The Missing Link: A Preliminary Typology for Understanding Link Decay in Social Media

Shawn Walker[1], Sheetal Agarwal[1,2]
[1]University of Washington
[2]Kwilt Strategy

Abstract
In this study we offer a preliminary typology of types of link ephemerality that can occur and affect content of social media data. The preliminary typology emerged while collecting and analyzing social media data, videos, and web links in three distinct projects. The Internet Archive was also queried to determine if each URL was archived and the timespan between tweet production and archiving. Since certain types of URL change may have a negative impact on our research, we need a more nuanced understanding of the ways in which URLs in social media data change. Understanding this change moves us beyond binary inclusion/exclusion categories of accessible ("404 Not Found") and inaccessible. Our analysis and typology informs researchers' data collection strategies as well as opening up the possibilities on setting more nuanced boundaries on data collection and inclusion.
Keywords: Social media; archives; ephemerality; methods; data collection
doi: 10.9776/16590

Contact: stw3@uw.edu, sheetal@kwiltstrategy.com

1 Introduction

One of the oft-discussed challenges of working with social media data, especially Twitter, is the ephemeral nature of the data (Bruns, 2012; Driscoll & Walker, 2014). If researchers do not collect social media data in real-time, it disappears. While historical data is available from GNIP; the cost is out of reach of most researchers. To combat the ephemerality of social media data, researchers often collect the data in real-time and archive it for use in one or more research projects. It's easy to think that we've dodged that bullet of ephemerality by constructing archives of tweets or other social media data about a specific event or topic. However, these archives only preserve the content of the social media post but does not preserve content within the dynamic components embedded or linked to in each post such as URLs, images, or videos. We contend that hyperlinks in social media posts should be conceptualized as an extension of the content of the post and thus merit inclusion in the analysis of social media content. Simply viewing the links is not enough though, link decay can be problematic, especially for researchers who often analyze social media data months after it is captured. In this study we offer a typology of types of link decay that can occur and affect content analysis of social media data.

2 The Importance of Hyperlinks

At the root of the internet is a hypertext system in which data is stored in a network of nodes connected by links (Smith et al, 1988). These links, commonly called hyperlinks (URLs), serve as the primary mechanism that connects nodes in the web to one another, and are technological affordances that allow seamless connection between one website and another (Park & Thelwall, 2003).

Estimates of link decay (also known as: half-life, death, accessibility, persistence, and link rot) mainly come from studies of links in journal articles and range 31% - 39% (Isfandyari-Moghaddam, & Saberi, 2011; Goh & Ng 2007; Dimitrova, & Bugeja, 2007). These studies use a combination of automated analysis of error codes returned by web servers or rely on researchers visiting each of the URLs. These methods only detect obvious cases where the destination of the URL returns an error. In addition, focus on "404 not found" errors of links in academic journals tell us little about the other ways links can decay, disappear, or erode. Additional research and more comprehensive frameworks are needed to understand the complexities of link decay -- especially in relationship to social media.

2.1 Links in Social Media

Users of the microblogging service, Twitter, often include hyperlinks in tweets, many times because they are limited to 140 characters that do not allow them to fully express or contextualize their thoughts (Gao, Zhang, Li & Hou, 2012). We ground this conceptualization of links as extensions of tweets by drawing from the concept of web spheres (Schneider & Foot, 2005). Web spheres, they argue, are not just the

websites but the “dynamically defined digital resources spanning multiple websites deemed relevant or related to a central event, concept or theme, and often connected by hyperlinks”. In understanding the communicative actions of Twitter users and the content of their tweets, hyperlinks provide important means of contextualizing and providing relevant information that cannot be captured in 140 characters.

3 Uncovering the need for a Link Decay Topology

A preliminary typology offered in this paper emerged while collecting and analyzing social media data, videos, and web links in three distinct projects. The first project examined YouTube videos and blog during the 2008 US Presidential Election to understand the role of blogs in propagating viral political videos (Nahon et al., 2011). The second project focused on the use of social media in the Occupy Wall Street movement, where 31,000 seed URLs embedded in tweets related to the Occupy Wall Street movement were coded based on the type of resource linked each URL linked to (Agarwal et al., 2014). Example coding categories for the URLs included mainstream media site, government site, celebrity site, and occupy-related site. The third used tweets and links embedded in tweets to examine rumor propagation after the Boston Marathon Bombings (Starbird et al, 2014). The experiences with URLs during the data collection and analysis process in each of these projects pointed to more complex types of URL decay -- not just either found or (404) not found.

Instead, link decay issues related to the loss of content due to broken links, a shortened URL would redirect to a new location, or domains expired and forwarded to new locations. We also encountered issues with page content quickly changing to reflect evolving information about the emergency event, as a result each time a page was accessed the information contained in it was no longer in sync with the tweet that referred to the content. Other times, we ran into large-scale deletions of rumor content by social media sites such as reddit.

From the experiences described above, patterns began to emerge, lending towards a preliminary typology of types of link decay that we offer below. These six types of link decay occurred across projects and early insights indicate are useful means to account for the ephemerality of link data.

