International Journal on Digital Libraries
https://doi.org/10.1007/s00799-024-00397-2

Robots still outnumber humans in web archives in 2019, but less than in 2015 and 2012

Himarsha R. Jayanetti[1] · Kritika Garg[1] · Sawood Alam[2] · Michael L. Nelson[1] · Michele C. Weigle[1]

Received: 9 January 2023 / Revised: 22 January 2024 / Accepted: 24 January 2024

Abstract
The significance of the web and the crucial role of web archives in its preservation highlight the necessity of understanding how users, both human and robot, access web archive content, and how best to satisfy this disparate needs of both types of users. To identify robots and humans in web archives and analyze their respective access patterns, we used the Internet Archive's (IA) Wayback Machine access logs from 2012, 2015, and 2019, as well as Arquivo.pt's (Portuguese Web Archive) access logs from 2019. We identified user sessions in the access logs and classified those sessions as human or robot based on their browsing behavior. To better understand how users navigate through the web archives, we evaluated these sessions to discover user access patterns. Based on the two archives and between the three years of IA access logs (2012 vs. 2015 vs. 2019), we present a comparison of detected robots vs. humans and their user access patterns and temporal preferences. The total number of robots detected in IA 2012 (91% of requests) and IA 2015 (88% of requests) is greater than in IA 2019 (70% of requests). Robots account for 98% of requests in Arquivo.pt (2019). We found that the robots are almost entirely limited to "Dip" and "Skim" access patterns in IA 2012 and 2015, but exhibit all the patterns and their combinations in IA 2019. Both humans and robots show a preference for web pages archived in the near past.

Keywords Web archiving · User access patterns · Web server logs · Web usage mining · Web robot detection

1 Introduction

The web has become ingrained in our lives, influencing our daily activities. Preserving the web through web archives is more important than before. With over 686 billion web pages archived [30] dating back to 1996, the Internet Archive (IA) is the largest and oldest of the web archives. The Wayback Machine, which can replay past versions of websites, is a public service provided by IA. Arquivo.pt [15, 23] has been archiving millions of files from the Internet since 1996. Both web archives contain information in a variety of languages and provide public search capabilities for historical content.

Previous research has predominantly concentrated on examining user behaviors within several domains of the live web, such as e-commerce platforms, search engine interactions, and general website users and usage [20, 52, 61, 65, 67]. A recent avenue gaining attention encompasses exploring user behaviors for security and intrusion detection [27, 29, 53, 57, 66]. This method of analysis serves multiple purposes. It assists in understanding user preferences toward products or events. Additionally, it aids in recognizing potentially suspicious behavior across various online platforms concerning security and privacy by evaluating their specific traits [18]. However, despite these focused studies, there remains a substantial gap in comprehensively exploring user interactions specifically within web archives. It is important to understand accesses to web archives as it provides invaluable insights for maximizing the use of limited web archive resources. It also helps efficient maintenance and organization of web archive data effectively for future use.

✉ Himarsha R. Jayanetti
hjaya002@odu.edu

Kritika Garg
kgarg001@odu.edu

Sawood Alam
sawood@archive.org

Michael L. Nelson
mln@cs.odu.edu

Michele C. Weigle
mweigle@cs.odu.edu

[1] Department of Computer Science, Old Dominion University, Norfolk, VA, USA

[2] Wayback Machine, Internet Archive, San Francisco, CA, USA

Published online: 07 March 2024

Our study is an extension of a previous study by AlNoamany et al. [7] that examined access patterns for robots and humans in web archives based on a web server log sample from 2012 from the Wayback Machine. By using several heuristics including browsing speed, image-to-HTML ratio, requests for robots.txt, and User-Agent strings to differentiate between robot and human sessions, AlNoamany et al. determined that in the IA access logs in 2012, humans were outnumbered by robots 10:1 in terms of sessions, 5:4 in terms of raw HTTP accesses, and 4:1 in terms of megabytes transferred. The four web archive user access patterns defined in the previous study are **Dip** (single access), **Slide** (the same page at different archive times), **Dive** (different pages at roughly the same archive time), and **Skim** (lists of what pages are archived, i.e., TimeMaps).

In our initial study [34], we revisited the work of AlNoamany et al. by examining user accesses to web archives using three different datasets from anonymized server access logs: 2012 Wayback Machine (**IA2012**), 2019 Wayback Machine (**IA2019**), and 2019 Arquivo.pt (**PT2019**). In this study, we examined a new dataset of 2015 Wayback Machine anonymized server access logs (**IA2015**). Using these datasets, we identify human and robot access, identify important web archive access patterns, and discover the temporal preference for web archive access. We add to AlNoamany et al.'s criteria for distinguishing robots from humans by making a few adjustments. These heuristics will be discussed in detail in Sect. 3.4.

The following are the primary contributions of our study:

1. We used a full-day's worth of four web archive access logs datasets (IA2012, IA2015, IA2019, PT2019) to distinguish between human and robot access. The total number of robots detected in IA2012 (91% of requests) and IA2015 (88% of requests) is greater than IA2019 (70% of requests). Robots account for 98% of requests in PT2019.
2. We looked at different access patterns exhibited by web archive users (humans and robots). We found out that the robots are almost entirely limited to Dip and Skim in IA2012 and IA2015, but exhibit all the established patterns and their combinations in IA2019.
3. We explored human and robot users' temporal preferences for web archive content. The majority of requests were for mementos [63] that were near to the datetime of each access log dataset, suggesting a preference for the archived content in the recent past.

In this paper, we are attempting to understand who accesses the web archives. To be clear, we are not making any value judgments about robots, because we recognize that not all bots are bad. For example, there are beneficial services like Internet Archive Scholar [50], ArchiveReady [9], TMVis [41], and MemGator [3] that are built on top of web archives. But the needs of interactive users are different from those of robots, and we can better design and implement API access for robots (e.g., [19, 42, 48, 55]) if we better understand how robots are using the interfaces designed for interactive users.

2 Background and related work

Web clients and servers communicate using the hypertext transfer protocol (HTTP) [21]. Web clients (such as a web browser or web crawler) make HTTP requests to web servers using a set of defined methods, such as GET, HEAD, and POST to interact with resources [2]. For instance, the GET method is used to request a resource, the POST method is employed to update a resource with specific information, and the HEAD method is similar to a GET request, but it exclusively requests for metadata without fetching the actual content (payload). Web servers respond using a set of defined HTTP status codes, headers, and payload (if any). The HTTP Status Codes convey the outcome of the request (200 OK, 404 Not Found, etc.), headers provide metadata about the response (content type, server details, etc.), and the payload contains the actual data being sent back to the client (HTML, JSON, images, etc.).

Web server logs are records containing information about requests, responses, and errors processed by a web server. Extracting useful data from web server logs and analyzing user navigation activity is referred to as web usage mining [47, 59, 64]. Numerous studies have been conducted for analyzing different web usage mining techniques as well as to identify user access patterns on the Internet [40, 44]. Web usage mining is used to increase the personalization of web-based applications [46, 51]. Mobasher et al. [45] developed an automatic personalization technique using multiple web usage mining approaches. Web usage mining has also been applied in user profiling [14, 25], web marketing initiatives [10], and enhancing learning management systems [68, 68].

The goal of web archives is to capture and preserve original web resources (URI-Rs). Each capture, or memento (URI-M), is a version of a URI-R that comes from a fixed moment in time (Memento-Datetime). The list of mementos for a particular URI-R is called a TimeMap (URI-T). All of these notions are outlined in the Memento Protocol [49, 63].

In this work, we look at web archive server access logs and perform web usage mining in the context of web archives. There has been past work in how users utilize and behave in web archives [6, 16, 17, 22, 24, 28], including the 2013 study [7] that we revisit. Web archives maintain their web server access logs as plain text files that record each request to the web archive. Most HTTP servers use the standard Common Log Format or the extended Combined Log Format to record their server access logs [8]. An example access log entry from Arquivo.pt web archive is shown in Fig. 1. A single log entry

consists of the IP address of the client, user identity, authenticated user's ID, date and time, HTTP method, request path, HTTP version, HTTP status code, and size of the response in bytes, referrer, and User-Agent (left to right). The request path on this log entry show that this is a request to a URI-M. The client IP address is anonymized in the access log datasets for privacy reasons. Alam has implemented an HTTP access log parser [1], with exclusive features for web archive access logs, which can be used to process such web archive access logs.

AlNoamany et al.'s previous work [7] in 2013 set the groundwork for this study. In addition to their analysis of the prevalence of robot and human users in the Internet Archive, they also proposed a set of basic user access patterns for users of web archives:

Dip—The user accesses only one URI (URI-M or URI-T).
Slide—The user accesses the same URI-R at different Memento-Datetimes.
Dive—The user accesses different URI-Rs at nearly the same Memento-Datetime (i.e., dives deeply into a memento by browsing links of URI-Ms).
Skim—The user accesses different TimeMaps (URI-T).

In a separate study, AlNoamany et al. looked into the Wayback Machine's access logs to understand who created links to URI-Ms and why [5, 6]. They found that web archives were more often used to visit pages no longer on the live web (as opposed to prior versions of pages still on the web), and much of the traffic came from sites like Wikipedia.

