Int J Digit Libr (2014) 14:101–115
DOI 10.1007/s00799-014-0111-5

Who and what links to the Internet Archive

Yasmin AlNoamany · Ahmed AlSum · Michele C. Weigle · Michael L. Nelson

Received: 30 October 2013 / Revised: 16 March 2014 / Accepted: 19 March 2014 / Published online: 23 April 2014

Abstract The Internet Archive's (IA) Wayback Machine is the largest and oldest public Web archive and has become a significant repository of our recent history and cultural heritage. Despite its importance, there has been little research about how it is discovered and used. Based on Web access logs, we analyze what users are looking for, why they come to IA, where they come from, and how pages link to IA. We find that users request English pages the most, followed by the European languages. Most human users come to Web archives because they do not find the requested pages on the live Web. About 65 % of the requested archived pages no longer exist on the live Web. We find that more than 82 % of human sessions connect to the Wayback Machine via referrals from other Web sites, while only 15 % of robots have referrers. Most of the links (86 %) from Websites are to individual archived pages at specific points in time, and of those 83 % no longer exist on the live Web. Finally, we find that users who come from search engines browse more pages than users who come from external Web sites.

Keywords Web archiving · Web server logs · Web usage mining · Robots detection · Language detection

Y. AlNoamany (✉) · A. AlSum · M. C. Weigle · M. L. Nelson
Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA
e-mail: yasmin@cs.odu.edu

A. AlSum
e-mail: aalsum@cs.odu.edu

M. C. Weigle
e-mail: mweigle@cs.odu.edu

M. L. Nelson
e-mail: mln@cs.odu.edu

1 Introduction

A variety of research has been conducted for studying Web archives to answer questions related to user needs and to present Web archive data to users [22,34,43,67]. However, no previous work has been carried out to answer these questions: What do Wayback Machine users look for in the context of languages of the pages they request? Why do users come to Web archives? Do they come because they cannot find the Web pages on the live Web, or do they come because they want a copy of a Web page at a specific time? Where do Web archive users come from? Who links to Web archives? How do sites link to Web archives? Do sites link deeply to specific archived pages or link to the repository? Why do sites link to the past? Is there a relationship between the referrer and the time spent in the archive? Is there a relationship between the referrer and the number of browsed archived pages?

The Internet Archive [39] is the first Web archiving initiative attempting global scope and currently holds over 360 billion Web pages with archives as far back as 1996 [46]. It allows traveling back in time for traversing archived versions of Web pages through the Wayback Machine [61]. In this paper, we examine the requests of Web archive users, both humans and robots, to gain insight into what users look for, in the context of the language of the requested pages, through an analysis of the server logs of the Internet Archive's Wayback Machine. We deduce the reason for using the Web archives by checking the status of requested Web pages on the live Web. We also provide an analysis of referring pages of human users to investigate how humans discover the Wayback Machine, why the referrers link to Web archives, and how they link to Web archives. Finally, we investigate if there is a relationship between the referrer and the session length and duration.

We found that users of Internet Archive's Wayback Machine request English pages the most, followed by several European languages. We also found that most human users come to the Wayback Machine via links or direct address presumably because they did not find the requested pages on the live Web. Of the requested archived pages, 65 % do not currently exist on the live Web, which we believe implies that many users come to Web archives because they do not find the Web pages on the live Web. From analyzing the referrers, we found that more than 82 % of human sessions have referrers, while only 15 % of robot sessions have referrers. We also found that 86 % of the referrers link deeply to archived pages at specific times, and of those 83 % no longer exist on the live Web. In terms of linking, there is an overall preference for linking to the recent past. Finally, we found that the users who come from search engines and the Internet Archive's home page have longer sessions in terms of the number of browsed pages than users who come from external Web sites. Most of the sessions that are composed of one request come from external web sites.

This paper is organized as follows. Definitions of important terms and a review of related work on Web usage mining and Web archive studies are presented in Sect. 2. Section 3 contains a description of the Wayback Machine's Web server logs, the sampling methodology, and the dataset we used in the analysis. The methodology of this study and the methodology of the analysis are presented in Sect. 4. Section 5 contains the results from analyzing the Wayback Machine access logs in terms of the content language of the requested pages and the existence of these pages on the live Web. Section 6 presents a detailed analysis of the referrers of human users. Future work and conclusions for the findings are presented in Sect. 7.

2 Background

2.1 Related work

To the best of our knowledge, no prior study has analyzed where Web archive users come from or what they look for in terms of the linguistic context. Furthermore, the usage of Web archives in general has not been widely studied. The characterization of search behavior and the information needs of Web archive users have been studied by Costa et al. [21,22] based on quantitative analysis of the Portuguese Web Archive (PWA) search logs.

In a previous study [9], we provided the first analysis of user access to a large Web archive. We discovered four basic access patterns for Web archives through analysis of Web server logs from the Internet Archive's Wayback Machine. In the study, we applied heuristics for robot detection after data filtering and found that robot sessions outnumber human sessions 10:1. Robots outnumber humans 5:4 in terms of raw, unfiltered requests, and 4:1 in terms of megabytes transferred.

Many studies have investigated what is missing from digital libraries and Web archives, in addition to the effect of this on the satisfaction of users' needs and expectations [17,51,60,68]. In [60], the Internet Archive's coverage of the Web was investigated. The results showed an unintentional international bias in the archive coverage through uneven representation of different countries in the archive. The reason of unbalanced representation of countries is visibility of the Websites (i.e., the number of inlinks of Websites). The results also showed that the language of a Website does not have an effect on how Internet Archive indexes it.

Ainsworth et al. [8] estimated the coverage of Web resources in Web archives in "How Much of the Web Is Archived?" They sampled 4,000 URIs from DMOZ, Delicious, Bitly, and search engines and measured their coverage in the public Web archives and the number and frequency of archived versions. They found that, according to the URI source, the archived percentage varies from 16 to 79 %.

AlSum et al. [10] studied the coverage of 12 Web archives on 3 data sets from the live Web, web server access logs of the archives, and fulltext search of the archives to create profiles for the 12 archives. They discovered that IA has the largest and widest coverage of all the archives. This finding matches our results of checking the coverage of other archives in Sect. 5.

Carmel et al. [17] suggest a tool to dynamically analyze the query logs of a digital library system, identify the missing content queries, and then direct the system to obtain the missing data.

We investigate what is missing through an analysis of requests with an HTTP 404 status in the Wayback Machine Web server logs.

2.2 Memento terminology

In this section, we explain the terminology we adopt in the rest of the paper. Memento [64] is an HTTP protocol extension which enables time travel on the Web by linking the current resources with their prior state. Memento defines the following terms:

- URI-R identifies the original resource, which is the resource on the live Web that is being archived. A URI-R may have 0 or more mementos (URI-Ms).
- URI-M identifies an archived snapshot of the URI-R at a specific datetime, which is called Memento-Datetime, e.g., URI-M$_i$= URI-R@t_i.
- URI-T identifies a TimeMap, a resource that provides a list of mementos (URI-Ms) for a URI-R with their Memento-Datetimes, e.g., URI-T(URI-R) = {URI-M$_1$, URI-M$_2$, ..., URI-M$_n$}.

```
0.247.222.86 - - [02/Feb/2012:07:03:46 +0000] "GET
http://wayback.archive.org/web/*/http://www.cnn.com HTTP/1.1" 200 96433
"http://www.archive.org/web/web.php" "Mozilla/5.0
(Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 - - [02/Feb/2012:07:03:55 +0000] "GET
 http://web.archive.org/web/20130318135600/http://www.cnn.com/ HTTP/1.1" 200 18875
"http://wayback.archive.org/web/*/http://www.cnn.com" "Mozilla/5.0
(Macintosh; Intel Mac OS X 10_6_8)  AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"}

0.179.81.310_0 - - [02/Feb/2012:13:46:16 +0000] "GET
http://wayback.archive.org/web/20071015000000*/http://9gag.com HTTP/1.1" 200 118819
"http://fr.wikipedia.org/wiki/9gag" "Mozilla/5.0
(Windows NT 5.1; rv:9.0.1) Gecko/20100101 Firefox/9.0.1"

0.251.197.1210_0 - - [02/Feb/2012:18:40:57 +0000] "GET
http://web.archive.org/web/20071008113630/http://www.filg.uj.edu.pl/ifa/przeklad/przeklad2/poezja2.html
HTTP/1.1" 200 25335
"http://info-poland.buffalo.edu/web/arts_culture/literature/poetry/szymborska/poems/link.shtml"
"Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; .NET CLR 1.1.4322; .NET CLR 2.0.50727)"

0.83.5.950_0 - - [02/Feb/2012:03:18:56 +0000] "GET
http://web.archive.org/ HTTP/1.1" 302 0
"http://www.google.co.uk/search?gcx=c&sourceid=chrome&ie=UTF-8&q=website+archiver"
"Mozilla/5.0 (X11; Linux i686) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/14.0.835.186 Safari/535.1"
```

