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

ABSTRACT

When retrieving archived copies of web resources (mementos) from web archives, the original resource's URI-R is typically used as the lookup key in the web archive. This is straightforward until the resource on the live web issues a redirect: $R \to \overline{R}$. Then it is not clear if R or $\overline{R}$ should be used as the lookup key to the web archive. In this paper, we report on a quantitative study to evaluate a set of policies to help the client discover the correct memento when faced with redirection. We studied the stability of 10,000 resources and found that 48% of the sample URIs tested were not stable, with respect to their status and redirection location. 27% of the resources were not perfectly reliable in terms of the number of mementos of successful responses over the total number of mementos, and 2% had a reliability score of less than 0.5. We tested two retrieval policies. The first policy covered the resources which currently issue redirects and successfully resolved 17 out of 77 URIs that did not have mementos of the original URI, but did of the resource that was being redirected to. The second policy covered archived copies with HTTP redirection and helped the client in 58% of the cases tested to discover the nearest memento to the requested datetime.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Design, Experimentation, Standardization

1. INTRODUCTION

HTTP [12] supports redirection using 3xx status codes, which indicate further action needs to be taken by the user-agent in order to fulfill the request. The resource has been moved temporarily (302, 307) or permanently (301) to another URI provided in the "`Location`" response header.

In web archiving, the user-agent must decide if the URI before or after the redirection should be used to access the web archive. For example, URI `http://bit.ly/r9kIfC` provides a redirection to `http://www.cs.odu.edu` via a 301 status code and a "`Location`" header.

Querying the ODU Memento Aggregator[1] with the shortened URI returns a 404 response because this URI has never been archived, while using `www.cs.odu.edu` as the lookup key returns 700+ mementos.

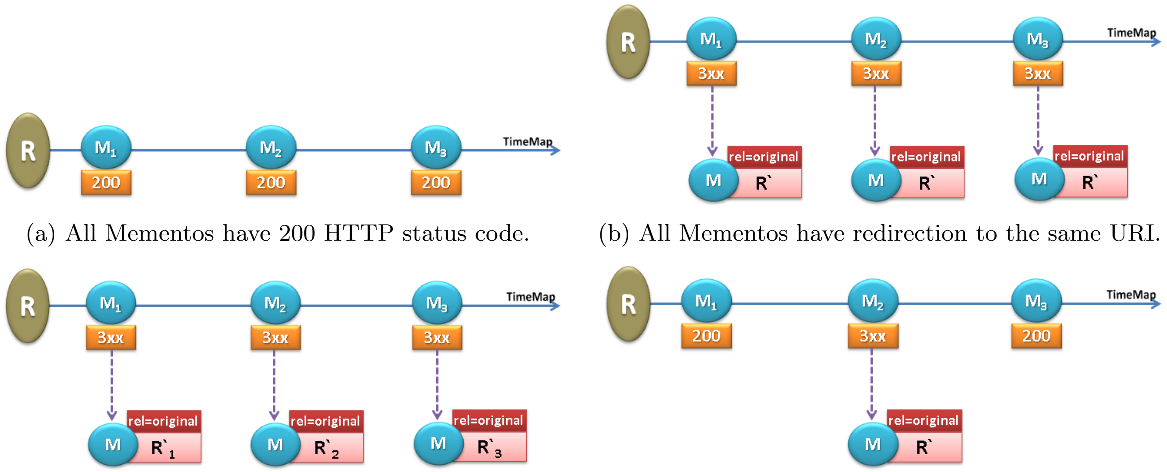
Another example, URI `www.draculathemusical.co.uk` has a redirection on the live web to `http://www.dracula-uk.com/index.html`. If we use URI-R as a lookup key, we can find a memento with HTTP redirection (i.e., `http://api.wayback.archive.org/memento/20020212194020/http://www.draculathemusical.co.uk/` redirects to `http://api.wayback.archive.org/memento/20020212194020/http://www.geocities.com/draculathemusical`). Now, we end up with three original URIs.

1. www.draculathemusical.co.uk
2. www.dracula-uk.com/index.html
3. www.geocities.com/draculathemusical