Alam et al. [4] describe archival voids, or portions of URI spaces that are not present in a web archive. They created multiple archival void profiles using Arquivo.pt access logs, and while doing so, identified and reported access patterns, status code distributions, and issues such as Soft-404 (when a web server responds with an HTTP `200 OK` status code for pages that are actually error pages [43]). While their research is very similar to ours, the mentioned access patterns differ from ours. Their study looks at which users are accessing the archive and what they request, whereas we explain how a user (robot or human) might traverse through an archive.

3 Methodology

In this work, we leverage cleaned access logs after preprocessing raw access logs to identify user sessions, detect robots, assess distinct access patterns used by web archive visitors, and finally check for any temporal preferences in user accesses. The steps of our analysis are shown in Fig. 2. The code [31] and visualizations [32] are published, and each step is explained in detail in this section.

3.1 Dataset

In this study, we are using four **full-day** access log datasets from two different web archives: February 2, 2012 access logs from the Internet Archive (IA2012); February 5, 2015 access logs from the Internet Archive (IA2015); and February 7, 2019 access logs from Internet Archive (IA2019) and Arquivo.pt (PT2019). We chose the first Thursday of February for our datasets to align with the prior analysis performed on a much smaller sample (2 million requests representing about 30 min) from the Wayback access logs from February 2, 2012 [7].

The characteristics of the raw datasets are listed in Table 1. We show the frequency of HTTP request methods and HTTP response codes, among other features. HTTP GET is the most prevalent request method (>98%) present in all three datasets, while the HTTP HEAD method accounts for less than 2% of requests.

Due to the practice of web archives redirecting from the requested Memento-Datetime to the nearest available memento, all four of our samples have numerous 3xx requests. IA2012 has about 53% 3xx requests, IA2015 has about 40% 3xx requests, and IA2019 has about 43% 3xx requests out of the total number of requests in the respective samples. About 20% of requests are 3xx in PT2019, due to the same behavior. IA2015 and IA2019 have a higher number of requests to embedded resources (about 63%) followed by IA2012 (44%), whereas PT2019 has only 20%. IA2015 has the highest percentage of requests with a null referrer field (78%) whereas IA2012 has around 48% requests with a null referrer field. The percentage number of requests with a null referrer field has reduced by nearly four times between IA2015 (78%) and IA2019 (20%). There is an increase in the percentage of self-identified robots (SI robots) from IA2012 (0.01%) and IA2015 (0.04%) to IA2019 (0.15%). The percentage of SI robots in PT2019 is as twice that in IA2019. We used some of these features (HEAD requests, embedded resources, and SI robots) in the bot identification process (covered in Sect. 3.4).

3.2 Data cleaning

An overview of our data cleaning process is shown in Fig. 2. In the Stage 1 data cleaning (S1), we removed the log entries that were either invalid or irrelevant to the analysis. We only kept legitimate requests to web archive content (mementos and TimeMaps) and requests to the web archive's robots.txt. The robots.txt requests were preserved since they will be utilized as a bot detection heuristic later on in our process.

After S1 data cleaning, we identified user sessions in each of our three datasets (Sect. 3.3) and conducted bot identification (Sect. 3.4). Stage 2 data cleaning (S2) takes place only after the requests were flagged as human or robot. Our study's

```
128.82.7.3 - - [07/Jul/2019:04:44:14 +0100] "GET/wayback/20091223043049/http://www.cs.odu.edu/
HTTP/1.1" 200 9593 "-" "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:48.0) Gecko/20100101 Fire-
fox/48.0"
```

Fig. 1 A sample access log entry from the PT2019 dataset (Fields: IP address of the client, user identity, authenticated user's ID, date and time, HTTP method, request path, HTTP version, HTTP status code, size of the response in bytes, referrer, and User-Agent)

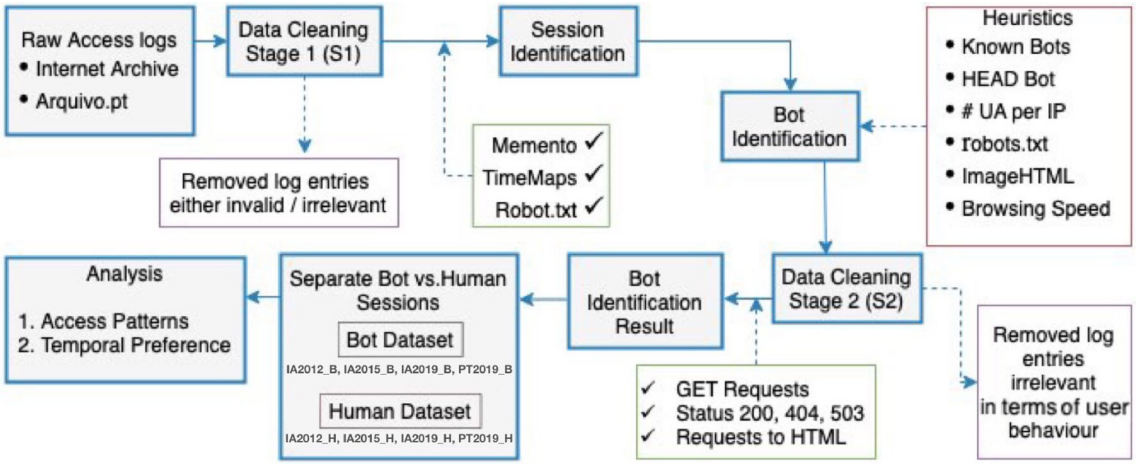

Fig. 2 A chart illustrating the phases in our analytical procedure

ultimate goal was to detect user access patterns of robots and humans in our datasets, and to do so, we must ensure that the refined datasets only included requests that a user would make. As a result, in S2, we purged log items that were unrelated in terms of user behavior. This includes the browser's automatic requests for embedded resources, any requests using a method other than HTTP GET, and requests generating responses with status codes other than 200, 404, and 503. Several of these requests, including embedded resources and HEAD requests, were necessary during the bot detection phase. Thus, we had to follow a two-step data cleaning approach.

Table 2 shows the number of requests for each dataset after each cleaning stage. The percentages are based on the raw dataset's initial number of requests. PT2019 had a higher percentage of requests remaining after S2 compared to IA2012, IA2015, and IA2019. This could be related to the raw dataset's low percentage of embedded resources (20%) in the PT2019 dataset (Table 1).

3.3 Session identification

After S1 data cleaning, the next phase in our study was session identification (Fig. 2). A session can be defined as a set of interactions by a particular user with the web server within a given time frame. We split the requests into different user sessions after S1 data cleaning. First, we sorted all of the requests by IP and User-Agent, then identified the user sessions based on a 10-minute timeout threshold similar to the prior study's process [7]. That is, if the interval between two consecutive requests with the same IP and User-Agent is longer than 10 min, the second request is considered as the start of the next session for that user.

3.4 Bot identification

As the next step in our process, we employed a heuristic-based strategy to identify robot requests (Fig. 2). We used the original five heuristics used in prior work [7] (User-Agent check, number of User-Agents per IP, robots.txt file, browsing speed, and Image-to-HTML ratio) with some minor adjustments to improve the performance of the robot detection. Additionally, we have introduced a new heuristic named "the Type of HTTP request method" to identify robot accesses. The following sub-sections will go through each heuristic in detail. The real-world examples for each heuristic taken from the web archive access logs are shown in the appendices.

3.4.1 Known bots

We created a list of User-Agents that are known to be used by bots. We first constructed UA_l, a list of all User-Agent strings from our three datasets. From this list, we compiled UA_m by filtering for User-Agent strings that contained robot keywords, such as "bot," "crawler," and "spider.". We compiled a separate bot User-Agent list UA_d by running our full list UA_l through DeviceDetector [12], a parser that fil-

Table 1 Features for each dataset: February 2, 2012 from IA (IA2012); February 5, 2015 from IA (IA2015); February 7, 2019 from IA (IA2019); and February 7, 2019 from Arquivo.pt (PT2019)

