INDICATORS THAT TWEETING MAY IMPROVE THE DETECTION OF NEWS ARTICLES FOR WEB ARCHIVES

SHAWN M. JONES, LOS ALAMOS NATIONAL LABORATORY

1. Introduction

The archival of news stories is important because of the need to reconstruct the history of what occurred before, during, and after a particular event. For example, the sentiment within the Republican Party at the beginning of the 2016 US Presidential Primary was much different than it is now. Someone attempting to study this change will need access to news stories because they embody the knowledge that was available at the time. As publishers change ownership and software, access to past news stories may no longer be possible, making web archives a more important component of cultural memory.

Currently, web archives employ the use of crawlers[2] that follow links from web page to web page in order to find pages to archive. Some of the pages discovered in this manner are indeed news stories, but they are only archived if discovered while the crawler is following links. There is almost no archiving triggered by those creating web pages, making crawling a best effort solution by current web archiving entities.

We hypothesize that, even though online news articles are promoted by their publishers, those that are shared via social media are more likely to be discovered by the archiving crawling software, thus such articles have a better chance of being archived and are also archived more quickly after publication.

This project seeks to find indicators to support further research on this hypothesis. Support for this hypothesis can inform the publication and outreach processes of organizations that want to ensure that web archives discover their content.

2. Methodology

Using the Local Memory Project[1] we discovered five different media publishers in the Los Angeles area. The publishers chosen were required to have an active web site, with news articles published throughout the week, complete with dates and times of publication, and an active twitter feed that shared output from the publisher. Twitter IDs were gathered from the Local Memory Project, but verified with the website of the publisher. We selected a variety of different types of news publishers, including a national newspaper, the *Los Angeles Times*, a local television station, *CBS Los Angeles*, a college newspaper, *the Daily Trojan*, and two Spanish language newspapers, *Hoy Los Angeles* and *La Opinión*. These publishers, their subscriber counts, and follower counts, are listed in Table 1.

[1] `http://www.localmemory.org/`

TABLE 1. Information for news publishers used in this experiment

Publisher	**URI**	**Twitter Username**	**Subscribers**	**Twitter Followers**[2]
CBS Los Angeles	http://losangeles.cbslocal.com/	@CBSLA	est. 500,000[3]	137,000
Daily Trojan	http://dailytrojan.com/	@dailytrojan	65,000[4]	20,400
Hoy Los Angeles	http://www.hoylosangeles.com/	@hoylosangeles	137,796 - 1.2 million[5]	6,290
La Opinión	http://laopinion.com/	@laopinionla	480,000 - 500,000[6]	36,000
Los Angeles Times	http://www.latimes.com/	@latimes	1.5 million - 2.4 million[7]	2.3 million

Once our five publishers were selected, we collected URIs and publication dates for articles published between the dates of October 10, 2016 and October 16, 2016, inclusive. After article URIs were gathered from each publisher's web site, the Memento framework[3, 5, 4] was used on Ocotober 20, 2016 and again on November 3, 2016 to discover archived versions of those articles in 22 web archives. These archived versions will hereafter be referred to as **mementos**.

Using twarc[8], we then gathered all URIs tweeted by these publishers on Twitter within the date range. Publishers occasionally retweet the same article, so duplicate article URIs from twitter were removed, keeping the earliest tweet. Publication times were converted to the GMT time zone to ease direct comparisons to tweet times and memento times.

If resolving an article URI, regardless of from where it was gathered, resulted in an HTTP redirect, then that redirect was followed to the final URI in the chain. This way we could compare URIs directly to each other. For URIs gathered from publisher web sites, the final URI in the redirect chain was used to discover mementos.

As shown in Figure 1, some articles had been archived, and then republished, resulting in publication dates after the datetime of their first memento. Because the gathered publication date did not reflect the original publication date, these articles were removed. Their counts are shown in Table 2.