In these examples, the client's awareness with the HTTP redirection status code provided a new approach to reach a nearest memento for the requested datetime. On the other hand, using $URI - \overline{R}$ directly could be misleading. For example, the department of architecture in Oxford Brookes university's URI-R (`http://www.brookes.ac.uk/schools/be/architecture/`) redirects to $\overline{R}$ (`http://www.brookes.ac.uk/about/faculties/tde`). Using $URI - R$ as a lookup key in this example, we reach 30+ mementos where $URI - \overline{R}$ has only one memento. It is difficult to know a priori which of these two URIs should be used to discover archived copies of the resource. In this paper, we study the stability of redirecting $URI - R$s across time. We present new policies that will help the client to use the HTTP redirection and obtain a closer Memento to the requested datetime. The first proposed policy (section 6.1) will discuss the different cases that enables the user to use the redirected URIs on the live web instead of the original URIs (i.e., select between 1 and 2 in the previous list). The second policy (section 6.2) will discuss the cases when the user-agent should use the redirected URIs on the archived web instead of the original URIs (i.e., select between 2 and 3 in the previous list). In section 3, we build an abstract model for the stability and reliability of the URI as a lookup key including redirection cases for the TimeMap and mementos. Section 4 describes the experiment with the detailed results in section 5. Section 6 discusses the retrieval policies for $URI - R$ and $URI - M$ that carry the HTTP redirection status codes.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] `http://mementoproxy.cs.odu.edu/aggr/timemap/link/`



(a) All Mementos have 200 HTTP status code.

(b) All Mementos have redirection to the same URI.

(c) All Mementos have redirection to different URIs.

(d) Mementos have different HTTP status code.

Figure 1: Timemap Redirection Categories.

2. RELATED WORK

The Library of Congress defined Web Archiving as "the process of creating an archival copy of a website. An archived site is a snapshot of how the original site looked at a particular point in time."[2] In 2006, Masanes [16] published a book about web archiving where he covered web preservation issues with the required methodologies and tools. Brown [8] in 2006, provided a practical guide for archiving the Web.

Heritrix [17] is an open source web crawler that is used by Internet Archive to take a periodic snapshots of the Web. Heritrix saves all the responses into WARC files. So it keeps a record of the "`Location`" header to be used later in the retrieval process. Wayback Machine [22] is an open source tool to replay the web page as it appeared in the past. Wayback Machine focused on the content crawled in ISO WARC [1] format.

The crawler depends on crawling strategy which determines what the order of page to be crawled. Cho et al. [10] proposed a policy to visit the most important page first based on re-ordering the visited URL. Baeza-Yates et al. [5] compared different strategies based on the available information about the crawling cycle (no-information, partial information, or all the information). Ben Saad and Gançarski [6, 7] focused on adapting new crawling strategies to increase the quality of the web archive for completeness and coherence.

Some research has been conducted to provide easier and more functional user interface. Jatowt et al. [14, 15] proposed different models to browse the past web. Adar et al. [2] proposed "Zoetrope", a system that enables interaction with the historical Web. They discussed different techniques for specifying interesting portions of the current page and visualizing the relevant historical information. Teevan et al. [21] proposed "DiffIE" an Internet Explorer browser plug-in that caches the pages a person visits and highlights how those pages have changed when the person returns to them.

The Memento Protocol [23] is an HTTP extension to allow the user to browse archived web resources seamlessly with the current web. Memento extends HTTP content negotiation [13] to include the datetime dimension using the "`Accept-Datetime`" and "`Memento-Datetime`" headers. The archived copy, a Memento (denoted by $URI-M$ or M) for an Original Resource ($URI-R$) is defined as a resource that encapsulates the state of $URI-R$ at time t_i. A TimeGate is a resource that supports negotiation to allow selective, datetime-based, access to an archived copy of $URI-R$. A TimeMap, denoted by $URI-T$ or TM, is a list of the URIs of Mementos of $URI-R$ is available. A memento aggregator [24] provides a single TimeGate and TimeMap for multiple archives.

3. ABSTRACT MODEL

Although a lot of research has been done on estimating the frequency of change of a web page [9, 11, 18], no one has focused on the change of the HTTP status code of the URI. In this section, we will discuss the change of HTTP status code through time and the relationship between the live web HTTP status code and the memento HTTP status code.

In this section, $URI-R$ and R denote the original resource; $URI-\overline{R}$ and $\overline{R}$ denote the redirected resources interchangeably. Memento defines the TimeMap TM as a list of the available mementos for $URI-R$:

$$TM(R) = \{M_1, M_2, \ldots M_n\}, \quad where\ M_i = M(R)\ at\ t_i$$

We extend the Memento TimeMap definition to include the HTTP status code for each memento. $Status(M_i(R))$ returns the HTTP status code for $M_i(R)$. $Location(M_i(R))$ returns the URI in the "`Location`" header for $M_i(R)$ with HTTP redirection status code. Also, we define $|TM(R)|$ as the number of mementos per TimeMap, and $[TM(R)]$ as the time span for the $TM(R)$, the minimum and the maximum memento datetime in the TimeMap.

3.1 URI Stability

We can determine a URI's stability by examining the HTTP responses across time, and then count the number of changes to the status code (200, 3xx, or 4xx) and the number of different URIs in the "`Location`" for 3xx status code as appeared in equation 1.

