization for Archive-It collections that allows the user to visualize the collections in each of the aforementioned visualization techniques with the new grouping.

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 web pages collected over a period of 20 days. The absence of

TABLE 3
Categories assigned to web pages in the Pakistan Floods collection after applying rule-based categorization.

Group	#of web pages
Blogs	35
News Websites	259
Pakistan News Sites	47
Relief Websites	5
Social Media	105
Television	3
Videos	18
Others	151

any grouping makes it difficult for the user to selectively explore the collection.

Prior to categorization, the Level 1 meta-data from Figure 16 (Chapter II) contains only one category, uncategorized, since there is no curator-defined grouping. When dealing with event-driven collections like the Pakistan Floods collection, it is not uncommon to find the absence of a curator-defined grouping since the primary focus when creating such collections is on seeding it with as many relevant URIs as possible, and the categorization of seed URIs can be dealt with at a later stage.

After running our rule-based categorization, the collection was organized into eight categories as shown in Table 3. Seven of these eight categories are meaningful, while the category “Others” consists of all those web pages which could not be categorized under any of the other seven categories.

Since this collection is concerned with a natural disaster, we added the following rules to our general rules in order to categorize news websites from Pakistan and websites discussing relief operations into their own groups:

- If the top level domain is .co.pk or .pk, put the URI under the Pakistan News Sites category.
- If the hostname contains the words “relief” or “aid”, put the URI under the Relief Websites category.

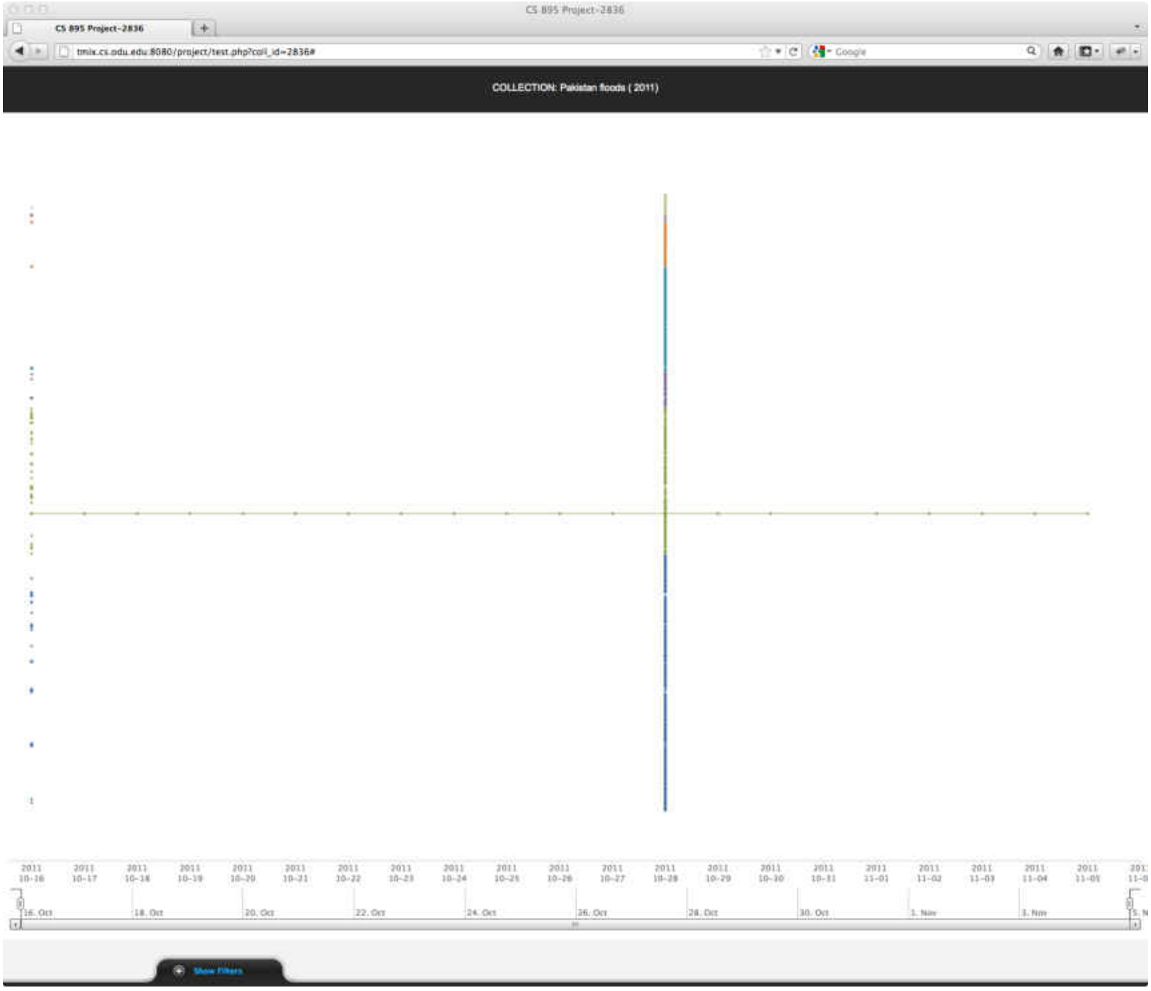


Fig. 37. Timeline visualization for the Pakistan Floods (2011) collection after rule-based categorization.

This categorization is shown in the bubble chart in Figure 36. After this categorization, the composition of the collection can be visualized more clearly by the user. Also by looking at the timeline visualization for this collection (Figure 37), the user can infer almost immediately that this collection is either still in its early stages because for most of the web pages there is only 1 archived copy, or that the collection is closed now, since no new web pages are added to the collection and also no new mementos have been captured. Figure 37 also shows that the timeline visualization has a bias for long running collection as opposed to an event-driven "snapshot" collection.

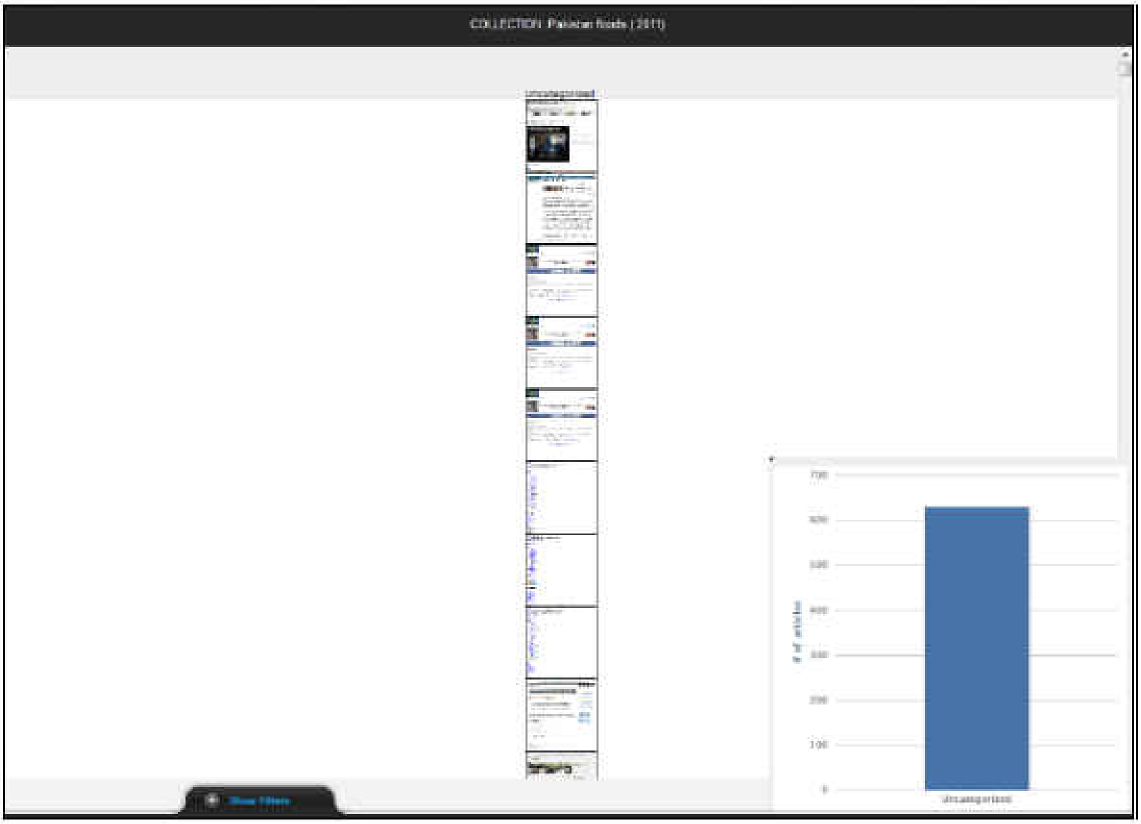


Fig. 38. Image plot visualization for the Pakistan Floods (2011) collection before rule-based categorization.