Fig. 1 Sample of the Wayback Machine access log. The first example is a request for a TimeMap, while the second one is a request for a memento. The last three examples are different cases for how the users linked to the Wayback Machine. In the third example, the referrer is Wikipedia, which links to a partial TimeMap (TimeMap for a year only). The fourth example shows an example of an external referrer. The fifth request shows an example of a Google referrer

Although we use Memento terminology, the logs we analyze are from the Internet Archive's Wayback Machine and not the Memento API.

3 Dataset

We used Internet Archive's Wayback Machine server logs in our analysis. The Internet Archive anonymized the client IP addresses, so it is not possible to geolocate the incoming requests. Furthermore, in the interest of further protecting the anonymity of their users, the Internet Archive recently announced that they were encrypting all traffic to their site [15,30,56].

3.1 Wayback Machine access logs

A Web server log file is a plain text file that records the activity of the submitted requests from users of the Web server. The Wayback Machine access logs contain the following fields[1]: client IP address, access time, HTTP request method (GET or HEAD), URI, the protocol (HTTP), HTTP status code (200, 404, etc.), bytes sent, referring URI, and User-Agent. A segment of five requests from the Wayback Machine server log is shown in Fig. 1. The first line shows a request for a URI-T. The second line shows a request for a URI-M. The last three requests show different cases for the referrer field. As we mentioned earlier, for privacy purposes, Internet Archive anonymized the client IP addresses.

3.2 Sampling the dataset

Before analyzing the dataset, we used mixed sampling strategies [32,59] to prove that the samples we used in the analysis are representative.

The Wayback access logs were sampled using two probability techniques: cluster sampling, which is choosing a cluster of data randomly, and random sampling, where each sampling unit has an equal chance of being included. We performed cluster sampling by choosing 7 days (Feb. 2–8, 2012) and random sampling by taking a random slice from each day. We then used purposive sampling [62,65] for choosing the samples we used in the analysis. Because we checked the language of the content of requested pages, we covered the peak times of Internet traffic for many countries with speakers of different languages to avoid biasing the results. We picked samples from the log file which are representative for the peak times of several cities around the world, as shown in Fig. 2. According to previous studies, the hours between 6 p.m. and 12 a.m. (i.e., midnight) are considered to be peak times for Internet traffic [25,45,66]. Home internet use has been well-studied, at least in the United States, and reveals that people engage in a wide range of activities, including commerce, entertainment, job and career enrichment, classes, and news [27,52]. Note that even though we focused on choosing samples that cover the peak times in multiple cities, each sam-

[1] Apache "Combined Log File Format".

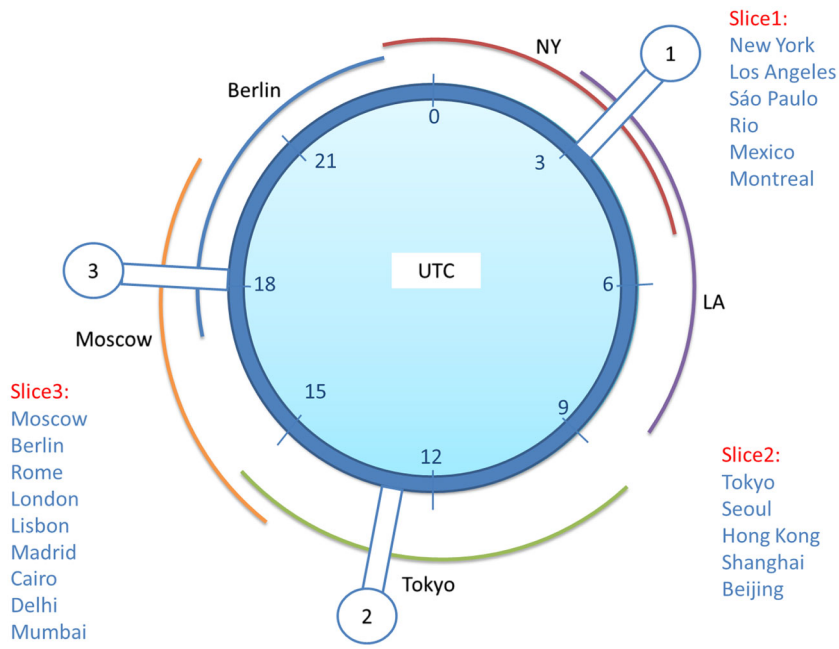

Fig. 2 The dataset of 6M HTTP requests is constructed from slices of 2M each from 03:00, 13:00, and 18:00 UTC on February 2, 2012. The peak hours in New York, Los Angeles, Tokyo, Moscow, and Berlin are indicated by *arcs*

Table 1 Features for each sample of 2M records, Feb. 2–8, 2012 and the three samples from different times we used in our analysis

Samples	Duration	GET (%)	Embedded (%)	s2xx (%)	s3xx (%)	s4xx (%)	s5xx (%)	SI robots (%)	NullRef (%)
2 Feb@03UTC	0:31:30	98.5	37.8	32.1	51.2	11.7	5.0	7.2	48.1
2 Feb@13UTC	0:26:42	99.3	41.6	33.0	51.7	12.4	2.9	5.0	50.3
2 Feb@18UTC	0:26:10	98.3	49.3	34.3	51.4	12.0	2.4	4.3	41.5
2 Feb@06UTC	0:31:30	98.4	37.8	33.7	51.8	11.7	2.8	4.5	42.6
3 Feb@07UTC	0:31:15	99.3	34.8	32.4	52.3	13.1	2.3	12.0	56.6
4 Feb@12UTC	0:40:34	97.7	43.7	34.2	50.8	12.0	3.0	7.7	47.5
5 Feb@09UTC	0:42:57	97.9	42.7	33.2	52.2	11.6	2.9	7.7	47.0
6 Feb@13UTC	0:29:35	99.4	41.9	34.1	51.7	11.2	3.0	2.9	49.4
7 Feb@21UTC	0:25:45	99.7	44.7	33.4	51.9	10.3	4.4	3.5	42.6
8 Feb@16UTC	0:24:33	99.8	46.8	33.6	53.2	10.1	3.1	3.8	43.9
Mean	0:31:03	98.8	42.1	33.4	51.8	11.6	3.2	5.9	46.9
STDDEV	0:05:37	0.7	4.0	0.7	0.6	0.8	0.8	2.5	4.1
STDERR	0:01:47	0.2	1.3	0.2	0.2	0.3	0.2	0.8	1.3

ple also covers work hours for other cities of the world. For example, the 13:00 UTC sample that covers the peak time of Moscow, Berlin, etc., will cover the work hours for New York City (8 a.m. Eastern Daylight Time).

Table 1 contains the following features for each sample:

- *Duration*: the difference between the last request time and the first request time of each sample in HH:MM:SS format.
- *GET*: the percentage of requests that used the GET method.
- *Embedded*: the percentage of requests that were for embedded resources of Web pages (such as images and CSS files, etc.).
- *s2xx*: the percentage of successful requests (2xx status code).
- *s3xx*: the percentage of redirections (3xx status code).
- *s4xx*: the percentage of client errors (4xx status code).
- *s5xx*: the percentage of server errors (5xx status code).
- *SI Robots*: the percentage of requests by self-identified robots based on the User-Agent field.
- *NullRef*: the percentage of requests that had an empty referral field.

