Archiving 2008

June 24–27, 2008
Bern, Switzerland

Final Program and Proceedings

General Chair:
Rudolf Gschwind, University of Basel

IS&T
imaging.org

Sponsored by the
Society for Imaging Science and Technology

In cooperation with

AIC American Institute for Conservation of Historic & Artistic Works
ALA ALCTS Association for Library Collections and Technical Services
CNI Coalition for Networked Information
DLF Digital Library Federation
DPC Digital Preservation Coalition
ECPA European Commission on Preservation and Access
ISCC Inter-Society Color Council
IOP Institute of Physics
MCN Museum Computer Network
OCLC Online Computer Library Center
RPS Royal Photographic Society

Capturing the Moment: Strategies for Selection and Collection of Web-Based Resources to Document Important Social Phenomena

Christopher A. Lee and Helen R. Tibbo; University of North Carolina; Chapel Hill, NC

Abstract

The VidArch project is capturing YouTube videos and web pages associated with the 2008 U.S. presidential election. We are also exploring strategies and building tools for curators of digital collections to appraise and describe such materials. Blogs are an increasingly important source for documenting online deliberations. Blogs can provide commentary, but they can also serve as "contextual information bridges" for identifying and capturing resources to which the pages link.

Web archiving literature usually defines collecting in terms of setting up a set of seeds for crawls based on specific URLs. However, a substantial portion of material on the Web is accessible through posing queries. Curators of digital collections will need tools and methods for combining information from queries and crawls to identify and collect materials. The VidArch project is developing and testing such approaches, in order to support what Hans Booms would call a "documentation plan" for reflecting the heterogeneous and interlinked conversation space surrounding contemporary events.

Introduction

The Web has become a vital forum of deliberation around issues of societal importance. The 2008 election for president of the United States, for example, is likely to be strongly influenced by materials posted to, shared and discussed on the Web. According to a December 2007 Pew survey, 24% of Americans report regularly learning about the campaign from the Internet, up from 13% in 2004 and 9% in 2000 [1]. According to the Pew survey, 24% of Americans report having seen something about the campaign in an online video.

In order to make sense of the 2008 electoral process, future researchers would benefit not only from perpetual access to Web materials but also contextual information to make meaningful use and sense of the materials. The VidArch project [2] 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.

Importance of Documenting Online Deliberation Spaces

YouTube allows for widespread dissemination of videos. According to a December 2007 Pew survey, 48% of American Internet users reported having "watch[ed] a video on a video-sharing site like YouTube or GoogleVideo," while 14% reported posting videos online that they had recorded [3]. These numbers are much higher among American "poli-fluentials," who are expected to be most active in the 2008 election [4]. 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..." [5]

YouTube is playing an increasingly important role in political discourse and may have a significant impact on voting behavior [6]. All major candidates for the U.S. presidential election created YouTube channels. YouTube and CNN jointly sponsored presidential debates, featuring video questions uploaded by YouTube users. Several candidates also posted videos to YouTube in which they posed specific questions and asked users to post video replies. Perhaps even more importantly, 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.

We use the term "blogosphere" to refer to the distributed and inter-linked body of blog (web log) pages. Blogs are based on software that allows for relatively easy additional of small entries to pages over time. As with YouTube, the blogosphere is a popular and influential space for political discourse. Like YouTube, the blogosphere is also provides space for extended discussion, speculation and agenda-pushing that might not happen in traditional media venues. 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" [7]. Blog pages 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" [8].

The above discussion suggests that, if a repository had the goal of documenting political deliberation surrounding the election, it would be well served by including in its collecting scope, not only "official" materials from the 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" [9].

Appraisal of Web Materials

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 [10]. In VidArch, we are addressing what we see as a gap between the literature on web archiving and established conceptions of archival appraisal.

Twenty years ago, the translation of writing by Hans Booms introduced a new perspective to North American archival thought: 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 [11]. 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.

Two assumptions underlying the work described in this paper are: online deliberation surrounding the U.S. presidential election process is important to document; and YouTube videos are playing a prominent role in the deliberation process, which warrants the preservation and contextualization of a subset of the videos.