Figure 38 shows the image plot visualization for the collection before applying the rule-based categorization. Without any defined grouping, all 623 web pages are represented on a single bar. On the other hand Figure 39 shows the image plot visualization for the collection after applying the rule-based categorization. It can be seen that the collection has a larger spread and the user can interact more efficiently with the web pages in the collection.

V.3 COLLECTION SYNOPSIS

Identifying the theme and summarizing the content to get a better understanding of a collection is a challenge for a user exploring Archive-It collections. Archive-It allows curators to provide a brief description of a collection to describe its nature. However, in many cases this description alone is not sufficient to describe the nature

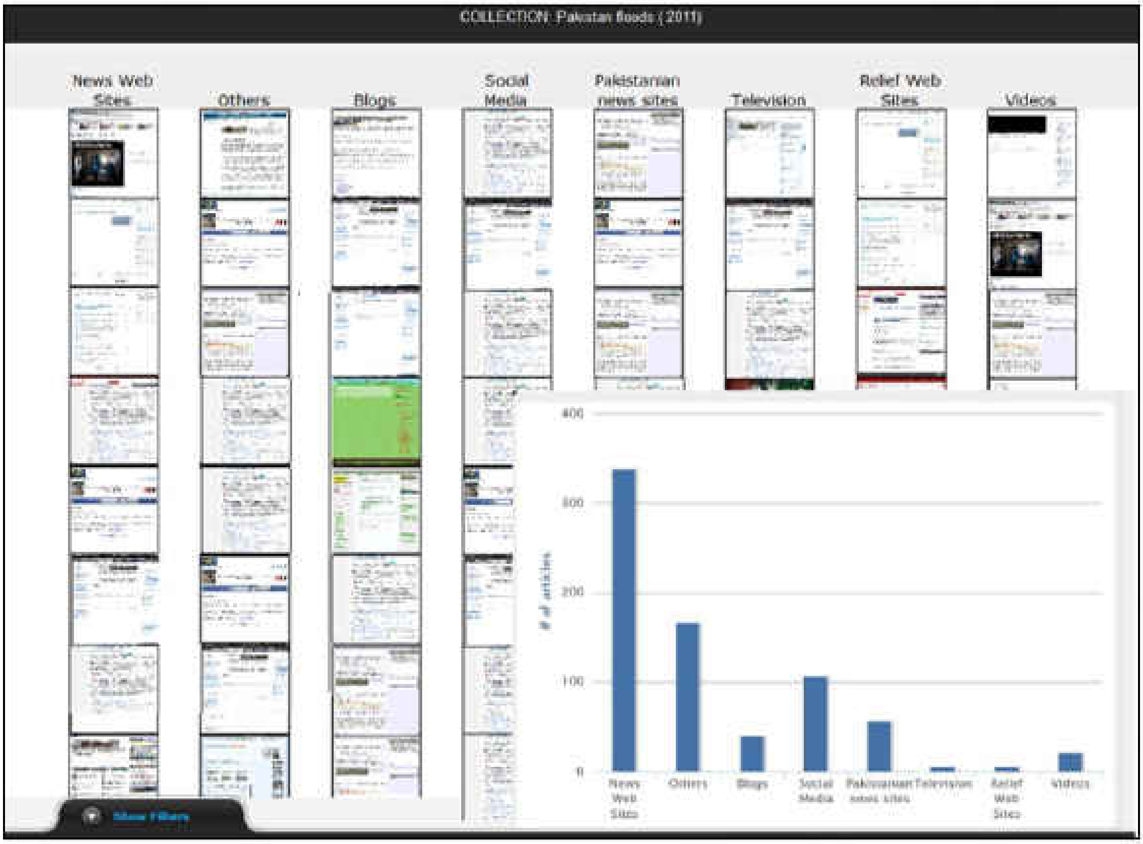
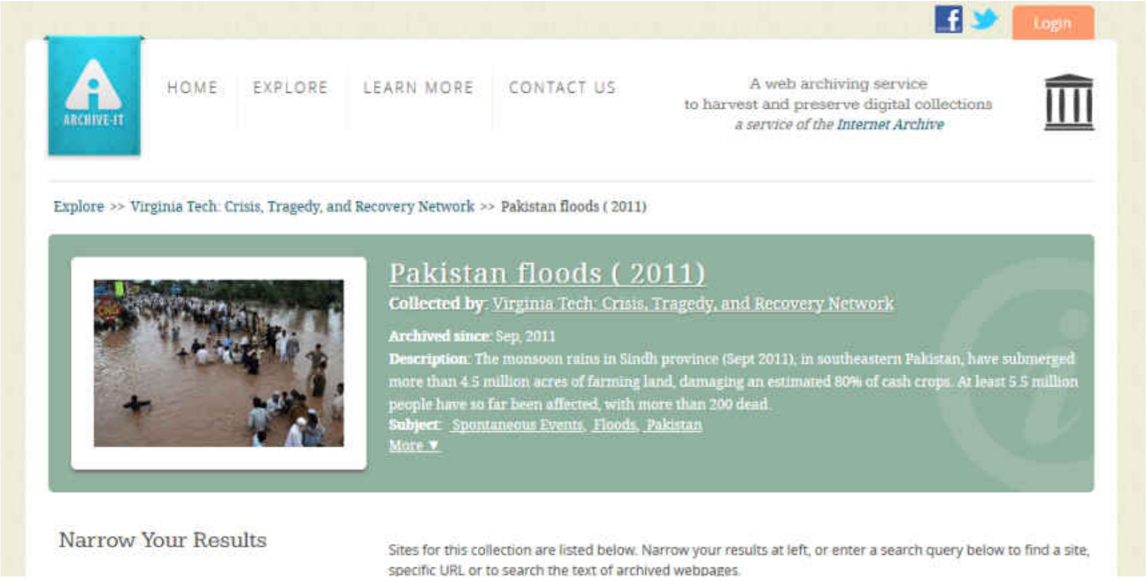


Fig. 39. Image plot visualization for the Pakistan Floods (2011) collection after rule-based categorization.

or content of the web pages in the collection. Figure 40 shows the description provided on the home page for Pakistan Floods collection at Archive-It. The description provided by the curators is detailed enough to explain the impact of the disaster to the user. However, it fails to inform the user about the cause of and measures taken to provide relief from the disaster. The absence of description for individual web pages and lack of categorization further adds to the difficulty in obtaining this information. This case study illustrates how wordles can provide more information about Pakistan floods by visually summarizing the web pages in this collection.

The web pages in the Pakistan Floods collection are not categorized into any group by the curator. Moreover most web pages do not have any title or description associated with the listed URI (Figure 41). This makes it difficult to quickly identify


Login
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)