Each sample consists of a slice of 2M requests to the Wayback Machine Web server. The last three columns of Table 1 show the mean, standard deviation, and corresponding standard error between the samples. We can see small standard

errors between the samples, from which we conclude the samples of Feb. 2, 2012 are representative and can be used for the analysis. There could be exceptional days that result in an increase in traffic in the Web archive, such as an anniversary of an important event (e.g., 9/11 collection[2]), or a decrease in traffic when something happens to the Web archive, such as the fire that occurred at the scan center of IA's headquarters in Nov. 2013 [5,6,29].

From Table 1, we notice that HTTP 3xx (e.g., the fifth example of Fig. 1) accounts for 51 % of the total number of requests. This is related to the default Wayback Machine behavior. First, the Wayback Machine rewrites all of the hyperlinks of a memento's embedded resources with the memento's timestamp. Second, in the resolution of these URIs, the Wayback Machine will redirect the request of the embedded resources and hyperlinks to the nearest (timestamp) available memento. Furthermore, the Wayback Machine responds with a 302 status first when the requested URI-R is not available on the Wayback Machine, and then responds with a 404 status.

We composed 6M requests from three slices of times, in which each slice is 2M records from the Wayback Machine Web server. We chose the samples from Feb. 2, 2012, which was representative for a normal traffic day, at different times.

4 Methodology

4.1 Preprocessing

The breadth and depth of studies on the topic of Web log preprocessing are massive and increasing [9,12,44,54,58]. Preparing the Wayback access logs for usage mining starts with transforming the raw log file into server sessions through Web log preprocessing (data cleaning, user identification, and session identification) [20].

Data cleaning is eliminating the irrelevant entries from the log file [37]. Because robots crawl Web archives intentionally, we did not eliminate their requests in the cleaning step. We eliminated the following items which were irrelevant in terms of user behavior:

1. Requests that were generated automatically by the Web browser for embedded resources of the requested Web page (such as graphic files, page style files, etc.).
2. Entries with an HTTP status code other than HTTP 200, 404, or 503. We kept only these because we considered them to be requests executed by the user.
3. Requests using the HEAD request method (as suggested by [37]).
4. Static resources of the Internet Archive Web site and the URIs of the liveWeb service, which the Internet Archive introduces to redirect the user to the live Web when the copy is not found on the Wayback Machine.
5. Invalid requests from Web sites which included a link for malformed URI-Rs (for example, `about:blank`) among their embedded resources, so that each request on their Web sites caused automatic requests to the Wayback Machine server. Similar behavior had been detected by Omodei [42].

There were 1,219,408 requests (20 % of the raw file) remaining after cleaning, which is consistent with our previous analysis of Web archives [9]. We then applied user identification by the IP and the User-Agent fields. For illegitimate users who change their User-Agent every request, we used 20 different User-Agents as a threshold for the number of User-Agents per IP to detect malicious robots, as suggested by [9].

A session is the group of consecutive requests that is performed by a user [38]. Session identification starts with user identification, then continues by applying a threshold timeout so that if the time elapsed between two consecutive requests is longer than this threshold, the second request is considered as the first request of the new session. There have been several suggested timeout thresholds including 10 min [37,53], 25.5 min [19], 30 min [36,57], and 60 min [11]. We used 10 min as a timeout threshold for session identification.

After identifying the sessions, we extracted features for each session to be used further in the analysis. A session, S, is 5-tuple.

$$S = \langle \text{URI}, S_\text{l}, S_\text{d}, \text{BS}, \text{IH} \rangle$$

The following is the description of each item:

- URI is the set of URIs that the user visited in the session. The set of URIs are defined as:

$$\text{URI} = \{\text{URI}_i \,|\, i \text{ is an integer}, 1 \leq i \leq S_\text{l} \text{ and URI} \in \{\text{URI-T}, \text{URI-M}\}\}$$

- S_l, session length, is the number of Webpages the user requested during the session.
- S_d, session duration, is calculated by subtracting the timestamp of the first request of the session from the timestamp of the last request of the session.
- BS is the browsing speed of each session. $\text{BS} = S_\text{l}/S_\text{d}$
- IH, image-to-HTML, is the ratio between the number of image files and the number of HTML files per session.

[2] https://archive.org/details/911.

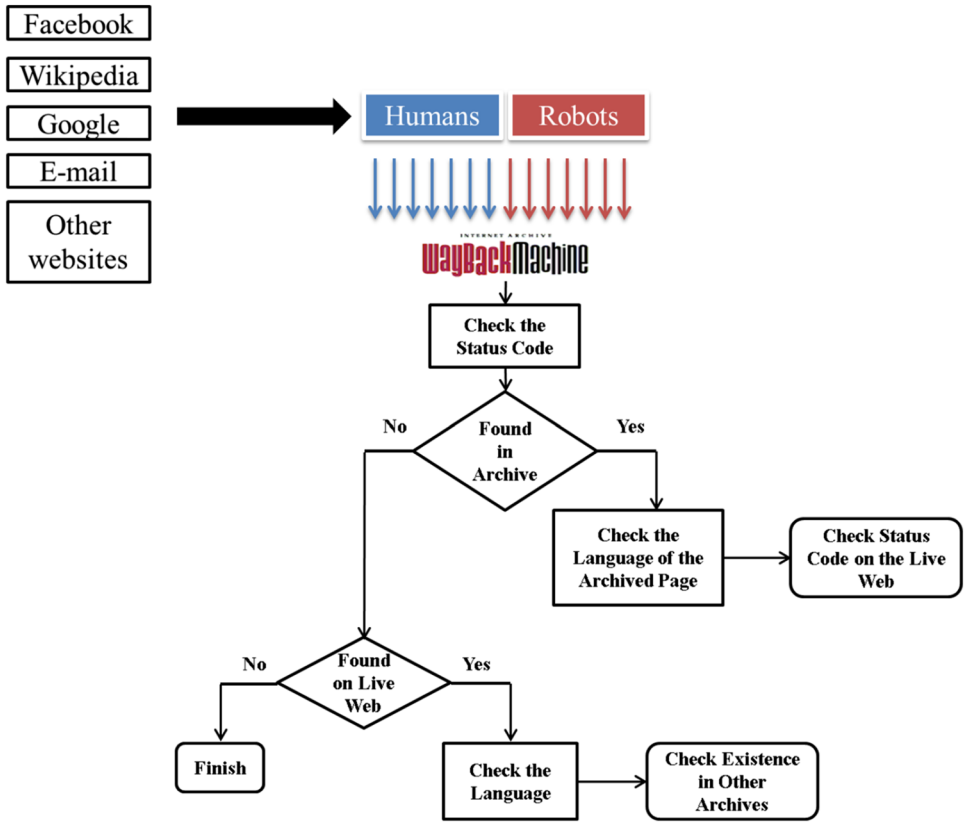

Fig. 3 Overview model for results of the paper

4.2 Detecting robots

We had the challenge of detecting the robots that had not identified themselves in the User-Agent field. The robot detection problem has been investigated in several studies [23,24,57]. We used several robot detection heuristics which have shown great effect in distinguishing robots from humans [9]:

- User-Agent check: a syntactical log analysis that is done by checking the User-Agent field. This heuristic is helpful for the robots which declare their identity to the Web server through the User-Agent field.
- Number of User-Agents per IP: as mentioned earlier, based on our previous study [9], IP addresses with more than 20 different User-Agents were classified as robots.
- Robots.txt file: the sessions in which users downloaded the robots.txt file for the Wayback Machine (http://web.archive.org/robots.txt) are labeled as robots.
- Browsing speed: based on various studies [18,41,44,58], we used BS $\leq$0.5 (i.e., no faster than one request every two seconds) as a threshold for human browsing speed.
- Image-to-HTML ratio: a session with IH $<$0.1 (less than one image file for every 10 HTML files) is labeled as a robot. This heuristic is the best predictor for robots [9,55,57].

After applying the robot detection heuristics, we found that robots contribute to 92.4 % (1,127,204 out of 1,219,408) of the filtered requests. The percentage is consistent with our previous analysis of Web archives [9].

