

Archival HTTP Redirection Retrieval Policies

Ahmed AlSum,
Michael L. Nelson

Old Dominion University
Norfolk VA, USA
{aalsum,mln}@cs.odu.edu

Robert Sanderson,
Herbert Van de Sompel

Los Alamos National Laboratory
Los Alamos NM, USA
{rsanderson,herbertv}@lanl.gov

Temporal Web Analytics Workshop 2013, Rio De Janiro

Agenda

- Introduction
- Abstract Model
- Experiment And Results
- Retrieval Policies

Memento Terminology



URI-R, R Original Resource → <http://www.amazon.com>

URI-M, M Memento → <http://web.archive.org/web/20110411070244/http://amazon.com>

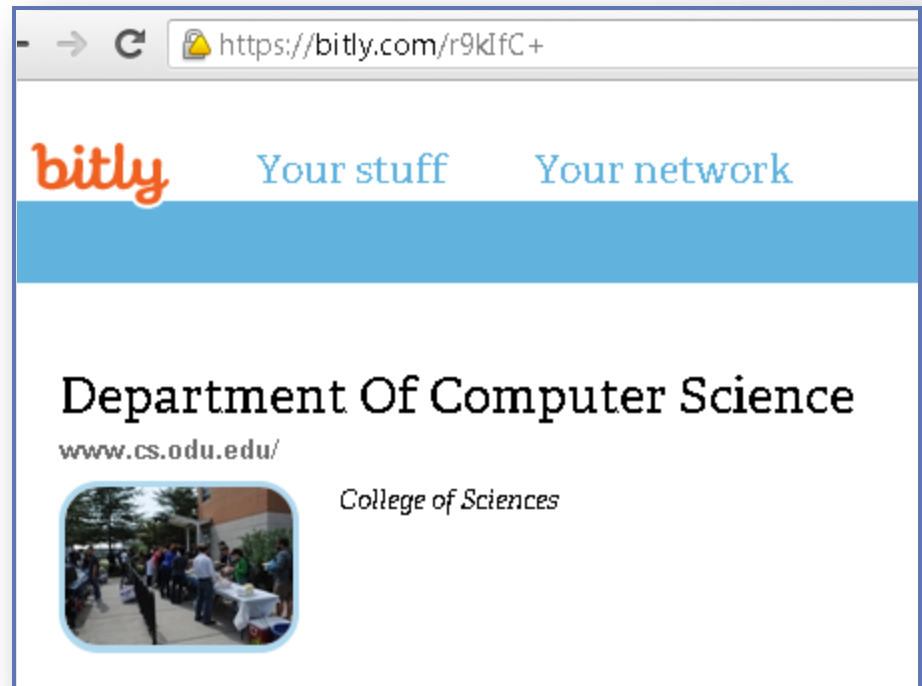
URI-T, TM TimeMap →

```
1 <http://http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.amazon.com>;  
  rel="self"; type="application/link-format",  
2 <http://mementoproxy.cs.odu.edu/aggr/timegate/http://www.amazon.com>;  
  rel="timegate",  
3 <http://www.amazon.com>; rel="original",  
4 <http://api.wayback.archive.org/memento/19981212012532/http://amazon.com/>;  
  rel="first memento"; datetime="Sat, 12 Dec 1998 01:25:32 GMT",  
5 <http://api.wayback.archive.org/memento/19990125093156/http://www.amazon.com/>;  
  rel="memento"; datetime="Mon, 25 Jan 1999 09:31:56 GMT",  
6 <http://api.wayback.archive.org/memento/19990208011311/http://www.amazon.com/>;  
  rel="memento"; datetime="Mon, 08 Feb 1999 01:13:11 GMT",  
7 <http://api.wayback.archive.org/memento/19990209063505/http://amazon.com/>;  
  rel="memento"; datetime="Tue, 09 Feb 1999 06:35:05 GMT",  
8 <http://api.wayback.archive.org/memento/19990218063448/http://www.amazon.com/>;  
  rel="memento"; datetime="Thu, 18 Feb 1999 06:34:48 GMT",
```

Live Redirect

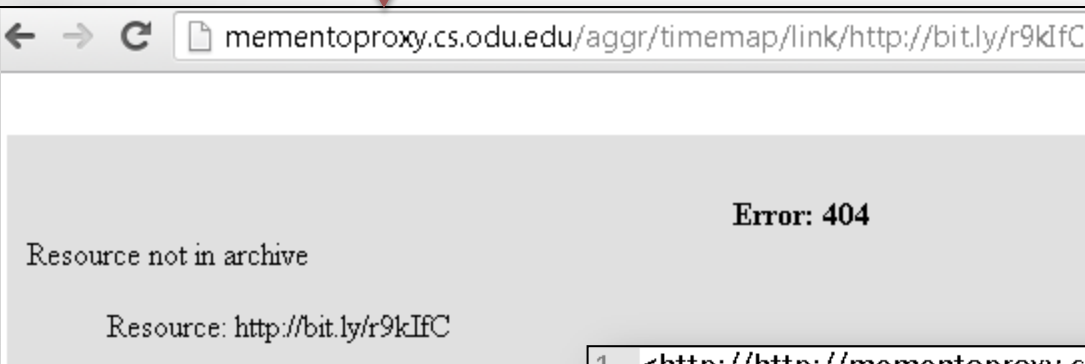
<http://bit.ly/r9kIfC> redirects to <http://www.cs.odu.edu>

```
% curl -I http://bit.ly/r9kIfC
HTTP/1.1 301 Moved
....
Location: http://www.cs.odu.edu/
...
```



Live Redirect

$\mathcal{R}$ <http://bit.ly/r9klfC> redirects to <http://www.cs.odu.edu> $\overline{\mathcal{R}}$



1 `<http://http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://cs.odu.edu>;rel="self"`
2 `";type="application/link-format",`
3 `<http://mementoproxy.cs.odu.edu/aggr/timegate/http://cs.odu.edu>;rel="timegate",`
4 `<http://cs.odu.edu>;rel="original",`
5 `<http://api.wayback.archive.org/memento/19970102130137/http://cs.odu.edu/>;rel="first memento";datetime="Thu, 02 Jan 1997 13:01:37 GMT",`
6 `<http://api.wayback.archive.org/memento/19970606105039/http://www.cs.odu.edu/>;rel="memento";datetime="Fri, 06 Jun 1997 10:50:39 GMT",`
7 `<http://api.wayback.archive.org/memento/19971010201632/http://www.cs.odu.edu/>;rel="memento";datetime="Fri, 10 Oct 1997 20:16:32 GMT",`
8 `<http://api.wayback.archive.org/memento/19971211124211/http://www.cs.odu.edu/>;rel="memento";datetime="Thu, 11 Dec 1997 12:42:11 GMT",`

