I Know It's Important, But What Am I Looking At?
Strategies for Using Blog Content to Contextualize YouTube Videos

CHRISTOPHER A. LEE, ROBERT CAPRA, RACHAEL CLEMENS, and LAURA SHEBLE
School of Information and Library Science, University of North Carolina at Chapel Hill

Abstract: Archivists have long recognized that item-level description is not viable for most collections of primary sources, and have instead focused on documenting aggregate-level units within which records were organized by their creators. The VidArch project is capturing YouTube videos associated with the 2008 U.S. presidential election, as well as exploring strategies for appraising and describing them. The impact of YouTube videos on public perception and election outcomes is likely to be significant. Unfortunately, their "origin order" often provides very limited contextual information that will be essential for users to make sense of them in the future (e.g., creators, events and individuals represented, how they were interpreted by citizens at the time). Blogs are also an important source for documenting online deliberations. Archivists can collect blogs directly, but there is also great potential to tap blogs as sources of documentation about the "story behind" YouTube videos, both through the text they present and through links that they provide to other sources (serving as "contextual information bridges"). We have been exploring and testing ways to systematically collect blog entries related to given sets of YouTube videos. The harvesting of contextual information from external sources will become increasingly important for archivists as (1) items in environments such as YouTube play a significant role in phenomena that should be documented, and (2) the environments themselves provide limited contextual information. We report on approaches to support what Hans Booms would call a "documentation plan" for reflecting the conversation space surrounding contemporary events.

Introduction

The Web has become a vital forum of deliberation around issues of societal importance. The 2008 U.S. presidential election, for example, is likely to be strongly influenced by materials posted to, shared and discussed on the Web. When building long-term digital collections, it is essential not only to ensure continuing access to "target digital objects" but also to create, capture and manage contextual information to allow future users to understand, make sense of, analyze and use the target digital objects (Lee, 2007). The VidArch project is capturing YouTube videos and web pages associated with the election, as well as exploring strategies and building tools for curators of digital collections to appraise and describe such materials.

In December 2007, 48% of American Internet users reported having "watch[ed] a video on a videosharing site like YouTube or GoogleVideo," while 14% reported posting videos online that they had recorded (Rainie, 2008). American "poli-fluentials" – who are the Americans "likeliest to volunteer, donate, promote candidates and join causes through both online and word-of-mouth advocacy" – are more likely than other respondents to report having seen or posted online political videos (Darr and Graf, 2008). This provides new opportunities for relatively open discourse, while also challenging control of traditional authorities over predominant messages. In the 2006 U.S. elections, YouTube and MySpace "weaken[ed] the level of control that campaigns have over the candidate's image and message since anybody, both supporters and opponents, can post a video and/or create a page on behalf of the candidates" (Guerguieva, 2008). In December 2007, 24% of Americans reported regularly learning about the campaign from the Internet, and 24% report having seen something about the campaign in an online video (Social Networking and Online Videos Take Off, 2008). YouTube is playing an increasingly important role in political discourse and may have a significant impact on voting behavior (Panagopoulos, 2007). Events that would have previously had only a very local impact can now attain widespread visibility and impact, because they are posted to YouTube. An even larger set of web sites provide links to and commentary about the content in YouTube.

As with YouTube, the blogosphere is a popular and influential space for political discourse, providing space for extended discussion, speculation and various forms of advocacy. Blogs (web logs) are web sites that are periodically updated with a series of self-contained entries, which are usually presented in reverse-chronological order (newest entries at the top). They are often maintained by only one person or select group of people, though the blog software usually allows visitors to post comments related to individual blog entries. The "blogosphere" is a common label for the entire network of blogs on the Internet, which often link to each other. Those who report daily use of political blogs are more likely to be at the ends of the political spectrum, and their political blog reading is strongly motivated by an interest in "news the mass media ignore" and a "different perspective on the news" (Graf, 2006). Blog pages[1] are more likely than other Web pages to provide out-links to "hubs," often as a result of bloggers copying material out of "news items from key blog hubs and adding their own comments to them; in most cases this is done to let friends within the local peer network know what is interesting in the wider Web, while giving credit to the source" (Kirchhoff, Bruns and Nicolai, 2007).