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

More ▼

Narrow Your Results

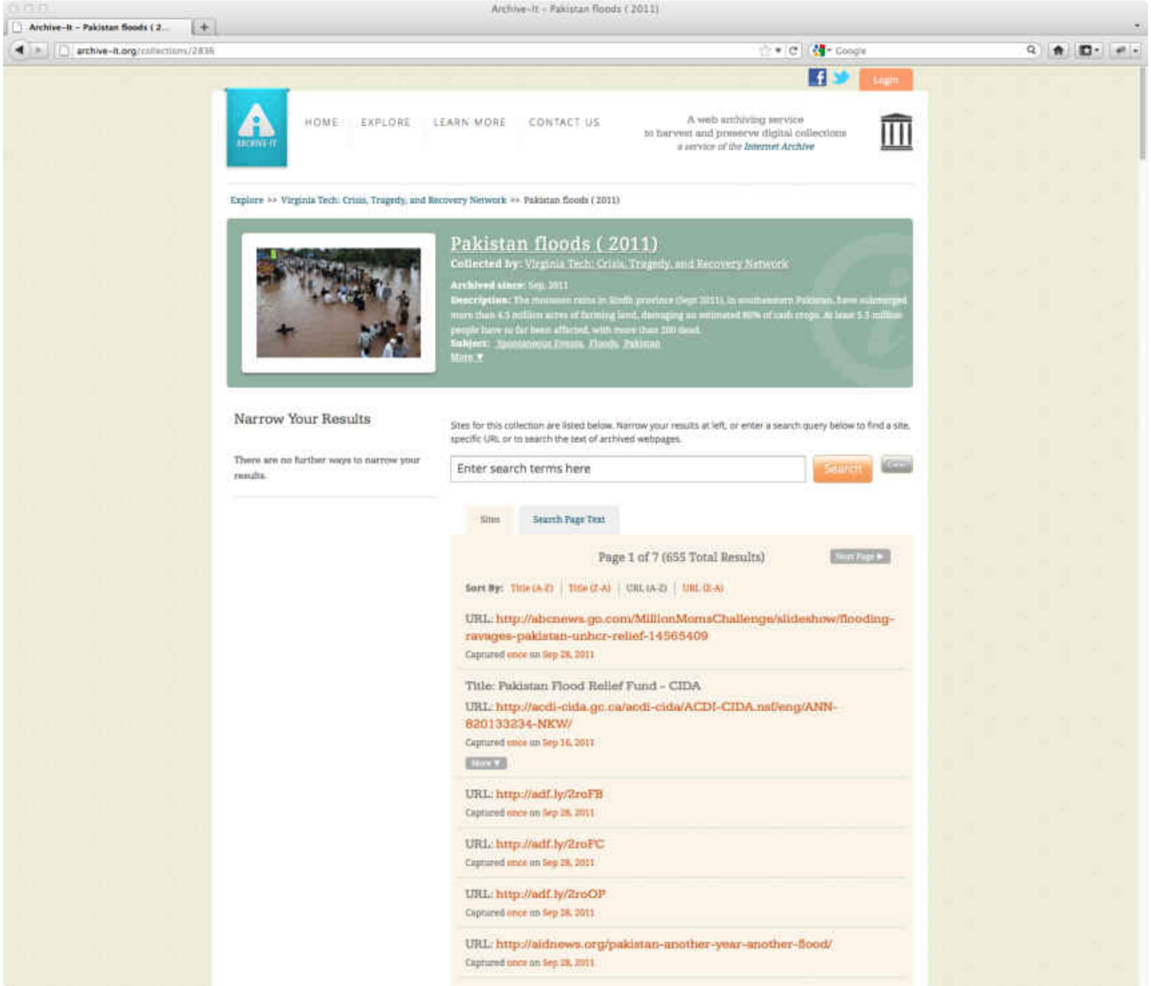
Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Fig. 40. Description for the Pakistan Floods collection as provided by the curator.

web pages relevant to the information the user is seeking. We thus applied our rule-based categorization to provide a naive classification prior to exploring the collection using image plot and wordle (Figure 39).

Figure 42 shows the wordle for a web page[1] under the News Websites category. Figure 43 shows a screenshot of the web page with key terms in the news article annotated within red boxes. The wordle provides a visual summary of the web page and highlights the following keywords accurately summarizing the news article on the web page: climate, change, pakistan, flood, disaster, himalaya, food water, sanitation, health. By studying this wordle, the user can infer that climatic changes in the Himalayan region is a probable cause for the floods. By associating the top terms with the web page title, *UN appeals for Pakistan flood aid*, the user can also infer that the flood aid includes providing food, water, sanitation and health facilities to the victims. Thus using the wordle alone, the user can quickly infer key points discussed in the web page without visiting it. This not only saves time but also viewing wordles for different web pages can help the user to quickly gain insight into

[1] UN Appeals for Pakistan flood aid - http://chimalaya.org/2011/09/21/un-appeals-for-pakistan-flood-aid/



A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)

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 5.3 million people have so far been affected, with more than 200 dead.

Subject: Spontaneous Events, Floods, Pakistan

More ▼

Narrow Your Results

There are no further ways to narrow your results.

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.

Page 1 of 7 (655 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409
Captured once on Sep 28, 2011

Title: Pakistan Flood Relief Fund - CIDA
URL: http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/
Captured once on Sep 16, 2011

URL: http://adf.ly/2roFB
Captured once on Sep 28, 2011

URL: http://adf.ly/2roFC
Captured once on Sep 28, 2011

URL: http://adf.ly/2roOP
Captured once on Sep 28, 2011

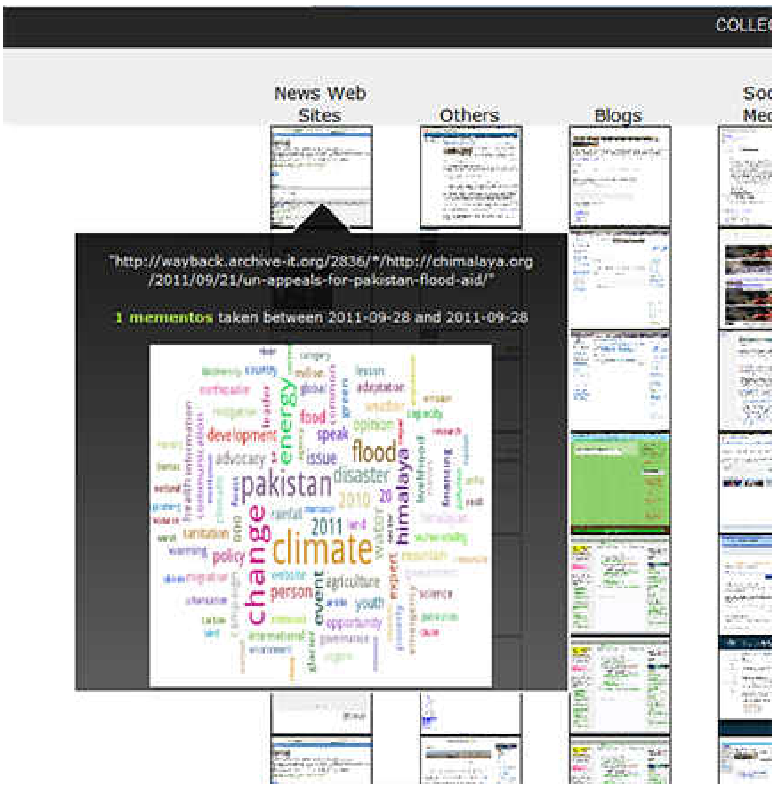
URL: http://aidnews.org/pakistan-another-year-another-flood/
Captured once on Sep 28, 2011

Fig. 41. Pakistan Floods (2011) collection at Archive-It. Note the missing title and description for the URIs listed in the collection.

the various aspects of the disaster.

Figure 44 displays a wordle for a web page[2] under the Others category. The important terms highlighted in this wordle are: canadian, charity, pakistan, flood, donation, fund, humanitarian, response. Similar to the previous example, this wordle is also an accurate summary of the content of the web page. By observing this wordle, the user can infer that Canadian charities have seeked donations to provide relief to the flood victims of Pakistan.

[2] Pakistan Flood Relief Fund CIDA - http://wayback.archive-it.org/2836/*/http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/



"http://wayback.archive-it.org/2836/*/http://chimalaya.org /2011/09/21/un-appeals-for-pakistan-flood-aid/"

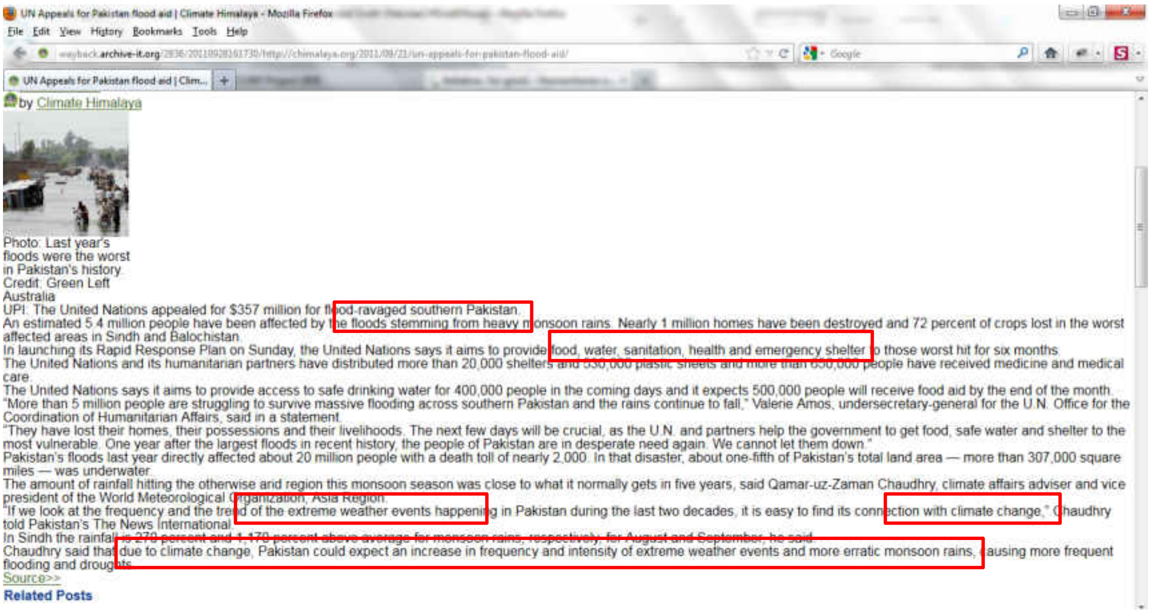
1 mementos taken between 2011-09-28 and 2011-09-28

Fig. 42. Wordle for the web page titled “UN Appeals for Pakistan flood aid” under News category.

