Why We Need Private Web Archives: Almost Two-Thirds of Web Traffic IS NOT Publicly Archivable

Hussam Hallak and Michael L. Nelson
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
{hhallak,mln}@cs.odu.edu

ABSTRACT

The limitations of public web archiving encouraged the development of private web archiving tools. In this paper, we show that private web archiving is crucial because, at least, 31.91% of the entire web traffic is not archivable by public web archives. This is due to the increase use of personalized/private web pages and the use of technologies that, unintentionally, prevent web archives' crawlers from crawling and archiving these pages. Our experiment shows that the percentage of not publicly archivable web traffic can be as high as 83.05%, but the more likely case is that around 65% of web traffic is not publicly archivable. Furthermore, no matter how good public web archives get at capturing web pages, there will always be a significant number of web pages that are not publicly archivable.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; **World Wide Web**.

KEYWORDS

Memento, Web Archiving, paywall

ACM Reference Format:
Hussam Hallak and Michael L. Nelson. 0000. Why We Need Private Web Archives: Almost Two-Thirds of Web Traffic IS NOT Publicly Archivable. In *Proceedings of A Boilerplate Conference (ABC '00).* ACM, New York, NY, USA, 4 pages. https://doi.org/00.000/000_0

1 INTRODUCTION

The importance of private web archiving has not received enough attention from research community. Furthermore, no one has tried to estimate the amount of web traffic that can be preserved using public web archives. In this work, we propose a method for estimating publicly archivable web traffic and classifying web pages into one of two classes, publicly archivable, or not publicly archivable. In terms of the ability to be archived in public web archives, web pages fall into one of two categories: publicly archivable, or not publicly archivable. The first category includes pages that can be accessed without login/authentication. In other words, the pages do not reside behind a paywall. Web pages with a minimum amount of content based on GeoIP and/or user-specific personalization are also publicly archivable. On the other hand, not publicly archivable web pages include search engines' result pages (SERPs) and web pages behind paywalls like banks' statements, email accounts' pages, and some social media pages. Furthermore, a decent portion of public web pages are not archivable because they use client-side scripts that fetch new content based on user interaction with the page like scrolling. Web pages with completely personalized content are not publicly archivable. There are few emerging tools that allow personal/private web archiving like WebRecorder, WAIL, WarCreate, and Memento Tracer. Even if/when public web archives use Brozzler and resolve the problems with Ajax, they cannot be expected to have pages requiring authentication, nor pages with effectively infinite inputs like search engines and maps.

2 PUBLICLY ARCHIVABLE WEB PAGES

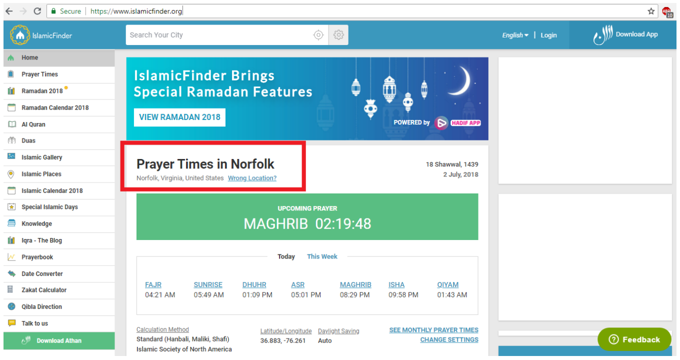
Web pages that can be accessed without login/authentication are archivable using public web archives. These pages do not reside behind a paywall. Grant Atkins examined paywalls in the Internet Archive for news sites and found that web pages behind paywalls may actually be redirecting to a login page at crawl time [1]. Examples of publicly archivable pages include those that do not require authentication to view the page and do not use client-side scripts (i.e., Ajax) to load additional content. What you see in the web browser and what you can replay from public web archives are exactly the same.

Some web pages provide "personalized" content depending on the GeoIP of the requester. In these cases, what you see in the browser and what you can replay from public web archives are nearly the same, except for some minor personalization/GeoIP related changes. For example, a user requesting https://www.islamicfinder.org from Suffolk, Virginia will see the prayer times for the closest major city (Norfolk, Virginia) Figure[1]. On the other hand, when the Internet Archive crawls the page, it sees the prayer times for San Bruno, California. This is likely because the crawling/archiving is happening from San Francisco, California Figure[1]. The two pages, otherwise, are exactly the same.

