Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1123–1130
DOI: 10.15439/2014F346
ACSIS, Vol. 2

Analysis of Aggregated Bot and Human Traffic on E-Commerce Site

Grażyna Suchacka
Opole University
ul. Oleska 48,
45-052 Opole, Poland
Email: gsuchacka@uni.opole.pl

***Abstract*—A significant volume of Web traffic nowadays can be attributed to robots. Although some of them, e.g., search-engine crawlers, perform useful tasks on a website, others may be malicious and should be banned. Consequently, there is a growing need to identify bots and to characterize their behavior. This paper investigates the share of bot-generated traffic on an e-commerce site and studies differences in bots' and humans' session-based traffic by analyzing data recorded in Web server log files. Results show that both kinds of sessions reveal different characteristics, including the session duration, the number of pages visited in session, the number of requests, the volume of data transferred, the mean time per page, the number of images per page, and the percentage of pages with unassigned referrers.**

I. Introduction

ALONG with the growing popularity of search engines and other Web-based applications there has been the growing need to develop advanced tools for retrieving information on the Web content, structure, and usage. Such tools are Web bots (also called Web robots, spiders, or crawlers). They can traverse the Web autonomously by following the structure of hyperlinks, collect different kinds of information, and perform specific tasks on websites.

The most common bots are search engine crawlers, which visit Web pages on a regular basis to build and maintain huge search indexes [1], [2]. Popular bots visiting e-commerce sites are shopping bots, which collect information on products in various Web stores on behalf of product search engines or price comparison services. SEO spybots and content scrapers can expeditiously scrape from websites large amounts of information which may be valuable for SEO professionals or competitive e-business companies [3]. Other examples of robots include resource archivers, link checkers, e-mail harvesters, chat bots [4], spambots [5], hacking bots, artificial actors in e-dating [6], or automatic online game players [7].

Robot traffic on a website should be identified and sometimes also banned for several reasons. The most obvious ones are connected with potential threats of malicious bot activities [8], [9], [10]. Bot-generated click frauds in pay-per-click advertising result in higher fees paid by the advertisers [11]. Content-stealing bots may gather valuable business intelligence knowledge from websites and thus indirectly harm e-business competitiveness. High bot activity may also negatively affect the position of a website in search-engine rankings. Besides, bots consume network bandwidth and server resources; thus, they may cause degradation of the server performance and the quality of service offered to human users. Especially dangerous are automated DoS (Denial of Service) attacks, which may even make the server stall or crash [9]. Lastly, identification of robot traffic is essential when analyzing behavior of human users, who are characterized by different navigational patterns than bots [1], [8], [9], [12], [13], [14], [15].

Although some studies have addressed the problem of robot traffic characterization based on Web server logs, very little research has been done for e-commerce sites ([16], [17], [18]). Our study aims to partially fill this gap by comparing key characteristics of bot- and human-generated traffic on a Web bookstore site. This issue is crucial for e-commerce sites where human users are potential buyers and their activity on the site is directly related to the profitability of the online store.

The paper is organized as follows. Section II presents Web server log data underlying our research and the research methodology. Section III discusses the share of bots in the analyzed Web traffic, whereas Section IV compares key characteristics of bot and human sessions. Section V concludes the paper and suggests prospective future work.

II. Research Methodology

A. Web Server Log Data Description

When an Internet user visits a website, their Web browser (which is a Web client, in fact) communicates via the HTTP protocol with the server hosting the site. For each Web page requested by the user, their client typically issues a series of HTTP requests to the server: one request for a page description file and the following requests for objects embedded in the page, such as images or video files. After

978-83-60810-58-3/$25.00

receiving HTTP responses the client assembles the page and displays it in a browser window. A Web client may represent not only a human user but it may also be a computer program, i.e. a Web bot.

Data concerning each incoming HTTP request is recorded in the access log file stored at the Web server. That data includes some client data (a client IP address, a client identifier, a user agent field, a user identifier, a referrer field), the requested resource data (an URI identifying the requested server resource, a transfer size), the HTTP-related data (a method, a protocol version, a status code), and a timestamp. As an example, let us consider the following log entry, representing one HTTP request:

```
66.249.66.52 - - [03/Dec/2013:08:55:59 +0100] "GET shopping/images/pict21.jpg HTTP/1.1" 200 242 "-" "Mozilla/5.0 (compatible;Googlebot/2.1;+http://www.google.com/bot.html)".
```

This line describes a request sent by a Web client with the IP address 66.249.66.52, whose user identifier is not available. The request was served 3 December 2013 at 8:55:59 (according to Central European Time) and it concerned downloading (by using the GET method) an image file identified by URI "shopping/images/pict21.jpg". The request was successfully served (a status code is 200) and the server sent to the client 242 bytes in response. A referrer field is unassigned. The client was Mozilla 5.0 which used the protocol HTTP/1.1. One can notice that the user was not a human but Google's web crawling bot (the user agent field contains the bot's name, "Googlebot").

Our analysis was based on access logs for an online store (the store name is not given in the paper due to a non-disclosure agreement). The data covered the period of one month, December 2013.