[2]As of November 26, 2016

[3]As it is the website associated with a local TV station, *CBS Los Angeles* does not have a subscription model like other publishers. The mediakit at http://wayne.cbslocal.com/about-us/CBSLocalMediaKit.pdf provides information for the entire cbslocal.com domain, but not individual local stations. It states that the cbslocal.com domain receives 60 million monthly unique visitors. This includes 146 radio and tv stations spread across the United States. Estimate is calculated by dividing 60 million by 146 and rounding up to the next hundred thousand, assuming that Los Angeles, being more populous, would spur more subscribers.

[4]http://dailytrojan.com/wp-content/uploads/2009/09/dtrates16-17.pdf

[5]http://mediakit.latimes.com/Media/LosAngelesTimesMediaKit/Toolkit/Fast%20Facts_Hoy.pdf

[6]https://www.mediabids.com/publication/mediakit/la-opinion/?pi=40549

[7]https://mediakit.latimes.com/Media/LosAngelesTimesMediaKit/Toolkit/Fast%20Facts.pdf

[8]https://github.com/DocNow/twarc

FIGURE 1. Example of an article republished from *Hoy Los Angeles*: the article on the left was collected from our dataset and published on October 15, but we see in the archived version on the right that it was published before on September 28, which is outside of the date range of the experiment.

TABLE 2. Disposition of articles used in experiment

	CBS Los Angeles	Daily Trojan	Hoy Los Angeles	La Opinión	Los Angeles Times	Total
Articles gathered from website	337	206	308	859	2255	3965
Removing Articles republished	1	0	8	0	139	148
Articles used for experiment	336	206	300	859	2116	3817

Based on an analysis of the URI structure for these publisher web sites, we found that URIs for articles on each of these sites do not contain a query string[1] (e.g., `http://example.com/page?key=value&key2=val2`). URIs that are tweeted sometimes contain a query string for tracking purposes, but their content is not any different from the same URI without the query string, so question marks and the text that followed were removed from each URI before comparisons were made between those that were captured from the publisher's website and those that were tweeted. To further assist in these comparisons, URI fragments[1] (e.g. `http://example.com/page#section`) were removed as well.

The Memento framework allowed us to query multiple web archives during the course of the experiment. Our 3,817 articles correspond to 10,169 mementos spread across 3

TABLE 3. Diversity of mementos across web archives per publisher

Archive \ Publisher	CBS Los Angeles	Daily Trojan	Hoy Los Angeles	La Opinión	Los Angeles Times	Total
archive.is	2	0	0	0	60	62
Archive-It	4	0	0	0	65	69
Internet Archive	293	133	248	469	8895	10,038
Total	299	133	248	469	9020	10,169

different web archives: archive.is, Archive-It, and the Internet Archive. Table 3 shows the diversity of these mementos across the web archives. As the largest web archive it is unsurprising that the Internet Archive holds most of the mementos for these articles. On-demand archives allow humans to directly submit URIs for archiving. The archives archive.is is a purely on-demand archive. Archive-It allows a user to submit initial URIs for archiving, but may also crawl connected documents, making it partially on-demand. This indicates that individuals actively chose to create 62 archival copies of articles in archive.is and potentially chose 69 archival copies in Archive-It. Our hypothesis states that exposure on social media improves the chances of an archive detecting and preserving articles, therefore mementos from these on-demand archives are included because we are interested in those articles discoverable by both machines and humans.

From this data, we were able to determine which articles published during the time period had been tweeted and which articles during the time period were archived. The intention is to determine if the percentage of tweeted and archived articles is higher than those that are not tweeted, which would indicate support for our hypothesis that tweeted URIs have a higher chance of being archived.

We also examined the difference between the time of first tweet and the time an article was first archived and the mean number of mementos per article.

3. RESULTS

The results for each publisher, shown in Tables 4a to 4e, are promising for our hypothesis. In each case, the percentage of tweeted and archived is indeed higher than those not tweeted, but still archived. In Table 4a we see that the *Los Angeles Times* shows that 98.5% of those articles tweeted are archived and 74.1% not tweeted are archived, giving a separation of 24.4% between the two categories. Referring back to Table 1, we see that the Los Angeles Times has the most subscribers and Twitter followers of any publisher in our dataset. Perhaps the high number of twitter followers results in a higher chance of articles being discovered and archived.

