SoLoGlo - A Service to Archive, Analyze, and Link Social, Local, and Global News

Martin Klein, Peter Broadwell, Todd Grappone, Sharon Farb; University of California Los Angeles, Los Angeles, California

Abstract

Web-based global news content, especially when delivered via social media, is highly dynamic and hence subject to the detriments of link rot and content drift immediately after publication. To address these issues, we are building SoLoGlo, a capture and archival service for social media collections that supports the longevity of multi-perspective histories of world events. SoLoGlo incorporates functions to collect and preserve content, proactively archiving embedded web resources, performing rapid analysis and data mining tasks, and linking related content between otherwise disjoint collections.

Introduction and Motivation

The UCLA Research Library has a long-standing history of collecting ephemera from areas of crisis around the world. Most recently, under the International Digitizing Ephemera Project [1], we have been collecting material from activists involved in the Iranian Green Movement in 2009 and the 2011 Egyptian Revolution based in Tahrir Square in Cairo. This material comes in various formats: analogue, e.g., flyers and handouts, as well as born-digital, e.g., images and cell phone videos. This variety of formats already poses a challenge to building digital and web-accessible collections. In addition, we aspire to incorporate related social media content into these collections as an additional, rich source of information and individual perspectives on these events.

Regarding the collection of social media content, however, we noticed a lack of orchestrated efforts within our institution. Individual scholars had repeatedly approached the Research Library with requests to host and preserve their Twitter collections, yet rather than one central campus unit being responsible for collection building and sharing of data, the current reality is that various individual researchers create their own ad-hoc solutions for social media collection that are often short lived, rarely aimed at data preservation, and usually are not suitable for collaboration with other scholars. While this state of affairs has produced valuable collections of social media data – for example, nearly 421,00 tweets related to the Tahrir Square uprising – it convinced us that the Research Library should assume the role of the central entity providing a service to collect Twitter data. Looking deeper into the acquired collections, we further realized that such a service also should include archival functions and offer rapid analysis and data mining capabilities. Because the Research Library holds a very rich set of special collections, finding and linking related content between social media data and otherwise disjoint special collections is a natural next evolutionary step in the process of building such a service.

In this proposal, we introduce SoLoGlo, our service for archiving, analyzing, and linsking **so**cial, **lo**cal, and **glo**bal news events. We briefly describe the service's three fundamental pillars and provide motivational examples of their use.

Assembling an Open-Source Capture System

The first pillar that forms the basis for SoLoGlo is the capture of topically relevant Twitter data, including the pro-active archiving of referenced web resources and embedded media. SoLoGlo is not the first platform to offer a comprehensive tool set for the capture of Twitter data. We have discovered a broad variety of software tools and libraries to assist in the collection of tweets, presumably the results of the numerous disconnected ad-hoc efforts we observed earlier. Rather than implementing yet another tool, we prefer to adapt, extend, and if necessary combine existing software packages while developing SoLoGlo. Our evaluation of the extant open-source tools in this space took into account their available features, interfaces, and extensibility, as well as the potential for contributing to and benefiting from the community of users and developers around each tool. At the moment, we have built most of SoLoGlo upon the functionality of the Social Feed Manager (SFM) [2], which we found to be one of the more robust and feature-rich software packages available. In addition, we have merged the real-time capture abilities of the SFM with the historical Twitter search functionality of twarc, another prominent open-source Twitter capture tool [3]. The resulting software package enables us to capture tweets filtered by keywords, user handles, location, and temporal boundaries.

Pro-Active Archiving of Web References

The second aspect of this first pillar, the pro-active archiving of web references, is the first novel contribution of SoLoGlo. Many link rot studies in the past have shown that the web is very dynamic; resources change and disappear very frequently with the effect that links are subject to rot as soon as they are published. The same holds true for web resources referenced from within Twitter [4] and thus archiving such web resources provides the most value if it happens in near real-time, meaning as soon as the referencing tweet is captured. Today's web archiving infrastructure allows for the pro-active archiving of a web resource simply by submitting its URI to an archival service. We utilize two of the most popular and widely used web archives for this pro-active service: the Internet Archive and archive.today. Cognizant, however, of the fact that it can be risky and also cumbersome to rely on third-party web archives to preserve these resources, especially when building local collections for use in active

research, we also store these resources whenever possible in our own local archive.

Adaptive Visualizations

The second pillar of SoLoGlo implements methods of visualizing and analyzing the collected social media data. Clearly, we cannot provide visualizations and tools to meet the needs of all scholars interested in using SoLoGlo. However, we see value in implementing functions to help users gain a high-level understanding of the data. For example, we offer a multi-component overview of the frequencies of prominent terms in a Twitter stream captured via SoLoGlo, combined with a display of a basic sentiment analysis of the terms in the tweets. During a live capture, this visualization updates in real-time, potentially allowing researchers to tweak their search parameters based on observed shifts in the recorded conversations. This evolution also can be reviewed by replaying the visualization after the capture has completed. Figure 1 provides a screenshot of this visualization.