Figure 3 shows a model for the flow of the paper and how we achieved the results of each section. After distinguishing the human requests (92,204) from robot requests (1,127,204), we performed a check of the status code. For the requested Web pages that were found in the archive, we checked the language of the URI-Ms of these requested pages. We also checked the status codes of the URI-Rs of the requested pages on the live Web to discover why people came to Web archives. For the requested Web pages that were not found in the archive, we first check their relative URI-Rs existence on the live Web, then for those that were found on the live Web, we detected the language. We also checked the existence of the missing pages on other archives. For discovering who links to Web archives, how people reach Web archives, why they link to Web archives, and how they link to Web archives (i.e., deep link or link to the whole repository), we analyzed the referrer field of the humans as detailed in Sect. 6.

Table 2 The top 10 languages for URI-Ms with HTTP 200 (on the left) and for the URI-Rs of unarchived requested pages (on the right)

URI-Ms with HTTP 200				URI-Rs with HTTP 404			
Language	Humans (%)	Language	Robots (%)	Language	Humans (%)	Language	Robots (%)
English	71.7	English	72.4	English	66.9	English	62.2
Japanese	5.5	Russian	7.0	Russian	7.9	Russian	11.1
German	3.6	German	3.1	German	5.4	German	3.8
Vietnamese	2.9	Spanish	1.9	Japanese	5.1	Indonesian	3.1
Russian	2.3	French	1.8	Spanish	2.5	Polish	2.5
Portuguese	2.1	Vietnamese	1.7	Polish	2.3	Vietnamese	2.2
French	2.1	Japanese	1.5	Romanian	1.6	Spanish	2.0
Spanish	1.9	Polish	1.5	French	1.2	Thai	1.9
Bengali	1.8	Portuguese	1.3	Italian	0.8	French	1.8
Italian	0.9	Thai	1.1	Portuguese	0.7	Dutch	1.1

Table 3 The existence of the requested archived pages on the live Web. Available represents the requests which ultimately return "HTTP 200", while missing represents the requests that return HTTP 4xx, HTTP 5xx, HTTP 3xx to others except 200, timeouts, and soft 404s

	Found in archive		Unarchived	
	Humans	Robots	Humans	Robots
URI-Rs available on live Web (%)	36.4	62.5	25.4	33.2
URI-Rs missing from live Web (%)	63.6	37.5	74.6	66.8
Unique URI-Rs	40,791	331,573	2,441	209,384

5 What do Wayback Machine users look for?

In this section, we give insight into what Web archive users look for in terms of the content language of requested pages. We used the language detection library created by Shuyo [50] for detecting the language with precision over 99 % for 53 languages [16,49]. We also check the existence of the requested Web pages on the live Web to discover why users access Web archives.

5.1 Archived Web pages

5.1.1 Distribution of languages used in the Wayback Machine

We extracted the successful requests (HTTP 200 status code) from humans and robots to detect the language distributions for the content of the requested pages. These successful requests represent 93.1 % (85,909 out of 92,204) of all human requests and 56.7 % (639,684 out of 1,127,204) of all robot requests. The request can be for a URI-T or a URI-M. For the URI-Ts, which represent 13 % of human-requested pages and 80.8 % of robot-requested pages, we estimated the language using the most recent URI-M from the TimeMap.

We identified 52 different languages from the successful requests. The left two columns of Table 2 show the top 10 languages which accounted for 94.8 % of human and 93.4 % of robot requests. For both human and robot users, English contributes the most to the successful requests, reflecting the high Web archive penetration rate in English-speaking countries. Japanese is the second most frequent language with 5.5 % for humans, and Russian is the second most frequent language for robots at 7.0 %. We also notice that despite the existence of Web archives in Europe, the requests to the IA from speakers of European languages contribute 13 % of the top 10 list for human requested pages and 18.5 % of the top 10 list for the robot requests. We assume that this is because of the popularity of the Internet Archive, so most of the people who know about Web archiving may only know about the Internet Archive.

5.1.2 Existence on the live Web

From all 85,909 successful human requests, we checked the existence of the 40,791 unique URI-Rs on the live Web. The robots generated 639,684 successful requests, in which there were 331,573 unique URI-Rs whose existence on the live Web was also checked. We also checked the pages that give "soft 404s", which return HTTP 200 but do not actually exist, based on the algorithm in [14]. Table 3 contains the results of checking the status of the Web pages that were found in the archive on the live Web. For example, the URI-R "http://err.lolipop.jp/404.html" is not available on the live Web but it is found in the archive:

Table 4 The number of found URI-Rs and the corresponding URI-Ms of the missing pages (211,825 unique URI-Rs) from the Web archives

Web Archive	Archive Web Site	#URI-R	#URI-M
Internet Archive (2013)	Web.archive.org	56,503	1,657,264
The National Archives	webarchive.nationalarchives.gov.uk	787	15,354
ArchiefWeb	www.archiefweb.eu	47	18,347
Archive-It	archive-it.org	41	4,682
UK Web Archive	www.webarchive.org.uk	38	12,277
Library of Congress	webarchive.loc.gov	35	1,092
WebCite	webcitation.org	29	1,104

```
curl -I "http://err.lolipop.jp/404.html"
HTTP/1.1 404 Not Found
...

curl -I "http://web.archive.org/web/*/
http://err.lolipop.jp/404.html"
HTTP/1.1 200 OK
...
```

We believe that humans access the Wayback Machine because they do not find Web pages on the live Web. Table 3 shows that for the requested pages that were found in the archive (returned HTTP 200 status), the percentage of the missing pages from the live Web for human requests is 63.6 %. On the other hand, the percentage of the missing pages from the live Web for robot requests is 37.5 %.

5.2 Unarchived Web pages

Of the 6M requests in our sample, 12 % returned an HTTP 404 status. Not all of these are actually unarchived; approximately 2 % of the unique URI-Rs are malformed (e.g., http://cnn.com) and were removed. We used the remaining valid URI-Rs (209,348 robots and 2,441 humans) to detect content language, check live Web status, and check existence in other archives.

5.2.1 Existence on the live Web

The current state of the requested URI-Rs that had HTTP 404 status was determined by testing their existence on the live Web. For example, the URI-R "http://SMALLPOTTER.RU" gives "404 Not Found" in the archive and also does not exist on the live Web:

```
curl -I "http://wayback.archive.org/web
/*/
http://SMALLPOTTER.RU/"
HTTP/1.1 404 Not Found
...

curl -I "http://SMALLPOTTER.RU"
curl: (6) Couldn't resolve host
'SMALLPOTTER.RU'
```

The results of the requested pages that were not found in the archive are shown in the right hand side of Table 3. Of the URI-Rs that were not found in the Wayback Machine, 66.8 % of those requested by robots and 74.6 % of those requested by humans do not exist on the live Web. To compensate for transient errors, we repeated the requests several times for a week before declaring a URI-R non-existent.

5.2.2 Distribution of the content language for unarchived Web pages

We detected the content language of available URI-Rs on the live Web, which represent 25.4 % (620 out of 2,441) of the unique URI-Rs for humans and 33.2 % (69,510 out of 209,384) for robots. The total number of requested URI-Rs is 227,450 for robots and 1,578 for humans. The two rightmost columns of Table 2 have the results for robots and humans separately. For the Web pages that were not archived in IA's Wayback Machine, English is the most requested language with 66.9 % of the human-requested Web pages and 62.2 % of the robot-requested Web pages. The top 10 languages comprised 94.5 % of all the content language of the requested pages. European languages made up 22.5 % of the human-requested pages and 22.4 % of the robot-requested pages.

5.2.3 Existence in other Web archives

We checked the 211,825 unarchived pages for existence in other archives at the time of the experiment. The existence in the Web archives was tested by querying Memento proxies and aggregator [63]. For completeness and fairness, we also included the results from IA's Wayback Machine in March 2013. This resulted in 56,503 out of 211,825 URI-Rs that were unarchived in Feb. 2012 now being available in the archive. Table 4 contains the number of URI-Rs found in the Web archives and the number of covered URI-Ms.

The Internet Archive has the most coverage at the time of experiment as it has increased its repository recently [31].

6 Where do Wayback Machine users come from?

We used the referrer field, which contains the Web page that links to the resource to determine how people discover the Wayback Machine. Table 5 contains the result of analyzing the referrer field. As we see in the table, in terms of sessions, 84.8 % of robot sessions do not have referrers while only 18.1 % of human sessions do not have referrers (i.e., they reached the Wayback Machine by a link in an email, direct address, or direct bookmark). An empty referral field is a strong indicator of a robot, which means that the behavior of robots in Web archives is similar to the behavior of robots on the live Web in terms of the referral field.