Enacting Appraisal Criteria through Crawling

Web archiving tools and techniques have matured dramatically in recent years, and numerous institutions have taken on web archiving initiatives [12] [13]. Web capture has usually been based on identifying a set of seed uniform resources locators (URLs) and then recursively following links within a specified set of constraints (e.g. number of hops, specific domains). The "Arizona Model" is an important attempt to operationalize the archival principles of provenance and original order by mapping web crawling criteria to hierarchical structure of sites [14]. Web archiving based on recursive link following, however, faces two major challenges. First, link paths often do not map cleanly or directly to long-standing criteria for appraisal and collection development, e.g. topics, provenance, genres, dates. Second, a substantial portion of material on the Web is accessible through posing queries to databases, rather than following links.

Several projects have demonstrated methods for scoping a topic-based crawl, based on automated analysis of the content of pages [15]. There have also been efforts to automatically populate web entry forms and collect pages that cannot be reached through link-following [13][16][17][18]. There has been relatively little investigation of combining link-following and queries to select complimentary sets of resources.

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.

VidArch Approach

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 [19]. The queries include 50 names of individual candidates and 6 queries related to the election in general (e.g. "election 2008").

We use the term "crawl" to indicate one instance of executing the following two sets of activities: 1) submitting all 57 queries to YouTube and collecting data from the top 100 results of each query based on YouTube's relevance ranking; and 2) collecting updated dynamic metadata for each video that has been "discovered" through an instance of step 1.

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 [20]. We are using data from YouTube and elsewhere on the Web (blogs, in-links identified by Web search engines, Web traffic data) to inform the appraisal of the YouTube videos and collect further contextual information associated with the videos. Collecting data from various sources allows us to identify likely strengths, gaps and complementarities.

Collecting and Analyzing Blog Pages

Beginning June 6, 2007, Fred Stutzman began a systematic collection of links from blog postings related to the 2008 U.S. presidential election. Queries related to 15 of the candidates were submitted through both Google Blogsearch and Technorati. For VidArch, a subset of blog postings were captured that either (1) included the name of a presidential candidate in their content or (2) linked to a candidate's web site. The queries were run three times per hour, for a total of 72 queries per term each day. Once a given query set was retrieved, a web crawler created a "profile" for each blog page, reflecting its out-links. Within the resulting data set, an "out-link" is any link from the blog page to a resource outside that page; thus including links to other postings within a blog, navigational links, and links to ads or related postings.

Contextual Information Bridges

As noted above, blog entries are often relatively short snippets of text that then provide links to other resources where readings can discover more information [8]. Drezner and Farrell argue that blogs "are important less because of their direct effects on politics than their indirect ones—they influence important actors within mainstream media who in turn frame issues for a wider public" [21]. McKenna and Pole report, "By far the most popular activity for all political bloggers is providing readers with links to reports and articles found elsewhere." [22]

We are exploring the strengths and limitations of blog pages as "contextual information bridges" (CIBs), allowing curators to mine the postings to identify and then capture other online resources to which the pages link (i.e. those who linked to a given YouTube video also tended to link to some other explanatory source such as a newspaper article).

A motivating example for considering blog pages as contextual information bridges is "Edwards Places Campaign Headquarters in NC" [23], which was produced and posted to YouTube by Carla Babb, a graduate student in journalism at the University of North Carolina. The video points out the irony of the campaign headquarters for John Edwards being located in an affluent neighborhood, given the Edwards campaign's strong focus on alleviating poverty. The video drew controversy and media attention when the Edwards campaign allegedly demanded it be