Hoy Los Angeles, shown in Table 4b, also has high percentages, with 90.9% of all articles tweeted being archived, compared to 68.8% for those not archived, giving a separation of 22.1% between the two. On Sundays, *Hoy Los Angeles* has 1.2 million subscribers, half as many as the *Los Angeles Times*, but *Hoy Los Angeles* also has the least number of Twitter

followers in our dataset. This indicates that the number of Twitter followers may not be the only indicator of the number of articles being both tweeted and archived. Because of its affiliation with the *Los Angeles Times*, it is conceivable that an archivist has taken a special interest in archiving this site.

While most of the publishers have larger number of non-tweeted than tweeted articles, *CBS Los Angeles*, in Table 4c, has the closest number of tweeted vs. not tweeted in the entire dataset. During the week of our study, 40.2% of all articles published were tweeted. Its percentage tweeted and archived is higher, and the separation between tweeted and archived vs. not tweeted and archived is 31.4%. It does have the second largest number of Twitter followers. If we consider the idea that more Twitter followers may lead to more articles archived and consider *Hoy Los Angeles* to be an outlier, then perhaps there is some point at which a Twitter account achieves a critical mass of followers that improves its articles' chances of being archived.

For *La Opinión*, in Table 4d, 41.4% of all articles published during the week of the study were tweeted. Even though this number is close to that for *CBS Los Angeles*, the similarities end there. Compared to all other publishers, we see the highest separation of 36.9% between between those tweeted and archived vs. those not tweeted and archived. The percentage of tweeted and archived is only 65.7%, which is lower than the other three reviewed so far. There definitely appears to be an indicator that tweeting is having an effect, seeing as only 28.8% of those not tweeted were archived. Revisiting the idea of a critical mass of Twitter followers, perhaps it is reached by *La Opinión* at $36,000$. This still does not explain why the percentage of tweeted and archived is so low.

Most unlike the others, however, is *The Daily Trojan*, shown in Table 4e. Its separation between percentage tweeted and archived vs. not tweeted and archived is only 5.5%. This publisher also tweets the least number of articles in our dataset. Is it possible that there is some threshold at which the number of articles tweeted is too low to provide any meaningful indicators between tweeting and archiving. Our hypothesis barely holds for *The Daily Trojan*.

In Table 5, we see that tweeted articles have slightly more mementos on average for most publishers. *Los Angeles Times* articles generated almost 3 more mementos on average per article for those tweeted. In contrast, *Hoy Los Angeles* articles generate 0.1 more mementos for those not tweeted. *Daily Trojan* shows no difference. It is possible that running the experiment for a longer period would result in more mementos being generated. It is also possible, based on the prior results, that *Daily Trojan* is merely an outlier.

Our hypothesis relies upon the idea that a crawler has a higher chance of finding a tweeted article than one merely published on the web site. One might expect such a crawler to find tweeted articles faster and hence archive them sooner. Figure 2 contrasts the amount of time between publication and archival for tweeted and non-tweeted articles. In most cases, the articles that were not tweeted were archived faster. The sole exception is the *Los Angeles Times*, which again has far more subscribers and Twitter followers than the other publishers. This outlier may indicate that this trend does not hold for all types of publishers. Both *Hoy Los Angeles* and *La Opinión* have similar times for memento creation after publication. Being both Spanish language newspapers, perhaps they share

TABLE 4. For each publisher: Numbers and Percentages of Articles Tweeted and Archived From Period of October 10, 2016 to October 16, 2016

(A) Los Angeles Times

	Tweeted	**Not Tweeted**	**Total**
Published	335	1781	2116
Archived	330	1319	1649
Percentage Archived	98.5%	74.1%	

(B) Hoy Los Angeles

	Tweeted	**Not Tweeted**	**Total**
Published	44	256	300
Archived	40	176	216
Percentage Archived	90.9%	68.8%	

(C) CBS Los Angeles

	Tweeted	**Not Tweeted**	**Total**
Published	135	201	336
Archived	119	114	233
Percentage Archived	88.1%	56.7%	