Some social media sites, like Twitter, are publicly archivable and the Internet Archive captures most of their content. Twitter's home page is personalized, so user-specific contents, like "Who to Follow" and "Trends for you" are not captured, but the tweets are. Although some Twitter services require authentication and the archived memento for users' accounts show a message that cookies

ABC '00, January 0000, Norfolk, Virginia, USA

ACM ISBN 000-0000-00-000/00/00…$0.00
https://doi.org/00.000/000_0


https://www.islamicfinder.org

IslamicFinder Brings Special Ramadan Features

Prayer Times in Norfolk
Norfolk, Virginia, United States Wrong Location?

18 Shawwal, 1439
2 July, 2018

UPCOMING PRAYER
MAGHRIB 02:19:48

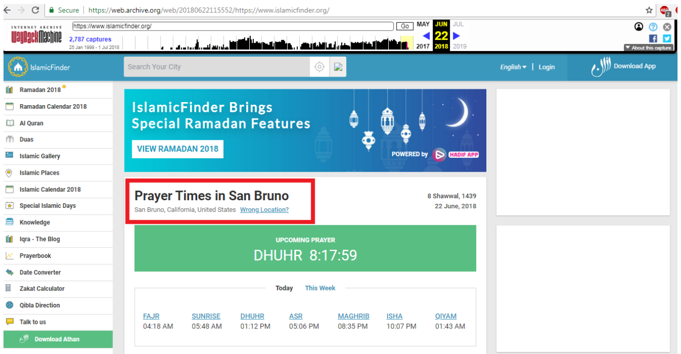
FAJR	SUNRISE	DHUHR	ASR	MAGHRIB	ISHA	QIYAM
04:21 AM	05:49 AM	01:09 PM	05:01 PM	08:29 PM	09:58 PM	01:43 AM

Calculation Method
Standard (Hanbali, Maliki, Shafi)
Islamic Society of North America

Latitude/Longitude
36.883, -76.261

Daylight Saving
Auto

Figure 1: The live version of https://www.islamicfinder.org for a user in Suffolk, VA on 2018-07-02


https://web.archive.org/web/20180622115552/https://www.islamicfinder.org/

https://www.islamicfinder.org/
2,787 captures
25 Jan 1999 - 1 Jul 2018

IslamicFinder Brings Special Ramadan Features

Prayer Times in San Bruno
San Bruno, California, United States Wrong Location?

8 Shawwal, 1439
22 June, 2018

UPCOMING PRAYER
DHUHR 8:17:59

FAJR	SUNRISE	DHUHR	ASR	MAGHRIB	ISHA	QIYAM
04:18 AM	05:48 AM	01:12 PM	05:06 PM	08:35 PM	10:07 PM	01:43 AM

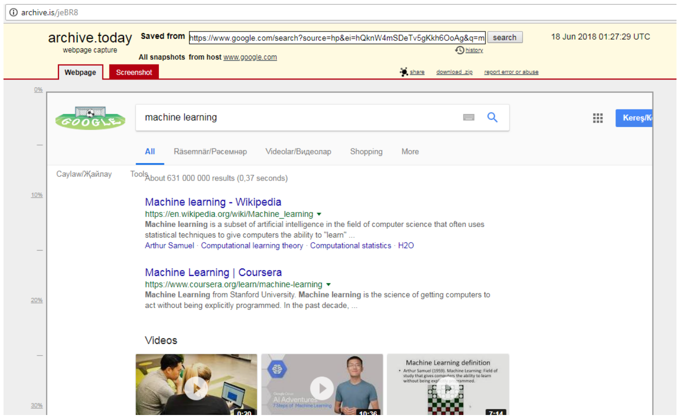
Figure 2: Memento for https://www.islamicfinder.org from the Internet Archive captured on 2018-06-22

are used and they are important for an enhanced user experience, the main content of the page, tweets, is preserved (or at least the top-k tweets, since the crawler does not automatically scroll at archive time to activate the Ajax-based pagination, which is considered Harmful to web archiving [3].

3 NOT PUBLICLY ARCHIVABLE WEB PAGES:

According to similarweb.com, search engines are at the top when it comes to the amount of traffic websites receive. Google is number one; its share is 10.10% of the entire web traffic. The Internet Archive crawls it on regular basis, and has over 0.5 million mementos as of 2018-05-01. The captured mementos are exact copies as far as the look, but obviously not a functioning search page, so we considered search engines to be not publicly archivable. Although it is possible to push a search result page from Google to a public web archive like archive.is, but that is not how web archives are normally used Figure[3].

