I Know it's Important, But What Am I Looking at? Strategies for using Blog Content to Contextualize YouTube Videos

Christopher (Cal) Lee
School of Information and Library Science
University of North Carolina, Chapel Hill

Society of American Archivists Research Forum
August 26, 2008
San Francisco, CA

Web is Vital Forum of Deliberation

- Election of public officials as example - likely to be strongly influenced by materials on Web
- Materials include various forms of textual documents & rapidly growing body of audio, images, video
- To make sense of electoral process, future researchers could benefit from:
  - Perpetual access to Web materials
  - Sufficient contextual information to make meaningful use & sense of them

Hans Booms & Mirror of Society

- Appraisal should be based on best (i.e. most informed by empirical evidence) judgments about what members of society judged most valuable or important at time documents were created
- Use various data sources to determine what is currently most influential, viewed, discussed & cited

The “Social Web” (e.g. YouTube, blogs) – What Literature Says

- Important source for documenting online deliberation about major processes & events
- Some sites & items will have major impact on events, perceptions & behaviors
- Significant variance in focus, depth of content & degree of exposure (traffic & in-links)

Appraisal, Selection & Capture of Web Materials

- Fundamental challenge: determining what to collect & preserve
- Need tools & methods for combining information from queries & crawls to identify & collect materials that document & contextualize socially important phenomena

VidArch Project

- Funded through NDIIPP (Library of Congress & National Science Foundation)
- Exploring strategies & building tools for appraisal & description of online digital video
- Capturing YouTube videos & web pages associated with 2008 U.S. presidential election & various other topics

VidArch Approach

- Text queries (e.g. “John Edwards”) as basis for crawls of YouTube, then use data from YouTube & elsewhere (e.g. blogs linking to videos, in-links identified by Web search engines) to:
  - Inform the appraisal of YouTube videos
  - **Collect further contextual information (focus of today’s talk)**

I Know It's Important, But What am I Looking at?

- Many YouTube videos inspire great deal of online discussion & attention
- Often very difficult to understand "what you're looking at" solely based on YouTube page itself
- Archival description may involve **capturing online discussion** (e.g. sampling from blogosphere), rather than archivist being primary creator of description

An Example

Sign Up | QuickList (0) | Help | Sign In | Site:

YouTube
Broadcast Yourself™

Home | Videos | Channels | Community

Videos | Search | advanced | Upload

Vote Different

Rate: ★★★★☆ 12,058 ratings

Views: 5,268,816

From: ParkRidge47
Joined: 1 year ago
Videos: 3

Subscribe

Added: **March 05, 2007** (More info)

Make up your own mind. Decide for yourself who ...

Embed: Customize

<object width="425" height="344"><param name="movie" value="http:

More From: ParkRidge47

Related Videos

Barack Obama Hillary Clinton - Umbrella
01:56 From: wolf084
Views: 11,179,757

The Shocking Video Hillary Does NOT Want You To See! (1of2)
10:28 From: NufffRespect
Views: 3,401,587

Obama Girl Returns for Iowa! (Why Obama Won)
02:19 From: barelypolitical
Views: 2,451,439

