Impact of URI Canonicalization on Memento Count

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

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

Motivation

- Memento TimeMaps [2] list identifiers for archival web captures (URI-Ms).
- Some URI-Ms redirect when dereferenced.
- Ergo, obtaining an accurate count of non-forwarding URI-Ms is not possible by solely using the contents of a TimeMap without dereferencing.
- Some archives have various endpoints beyond Memento.
- Endpoints' counts may differ depending on which is consulted.
- Therefore a standard-based mechanism must be used to obtain an accurate count.

Canonicalization

URI canonicalization associates differently formatted URIs.
For example, `http://example.com` might be associated with:

`http://www.example.com`	`https://www.example.com`
`http://example.com/`	`http://example.com/index.html`
`http://example.com/#articles`	`//example.com`

Direct-to-Indirect Metric

$$M = \text{URI-M} \subset \text{TimeMap}, \text{if } \text{“}memento\text{”} \subset \text{valuesOf}(rel) \tag{1}$$

$$|TM|_{rel} = \sum_{m=1}^{len(M)} 1 \tag{2}$$

$$|TM|_D = \sum_{m=1}^{len(M)} \begin{cases} 0 & 300 \geq httpStatus(m) < 400, \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

$$|TM|_I = |TM|_{rel} - |TM|_D \tag{4}$$

$$DI = \begin{cases} \frac{|TM|_D}{|TM|_I} & |TM|_I > 0, \\ \infty & \text{otherwise.} \end{cases} \tag{5}$$

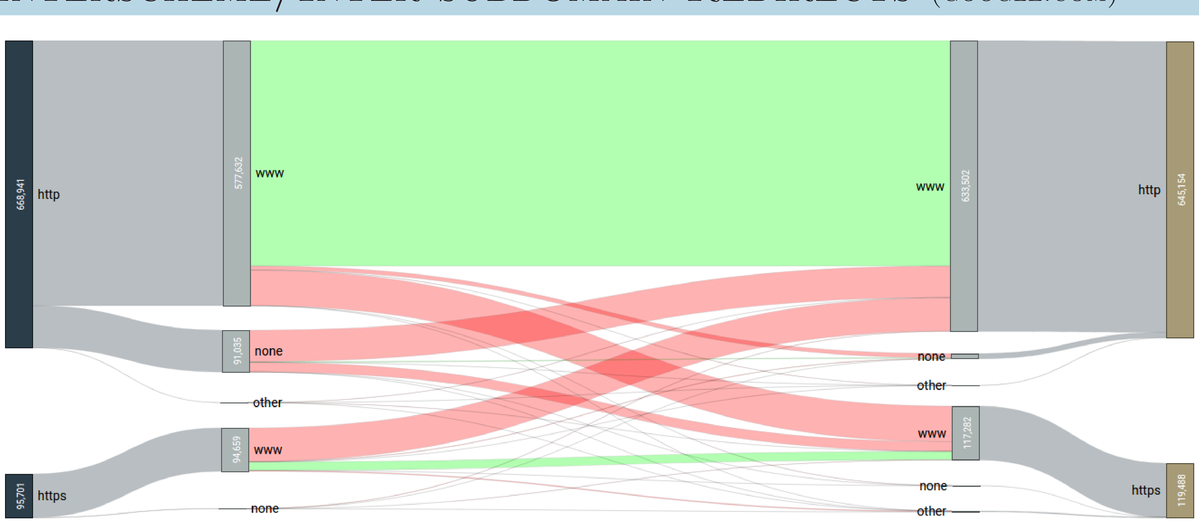
Redirection for Archived URI-Rs

host	% 3XX	% 200	$\mathbf{M}_{TM}$	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
stanford	62.14	37.84	19,309	0.609
princeton	60.10	39.88	9,355	0.663
columbia	48.01	51.88	9,882	1.082
harvard	33.91	65.96	7,699	1.948
caltech	33.13	66.86	5,474	2.017
mit	26.57	73.24	6,379	2.763
gatech	26.03	73.94	3,907	2.841
ufl	24.76	75.23	4,927	3.038
vt	23.07	76.92	4,061	3.334
lsu	15.06	84.93	2,974	5.638
nsu	13.82	86.00	1,208	6.233
odu	9.727	90.27	1,727	9.279
tcc	5.429	94.57	884	17.41