Feature	IA2012 Feb 2, 2012	IA2015 Feb 5, 2015	IA2019 Feb 7, 2019	PT2019 Feb 7, 2019
No. of requests	99,173,542	143,517,254	308,194,916	1,046,855
	(100.00%)	*(100.00%)*	*(100.00%)*	*(100.00%)*
GET	97,987,295	141,056,534	304,125,661	1,025,132
	(98.80%)	*(98.29%)*	*(98.68%)*	*(97.92%)*
HEAD	1,109,810	2,179,741	2,578,735	14,330
	(1.12%)	*(1.52%)*	*(0.84%)*	*(1.37%)*
PROPFIND	2,092	6,482	27,896	0
	(0.00%)	*(0.00%)*	*(0.01%)*	*(0.00%)*
POST	32,557	265,340	1,368,941	222
	(0.03%)	*(0.18%)*	*(0.44%)*	*(0.02%)*
OPTIONS	1,925	5,631	7,982	0
	(0.00%)	*(0.00%)*	*(0.00%)*	*(0.00%)*
Status Code 2xx	32,460,590	66,584,755	148,742,768	272,467
	(32.73%)	*(46.39%)*	*(48.26%)*	*(26.03%)*
Status Code 3xx	52,131,835	56,778,772	131,729,104	211,709
	(52.57%)	*(39.56%)*	*(42.74%)*	*(20.22%)*
Status Code 4xx	11,614,387	19,701,106	27,099,599	560,913
	(11.71%)	*(13.73%)*	*(8.79%)*	*(53.58%)*
Status code 5xx	2,964,146	451,406	614,502	1,764
	(2.99%)	*(0.31%)*	*(0.20%)*	*(0.17%)*
Embedded resources	43,260,926	91,154,904	195,287,060	205,976
	(43.62%)	*(63.51%)*	*(63.36%)*	*(19.68%)*
Null referrer	47,625,026	111,793,899	60,935,472	265,515
	(48.02%)	*(77.89%)*	*(19.77%)*	*(25.36%)*
SI Robots	8,867	55,163	476,367	3,602
	(0.01%)	*(0.04%)*	*(0.15%)*	*(0.34%)*

Table 2 Number of requests in each of the four datasets (IA2012, IA2015, IA2019, and PT2019): Initial raw data, after stage 1 cleaning, and after stage 2 cleaning

Dataset	Raw dataset	Stage 1 cleaning	Stage 2 cleaning
IA2012	99,173,542	84,512,394 *(85.22%)*	18,432,398 *(18.58%)*
IA2015	143,517,254	125,888,693 *(87.71%)*	27,424,389 *(19.11%)*
IA2019	308,194,916	237,901,926 *(77.19%)*	35,015,776 *(11.36%)*
PT2019	1,046,855	904,515 *(86.40%)*	604,762 *(57.77%)*

ters on known bot User-Agent strings. Our final list [33] of bot User-Agents UA_{K_b} was constructed by combining UA_d with our keyword set UA_m. Any request with a User-Agent found in UA_{K_b} was classified as a robot. This heuristic is an adapted iteration of the User-Agent check heuristic from previous work. AlNoamany et al. considered that if a request's User-Agent matched any of the browsers it was classified as a human request. To ensure the recognition of known bots present within our datasets, we developed UA_{K_b} specifically to retrieve the most current and updated User-Agents. Appendix A provides a real-world example where the "bot" keyword is available on the User-Agent itself.

3.4.2 Type of HTTP request method

Web browsers, which are assumed to be operated by humans, send GET requests for web pages. Therefore, we used HEAD requests as an indicator of robot behavior and integrated this approach as a new heuristic in our work. If the request made is a HEAD request, it is considered a robot request, and the session to which it belongs is counted as a robot session. Appendix B provides a real-world example where HEAD requests are made.

3.4.3 Number of user-agent per IP (UA/IP)

There are robots that repeatedly change their User-Agent (UA) between requests to avoid being detected. The previous study [7] found that a threshold of 20 UAs per IP was effective in distinguishing robots from humans. This allows for some human requests behind a proxy or NAT that may have the same IP address but different User-Agents, representing different users sharing a single IP. As discussed in Sect. 3.3, we sorted the access logs from the three datasets based on IP first and then User-Agent. We marked any requests from IPs that update their User-Agent field more than 20 times as robots. Appendix C provides a real-world example where the IP address is changed for each request.

3.4.4 Requests to robots.txt file

A robots.txt [37, 69] file contains information on how to crawl pages on a website. It helps web crawlers control their actions so that they do not overburden the web server or crawl web pages that are not intended for public viewing. As a result, a request for the robots.txt file can be considered an indication of a robot request. We identified any user who made a request for robots.txt as a robot. Appendix D provides a real-world example where requests are made to the robots.txt file.

3.4.5 Browsing speed (BS)

We used browsing speed as a criterion to distinguish robots from humans. Robots can navigate the web far faster than humans. Castellano et al. [13] found that a human would only make a maximum of one request for a new web page every two seconds. Similar to the previous study [7], we classified any session with a browsing speed faster than one HTML request every two seconds (or, $BS >= 0.5$ requests per second) as a robot. We experimented with an alternate approach involving browsing speed using a three-way criterion (visit duration exceeding 60 s, surpassing a threshold of 10 pages, and a browsing speed threshold of 0.25 pages/s), which was proposed by Tanasa et al. [62] in 2004. However, this approach resulted in a significantly lower detection rate of bots. Therefore, we opted to maintain the threshold set by Castellano et al., which had also been used in previous work by AlNoamany et al. Appendix E provides a real-world example where we can see several requests within a couple of seconds, which is unusual for human behavior.

3.4.6 Image-to-HTML ratio (IH)

Robots tend to retrieve only HTML pages, therefore requests for images can be regarded as a sign of a human user. A ratio of 1:10 images to HTML was proposed by Stassopoulou and Dikaiakos [60] and used in the prior study [7] as a threshold for distinguishing robots from humans. We flagged a session requesting less than one image file for every 10 HTML files as a robot session. IH was found to have the largest effect in detecting robots in the prior study's dataset, and this holds true for our three datasets as well. Appendix F provides a real-world example where a session is marked as a robot using the IH ratio.

We used the aforementioned heuristics on our three datasets to classify each request as human or robot. If a request/session has been marked as a robot at least by one of the heuristics, we have classified it as a robot. After bot identification but before reporting the final results, we performed S2 as described in Sect. 3.2.

4 Results and analysis

In order to investigate the data further after S2 data cleaning, we divided the dataset into two subsets, human sessions, and bot sessions. For each dataset, we used these two subsets to determine user access patterns and compare them to robot access patterns. Finally, we conducted a temporal analysis of the requests in both subsets for each dataset.

4.1 Robots versus humans

Table 3 reports the number of detected robots for each dataset based on the total number of sessions and the total number of requests. We counted the number of requests classified as robots based on each heuristic independently (as mentioned earlier, the heuristics are not mutually exclusive, so these numbers across a column do not need to add to exactly 100%). The final row in the table represents the total number of sessions and requests that are marked as robots after applying all the heuristics together.

The image-to-HTML ratio (IH) had the largest effect on detecting robots across all four datasets. The impact of IH was ≈85-90% in IA2012 and ≈75-80% in IA2015, but only around $\approx 55-65\%$ in IA2019. In PT2019, $\approx 80-96\%$ of robots were detected using the IH ratio, which is higher compared to IA2019. In PT2019, we were able to detect almost all the robots through this one heuristic, IH. We found that $\approx$ 90% of requests were robots in IA2012, $\approx$ 88% of requests were robots in IA2015, $\approx$ 70% of requests were robots in IA2019, and $\approx$ 98% of requests were robots in PT2019.

The reason for this increase in human sessions in 2019 than in 2012 and 2015 could be the increase in awareness of web archives among human users over the years. In addition, headless browsers, such as Headless Chromium [11], PhantomJS [26], and Selenium [56], that provide automated web page control have also become popular in recent years. Their functionality simulates a more human-like behavior that may

Table 3 Bot identification results based on the total number of sessions and the total number of requests for each dataset: IA2012, IA2015, IA2019, and PT2019 (the header for each column displays the total number of sessions and requests). The heuristics are not mutually exclusive

Heuristics	IA2012 Sessions 1,527,340	Requests 22,302,090	IA2015 Sessions 1,355,286	Requests 27,424,389	IA2019 Sessions 2,658,637	Requests 42,868,048	PT2019 Sessions 3,680	Requests 613,672
Known	21,423	398,053	19,441	639,335	322,379	4,969,187	884	67,453
Bots	*(1.40%)*	*(1.78%)*	*(1.43%)*	*(2.33%)*	*(12.13%)*	*(11.59%)*	*(24.02%)*	*(10.99%)*
#UA	5,050	756,801	1,824	683,138	5,475	1,442,574	3	2,636
per IP	*(0.33%)*	*(3.39%)*	*(0.13%)*	*(2.49%)*	*(0.21%)*	*(3.37%)*	*(0.08%)*	*(0.43%)*
robots.txt	1,958	11,074	2,992	11,061	9,296	31,452	404	4,236
	(0.13%)	*(0.05%)*	*(0.22%)*	*(0.04%)*	*(0.35%)*	*(0.07%)*	*(10.98%)*	*(0.69%)*
IH Ratio	1,327,896	19,893,394	1,034,404	22,308,925	1,746,989	24,056,112	2,916	589,363
	(86.94%)	*(89.20%)*	*(76.32%)*	*(81.35%)*	*(65.71%)*	*(56.12%)*	*(79.24%)*	*(96.04%)*
Browsing	237,271	4,563,851	239,120	8,108,851	514,878	21,176,163	1,694	162,068
Speed	*(15.53%)*	*(20.46%)*	*(17.64%)*	*(29.57%)*	*(19.37%)*	*(49.40%)*	*(46.03%)*	*(26.41%)*
Total	**1,340,318**	**20,281,301**	**1,083,830**	**24,132,614**	**1,854,282**	**29,968,059**	**3,584**	**603,654**
Robots	*(87.76%)*	*(90.94%)*	*(79.97%)*	*(87.99%)*	*(69.75%)*	*(69.91%)*	*(97.39%)*	*(98.37%)*

not be caught easily by bot detection techniques. For instance, applications like the work of Ayala [54] and tools like the oldweb.today [38, 39], DSA Toolkit [35, 36], TMVis [41], and Memento-Damage service [58] that replicate human behavior make things challenging for detection algorithms. Between IA2019 and PT2019, PT2019 has $\approx$ 30% more robots present. Based on our PT2019 dataset, only 2% of all requests coming into the Arquivo.pt are potential human requests.