Archived Redirect

R_1

www.draculathemusical.co.uk

redirects

R_2

www.dracula-uk.com/index.html



<http://api.wayback.archive.org/memento/20020212194020/http://www.draculathemusical.co.uk/>

Archived redirects

R_3

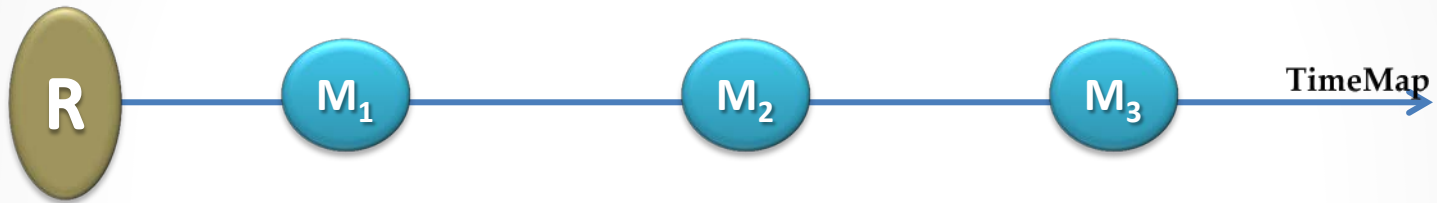
<http://api.wayback.archive.org/memento/20020212194020/http://www.geocities.com/draculathemusical>

Abstract Model

...

Abstract Model

- TimeMap for R



$$TM(R) = \{M_1, M_2, \dots, M_n\}; \text{ where } M_i = M(R) \text{ at } t_i$$

URI Stability

- URI's stability is a count of the change in HTTP responses across time (200, 3xx, or 4xx) and the number of different URIs in the "Location" for 3xx status code.

High Stability = 1

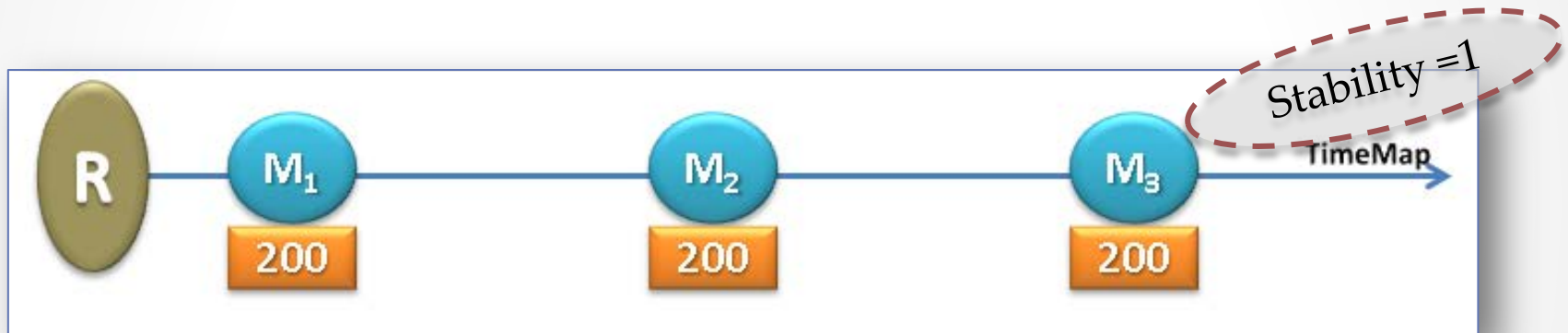
No Stability = 0

$$Stability(R) = 1 - \frac{\sum_{M \in TM} Change(M_i, M_{i-1})}{|TM|}$$

$$Change(M_i, M_{i-1}) = \begin{cases} 1 & \text{Status}(M_i) \neq \text{Status}(M_{i-1}) \\ & \text{or Location}(M_i) \neq \text{Location}(M_{i-1}) \\ 0 & \text{Otherwise} \end{cases}$$