In this section, we provide a detailed analysis of the referrer field of human users to gain insight into who links to the Wayback Machine and how they link to it. Robots are not included in the analysis of referrers because the majority of robots do not have referrers and for those that do, we do not necessarily trust their values.

6.1 Who links to the Wayback Machine?

The percentage of human sessions with referrers is 81.9 %. We eliminated the sessions that were referred by a URI-M or URI-T because their sessions started prior to our sample. Of the sessions that started with an external referrer, 9.6 % came from Google. The users who came from the home page of the IA contributed to 11.9 % of the sessions with referrers. That means many people start at the IA to access the Wayback Machine.

Table 5 The referrer statistics for the robots and humans

Sessions	Referrer (%)	Null referrer (%)
Human	81.9	18.1
Robot	15.2	84.8

Table 6 The top 10 referrers

Web Site	Percentage (%)	Description
en.wikipedia.org	12.9	Wikipedia
archive.org	11.9	IA Home Page
reddit.com	10.2	Social News Web Site
google.TLD	9.9	Google Search Engine
info-poland.buffalo.edu	1.5	Polish Studies
de.wikipedia.org	1.4	Wikipedia
cracked.com	1.2	Humor Site
snopes.com	1.1	Urban Legends Reference Pages
facebook.com	0.9	Social Media
crochetpatterncentral.com	0.9	Crocheting Hobbies

6.1.1 Top referrers

Table 6 contains the top 10 referrers that link to IA's Wayback Machine. The list of top 10 referrers represents 51.9 % of all the referrers. As the table shows, en.wikipedia.org outnumbers all other sites including Google search and the home page of Internet Archive (archive.org). Note that Google search is 99.3 % of all search engines that link to the archive.

Wikipedia is one of the most popular information resources in the world with more than 500 million unique visitors monthly [7]. Wikipedia typically links to external references in each article, but there are about 128,604 articles with dead links [2]. Wikipedia editors have more knowledge about citing Web pages and are aware of linkrot (link death or breaking) [3] and citing archived pages of the Web pages they refer to [4]. There is an initiative from the Internet Archive for fixing broken Wikipedia links which will cause an increase in the usage of the archive among the Wikipedia community editors [46]. A real example from the logs that reflects how Wikipedia editors cite archived pages is shown in the third line of Fig. 1. The referred-to page is a partial TimeMap, the TimeMap of a Website in specific year.

Many of the people who came to the Wayback Machine via a referrer, came through a search engine, as shown in Table 6. Note that "google.TLD" represents Google search and 24 other Web pages from Google's Website (e.g., http://www.google.com/about/company/history.html). Since the majority are from Google search, we describe it as Google search engine. An example that shows how people find the archive is presented in the fourth sample of Fig. 1. Facebook also appears as a top referrer, which indicates that many people share links to the past. The data in Table 6 suggest that the top 10 list of referrers is not dominated only by popular Web sites, such as Wikipedia and Facebook. There are non-popular Web sites (e.g., crochetpatterncentral.com, info-poland.buffalo.edu) as well, but with small percentages. We assume that the reason for having non-popular Web sites among the top 10 referrers is that the results represent the long tail of access. If a few people know about the Internet Archive they will be the reason for traffic from non-popular Web sites. For example, if even one or two users in a presumably small community, such as crocheting, are aware of the Internet Archive and create links to archived Web pages in it, then this will produce a noticeable spike in referrals. It would not be surprising to find that on other days, different non-popular Web sites appear in the top 10 referrers with small percentages.

Table 7 The top 10 TLDs of the referrers

TLD	.com (%)	.org (%)	.net (%)	.jp (%)	.ru (%)	.de (%)	.edu (%)	.to (%)	.uk (%)	.info (%)
Percentage	45.4	33.9	8.4	1.8	1.4	1.4	1.1	0.7	0.6	0.5

Table 8 The top 10 ccTLDs of Google search referrers

ccTLD	.com (%)	.uk (%)	.de (%)	.ca (%)	.jp (%)	.pl (%)	.nl (%)	.ru (%)	.fr (%)	.br (%)
Percentage	56.7	6.0	5.3	4.8	3.7	2.2	1.9	1.7	1.5	1.4

6.1.2 Classification of referrers

Table 7 presents the distribution of Top-Level Domains (TLD) for the URIs that link to the IA's Wayback Machine (only the top 10 are shown). The data in the table suggest that most of the connections are from the .com, .org, .net, .jp, .edu, and .ru domains. Despite the existence of many Web archives in Europe, there are many European domains linking to the IA, such as .ru (Russia), .de (Germany), .fr (France), and .it (Italy). Note that .to is the TLD for a Russian language site (http://lurkmore.to/).

For the referrers from Google search, we extracted the country code top-level domain (ccTLD) of the URIs to discover the countries of the users who came to the Wayback Machine through the search engine. The results are shown in Table 8. English-speaking countries are in the lead, followed by the European language countries.

6.2 Inter-linking between languages

From the analysis of the content languages of the referrers and the archived pages which have been linked by the referrers, English represents 80.7 % of the referrers' content languages and 80.2 % of all referred pages. English referrers link to English archived pages 92 % of the time. A small percentage of English referrers link to pages in other languages. The top five languages that English pages link to are (in decreasing order) Portuguese, Vietnamese, French, and German. Figure 4 contains a directed weighted graph, which is created using Circos [35], to show the relationship between the languages of the referrers and referees. We exclude English from the graph to be able to analyze the rest of the languages and see what they are linking to. For a particular language, the length of the outer arc represents the sum of the number of referrer pages and the number of referee pages in that language. Moving toward the center, the next arc represents the percentage of referees, and the third arc represents the percentage of referrers in that language. For example, links to Japanese archived Web pages denote 46 % (160 out of 357) of all the Japanese language pages for referrers and archived pages all together. The inner circle shows the relationships between languages of the referrers and referees. Ribbons of different widths connect the languages. The direction is represented by a gap between the line and the incoming language (referrer language). For example, there are 30 links from Japanese (ja) pages to Bengali (bn) pages, which are shown as a fairly broad blue line. The languages where the relative number of referrers and referees together is <20 have been excluded to remove noise from the graph.

The figure shows that languages primarily self-link with a few inter-language links (recall that we have excluded English from the figure). Many of the top ranked languages of human-requested pages appear in the top ranked list of referrers, such as Japanese, German, Russian, Spanish, French, Polish, Dutch, Bengali, etc. It is surprising to find many European referrers to IA's Wayback Machine despite the existence of European Web archives. However, many of the European archives are dark (i.e., not publicly accessible) and those that are publicly accessible often have comparatively smaller holdings as a result of their collection policy [10].

6.3 How do Web pages link to the Wayback Machine?

We found that 86.4 % of the Web pages that link to the Wayback Machine point to mementos, which means that they link to Web pages at a specific time, 12.8 % of Web pages point to TimeMaps, and 0.8 % point to the repository (e.g., http://web.archive.org). Google search links to the repository, because Google does not crawl the archive based on the robots.txt exclusion protocol.