[2] `http://www.loc.gov/webarchiving/faq.html#faqs_02`

For example, if $URI - R$ has a TimeMap and all the mementos have HTTP status code 200, we consider $URI - R$ stable through time (Stability = 1). Also, we can consider $URI - R$ stable if its TimeMap has mementos with status code 3xx to the same "`Location`", in other words, it always redirects to the same $URI - \overline{R}$ through time with stability of 1.0. On the other hand, if $URI - R$ has a TimeMap and each memento has a redirection to a different "`Location`", in this case we consider this $URI - R$ as unstable (Stability $\simeq 0$ for large TimeMap), because $URI - R$ redirects to different $URI - \overline{R}$ through time.

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

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

$$where\ |TM| > 0$$

We present four categories of TimeMaps, and discuss the stability for each one. These categories focus only on the HTTP status codes of the mementos excluding the HTTP status code of the original URI on the current Web.

Figure 1 illustrates the different categories. The horizontal line represents the TimeMap for the resource R in the golden oval. The blue circle represents a memento; the attached orange rectangle represents the HTTP status code of this memento. If the status code is 3xx, a dashed arrow will represent the redirection to another memento. The red rectangle carried the original resource that belongs to this memento.

3.1.1 All Mementos have 200 HTTP status code

The TimeMap $TM_1(R)$ is a list of available mementos M_i for the resource R where each memento carried HTTP response code 200.

$$TM_1(R) = \{M_1, M_2, \ldots M_n\} \quad where\ Status(M_i) = 200$$

For this TimeMap category, we calculate the stability as 1.0 because $URI - R$ did not change through time.

3.1.2 All Mementos have redirection to the same URI

The TimeMap $TM_2(R)$ is a list of available mementos M_i for the resource R where each memento has HTTP redirection status code. Each $M(R)$ redirects to $M(\overline{R})$ for all the mementos in TimeMap.

$$TM_2(R) = \{M_1, M_2, \ldots M_n\} \quad where\ Status(M_i) = 3xx$$
$$\forall\ M_i(R)\ \exists M_j(\overline{R})\ where\ M_i(R) \rightarrow M_j(\overline{R})$$

This category describes this set of URIs that have a redirection status code that have not changed over time. For example, bit.ly/xxx URIs do not change over time. The stability for such $URI - R$ is 1.0 because it redirects to one $\overline{R}$ through time. Stability is a function of redirection so it is possible to have a stable TimeMap that never returns 200 response code.

3.1.3 All Mementos have redirection to different URIs

The TimeMap $TM_3(R)$ is a list of available mementos M_i for the resource R where each memento has a redirection status code to mementos that belong to the same or different $\overline{R}$.

$$TM_3(R) = \{M_1, M_2, \ldots M_n\} \quad where\ Status(M_i) = 3xx$$
$$\forall\ M_i(R)\ [status: 3xx]\ \exists M_j(\overline{R})\ where\ \overline{R} \in \{\overline{R}_1, \overline{R}_2, \ldots \overline{R}_n\}$$

In this case, $URI - R$ was not stable over time, as $URI - R$ redirects to various $URI - \overline{R}$ through time. Here, stability will asymptotically approach 0.

3.1.4 Mementos have different HTTP status codes

The TimeMap $TM_4(R)$ is a list of available mementos M_i for the resource R where each memento may or may not have a redirection status code. In the existence of the HTTP redirection status code, it could be to the same or different $URI - \overline{R}$.

$$TM_4(R) = \{M_1, M_2, \ldots M_n\} \quad where\ Status(M_i) = xxx$$
$$where\ xxx\ \text{is a valid HTTP response code}$$

3.2 URI Reliability

Even though the stability gave us a good indication about the status code change of the URI-R through time, it does not necessary indicate the ability to retrieve the mementos successfully. We can categorize the $M_i(R)$ into two categories: successful retrieval, where the memento has HTTP status code 200 or a redirection chain ends with 200, and unsuccessful retrieval, where the memento has 4xx/5xx or a redirection chain that ends with 4xx/5xx. We define URI reliability as the ratio between the number of successful mementos to the total number of mementos per TimeMap.