Timemap Redirection Categories

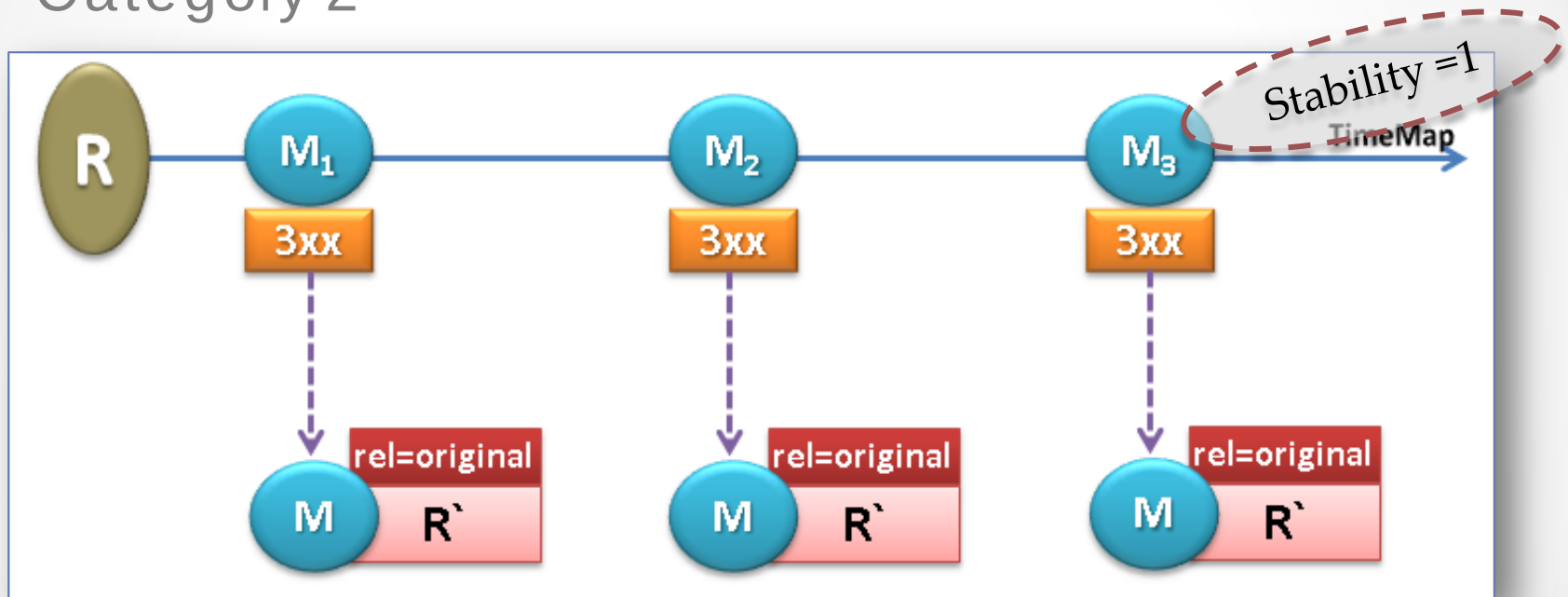
- Category 1



All Mementos have 200 HTTP status code

Timemap Redirection Categories

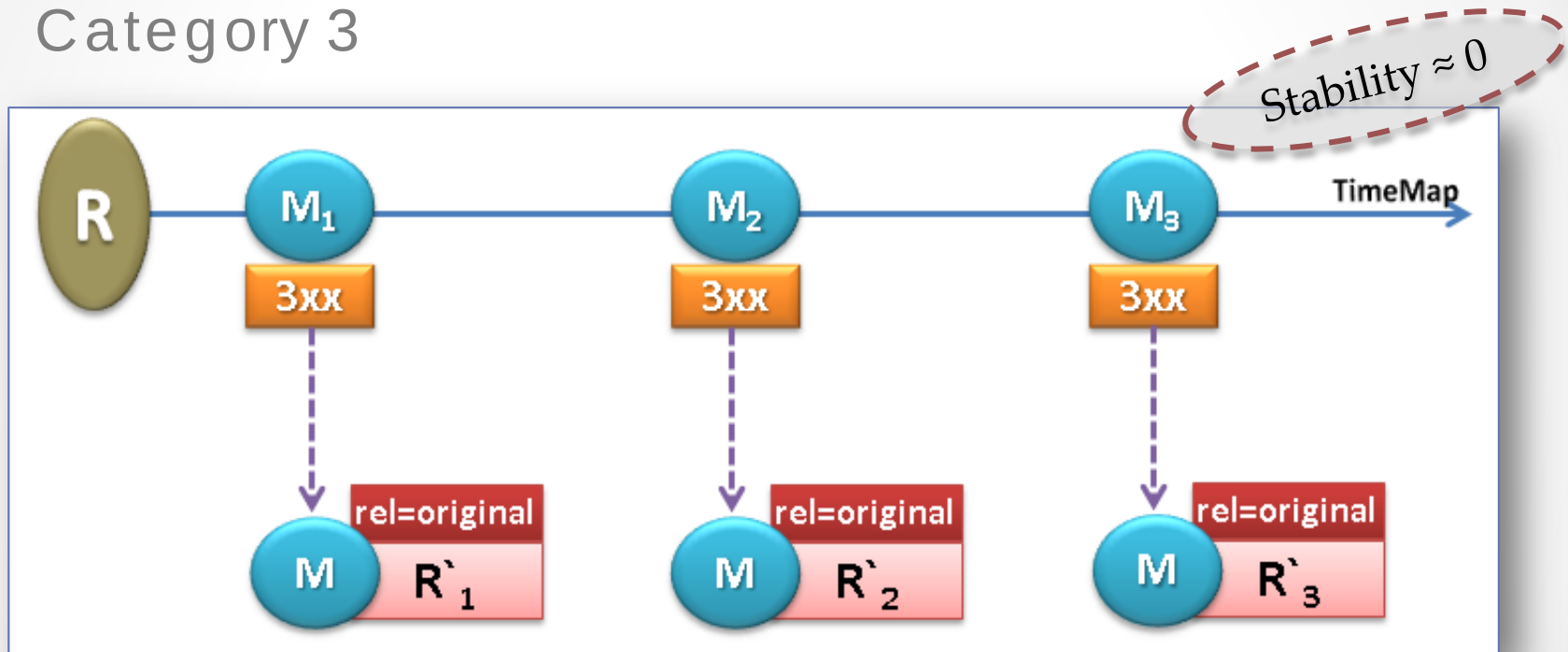
- Category 2



All Mementos have redirection to the same URI.

Timemap Redirection Categories

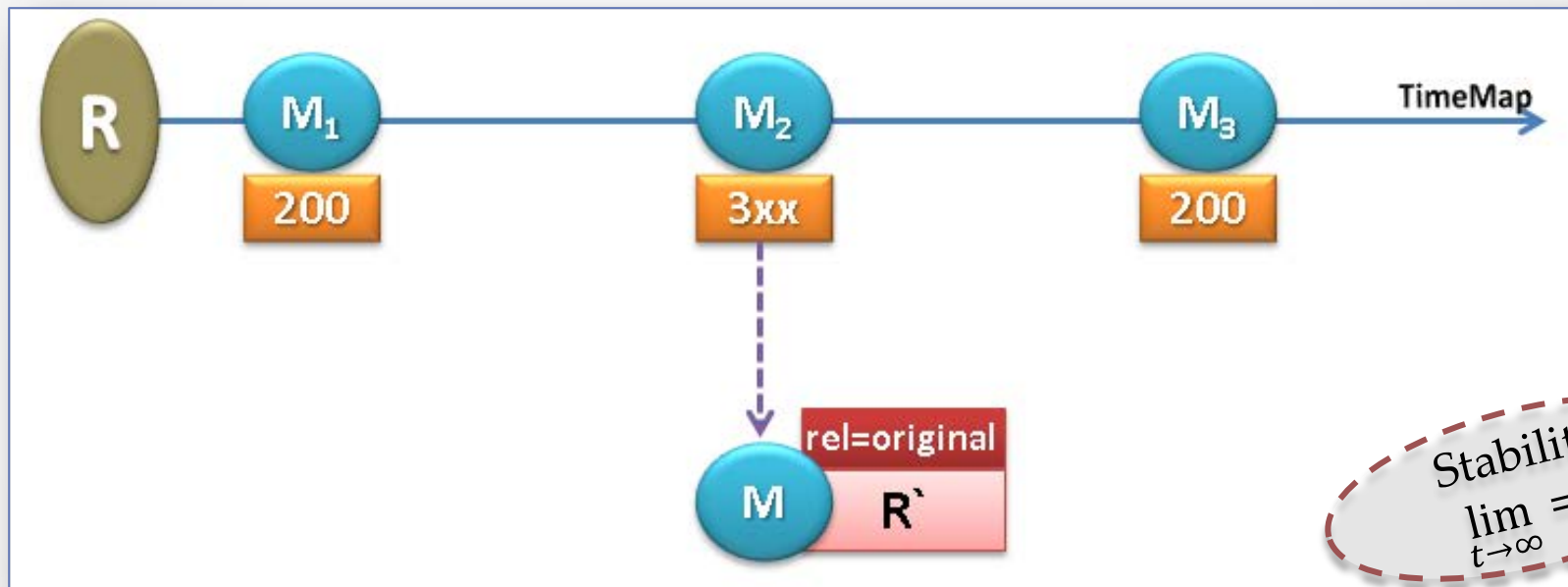
- Category 3



All Mementos have redirection to different URIs.