6.3.1 Temporal distribution of the referred URI-Ms

Figure 5 shows the total number of mementos which were pointed to by the referrers, grouped by the year of their Memento-Datetime. There is a significant bias toward 2008, then 2007, and then a bias against the more distant past. We found 14 URI-Ms all from a single Web site that link to a datetime in 2099. We assume that the referrer wants to

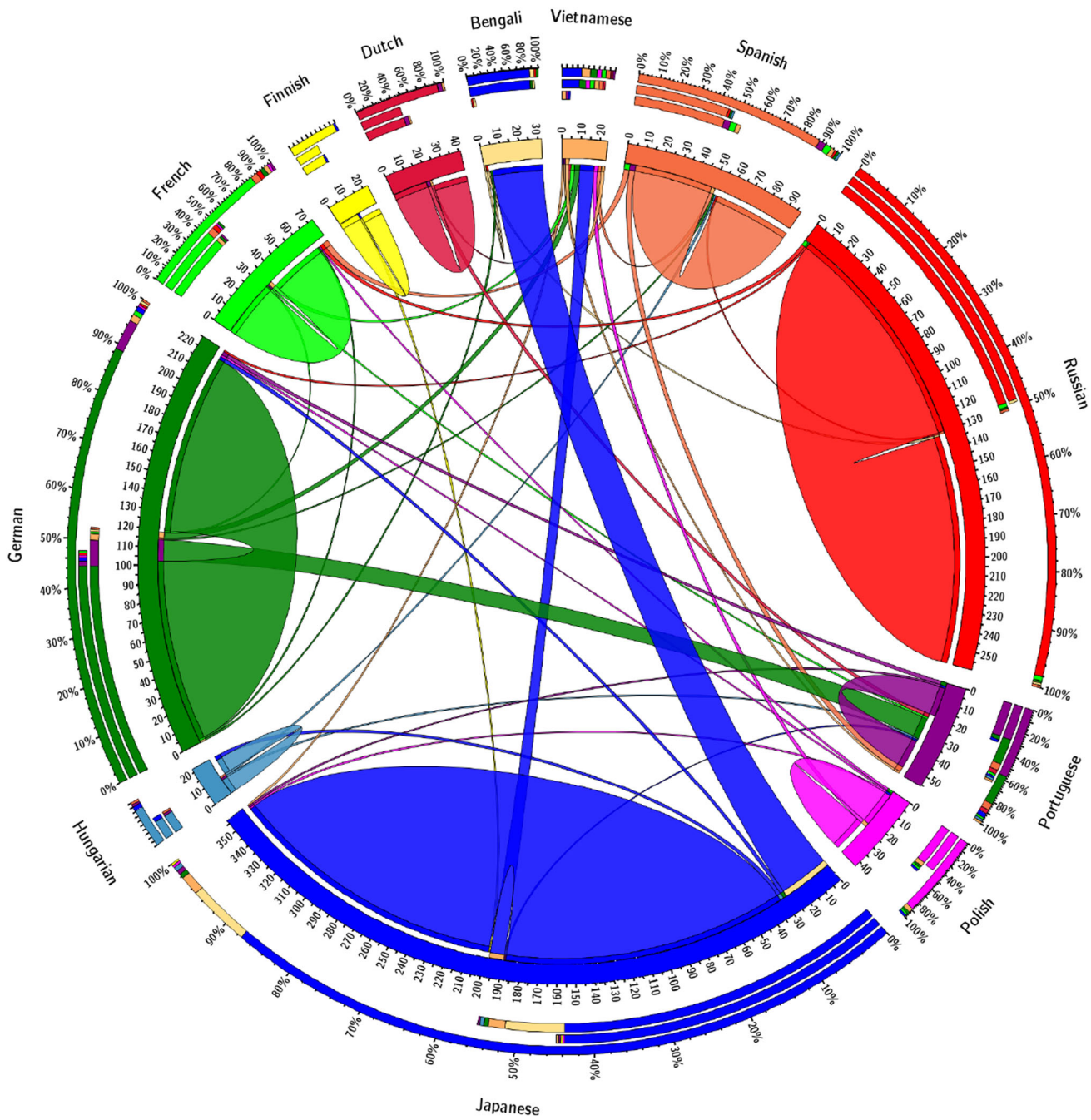

Fig. 4 Most languages self-link, with the notable exceptions of Japanese → {Bengali, Vietnamese} and German → Portuguese

redirect the site's visitors to the most recent copy of the linked Web page.

6.3.2 Why do Web sites link to the Wayback Machine?

The nature of the Web is ephemeral and the expected lifetime of a Web pages is short [26,28,40,48]. Many of the changes that frequently occur on the Web are actually the loss of the Web pages themselves [33,47]. So, Web archives are important to Webmasters and third parties for preserving and saving many Web sites [13]. Figure 5b clarifies that most people link to the Wayback Machine because they did not find the pages on the live Web. The figure shows that for most years, more than 70 % of the referred pages in the archive no longer exist on the live Web. About 83 % of all referred-to URI-Rs do not currently exist on the live Web. That means there are people who are aware of the archive's existence and link to archived pages of articles that no longer exist on the live Web.

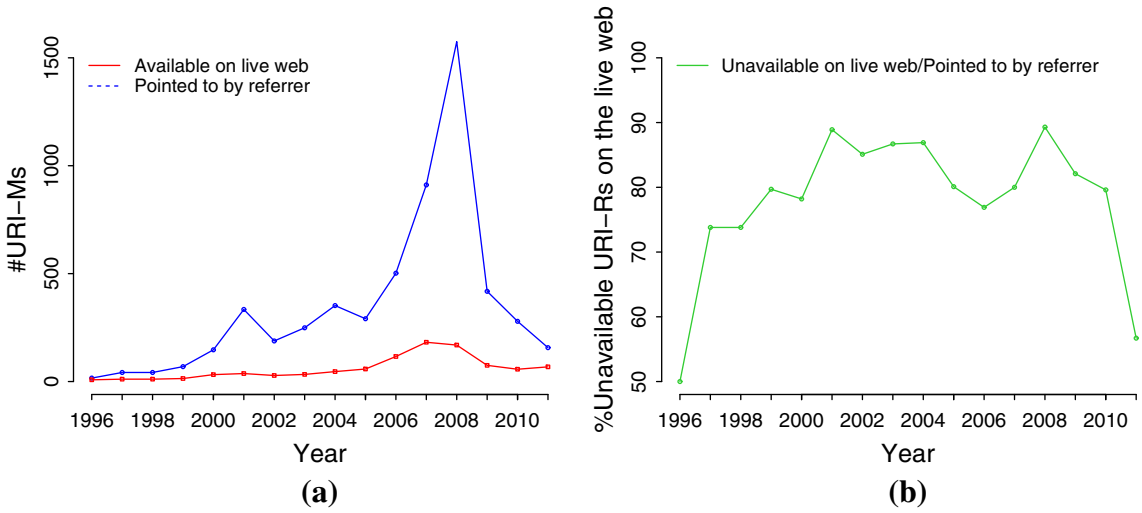

Fig. 5 **a** The temporal distribution of URI-Ms pointed to by the referrers and the number of relative URI-Rs of these URI-Ms that are currently available on the live Web. **b** The percentage of unavailable URI-Rs of these URI-Ms on the live Web

Table 9 The median and average session length and session duration, divided based on the referrer

	Session length		Session duration in seconds	
Referrer	Median	Average	Median	Average
External sites	1	2.9	74	171.2
Search engines	6	11.4	92	190.3
Archive home page	6	11.3	95	199.9
Direct address	2	7.2	136	326.2

Note that session duration does not include the one-request sessions

6.4 Is there a relationship between the referrer and the session length and duration?

In this section, we check if the referrer type relates to the time that the user spent in the archive. We also check if there is a relationship between the referrer type and the number of browsed archived pages. Of all the sessions, 50 % were composed of one request only. In the terms of patterns, which we have identified in our prior work [9], those one-request sessions are called Dips. We investigated if the type of the referrer could be a reason for these Dips.

In this section, we give a detailed analysis of the sessions after dividing them based on their source (i.e., the referrer field) into four categories: sessions from external Web sites, sessions from search engines, sessions from the archive home page, sessions with no referrer (e.g., sessions that used a direct address such as a link in an email). For the sessions with no referrer, which we call "direct address" through the rest of the paper, there is not much information about how they link to the archive (e.g., link in an email or bookmarking) from the logs.

6.4.1 Session length

We found that 77 % of the one-request sessions came from external Web sites. Table 9 shows a summary of median and average values for the session lengths and durations of the four categories of sessions. The left two columns of Table 9 show that the median and average values of session length for the sessions that came from search engines and the archive home page are much larger than median and average values that came from external Web sites. This means that the people who know about the archive browse more pages than the users who come from external Web sites. The sessions that come from a direct address also have longer session lengths than the sessions from the external Web sites.

Figure 6 shows that Dips (sessions of length one) represented 64 % of all the sessions that came from the external Web sites. Also, it is rare to have a long session length when referred by an external Web site. However, the three-request sessions represent the highest percentage, with 12 % of the sessions that came from search engines. The Dips represent only 7 % of all the sessions that came through the search engines. Of the sessions that started with the archive home page, 11 % are Dips. From these results, we notice that the sessions that came from search engines and the sessions that started with the archive home page have similar behavior, in which both browsed more pages than the sessions that came from external Web sites and the null referrer sessions.