1. D*ead on Arrival Links.*
   Links that are dead at the time of creation of the tweet leading to an error page. For example, the URL may link to http://www.washington.ed instead of http://www.washington.edu, missing the “u” in the “.edu”. In most cases, the link seemed to be broken due to a typo.
2. *Changed content.*
   Content that changes after the link has been posted. For example, the content of the BBC homepage is updated as news develops so the content of these pages may change after the social media post was created.
3. *Deleted content.*
   A user or service provider deleted an account or specific piece of content. For example protesters tweeted links to images of police activity during a protest, deleting that content after the protest had concluded. We also ran into numerous YouTube videos that were removed for Terms of Service (Tos) violations.
4. *URL shortener decay.*
   A shortened URL is either recycled to point to a different link or the shortner service is shutdown. We found cases where a shortened URL expanded to URL over time. In other cases, we’ve seen URL shortening services such as Tr.im and litturl.com shutdown leaving thier shortneted links unresolveable.
5. *Redirected Link.*
   A page loads, but now automatically forwards to a new page. For example, when a domain had expired, we found that the URL was redirected to a temporary landing page stating that the domain was for sale instead of the original content.
6. *Embedded content decay.*
   The content embedded inside of the page changes or is removed. We saw this happen in blog posts when YouTube videos were deleted. The embedded video no longer played, displaying the message that this content was no longer available.

3.1 Next Steps

To expand on the preliminary data and typology we gathered through experiences in collecting social media data, a rigorous coding process will be applied. To measure the stability of linked URLs, copies of the URLS will be archived at tweet inception during real-time data collection and at regular weekly intervals over a three month period after data collection. A random sample of these URLs will be coded to document the types of changes that occur over that time period. Through emergent coding processes we will develop a more robust typology.

4 Conclusion

In developing this typology important questions emerge about link decay in social media datasets and how we address them as researchers. Existing web archives such as the Internet Archive and WebCite may seem like possible solutions, but they crawl and archive sites on their own time schedule. The timing of their crawls and the fact that tweets contain links to specific pages make these archives poor tools for accurately capturing the context of a tweet or Facebook post.

As a result, the solution is to archive links mentioned in social media posts at or very close to the time of production. This brings a whole new set of challenges to social media data collection since the current tools we use do not facilitate this. Also, web archives are much larger than their social media cousins. Social media data is primarily text, so it compresses well and doesn't take up much space. Web archives include not just the text (HTML) of the link, but all of the elements within a page such as images and videos. Since certain types of URL change may have a negative impact on our research, we need a more nuanced understanding of the ways in which URLs in social media data change. Understanding this change moves us beyond binary inclusion/exclusion categories of accessible ("404 Not Found") and inaccessible. Our analysis and typology informs researchers' data collection strategies as well as opens up the possibilities on setting more nuanced boundaries on data collection and inclusion.

5 References

Agarwal, S. D., Bennett, W. L., Johnson, C. N., & Walker, S. (2014). A Model of crowd enabled organization: Theory and methods for understanding the role of Twitter in the occupy protests. *International Journal of Communication*, 8(0), 27.

Bennett, W. L., Segerberg, A., & Walker, S. (2014). Organization in the crowd: Peer production in large-scale networked protests. *Information, Communication & Society*, 17(2), 232–260. http://doi.org/10.1080/1369118X.2013.870379

Bruns, A. (2011). How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Information, Communication & Society*, 15(9), 1323-1351. http://doi.org/10.1080/1369118X.2011.635214

Bruns, A., & Stieglitz, S. (2014). Twitter data: What do they represent? *It-Information Technology*, *56*(5).

Dimitrova, D. V., & Bugeja, M. (2007). The half-life of internet references cited in communication journals. *New Media & Society*, *9*(5), 811–826. http://doi.org/10.1177/1461444807081226

Driscoll, K., & Walker, S. (2014). Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, *8*(0), 20.

Foot, K., & Schneider, S. M. (2006). *Web Campaigning*. The MIT Press.

Goh, D. H.-L., & Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, *58*(1), 15–24. http://doi.org/10.1002/asi.20513

Isfandyari-Moghaddam, A., & Saberi, M.-K. (2011). The Life and Death of URLs: The Case of Journal of the Medical Library Association. *Library Philosophy and Pracitce*, *7*. Retrieved from http://digitalcommons.unl.edu/libphilprac/592/

Nahon, K., Hemsley, J., Walker, S., & Hussain, M. (2011). Fifteen minutes of fame: The power of blogs in the lifecycle of viral political information. *Policy & Internet*, *3*(1), 1-28.

Park, H. W., & Thelwall, M. (2003). Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer-Mediated Communication*, *8*(4), 0–0. http://doi.org/10.1111/j.1083-6101.2003.tb00223.x

Smith, J., Weiss, S., & others. (1988). Hypertext. *Communications of the ACM*, *31*(7), 816–819.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In iConference 2014 Proceedings (p. 654–662). doi:10.9776/14308