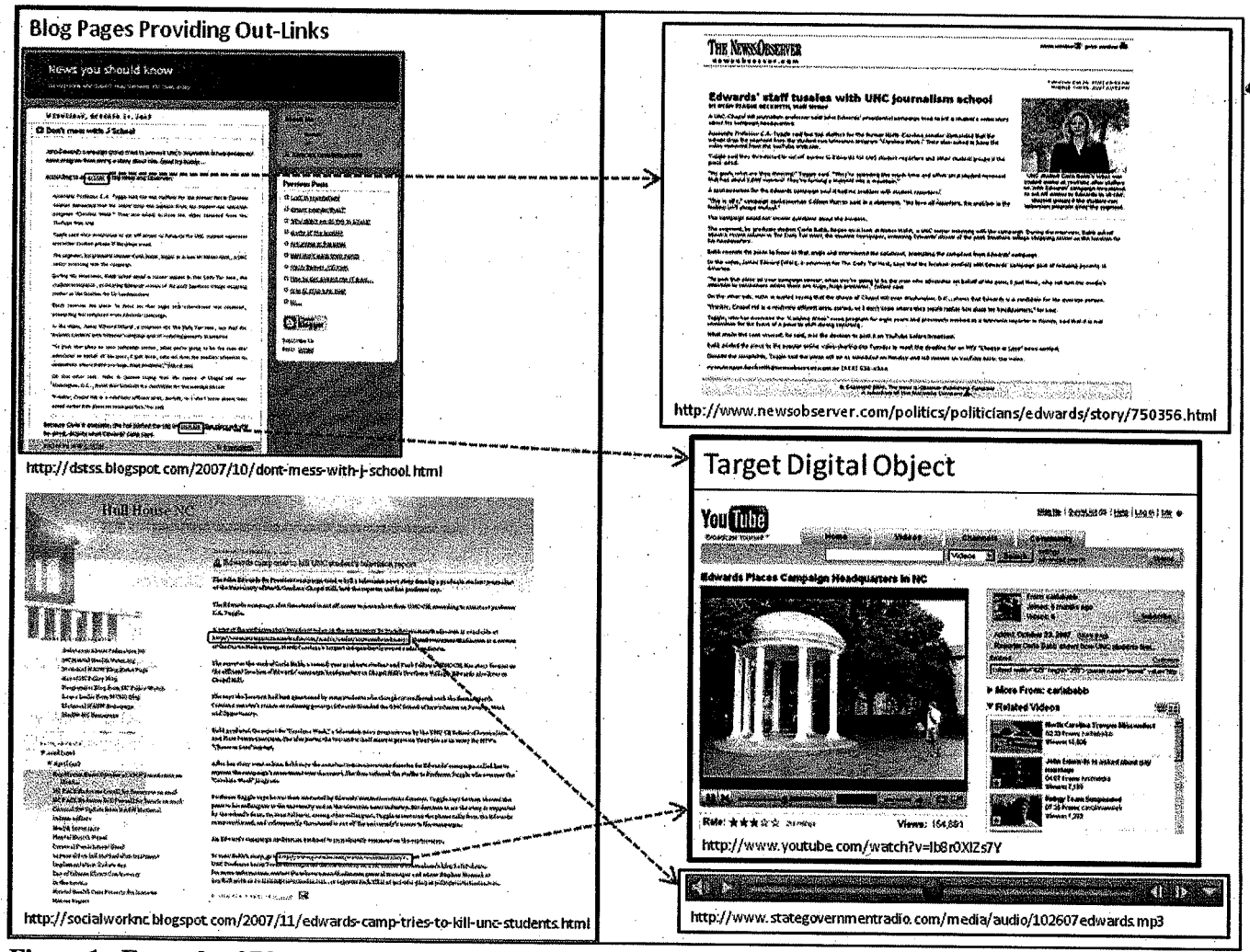

Figure 1 - Example of Blogs as Contextual Information Bridges

taken down from YouTube. It never appeared in the top 100 results for a "John Edwards" query to YouTube, which is an important reminder that videos that are influential and popular within YouTube might be missed by simple queries based solely on relevance rank. YouTube did list this video as #8 for most viewed videos in "News & Politics" for this week of 2007-10-30.

We conducted a query within Google Blog Search for blog pages that link to this video. In addition to any information that the blog pages themselves provide about the video (which varies from a simple link with no further explanation, to a fairly detailed 319-word explanation of the controversy surrounding the video), we noted that many of the blog pages also linked to other online sources that provided further contextual information. See Figure 1 for an illustration of blog pages that link to this video serving as bridges to contextual information in other online sources.