A repository attempting to document political deliberation surrounding the election would be well served by including in its collecting scope, not only "official" materials from campaigns and mainstream media, but also content from these popular online interaction spaces, especially when repositories intend to serve as "curators of the experience as well as the record" (Hinding, 1993).

Many YouTube videos inspire a great deal of online discussion and attention, which suggests that they can be important to preserve. However, it is often very difficult to understand "what you're looking at" solely based on the contents of a YouTube page itself. In this environment, archival description may involve capturing online discussion (e.g., sampling from blogosphere), rather than archivists being the primary creators of descriptive materials. This paper reports on an investigation of the utility and value of capturing blog pages that link to YouTube videos, as a means of collecting contextual information associated with (and thus supporting more meaningful access in the future of) those videos.

Appraisal as Mirror of Society

A fundamental challenge for curators of digital collections is appraisal, i.e. determining what segments of the documentary universe should be obtained and preserved. In a Web environment, appraisal can inform rules for crawls (sources, access points, filtering rules, and relevance criteria). Appraisal should be guided by notions of what one ultimately is trying to document. Documenting a contemporary phenomenon often requires cutting across numerous institutions and media (Samuels, 1986). In VidArch, we are

[1] We use the term "blog page" to refer to an entire page rendered in a browser when visiting a specific blog URL. A "blog entry" is the subset of content on that page that was posted by the original contributor and excludes comments posted by others and any navigational elements on the page.

addressing what we see as a gap between the literature on web archiving and established conceptions of archival appraisal.

Hans Booms argued that appraisal should be based on best (i.e. most informed by empirical evidence) judgments of the "value ascribed by those contemporary to the material," i.e. what members of society judged most valuable or important at the time documents were created (Booms, 1987). If one accepts this approach, then a natural next question is how best to reflect the emphasis that people were placing on issues or materials at a given time. There is no single monolithic set of values or perceptions of "society" but one can use various data sources to what is most influential, viewed, discussed, and cited. Curators of digital collections will need tools and methods for combining information from both queries and crawls to identify and collect Web materials that document and contextualize phenomena. VidArch is developing and testing such approaches, in order to support what Booms would call a "documentation plan" for reflecting the heterogeneous and interlinked conversation space surrounding contemporary events.

Research Design

The VidArch team has used the YouTube application program interfaces (APIs) to collect videos related to the 2008 U.S. presidential election, along with associated comments and other metadata, based on 57 queries to YouTube every day (except for days of maintenance), since May 2007 (Shah and Marchionini, 2007). The queries[2] include 50 names of individual candidates and 6 queries related to the election in general (e.g., "election 2008").There are two sources of data for the study reported in this paper:

1. YouTube videos and associated metadata that the VidArch project has been capturing related to the U.S. presidential candidates for the 2008 election.

2. Beginning June 6, 2007, Fred Stutzman, doctoral student at the University of North Carolina, began a systematic collection of links from blog pages related to the 2008 U.S. presidential election, as part of a separate effort to investigate online political deliberation. Queries related to 15 of the presidential candidates were submitted through both Google Blogsearch and Technorati. Software then captured a subset of blog pages that either included the name of a presidential candidate in their content or provided one or more links to a candidate's web site. Once the query set was retrieved from the search engine, a web crawler was dispatched to matching pages. This crawler created "profiles" of those pages, collecting the outbound links. The queries were run three times an hour, every hour of the day. Results were limited to 10 per query due to service limitations of using the RSS functions of Google Blogsearch and Technorati.

For the present study, we generated a subset of blog pages from #2 that linked to at least one video from #1. We examined and coded a statistical sample[3] of blog pages for four candidates: Barack Obama, Mitt Romney, Tom Tancredo, and Tom Vilsack.[4] All four of the co-authors served as coders. We engaged in four rounds of preliminary coding using a candidate whom we did not plan to include in the analysis, in order to clarify coding categories and ensure inter-coder reliability. Each blog page was coded by two different people, and the lower of the two codes was used for the analysis shown here. Table 1 summarizes the questions and codes. We excluded comments posted to blogs from our analysis. In order to facilitate efficient coding, we relied on the live web for viewing the YouTube videos and blog pages. For this reason, we coded some items as "Not Available": specifically, videos that were removed from YouTube and blog entries that were no longer available at the URL identified in the original crawl.

[2] A full list of the queries is available at http://www.ils.unc.edu/vidarch/.