4.2 Discovering access patterns

Upon distinguishing robots from humans, we divided all four of our datasets into human and bot subdatasets (IA2012_H, IA2012_B, IA2015_H, IA2015_B, IA2019_H, IA2019_B, PT2019_H, PT2019_B). We used these datasets to identify different access patterns that are followed by both human and robot sessions. As introduced in Sect. 2, there were four different user access patterns established by AlNoamany et al. [7]. We looked into each of these patterns and identified their prevalence in our three datasets. We discovered the prevalence of sessions that followed each of the four patterns (Dip, Dive, Slide, Skim), as well as sessions that followed a hybrid of those patterns ("Dive and Slide," "Dive and Skim," "Skim and Slide," and "Dive, Slide, and Skim"). We categorized requests that do not fall into any pattern as **Unknown**.

Figure 3 shows a chart for each subdataset. The horizontal (x) axis represents the percentage of the number of requests and the vertical (y) axis represents the different patterns or a hybrid of patterns. The percentages are based on the total number of requests for each subdataset. According to AlNoamany et al.'s findings based on the IA2012 dataset, **Dips** were the most common pattern in both human and robot sessions. However in our IA2012 and IA2015 datasets (full-day), **Dive** and **Dip** account for about the same percentage of human sessions and **Skim** is the most common pattern among robot sessions. **Dip** is the most common pattern in IA2019, followed by **Dive, Slide** for both human and robot sessions. The human **Dips** have doubled from IA2012 (24%) and IA2015 (26%) to IA2019 (51%) indicating that more humans are accessing web archives to access a single URI-M or URI-T in 2019 than the previous years. There are a high number of robot **Skims** in IA2012 and IA2015 compared to IA2019. In IA2012 robot sessions, it is over 90% **Skims** and in IA2015 robot sessions, it is around 80% **Skims**. We could see that the long-running robot sessions that request URI-Ts account for most of the **Skim** percentage. In contrast to IA2019, PT2019 humans exhibit a higher percentage of **Dive** and **Slide** (45%) than **Dips** (29%). Even in robot sessions, **Dive** (70%) and **Dive** and **Slide** (24%) percentage is higher than **Dip** (6%).

The percentage of accesses by humans and robots to TimeMaps and Mementos over the four datasets (IA2012, IA2015, IA2019, and PT2019) is shown in Table 4. In IA2012, robots almost always access TimeMaps (95%) and humans access mementos (82%). This trend continued in IA2015 with robots accessing TimeMaps 83% of the time, while humans accessed mementos 88% of the time. However, in IA2019, humans and robots almost always access mementos (96%), whereas only 4% of those accesses are to TimeMaps. When looking at the hybrid patterns, PT2019 bot sessions only have a maximum of two patterns while the rest have a small percentage of all three patterns (**Dive, Skim, and Slide**). For each dataset in IA, there is a very small percentage of requests (4.22% in IA2012, 3.75% in IA2015, and 0.97% in IA2019) that do not belong to any of the patterns. We were able to identify all the different patterns

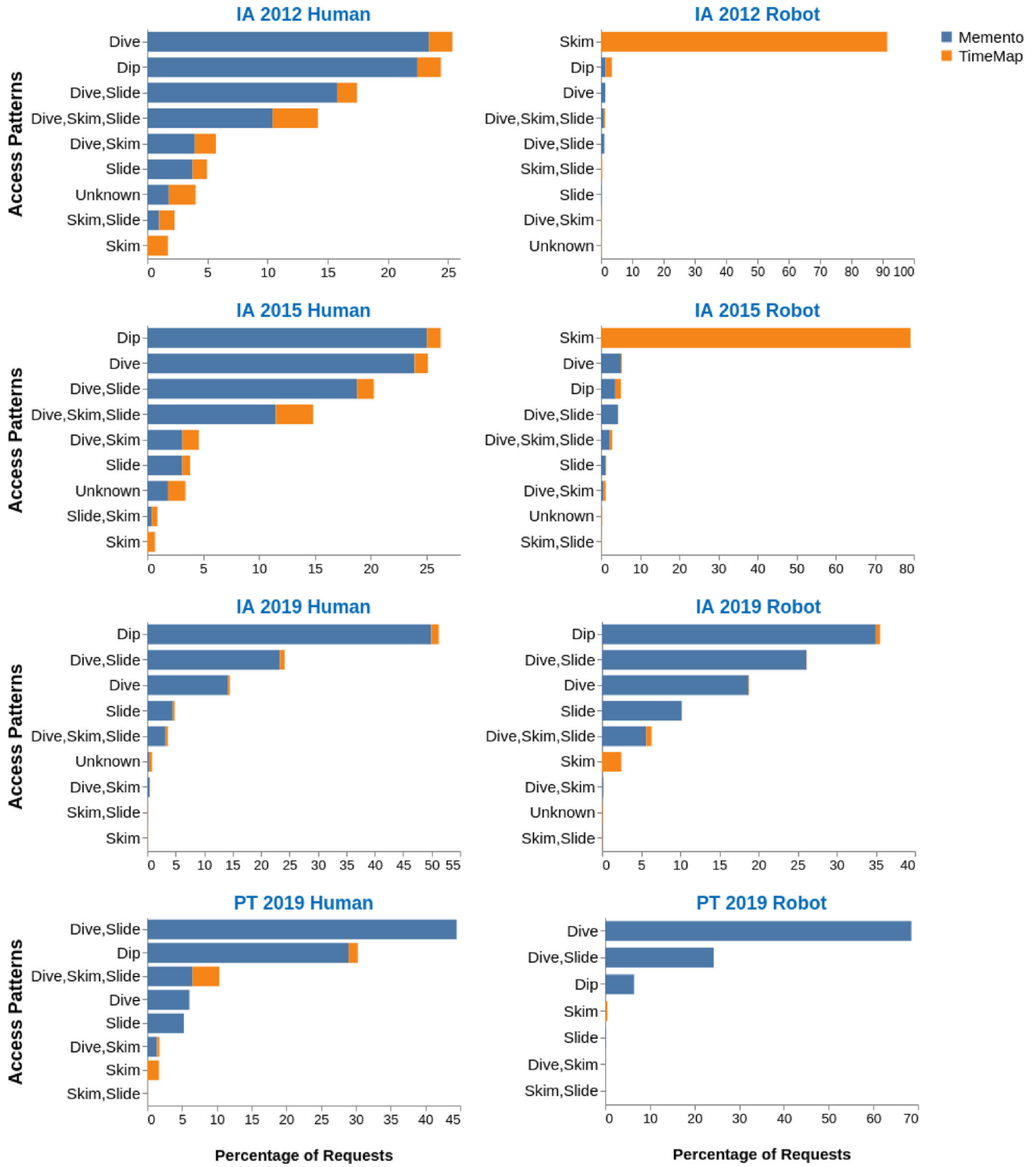

Fig. 3 Access patterns of robots and humans in our subdatasets (IA2012_H, IA2012_B, IA2015_H, IA2015_B, IA2019_H, IA2019_B, PT2019_H, PT2019_B). The color of the stacked bar distinguishes between requests for mementos (URI-Ms) and TimeMaps (URI-Ts). Each chart is sorted in descending order by x-axis value (request percentage). Note that the x-axes in the charts are not the same

in the PT2019 dataset. The percentage of human requests falling under the **Unknown** category in IA2012 (4.02%) and IA2015 (3.42%) is higher compared to the IA2012 robot requests (0.2%), IA2015 robot requests (0.33%), IA2019 human requests (0.85%), and IA2019 robot requests (0.12%).

Table 4 Proportion of robot and human accesses in each of the four datasets (IA2012, IA2015, IA2019, and PT2019) for TimeMaps and Mementos

Dataset	Human		Robots	
	TimeMap (%)	Mementos (%)	TimeMap (%)	Mementos (%)
IA2012	17.54	82.46	94.61	5.39
IA2015	12.26	87.74	82.82	17.18
IA2019	3.89	96.11	4.04	95.96
PT2019	7.28	92.72	0.49	99.51

4.3 Identifying temporal preferences

We also explored the requested Memento-Datetime in our subdatasets to see if there was any temporal preference by web archive users. Figure 4 illustrates the temporal preference of robots and humans in our datasets. The *x*-axis represents the number of years prior, meaning the number of years passed relative to the datetime of the access logs (e.g., for IA2012, 2 years prior is 2010) and the *y*-axis represents the number of requests. Note that the *y*-axis in each chart is different.