Obama Collection

VidArch has analyzed relationships between YouTube and blog data, including overlap, consistency and relative relevance to the intended collecting scope, which are reported elsewhere [23]. In order to further investigate the potential role of blogs as sources of contextual information for online videos, we have more closely analyzed the blog and YouTube data related to Barack Obama.

The blog crawler collected data for 136,687 blog pages in response to the Obama query. Those pages contain 1,468,533 out-links, for an average of 10.74 out-links per page. Of those out-links, 10,285 (.7%) are to YouTube videos, of which there are 6,903 unique videos. There are 4135 pages (3% of all blog pages generated from the Obama crawl) that represent potential contextual information bridges, because they contain out-links to at least one YouTube video. The 4135 pages contain a total of 170,990 out-links. Table 1 lists the 20 YouTube videos that received the most in-links from the blog pages, 10 of which are also in our collection based on crawling YouTube directly.

Out-links from the crawls of blog pages are a less precise indicator of a video being "about Barack Obama" than is YouTube's relevance ranking, which is consistent with our earlier findings across several of the candidates [25]. Of the videos in Table 1, four are focused on other candidates besides Obama, one is about a politician who was not a presidential candidate, and two that are not about election campaigns at all, none of which appeared in our collection resulting from crawls of YouTube. If one's primary task is identifying a very focused set of YouTube videos to serve as the target digital objects to collect on a given topic, it appears that direct crawls of YouTube may be more effective than relying on links from crawled web pages, particularly if the topic can easily be translated into a simple query (e.g. name of a candidate). If, however, one has already identified a set of target YouTube videos to collect, and would like to identify further online resources that can provide contextual information associated with those videos, the blog pages could be much more valuable. We identified many cases of blog pages

providing useful contextual information about a video either in the content of the blog entry itself or through a link to an informative textual, audio or video source.

Table 1 - Top 20 YouTube Videos Linked from "Barack Obama" Blog Pages (* = also in YouTube Crawl Set)

Title	YouTube ID	Links from Blogs
Vote Different	6h3G-IMZxjo	244*
"I Got a Crush...On Obama" By Obama Girl	wKsoXHYICqU	173*
Barack Obama 2004 Democratic National Convention Part 1	MNCLomrqIN8	40*
John Edwards Feeling Pretty	2AE847UXu3Q	39
1984 Apple's Macintosh Commercial	OYecfV3ubP8	29
Mike Gravel - Rock -	0rZdAB4V_j8	28
1984	cWvHbOoG3tI	27*
Barack Obama on Saturday Night Live	ndQM0X5rhfE	26*
Rudy Giuliani in Drag Smooching Donald Trump	4IrE6FMpai8	26
Barack Obama 2004 Democratic National Convention Part 2	56-m8wx1mwo	25*
Barack Obama: My Plans for 2008	H5h95s0OuEg	25*
Joel Surnow's "The 1/2 Hour News Hour"	YjlfaMwlFxU	23
Tomorrow Begins Today	1etlZaf6zUw	23
FOX ATTACKS OBAMA	ouKJixL--ms	23*
Debate '08: Obama Girl vs Giuliani Girl	ekSxxlj6rGE	21
Barack Obama's Speech at the Jefferson Jackson Dinner	tydfsfSQiYc	20*
Hott 4 Hill feat. Taryn Southern	-Sudw4ghVe8	20
George Allen introduces "Macaca"	r90z0PMnKwI	20
Meet Barack Obama	WGGIHqloP2k	20*
Hillary Clinton Sings National Anthem	bfZ_gXCHaMw	18

An illustrative example is the most in-linked video in Table 1: "Vote Different" with 271 in-links (including "1984," the same video content that was re-posted to YouTube by someone else). The set of blog pages linking to this video lead to several vital elements of contextual information about it, including that it is based on the earlier "Think Different" advertisement from Apple (the fifth video in Table 1), it was often falsely assumed to be a product of the Obama campaign, Phil de Vellis created it, and he did so "to show that an individual citizen can affect the process" [23]. This important contextual information does not appear to be represented in our collection resulting solely from submitting daily "Barack Obama" queries and capturing the top 100 results. Such retrospective discoveries are compelling, but systematically exploiting this type of contextual information will require more detailed rules and heuristics for determining the online resources that are most likely to contain it.