[3] Since we were using a fixed-size data set, we based our sampling on the hyper-geometric distribution and chose sample sizes designed to generate results with 95% confidence intervals and a range of ±4%.

[4] The Vilsack data are not reported in this paper, due to a small sample size.

Table 1. Coding categories applied to YouTube Videos and Blog Entries

Question	Object of Analysis	Coding Categories
Is the video about the candidate?	YouTube Video	3 = about the candidate 2 = about the election but not the candidate 1 = about neither the candidate nor the election
What **portion** of the blog entry is about the video?	Blog entry that links to a YouTube Video	3 = entire entry 2 = part of the entry 1 = not part of the entry (e.g., in page sidebar) C = only in comments
To what extent does the blog entry provide **contextual information** related to the video?	Blog entry that links to a YouTube 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)

Findings

Our inter-coder agreement ratings indicated substantial agreement. Cohen’s kappa for the coders averaged 0.76 for the question about what portion of the blog entry is about the video and 0.67 for the question about the extent of contextual information provided. In the results presented in this section, we combined the two ratings for each item into a single rating. In cases where the coders agreed, we used the agreed upon rating. In cases of disagreement, we used the lower of the two ratings.

Consistent with our previous analysis (Capra et al, 2008), videos that the YouTube ranking algorithm has identified as falling within the top 100 results for given election-related queries are, in fact, related to the election. The average number of blog pages per video for the candidates is: Romney = 5.44, Obama = 5.31, Tancredo = 3.74. Of the 182 videos that we analyzed, 16% were not available, 9% were relevant to the election in general, and 75% were directly relevant to the candidate named in the query. Figure 1 provides a further breakdown of video relevance by individual candidate.