(D) La Opinión

	Tweeted	**Not Tweeted**	**Total**
Published	356	503	859
Archived	234	145	379
Percentage Archived	65.7%	28.8%	

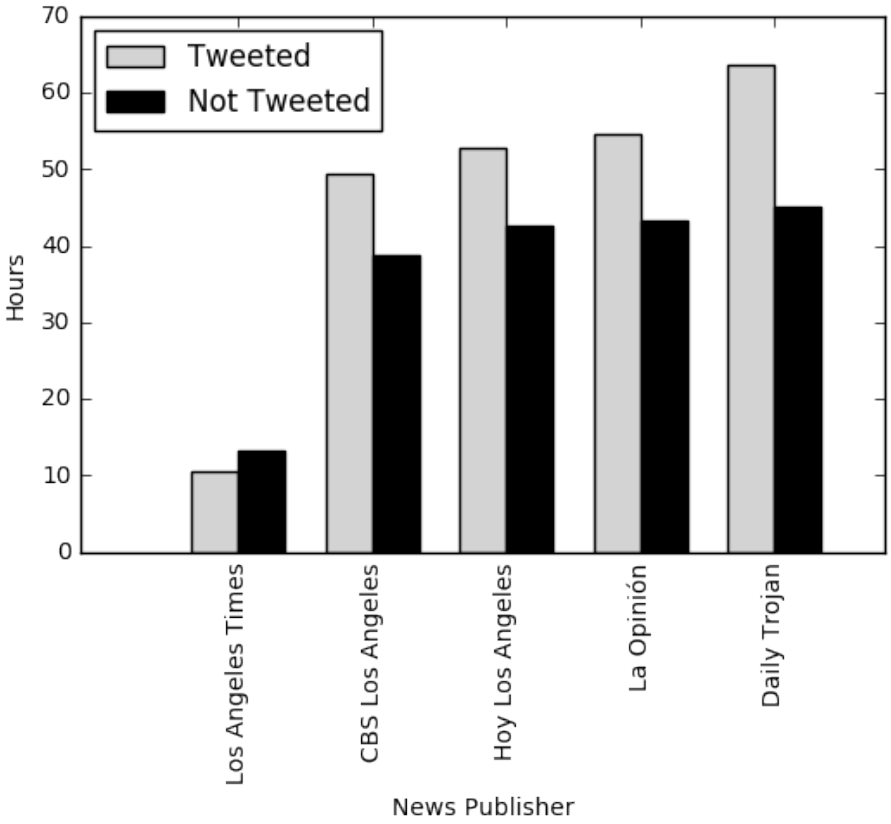
(E) Daily Trojan

	Tweeted	**Not Tweeted**	**Total**
Published	28	178	206
Archived	19	111	130
Percentage Archived	67.9%	62.4%	

TABLE 5. Mean number of mementos per article per publisher

Publisher	**Tweeted**	**Not Tweeted**
CBS Los Angeles	1.4	1.2
Daily Trojan	1.0	1.0
Hoy Los Angeles	1.1	1.2
La Opinión	1.3	1.2
Los Angeles Times	7.8	4.9

FIGURE 2. Mean time between article publication time and first memento



a subscriber base that is influencing the archiving of articles in a similar way. It is also possible that our sample size of one week is too small and that a longer study will result in the expected results.

4. Future Work

To examine potential different patterns, we attempted to use a wide variety of news outlets in the Los Angeles area, including a national newspaper, a local television station, a college newspaper, and local Spanish language newspapers. Future work could classify these media outlets into categories, such as by number of subscribers, number of social media followers, language of content, intended audience, or to which locality or region the content applies. Once categorized, one could then analyze multiple publishers of the same class. For example, based on number of Twitter followers, the *Los Angeles Times* may be in a completely different class than the other media outlets from this project. One could theoretically conduct this same study using publications with more than 1 million Twitter followers like *Los Angeles Times*, *The New York Times*[9] (31.8 million followers), *The Washington Post*[10] (8 million followers), the *Wall Street Journal*[11] (12.3 million followers), and USA Today[12] (2.8 million followers) to determine if the trends seen here apply to that class of publisher[13]. Alternatively, one could classify these publishers in terms of type of media. For example, *CBS Los Angeles* is a television station providing news for the local Los Angeles area. Perhaps conducting the study based on other local television station web sites will yield similar results to that class of publisher.