A dedicated computer program was used to read, preprocess, clean, and analyze the data. The program was implemented in C++ using MS Visual Studio. Its most important modules include:

- Input/Output Module containing functions for reading raw data from the input log files and saving the results to the output files;
- Basic Functions Module with functions for parsing each HTTP request's line in order to distinguish individual data describing the request and transform it to the format suitable for the analysis;
- Request Module for managing and processing HTTP requests, e.g., checking whether a request was generated by bot;
- Session Module for reconstructing and processing user sessions;
- Robot Module for identifying and processing sessions generated by bots;
- Statistics Module containing functions for computing all the necessary statistics;
- other modules implementing the operation of visual forms.

B. Reconstruction and Characterization of User Sessions

Based on HTTP requests user sessions were reconstructed. A `user session` means a sequence of requests issued by a Web client during the single visit to the Web store. Each individual user was identified based on two data fields describing HTTP requests: the client IP address and the user agent field. Consecutive user sessions were reconstructed based on the requests' timestamps, assuming a minimum 30-minute interval between two subsequent sessions of a given user (the value of 30 minutes has been commonly applied in previous Web traffic analyses, e.g. in [9], [19]).

Afterwards, each user session was described with a number of attributes:

- session length – the number of pages visited in session;
- session duration – time interval (in seconds) between the times of the last and the first requests in session (session duration is shorter than the actual time of the user-site interaction because the time of browsing the last page in session by the user in unknown at the server side; for the same reason this attribute cannot be determined for sessions containing only one page);
- mean time per page – the average time (in seconds) the user browsed a single page in session (this attribute may be derived only for sessions containing more than one page);
- volume of data transferred to the Web client (in MB);
- number of HTTP requests;
- image-to-page ratio – the average number of image file requests over the number of page requests in session;
- percentage of pages with unassigned referrers – the percentage of page requests with unassigned or blank referrer fields;
- percentage of requests with unassigned referrers – the percentage of HTTP requests with unassigned or blank referrer fields;
- percentage of requests of type HEAD – the percentage of HTTP requests with HEAD method;
- percentage of 4xx responses – the percentage of erroneous HTTP requests in session (i.e. requests with status codes starting with "4").

We decided to compute the aforementioned attributes because some previous user session analyses for non-e-commerce environments reported that these session features may be useful in distinguishing Web robots from human users [1], [8], [9].

Some sessions contained no Web page request and only one request for an image file (such a situation is often connected with displaying a banner advertisement of the store on another Web page). As these sessions cannot be regarded as intended visits to the store, we did not take them into consideration in our analysis.

C. Identification of Bot Sessions

There are a few ways to identify at least some part of user sessions issued by Web robots.

First, one should check if the file "robots.txt" was accessed in a session. Cooperative robots should request this file at the beginning of each visit to a site in order to read which parts of the site they can access.

Second, "ethical" bots should inform a Web server about their identities via their user agent fields, containing the name of the robot. We implemented a function verifying HTTP requests' user agent fields for compliance with user agents of known robots, available on online databases [20] and [21]. Moreover, some robots not included in these databases were identified based on keywords contained in user agent fields ("bot", "spider", "crawler", "worm", "search", "track", "harvest", "dig", "hack", "trap", "archive", or "scrap"), as well as through a semi-automatic inspection of user agent fields.

In practice, not all robots access the file "robots.txt" or declare their identities in user agent fields. However, some of such bots may be still identified based on the character of their interaction with the site, which proceeds differently from the interaction of human users. Humans usually communicate with the site via the Web interface and follow navigation paths according to the site topology. Each Web page request is typically followed by a group of requests for embedded objects (usually images). Moreover, the successive page requests are separated with some time intervals called "user think times". In contrast, robots tend to reveal navigational patterns incompatible with the site topology and have unintuitive session characteristics, e.g., the extremely low mean time per page. We assumed the following three groups of session characteristics that indicate Web robots:

- the mean time per page shorter than 0.5 second;
- an unassigned referrer field in the first request in session, the percentage of pages or requests with unassigned referrers equal to 100, and the percentage of requests of type HEAD equal to 100;
- an unassigned referrer field in the first request in session, the percentage of pages or requests with unassigned referrers equal to 100, the percentage of 4xx responses equal to 100, and the image-to-page ratio equal to 0.

All sessions which were not classified as performed by robots, after excluding sessions connected with executing administrative tasks on the site, were assumed to be performed by human users.

III. Bot Share in Overall Web Server Traffic

According to the results of earlier analyses for e-business workloads the share of robot requests has differed from several (3.2% in [17]) to a dozen or so (15% in [22], 16% in [23]) percent. In our data set 22.3% of all HTTP requests were identified as generated by bots (Fig. 1). However, as regards the number of user sessions, as many as 79.3% were performed by bots. Bot sessions seem to contain on average less requests and consume less server resources than human sessions (the volume of data transferred to bot clients comprised 38% of the overall data transfer).

