EFFICIENCY ANALYSIS OF RESOURCE REQUEST PATTERNS IN CLASSIFICATION OF WEB ROBOTS AND HUMANS

Grażyna Suchacka
Institute of Mathematics and Informatics
Opole University
ul. Oleska 48
45-052 Opole, Poland
E-mail: gsuchacka@uni.opole.pl

Igor Motyka
Institute of Mathematics and Informatics
Opole University
ul. Oleska 48
45-052 Opole, Poland
E-mail: igor_motyka@mail.com

KEYWORDS
Internet Robot, Web Bot, Web Crawler, Web Server, Web Traffic, HTTP Traffic, Classification

ABSTRACT

The paper deals with the problem of classification of Web traffic generated by robots and humans on e-commerce websites. Due to the still growing proliferation and specialization of bots, a large body of research into characterization and recognition of their traffic has been conducted so far. In particular, some approaches to classify bot and human sessions on websites have been proposed in the literature. In this paper we verify and discuss the efficiency of such recently proposed approach, which uses differences in resource request patterns of bots and humans. We reconstructed Web sessions from actual HTTP log data for three different e-commerce sites, varying in the traffic intensity and proportions of bot sessions in the overall traffic. Two heuristic procedures for labeling sessions as driven by a bot or a human were proposed and implemented. Resource request patterns for both session classes, using both session labeling procedures, were analyzed and their potential to differentiate between bot and human sessions was investigated. Results show that the broader session labeling procedure allows one to capture more bot sessions and that resource requests patterns are a good discriminant of bots and humans on e-commerce sites.

INTRODUCTION

Nowadays the traffic on many Web servers is dominated by Web robots. A Web robot (bot, crawler) is an autonomous software tool that can traverse the Web using the structure of hyperlinks and carry out specific tasks on visited sites. According to the Bot Traffic Report 2016 (Zeifman 2017), bots comprise the majority of online traffic, with an average of almost 52%. About 23% of the overall traffic may be attributed to good bots (feed fetchers, search engine bots, commercial crawlers, monitoring bots) and 29% – to bad ones (impersonators, hacking tools, content scrapers, spammers). The most active bots are impersonators, i.e., bots which assume false identities to circumvent security solutions and perform attacks – typically DDoS (distributed denial of service) attacks. It is common for such bots to hide behind user agents in HTTP/s headers to present themselves as legitimate Web clients (i.e., Web browsers driven by human users).

To cope with the undesirable presence of bots on Web servers, much research into the analysis and characterization of their traffic has been conducted so far (Almeida et al. 2001; Doran et al. 2013; Suchacka 2014). The goal has been to distinguish specific bots' features and to develop efficient bot recognition strategies. In (Doran and Gokhale 2011) four types of bot recognition approaches were distinguished taking into consideration the information used and techniques applied: syntactic log analysis, traffic pattern analysis, analytical learning techniques, and Turing test systems. In general, one can distinguish bot detection methods operating offline (Doran and Gokhale 2016; Lee et al. 2009; Saputra et al. 2013; Stassopoulou and Dikaiakos 2009; Stevanovic et al. 2011; Suchacka and Sobków 2015) and online (Balla et al. 2011; Doran and Gokhale 2016).

In this paper we focus on offline bot detection at a Web server, when a decision on session classification (bot or human) is made given a description of the whole session (as a sequence of HTTP requests). The key motivation for the offline bot detection is gaining the possibility of assessing an impact of bots on server performance and security, as well as gaining an insight into properties and behavioral patterns of bot traffic. This is useful to develop online bot detection methods.

In this study we implemented and tested the offline bot classification method proposed in (Doran and Gokhale 2016). This approach classifies bot and human sessions on a Web server based on the information on resource request patterns in session. In contrast to most of other classification methods, this approach is conceptually simple and relatively easy to implement (it does not involve a time-consuming learning phase to infer traffic patterns). Doran and Gokhale (2016) argued that their approach is effective because resource request patterns of robots and humans constitute an intrinsic distinction between these two types of sessions which is not expected to change over time. They evaluated the efficiency of their approach using log data from three various Web servers and compared it against some other analytical learning classifiers, achieving very good classification results of their approach in terms of recall, precision, and F1.