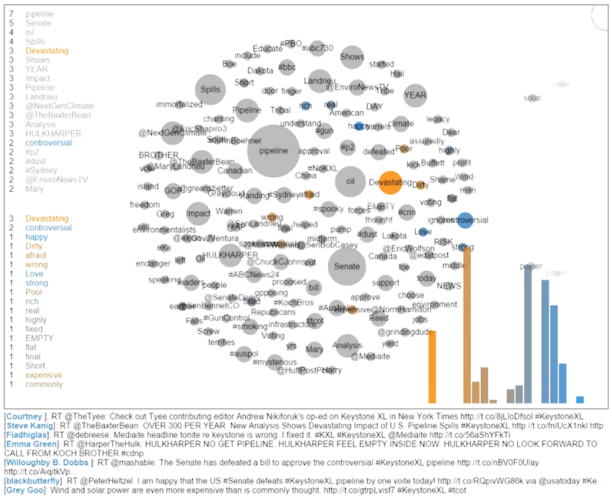

***Figure 1**. Term frequencies from tweets related to the U.S. Senate's vote on the Keystone XL oil pipeline in November 2014*

Discovering and Linking Related Resources

The third pillar of SoLoGlo provides a means to link between related resources from collections that otherwise would remain disjoint. To demonstrate this functionality, we present the following example: UCLA has been archiving broadcast television news, including closed captions, for the last 10 years and has built a large collection of mostly US-based national news coverage. UCLA also has created a collection of Twitter data related to the Egyptian revolution based around Tahrir Square in 2011. The collected tweets originated from within a 200-mile radius around Cairo and were therefore much more likely to have been sent from activists on the ground rather than by members of the American mass media. Figure 2 shows an example of such a tweet that also references a web resource. Using the text of the tweet as the input parameters to a search against UCLA's indexed news collection, we are able to discover the originally broadcast news coverage referenced in the tweet (Figure 3). This link therefore establishes a novel connection between directly related resources and highlights one of the primary motivations for discovering such connections: to facilitate the construction and preservation of multi-narrative histories of an event that incorporate alternate perspectives to those of "mainstream" media and political authorities.

***Figure 2.** Tweet containing a web reference*

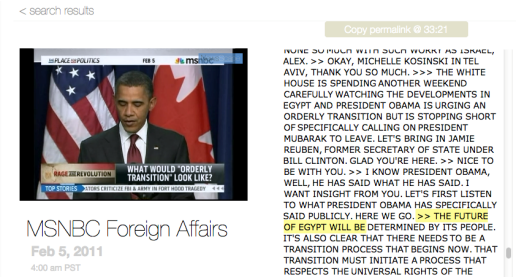

***Figure 3.** Referenced resource discovered in a television news archive*

This example also further demonstrates the power of enhancing social media collections with additional related data such as the contents of embedded links: the web reference in the tweet is now rotten and returns a 404 "Page not found" error. SoLoGlo's nascent inter-collection linking feature allowed us to find the referenced television broadcast in this case, but the media item's immediate context within the web resource referenced in the tweet has been lost. Full deployment of SoLoGlo – and in particular its first "pillar," the pro-active archiving of web references in social media – would have prevented this loss.

Figure 4 provides an example of another capability that the inter-archive linking feature of SoLoGlo can provide: the direct comparison of materials related to a given event across different archives and media types. In this example, we have generated an interactive plot of the volumes of the tweets in UCLA's Twitter collection from the Egyptian revolution of 2011 described above (blue series, appearing slightly later), and the number of on-air appearances, spoken or on-screen, of terms related to the same events (e.g., "Egypt," "Tahrir") in UCLA's television news archive during the same period (red series). The scales of the series are quite different: the frequency of Twitter messages peaks at 10,000 tweets per hour, while the detected televised appearances of the search terms range from 0 to 500 per hour. Other discrepancies due to the nature of each media type also are evident, such as a natural decrease in Twitter conversations in the

early morning, as well as the customary cycles of television news coverage. And even though it is generally impossible to find a truly "representative" sample with such collections, the simple comparison graph in Figure 4 still reveals tantalizing hints of significant phenomena within the social media and television news coverage of the Egyptian revolution, as well as the events themselves. For example, it is possible to detect in the graphs such happenings as the Internet blackout in Egypt beginning Friday, January 28, the initial surge in U.S. television coverage of the protests on January 29, and the peaks in both media types around Hosni Mubarak's resignation on February 11.