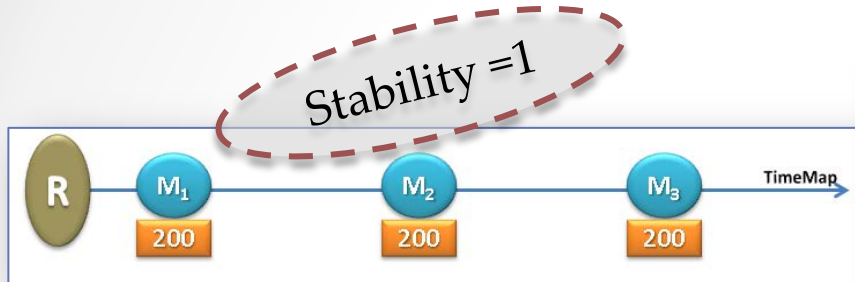
Timemap Redirection Categories

- Category 4

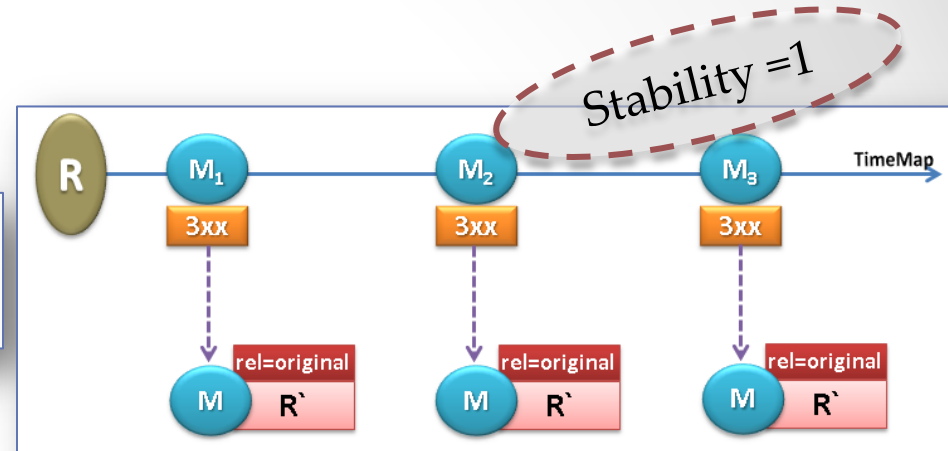


Mementos have different HTTP status code.

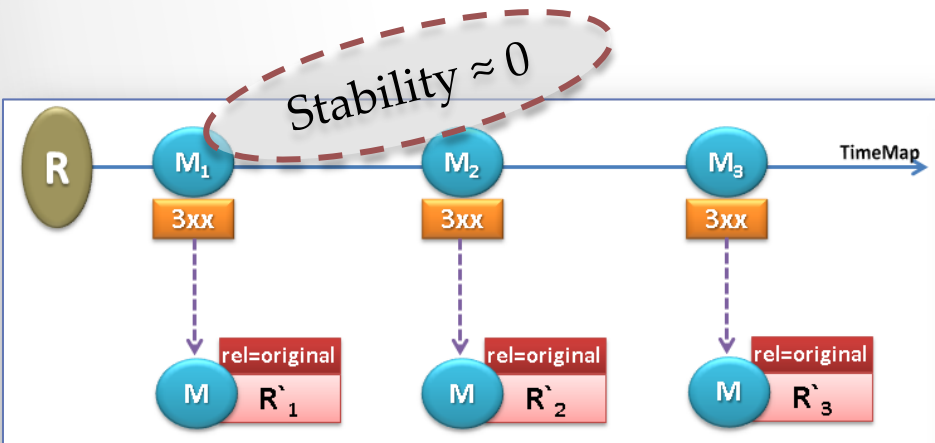
Timemap Redirection Categories



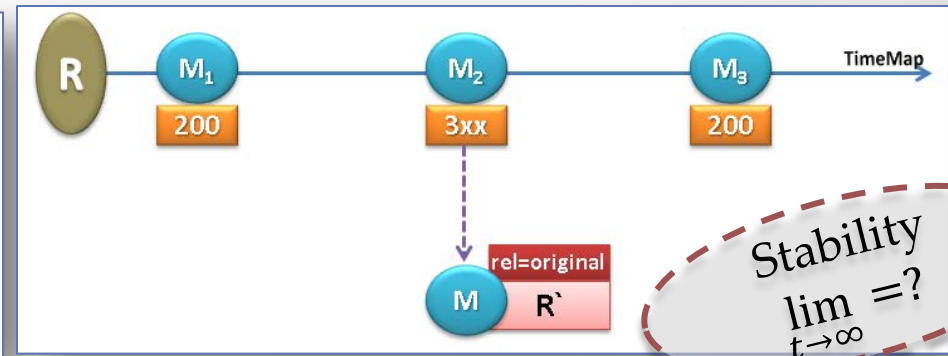
All Mementos have 200 HTTP status code



All Mementos have redirection to the same URI.