Furthermore, it is not viable for web archives to try to archive search engines' result pages (SERPs) because there is an infinite number of possible URIs due to an infinite number of search queries and syntax, so even if we preserve an SERP from july 2018, we are unable to issue new queries against a July, 2018 version of Google. Similarly, Maps and other applications that depend on user interaction can be archived, but we typically don't consider the entire application "archived".


archive.is/jeBR8

archive.today
webpage capture

Saved from https://www.google.com/search?source=hp&ei=hQknW4mSDeTv5gKkh6OoAg&q=m
18 Jun 2018 01:27:29 UTC

All snapshots from host www.google.com

machine learning

About 631 000 000 results (0,37 seconds)

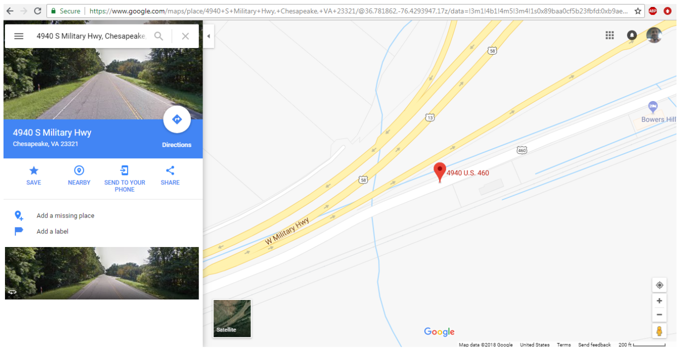
Machine learning - Wikipedia
https://en.wikipedia.org/wiki/Machine_learning
Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" ...
Arthur Samuel · Computational learning theory · Computational statistics · H2O

Machine Learning | Coursera
https://www.coursera.org/learn/machine-learning
Machine Learning from Stanford University. **Machine learning** is the science of getting computers to act without being explicitly programmed. In the past decade, ...

Videos

Figure 3: A Google search query for "Machine Learning" on 2018-06-18 archived in archive.is

Even when web archives use headless browsers to overcome the Ajax problem, there can be additional challenges. For example, we pushed a page from Google Maps with an address in Chesapeake, Virginia to archive.today and the result was a page from Google support (in Russian) telling me that I (or more accurately, archive.today) need to update the browser in order to use Google Maps Figures[4, 5]. While technically not a paywall, there is now something in the web archive corresponding to that Google Maps URI, but it does not match the users' expectations. It also reveals a clue about the GeoIP of archive.today.

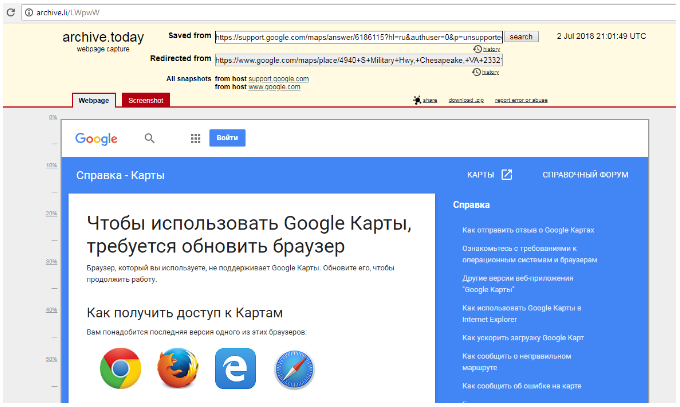

https://www.google.com/maps/place/4940+S+Military+Hwy,+Chesapeake,+VA+23321/@36.781862,-76.4293947,17z/data=!3m1!4b1!4m5!3m4!1s0x89baa0cf5b23fbfd:0xb9ae...

4940 S Military Hwy, Chesapeake,

4940 S Military Hwy
Chesapeake, VA 23321

Figure 4: Google Maps page for the address 4940 S Military HWY, Chesapeake, VA 23321 pushed to archive.today on 2018-07-02

Different social media pages respond differently when web archives' crawlers try to crawl and archive them. Public web archives might have mementos of some social media pages, however, they often require a login to allow the download of the pages' representation. Otherwise, a redirection takes place. Another obstacle faces archiving social media pages is their heavy use of client-side executed scripts that will, for example, fetch new content when the page is scrolled or when hiding/showing comments with no change in the



archive.li/LWpwW

archive.today
webpage capture

Saved from https://support.google.com/maps/answer/6186115?hl=ru&authuser=0&p=unsupporte

Redirected from https://www.google.com/maps/place/4940+S+Military+Hwy,+Chesapeake,+VA+2332

2 Jul 2018 21:01:49 UTC

All snapshots from host support.google.com
from host www.google.com

Чтобы использовать Google Карты, требуется обновить браузер

Браузер, который вы используете, не поддерживает Google Карты. Обновите его, чтобы продолжить работу.