Automating the discovery and linking of related resources in disparate collections will require considerable future development, in some cases incorporating still-emergent technologies for information retrieval, such as advanced automated natural language processing, audio analysis, and image and video parsing. Although this component remains under active development, we are configuring SoLoGlo with a modular architecture to enable it to incorporate new data mining tools and archive interoperability standards as they continue to emerge.

Conclusions

We introduce SoLoGlo, a global news capture service with the ability to archive web resources embedded in social media. SoLoGlo makes a novel contribution to current web archiving practices by proactively archiving linked resources from social media via an adaptive, user-friendly interface that includes rapid analysis and data mining of materials as they are ingested. The service also enables linkages to related materials in separate archives, providing automatic description and summary services that facilitate the discovery of relevant records in other collections. SoLoGlo thus actively discourages the creation of siloed data collections; it also seeks to deter the development of standalone, one-off tools by using pre-established frameworks and software and encouraging a collaborative, incremental approach to tool creation and collection development. SoLoGlo therefore supports the preservation and analysis of multi-perspective histories of world events, revealing the interconnections of related media records to facilitate the comparison and juxtaposition of disparate perspectives on the same event, generating multi-narrative histories that will extend present scholarly perspectives and enrich future scholarship.

References

[1] International Digitizing Ephemera Project, http://digital.library.ucla.edu/dep

[2] Social Feed Manager, http://social-feed-manager.readthedocs.org/

[3] twarc, https://github.com/edsu/twarc

[4] Salah Eldeen H, Nelson ML "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?" in Proceedings of TPDL 2012, 125-137.

Author Biographies

Martin Klein is a scientist in the Research Library at UCLA. He has conducted research and developed software in the realm of web archiving and digital libraries for the past 10 years. Prior to his tenure at UCLA, Martin was a postdoctoral scholar in the Research Library at the Los Alamos National Laboratory.

Peter M. Broadwell received a Ph.D. from UCLA in 2010, where he applied his background in computer science to several experimental Digital Humanities projects. He subsequently worked for three years as a Council on Library and Information Resources postdoctoral fellow at the UCLA Library, and is now an Academic Projects Developer in UCLA Library Digital Initiatives, exploring the intersection of archives, technology, and scholarly inquiry.

Todd Grappone is the Associate University Librarian for Digital Initiatives, Information Technology, the College Library, Teaching and Learning Services and the East Asian Library and a senior administrator in the UCLA Library. Todd has extensive experience in developing digital libraries, discovery systems and interoperable digital archives.

Sharon E. Farb is UCLA's Associate University Librarian for Collection Management and Scholarly Communication. She specializes in legal and policy issues that impact libraries, archives, and museums in the areas of intellectual property, copyright, licensing, privacy, intellectual freedom, stewardship, and data curation, with an overall focus on providing the broadest possible access to scholarly information and recorded knowledge. .

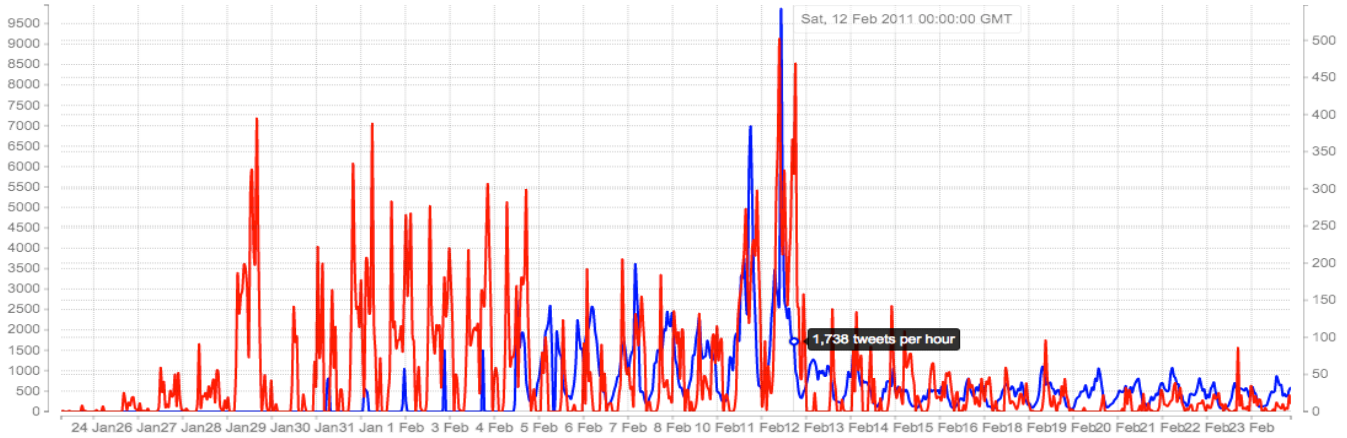

***Figure 4**. Comparative volumes of Twitter messages and television news coverage related to the Egyptian revolution of 2011*