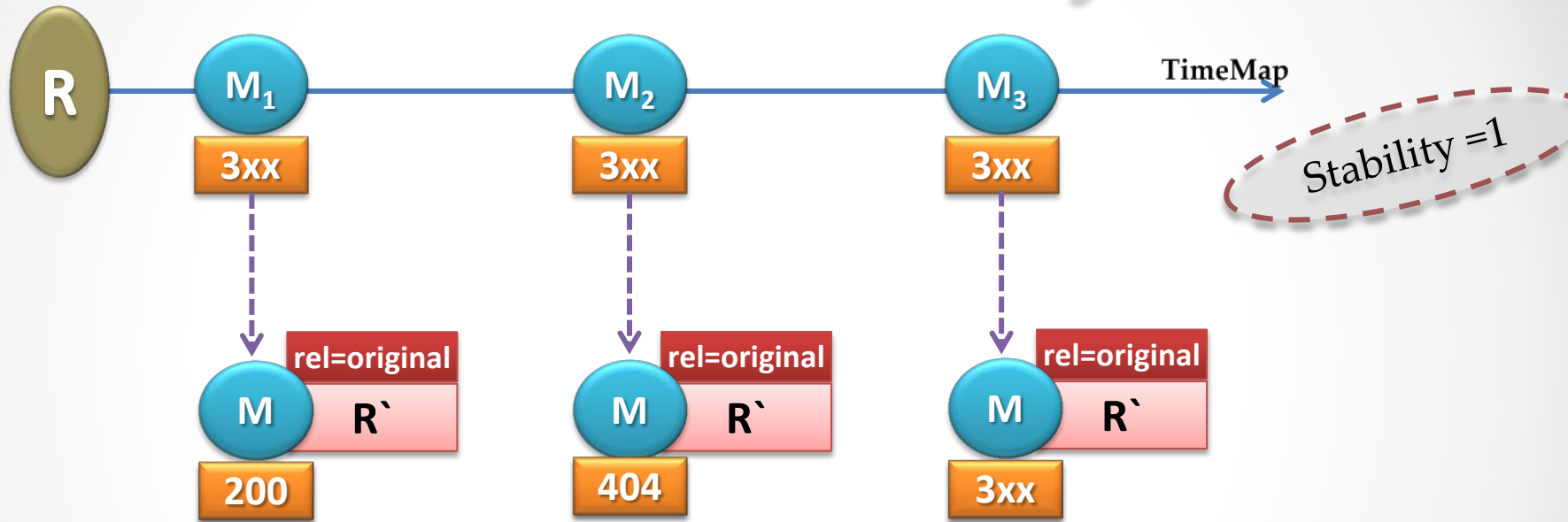


All Mementos have redirection to different URIs.



Mementos have different HTTP status code.

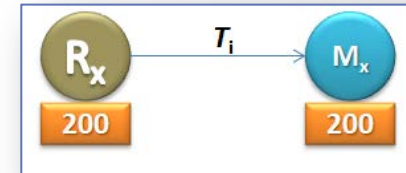
URI Reliability



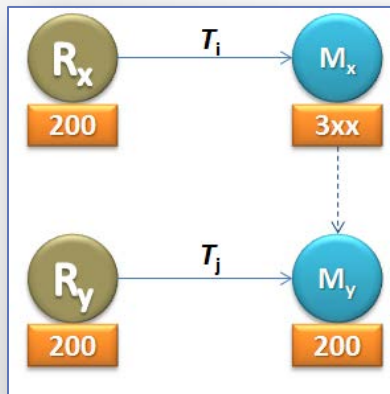
$$Reliability = \frac{\#Mementos\ end\ 200}{|TM|}$$

HTTP Redirection Relationship between URI-R & URI-M

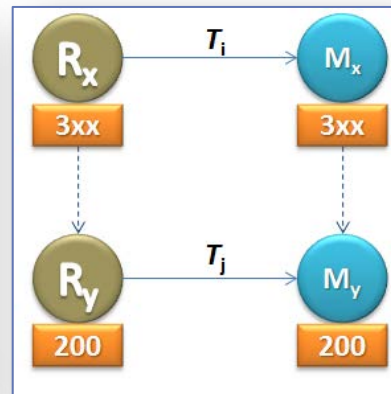
		Live Web URI – R	
		OK	Redirection
Web Archive URI-M	OK	Case 1	5
	Redirection	2	3,4



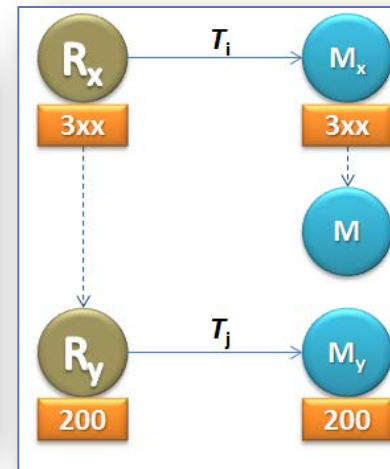
Case 1



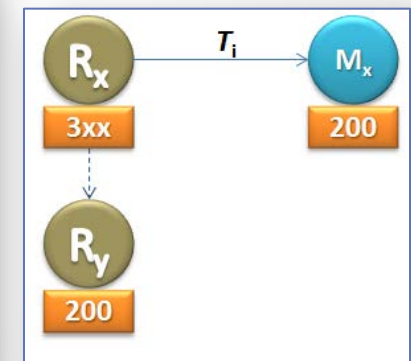
Case 2



Case 3



Case 4



Case 5

Experiment & Results

...

Experiment

- Dataset: 10,000 sample URIs from [dmoz](#) open directory project
- Dataset doesn't have bit.ly nor doi.
- Experiment focused on the root page (no embedded resources)

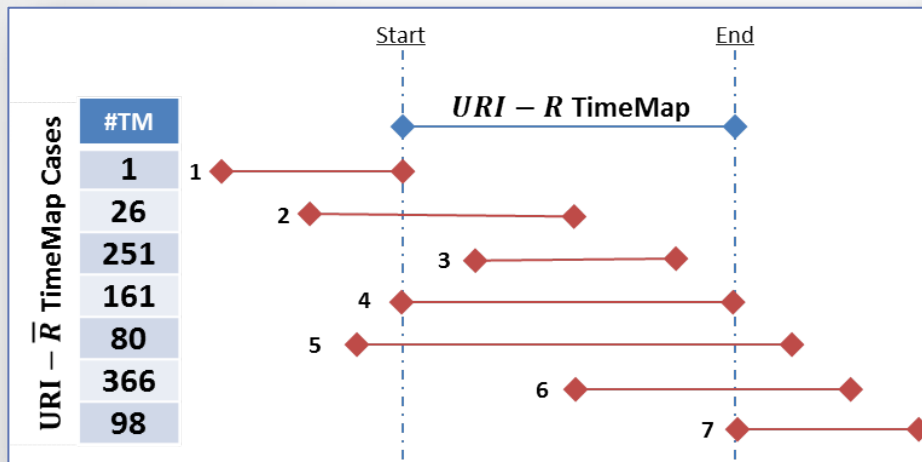
URIs Live HTTP status code

HTTP Status/Code	(10,000 URI-R)
OK (200)	82.83%
Redirection (3xx)	14.71%
<i>Redirection (301)</i>	8.4%
<i>Redirection (302)</i>	6.1%
<i>Redirection (others)</i>	0.2%
Not-Found (4xx)	1.18%
Others	1.28%