6.4.2 Session duration

Table 9 shows a summary of the relationship between the session duration and the referrer. It shows the medians and

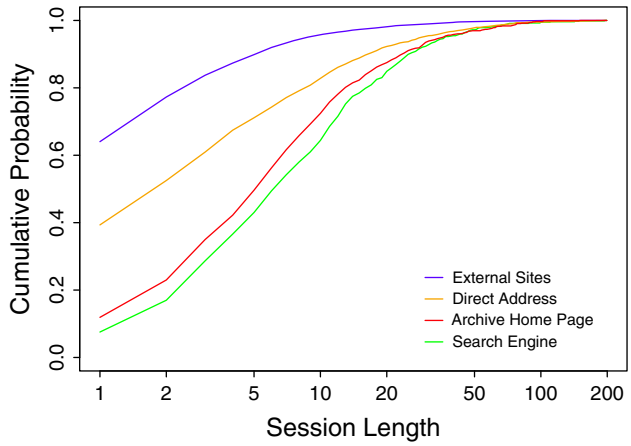

Fig. 6 The CDF for session length based on the referrer types. Note that direct address includes the sessions with null referrers

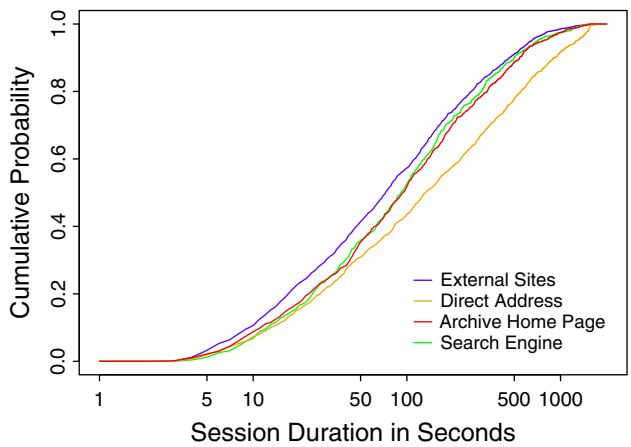

Fig. 7 The CDF for session duration based on the referrer types. Note that direct address includes the sessions with null referrers

averages for each group of sessions, excluding the sessions that were composed of one request only. Although the one-request sessions represent a large portion of the sessions that came from external Web sites (64 %), there is no indication in the logs how long users remained on those pages. In Web analytics, this is termed the "bounce rate" [1]. We notice that the sessions that came from a direct address had longer durations than the rest of the groups, furthermore, the smallest median and average are for the sessions that came from external referrers.

Figure 7 shows the cumulative distribution function (CDF) for the session durations in seconds. We can see from the figure that the sessions that came from a direct address tended to have longer durations. Next are the sessions that came from search engines and the archive home page. The sessions from external Web sites have the smallest durations. We can see that the sessions from the archive home page and from the search engines overlap.

7 Future work and conclusions

In our future work, we plan to extend our analysis of the behavior of robots in Web archives and contrast it with the behavior of robots on the live Web to distinguish their respective behaviors.

From the analysis of Internet Archive's Wayback Machine server logs, we conclude that most humans come to the Wayback Machine to find missing pages from the live Web. The percentage of the requested archived pages which currently do not exist on the live Web is 65 %, which means that many people come to the Web archives because they do not find the Web pages on the live Web. We provided analysis for the distributions of languages to gain insight about what users look for in terms of the languages of the pages they browsed. We found that English is the most used language on the Wayback Machine, followed by many European languages. European languages represent about 22 % of the Web pages that were not found on the Wayback Machine, for both human and robot requests. The large percentage of European languages among the unarchived pages can be a good indicator for archival demand for European Web pages.

We also provided analysis for the referrers of human users to discover where Wayback Machine users come from. We discovered that Wikipedia is the most frequent referrer of pages to IA's Wayback Machine. From analyzing the TLDs of the referrers, we found many European domains (.ru, .de, .fr, etc.) in the top list of the referrers. English represents 80.2 % of the referrer languages, followed by European languages. We found that the languages are linking mainly to themselves and to English.

We found that 86 % of the referrer Web pages link deeply to mementos. More than 82 % of the links to these mementos are because their corresponding URI-Rs do not exist on the live Web. There is a bias toward the recent past in terms of linking. We also found that the users who come through search engines make long sessions in terms of the number of requests per session, while the users who come from external Web sites tend to have short sessions. There are some questions that cannot be answered from the data we have. For instance, where do IA users come from, in terms of their countries? Because Internet Archive anonymized the client IPs, the GeoIP information cannot be retrieved. But we were able to get insight about the languages that have been used by detecting the languages of the Web pages that the users requested.

Acknowledgments This work was supported in part by the NSF (IIS 1009392) and the Library of Congress. We thank Kris Carpenter Negulescu (Internet Archive) for access to the anonymized Wayback Machine logs.