It is also possible that articles from these publications are shared over other social media platforms, and hence discovered by web archiving crawlers at those locations instead. To uncover these connections, additional studies could examine posts from sites such as Facebook[14], Pinterest[15], Google+[16], Tumblr[17], and Reddit[18]. It would also be useful, but not required, if such a study had access to an API for acquiring content from posts, especially URIs, as well as the date and time the posts were published.

Of course, there are also additional questions to answer if the hypothesis continues to gain support. Which classes of media publishers achieve better archiving coverage from posting to social media? Which social media sites correlate to better archiving coverage? Does posting to social media appear to inspire on-demand web archiving? Why are some tweeted articles archived faster than others? Which web archives contain news content? Is there some point at which a publication has achieved enough followers to build enough critical mass such that it is archived more frequently? What is the lower bound on the number of articles tweeted by a publisher for our hypothesis to hold? How are those not tweeted being discovered?

---

[9] `http://www.nytimes.com/`
[10] `https://www.washingtonpost.com/`
[11] `http://www.wsj.com/`
[12] `http://www.usatoday.com/`
[13] Twitter follower counts gathered on November 26, 2016
[14] `https://www.facebook.com/`
[15] `https://www.pinterest.com/`
[16] `https://plus.google.com/`
[17] `https://www.tumblr.com/`
[18] `https://www.reddit.com/`

5. Conclusions

Our project sought to find indicators to support further research on the hypothesis that articles that are shared via social media are more likely to be discovered by the archiving crawling software, thus such articles have a better chance of being archived and are also archived more quickly after publication.

We did discover that a higher percentage of articles tweeted are archived. We also found that, in most cases, slightly more mementos were created for those articles tweeted. This provides some support for the first part of our hypothesis, but also raises additional questions, as covered in the Future Work section.

For the second part of our hypothesis, we discovered that, for the most part, tweeted articles are not archived faster. There is only one case, the *Los Angeles Times*, where tweeted articles are archived faster. This raises similar questions about the nature of the publisher, as the *Los Angeles Times* has the most most Twitter followers in our dataset by a factor of 10. Because our study used a variety of different publishers, it is possible that our hypothesis will be better supported if publishers are classified and our experiment is performed again on one class of publisher for a longer period of time than one week.

Because of partial support for our hypothesis and outstanding questions, these results can serve as the start for more research into the connection between articles posted in social media posts and their mementos appearing in web archives.

References

[1] Berners-Lee, T., Fielding, R., and Masinter, L. RFC 3986: Uniform Resource Idntifier (URI): Generic Syntax, 2005. `https://tools.ietf.org/html/rfc3986` Accessed: 18 Nov 2016.

[2] Mohr, G., Kimpton, M., Stack, M., and Ranitovic, I. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop* (September 2004). `http://iwaw.europarchive.org/04/Mohr.pdf` Accessed: 18 Nov 2016.

[3] Van de Sompel, H., Nelson, M. L., and Sanderson, R. RFC 7089: HTTP Framework for Time-Based Access to Resource States – Memento, 2013. `http://tools.ietf.org/rfc/rfc7089.txt` Accessed: 03 Nov 2016.

[4] Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L., Ainsworth, S., and Shankar, H. Memento: Time travel for the web. Tech. Rep. arXiv:0911.1112, Los Alamos National Laboratories and Old Dominion University, 2009. `https://arxiv.org/abs/0911.1112`.

[5] Van de Sompel, H., Sanderson, R., Nelson, M., Balakireva, L., Shankar, H., and Ainsworth, S. An HTTP-based versioning mechanism for linked data. In *LDOW* (April 2010). `http://events.linkeddata.org/ldow2010/papers/ldow2010_paper13.pdf`.