Team for Study Reported Today

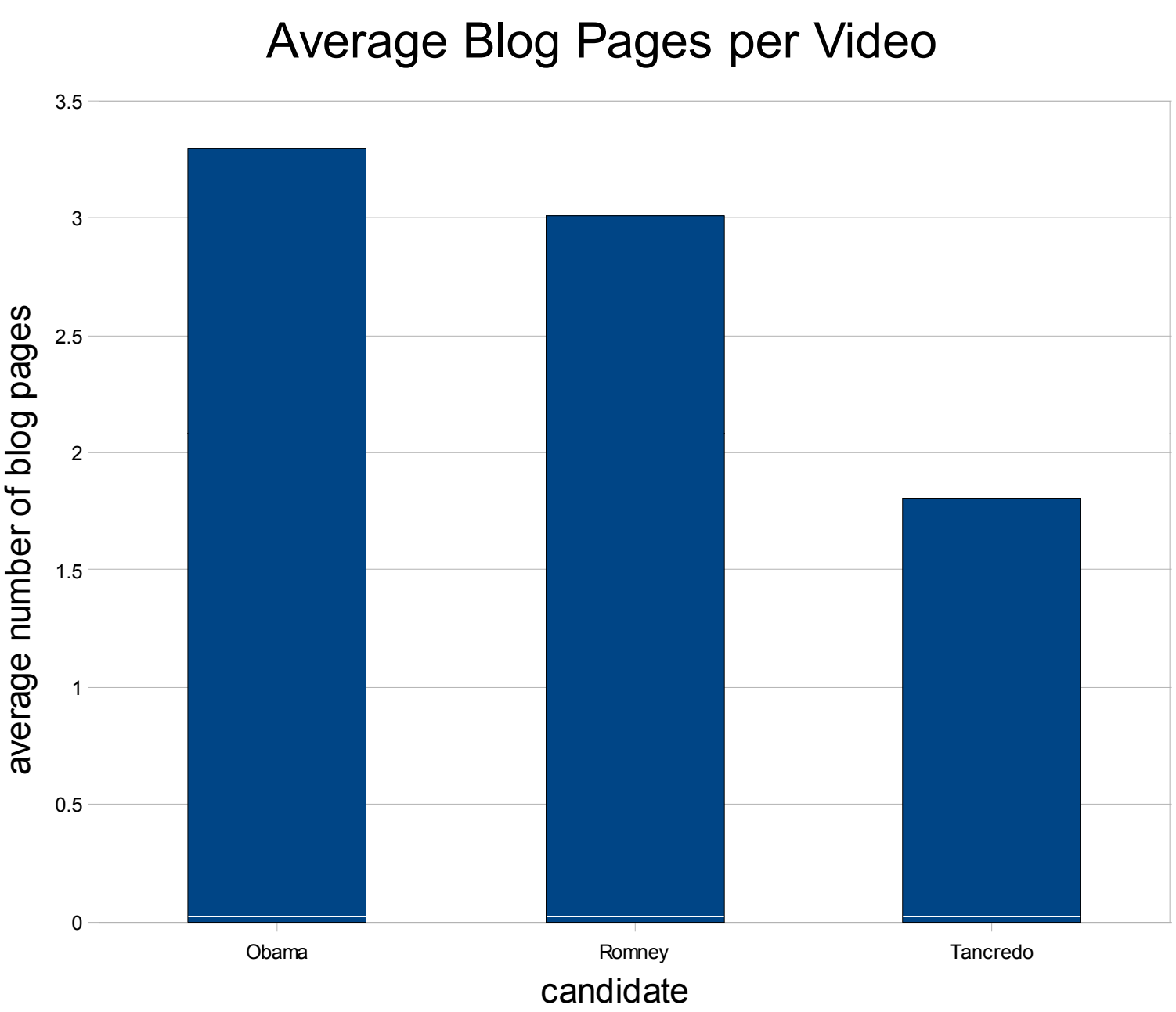
- Rob Capra, Postdoctoral Fellow at UNC-CH
- Rachael Clemens, doctoral student at UNC-CH
- Cal Lee, Assistant Professor at UNC-CH
- Laura Sheble, doctoral student at UNC-CH

What We Did

1. Crawled YouTube for videos about presidential candidates (based on YouTube relevance ranking)
2. Searched Technorati & Google Blog Search for blog pages that mention presidential candidates
3. Generated subset of blog pages from #2 that linked to at least one video from #1

What We Did

- Examined & coded statistical sample of blog pages for four candidates: Obama, Romney, Trancredo, Vilsack
  - Four separate coders
  - Four rounds of preliminary coding to clarify coding categories & ensure inter-code reliability
  - Two coders per blog page


Average Blog Pages per Video
average number of blog pages
3.5
3
2.5
2
1.5
1
0.5
0
Obama
Romney
Tancredo
candidate

How We Coded Each Blog Entry

Watch & Code Video to Which Blog Entry Links

- Is video about the candidate?
  - 3 = about the candidate
  - 2 = about election but not candidate
  - 1 = not about candidate or election

Examine & Code Blog Entry Itself

What **portion** of entry is about the video?

- 3 = entire entry
- 2 = part of entry
- 1 = not part of entry (e.g. in page sidebar)
- C = only in comments

To what extent does entry provide **contextual information** related to video?