In regards to known bots, possible to recognize by checking requests' user agent fields, 76 different robots were identified. The most active of them were popular search engine crawlers (Bingbot, MJ12bot, Googlebot, Google AdsBot, Yandex bot, MSNBot, Baidu spider). Large part of bot traffic was also generated by SEO and e-commerce crawlers (AhrefsBot, ShopWiki, WillyBot), link checkers (SEOkicks robot, SpBot), and social media agent FacebookExternalHit.

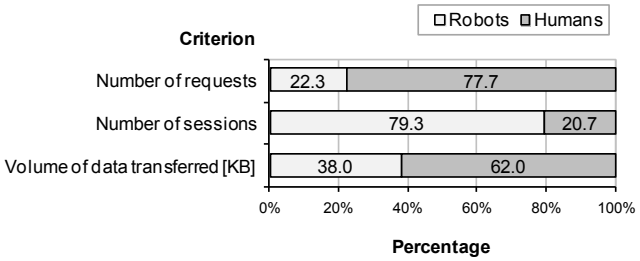

Fig. 1 Percentage breakdown of robots' and humans' data

It is worth noting that only 11% of all bot sessions (including 22.7% of known bots' sessions) accessed the file "robots.txt".

IV. Comparison of Bot and Human Sessions

An insight into the session characteristics revealed that a large number of all user sessions contained only one page and/or lasted only one second. One of the justifications of such user behavior may be that some users were referred to the store site by following a search engine link or by clicking a store advertisement placed on another site but the content of the store website was not what they had searched for. Such users left the site immediately after entering it. We decided to exclude from the statistical analysis sessions containing only one page and sessions lasting only one second.

Furthermore, using the graphical method, we excluded from the analysis a few outlying sessions which were extremely long and lasted for an extremely long time compared to other sessions. The outliers were two human sessions (containing 202 and 312 pages) and eight bot sessions (containing from 1 453 to 6 029 pages and/or lasting longer than 48 hours) – the reason for the exclusion of these several sessions was that they strongly distorted the statistical results for all human and bot sessions, respectively. Finally, only 27% of sessions were left in our dataset.

A. Session Length

An important aspect of user session characterization in the context of distinguishing bots from humans is the session length in the number of pages visited in session. Intuitively, there is some upper limit on the maximum number of human user's clicks, i.e. the number of pages a human user can open and browse during a single visit to a website. This limitation does not apply to automatic computer programs, such as bots, which are able to automatically traverse all pages belonging to a site in a relatively short time. For the same reasons the maximum time of a human-website interaction is limited as well, so bot-generated sessions tend to last much longer than human ones.

Our results achieved for the e-commerce site confirm these observations. Session length statistics presented in Table I show that robots requested on average above four times more pages in session than humans and the maximum session length was an order of magnitude higher for bots than for humans. (Taking into consideration the outlying sessions excluded from the analysis one can notice that the longest human session contained 312 pages whereas the longest bot session contained as many as 6 029 pages.) However, session length distributions presented in Fig. 2 are similar for robots and humans: both histograms illustrate a strong right-skew of session length distribution. Over 95% of human users opened less than 26 pages during their visits in the store and 97% of bots requested less than 176 pages.

B. Session Duration

Session duration statistics presented in Table II show that Web bots tend to spend much more time on the website than human users. The mean session duration is fifteen times longer for bots than for humans. The maximum session duration for bots is 39 hours (and up to 181 hours taking into consideration the excluded outlying bot sessions!) whereas for humans it is less than two hours. Distributions of session durations, shown in Fig. 3, are very similar to the distributions of session lengths in Fig. 2. For both kinds of sessions they are strongly right-skewed and heavy-tailed.

Intuitively, bigger numbers of pages in session should correspond to longer session durations so it is worth graphically examining this relationship. Fig. 4 presents a two-dimensional scatter plot of the session duration against the session length (to improve the graph readability robot sessions with lengths exceeding 800 pages were not shown in the figure). One can see the correlation between the number of pages visited in session and the duration of a visit for both kinds of session: as the session length increases, the session duration tends to increase as well. Human sessions form a quite well-knit group in the two-dimensional area whereas robot sessions are rather dispersed and seem to form a few (at least five) separate clusters. Fig. 4 suggests that different kinds of bots may reveal different behavior so it would make sense to separately characterize behavior of various bots (search engine crawlers, image indexers, link checkers, e-

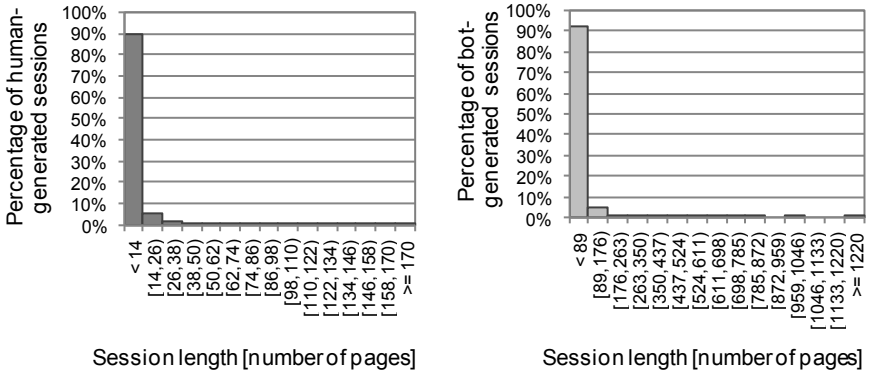

Fig. 2 Histogram of session lengths: (left) for humans, (right) for bots

TABLE I.
SESSION LENGTH STATISTICS
(IN NUMBER OF PAGES VISITED)