References

1. Bounce rate. http://en.wikipedia.org/wiki/Bounce_rate
2. Category: All articles with dead external links. http://en.wikipedia.org/w/index.php?title=Category:All_articles_with_dead_external_links
3. Wikipedia: Link rot. http://en.wikipedia.org/wiki/Wikipedia:Link_rot
4. Wikipedia: Using the Wayback Machine. http://en.wikipedia.org/wiki/Wikipedia:Using_the_Wayback_Machine
5. Internet Archive appeals for donations after 600,000 in fire damage. http://www.theguardian.com/technology/2013/nov/08/internet-archive-appeals-donations-fire-damage (2013)
6. Internet Archive building damaged by fire. http://www.bbc.co.uk/news/technology-24848907 (2013)
7. Wikimedia Report Card. http://reportcard.wmflabs.org/ (2014)
8. Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How much of the web is archived? In: Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '11, p. 133. ACM Press, USA (2011)
9. AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Access patterns for robots and humans in web archives. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, pp. 339–348. ACM, USA. doi:10.1145/2467696.2467722. http://doi.acm.org/10.1145/2467696.2467722 (2013) ISBN 978-1-4503-2077-1
10. Alsum, A., Weigle, M., Nelson, M., Sompel, H.: Profiling web archive coverage for top-level domain and content language. Research and advanced technology for digital libraries. Lecture notes in computer science, pp. 60–71. Springer, Berlin (2013)
11. Anick, P.: Using terminological feedback for web search refinement—a log-based study. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03, pp. 88–95. ACM, USA (2003)
12. Aye, T.T.: Web log cleaning for mining of web usage patterns. In: IEEE 3rd International Conference on Computer Research and Development, ICCRD, pp. 490–494 (2011)
13. Banos, V., Kim, Y., Ross, S., Manolopoulos, Y.: CLEAR: a credible method to evaluate website archivability. In: Proceedings of the 9th International Conference on Preservation of Digital Objects, iPRES (2013)
14. Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic Transit Gloria Telae: towards an understanding of the web's decay. In: Proceedings of the 13th International Conference on World Wide Web, WWW '04, pp. 328–337. ACM, USA (2004)
15. Broache, A.: FBI rescinds secret order for Internet Archive records. http://news.cnet.com/8301-10784_3-9938603-7.html (2008)
16. Brown, R.: Selecting and weighting N-grams to identify 1100 languages. Text, speech, and dialogue. Lecture notes in computer science, pp. 475–483. Springer, Berlin (2013)
17. Carmel, D., Yom-Tov, E., Roitman, H.: Enhancing digital libraries using missing content analysis. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '08, pp. 1–10. ACM, USA (2008)
18. Castellano, G., Fanelli, A.M., Torsello, M.A.: LODAP: A LOg DAta Preprocessor for mining Web browsing patterns. In: Proceedings of the 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED '07, vol. 6, pp. 12–17 (2007)
19. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the world-wide web. Comput. Netw. ISDN Syst. **27**(6), 1065–1073 (1995)
20. Cooley, R., Mobasher, B.: Data preparation for mining World Wide Web browsing patterns. Knowl. Inf. Syst. **1**, 5–32 (1999)
21. Costa, M., J. Silva, M.: Characterizing search behavior in Web Archives. In: Proceedings of Temporal Web Analytics Workshop, TWAW (2011)
22. Costa, M., Silva, M.J.: Understanding the information needs of web archive users. In: Proceedings of the 10th International Web Archiving, Workshop, pp. 9–16 (2010)
23. Dikaiakos, M.D., Stassopoulou, A., Papageorgiou, L.: An investigation of web crawler behavior: characterization and metrics. Comput. Commun. **28**(8), 880–897 (2005)
24. Doran, D., Gokhale, S.S.: Web robot detection techniques: overview and limitations. Data Min. Knowl. Discov. **22**(1–2), 183–210 (2010)
25. Fukuda, K., Cho, K., Esaki, H.: The impact of residential broadband traffic on Japanese ISP backbones. SIGCOMM Comput. Commun. Rev. **35**(1), 15–22 (2005)
26. Harrison, T.L., Nelson, M.L.: Just-in-time recovery of missing Web Pages. In: Proceedings of the 17th Conference on Hypertext and Hypermedia, HYPERTEXT '06, pp. 145–156. ACM, USA (2006)
27. Horrigan, J.: Broadband adoption and use in America. Federal Commun. Comm. (2010)
28. Kahle, B.: Preserving the Internet. Sci. Am. **276**(3), 82–83 (1997)
29. Kahle, B.: Fire update: lost many cameras, 20 boxes. No one hurt. https://blog.archive.org/2013/11/06/scanning-center-fire-please-help-rebuild/ (2013)
30. Kahle, B.: Reader privacy at the Internet Archive. http://blog.archive.org/2013/10/25/reader-privacy-at-the-internet-archive/ (2013)
31. Kahle, B.: Wayback Machine: now with 240,000,000,000 URLs. http://blog.archive.org/2013/01/09/updated-wayback/ (2013)
32. Kemper, E.A., Stringfield, S., Teddlie, C.: Mixed methods sampling strategies in social science research. Handbook of mixed methods in social and behavioral research, pp. 273–296 (2003)
33. Koehler, W.: Web page change and persistence-a four-year longitudinal study. J. Am. Soc. Inf. Sci. Technol. **53**(2), 162–171 (2002)
34. Kramer-Smyth, J., Nishigaki, M., Anglade, T.: ArchivesZ: Visualizing archival collections. http://archivesz.com/ArchivesZ.pdf (2007)
35. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A.: Circos: an information aesthetic for comparative genomics. Genome Res. **19**(9), 1639–1645 (2009)
36. Kumar, R., Tomkins, A.: A characterization of online browsing behavior. In: Proceedings of the 19th International World Wide Web Conference, WWW '10, pp. 561–570. ACM, USA (2010)
37. Liu, H., Kešelj, V.: Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data Knowl. Eng. **61**(2), 304–330 (2007)
38. Markov, Z., Larose, D.T.: Data mining the web: uncovering patterns in wWeb content, structure, and usage. Wiley, New York (2007)
39. Negulescu, K.C.: Web archiving @ the Internet Archive. In: Presentation at the 2010 Digital Preservation Partners Meeting, http://www.digitalpreservation.gov/meetings/documents/ndiipp10/NDIIPP072110FinalIA.ppt (2010)
40. Nelson, M.L., Allen, B.D.: Object persistence and availability in digital libraries. D-Lib Mag. 8(1) (2002).10.1045/january2002-nelson
41. Nithya, P., Sumathi, P.: Novel pre-processing technique for web log mining by removing global noise, cookies and web robots. Int. J. Comput. Appl. **53**(17), 1–6 (2012)
42. Omodei, M.: Trends in use of Pandora Archive. International Internet Preservation Consortium. http://netpreserve.org/sites/default/files/resources/IIPC-GA-NLA-presentation_m.pdf (2012)

43. Padia, K., AlNoamany, Y., Weigle, M.C.: Visualizing digital collections at Archive-It. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, pp. 15–18. ACM, USA (2012)
44. Reddy, K.S., Varma, G.P.S., Babu, I.R.: Preprocessing the Web Server Logs: an illustrative approach for effective usage mining. ACM SIGSOFT Softw. Eng. Notes **37**(3), 1–5 (2012)
45. Reisinger, D.: Netflix gobbles a third of peak Internet traffic in North America. CNET, http://news.cnet.com/8301-1023_3-57546405-93/netflix-gobbles-a-third-of-peak-internet-traffic-in-north-america/ (2012)
46. Rossi, A.: Fixing broken links on the Internet | Internet Archive Blogs. https://blog.archive.org/2013/10/25/fixing-broken-links/ (2013)
47. SalahEldeen, H.M., Nelson, M.L.: Carbon dating the web: estimating the age of web resources. In: Proceedings of 3rd Temporal Web Analytics Workshop, TempWeb '13, pp. 1075–1082 (2013)
48. Sanderson, R., Phillips, M., Van de Sompel, H.: Analyzing the Persistence of referenced web resources with Memento. Tech. Rep. arXiv:1105.3459 (2011)
49. Shuyo, N.: Language Detection Library - 99% over precision for 49 languages. http://www.slideshare.net/shuyo/language-detection-library-for-java (2010)
50. Shuyo, N.: Language Detection Library for Java. http://code.google.com/p/language-detection/ (2012)
51. Silva, A.J.C., Gonçalves, M.A., Laender, A.H.F., Modesto, M.A.B., Cristo, M., Ziviani, N.: Finding what is missing from a digital library: a case study in the computer science field. Inf. Process. Manage. **45**(3), 380–391 (2009)
52. Smith, A.: Home broadband 2010. Technical report, Pew Internet & American Life Project, An initiative of the Pew Research Center (2010)
53. Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M.: A framework for the evaluation of session reconstruction heuristics in web-usage analysis. INFORMS J. Comput. **15**(2), 171–190 (2003)
54. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explor. Newslett. **1**(2), 12 (2000). doi:10.1145/846183.846188
55. Stassopoulou, A., Dikaiakos, M.D.: Web robot detection: a probabilistic reasoning approach. Comput. Netw. **53**(3), 265–278 (2009)
56. Streitfeld, A.: Internet Archive will shield visitors - NYTimes.com. http://bits.blogs.nytimes.com/2013/10/24/internet-archive-will-shield-visitors/ (2013)
57. Tan, P.N., Kumar, V.: Discovery of web robot sessions based on their navigational patterns. Data Min. Knowl. Discov. **6**(1), 9–35 (2002)
58. Tanasa, D., Trousse, B.: Advanced data preprocessing for intersites Web usage mining. IEEE Intell. Syst. **19**(2), 59–65 (2004)
59. Teddlie, C., Yu, F.: Mixed methods sampling: a typology with examples. J. Mixed Methods Res. **1**(1), 77–100 (2007)
60. Thelwall, M., Vaughan, L.: A fair history of the Web? Examining country balance in the Internet Archive. Libr. Inf. Sci. Res. **26**(2), 162–176 (2004)
61. Tofel, B.: Wayback for accessing Web Archives. In: Proceedings of International Web Archiving Workshop, IWAW (2007)
62. Tongco, M., Dolores, C.: Purposive sampling as a tool for informant selection. Ethnobot. Res. Appl. **5**, 147–158 (2008)
63. Van de Sompel, H., Nelson, M.L., Sanderson, R.: RFC 7089—HTTP framework for time-based access to resource states—Memento. http://tools.ietf.org/html/rfc7089 (2013)
64. Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: Time Travel for the Web. Tech. Rep. arXiv:0911.1112 (2009)
65. Van Ryzin, G.G.: Cluster analysis as a basis for purposive sampling of projects in case study evaluations. Am. J. Eval. **16**(2), 109–119 (1995)
66. Wasserman, T.: Netflix takes up 32.7% of Internet bandwidth. Mashable, http://mashable.com/2011/10/27/netflix-takes-up-32-7-of-internet-bandwidth-study/ (2011)
67. Whitelaw, M.: Exploring archival collections with interactive visualisation. In: Proceedings of E-Research Australasia Conference (2009)
68. Zhuang, Z., Wagle, R., Giles, C.: What's there and what's not? Focused crawling for missing documents in digital libraries. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '05, pp. 301–310 (2005)