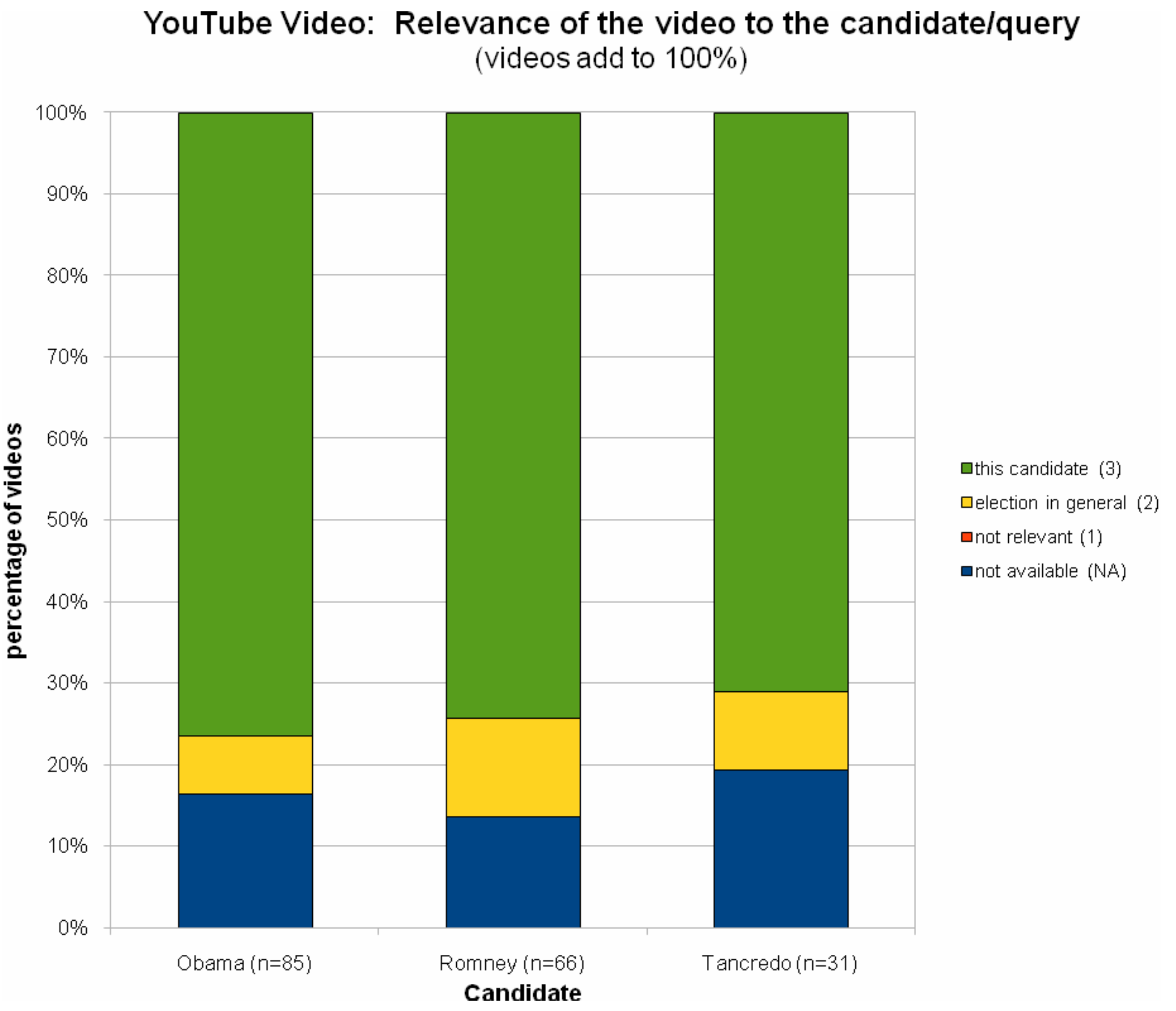

Figure 1. Video Relevance

Our analysis of the blog entries that link to the YouTube videos revealed that most of the entries do discuss the videos to which they link. However, most of the blog entries also contain discussions of other things that are not directly related to the video. Overall, 15% of the blog entries were *entirely* about the linked YouTube video, while 52% were determined by the coders to be only *partially* about the linked YouTube video. Twenty-nine percent (29%) were not available and only 3% contained content linking to the YouTube video only in a sidebar on the page. Figure 2 shows the percentages for the individual candidates.