Как получить доступ к Картам

Вам понадобится последняя версия одного из этих браузеров:

search
history
history
Webpage
Screenshot
share
download .zip
report error or abuse
0%
10%
20%
30%
40%
50%
Google
Войти
Справка - Карты
КАРТЫ
СПРАВОЧНЫЙ ФОРУМ
Справка
Как отправить отзыв о Google Картах
Ознакомьтесь с требованиями к операционным системам и браузерам
Другие версии веб-приложения "Google Карты"
Как использовать Google Карты в Internet Explorer
Как ускорить загрузку Google Карт
Как сообщить о неправильном маршруте
Как сообщить об ошибке на карте

Figure 5: Memento for the Google Maps page I pushed to archive.today on 2018-07-02

URI. Facebook, for example, does not allow web archives’ crawlers to access the majority of its pages. The Internet Archive’s Wayback Machine returned 1,699 mementos for the former president’s official Facebook page, but when we opened one of these mementos, it returned the infamous Facebook login or register page.

There are few exceptions where the Internet Archive is able to archive some user-contributed Facebook pages. With these exceptions in mind, it is still safe to say that Facebook pages are not publicly archivable.

Linkedin shares the same behavior with Facebook. The notifications page has 46 mementos as of 2018-05-29, but they are entirely empty. The live page contains notifications from contacts such as who is having a birthday, job anniversary, got a new job, and so on. This page is completely personalized and requires a cookie or login to display information that is related to the user, and therefore, the Internet Archive has no way of downloading its representation.

We consider Amazon and other E-commerce websites to be not publicly archivable. Amazon’s "yourstore" page contains recommended items that are based the user’s behavior. We found 111 mementos for "my Amazon’s your store page" in the Internet Archive, and opened one of them to see what has been captured. As expected, the page has a redirect to another page that asks for a login. It returned a 302 response code when it was crawled by the Internet Archive. The actual content of the original page was not archived because the IA crawler does not provide credentials to download the content of the page. The representation saved to the Internet Archive is for a resource different from the originally crawled page. There are many web sites with this behavior, so it is safe to assume that for some web sites, even when there are plenty of mementos, they all might return a soft 404 [5].

4 PERSONAL/PRIVATE WEB ARCHIVES:

There are emerging tools like Webrecorder, WARCreate, WAIL, and Memento Tracer for personal web archiving (or community tools in the case of Tracer), but even if/when the Internet Archive replaces Heritrix with Brozzler and resolves the problems with Ajax, their Wayback Machine cannot be expected to have pages requiring authentication, nor pages with effectively infinite inputs like search engines and maps.

5 ESTIMATING THE AMOUNT OF ARCHIVABLE WEB TRAFFIC:

5.1 Dataset:

To explore the amount of web traffic that is archivable, we examined the top 100 sites as ranked by alexa.com, and manually constructed a data set of those 100 sites using traffic analysis services from similarweb and semrush.com. The data was collected on 2018-02-23 and we captured three web traffic measures offered by both websites, total visits, unique visits, and pages per visit.

- Total visits is the total number of non-unique visits from last month.
- Unique visits is the number of unique visits from last month.
- Pages per visit is the average number of visited pages per user’s visit.

We determined whether or not a website is archivable based on the discussion provided earlier, and put it all together in a csv file to use it later as input for the script we wrote and uploaded the dataset with the script to Github.com [2].

5.2 Methodology:

Using Python 3, we wrote a simple script that calculates the percentage of web traffic that is publicly archivable. We are assuming that the top 100 sites is a good representative of the whole web. We are aware that 100 sites is a small number compared to 1.7 billion live websites on the Internet [4], but according to similarweb.com, the top 100 sites receive 48.86% of the entire traffic on the web which is consistent with a Pareto distribution [6]. The program offers six different results, each of which is based on a certain measure or a combination of measures, total visits, unique visits, and pages per visit. Flags can be set to control what measures are used in the calculation. If no flags are set, the program shows all the results using all three measures and their combination. We came up with this formula to calculate the percentage of publicly archivable websites based on all three measures combined:

(1) Multiply the pages/visit by visits for each web site from both SimilarWeb and SemRush
(2) Take the average for both sources, SimilarWeb and SemRush
(3) Take the average of unique visits for each website from SimilarWeb and SemRush
(4) Add the numbers obtained in 2 and 3
(5) Add the number obtained in 4 for all archivable websites
(6) Add the number obtained in 4 for all non-archivable websites
(7) Add the numbers obtained in 5 and 6 to get the total
(8) Calculate the percentage of the numbers obtained in 5 and 6 from the total, obtained in 7