Potential Selection Rules or Heuristics

One important factor in defining crawling and selection rules is domain name. The blog pages in our Obama data set come from 2564 domains. Table 2 lists the 20 domains that account for 741 (18%) of the blog pages. The relatively large number of blog pages from these domains suggests that it would be worthwhile to weigh them more heavily in future crawls.

Table 2 - Top 20 Domains of "Barack Obama" Blog Pages with at Least One Link to YouTube Video

Domain Name	Pages in Domain
blog.myspace.com	201
seattleforbarackobama.blogspot.com	69
my.barackobama.com	60
www.mydd.com	51
howieinseattle.blogspot.com	42
www.personaldemocracy.com	40
www.techpresident.com	40
www.huffingtonpost.com	33
www.thecarpetbaggerreport.com	27
blogometer.nationaljournal.com	20
whitehouser.com	20
digg.com	19
newsbusters.org	17
www.firedoglake.com	16
pseudomanitou.livejournal.com	15
www.dailykos.com	15
blog.washingtonpost.com	14
blogs.abcnews.com	14
thinkonthesethings.wordpress.com	14
proctoringcongress.blogspot.com	14

Perhaps even more important are domains of pages that receive in-links from blog pages. There are a large number (53370) of in-linked domains, but the top 20 domains (see Table 3) account for 29060 (17%) of all in-links. The high numbers for YouTube were expected, because this data set is specifically pages that include at least one link to YouTube. We plan to further investigate whether some of the other highly represented domains provide a disproportionately high amount of information relevant to the videos (meaning the domains should given high priority when attempting to get contextual information), or because there are simply many extraneous links from blogs to those domains (meaning the domains should be given low priority when attempting to get contextual information). We also plan to analyze a sample of the 38,760 (73%) of domains associated with only one in-linked page. If a considerable amount of useful contextual information resides in singleton domains, domain-based filtering to get contextual information will not be a useful approach.

Table 3 - 20 Domains with Most In-Links from Blogs

Domain Name	In-Link Count
www.youtube.com	9433
en.wikipedia.org	2186
www.washingtonpost.com	1821
www.nytimes.com	1702
www.amazon.com	1587
YouTube.com	1378
www.dailykos.com	1231
news.yahoo.com	1016
www.barackobama.com	885
news.bbc.co.uk	848
icestationtango.blogspot.com	777
mediamatters.org	753
www.msnbc.msn.com	731
www.cnn.com	697
video.google.com	684
del.icio.us	621
www.myspace.com	610
www.huffingtonpost.com	596
feeds.wired.com	533
www.counterpunch.org	518

Another potentially useful factor is depth of links. Shallow links - the most extreme case being pointers to entire domains (e.g. http://www.cnn.com) – may be less useful than deep links in yielding contextual information for three reasons: (1) they are less likely to lead to contextual information about a particular video; (2) their content is subject to change, so that whatever someone linked to at the time is not likely to appear there anymore (having either been taken down entirely or moved to another URL); and (3) the logistics of capturing all (and only) the files needed to render the front pages of sites is challenging, in comparison to capturing specific pages at deeper URLs. This suggests that there may be utility in placing higher priority on crawling pages that have longer URLs, as measured by the number of "/" characters that appear after the initial http://. Of all the blog pages, there are 2505 (61%) that contain three or more slashes; of all the out-links from the blog pages, 60875 (36%) contain three or more slashes. This is still a larger number of pages than a repository may want to crawl simply to provide further contextual information about the videos, but it could considerably narrow the likely candidates for crawling to a subset of pages that are more likely to be (a) focused on a specific topic and (2) feasible to crawl after the fact.