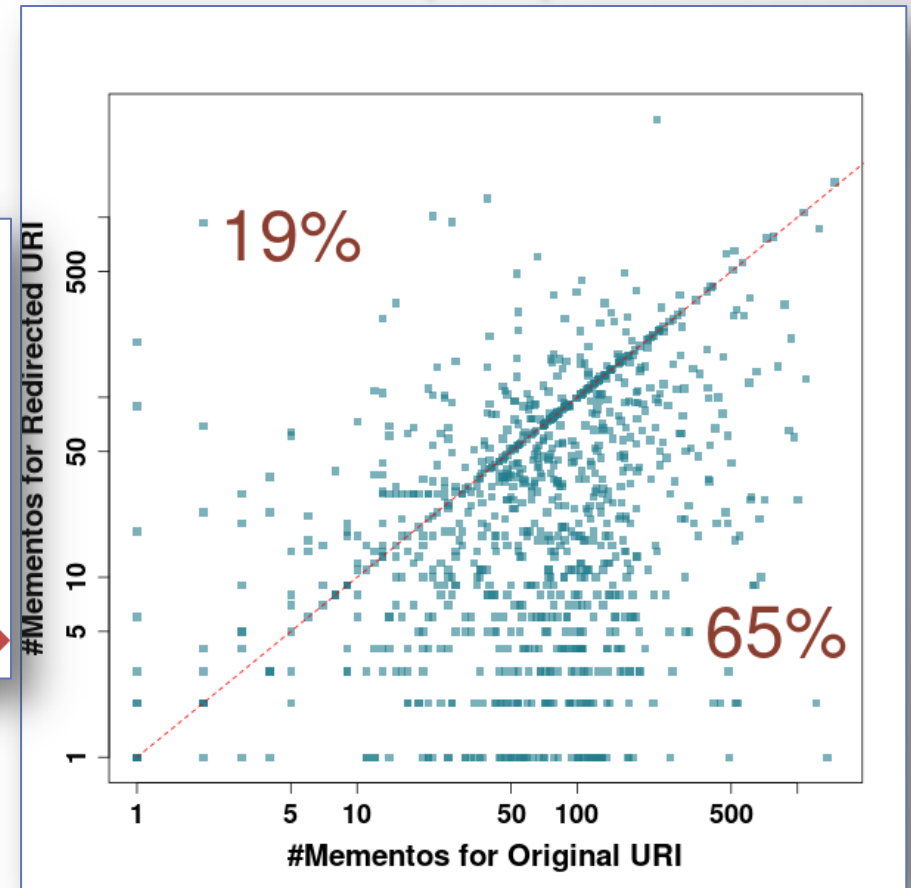
Memento HTTP status code

HTTP Status/Code	(894,717 URI-M)
OK (200)	93.46%
Redirection (3xx)	5.69%
Not-Found (4xx)	0.26%
Others	0.59%

Relationship between $TM(R)$ and $TM(\bar{R})$



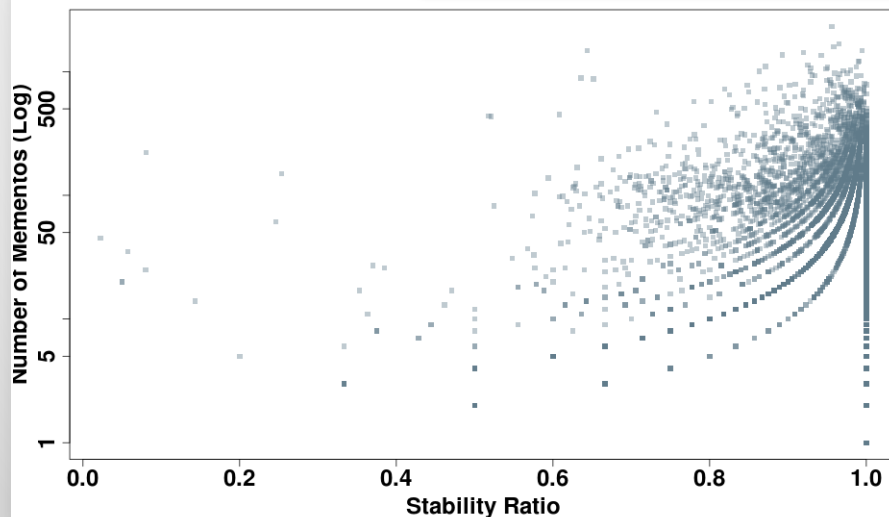
Time span



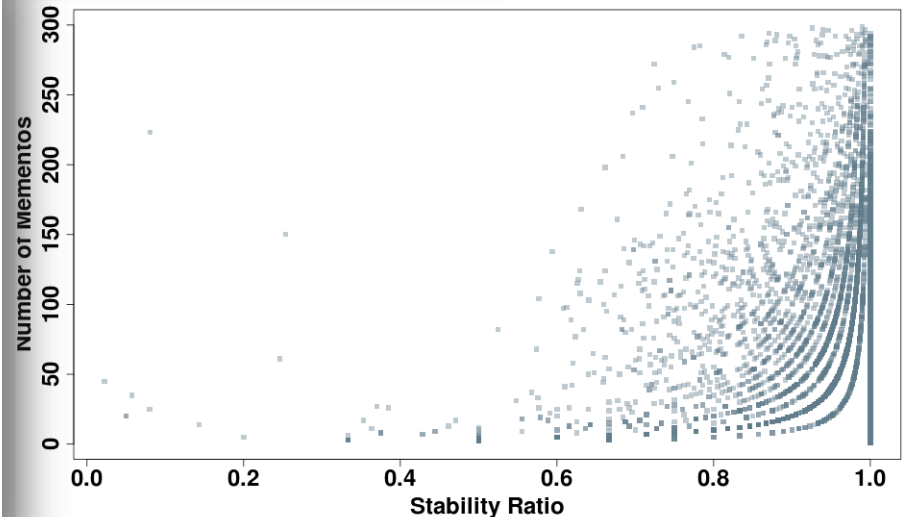
Number of Mementos

URI Stability

TimeMap Category	Percentage	Stability
All Mementos have OK	52%	1
Mementos have mix status code	36%	0.91
All Mementos have Redirection	0.92%	0.85
<i>Redirection to the same URI</i>	0.62%	
<i>Redirection to different URIs</i>	0.30%	
URI has no Mementos at all	10.97%	0