5.3 Equation:

This is the formula using all measures:

For each website of the 100 top websites:

$$V = \frac{\text{Pages/Visit}_{\text{sem}} \times \text{Visits}_{\text{sem}} + \text{Pages/Visit}_{\text{sim}} \times \text{Visits}_{\text{sim}}}{2}$$

$$U = \frac{\text{UniqueVisit}_{\text{sem}} + \text{UniqueVisit}_{\text{sim}}}{2}$$

$$\tau = U + V$$

We use a, n to represent the number of publicly archivable websites, and the number of not publicly archivable websites respectively.

For our dataset: a + n = 100

$$T = \sum_{i=0}^{a} \tau_i + \sum_{j=0}^{n} \tau_j$$

The percentage of publicly archivable traffic is:

$$P_a = \frac{\sum_{i=0}^{a} \tau_i}{T}$$

The percentage of not publicly archivable traffic is:

$$P_n = \frac{\sum_{j=0}^{n} \tau_j}{T}$$

Using all measures, I found that 65.30% of the traffic of the top 100 sites is not archivable by public web archives.

Now, it is possible to discuss three different scenarios and compute a range. If the top 100 sites receive 48.86% of the traffic, and 65.30% of that traffic is not publicly archivable, therefore:

- If all of the remaining web traffic is publicly archivable, then 31.91% of the entire web traffic is not publicly archivable. 65.30 * 0.4886 = 31.91.
- If the remaining web traffic is similar to the traffic from the top 100 sites, then 65.30% of the entire web traffic is not publicly archivable.
- If all of the remaining web traffic is not publicly archivable, then only 16.95% of the entire web traffic is archivable. 34.7 * 0.4886 = 16.95. This means that 83.05% of the entire web traffic is not publicly archivable.

5.4 Results:

So the percentage of not publicly archivable web traffic is between 31.91% and 83.05%. More likely, it is close to 65.30% (the second case).

I would like to emphasize that since the top 100 websites are mainly Google, Bing, Yahoo, etc, and their derivatives, the nature of these top sites is the determining factor of my results. However, since the range has been calculated, it is safe to say that, at least, 1/3 of the entire web traffic is not publicly archivable. This percentage constitutes the necessity of private web archives. There are few available tools to solve this problem, Web Recorder, Warcreate, and WAIL. Public web archiving sites like the Internet Archive, archive.is, and others will never be able to preserve personalized or private web pages like emails, bank accounts, etc.

6 CONCLUSIONS AND FUTURE WORK

Personal web archiving is crucial since, at least, 31.91% of the entire web traffic is not archivable by public web archives. This is due to the increase use of personalized/private web pages and the use of technologies hindering the ability of web archives’ crawlers to crawl and archive these pages. The experiment shows that the percentage of not publicly archivable web traffic can be as high as 83.05%, but the more likely case is that around 65% of web traffic is not publicly archivable. Unfortunately, no matter how good public web archives get at capturing web pages, there will always be a significant number of web pages that are not publicly archivable. This emphasizes the need for personal web archiving tools, such as Web Recorder, Warcreate, and WAIL - possibly combined with a collaboratively-maintained repository of how to interact with complex sites, as introduced by Memento Tracer. Even if Ajax-related web archiving problems were eliminated, no less than 1/3 of web traffic is to sites that will otherwise never appear in public web archives.

7 ACKNOWLEDGEMENTS

TODO: To be updated later!

REFERENCES

[1] Grant Atkins. 2018. Paywalls in the Internet Archive. http://ws-dl.blogspot.com/2018/03/2018-03-15-paywalls-in-internet-archive.html.
[2] Hussam Hallak. 2018. ArchivableWeb. https://github.com/HussamHallak/ArchivableWeb.
[3] Corren Mccoy. 2017. Pagination Considered Harmful to Archiving. https://ws-dl.blogspot.com/2017/09/2017-09-13-pagination-considered.html.
[4] Internet Live Stats. 2019. Total number of Websites. https://www.internetlivestats.com/total-number-of-websites/.
[5] Google Support. 2019. Soft 404 errors. https://support.google.com/webmasters/answer/181708?hl=en.
[6] Wikipedia. 2019. Pareto Distribution. https://en.wikipedia.org/wiki/Pareto_distribution.