Statistics	Humans	Bots
Mean	7.2	30.1
Median	3	8
Mode	2	2
Std. dev.	13.1	73.7
Minimum	2	2
Maximum	173	1 253

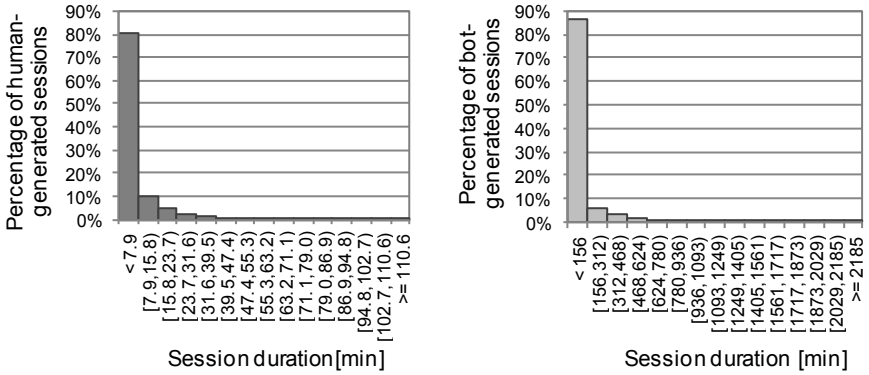

Fig. 3 Histogram of session durations: (left) for humans, (right) for bots

TABLE II.
SESSION DURATION STATISTICS
(IN MINUTES)

Statistics	Humans	Bots
Mean	5.4	82.6
Median	1.4	11.4
Mode	0.05	0.03
Std. dev.	10.2	213
Minimum	0.03	0.03
Maximum	1 18.4	2 341

mail collectors, etc.). One could also apply classification or clustering methods to determine classes or clusters of robot sessions. We leave these issues for our future work.

C. Mean Time per Page

Based on the session length and the session duration one can determine the mean time per page for each user session containing more than one page. The mean time per page in session was computed according to the formula:

$$\bar{d}_s = \frac{d_s}{l_s - 1} \quad (1)$$

where d_s is the session duration (in seconds) and l_s is the session length (in number of pages), $l_s > 1$. Unlike the session duration, which does not include the time for the last page visited in session, the mean time per page is not underrepresented as it is computed for all visited pages except one.

Mean time per page statistics, presented in Table III, show significant differences between bot and human sessions. It may be surprising that bots spend more time analyzing Web pages than human users and the average is equal to as much as 5.3 minutes. However, the median equal to 1.9 minute and the mode equal to 2 seconds are much lower. Besides, a relatively high value of the standard deviation, 8 minutes, indicates that the mean time per page is rather differentiated for bots. Histograms in Fig. 5 also show a bigger differentiation of mean times per page for bots than for humans. For robots the distribution of mean times per page is not so strongly right-skewed and does not include such a long tail as for human users. These results also indicate that it may be worth performing the statistical characterization of various kinds of robots visiting the Web store site.

D. Image-to-Page Ratio

Some previous studies reported that robots (especially crawlers) request mostly Web page files and ignore image files [8], [9], [24]. In contrast, human users navigate through the website following the structure of hyperlinks and when they open a new page, they usually download the page description file along with image files embedded in the page. Hence, such metrics as image-to-page ratio or percentage of image requests in session belong to strongly distinguishable characteristics between bots and humans. Image-to-page ratio statistics in Table IV, as well as histograms in Fig. 6 confirm these results. Humans request more than 22 images per page; the median is equal to 19 and the mode is equal to 36. In contrast, robot requests for image files are negligible: the mean number of images per page is only 0.3 and what is more, both the median and the mode are equal to 0. Among all robot sessions almost 50% did not request any image at all. Surprisingly, 0.6% of human sessions also contained no image request (it may indicate that some sessions considered as generated by humans were performed by bots, in fact).