Figure 45 shows a wordle for a web page[3] under the Blogs category. This wordle highlights the following terms: austrailium - stemming of the world Australia prior to generation of TF and TF-IDF scores is responsible for this, relief, 5 million - as separate words, food, water, sanitation, pakistan, sindh. This wordle again is an effective summary of the web page and highlights the relief efforts of the Australian government. By aggregating the insight gained from wordles created for various web pages, the user can infer the cause and impact of the disaster, and the steps taken to provide relief to the victims.

The above examples demonstrate how wordles can efficiently summarize the key concepts in a web page. By studying wordles for various web pages in the collection

[3] Australia responds to the monsoon flood disaster in Pakistan - http://wayback.archive-it.org/2836/*/http://foreignminister.gov.au/releases/2011/kr_mr_110926.html?utm_source=twitterfeed&utm_medium=twitter


UN Appeals for Pakistan flood aid | Climate Himalaya - Mozilla Firefox
File Edit View History Bookmarks Tools Help
wayback.archive-it.org/2836/20110928161730/http://chimalaya.org/2011/09/21/un-appeals-for-pakistan-flood-aid/
Google
UN Appeals for Pakistan flood aid | Clim...

by Climate Himalaya

Photo: Last year's floods were the worst in Pakistan's history. Credit: Green Left Australia

UPI: The United Nations appealed for $357 million for flood-ravaged southern Pakistan.
An estimated 5.4 million people have been affected by the floods stemming from heavy monsoon rains. Nearly 1 million homes have been destroyed and 72 percent of crops lost in the worst affected areas in Sindh and Balochistan.
In launching its Rapid Response Plan on Sunday, the United Nations says it aims to provide food, water, sanitation, health and emergency shelter to those worst hit for six months.
The United Nations and its humanitarian partners have distributed more than 20,000 shelters and 530,000 plastic sheets and more than 650,000 people have received medicine and medical care.
The United Nations says it aims to provide access to safe drinking water for 400,000 people in the coming days and it expects 500,000 people will receive food aid by the end of the month.
"More than 5 million people are struggling to survive massive flooding across southern Pakistan and the rains continue to fall," Valerie Amos, undersecretary-general for the U.N. Office for the Coordination of Humanitarian Affairs, said in a statement.
"They have lost their homes, their possessions and their livelihoods. The next few days will be crucial, as the U.N. and partners help the government to get food, safe water and shelter to the most vulnerable. One year after the largest floods in recent history, the people of Pakistan are in desperate need again. We cannot let them down."
Pakistan's floods last year directly affected about 20 million people with a death toll of nearly 2,000. In that disaster, about one-fifth of Pakistan's total land area — more than 307,000 square miles — was underwater.
The amount of rainfall hitting the otherwise arid region this monsoon season was close to what it normally gets in five years, said Qamar-uz-Zaman Chaudhry, climate affairs adviser and vice president of the World Meteorological Organization, Asia Region.
"If we look at the frequency and the trend of the extreme weather events happening in Pakistan during the last two decades, it is easy to find its connection with climate change," Chaudhry told Pakistan's The News International.
In Sindh the rainfall is 270 percent and 1,170 percent above average for monsoon rains, respectively, for August and September, he said.
Chaudhry said that due to climate change, Pakistan could expect an increase in frequency and intensity of extreme weather events and more erratic monsoon rains, causing more frequent flooding and droughts.
Source>>
Related Posts

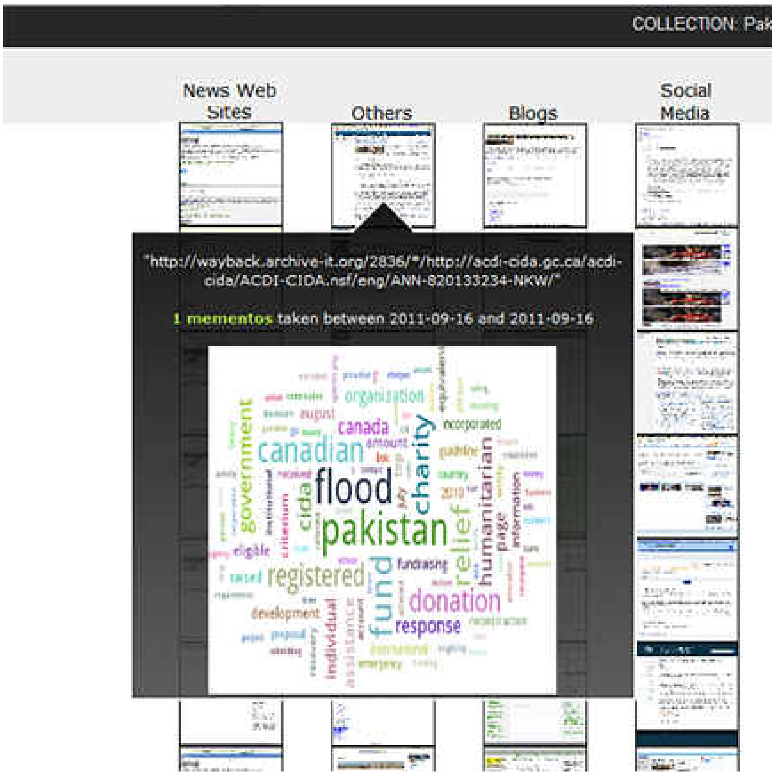
Fig. 43. Archived version of web page titled "UN Appeals for Pakistan flood aid".

a user can quickly grasp the synopsis of the collection, a feature absent in the textual interface.

V.4 THEME TRACKING

As web pages are added to the collection and mementos are captured, the theme of the collection changes over time due to introduction of new content. We have already discussed how the collection synopsis can be obtained by visualizing the content of web pages in the collection as wordles. However, wordles fail to track the change in the theme of a collection over time. Time clouds use the temporal aspect to create visualizations and thus can be used to depict a change in the content (and thus the theme) of an archived collection over time.

Figures 46 and 47 depict time clouds for the North Carolina State Government Web Site collection, archived between 1999 and 2011. Each figure displays the time clouds created using top terms appearing in the web pages archived during a timespan of 1 year, starting from 1997. Figure 46 shows 6 time clouds, 1 for each year, between 1999 and 2003. Figure 47 shows 8 time clouds, 1 for each year, between 2003 and 2011.