It is evident that the majority of the requests are for mementos that are close to the datetime of each access log sample and gradually diminish as we go further back in time. There is no significant difference in temporal preference in IA2012, IA2015, and IA2019. IA2019 humans, IA2019 bots, and PT2019 bots exhibit the same trend; however, it is difficult to see a trend in PT2019 humans due to the fewer number of humans in the dataset. For PT2019 humans, there is a spike around 4–5 years prior which implies PT human accesses were mostly for mementos around 2015–2016. There is an advantage to knowing the temporal preferences of web archive users. Web archives can prioritize or store data in memory for the most recent years to speed up disk access.

5 Future work

AlNoamany et al. [7] observed four different user access patterns in 2013. In our datasets combined, 0.48% of requests were outside of any of these patterns or their combinations. One may look into if the percentage of requests that fell into the **Unknown** category have any other generally applicable patterns, or if they are completely random. The overall number of robots identified in IA2019 is much lower than in IA2012 and IA2015. We would like to repeat this study on more distinct full-day datasets to see if the reduction in robots is a general behavior from 2012 and 2015 to 2019 or specific to the day we chose. Additionally, the IH [60] and BS thresholds [13] in our bot identification heuristics are based on the behavior of conventional web servers; however, it remains to be determined if the same thresholds apply to web archival replay systems, as the dynamics of web archival replay systems differ (e.g., the Wayback Machine is typically slower than a typical web server).

6 Conclusions

We used a full-day access logs sample of Internet Archive's (IA) Wayback Machine from 2012, 2015 and 2019, as well as Arquivo.pt's from 2019, to distinguish between robot and human users in web archives. The total number of robots request detected for IA2012 (90.94%) and IA2015 (87.99%) datasets is higher than the overall number of robots discovered in IA2019 (69.91%). We discovered that robot accesses account for 98% of requests (97% of sessions) based on 2019 server logs from Arquivo.pt. We also discovered that in IA2012 and IA2015, the most common pattern for robots were almost exclusively Skim, but that in IA2019, they exhibit all of the patterns and their combinations. Regardless of whether it is a robot or a human user, the majority of requests were for mementos that are close to the datetime of each access log dataset, demonstrating a preference for the recent past. In summary, these insights into users' behaviors and temporal preferences can be leveraged to improve the efficiency of web archives by tailoring resource allocation accordingly. We believe that this will further strengthen web archives, enhancing accessibility, and preserving invaluable historical web content for diverse purposes.

Acknowledgements We thank Mark Graham, director of the Wayback Machine, for sharing access log data from the Wayback Machine at the Internet Archive. We are grateful to Daniel Gomes and Fernando Melo of Arquivo.pt for sharing access log data from the Arquivo.pt web archive with us.

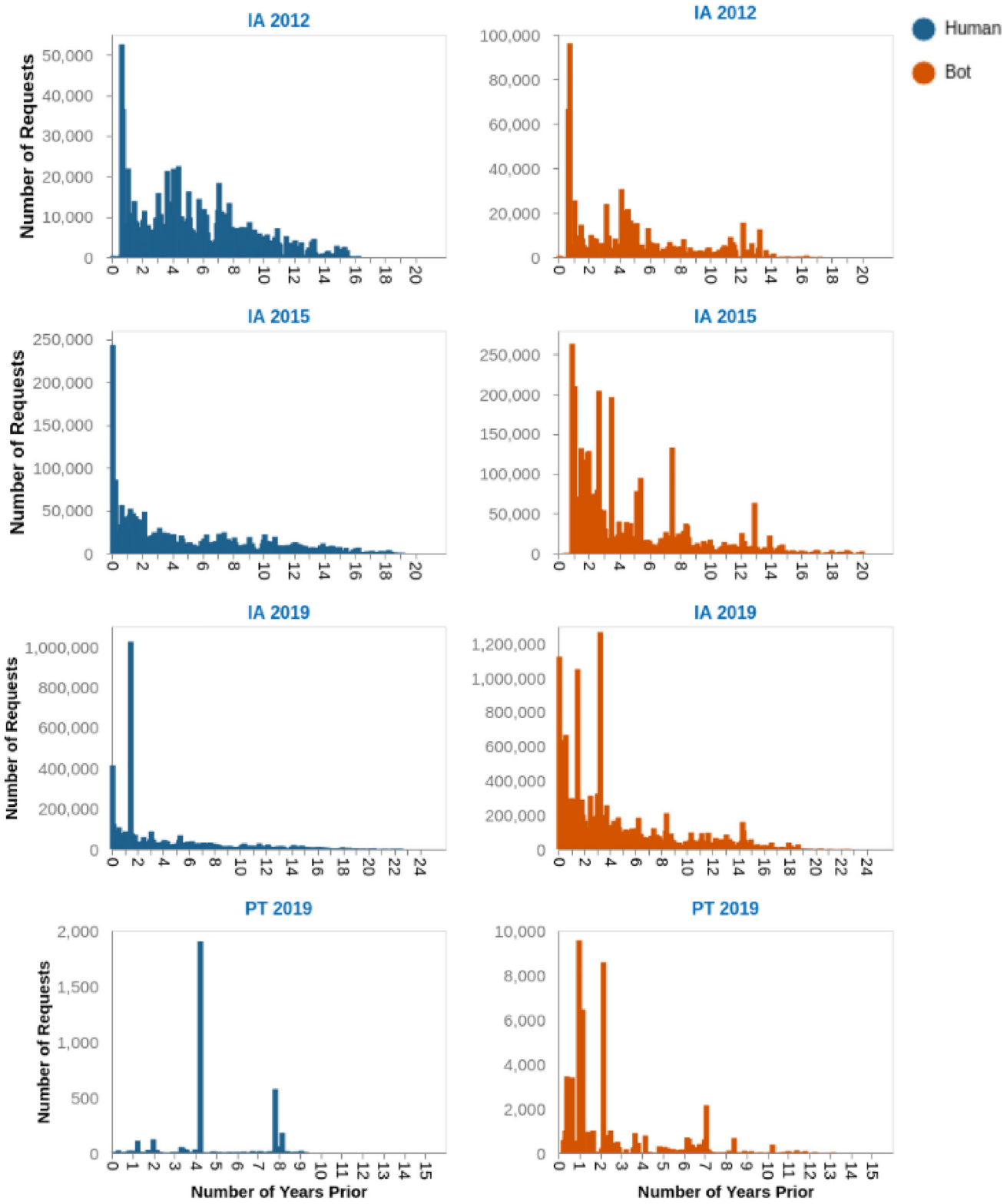

Fig. 4 Temporal preference of bots and humans in IA2012, IA2019, PT2019 datasets

Appendix A: Known bots

This heuristic makes use of a list of User-Agents that are recognized as being used by bots. We first constructed a list of all User-Agent strings from our three datasets. From this list, we compiled a list by filtering for User-Agent strings that contained robot keywords, such as "bot," "crawler," and "spider." Our GitHub repository hosts the comprehensive list of known bots [33] that was created. Below is an example where "Twitterbot/1.0" is the User-Agent. Section 3.4.1 discusses this heuristic in more detail.

```
199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:01 +0100] "GET /robots.
    txt HTTP/1.1" 200 1414 "-" "
    Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:01 +0100] "GET /robots.
    txt HTTP/1.1" 200 1414 "-" "
    Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:02 +0100] "HEAD /wayback
    /20170625001353/http://www.
    fabricadochocolate.com HTTP/1.1" 200
     - "-" "Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:02 +0100] "HEAD /wayback
    /20170625001353/http://www.
    fabricadochocolate.com HTTP/1.1" 200
     - "-" "Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:05 +0100] "HEAD /wayback
    /20170625001353/http://www.
    fabricadochocolate.com/ HTTP/1.1"
    200 - "-" "Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:05 +0100] "HEAD /wayback
    /20170625001353/http://www.
    fabricadochocolate.com/ HTTP/1.1"
    200 - "-" "Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:07 +0100] "HEAD /wayback
    /20170625001353/http://www.
    fabricadochocolate.com/ HTTP/1.1"
    200 - "-" "Twitterbot/1.0"

199.16.157.100_0_0 - - [07/Jul
    /2019:14:00:07 +0100] "HEAD /wayback
    /20170625001353/http://www.
    fabricadochocolate.com/ HTTP/1.1"
    200 - "-" "Twitterbot/1.0"
```