Future Directions

In this paper, we report preliminary findings from only one collecting area (U.S. presidential election campaign), and within that area, only one subset of the materials (related to Barack Obama). We will analyze materials related to other candidates, as well as materials related to other socially important collecting areas. For almost a year, VidArch has 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 preliminary 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 postings that link to the video, that it might not be worthwhile to expend resources on capturing even more blog pages.

Future research will be needed, in order to determine more specific implications for crawling parameters. Building collections of material from the Web requires a translation of appraisal and collection development principles into processes that can be carried out by machines. Potential approaches can vary across at least three dimensions: environments crawled (e.g. blogosphere, YouTube), access points from those environments used as crawling or selection criteria (e.g. number of views, primary relevance based on term matching, number of in-links, channel or account from which an item was submitted), and threshold values for scoping capture within given access points (e.g. 100 most relevant query results, at least 5 in-links) [25]. We see great promise in further investigation of crawling approaches that integrate environments, access points and thresholds in different ways, in order to best meet that needs of those who are building and managing digital collections.

Within the context of collecting blog pages – as sources of contextual information, as contextual information bridges, and as sources to be collected in themselves – selection can be informed by analysis of link and use patterns. These data can help curators to determine what blogs or blog pages to select. There is a growing body of research that 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 [26]. More generally, "hyperlinks and search engines play a key role in funneling Web users to a handful of sites," [27] so that "although the range of online content is vast, the range of sites that users actually visit is small" [28]. Our data from the blog pages also suggests that a small number of domains account for a large number of the out-links from blogs. Further analysis may allow us to determine that some 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 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, such as other entries on the same page, navigational elements, blogrolls (links to other favorite blogs), and comments. A parser or crawler cannot do this easily, because there is little consistency in underlying tags. A second question is how best to detect and systematically filter the large number of unrelated out-links in some blog pages. Our investigation has revealed a small number of blogs that contributed substantially to the set of "false positive" links, based on including long lists of unrelated resources in their navigation areas. Finally, selection activities could be greatly facilitated by automated or semi-automated detection of unrelated blogs that systematically provide distractor links. One example is a musical fan blog that was mirrored in many locations, which provided many links to music videos unrelated to the election, but appeared in the blog crawls, because each page included a link to the Barack Obama campaign site.

Our initial investigation suggests there is great promise in using blog content to provide contextual information about and to inform the selection of videos from YouTube. We have identified numerous possibilities for further investigation. The time is ripe to formulate crawling and filtering rules that can operationalize selection criteria in an online environment.

Acknowledgements

The authors wish to thank Rob Capra, Paul Jones, Gary Marchionini, Terrell Russell, Chirag Shah, and Yaxiao Song for their insights and input on various aspects of this project, and Fred Stutzman for sharing and helping us to use the blog crawl data. This work 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

[1] Pew Research Center for the People & the Press, Social Networking and Online Videos Take Off: Internet's Broader Role in Campaign 2008. (Washington, DC, 2008).

[2] Helen R. Tibbo, Christopher A. Lee, et al, VidArch: Preserving Meaning of Digital Video over Time through Creating and Capture of Contextual Documentation, Proc. IS&T Archiving, pg. 210-215 (2006).

[3] Lee Rainie, Pew Internet Project Data Memo: Video Sharing Websites (Pew Internet & American Life Project, Washington, DC, 2008).

[4] Carol Darr & Joseph Graf, Poli-Fluentials: The New Political Kingmakers (Washington, DC, Institute for Politics, Democracy & the Internet, 2007).

[5] Vassia Guerguieva, "Voters, MySpace and YouTube: the Impact of Alternative Communication Channels in the 2006 Election Cycle and Beyond," Social Science Computer Review, 26, 3 (2008).

[6] Costas Panagopoulos, "Technology and the Transformation of Political Campaign Communications," Social Science Computer Review, 25, 4 (2007) pg. 423-424.

[7] Joseph Graf, The Audience for Political Blogs: New Research on Blog Readership (Washington, DC, Institute for Politics, Democracy & the Internet, Georg Washington University, 2006).