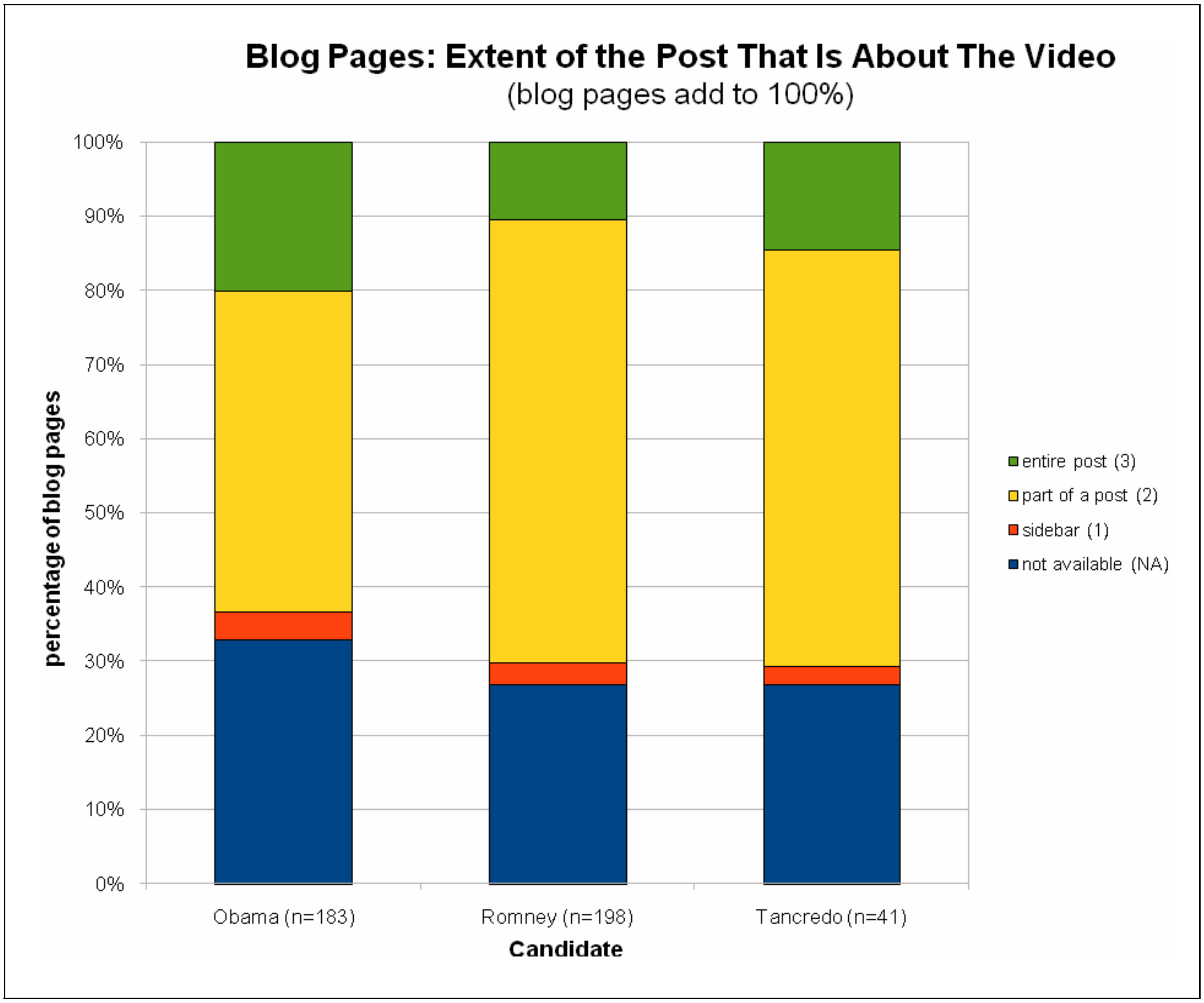

Figure 2. Extent of Blog Entry that is about YouTube Video to which it Links

Most of the blog entries that we analyzed do provide some additional contextual information, which suggests that crawling and capturing blog pages that link to YouTube videos can, in fact, serve as a means to gather contextual information about those videos (see Figure 3). Overall, 47% of the blog entries were determined by the coders to provide some contextual information beyond that provided by title of video itself. However, only a small subset (7%) of the blog entries were determined to provide substantial contextual information. Twenty-nine percent (29%) of the blog entries were not available, and 17% provided only a few words of contextual information. We also observed that the same information (similar in factual content and often even using identical text) about a YouTube video is often repeated across multiple blog entries.