Appendix B: Type of HTTP request method

We used HEAD requests as an indication of robot behavior. If the request made is a HEAD request, it is considered a robot request, and the session to which it belongs is counted as a robot session. Below is an example where HTTP HEAD requests are made to different mementos. The User-Agent is "Twitterbot" in these request logs, which is another indication that they are robot requests. Section 3.4.2 discusses this heuristic in more detail.

```
0.77.87.100 - - [02/Feb/2012:03:46:54
    +0000] "POST http://web.archive.org/
    web/20070211155651/http
    ://212.227.83.57/cproc.aspx HTTP
    /1.0" 302 0 "http://www.vbleisure.co
    .uk/guest\_book.html" "Mozilla/4.0 (
    compatible; MSIE 5.5; Windows NT
    4.0)"

0.77.87.100 - - [02/Feb/2012:04:06:29
    +0000] "POST http://web.archive.org/
    web/20070211155651/http
    ://212.227.83.57/cproc.aspx HTTP
    /1.0" 302 - "http://www.vbleisure.co
    .uk/guest\_book.html" "Mozilla/4.0 (
    compatible; MSIE 6.0; Windows NT
    5.1; SV1; .NET CLR 1.1.4322)"

0.77.87.100 - - [02/Feb/2012:05:09:30
    +0000] "POST http://web.archive.org/
    web/20070211155651/http
    ://212.227.83.57/cproc.aspx HTTP
    /1.0" 302 - "http://www.vbleisure.co
    .uk/guest\_book.html" "Mozilla/4.0 (
    compatible; MSIE 6.0; Windows NT
    5.0)"

0.77.87.100 - - [02/Feb/2012:07:59:43
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\\_stub/
    formmailer.php HTTP/1.0" 302 0 "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Mozilla/4.0 (compatible; MSIE
```

```
    6.0; Windows NT 5.1; ru) Opera 8.50
    "

. . .

. . .

0.77.87.100 - - [02/Feb/2012:22:08:02
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\_stub/
    formmailer.php HTTP/1.0" 503 - "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Mozilla/4.0 (compatible; MSIE
     6.0; Windows NT 5.1; .NET CLR
    1.0.3705)"

0.77.87.100 - - [02/Feb/2012:23:40:31
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\_stub/
    formmailer.php HTTP/1.0" 503 - "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Mozilla/4.0 (compatible; MSIE
     6.0; Windows NT 5.1; en) Opera 9.0"

0.77.87.100 - - [02/Feb/2012:23:40:32
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\_stub/
    formmailer.php HTTP/1.0" 503 - "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Mozilla/4.0 (compatible; MSIE
     6.0; Windows NT 5.1; MRA 4.6 (build
     01425))"

0.77.87.100 - - [02/Feb/2012:23:59:34
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\_stub/
    formmailer.php HTTP/1.0" 503 - "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Opera/7.60 (Windows NT 5.2; U
    ) [en] (IBM EVV/3.0/EAK01AG9/LE)"
```

Appendix C: Number of user-agents per IP (UA/IP)

There are robots that repeatedly change their User-Agent (UA) between requests to avoid being detected. We marked any requests from IPs that update their User-Agent field more than 20 times as robots. Below is an example where the IP address is changed for each request. Section 3.4.3 discusses this heuristic in more detail.

```
0.77.87.100 - - [02/Feb/2012:03:46:54
    +0000] "POST http://web.archive.org/
    web/20070211155651/http
    ://212.227.83.57/cproc.aspx HTTP
    /1.0" 302 0 "http://www.vbleisure.co
    .uk/guest\_book.html" "Mozilla/4.0 (
    compatible; MSIE 5.5; Windows NT
    4.0)"

0.77.87.100 - - [02/Feb/2012:04:06:29
    +0000] "POST http://web.archive.org/
    web/20070211155651/http
    ://212.227.83.57/cproc.aspx HTTP
    /1.0" 302 - "http://www.vbleisure.co
    .uk/guest\_book.html" "Mozilla/4.0 (
    compatible; MSIE 6.0; Windows NT
    5.1; SV1; .NET CLR 1.1.4322)"

0.77.87.100 - - [02/Feb/2012:05:09:30
    +0000] "POST http://web.archive.org/
    web/20070211155651/http
    ://212.227.83.57/cproc.aspx HTTP
    /1.0" 302 - "http://www.vbleisure.co
    .uk/guest\_book.html" "Mozilla/4.0 (
    compatible; MSIE 6.0; Windows NT
    5.0)"

0.77.87.100 - - [02/Feb/2012:07:59:43
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\\_stub/
    formmailer.php HTTP/1.0" 302 0 "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Mozilla/4.0 (compatible; MSIE
     6.0; Windows NT 5.1; ru) Opera 8.50
    "

. . .

. . .

0.77.87.100 - - [02/Feb/2012:22:08:02
    +0000] "POST http://web.archive.org/
    web/20070501120942/http://www.
    ibcmemorial.org.way\_back\_stub/
    formmailer.php HTTP/1.0" 503 - "http
    ://ibcmemorial.org/sign-guestbook.
    html" "Mozilla/4.0 (compatible; MSIE
```

```
    6.0; Windows NT 5.1; .NET CLR
   1.0.3705)"

0.77.87.100 - - [02/Feb/2012:23:40:31
   +0000] "POST http://web.archive.org/
   web/20070501120942/http://www.
   ibcmemorial.org.way\_back\_stub/
   formmailer.php HTTP/1.0" 503 - "http
   ://ibcmemorial.org/sign-guestbook.
   html" "Mozilla/4.0 (compatible; MSIE
    6.0; Windows NT 5.1; en) Opera 9.0"

0.77.87.100 - - [02/Feb/2012:23:40:32
   +0000] "POST http://web.archive.org/
   web/20070501120942/http://www.
   ibcmemorial.org.way\_back\_stub/
   formmailer.php HTTP/1.0" 503 - "http
   ://ibcmemorial.org/sign-guestbook.
   html" "Mozilla/4.0 (compatible; MSIE
    6.0; Windows NT 5.1; MRA 4.6 (build
    01425))"

0.77.87.100 - - [02/Feb/2012:23:59:34
   +0000] "POST http://web.archive.org/
   web/20070501120942/http://www.
   ibcmemorial.org.way\_back\_stub/
   formmailer.php HTTP/1.0" 503 - "http
   ://ibcmemorial.org/sign-guestbook.
   html" "Opera/7.60 (Windows NT 5.2; U
   ) [en] (IBM EVV/3.0/EAK01AG9/LE)"
```

Appendix D: Requests to robots.txt

A robots.txt file contains information on how to crawl pages on a website. As a result, a request for the robots.txt file can be considered an indication of a robot request. The requests made to the web archives' robots.txt file are demonstrated in the examples that follow. Section 3.4.4 discusses this heuristic in more detail.

```
0.139.100.213_2_2 - - [02/Feb
   /2012:17:03:22 +0000] "GET http://
   web.archive.org/robots.txt HTTP/1.1"
    200 125 "-" "RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:06:30 +0000] "GET http://
   web.archive.org/web/*/http://
   c00lbookmarks.com/story.php?title=
   best-door-blinds-inside HTTP/1.1"
   302 0 "-" "RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:06:32 +0000] "GET http://
   wayback.archive.org/web/*/http://
   c00lbookmarks.com/story.php?title=
   best-door-blinds-inside HTTP/1.1"
   404 2409 "http://web.archive.org/web
   /*/http://c00lbookmarks.com/story.
   php?title=best-door-blinds-inside" "
   RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:07:38 +0000] "GET http://
   web.archive.org/robots.txt HTTP/1.1"
    200 125 "-" "RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:10:44 +0000] "GET http://
   web.archive.org/web/*/http://www.
   goloco.org/users/D5EWwXI HTTP/1.1"
   302 0 "-" "RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:10:45 +0000] "GET http://
   wayback.archive.org/web/*/http://www
   .goloco.org/users/D5EWwXI HTTP/1.1"
   404 2385 "http://web.archive.org/web
   /*/http://www.goloco.org/users/
   D5EWwXI" "RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:14:50 +0000] "GET http://
   web.archive.org/robots.txt HTTP/1.1"
    200 125 "-" "RSS Scout 0.9.2"

0.139.100.213_2_2 - - [02/Feb
   /2012:17:19:54 +0000] "GET http://
   web.archive.org/robots.txt HTTP/1.1"
    200 125 "-" "RSS Scout 0.9.2"
```

Appendix E: Browsing speed (BS)