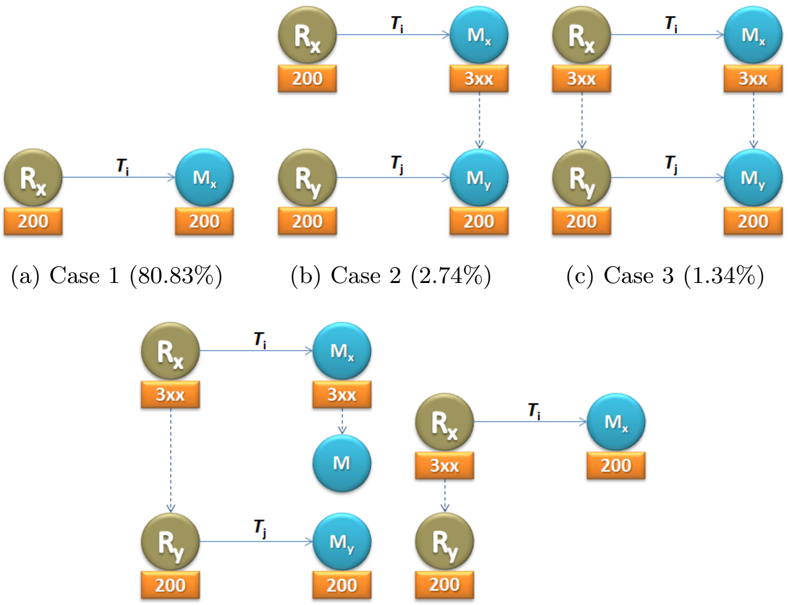
$$Reliability(R) = \frac{\#Mementos\ end\ 200}{|TM|} \tag{2}$$

$$where\ |TM| > 0$$

3.3 HTTP Redirection Relationship between URI-R & URI-M

In this section, we study the relationship between the HTTP status code for the original resource ($URI - R$) and the memento ($URI - M$) which we classify into five cases, shown in Table 1. The column represents the status code on the live web for $URI - R$ and the row represents the status code on the web archive for $URI - M$. Both of cases three and four have redirection for $URI - R$ and $URI - M$. If both of Original and Memento redirect to the same $URI - \overline{R}$, it will be case 3, otherwise it is case 4.

Figure 2 illustrates these cases. The golden circle represents the URI-R in the current Web, and the blue circle represents its memento at time t_i. The orange rectangle represents the HTTP status code. The dashed arrow represents the redirection between two URI-Rs or two mementos. Table 2 shows an example for each case.



(a) Case 1 (80.83%) (b) Case 2 (2.74%) (c) Case 3 (1.34%)

(d) Case 4 (1.33%) (e) Case 5 (13.73%)

Figure 2: URI-R & URI-M HTTP Redirection relationship cases.

Table 1: URI-R & UR-M Relationship.

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

4. EXPERIMENT

To quantify our abstract model (section 3), we sampled 10,000 URIs randomly from Open Directory Project (DMOZ)[3] on January 2012. We selected DMOZ because it is well-represented in web archives [3]. Table 3 shows the distribution of HTTP status code on the current web for our 10,000 sampled URIs. The sample set does not include any shortened [4] nor DOI URIs [19], we consider them as a special case and we will include them in future research.

First, we used the Memento Aggregator to retrieve the TimeMap ($TM(R)$) for each $URI - R$. For each $R \rightarrow \overline{R}$, we also retrieved the TimeMap ($TM(\overline{R})$) for $\overline{R}$.

Secondly, for each M in $TM(R)$, we retrieved its HTTP status code ($Status(M)$). For the mementos with redirection (i.e., $M(R_x) \rightarrow M(R_y)$)), we followed the redirection location and recorded the destination $M(R_y)$, then extracted its original resource R_y. In order to compare the URIs, we performed a canonicalization routine to ensure consistency.

5. RESULTS

From 10,000 URIs sampled from DMOZ, we found 8903 URIs returned TimeMap with total 894,717 mementos. The HTTP status codes distribution for the memento list is shown in Table 4. The table shows that nearly 6% of the mementos have archived redirects (i.e., $URI - M$ with 3xx HTTP status code) where $URI - R$ had a redirection status code at the crawling time.

5.1 Relationship between $TM(R)$ and $TM(\overline{R})$

Assume that $URI - R$ redirects to $URI - \overline{R}$ on the live web. In this section, we will compare the TimeMaps for $URI - R$ and $URI - \overline{R}$.

Figure 3(a) illustrates the relationship between $[TM(R)]$ and $[TM(\overline{R})]$. Each case is defined based on the first and the last Memento-Datetime in the TimeMap for R and $\overline{R}$, it may indicate the lifetime of the URI. For example, in case 1, $TM(\overline{R})$ started and ended before the beginning of $TM(R)$. The figure lists the number of TimeMaps that occurred in each of the seven cases. In case 4, both of $TM(R)$ and $TM(\overline{R})$ are the same. This case occurs when the redirection does not affect the canonicalized form of R (i.e., `http://example.org` redirects to `http://www.example.org`), the web crawler considers both of them as one URI. Case 1, 2 and 5 have low numbers, which means that the existence of $\overline{R}$ was related to the existence of R first. Cases 5, 6, and 7 showed the continuous existence of the $\overline{R}$ on the web even after the disappear of the R.