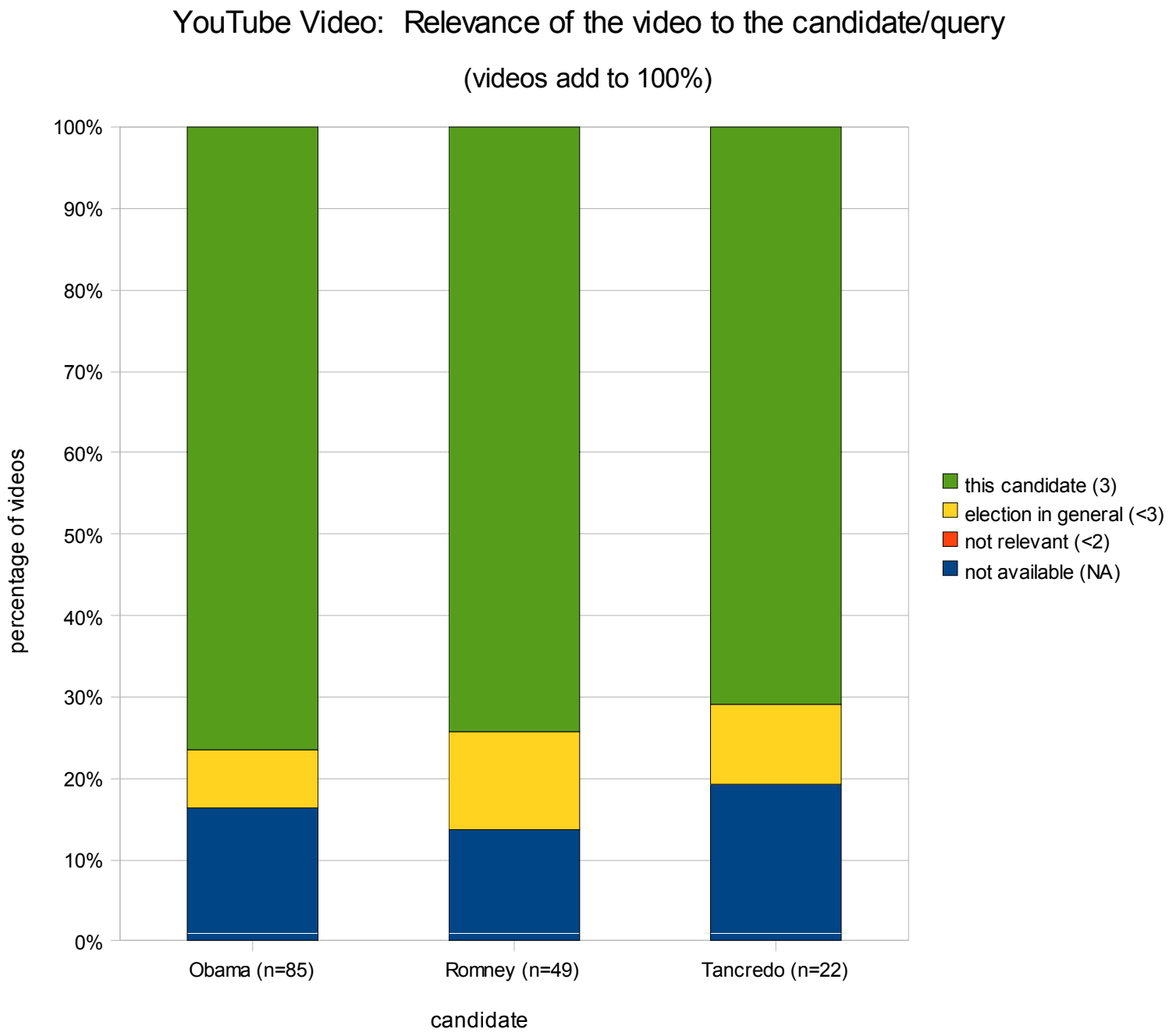
- 3 = provides substantial amount of contextual information (like a news item about the video)
- 2 = some contextual information beyond that provided by title of video itself
- 1 = no real contextual information (e.g. only “click here,” video title, or URL)