We used browsing speed as a criterion to distinguish robots from humans. Robots can navigate the web far faster than humans. It was found that a human would only make a maximum of one request for a new web page every 2 s. We classified any session with a browsing speed faster than one HTML request every two seconds (or, $BS >= 0.5$ requests per second) as a robot. The example below demonstrates how a single IP made several requests within couple of seconds (Browsing. Section 3.4.5 discusses this heuristic in more detail.

```
0.0.115.10_0_0 web.archive.org - [07/
    Feb/2019:04:41:46 +0000] "GET /web
    /20070524115946/http://www.moviehole
    .net/interviews/20070521
    _exclusive_interview_jerry_bruc.html
     HTTP/2.0" 200 11994 "-" "Mozilla
    /5.0 (Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/71.0.3578.98 Safari
    /537.36" 0.000 HIT - "text/html;
    charset=iso-8859-1" - "-"

0.0.115.10_0_0 web.archive.org - [07/
    Feb/2019:04:41:46 +0000] "GET /web
    /20070524115946/http://www.moviehole
    .net/interviews/20070521
    _exclusive_interview_jerry_bruc.html
     HTTP/2.0" 200 11994 "-" "Mozilla
    /5.0 (Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/71.0.3578.98 Safari
    /537.36" 0.948 MISS 0.948 "text/html
    ; charset=iso-8859-1" - "-"

0.0.115.10_0_0 web.archive.org - [07/
    Feb/2019:04:41:47 +0000] "GET /web
    /20120118050811/http://uk.movies.ign
    .com/articles/455/455825p1.html HTTP
    /2.0" 200 25303 "-" "Mozilla/5.0 (
    Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/71.0.3578.98 Safari
    /537.36" 0.000 HIT - "text/html;
    charset=UTF-8" - "-"

0.0.115.10_0_0 web.archive.org - [07/
    Feb/2019:04:41:47 +0000] "GET /web
    /20120118050811/http://uk.movies.ign
    .com/articles/455/455825p1.html HTTP
    /2.0" 200 25303 "-" "Mozilla/5.0 (
    Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/71.0.3578.98 Safari
    /537.36" 1.439 MISS 1.440 "text/html
    ;charset=UTF-8" - "-"

0.0.115.10_0_0 web.archive.org - [07/
    Feb/2019:04:41:47 +0000] "GET /web
    /20120714014937/http://uk.movies.ign
    .com/articles/425/425848p1.html HTTP
    /2.0" 200 18620 "-" "Mozilla/5.0 (
    Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/71.0.3578.98 Safari
    /537.36" 0.000 HIT - "text/html;
    charset=UTF-8" - "-"

0.0.115.10_0_0 web.archive.org - [07/
    Feb/2019:04:41:47 +0000] "GET /web
    /20120714014937/http://uk.movies.ign
    .com/articles/425/425848p1.html HTTP
    /2.0" 200 18620 "-" "Mozilla/5.0 (
    Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/71.0.3578.98 Safari
    /537.36" 1.584 MISS 1.584 "text/html
    ;charset=UTF-8" - "-"
```

Appendix F: Image-to-HTML ratio (IH)

Robots tend to retrieve only HTML pages, therefore requests for images can be regarded as a sign of a human user. We flagged a session requesting less than ne image file for every 10 HTML files as a robot session. The below is an example where only requests for HTML files are made without any images or other embedded resources. Section 3.4.6 discusses this heuristic in more detail.

```
0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:55:22 +0000] "GET /web
    /*/http://maestro.haarp.alaska.edu/
    HTTP/2.0" 200 9002 "https://archive.
    org/search.php?query=http

0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:56:15 +0000] "GET /web
    /20130304102141/http://maestro.haarp
    .alaska.edu/ HTTP/2.0" 404 0 "https
    ://web.archive.org/web
    /20130715000000*/http://maestro.
    haarp.alaska.edu/"

0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:56:15 +0000] "GET /web
    /20130304102141/http://maestro.haarp
    .alaska.edu/ HTTP/2.0" 404 0 "https
    ://web.archive.org/web
    /20130715000000*/http://maestro.
    haarp.alaska.edu/"

0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:56:15 +0000] "GET /web
    /20130304102141/http://maestro.haarp
    .alaska.edu/ HTTP/2.0" 404 0 "https
```

```
    ://web.archive.org/web
    /20130715000000*/http://maestro.
    haarp.alaska.edu/"

. . .

. . .

. . .

0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:56:23 +0000] "GET /web
    /20130304102141/http://maestro.haarp
    .alaska.edu/ HTTP/2.0" 404 8274 "
    https://web.archive.org/web
    /20130715000000*/http://maestro.
    haarp.alaska.edu/"

0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:56:23 +0000] "GET /web
    /20130304102141/http://maestro.haarp
    .alaska.edu/ HTTP/2.0" 404 8274 "
    https://web.archive.org/web
    /20130715000000*/http://maestro.
    haarp.alaska.edu/"

0.0.122.100_1_0 web.archive.org - [07/
    Feb/2019:16:56:29 +0000] "GET /web
    /*/http://maestro.haarp.alaska.edu/*
     HTTP/2.0" 200 8341 "https://web.
    archive.org/web/20130304102141/http
    ://maestro.haarp.alaska.edu/"
```

References

1. Alam, S.: AccessLog Parser and CLI. https://github.com/oduwsdl/accesslog-parser (2019)
2. Alam, S., Cartledge, C.L., Nelson, M.L.: Support for Various HTTP Methods on the Web. Tech. Rep. arXiv:1405.2330, Old Dominion University (2014), arxiv:1405.2330
3. Alam, S., Nelson, M.L.: MemGator - a portable concurrent Memento aggregator: Cross-platform CLI and server binaries in Go. In: JCDL '16: Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 243–244 (2016), https://doi.org/10.1145/2910896.2925452
4. Alam, S., Weigle, M.C., Nelson, M.L.: Profiling Web Archival Voids for Memento Routing. In: Proceedings of the 21st ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 150–159. JCDL '21 (2021), https://doi.org/10.1109/JCDL52503.2021.00027
5. AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L.: Who and what links to the Internet Archive. In: Proceedings of Theory and Practice of Digital Libraries (TPDL). pp. 346–357 (2013), https://doi.org/10.1007/978-3-642-40501-3_35
6. AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L.: Who and what links to the Internet Archive. Int. J. Digit. Libr. **14**(3), 101–115 (2014). https://doi.org/10.1007/s00799-014-0111-5
7. AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Access patterns for robots and humans in web archives. In: JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 339–348 (2013). https://doi.org/10.1145/2467696.2467722
8. Apache HTTP Server: Common Log Format and Combined Log Format. https://httpd.apache.org/docs/trunk/logs.html (2013)
9. Banos, V., Manolopoulos, Y.: A quantitative approach to evaluate Website Archivability using the CLEAR+ method. Int. J. Digit. Libr. **17**(2), 119–141 (2016). https://doi.org/10.1007/s00799-015-0144-4
10. Berendt, B., Mobasher, B., Spiliopoulou, M., Wiltshire, J.: Measuring the accuracy of sessionizers for web usage analysis. In: Workshop on Web Mining at the First SIAM International Conference on Data Mining. pp. 7–14. SIAM Philadelphia PA (2001), http://facweb.cs.depaul.edu/research/TechReports/TR01-006.pdf
11. Bidelman, E.: Getting Started with Headless Chrome. https://developer.chrome.com/blog/headless-chrome/ (2018)
12. Burkholder, D.: DeviceDetector. https://github.com/thinkwelltwd/device_detector (2022)
13. Castellano, G., Fanelli, A.M., Torsello, M.A.: LODAP: A LOg DAta Preprocessor for mining web browsing patterns. In: Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases. pp. 12–17 (Feb 2007), https://dl.acm.org/doi/10.5555/1348485.1348488
14. Castellano, G., Mesto, F., Minunno, M., Torsello, M.A.: Web user profiling using fuzzy clustering. In: International Workshop on Fuzzy Logic and Applications. pp. 94–101. Springer (2007), https://doi.org/10.1007/978-3-540-73400-0_12
15. Costa, M., Gomes, D., Couto, F.M., Silva, M.J.: A Survey of Web Archive Search Architectures. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 1045–1050. TempWeb '13 (2013), https://doi.org/10.1145/2487788.2488116
16. Costa, M., Miranda, J., Cruz, D., Gomes, D.: Query suggestion for web archive search. In: International Conference on Digital Preservation (2013), https://sobre.arquivo.pt/wp-content/uploads/query-suggestion-for-web-archive-search-1.pdf
17. Costa, M., Silva, M.J.: Characterizing search behavior in web archives. In: International Temporal Web Analytics Workshop (2011), https://sobre.arquivo.pt/wp-content/uploads/characterizing-search-behavior-in-web-archives.pdf
18. Dash, S., Luhach, A.K., Chilamkurti, N., Baek, S., Nam, Y., et al.: A neuro-fuzzy approach for user behaviour classification and prediction. J. Cloud Comput. **8**(1), 1–15 (2019). https://doi.org/10.1186/s13677-019-0144-9
19. Deschamps, R., Fritz, S., Lin, J., Milligan, I., Ruest, N.i.: The cost of a WARC: Analyzing web archives in the cloud. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 261–264. IEEE (2019)
20. Dumais, S., Jeffries, R., Russell, D.M., Tang, D., Teevan, J.: Understanding user behavior through log data and analysis. Ways of knowing in HCI pp. 349–372 (2014), https://doi.org/10.1007/978-1-4939-0378-8_14
21. Fielding, R.T., Nottingham, M., Reschke, J.F.: HTTP Semantics - RFC 9110. https://tools.ietf.org/html/rfc9110 (2022)
22. Gomes, D., Costa, M.: The importance of web archives for humanities. Int. J. Human. Arts Comput. **8**(1), 106–123 (2014). https://doi.org/10.3366/ijhac.2014.0122
23. Gomes, D., Costa, M., Cruz, D., Miranda, J., Fontes, S.: Creating a billion-scale searchable web archive. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 1059–1066 (2013), https://doi.org/10.1145/2487788.2488118
24. Gomes, D., Cruz, D., Miranda, J., Costa, M., Fontes, S.: Search the past with the portuguese web archive. In: Proceedings of the