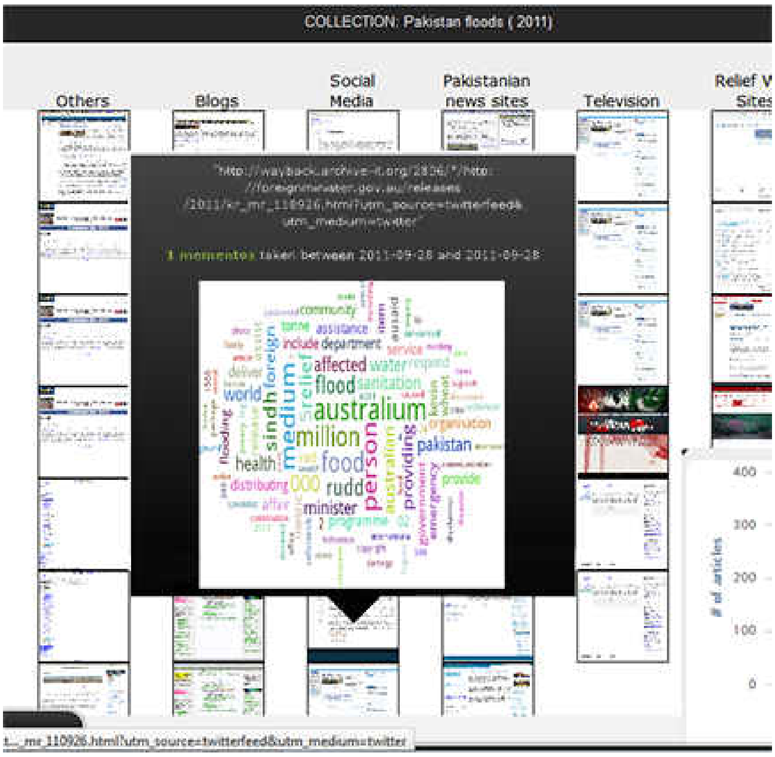
"http://wayback.archive-it.org/2836/*/http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/"

1 mementos taken between 2011-09-16 and 2011-09-16

Fig. 44. Wordle for the web page titled "Pakistan Flood Relief Fund CIDA" under Others category.

Below each time cloud, the year is marked along with the primary and secondary themes of the web pages (described later) archived during that year.

Each time cloud contains two word clouds, one created using top 20 terms based on TF score and the other using top 20 terms based on TF-IDF score. While TF can be directly used to find the most frequently occurring terms, TF-IDF is used to identify less frequently occurring, but potentially more relevant terms. The relative weight of terms appearing in each cloud is used to determine the primary and secondary theme of the collection. The terms *north* and *carolina* appear in every tag cloud, re-enforcing the fact that the web pages in the collection aptly belong to the North Carolina State Government Websites collection. For this reason, we ignore them while determining primary and secondary themes from the tag clouds. Also, we do not consider numbers for theme determination. Since we perform stemming of all terms before calculating their TF and TF-IDF scores, few words such as *business*



COLLECTION: Pakistan floods (2011)

"http://wayback.archive-it.org/2836/*/http://foreignminister.gov.au/releases/2011/kr_mr_110926.html?utm_source=twitterfeed&utm_medium=twitter"

1 mementos taken between 2011-09-28 and 2011-09-28

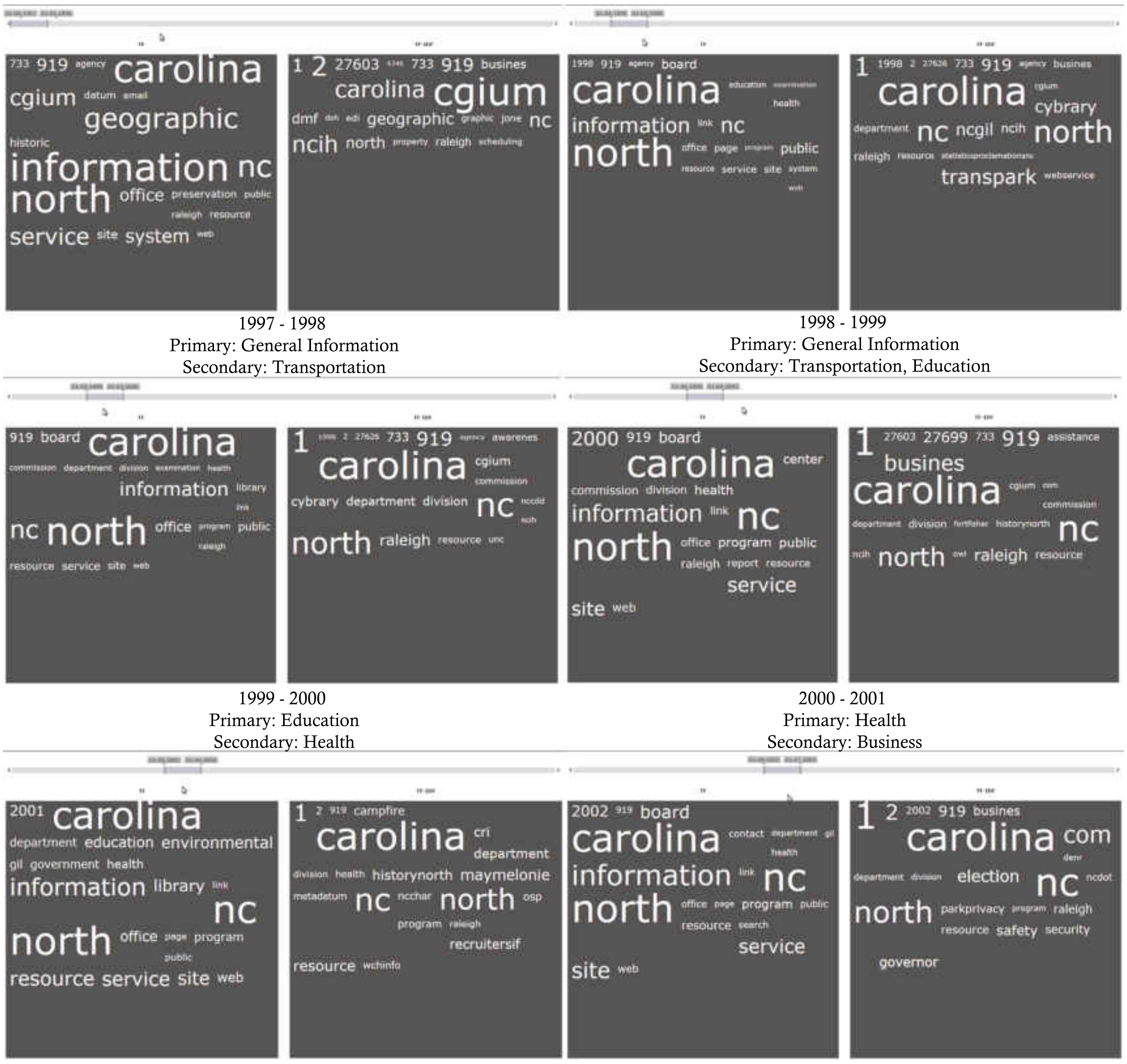
t..._mr_110926.html?utm_source=twitterfeed&utm_medium=twitter

Fig. 45. Wordle for the web page titled “Australia responds to the monsoon flood disaster in Pakistan” under Blogs category.

appear as *busines*, however, their effectiveness in theme determination is not affected.

Tables 4 and 5 tabulate the primary and secondary themes during different years and the terms determining them. It can be seen that the primary and secondary themes during the initial years of the collection tend to be unstable, varying between Health, Education, General Information, Transportation and Business. It is only after 2006 that the themes are stabilized in the collection, with little change. Another observation is the appearance of Election as a secondary theme during the years 2002-2004 and appearance of a related term “governor” during the years 2008-2011.

Time clouds are thus effective in tracking the changing theme of a collection as it develops over time. A user can leverage this information to selectively browse through the collection. Such an ability is not provided in the default textual interface of Archive-It.


733 919 agency carolina
cgium datum email
geographic
historic
information nc
north office preservation public
raleigh resource
service site system web
1 2 27603 733 919 busines
carolina cgium
dmf edi geographic graphic jone nc
ncih north property raleigh scheduling
1998 919 agency board
carolina education health
information link nc
north office page program public
resource service site system
web
1 1998 2 27626 733 919 agency busines
carolina cgium
cybrary
department nc ncgil ncih north
raleigh resource
transpark webservice
919 board carolina
commission department division examination health
information library
link
nc north office program public
raleigh
resource service site web
1 27626 733 919 agency awarenes
carolina cgium
commission
cybrary department division nc ncdid
ncih
north raleigh resource unc
2000 919 board
carolina center
commission division health
information link nc
north office program public
raleigh report resource
service
site web
1 27603 27699 733 919 assistance
busines
carolina cgium com
commission
department division historynorth nc
ncih north raleigh resource
2001 carolina
department education environmental
gil government health
information library link
nc
north office page program
public
resource service site web
1 2 919 campfire
carolina cri
department
division health historynorth maymelonie
metadatum nc ncchar north osp
program raleigh
recruitersif
resource wchinfo
2002 919 board
carolina contact department gil
health
information link nc
north office page program public
resource search
service
site web
1 2 2002 919 busines
carolina com
denr
department division election nc ncdot
north parkprivacy program raleigh
resource safety security
governor