Figure 3(b) shows the relationship between the number of mementos for the original resource $TM(R)$ (x-axis) and the number of mementos for the redirected resource $TM(\overline{R})$ (y-axis). The red dashed line shows the cases when $|TM(R)| = |TM(\overline{R})|$, it appeared on 16%. In 65% of the cases, the number of mementos $|TM(\overline{R})|$ is less than the number in $|TM(R)|$, and thus the original TimeMap has more Mementos.

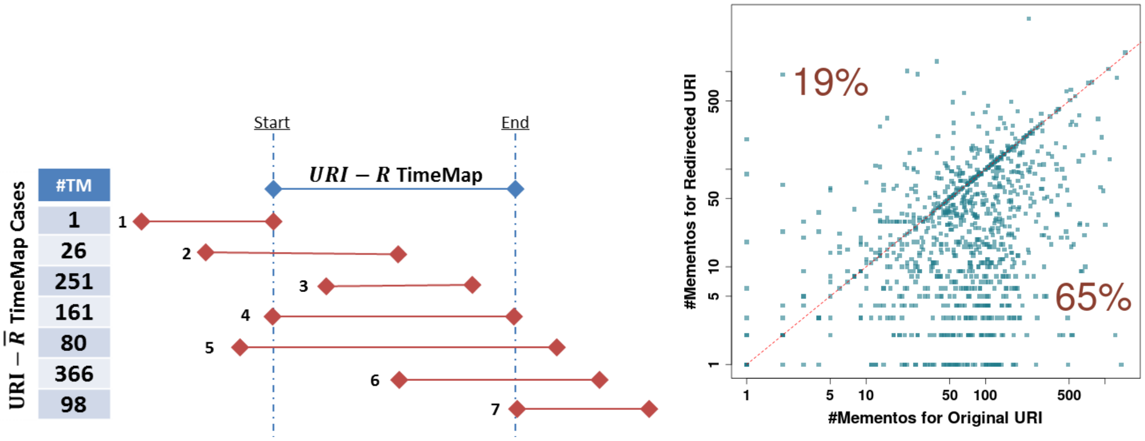
5.2 URI Stability

Figure 4 shows the relationship between the stability of $URI{-}R$ (x-axis) with the number of mementos in its TimeMap (y-axis). The results show that 48% of the URIs are not per-

[3] http://www.dmoz.org

Table 2: URI-R & URI-M Relationship examples (ARCBASE=http://api.wayback.archive.org/memento).

Case	URI Example	Memento Example
Case 1	R_x www.cnn.com	M_x [ARCBASE]/20110729013512/http://www.cnn.com/
Case 2	R_x www.abcsystems.com/ R_y http://www.abcsystems.com/content/abc/	M_x [ARCBASE]/20040612004302/http://www.abcsystems.com/ M_y [ARCBASE]/20040612004302/http://www.abcsystems.com/content/abc/
Case 3	R_x bit.ly/2EEjBl R_y www.cnn.com	M_x [ARCBASE]/20101109032705/http://bit.ly/2EEjBl M_y [ARCBASE]/20101109032705/http://www.cnn.com
Case 4	R_x draculathemusical.co.uk R_y www.dracula-uk.com/index.html	M_x [ARCBASE]/20020212194020/http://www.draculathemusical.co.uk M [ARCBASE]/20020212194020/http://www.geocities.com/draculathemusical M_y [ARCBASE]/20060615010730/http://www.dracula-uk.com/index.html
Case 5	R_x www.emsetal.com/ R_y www.emsetal.com/de/index.php	M_x [ARCBASE]/20080703195602/http://www.emsetal.com/



(a) Time span of TimeMap.

(b) Number of Mementos (16% on the diagonal).

Figure 3: The relationship between TimeMap for the Original ($URI - R$) and the Redirected ($URI - \overline{R}$).

Table 3: Sample URI Current HTTP status code

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

fectly stable across time. The figure shows that large number of mementos have high stability $\simeq 1$.

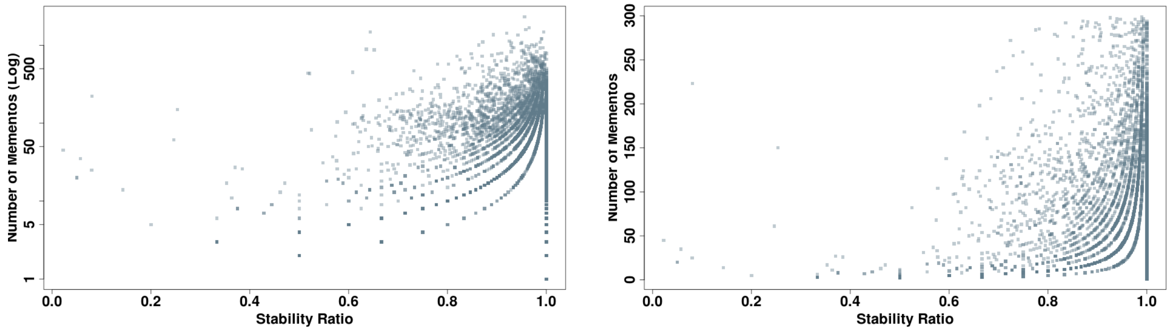
By grouping the memento's status code per TimeMap, we can quantify the different categories and calculate the average stability for each category. Table 5 shows that 52% had 200 status code for all mementos with stability 1.0. Also, 0.62% of the URIs have redirection to the same original URI with stability 1.0.