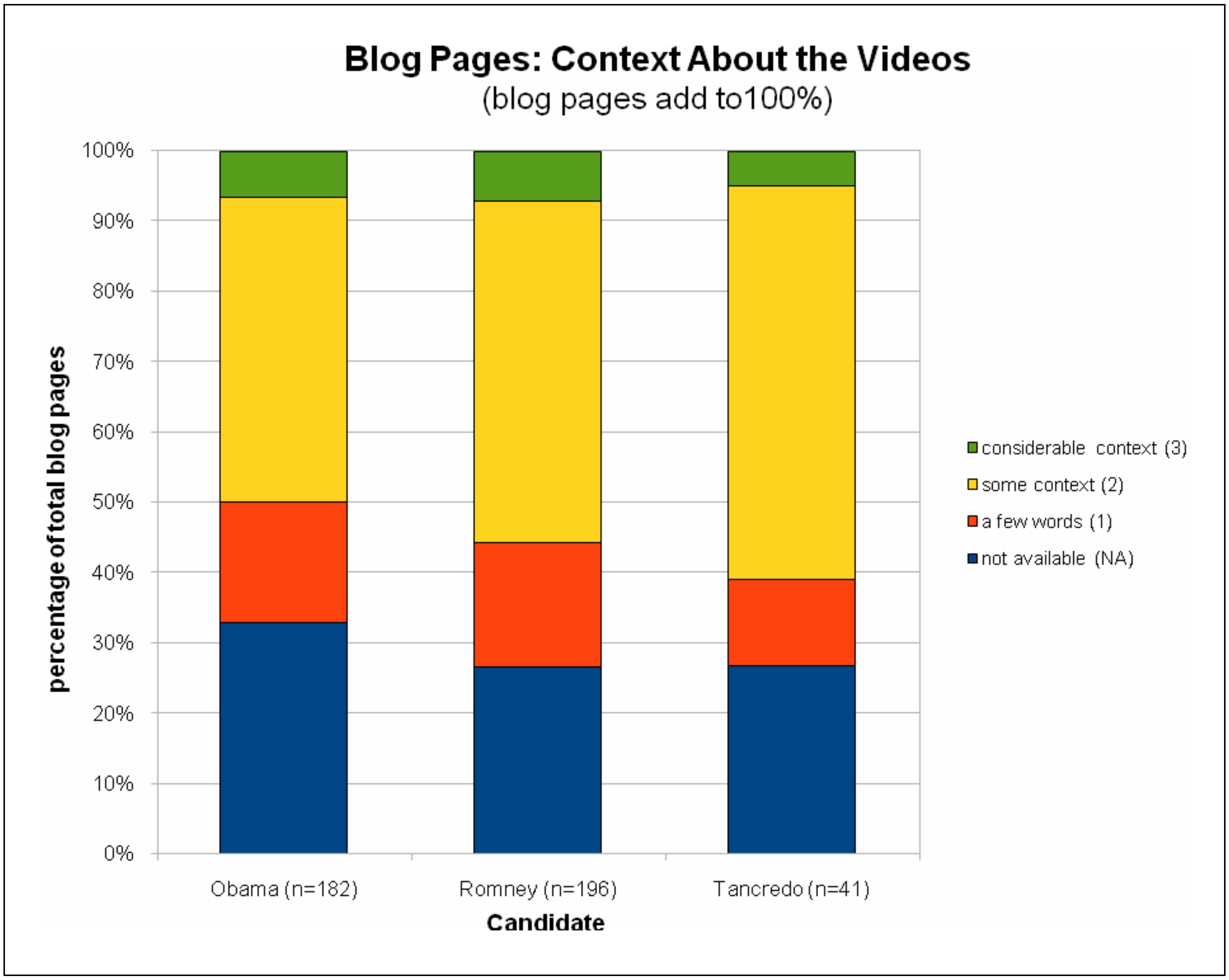

Figure 3. Amount of Contextual Information about YouTube Video Provided by Blog Entry that Links to that Video

In short, we have identified many instances in which gathering of contextual information from blog pages would add value to a collection of YouTube videos, but there are many remaining questions about how best to configure blog crawls to ensure an appropriate balance of precision and recall for a given set of collecting objectives.

Future Directions

In this paper, we report preliminary findings from only one collecting area (U.S. presidential election campaign), but VidArch has also been running crawls on many other topics (e.g., energy, epidemics, health, natural disasters, truth commissions). We do not yet know the extent to which the findings in this paper will be relevant to efforts to document non-election phenomena. We also do not yet have generalizable findings about the likely rate of diminishing returns when collecting additional contextual information. For example, if one could collect many salient contextual details about a video from the first 8–10 blog pages that link to the video, that it might not be worthwhile to expend resources on capturing even more blog pages.

We would also like to determine more specific implications for crawling parameters across at least three dimensions: environments crawled, access points from those environments used as crawling or selection criteria, and threshold values for scoping capture within given access points (Capra et al, 2008). When collecting blog pages, selection can be informed by analysis of link and use patterns. Other research suggests that there are distinct online "communities" within the blogosphere that tend to link to each other on given topics, as well as a small set of "A-List" blogs that are frequently consulted (Adamic and Glance, 2005). More generally, "hyperlinks and search engines play a key role in funneling Web users to a handful of sites," (Hindman, Tsioutsiouliklis and Johnson, 2005) so that "although the range of online content is vast, the range of sites that users actually visit is small" (Hindman, 2006). Further analysis could determine that some of those highly in-linked domains provide a substantial amount of information that can help to contextualize the YouTube videos, as opposed to those that serve as "hubs" for reasons unrelated to the contextual information that they provide.

There are also many open questions related to parsing pages and extracting appropriate links. First, it will often be helpful to identify the specific part of a page that represents an individual blog entry as distinct from other surrounding content. A second question is how best to detect and systematically filter the large number of unrelated out-links in some blog pages. Finally, selection activities could be greatly facilitated by automated or semi-automated detection of unrelated blogs that systematically provide distractor links.

Acknowledgements

This research is supported by a grant from the National Science Foundation and National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress.

References

Adamic, Lada A., and Natalie Glance. "The Political Blogosphere and the 2004 US Election: Divided They Blog." In *Proceedings of the 3rd International Workshop on Link Discovery*, 36-43. New York, NY: ACM Press, 2005.

Booms, Hans. "Society and the Formation of a Documentary Heritage: Issues in the Appraisal of Archival Sources." *Archivaria* 24 (1987): 69-107.