[8] Lars Kirchhoff, Axel Bruns & Thomas Nicolai, Investigating the Impact of the Blogosphere: Using PageRank to Determine the Distribution of Attention, Proc. AoIR (2007).

[9] Andrea Hinding, "Inventing a Concept of Documentation," Journal of American History, 80, 1 (1993) pg. 168-178.

[10] Helen Samuels, "Who Controls the Past: Documentation Strategies Used to Select What Is Preserved," American Archivist 49 (1986) pg. 109-24.

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

[12] Adrian Brown, Archiving Websites: A Practical Guide for Information Management Professionals (Facet, London, 2006).

[13] Julien Masànes, Web Archiving (Springer, New York, 2006).

[14] Pearce-Moses, Richard, and Joanne Kaczmarek. "An Arizona Model for Preservation and Access of Web Documents." DttP 33, 1 (2005) pg. 17-24.

[15] Donna Bergman, Collection Synthesis, Proc. JCDL, pg. 253-262 (2006).

[16] Alexandros Ntoulas, Petros Zerfos & Junghoo Cho, Downloading Textual Hidden Web Content through Keyword Queries, Proc. JCDL, pg. 100-109 (2005).

[17] Sriram Raghavan & Hector Garcia-Molina, Crawling the Hidden Web, Proc.VLDB, pg. 129-138 (2001).

[18] Jayant Madhavan & Alon Halevy, Crawling through HTML forms, Google Webmaster Central Blog, April 11 (2008), http://googlewebmastercentral.blogspot.com/2008/04/crawling-through-html-forms.html

[19] Chirag Shah & Gary Marchionini, Preserving 2008 US Presidential Election Videos, Proc. IWAW (2007).

[20] Christopher A. Lee, 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.

[21] Daniel Drezner & Henry Farrell, "The Power and Politics of Blogs," Public Choice, 134, 1-2 (2008) pg. 15-30.

[22] Laura McKenna & Antoinette Pole, "What Do Bloggers Do: An Average Day on an Average Political Blog," Public Choice 134, 1-2 (2008) pg. 97-108.

[23] Edwards Places Campaign Headquarters in NC, October 23 (2007), http://www.youtube.com/watch?v=lb8r0XIZs7Y

[24] Phil de Vellis, I Made the "Vote Different" Ad, The Huffington Post, March 21 (2007), http://www.huffingtonpost.com/phil-de-vellis-aka-parkridge/i-made-the-vote-differen_b_43989.html

[25] Robert Capra, Christopher A. Lee, et al, Selection and Context Scoping for Digital Video Collections: An Investigation of YouTube and Blogs, Proc. JCDL (2008).

[26] Lada A. Adamic & Natalie Glance, The political blogosphere and the 2004 US election: divided they blog, Proc. 3rd international workshop on Link discovery, pg. 36-43. (2005).

[27] Matthew Hindman, Kostas Tsioutsiouliklis & Judy A. Johnson, Measuring Media Diversity Online and Offline: Evidence from Political Websites, Research Conference on Communication, Information and Internet Policy (2005).

[28] Matthew Hindman, 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 (Mahwah, NJ: Lawrence Erlbaum Associates, 2006) pg. 327-347.

Author Biography

Christopher (Cal) Lee is Assistant Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill, where he teaches classes in digital curation, archives, records management and collection development. His research interests include digital preservation, standardization, significant properties of digital objects, selection, contextual information in digital collections, and professional education for digital curation. He has an MSI and PhD from the University of Michigan School of Information.

Helen Tibbo is Professor in the School of Information and Library Science at the University of North Carolina - Chapel Hill. She teaches in the areas of archival and records management, and digital curation. Dr. Tibbo is a co-investigator on the NSF-sponsored VidArch project with Gary Marchionini, Cal Lee, and Paul Jones. Dr. Tibbo is also co-principle investigator for the Mellon-sponsored Developing Standardized Metrics Project, and directs the NHPRC Electronic Records Research Fellowships. She is PI for the DigCCurr project.