

Impact of URI Canonicalization on Memento Count

Mat Kelly, Lulwah M. Alkwai, Sawood Alam,
 Michael L. Nelson, and Michele C. Weigle
 Old Dominion University
 Department of Computer Science
 Norfolk, Virginia, USA
 {mkelly,lalkwai,salam,mln,mweigle}@cs.odu.edu

Herbert Van de Sompel
 Los Alamos National Laboratory
 Los Alamos, New Mexico, USA
 herbertv@lanl.gov

Web archives return Memento TimeMaps with a list of URI-Ms for the HTTP transactions observed at archival time. TimeMaps have generally been used as a count of the number of representations of a URI-R present in an archive. However, TimeMaps may include URI-Ms for archived representations, redirections, and errors (See RFC7231). For example, 57% of the URI-Ms for `http://vimeo.com` produce an HTTP Redirect to another URI-M in the TimeMap that returns an HTTP Status OK. TimeMaps do not explicitly return a “count” value to indicate the number of mementos listed in the TimeMap that produce a non-redirecting HTTP status code when dereferenced. The heuristic of determining how many captures are represented by URI-Ms in a TimeMap cannot be completed without dereferencing.

Redirection in a Web archive can be attributed to a variety of canonicalization rules [1]. Preserving and replaying these redirects allows an archive to accurately reproduce the HTTP transactions that would have occurred when the URI being accessed resided on the live Web. Because of the potential for redirection, the heuristic of counting URI-Ms with Link relation values of “memento” is an inaccurate means of determining the number of unique representations inferred from a TimeMap. We further emphasize the distinction per the Memento specification that the identifiers for mementos (URI-Ms) in a TimeMap are identifiers for archived HTTP transactions (e.g., transmission of HTTP 2XX, 3XX, 4XX, etc.) rather than identifiers for representations.

Based on the number of URI-Ms in a TimeMap not necessarily resolving to unique mementos when archival redirects are followed, we examined the mementos from contemporarily large TimeMaps to evaluate the patterns and schemes used in Memento canonicalization. Through this, we identify the difference between the number of mementos available as reported by the TimeMap through naive “rel” counting heuristics to the temporally unique mementos identified once these mementos are dereferenced.

Table 1: Google over time (abbreviated), bucketed by year, based on IA mementos extracted from the TimeMap. M_{TM} is the memento count based solely on the data in the TimeMap, M_{RC} is the count based on exclusion of redirects when dereferenced, and DI is the ratio of non-redirecting mementos to redirecting mementos.

year	M_{TM}	M_{RC}	DI
2006	735	483	1.917
2007	1,055	842	3.953
2008	1,376	894	1.855
2009	6,074	4,335	2.493
2010	9,326	6,530	2.335
2011	20,634	9,279	0.817
2012	102,533	16,240	0.188
2013	228,405	25,203	0.124
2014	164,865	22,738	0.160
2015	17,978	11,286	1.686
2016	139,520	5,805	0.043

URI canonicalization associates differently formatted URIs (See RFC6596) and allows after-the-fact clustering of URIs that likely reference the same resource. As URI schemes from a Web site change over time, canonicalization is critical for retaining a cohesive, comprehensive listing of the mementos available for a Web page.

Table 2: Scheme and subdomain for redirects when dereferencing URI-Ms for google.com.

orig \ dest		http			https		
		none	www	other	none	www	other
http	none	1,279	68,837	55	12	20,825	27
	www	8,934	490,836	204	32	77,610	16
	other	0	224	22	0	26	2
https	none	14	731	0	0	296	1
	www	1,117	72,874	27	15	18,525	2,101
	other	0	0	0	0	0	0

Table 3: Dereferencing 7 other large Web sites’ TimeMaps from Internet Archive produces the above distribution of status codes for each site.

host	% 3XX	% 200	M_{TM}	DI
google	84.89	15.11	695,525	0.178
yahoo	88.16	11.83	418,896	0.134
sourceforge	73.34	26.63	31,408	0.363
instagram	67.32	32.65	55,228	0.485
vimeo	57.04	42.94	199,262	0.752
cnn	49.97	50.01	87,148	1.001
wikipedia	44.62	55.19	25,973	1.240
whitehouse	44.57	55.24	26,006	1.243

We acquired HTTP response headers for all URI-Ms accumulated from multiple Web archives for various URI-Rs. We obtained a TimeMap for google.com from our locally deployed Memento aggregator containing 714,470 URI-Ms from 8 different Memento-compliant archives. 89.1% of the URI-Ms returned were from Internet Archive.

Equation 1 excludes mementos that resolve to HTTP 3XX status codes. $|TM|_D$ represents the count of mementos that result in non-3XX statuses based on the URI-Ms in a TimeMap.

$$|TM|_D = \sum_{m=1}^{len(M)} \begin{cases} 0 & 300 \geq httpStatus(m) < 400, \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

$$|TM|_I = |TM|_{rel} - |TM|_D \quad (2)$$

We quantify the ratio of mementos with non-redirecting HTTP status codes (Equation 1) to those with redirects (Equation 2) in Equation 3 as DI .

$$DI = \begin{cases} \frac{|TM|_D}{|TM|_I} & |TM|_I > 0, \\ \infty & \text{otherwise.} \end{cases} \quad (3)$$

If an archive reports revisit records as an HTTP redirect based on the CDX listing, and this redirect is propagated to the archive’s Memento endpoint thus producing a unique URI-M, the DI ’s value for the URI-R decreases. Requesting the URI-M using the Accept-Datetime HTTP header then observing the Memento-Datetime response header’s presence often reveals this nuance, but by relying on the TimeMap data without requesting each URI-M, the DI for the URI-R is unknown. Table 2 shows the scheme and subdomain breakdown for google.com URI-Ms where a redirect is encountered upon dereferencing the URI-M. Table 3 shows the DI values based on the TimeMaps of 8 large Web sites.

REFERENCES

- [1] Mat Kelly, Lulwah M. Alkwai, Michael L. Nelson, Michele C. Weigle, and Herbert Van de Sompel. 2017. Impact of URI Canonicalization on Memento Count. (March 2017). arXiv:1703.03302