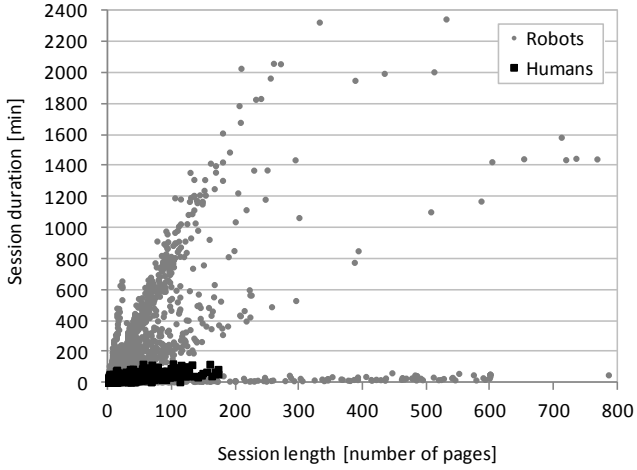

Fig. 4 Scatter plot of session duration vs. session length for robots and humans

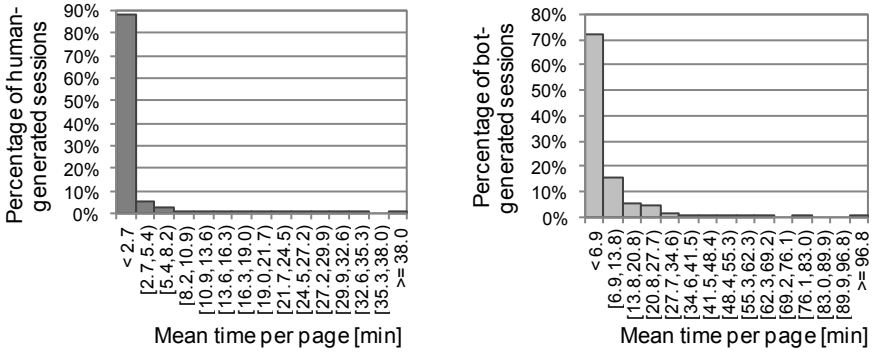

Fig. 5 Histogram of mean times per page: (left) for humans, (right) for bots

TABLE III.
MEAN TIME PER PAGE STATISTICS
(IN MINUTES)

Statistics	Humans	Bots
Mean	1.4	5.3
Median	0.4	1.9
Mode	0.1	0.03
Std. dev.	3.2	8.0
Minimum	0.01	0.001
Maximum	40.6	103.6

E. Volume of Data Transferred to Web Clients

We decided to examine if large numbers of image requests correspond to large volumes of data transferred to Web clients, i.e. whether data transfers are bigger for humans than for bots. Typically, a significant part of Web traffic concerns transmitting small files and messages (e.g. messages that the requested resource has not been modified or that resource could not be found on the server). In contrast, graphical and multimedia Web resources are relatively big files.

As can be seen in Table V, volumes of data transferred to human users tend to be much bigger than for robots. Although this metric in both cases ranges from 0 to over 14, distributions of data transfer volumes are quite different (Fig. 7). For robot sessions the histogram is extremely heavy-tailed. Average bot data transfer is 227 KB, however the median transfer is only 65 KB and the mode is merely 1 KB. For 98% of bot sessions the transmitted data did not exceed 1.7 MB. On the contrary, average data transfer for humans is 1.2 MB, the median is 765 KB and the mode is 263 KB. Data sent in 98% of human sessions were up to 5.1 MB.

It is interesting to observe that for both kinds of sessions the distributions of data transfer volumes (Fig. 7) correspond very accurately to the distributions of the numbers of HTTP requests in session (Fig. 8). Also the range of the numbers of HTTP requests in session is almost the same for bots and humans (Table VI).

F. Percentage of Pages with Unassigned Referrers

Users may reach the store website in many different ways, e.g. by following a search engine link (one of organic search engine results or sponsored links), or by clicking a banner ad on another website. In such cases an address of the referring

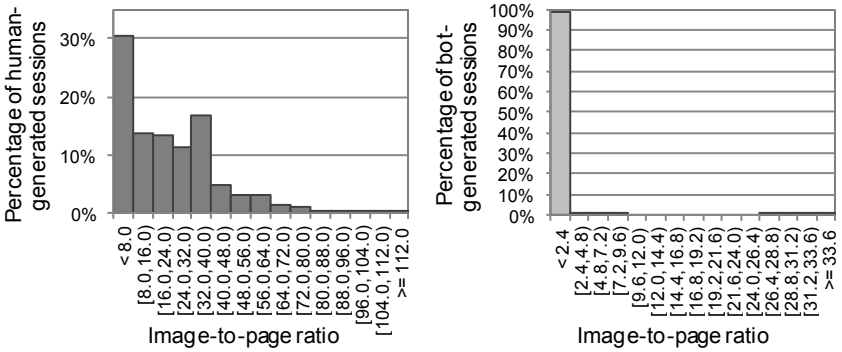

Fig. 6 Histogram of image-to-page ratios: (left) for humans, (right) for bots

TABLE IV.
IMAGE-TO-PAGE RATIO STATISTICS

Statistics	Humans	Bots
Mean	22.4	0.3
Median	19.1	0
Mode	36	0
Std. dev.	19	1.8
Minimum	0	0
Maximum	119.5	35.5

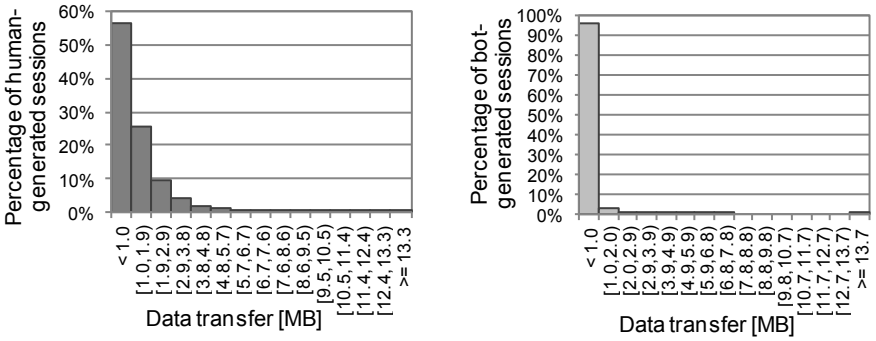

Fig. 7 Histogram of data transfer volumes: (left) for humans, (right) for bots

TABLE V.
DATA TRANSFER VOLUME STATISTICS (IN MEGABYTES)