1997 - 1998
Primary: General Information
Secondary: Transportation

1998 - 1999
Primary: General Information
Secondary: Transportation, Education

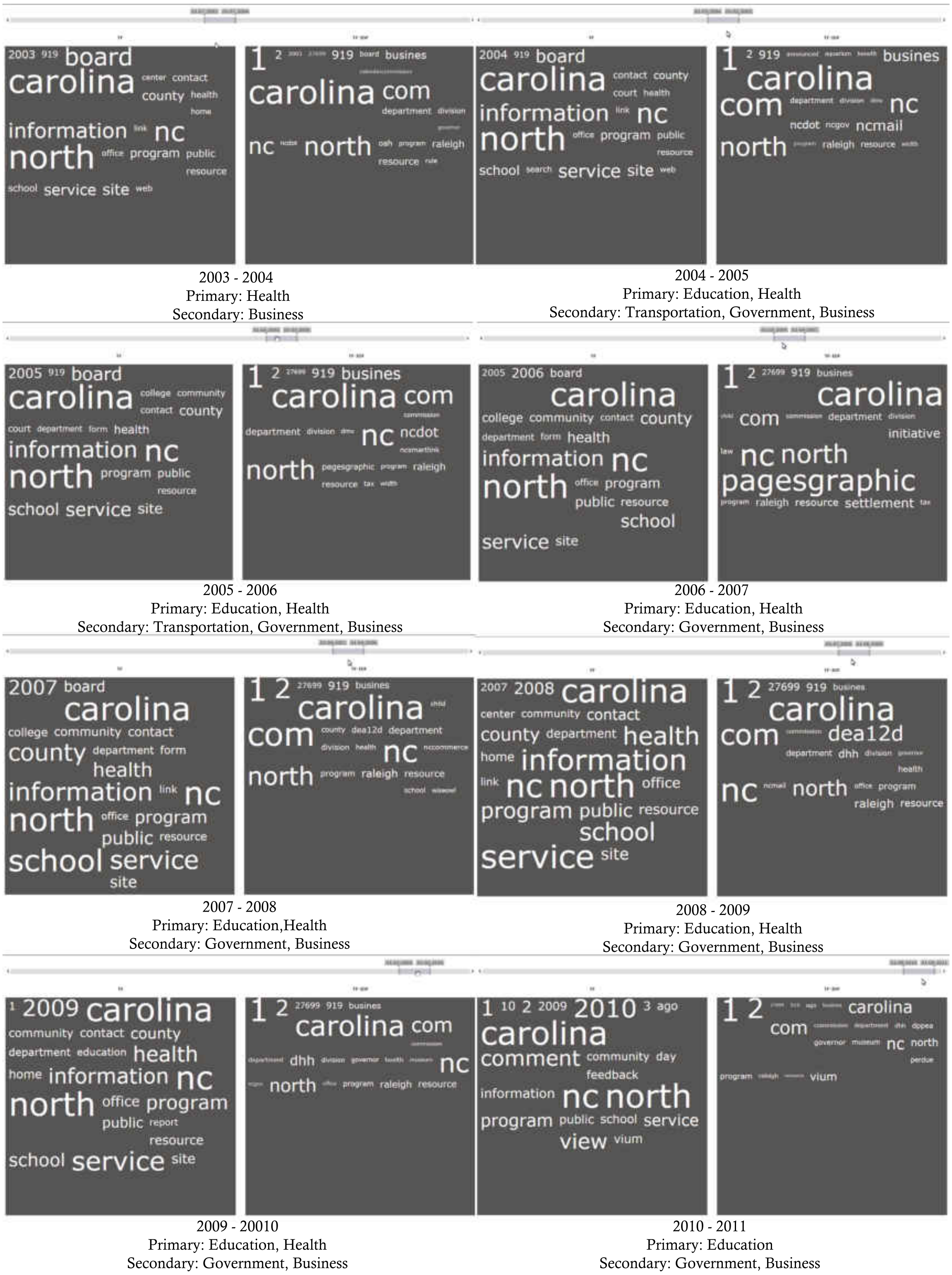
1999 - 2000
Primary: Education
Secondary: Health

2000 - 2001
Primary: Health
Secondary: Business

2001 - 2002
Primary: Education
Secondary: General Information

2002 - 2003
Primary: General Information
Secondary: Election

Fig. 46. Time clouds for the North Carolina State Government Web Site collection - 1997 to 2003.


2003 919 board carolina center contact county health home information link nc north office program public resource school service site web
1 2 2003 27699 919 board busines carolina com department division nc north osh program raleigh resource rule
2004 919 board carolina contact county court health information link nc north office program public resource school search service site web
1 2 919 busines carolina com department division nc ncdot ncgov ncmail north raleigh resource
2005 919 board carolina college community contact county court department form health information nc north program public resource school service site
1 2 27699 919 busines carolina com commission department division nc ncdot ncsmartlink north pagesgraphic program raleigh resource tax width
2005 2006 board carolina college community contact county department form health information nc north office program public resource school service site
1 2 27699 919 busines carolina child com commission department division initiative law nc north pagesgraphic program raleigh resource settlement tax
2007 board carolina college community contact county department form health information link nc north office program public resource school service site
1 2 27699 919 busines carolina child com county dea12d department division health nc nccommerce north program raleigh resource school
2007 2008 carolina center community contact county department health home information link nc north office program public resource school service site
1 2 27699 919 busines carolina com commission dea12d department dhh division health nc ncmail north office program raleigh resource
1 2009 carolina community contact county department education health home information nc north office program public report resource school service site
1 2 27699 919 busines carolina com commission department dhh division governor health museum nc north office program raleigh resource
1 10 2 2009 2010 3 ago carolina comment community day feedback information nc north program public school service view vium
1 2 carolina com department dhh dppea governor museum nc north perdue program raleigh vium

2003 - 2004
Primary: Health
Secondary: Business

2004 - 2005
Primary: Education, Health
Secondary: Transportation, Government, Business

2005 - 2006
Primary: Education, Health
Secondary: Transportation, Government, Business

2006 - 2007
Primary: Education, Health
Secondary: Government, Business

2007 - 2008
Primary: Education,Health
Secondary: Government, Business

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

2009 - 20010
Primary: Education, Health
Secondary: Government, Business

2010 - 2011
Primary: Education
Secondary: Government, Business

Fig. 47. Time clouds for the North Carolina State Government Web Site collection - 2003 to 2011.

TABLE 4
Primary and secondary themes appearing in the collection "North Carolina State Government Website Archive" between 1997-2005.

Year	Assigned Theme		Determining Terms
1997 - 1998	Primary	General Information	information, geographic, service, system
	Secondary	Transportaion	ncih
1998 - 1999	Primary	General Information	information, service, agency
	Secondary	Transporation, Eduation	ncih, transpark, education, examination, cybrary
1999 - 2000	Primary	Education	education, examination, cybrary
	Secondary	Health	health
2000 - 2001	Primary	Health	health, board, center
	Secondary	Business	business, commission, resource
2001 - 2002	Primary	Education	department, education, environmental, library
	Secondary	General Information	information, service, agency
2002 - 2003	Primary	General Information	information, service, agency
	Secondary	Election	election, governor
2003 - 2004	Primary	Education, Health	public, school, health
	Secondary	Business, Government	business, resource, governor
2004 - 2005	Primary	Education, Health	college, community, public, school, health
	Secondary	Business, Transportation, Government	business, ncdot, dmv, resource, tax

TABLE 5
Primary and secondary themes appearing in the collection "North Carolina State Government Website Archive" between 2005 - 2011.