22nd International Conference on World Wide Web. pp. 321–324 (2013), https://doi.org/10.1145/2487788.2487934
25. Grcar, M.: User profiling: web usage mining. In: Proceedings of the 7th International Multiconference Information Society IS (2004)
26. Hidayat, A.: PhantomJS. https://phantomjs.org/ (2011)
27. Hoang, X.D.: Detecting common web attacks based on machine learning using web log. In: Advances in Engineering Research and Application: Proceedings of the International Conference on Engineering Research and Applications, ICERA 2020. pp. 311–318. Springer (2021), https://doi.org/10.1007/978-3-030-64719-3_35
28. Hockx-Yu, H.: Access and scholarly use of web archives. Alexandria **25**(1–2), 113–127 (2014). https://doi.org/10.7227/alx.0023
29. Hosseini, N., Fakhar, F., Kiani, B., Eslami, S.: Enhancing the security of patients' portals and websites by detecting malicious web crawlers using machine learning techniques. Int. J. Med. Inf. **132**, 103976 (2019). https://doi.org/10.1016/j.ijmedinf.2019.103976
30. Internet Archive: Wayback machine homepage. https://web.archive.org/web/20220527205606/https://web.archive.org/ (2022)
31. Jayanetti, H., Garg, K.: Access patterns. https://github.com/oduwsdl/access-patterns/ (2022)
32. Jayanetti, H.R.: Visualizations for web archive access log datasets. https://observablehq.com/@himarshaj/visualizations-for-web-archive-access-log-datasets (2022)
33. Jayanetti, H.R., Garg, K.: Known bot list. https://github.com/oduwsdl/access-patterns/tree/main/Known_Bot_List (2022)
34. Jayanetti, H.R., Garg, K., Alam, S., Nelson, M.L., Weigle, M.C.: Robots still outnumber humans in web archives, but less than before. In: Proceedings of the Theory and Practice of Digital Libraries Conference (TPDL) (Sep 2022), https://doi.org/10.1007/978-3-031-16802-4_19
35. Jones, S.M.: Improving Collection Understanding for Web Archives with Storytelling: Shining Light Into Dark and Stormy Archives. Ph.D. thesis, Old Dominion University (2021), https://doi.org/10.25777/zts6-v512
36. Jones, S.M., Jayanetti, H.R., Osborne, A., Koerbin, P., Klein, M., Weigle, M.C., Nelson, M.L.: The DSA Toolkit Shines Light Into Dark and Stormy Archives. Code4Lib Journal (2022), https://journal.code4lib.org/articles/16441
37. Koster, M., Illyes, G., Zeller, H., Sassman, L.: Robots exclusion protocol- RFC 9309. https://www.rfc-editor.org/rfc/rfc9309 (2022). https://doi.org/10.17487/RFC9309
38. Kreymer, I., Rosenthal, D.S.H.: Guest Post: Ilya Kreymer on oldweb.today. https://blog.dshr.org/2016/01/guest-post-ilya-kreymer-on-oldwebtoday.html (2016)
39. Kreymer, I., Rosenthal, D.S.H.: Announcing the new OldWeb.today. https://webrecorder.net/2020/12/23/new-oldweb-today.html (2020)
40. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, vol. 1. Springer, Cham (2011)
41. Mabe, A., Patel, D., Gunnam, M., Shankar, S., Kelly, M., Alam, S., Nelson, M.L., Weigle, M.C.: Visualizing Webpage Changes Over Time. Technical Report Old Dominion University (Jun 2020), arxiv:2006.02487
42. Maemura, E.: All WARC and no playback: the materialities of data-centered web archives research. Big Data Soc. **10**(1), 20539517231163172 (2023)
43. Meneses, L., Furuta, R., Shipman, F.: Identifying 'Soft 404' Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections. In: Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries. TPDL '12, vol. 7489, pp. 197–208. Springer (2012), https://doi.org/10.1007/978-3-642-33290-6_22
44. Mobasher, B.: Web usage mining. In: Encyclopedia of Data Warehousing and Mining, pp. 1216–1220. IGI Globa (2005)
45. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on web usage mining. Commun. ACM **43**(8), 142–151 (2000). https://doi.org/10.1145/345124.345169
46. Mobasher, B., Dai, H., Luo, T., Sun, Y., Zhu, J.: Integrating web usage and content mining for more effective personalization. In: International conference on electronic commerce and web technologies. pp. 165–176. Springer (2000), https://doi.org/10.1007/3-540-44463-7_15
47. Mughal, M.J.H.: Data mining: Web data mining techniques, tools and algorithms: An overview. Int. J. Adv. Comput. Sci. Appl. (2018). https://doi.org/10.14569/ijacsa.2018.090630
48. Mumma, C., Phillips, M.: Systems interoperability and collaborative development for web archiving-filling gaps in the IMLS National Digital Platform. http://hdl.handle.net/2249.1/76270 (2016)
49. Nelson, M.L., Van de Sompel, H.: Adding the dimension of time to HTTP. In: SAGE Handbook of Web History, pp. 191–214. SAGE Publishing (2019)
50. Newbold, B.: Search scholarly materials preserved in the internet archive. https://blog.archive.org/2021/03/09/search-scholarly-materials-preserved-in-the-internet-archive/ (2021)
51. Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D.: Web usage mining as a tool for personalization: a survey. User Model. User-Adap. Inter. **13**(4), 311–372 (2003). https://doi.org/10.1023/A:1026238916441
52. Poggi, N., Muthusamy, V., Carrera, D., Khalaf, R.: Business process mining from E-commerce web logs. In: Business Process Management: 11th International Conference, BPM 2013, Beijing, China, August 26-30, 2013. Proceedings. pp. 65–80. Springer (2013), https://doi.org/10.1007/978-3-642-40176-3_7
53. Qbea'h, M., Alrabaee, S., Alshraideh, M., Sabri, K.E.: Diverse approaches have been presented to mitigate sql injection attack, but it is still alive: a review. In: 2022 International Conference on Computer and Applications (ICCA). pp. 1–5 (2022), https://doi.org/10.1109/ICCA56443.2022.10039611
54. Reyes Ayala, B.: Correspondence as the primary measure of information quality for web archives: a human-centered grounded theory study. Int. J. Digit. Libr. **23**(1), 19–31 (2022). https://doi.org/10.1007/s00799-021-00314-x
55. Ruest, N., Fritz, S., Deschamps, R., Lin, J., Milligan, I.: From archive to analysis: accessing web archives at scale through a cloud-based interface. Int. J. Dig. Humanit. **2**(1), 5–24 (2021)
56. Selenium: selenium client driver. https://selenium.dev/selenium/docs/api/py/ (2018)
57. Shaheed, A., Kurdy, M.: Web application firewall using machine learning and features engineering. Secur. Commun. Netw. (2022). https://doi.org/10.1155/2022/5280158
58. Siregar, E.: Deploying the memento-damage service. https://ws-dl.blogspot.com/2017/11/2017-11-22-deploying-memento-damage.html (2017)
59. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD Explor. Newsl. **1**(2), 12–23 (2000). https://doi.org/10.1145/846183.846188
60. Stassopoulou, A., Dikaiakos, M.D.: Web robot detection: A probabilistic reasoning approach. Comput. Netw. **53**(3), 265–278 (2009). https://doi.org/10.1016/j.comnet.2008.09.021
61. Suneetha, K., Krishnamoorthi, R.: Identifying user behavior by analyzing web server access log file. IJCSNS Int. J. Comput. Sci. Netw. Secur. **9**(4), 327–332 (2009)
62. Tanasa, D., Trousse, B.: Advanced data preprocessing for intersites web usage mining. IEEE Intell. Syst. **19**, 59–65 (2004). https://doi.org/10.1109/MIS.2004.1274912
63. Van de Sompel, H., Nelson, M.L., Sanderson, R.: HTTP Framework for Time-Based Access to Resource States - Memento - RFC 7089. http://tools.ietf.org/html/rfc7089 (2013)

64. Varnagar, C.R., Madhak, N.N., Kodinariya, T.M., Rathod, J.N.: Web usage mining: a review on process, methods and techniques. In: 2013 International Conference on Information Communication and Embedded Systems (ICICES). pp. 40–46. IEEE (2013), https://doi.org/10.1109/icices.2013.6508399
65. Wu, F., Qiao, Y., Chen, J.H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., Zhou, M.: MIND: A large-scale dataset for news recommendation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3597–3606. Association for Computational Linguistics (2020), https://doi.org/10.18653/v1/2020.acl-main.331
66. Wu, Y., Sun, Y., Huang, C., Jia, P., Liu, L.: Session-based webshell detection using machine learning in web logs. Secur. Communi. Netw. **2019**, 1–11 (2019). https://doi.org/10.1155/2019/3093809
67. Wu, Z., Sanderson, M., Cambazoglu, B.B., Croft, W.B., Scholer, F.: Providing direct answers in search results: A study of user behavior. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. p. 1635-1644. CIKM '20, Association for Computing Machinery (2020), https://doi.org/10.1145/3340531.3412017
68. Zaiane, O.: Web usage mining for a better web-based learning environment. Technical Report TR01-05, University of Alberta (2001), https://doi.org/10.7939/R3736M20P
69. Zeller, H., Harvey, L., Illyes, G.: Formalizing the robots exclusion protocol specification. https://webmasters.googleblog.com/2019/07/rep-id.html (2019)