Statistics	Humans	Bots
Mean	1.2	0.2
Median	0.7	0.1
Mode	0.3	0.001
Std. dev.	1.3	0.6
Minimum	0	0
Maximum	14.3	14.6

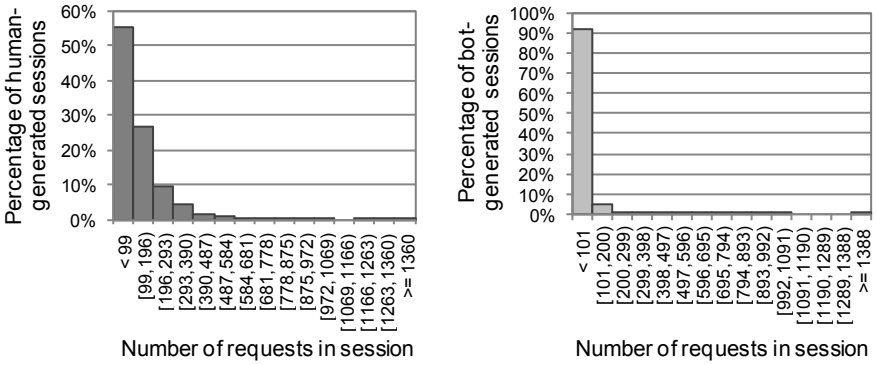

Fig. 8 Histogram of the numbers of HTTP requests in session: (left) for humans, (right) for bots

TABLE VI.
NUMBER OF REQUESTS IN SESSION STATISTICS

Statistics	Humans	Bots
Mean	120.9	35.8
Median	80	10
Mode	76	2
Std. dev.	130.4	80.8
Minimum	2	2
Maximum	1 450	1 484

page is contained in the referrer field of the first HTTP request in session. Sometimes this field may be empty, e.g. when a user enters the site directly by typing the site address in a browser's address bar or clicking on a bookmarked page. However, as a human user navigates through the website, each newly opened Web page request will contain in its referrer field the address of the previously browsed page. On the contrary, the vast majority of Web robots initiate their sessions (or even all HTTP requests in session) with unassigned referrer fields, so it may be a good indicator of a bot-generated session. This was confirmed in previous Web characterization studies, e.g. in [10] and [25].

We computed the percentage of page requests with unassigned referrer fields in each session. Our results, presented in Fig. 9 and Table VIII, are similar to observations reported in earlier studies. In fact, 98.4% of all robot sessions had all pages with unassigned referrers. For comparison, among human sessions there were only 2.1% of such sessions (it is very likely that they were actually unidentified bot-generated sessions).

G. Percentage of Requests of Type HEAD

The most common HTTP method is GET, which is used do download contents from Web servers. By default, when a human user browses a website via a browser, requests sent to the server by the browser will be of type GET. Other possible HTTP method is HEAD, used to retrieve only Web metadata. In contrast to humans, robots are expected to use HEAD method instead of GET when possible (e.g. to download only recently updated contents) in order to reduce the amount of data downloaded from servers and to minimize the consumption of server resources.

Some previous workload characterization studies showed that the percentage of requests of type HEAD is higher for bots than for humans [25]. Other studies reported that nearly all crawler requests were of type GET [24]. Our results signalize an advantage of bots over humans in this respect, however the percentage of requests of type HEAD for bot sessions was not very high (Table VII). Only 0.4% of bot sessions had some requests of type HEAD, compared to 0.1% of human sessions. After taking into consideration all robot sessions (even those containing only one request, excluded from our statistical analysis), the mean percentage of HEAD requests for bots increases to 0.8 and 0.9% of bots sessions contain only HEAD requests.

TABLE VII.
PERCENTAGE OF REQUESTS OF TYPE HEAD STATISTICS

Statistics	Humans	Bots
Mean	0.004	0.2
Median	0	0
Mode	0	0
Std. dev.	0.15	4.1
Minimum	0	0
Maximum	8.3	100