Year	Assigned Theme		Determining Terms
2005 - 2006	Primary	Education, Health	college, community, public, school, health
	Secondary	Business, Transportation, Government	business, ncdot, dmv, ncsmartlink, resource, tax
2006 - 2007	Primary	Education, Health	college, community, public, school, health
	Secondary	Business, Government	business, resource, department, office, division, tax, law
2007 - 2008	Primary	Education, Health	college, community, public, school, health
	Secondary	Business, Government	business, resource, department, office, governor, division, dhh, office, raleigh
2008 - 2009	Primary	Education, Health	college, community, public, school, health
	Secondary	Business, Government	business, resource, department, office, governor, division, dhh, office, raleigh
2009 - 2010	Primary	Education, Health	college, community, public, school, health
	Secondary	Business, Government	business, resource, department, office, governor, division, dhh, office, raleigh, ncgov
2010 - 2011	Primary	Education	community, public, school
	Secondary	Government	resource, department, governor, dhh, raleigh

V.5 USER EVALUATION

Researchers at Columbia University Libraries[4] have archived multiple collections at Archive-It, including the Human Rights collection. As a part of this work, we contacted them to conduct an informal user evaluation of our interface. The researchers were asked to evaluate the various interfaces and provide their feedback on the following aspects:

- 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

We received a positive feedback for our treemap, image plot and timeline visualizations. Following are some comments from the feedback [45]:

> “I like the image plot display best, as it allows the best navigation down to captures, which is critical. I like having the screenshots, and the idea of combining the screenshots with mouseover word clouds from the site in question.”
>
> “The timeline display is quite interesting as it shows development of groups in our collection (we started with type x, then added type y, etc.), and I like being able to hover over specific points on the timeline and adjust the resolution of the time scales.”
>
> “Color coding for date range of capture is useful for quickly seeing newer additions to collection.”
>
> “I would say that the tree map interface and the image plot display would be preferable to the older Archive-It.org design of a list of links, but the redesigned

[4] http://library.columbia.edu/

> archiveit.org that allows for faceting by meta-data fields is much improved. Adding elements from this project like screenshots and color coding related to length of capture dates could be useful, as long as all faceted meta-data was still included."

We also received the following suggestions for improving our visualizations:

- Widen the time scale for treemaps to include larger timespans
- Expand the list of stop words to include stop words from languages other than English, like French
- Timeline should provide an option to drill down into individual captures

V.6 SUMMARY

This chapter discussed how different visualizations can be used to explore the collections with the help of four case studies. Each case study discussed the use of a certain visualization to gain deeper insight about the collection, which would not have been possible using the default textual interface. The use of rule-based categorization to group web pages in uncategorized collections, thus making it easier for the user to explore the collection, was demonstrated in this chapter. This chapter also discussed the feedback and suggestions received from our partners at Columbia University as a part of an informal user evaluation.

CHAPTER VI

FUTURE WORK

While we have demonstrated the effectiveness of visualization as an alternative technique for exploring Archive-It collections, a great deal of work remains. Exploring novel and more efficient visual representations for digital collections is our primary future work. We also intend to perform a formal user study to evaluate the effectiveness of our visualization framework, and integrate it with Archive-It and other archives as a part of our future work.

VI.1 N-GRAM WORDLES

We have implemented wordles to provide a visual summary of web pages in a collection and demonstrated their effectiveness in identifying the underlying theme of a collection. However these wordles are unigram wordles, where each term in the wordle consists of only one word. Such wordles though simple to create, may not effectively convey an underlying idea, theme or message due to the spatial separation of consecutive terms. Figure 42 (Chapter V) shows the wordle for a web page in the Pakistan Floods collection. Here the words "climate" and "change" appear as two individual terms, and are less effective than a bigram "climate change" in conveying the idea that climatic changes are a probable cause for the recent floods in Pakistan. We thus intend to study the effectiveness of n-gram wordles for summarizing web pages in a collection as a part of our future work.

VI.2 DICTIONARY BASED STEMMING

In this work we have implemented the Porter stemming algorithm, which is run on all terms appearing in a web page prior to TF and TF-IDF calculation. This algorithm is primarily a suffix stripping algorithm and attempts to reduce a term to its probable root term. While this allows for normalization of terms, it also results in incorrect conflations like “business” stemmed to “busines’, “australia” stemmed to “australiam”. Such incorrect conflations add unwanted noise to the set of terms appearing in a web page and decrease the accuracy in calculation of top terms based on TF and TF-IDF.

Dictionary based stemmers like the Krovetz Stemmer [46] on the other hand preserve the linguistic morphology and meaning of a word even after stemming. We plan to implement Krovetz stemming algorithm as a replacement for the Porter stemmer as a part of our future work.

VI.3 INTEGRATION WITH ARCHIVE-IT AND OTHER ARCHIVES.

Archive-It exposes its collections as WARC and the meta-data of its collections in the form of WAT files. However, for this proof of concept visualization framework we relied on web scraping for obtaining collection data and meta-data due to its ease of implementation and our need to generate screenshots of web pages. As our future work, we plan to directly use WARC and WAT files as our data source. We intend to develop a system that can read the WARC and WAT files and create visualizations with minimum effort, thus enabling us to visualize any archive that stores its data in these formats.

CHAPTER VII

CONCLUSIONS

Digital archives are crucial in preserving and providing access to user-defined sets of digital resources on the web. Traditionally users interact with these archives through a text-based interface optimized for searching, but lacking in efficient browsing capabilities. In this work we have implemented novel visualizations for Archive-It collections in order to provide an alternative interface for collection browsing.

We have identified the various metrics/properties which define an Archive-It collection and implemented visualizations that leverage these properties to create efficient visual representations of the collection. The different visualizations we have implemented are treemap, time cloud, bubble chart, image plot, timeline and wordle. While treemaps are an efficient visualization for viewing the collection hierarchy, the bubble charts are a simple statistical representation of the collection. The time cloud and timeline visualizations allow the user to visualize the temporal aspect of the digital collections. Time clouds are an efficient way of visualize the change in collection theme over time, and the timeline visualization allows the user to gain insight about the collection structure and how it changes over time. Image plots represent each web page in the collection by its screenshot, there by allowing users to visually browse through the collection. Wordles show the user the summarized content of a web page in the form of a word cloud, thus providing the user with a quick visual summary of the web page.

While creating a collection at Archive-It, curators may group web pages into categories which allow for easier filtering and browsing. However, many collections lack such grouping making it very cumbersome to find related web pages. For such collections, we introduce rule-based categorization which provides structure in unstructured collections. For each of the visualization techniques and for the rule based categorization, we have demonstrated their effectiveness in browsing and gaining deeper insight into the collection.

BIBLIOGRAPHY

[1] "The size of the World Wide Web (The Internet)." [Online]. Available: http://www.worldwidewebsize.com/

[2] S. Lawrence, F. Coetzee, E. Glover, D. Pennock, G. Flake, F. Nielsen, B. Krovetz, A. Kruger, and L. Giles, "Persistence of Web references in scientific research," *IEEE Computer*, vol. 34, no. 2, pp. 26–31, 2001.

[3] R. Weiss, "On the Web, Research Work Proves Ephemeral," *Washington Post*, p. A08, 24 November 2003.

[4] "Library of Congress Web Archives." [Online]. Available: http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[5] S. Bailey and D. Thompson, "UKWAC: Building the UK's first public web archive," *D-Lib Magazine*, vol. 12, no. 1, 2006. [Online]. Available: http://dx.doi.org/10.1045/january2006-thompson

[6] W. Cathro, C. Webb, and J. Whiting, "Archiving the Web: The PANDORA Archive at the National Library of Australia," in *Proceedings of the Preserving the Present for the Future Web Archiving Conference*, June 2001, pp. 105–118.

[7] C. Lampos, M. Eirinaki, D. Jevtuchova, and M. Vazirgiannis, "Archiving the Greek Web," in *Proceedings of the 4th International Web Archiving Workshop (IWAW'04)*, September 2004.

[8] A. Arvidson, K. Persson, and J. Mannerheim, "The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages," in *Proceedings of the 66th IFLA Council and General Conference*, August 2000. [Online]. Available: http://www.ifla.org/IV/ifla66/papers/154-157e.htm

[9] S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati, "A First Experience in Archiving the French Web," in *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, 2002, pp. 1–15. [Online]. Available: http://dx.doi.org/10.1007/3-540-45747-X_1

[10] "California Digital Library." [Online]. Available: http://www.cdlib.org/

[11] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke, “Webbase: A repository of web pages,” *Computer Networks*, vol. 33, no. 16, pp. 277 – 293, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128600000633

[12] “Archive-It - Learn More.” [Online]. Available: http://archive-it.org/learn-more

[13] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar, “Memento: Time travel for the web,” Tech. Rep. arXiv:0911.1112v2, November 2009. [Online]. Available: http://arxiv.org/abs/0911.1112

[14] “Internet Archive’s Wayback Machine.” [Online]. Available: http://wa.archive.org/aroundtheworld/index.new.html

[15] K. Padia, Y. AlNoamany, and M. C. Weigle, “Visualizing Digital Collections at Archive-It,” in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, June 2012, pp. 15–18. [Online]. Available: http://doi.acm.org/10.1145/2232817.2232821