Table 4: Mementos HTTP status code

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

5.3 URI Reliability

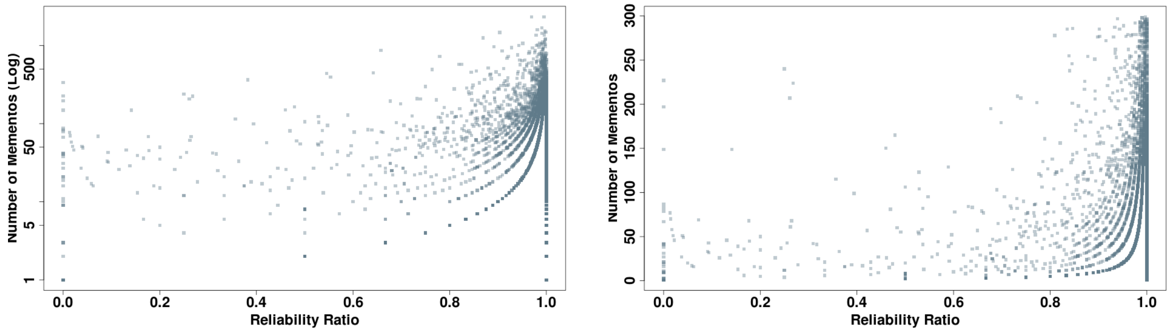
Figure 5 shows the relationship between the URI-R reliability (x-axis) and the number of mementos for each TimeMap (y-axis) that contains at least one memento that has a redirection status code (2890 URIs out of 10,000 URIs). The figure shows the distribution of the reliability, we did not find a strong correlation between the reliability and the number of mementos. Additionally, we calculated the redirection chain (the number of URI that should be followed before reaching 200 status code), and found that 23% did not lead to a successful memento at the end. A few mementos (0.63%) have infinite redirection chains (50+ redirections).



(a) Stability in semi-log scale.

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

Figure 4: URI Stability



(a) Reliability in semi-log scale.

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

Figure 5: URI Reliability

Table 5: Temporal TimeMap Redirection categories.

Timemap Category	%	Stability
All Mementos have OK	52%	1.0
Mementos have mix status code	36%	0.91
All Mementos have Redirection	0.92%	0.85
Redirection to the same URI	0.62%	
Redirection to different URI	0.30%	
URI has no Mementos at all	10.97%	0.0

5.4 HTTP Redirection Relationship between URI-R & URI-M

In this section, we compare between the live web (URI-R status code) and the archived web (TimeMap and mementos status codes). Table 6 shows the distribution of the cases of the relationship between the URI-R and its mementos as illustrated in figure 2. In 19% of the mementos, the client will face HTTP redirection that requires an advanced mechanism to deal with the existence of HTTP redirection status code in both live and archived Web.

Table 7 shows the relationship between the status code on the current web and the status code of the TimeMap. Even though 1471 URIs have HTTP redirection in the current web, only 83 TimeMaps had HTTP redirection status code for all the mementos, while there were 425 TimeMaps with 200 HTTP status code in all the mementos. We can conclude that the HTTP status code on the current web could not give us an indication about the status code of the TimeMap because the URI's HTTP status code could change through time without any rules. During the experiment, we were not able to conclude a pattern for the URI's HTTP status code change.

This quantitative analysis shows the importance of finding new policies, instead of the straightforward URI-R lookup.

6. ARCHIVED HTTP REDIRECTION RETRIEVAL POLICIES

In this section, we develop new policies to query the archive with a URI carried HTTP redirection status code. We will give two policies: policy one, $URI-R$ with an HTTP redirection status code, and policy two, $URI-M$ with an HTTP redirection status code.

6.1 Policy one: URI-R with HTTP redirection

In this case, we have $URI-R$ that redirects to another $URI-\overline{R}$ (for simplicity, $R \rightarrow \overline{R}$); it appeared in 1,471/10,000 URIs in our sample data, and covered with three cases: three, four and five in table 6. The proposed policy is as following:

Required: R and "`Accept-Datetime`" header.