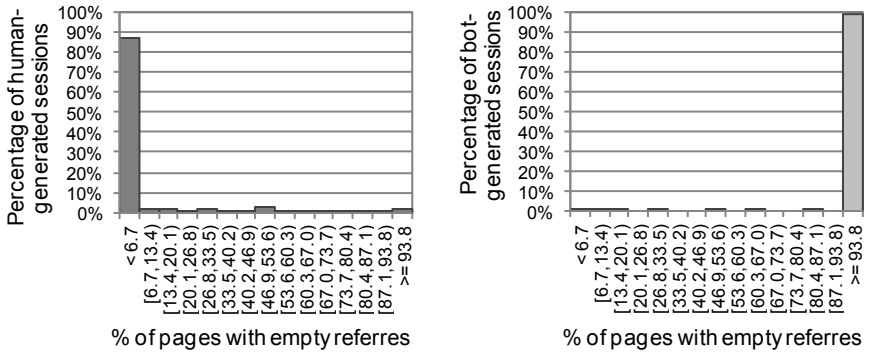

TABLE VIII.
PERCENTAGE OF PAGES WITH UNASSIGNED REFERRERS STATISTICS

Statistics	Humans	Bots
Mean	5.9	99
Median	0	100
Mode	0	100
Std. dev.	18.4	9.8
Minimum	0	0
Maximum	100	100

Fig. 9 Histogram of percentages of pages with unassigned referrer fields: (left) for humans, (right) for bots

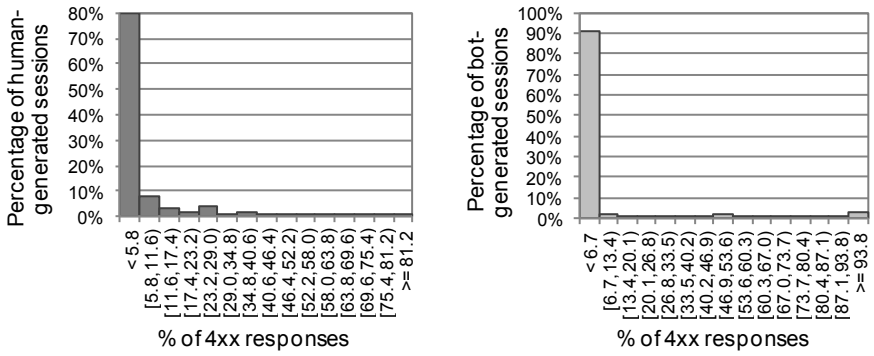

TABLE IX.
PERCENTAGE OF 4XX RESPONSES STATISTICS

Statistics	Humans	Bots
Mean	5	4.9
Median	1.3	0
Mode	0	0
Std. dev.	9	18.6
Minimum	0	0
Maximum	87	100

Fig. 10 Histogram of percentages of 4xx responses: (left) for humans, (right) for bots

H. Percentage of 4xx Responses

Web robots tend to have a higher rate of erroneous requests (i.e. requests with status codes of type 4xx), because it is more likely that they request outdated or deleted files [1], [24], [25]. However, we observed that for our data set the rate of erroneous requests was a little bit higher for human sessions (the mean equal to 5) than for bot ones (the mean equal to 4.9) (Table IX). For robots the mean is a bit lower but more variable. Moreover, 2.8% of bot sessions had 100% of erroneous responses (compared to 0% of such human sessions). After taking into consideration also sessions containing only one request, the statistics for humans increase insignificantly whereas for bots they increase notably: the mean is equal to 10.8, the standard deviation is equal to 22.5, and 3.8% of bot sessions have 100% of erroneous responses.

V. Conclusion

The paper discusses key characteristics of sessions realized by Web robots and human users on the e-commerce site. Our results confirm some earlier findings of Web workload analyses, concerning differences between bots and humans in the following session characteristics: the session duration, the number of pages visited in session, the number of HTTP requests in session, the volume of data transferred, the percentage of pages with unassigned referrers, and the number of images per page. However, such characteristics as the percentage of requests of type HEAD and the percentage of erroneous responses turned out not to be as good indicators of bot sessions as reported in previous studies.

The analysis was done for the aggregated Web robot traffic. However, our observations suggest that the behavior of bots is not homogenous and various kinds of bots may reveal different navigational patterns. In our future work we plan to address this issue. We also plan to extend our research to another Web stores of different sizes and branches to verify the reliability of our conclusions for other e-commerce scenarios. Our findings may be applied in classification and segmentation methods aiming at identifying sessions of unknown bots on the e-commerce website.

References

[1] A. Balla, A. Stassopoulou, M. D. Dikaiakos, "Real-time Web crawler detection," in Proc. 18th ICT, Ayia Napa, Cyprus, 2011, pp. 428-432, http://dx.doi.org/10.1109/CTS.2011.5898963.

[2] H. Kang, K. Wang, D. Soukal, F. Behr, Z. Zheng, "Large-scale bot detection for search engines," in Proc. 19th WWW, Raleigh, NC, USA, 2010, pp. 501-510, http://dx.doi.org/10.1145/1772690.1772742.

[3] N. Poggi, J. L. Berral, T. Moreno, R. Gavaldà, J. Torres, "Automatic detection and banning of content stealing bots for e-commerce", in Proc. NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security, British Columbia, Canada, 2007, pp. 7-8.

[4] S. Gianvecchio, M. Xie, Z. Wu, H. Wang, "Humans and bots in Internet chat: measurement, analysis, and automated classification," IEEE/ACM Trans. Netw. 19(5), 2011, pp. 1557-1571, http://dx.doi.org/10.1109/TNET.2011.2126591.