[16] C. Ahlberg and B. Shneiderman, “Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994, pp. 313–317. [Online]. Available: http://doi.acm.org/10.1145/191666.191775

[17] J. Stasko, C. Görg, and Z. Liu, “Jigsaw: Supporting Investigative Analysis through Interactive Visualization,” *Information Visualization*, vol. 7, no. 2, pp. 118–132, April 2008. [Online]. Available: http://dx.doi.org/10.1145/1466620.1466622

[18] S. Havre, B. Hetzler, and L. Nowell, “ThemeRiver: Visualizing Theme Changes over Time,” in *Proceedings of the IEEE Symposium on Information Vizualization*, 2000, pp. 115–123. [Online]. Available: http://dx.doi.org/10.1109/INFVIS.2000.885098

[19] E. Reid, J. Qin, Y. Zhou, G. Lai, M. Sageman, G. Weimann, and H. Chen, “Collecting and Analyzing the Presence of Terrorists on the Web: A Case Study of Jihad Websites,” in *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, 2005, pp. 402–411. [Online]. Available: http://dx.doi.org/10.1007/11427995_35

[20] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "TIARA: A Visual Exploratory Text Analytic System," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 153–162. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835827

[21] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K. Yee, "Finding the Flow in Website Search," *ACM Communications*, vol. 45, no. 9, pp. 42–49, Sep. 2002. [Online]. Available: http://doi.acm.org/10.1145/567498.567525

[22] A. Jatowt, Y. Kawai, and K. Tanaka, "Visualizing Historical Content Of Web Pages," in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 1221–1222. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367736

[23] "Microsoft Pivot." [Online]. Available: http://www.microsoft.com/silverlight/pivotviewer/

[24] M. Whitelaw, "Visualising the Digital Archive: the Visible Archive project," *Archives and Manuscripts*, vol. 37, pp. 22–40, 2009.

[25] A. Scharnhorst, O. ten Bosch, and P. Doorn, "Looking at a digital research data archive - Visual interfaces to EASY," Tech. Rep. arXiv:1204.3200v1, 2012. [Online]. Available: http://arxiv.org/abs/1204.3200

[26] B. Johnson and B. Shneiderman, "Tree-Maps: A Space-filling Approach to the Visualization of Hierarchical Information Structures," in *Proceedings of the 2nd IEEE Conference on Visualization '91*, 1991, pp. 284–291. [Online]. Available: http://dx.doi.org/10.1109/VISUAL.1991.175815

[27] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds," in *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, 2007, pp. 995–998. [Online]. Available: http://doi.acm.org/10.1145/1240624.1240775

[28] M. A. Hearst and D. Rosner, "Tag Clouds: Data Analysis Tool or Social Signaller?" in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, 2008, pp. 160–162. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2008.422

[29] J. Feinberg, "Wordle," in *Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice)*, 1st ed. O'Reilly Media, 2010, pp. 37–58.

[30] F. B. Viegas, M. Wattenberg, and J. Feinberg, "Participatory Visualization with Wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1137–1144, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2009.171

[31] Federal Deposit Insurance Corporation, "FDIC Outlook, Winter 2003." [Online]. Available: http://www.fdic.gov/bank/analytical/regional/ro20034q/na/t4q2003.pdf

[32] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[33] S. Robertson, "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[34] *ISO 28500:2009 Information and Documentation – WARC File Format.* [Online]. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

[35] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic, "Introduction to Heritrix, an Archival Quality Web Crawler," in *Proceedings of the 4th International Web Archiving Workshop (IWAW04)*, September 2004.

[36] M. Burner, "Crawling towards Eternity - Building An Archive of The World Wide Web," *Web Techniques*, vol. 2, no. 5, pp. 37–40, May 1997. [Online]. Available: http://www.webtechniques.com/archives/1997/05/burner/

[37] "Web Archive Transformation (WAT) Specification, Utilities, and Usage Overview." [Online]. Available: https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Metadata+File+Specification

[38] "Document Object Model (DOM)." [Online]. Available: http://www.w3.org/DOM/

[39] "Onix Text Retrieval Toolkit." [Online]. Available: http://www.lextek.com/manuals/onix/index.html

[40] M. F. Porter, "An Algorithm for Suffix Stripping," in *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., 1997, pp. 313–316.

[41] M. Klein and M. L. Nelson, "Correlation of Term Count and Document Frequency for Google N-Grams," in *Proceedings of the 31st European Conference on Information Retrieval*, April 2009, pp. 620–627. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-00958-7_58

[42] M. Klein, "Document Frequencies from Yahoo search." [Online]. Available: http://blanche-04.cs.odu.edu/cgi-bin/getdf/getdf.cgi

[43] M. Kan and H. O. N. Thi, "Fast Webpage Classification using URL Features," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 325–326. [Online]. Available: http://doi.acm.org/10.1145/1099554.1099649

[44] M. Costa and M. J. Silva, "Understanding the Information Needs of Web Archive Users," in *Proceedings of the 10th International Web Archiving Workshop*, September 2010, pp. 9–16.

[45] A. Thurman, "Personal communication," January 2012.

[46] R. Krovetz, "Viewing Morphology as an Inference Process," in *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1993, pp. 191–202. [Online]. Available: http://doi.acm.org/10.1145/160688.160718

APPENDIX A

LIST OF STOP WORDS

Below is a list of words ignored for TF calculation.

a able about above across after again against all almost alone along already also although always am among an and another any anybody anyone anything anywhere are area areas around as ask asked asking asks at away b back backed backing backs be became because become becomes been before began behind being beings best better between big both but by c came can cannot case cases certain certainly clear clearly come could d dear did differ different differently do does done down downed downing downs during e each early either else end ended ending ends enough even evenly ever every everybody everyone everything everywhere f face faces fact facts far felt few find finds first for four from full fully further furthered furthering furthers g gave general generally get gets give given gives go going good goods got great greater greatest group grouped grouping groups h had has have having he her here hers herself high higher highest him himself his how however i if important in interest interested interesting interests into is it its itself j just k keep keeps kind knew know known knows l large largely last later latest least less let lets like likely long longer longest m made make making man many may me member members men might more most mostly mr mrs much must my myself n necessary need needed needing needs neither never new newer newest next no nobody non noone nor not nothing now nowhere number numbers o of off often old older oldest on once one only open opened opening opens or order ordered ordering orders other others our out over own p part parted parting parts per perhaps place places point pointed pointing points possible present presented presenting presents problem problems put puts q quite r rather really right room rooms s said same saw say says second seconds see seem seemed seeming seems sees several shall she should show showed showing shows side sides since small smaller smallest so some somebody someone something somewhere state states still such sure t take taken than that the their them then there therefore these they thing things think thinks this those though thought thoughts three through thus tis to today together too took toward turn turned turning turns twas two u under until up upon us use used uses v very w want wanted wanting wants was way ways we well wells went were what when where whether which while who whole whom whose why will with within without work worked working works would x y year years yet you young younger youngest your yours z

VITA

Kalpesh Padia
Department of Computer Science
Old Dominion University
Norfolk, VA 23529

EDUCATION

- M.S., 2012, Computer Science, Old Dominion University.
- B.E., 2009, Computer Science, Visvesvaraya Technological University, INDIA.

PUBLICATIONS

1. K. Padia, Y. AlNoamany, M.C. Weigle, "Visualizing Digital Collections at Archive-It", *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL).* Washington D.C., June 2012.
2. G.A. Vikas, K. Padia, H.S. Iyer, V.R. Darshan, N.P. Ganesh Prasad, A. Srinivas, "A Dynamic GPS-Free Localization Technique using Progressive Interpolation", *Proceedings of World Congress on Science, Engineering and Technology.* Singapore, August 2009.
3. K. Padia, G.A. H.S. Iyer, V.R. Darshan, N.P. Ganesh Prasad, A. Srinivas, "A localization algorithm for a GPS-free system with static parameter tuning", *Proceedings of ICCNT.* Chennai, India, July 2009, pp. 37-41.
4. K. Navuluri, K. Padia, A. Gupta and T. Nadeem, "What's on your mind? A mind-based driving alert system", *Proceedings of the 9th international conference on Mobile systems, applications, and services, MobiSys.* Washington D.C., June 2011.

PROFESSIONAL ACTIVITIES

- Reviewer for Joint Conference on Digital Libraries, 2012
- Member, IEEE, 2008–Present.

Typeset using LaTeX.