1. Retrieve the memento for R.

Table 6: URI-R - Memento HTTP Redirection relationship cases.

Case Number	URI-R HTTP status code	URI-M HTTP status code	Percentage
Case 1	Non-Redirection	Non-Redirection	80.83%
Case 2	Non-Redirection	Redirection	2.74%
Case 3	Redirection to R$_x$	Redirection to R$_x$ Mementos	1.34%
Case 4	Redirection to R$_x$	Redirection to R$_y$ Mementos	1.33%
Case 5	Redirection	Non-Redirection	13.73%

Table 7: Timemap status compared to the URI-R status on the current web.

URI Status	Count	Timemap Status	Count
OK	8283	Mix status	1849
		All 200 status	5886
		All Redirect status	14
		No Mementos	534
Redirection	1471	Mix status	880
		All 200 status	425
		All Redirect status	83
		No Mementos	83
Not-found	118	Mix status	32
		All 200 status	75
		All Redirect status	2
		No Mementos	9
Others	128	All 200 status	79
		Mix status	30
		No Mementos	19

2. (a) If the retrieved memento has (OK 200) HTTP status code, then return this memento. (Stop)

   (b) Else if the retrieved memento has (Redirection 3xx) HTTP status code, go to policy two.

   (c) If the retrieved memento is unavailable (4xx/5xx) HTTP status code and R has a redirection to $\overline{R}$, use $\overline{R}$ instead of R then go to step 1.

This policy is already implemented in the MementoFox client [20].

6.2 Policy two: URI-M with HTTP redirection

Here, we address case two (see table 6). Assume memento $M(R_x)$ redirects to another memento with a different original $\overline{M}(R_y)$. For example, redirection from a memento for `http://bit.ly/2EEjBl` on 2010-11-09 to another memento for `http://www.cnn.com` on the same date.

```
curl -I http://api.wayback.archive.org/memento/
   20101109032705/http://bit.ly/2EEjBl
HTTP/1.1 301 Moved Permanently
Memento-Datetime: Tue, 09 Nov 2010 03:27:05 GMT
...
Location: http://api.wayback.archive.org/memento/
   20101109032705/http://www.cnn.com/
...
```

In this case, the client will repeat the content negotiation in the datetime dimension for the "`rel=original`" R_y extracted from "`Link`" header for $\overline{M}(R_y)$. For the previous example, the client should repeat the content negotiation with `www.cnn.com` with the requested datetime on 2010-11-09.

Some web archives do not rewrite the memento "`Location`" header, so the memento could redirect to another original resource on the live Web. In this case, policy two will redo the content negotiation using the new original resource instead of redirecting to the live Web.

The new policy extends the default Wayback Machine behavior by retrieving the nearest memento to the redirected $URI-\overline{R}$ which may not be available on the original $URI-R$. Also, applying the new policy on the Memento Aggregator will benefit from the multi-archive environment which may find a better copy in another archive [3].

6.3 Evaluation

6.3.1 *Policy one: URI-R with HTTP Redirection*

This policy focused on 1471 URIs from our sample that had HTTP redirection on the live web. We found 77 URIs that have no mementos at all ($|TM(R)| = 0$). Based on this policy, we were able to retrieve mementos for 17 URIs out of that 77 URIs where $|TM(\overline{R})| > 0|$.

6.3.2 *Policy two: URI-M with HTTP Redirection*

We have 2980 TimeMaps that showed HTTP redirection status code in at least one memento. For these TimeMaps, we followed the memento redirection and extracted the original URIs. We extracted 7115 URIs. The evaluation criteria for this policy is determined by the number of the cases that the policy will contribute to the TimeMap.

Assume that $TM(R) = \{M_1(R), M_2(R), \ldots M_n(R)\}$. For each $M(R)$ that carried HTTP redirection status code, we have $M(\overline{R})$ where $M(R) \rightarrow M(\overline{R})$. In this case, the policy will contribute to the $TM(R)$ if the $TM(\overline{R})$ covers a larger time frame (i.e., $\left[TM(\overline{R})\right] > [TM(R)]$).

From our sample, the policy contributed more mementos to the original TimeMap in 58% of the cases. The rest of the cases, the redirected TimeMap $TM(\overline{R})$ has a less coverage than the original $TM(R)$.

6.3.3 *Discussion*

The existence of the HTTP redirection supported the retrieval process with the required information to reach a better estimation of the presentation of this URI in the past. The policy evaluation showed the ability for the new policies to deliver new mementos that were unreachable using the regular methods. Both policies redo the content negotiation for the redirected URIs (on live or archived web). Policy one uses the live redirect if there is no mementos for the original resource. If there are mementos, the policy two will give the priority to the archived redirected because this is what has been recorded by the web archive in the past. Policy one succeeded in 17/77 of the cases. The second policy extends the TimeMap time span to include mementos from the archived redirected URI. So using the preserved redirection information helps the client to find the nearest memento to the requested datetime. These policies could be implemented in the client side. The client should give the user the ability to optionally select between the different policies.