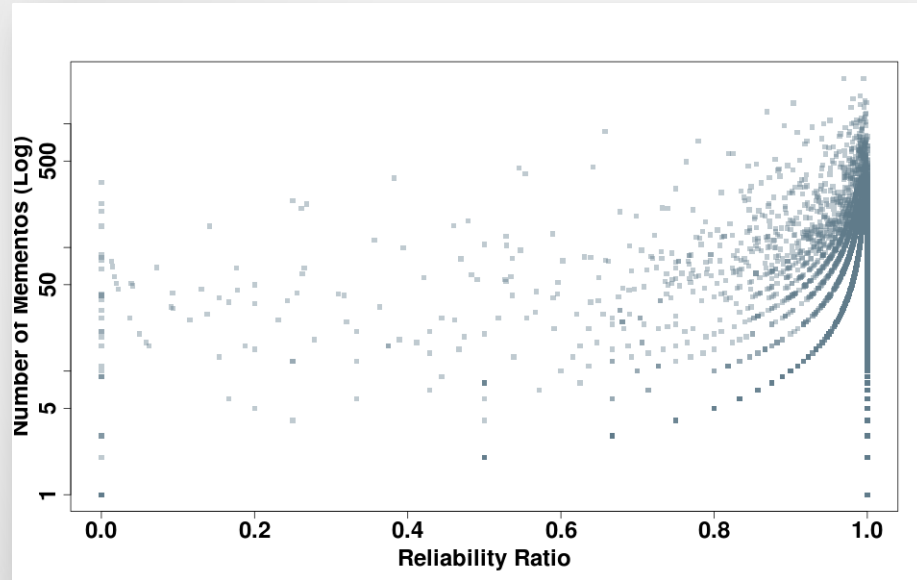
Stability in semi-log scale



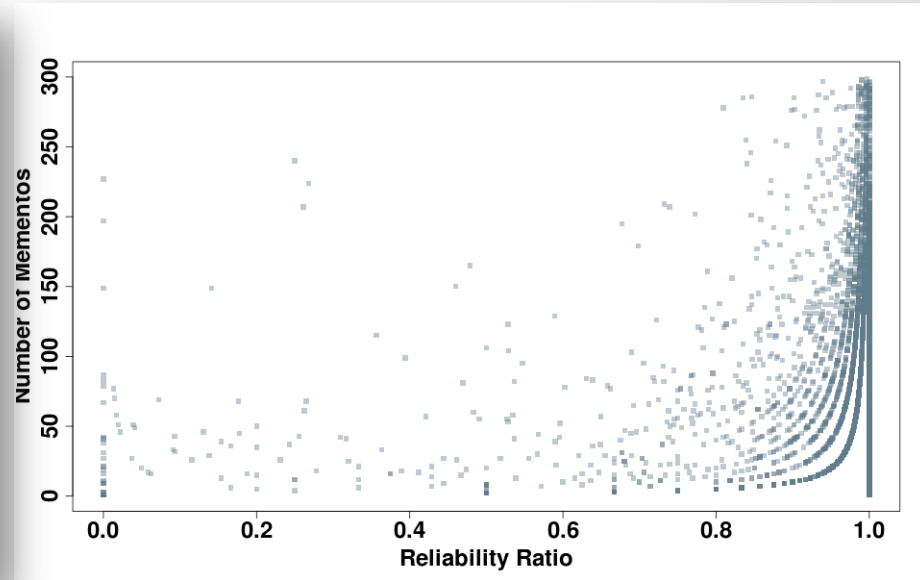
Stability for $|TM(R)| < 300$

URI Reliability

- 23% of the mementos did not lead to a successful memento at the end.



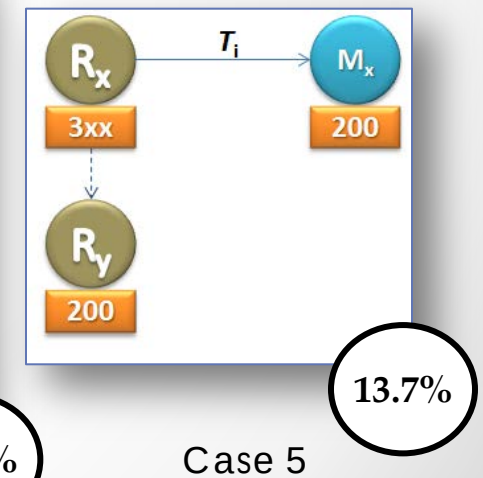
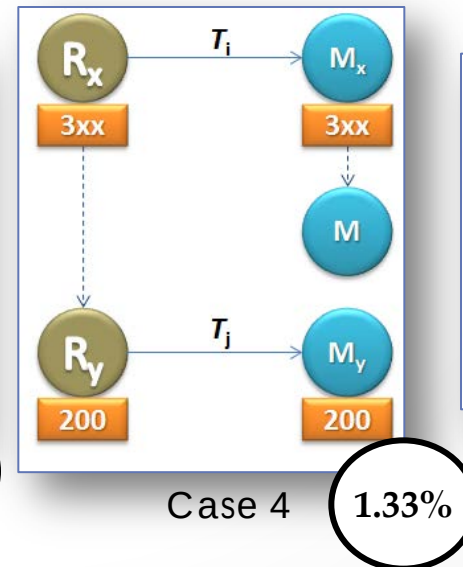
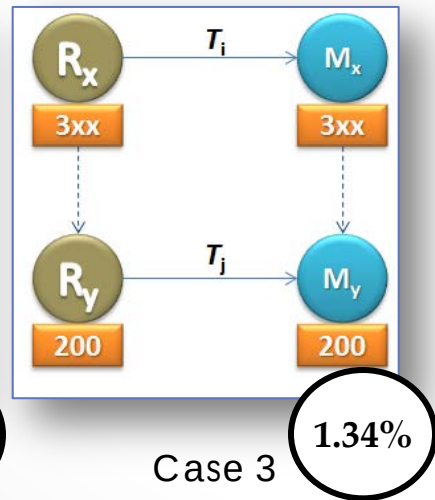
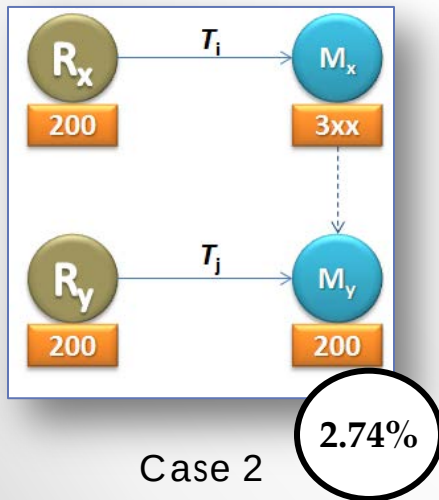
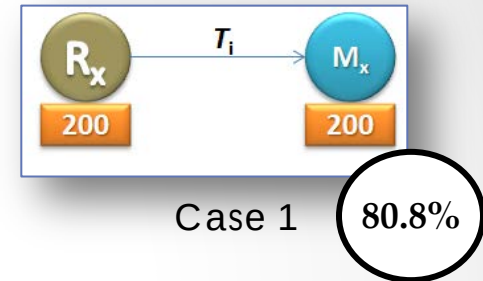
Reliability in semi-log scale