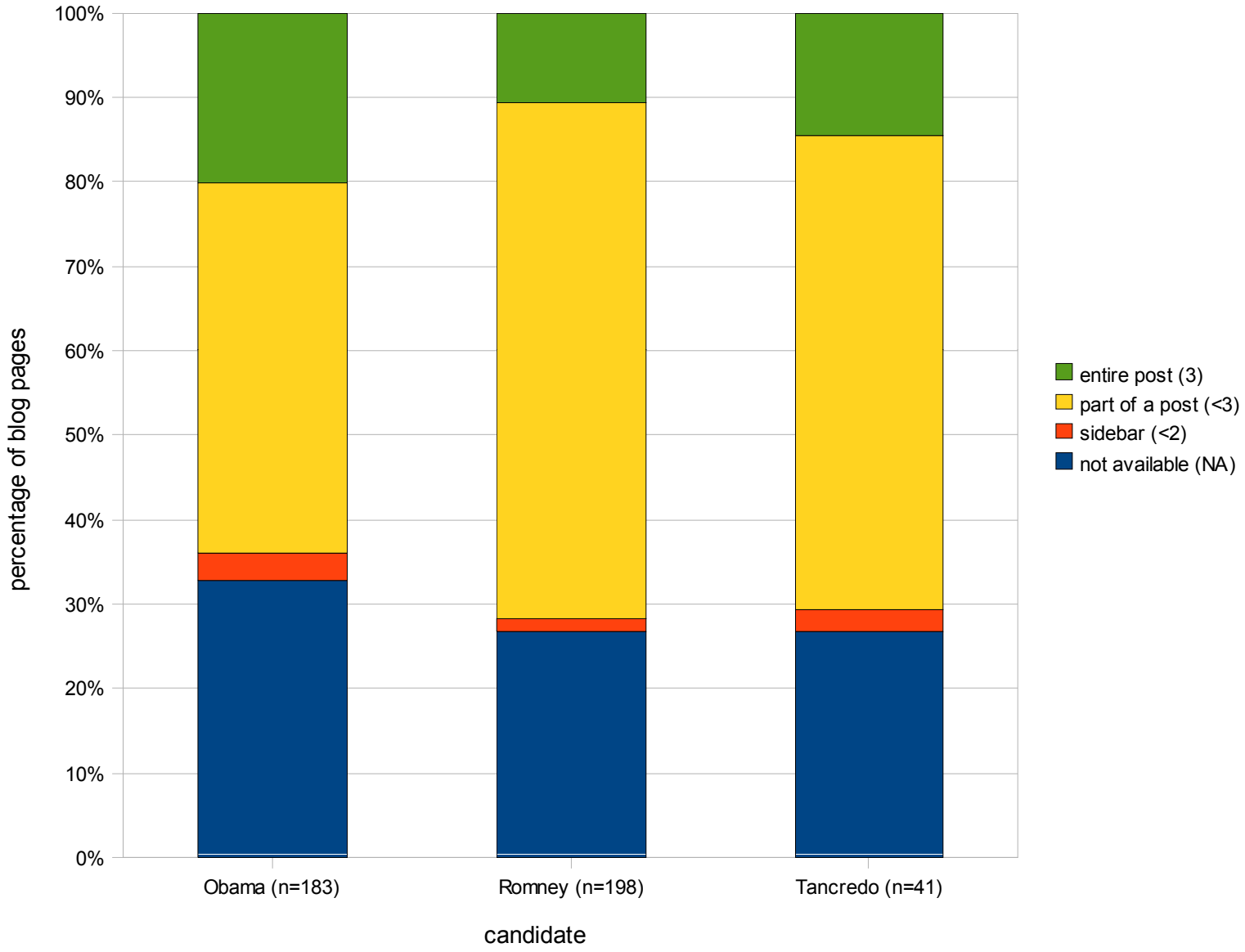
Excluded from this Analysis

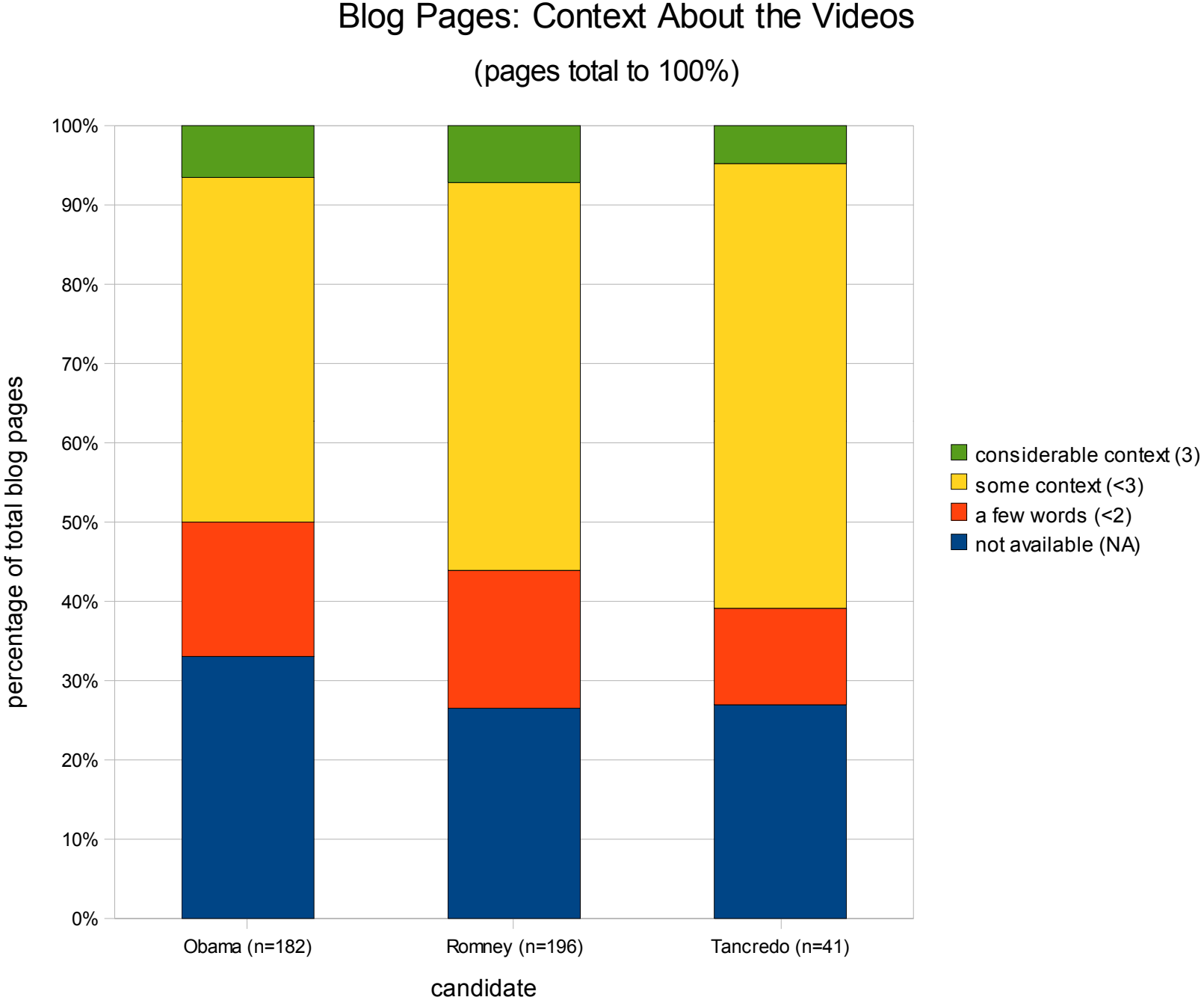
- Videos removed from YouTube
- Blog pages no longer available
- Comments posted to blogs

Provisional Findings

Note: Vilsack data not reported, due to small sample size


YouTube Video: Relevance of the video to the candidate/query
(videos add to 100%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
percentage of videos
Obama (n=85)
Romney (n=49)
Tancredo (n=22)
candidate
this candidate (3)
election in general (<3)
not relevant (<2)
not available (NA)


Blog Pages: Extent of the Post that is about the Video
(pages add to 100%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
percentage of blog pages
Obama (n=183)
Romney (n=198)
Tancredo (n=41)
candidate
entire post (3)
part of a post (<3)
sidebar (<2)
not available (NA)


Blog Pages: Context About the Videos
(pages total to 100%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
percentage of total blog pages
Obama (n=182)
Romney (n=196)
Tancredo (n=41)
candidate
considerable context (3)
some context (<3)
a few words (<2)
not available (NA)

Findings Highlights

- Most videos were relevant
- Most blog entries about videos are also about other things
- Most blogs provide some additional information beyond video's title
- 5-7% provide substantial contextual information
- Information often repeated across blog entries

Future Directions – Open Issues

- Materials related to our other collecting areas (not just election)
- Crawl parameters along 3 dimensions: environments crawled, access points, threshold values
- Predicting diminishing returns when collecting contextual information
- Systematic use of traffic & in-link data for selection
- Parsing pages & extracting appropriate links - specific part of page that represents individual blog entry, filtering unrelated sites & out-links
- Blog pages as “contextual information bridges”
- Predicting types of pages most likely to provide contextual information

Thank you!

http://ils.unc.edu/vidarch