7. CONCLUSIONS

In this paper, we studied the change of HTTP status code of the URI through the time with focus on the HTTP redirection. Two novel measurements have been proposed, the stability of the URI and the reliability of the URI as a lookup key. Our experiments showed that URIs are not stable through time. We studied the different categories of the TimeMaps with focus on HTTP status code. We found that in 36% of the cases the TimeMap are not fully stable through time. Based on this quantitative study, we concluded two retrieval policies to handle HTTP redirect. The first policy focused on a resource that redirects currently on with redirection on the live Web; it was successful with 22% of the applicable cases. The second policy focused on the mementos with HTTP redirection status code; it extended the original TimeMap in 58% of the applicable cases.

8. ACKNOWLEDGMENTS

This work is supported in part by the Library of Congress and NSF IIS-1009392.

9. REFERENCES

[1] ISO 28500:2009 Information and documentation – WARC file format. `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717`, 2009.

[2] E. Adar, M. Dontcheva, J. Fogarty, and D. S. Weld. Zoetrope: interacting with the ephemeral web. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, UIST '08, pages 239–248, 2008.

[3] S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is Archived? In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 133–136, 2011.

[4] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis. we. b: The web of short URLs. In *Proceedings of the 20th international conference on World Wide Web*, pages 715–724, 2011.

[5] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a country: better strategies than breadth-first for Web page ordering. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, pages 864–872, 2005.

[6] M. Ben Saad and S. Gançarski. Archiving the web using page changes patterns: a case study. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 113–122, 2011.

[7] M. Ben Saad, S. Gançarski, and M. B. Saad. Improving the Quality of Web Archives through the Importance of Changes. In *Proceedings of the 22nd international conference on Database and expert systems applications*, DEXA'11, pages 394–409, Toulouse, France, 2011.

[8] A. Brown. *Archiving websites: a practical guide for information management professionals*. Facet, London, 1st edition, 2006.

[9] J. Cho and H. Garcia-Molina. Effective page refresh policies for Web crawlers. *ACM Transactions on Database Systems (TODS)*, 28(4):390–426, 2003.

[10] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1-7):161–172, Apr. 1998.

[11] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. *Software: Practice and Experience*, 34(2):213–237, 2004.

[12] R. T. Fielding, J. Gettys, J. C. Mogul, H. Frystyk, L. Masinter, P. J. Leach, and T. Berners-Lee. RFC 2616 - Hypertext Transfer Protocol, 1999.

[13] K. Holtman and A. Mutz. RFC 2295 - Transparent Content Negotiation in HTTP. Technical report, W3C, United States, 1998.

[14] A. Jatowt, Y. Kawai, H. Ohshima, and K. Tanaka. What can history tell us?: towards different models of interaction with document histories. In *Proceedings of the 19th ACM conference on Hypertext and hypermedia*, HT '08, pages 5–14. ACM, 2008.

[15] A. Jatowt, Y. Kawai, and K. Tanaka. Visualizing historical content of web pages. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 1221–1222. ACM, 2008.

[16] J. Masanès. *Web archiving*. Springer, Berlin, Heidelberg, 2006.

[17] G. Mohr, M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. An Introduction to Heritrix An open source archival quality web crawler. In *Workshop Proceedings of the 4th International Web Archiving Workshop (IWAW04)*, pages 43–49, 2004.

[18] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? The Evolution of theWeb from a Search Engine Perspective. In *Proceedings of the 13th conference on World Wide Web*, WWW '04, pages 1–12, May 2004.

[19] N. Paskin. Digital object identifiers. *Information Services and Use*, 22(2-3):97–112, 2002.

[20] R. Sanderson, H. Shankar, S. Ainsworth, F. McCown, and S. Adams. Implementing Time Travel for the Web. *Code4Lib Journal*, (13), 2011.

[21] J. Teevan, S. T. Dumais, D. J. Liebling, and R. L. Hughes. Changing how people view changes on the web. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, pages 237–246, 2009.

[22] B. Tofel. 'Wayback' for Accessing Web Archives. In *Proceedings of 7th International Web Archiving Workshop*, IWAW '07, 2007.

[23] H. Van de Sompel, M. L. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states. Technical report, Internet Engineering Task Force (IETF), 2011.

[24] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar. Memento: TimeMap API for Web Archives. `http://www.mementoweb.org/events/IA201002/slides/memento_201002_TimeMap.pdf`, 2010.