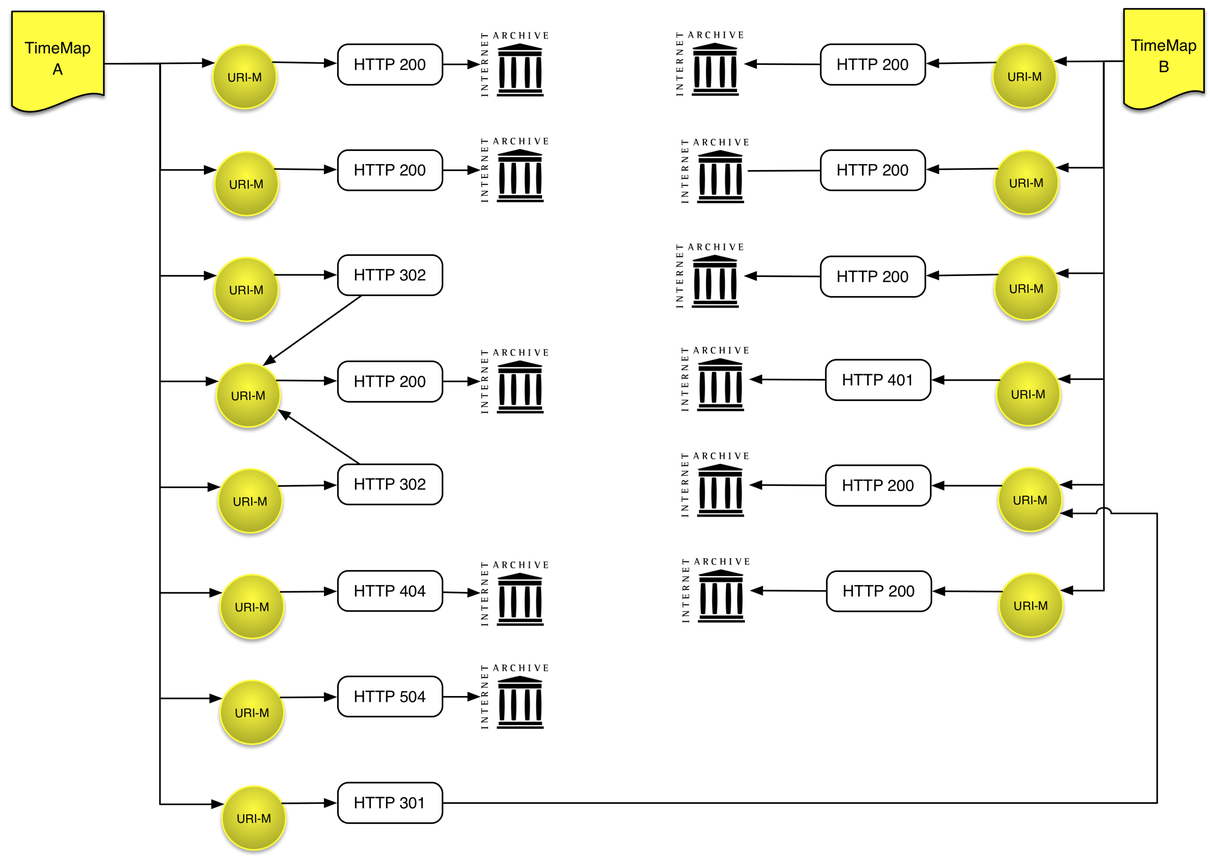
year	$\mathbf{M}_{TM}$	$\mathbf{M}_{RC}$	DI
1998	4	4	∞
1999	19	19	∞
2000	132	87	1.933
2001	1,185	579	0.955
2002	176	137	3.513
2003	75	55	2.750
2004	197	143	2.648
2005	1,236	414	0.504
2006	735	483	1.917
2007	1,055	842	3.953
2008	1,376	894	1.855
2009	6,074	4,335	2.493
2010	9,326	6,530	2.335
2011	20,634	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

`google.com` *redirection over time*
* *Data collected May 2016*

Interscheme/Inter-subdomain Redirects (google.com)

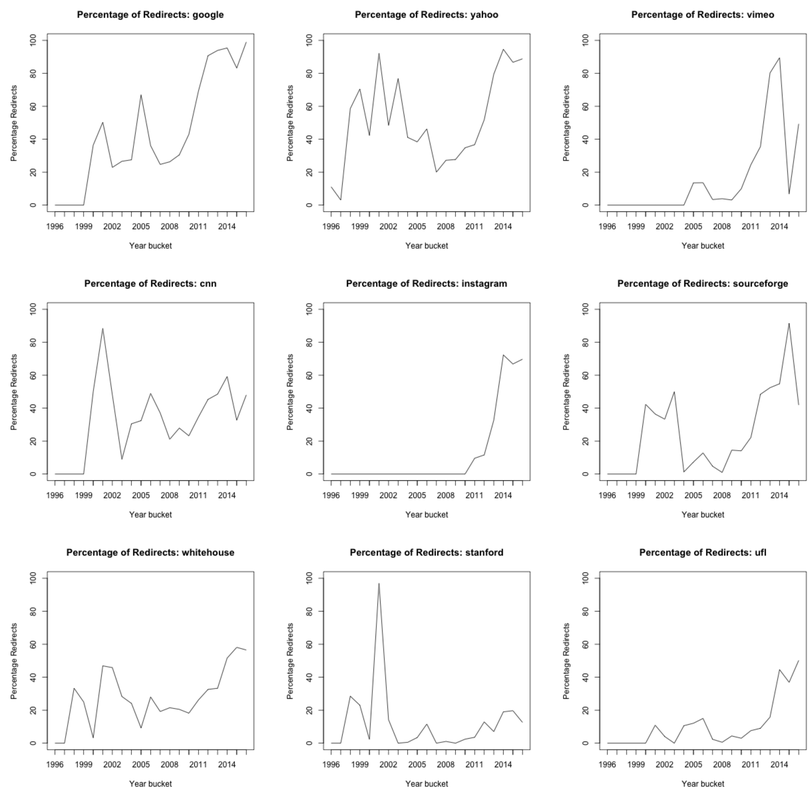


Dereferencing URI-Ms from a TimeMap



Intra-Archive Variance in "Count"

Redirection Over Time for Nine URI-Rs



Conclusions

- Counting captures is impossible using only a TimeMap.
- Non-standard capture endpoint listings differ in count interpretation intra-archive.
- TimeMaps need more info about mementos (e.g., status code) for accurate counting.
- Alluding to non-standard endpoints (e.g., CDX) may be N/A when aggregating with archives without these endpoints.

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

[2] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. 2013. HTTP Framework for Time-Based Access to Resource States – Memento. IETF RFC 7089. (December 2013)

JCDL 2016 and WADL 2017; Toronto, Canada; June 19-23, 2017

Comprehensive technical report [1] available at
`https://arxiv.org/abs/1703.03302`