Capra, Robert, Christopher A. Lee, Gary Marchionini, Terrell Russell, Chirag Shah, and Fred Stutzman. "Selection of Context Scoping for Digital Video Collections: An Investigation of Youtube and Blogs." In *JCDL 2008: Proceedings of the 8th ACM/IEEE Joint Conference on Digital Libraries: Pittsburgh, Pennsylvania, June 15–20, 2008,* edited by Ronald L. Larsen, Andreas Paepcke, José Luis Borbinha and Mor Naaman, 211–20. New York, NY: ACM Press, 2008.

Darr, Carol, and Joseph Graf. "Poli-Fluentials: The New Political Kingmakers." Washington, DC: Institute for Politics, Democracy and the Internet, 2007.

Drezner, Daniel W., and Henry Farrell. "The Power and Politics of Blogs." *Public Choice* 134, no. 1-2 (2008): 15–30.

Graf, Joseph. "The Audience for Political Blogs: New Research on Blog Readership." Washington, DC: Institute for Politics, Democracy and the Internet, George Washington University, 2006.

Gueorguieva, Vassia. “Voters, Myspace, and Youtube: The Impact of Alternative Communication Channels on the 2006 Election Cycle and Beyond.” *Social Science Computer Review* 26, no. 3 (2008).

Hinding, Andrea. “Inventing a Concept of Documentation.” *Journal of American History* 80, no. 1 (1993): 168V78.

Hindman, Matthew, Kostas Tsioutsiouliklis, and Judy A. Johnson. “Measuring Media Diversity Online and Offline: Evidence from Political Websites.” Paper presented at the 32nd Research Conference on Communication, Information and Internet Policy 2005.

Hindman, Matthew. “A Mile Wide and an Inch Deep: Measuring Media Diversity Online and Offline.” In *Localism and Media Diversity: Meaning and Metrics*, edited by Philip Napoli, 327–47. Mahwah, NJ: Lawrence Erlbaum Associates, 2006.

Kirchhoff, Lars, Axel Bruns, and Thomas Nicolai. “Investigating the Impact of the Blogosphere: Using PageRank to Determine the Distribution of Attention.” Paper presented at the Annual Conference of the Association of Internet Researchers, Vancouver, Canada 2007.

Lee, Christopher A. “Taking Context Seriously: A Framework for Contextual Information in Digital Collections.” UNC SILS Technical Report 2007-04. 2007. http://sils.unc.edu/research/publications/reports/TR_2007_04.pdf.

Lee, Christopher A., and Helen R. Tibbo. “Capturing the Moment: Strategies for Selection and Collection of Web-Based Resources to Document Important Social Phenomena.” In *Archiving 2008: Final Program and Proceedings, June 24–27, 2008, Bern, Switzerland,* 300–305. Springfield, VA: Society for Imaging Science and Technology, 2008.

McKenna, Laura, and Antoinette Pole. “What Do Bloggers Do: An Average Day on an Average Political Blog.” *Public Choice* 134 (2008): 97–108.

Panagopoulos, Costas. “Technology and the Transformation of Political Campaign Communications.” *Social Science Computer Review* 25, no. 4 (2007): 423–24.

Rainie, Lee. “Pew Internet Project Data Memo: Video Sharing Websites.” Washington, DC: Pew Internet and American Life Project, 2008.

Samuels, Helen Willa. “Who Controls the Past: Documentation Strategies Used to Select What Is Preserved.” *American Archivist* 49 (1986): 109–24.

Shah, Chirag, and Gary Marchionini. “Preserving 2008 US Presidential Election Videos.” Paper presented at the 7th International Web Archiving Workshop, Vancouver, Canada, June 23, 2007.

“Social Networking and Online Videos Take Off: Internet’s Broader Role in Campaign 2008.” Washington, DC: Pew Research Center for the People and the Press, 2008.

Tibbo, Helen R., Chirstopher A. Lee, Gary Marchionini, and Dawne Howard. "VidArch: Preserving Meaning of Digital Video over Time through Creating and Capture of Contextual Documentation." In *Archiving 2006: Final Program and Proceedings, May 23–26, 2006, Ottawa, Canada*, edited by Stephen Chapman and Scott A. Stovall, 210–15. Springfield, VA: Society for Imaging Science and Technology, 2006.