Storytelling for Summarizing Collections in Web Archives

Yasmin AlNoamany
Old Dominion University

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Follow this and additional works at: http://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons

Recommended Citation
http://digitalcommons.odu.edu/computerscience_presentations/3

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
Storytelling for Summarizing Collections in Web Archives

Yasmin AlNoamany
Michele C. Weigle
Michael L. Nelson

Old Dominion University
Web Science and Digital Libraries Group
@WebSciDL

This work is supported in part by IMLS LG-71-15-0077

CNI Spring 2016
2016-04-05
IMLS-Funded Research

1. Use small “stories” to summarize much larger collections of archived web pages
   - big $\rightarrow$ small

2. Generate web archive collections by mining user-generated stories for seed URIs
   - small $\rightarrow$ big

Archive-It, a subscription-based service, hosts curated web collections

- > 3,000 collections
- > 400 partners
- > 10B archived pages
The group that the resource belongs to

Collection title

Collection categorization according to the curator

Seed URI

Timespan of the resource and the number of times it has been captured

Metadata about the collection

Text search box

List of the seed URIs
Problem:
Collection understanding and collection summarization are not currently supported

Not easy to answer “what’s in that collection?”
There is more than one collection about the Egyptian Revolution

- “2010-2011 Arab Spring” https://archive-it.org/collections/3101
- “Egypt Revolution and Politics” https://archive-it.org/collections/2358
(1000s of Seeds X 1000s of Mementos) + Dimension of Time == Conventional Vis Methods Not Applicable

Using Timelines, Treemaps, etc.:
http://ws-dl.blogspot.com/2012/08/2012-08-10-ms-thesis-visualizing.html
Idea: Storytelling
Stories in Literature

Once upon a time…

Story elements: setting, characters, sequence, exposition, conflict, climax, resolution

http://www.learner.org/interactives/story/
Stories in social media

“It's hard to define a story, but I know it when I see it” (Alexander, 2008)

A sampling and arrangement of web resources for summarization.
Collection == thematic sample from the Web
Story == arranged sample from the collection

We sample $k$ mementos from $N$ pages of the collection to create a summary story
Collections have two dimensions

**Time**

**URI**

- **R1n**
- **R2n**
- **R3n**
- **R4n**
- **R5n**
- **R6n**

- **R12**
- **R22**
- **R32**
- **R42**
- **R52**
- **R62**

- **R11**
- **R21**
- **R31**
- **R41**
- **R51**
- **R61**

- **R1**
- **R2**
- **R3**
- **R4**
- **R5**
- **R6**

- **t1**
- **t2**
- **t3**
- **t4**
- **t5**
- **t6**
Fixed Pages, Fixed Time

R1

R1

R1

R1

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_6 \]
Fixed Page, Fixed Time

Andriod Chrome user-agent

A desktop Chrome user-agent

Sliding Page, Fixed Time

R4

R3

R2

R1

t1  t2  t3  t4  t5  t6
Feb. 11, 2011
Mubarak resigns
Sliding Page, Sliding Time
What do stories in Storify look like?

“Characteristics of Social Media Stories”, TPDL 2015
What is the length of a story (the number of resources per story)?

- This story has 31 resources.
What are the types of resources that compose a story?

• This story has
  – 19 quotes
  – 8 images
  – 4 videos
What are the most frequently used domains?

- This story uses:
  - 90% twitter.com
  - 7% instagram.com
  - 3% facebook.com
What differentiates a popular story?

Va. reporter, cameraman killed during live broadcast

WDBJ7's photographer Adam Ward and reporter Alison Parker were killed Aug. 26, 2015. Updates from social media that morning:

19,795 views

Here's my Facebook movie. Find yours at https://facebook.com/lookback/

#FacebookIs10

lookback in Facebook's 10th Birthday

Reactions and coverage of lookback...

64 views
(skipping many details, see TPDL 2015 paper)
We should create stories with:

• ~28 pages
• moar images!
• where possible, select pages from social media, news, blogs
• additional dimensions of quality:
  – are well archived (e.g., not missing images, stylesheets)
  – generate nice summaries in the Storify interface
Stories from collections about the Egyptian Revolution

https://storify.com/yasmina85/auto-stories-from-archived-collections-56fbc3d1b8d27c6f6571c647
https://storify.com/yasmina85/auto-stories-from-archived-collections-5702ff8f228eede273d49c21
https://storify.com/yasmina85/auto-stories-from-archived-collections-5702c7f1228eede273d48ddf
Evaluation: can humans tell human generated stories from machine generated?

https://storify.com/yasmina85/this-is-manually-generated-story-from-archive-it-c-56b25ae72c0664474ee34f13
https://storify.com/yasmina85/auto-stories-from-archived-collections-56f1cfd36bc660f47f1b9f5e
Use an interface people *already know how to use* to summarize collections.

**Storytelling services**
- curata
- Storify
- PearlTrees
- Scoop.it
- Bundlr
- Pinterest
- Roojooin
- Memolane
- Paper.li
- Snip.it

**Archived collections**

**Archived enriched stories**

more info:
https://github.com/yasmina85/OffTopic-Detection