[5] P. Hayati, V. Potdar, K. Chai, A. Talevski, "Web spambot detection based on Web navigation behaviour," in Proc. AINA, Perth, 2010, pp. 797-803, http://dx.doi.org/10.1109/AINA.2010.92.

[6] A. Schmitz, O. Yanenko, M. Hebing, "Identifying artificial actors in e-dating: a probabilistic segmentation based on interactional pattern analysis," Challenges at the Interface of Data Analysis, Computer Science, and Optimization - Studies in Classification, Data Analysis, and Knowledge Organization, 2012, pp 319-327, http://dx.doi.org/10.1007/978-3-642-24466-7_33.

[7] A. R. Kang, H. K. Kim, J. Woo, "Chatting pattern based game BOT detection: do they talk like us?," KSII TIIS 6(11), 2012, pp. 2866-2879.

[8] A. Stassopoulou, M. D. Dikaiakos, "Web robot detection: a probabilistic reasoning approach," Comput. Netw. 53(3), 2009, pp. 265-278, http://dx.doi.org/10.1016/j.comnet.2008.09.021.

[9] D. Stevanovic, N. Vlajic, A. An, "Unsupervised clustering of Web sessions to detect malicious and non-malicious website users," Procedia Computer Science 5, Elsevier, 2011, pp. 123–131, http://dx.doi.org/10.1016/j.procs.2011.07.018.

[10] P.-N. Tan, V. Kumar, "Discovery of web robot sessions based on their navigational patterns," Data Min. Knowl. Discov. 6 (1), 2002, pp. 9–35, http://dx.doi.org/10.1023/A:1013228602957.

[11] K. Springborn, P. Barford, "Impression fraud in online advertising via pay-per-view networks," in Proc. 22nd USENIX Conference on Security, Washington, D.C., 2013, pp. 211-226.

[12] S. Kwon, M. Oh, D. Kim, J. Lee, Y.-G. Kim, S. Cha, "Web robot detection based on monotonous behavior," ASTL 4, Springer-Verlag, 2012, pp. 43-48.

[13] C. H. Saputra, E. Adi, S. Revina, "Comparison of classification algorithms to tell bots and humans apart," JNIT 4(7), 2013, pp. 23-32.

[14] G. Suchacka, G. Chodak, "Practical aspects of log file analysis for e-commerce," CCIS 370, Springer, 2013, pp. 562-572, http://dx.doi.org/10.1007/978-3-642-38865-1_56.

[15] G. Suchacka, "Statistical analysis of buying and non-buying user sessions in a Web store," Information Systems Architecture and Technology – Network Architecture and Applications, Wroclaw, Poland, 2013, pp. 163-172.

[16] V. Almeida, D. Menascé, R. Riedi, F. Peligrinelli, R. Fonseca, W. Meira Jr., "Analyzing robot behavior in e-business sites," in Proc. ACM SIGMETRICS, Cambridge, MA, USA, 2001, pp. 338–339, http://dx.doi.org/10.1145/378420.378838.

[17] D. Doran, S. S. Gokhale, "Long range dependence (LRD) in the arrival process of Web robots," in Proc. ICCTS, New Delhi, India, 2012, pp. 176-180, http://dx.doi.org/10.7763/IPCSIT.2012.V47.33.

[18] N. Poggi, J. L. Berral, T. Moreno, R. Gavaldà, J. Torres, "Automatic detection and banning of content stealing bots for e-commerce," in Workshop on Machine Learning in Adversarial Environments for Computer Security, British Columbia, Canada, 2007.

[19] D. Doran, S. S. Gokhale, "Searching for heavy tails in Web robot traffic," in Proc. 7th Int. Conf. QEST, Williamsburg, Virginia, USA, 2010, pp. 282-291, http://dx.doi.org/10.1109/QEST.2010.42.

[20] List of user-agents (spiders, robots, crawler, browser), http://www.user-agents.org.

[21] List of user agent strings – Robots (crawlers), http://user-agent-string.info/list-of-ua/bots.

[22] N. Poggi, D. Carrera, R. Gavaldà, E. Ayguadé, J. Torres, "A methodology for the evaluation of high response time on e-commerce users and sales," Inform. Syst. Front., 2012, http://dx.doi.org/10.1007/s10796-012-9387-4.

[23] D. A. Menascé, V. Almeida, R. H. Riedi, F. Ribeiro, R. C. Fonseca, W. Meira Jr., "In search of invariants for e-business workloads," in Proc. 2nd ACM-EC Conf., Minneapolis, MN, USA, 2000, pp. 56-65, http://dx.doi.org/10.1145/352871.352878.

[24] M. D. Dikaiakos, A. Stassopoulou, L. Papageorgiou, "An investigation of Web crawler behavior: characterization and metrics," Comput. Commun. 28(8), 2005, pp. 880-897, http://dx.doi.org/10.1016/j.comcom.2005.01.003.

[25] C. Bomhardt, W. Gaul, L. Schmidt-Thieme, "Web robot detection - preprocessing Web logfiles for robot detection," New Developments in Classification and Data Analysis - Studies in Classification, Data Analysis, and Knowledge Organization, 2005, pp. 113-124, http://dx.doi.org/10.1007/3-540-27373-5_14.