Reliability for $|TM(R)| < 300$

HTTP Redirection Relationship between URI-R & URI-M

		Live Web URI – R	
		OK	Redirection
Web Archive URI-M	OK	Case 1	5
	Redirection	2	3,4



RETRIEVAL POLICIES

• • •

ARCHIVED HTTP REDIRECTION RETRIEVAL POLICIES

Current Wayback Machine Policy

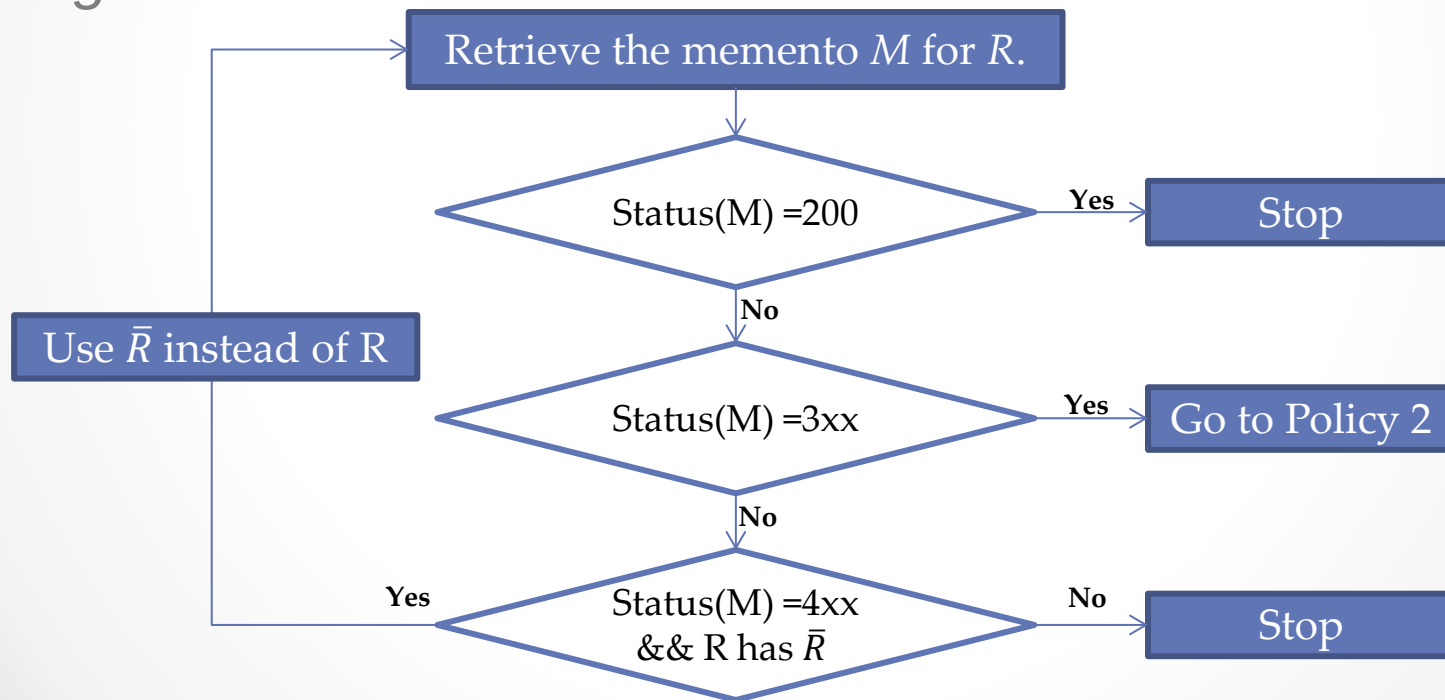
- *Live Redirect*: Wayback Machine ignores the live redirects. Use R instead of $\bar{R}$.
- *Archived Redirect*: Wayback Machine follows the redirection.



Policy one:

URI-R with HTTP redirection

- *Scope*: Selection between $R \rightarrow \bar{R}$ on the live web.
- *Example*: <http://bit.ly/r9klfC> $\rightarrow$ <http://www.cs.odu.edu>
- *Algorithm*:



Policy one: URI-R with HTTP redirection

- Evaluation:
 - o Policy scope has: 1471 URIs (*that have live redirection*)
 - o 77 out of 1471 have no mementos at all
 - o 17 out of 77 have been retrieved mementos based on live redirection
- Implementation

Tool	Comment
IA Wayback Machine	For bit.ly URIs only
MementoFox	v 0.9.6+
mcurl	v 1.0

Policy two:

URI-M with HTTP redirection

- *Scope*: Selection between $M \rightarrow \bar{M}$ in web archive.
- *Example*: <http://api.wayback.archive.org/memento/20101109032705/http://bit.ly/2EEjBI> → <http://api.wayback.archive.org/memento/20101109032705/http://www.cnn.com/>
- *Algorithm*:

$M \rightarrow \bar{M}$

<http://api.wayback.archive.org/memento/20101109032705/http://bit.ly/2EEjBI> → <http://api.wayback.archive.org/memento/20101109032705/http://www.cnn.com/>

Extract original from $\bar{M}$

<http://www.cnn.com/>

Repeat content-negotiation in datetime
for original($\bar{M}$)

Accept-Datetime: Sun, 13 May 2006
<http://www.cnn.com/>

Policy two: URI-M with HTTP redirection

- Evaluation:
 - o Policy scope: 2980 TimeMap (*that showed HTTP redirection status code in at least one memento*)
 - o Success criteria: Using policy two contributed to the original TimeMap
 - o Success percentage: 58% of the cases

Conclusion

- Quantitative study with 10,000 URIs.
- 48% were not fully stable through time.
- 27% were not perfectly reliable through time.
- New archival retrieval policy:
 - o Policy one: successfully retrieved mementos for 17 out of 77
 - o Policy two: Expanded the timemap for 58% of cases.



aalsum@cs.odu.edu

@aalsum