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

The motivation for our experimental analysis of the approach proposed by Doran and Gokhale (2016) were some shortcomings of their experimental study. First, they used relatively old log data, which dated from 2008 to 2011, depending on a dataset. Second, only one out of three datasets was for an e-commerce website. Third, their procedure for session labeling resulted in a significant percentage of unlabeled sessions (20-36%, depending on a dataset), which were therefore excluded from the analysis.

To address the aforementioned shortcomings, we implemented the offline bot recognition method based on resource request patterns and carried out an experimental analysis of its performance for multiple e-commerce sites. We used actual log data for three online stores (data had been recorded in December 2015, January 2016, and May 2016) from various domains of the Web. The analyzed stores offered various types of products/services online and applied various online marketing techniques to attract potential customers. They were also differentiated in the software used to implement the store, the website structure, and consequently, the type of server resources accessed. Furthermore, they were differentiated in terms of the website popularity, the Web traffic intensity, and the share of bot requests in the overall traffic. This allows us to generalize results of our experimental analysis to some extent.

The main contribution of the paper is:

- the experimental analysis of resource request patterns on various e-commerce sites,
- a proposal and verification of a broad heuristics for bot session labeling.

The rest of the paper is organized as follows. The next Section discusses the experimental methodology, including the implemented classification approach, our procedures for session labeling, and measures applied to evaluate classification results. Then we describe our datasets, basic statistics on the reconstructed sessions and distributions of resource requests. Afterwards we discuss classification results obtained using both labeling procedures. The last Section concludes the paper.

METHODOLOGY

Basic Concepts

We consider Web traffic incoming and being processed on a typical Web server according to the HTTP protocol. Typically, HTTP requests are issued by Web browsers used by human users to access consecutive pages of a website. For each page demand a browser generates a sequence of requests, in which a page request is followed by hits for objects embedded in the page (images, text documents, compressed files, etc.). Some part of the incoming traffic, however, is generated not by human users' browsers but by Web robots. A sequence of bot's requests in session does not reflect the logical structure of the site but it depends solely on the algorithm underlying the bot implementation.

A *session* is defined as a sequence of HTTP requests (containing more than one request) received from a Web client (identified with an IP address and a user agent field combined), assuming that the time between any two consecutive requests does not exceed 30 minutes. Each request corresponds to a specific server resource, uniquely identified by its URI (Unified Resource Identifier). Each resource may be assigned to some resource type depending on the file extension in URI. We followed the partition of resource types proposed by Doran and Gokhale (2016), except the type corresponding to requests for directory contents, because such requests are not typical for e-commerce sites. Eight resource types were distinguished:

- *text* – text-formatted files (e.g., txt, xml, sty),
- *web* – page files and scripts (e.g., html, php, cgi),
- *img* – graphic files (e.g., jpg, png, tiff),
- *doc* – rich-text documents (e.g., doc, pdf, dvi),
- *av* – multimedia files (e.g., avi, mp3, mpg),
- *prog* – program files (e.g., exe, dll, dat),
- *compressed* (*zip*) – compressed files (e.g., zip, gzip, rar, 7z),
- *malformed* – malformed requests or unknown file extensions.

Thus, each session being a sequence of requests may be represented as a sequence of resource types, i.e., a *resource request pattern*.

Session Classification Based on Resource Request Patterns

We apply the offline classification approach as it was proposed in (Doran and Gokhale 2016). In this approach resource request patterns in sessions are represented as a model based on a first-order discrete time Markov chain (DTMC). A DTMC model is a tuple (**s**, **P**). **s** is a vector of probabilities of starting a session at individual resource types so that the i^{th} element of vector **s** is the probability that a session starts at resource type i. **P** is a matrix of transition probabilities between the resource types in session so that $\boldsymbol{p}_{i,j}$ at the position (i, j) is the probability that after the request for resource type i there will occur a request for resource type j. A DTCM model is trained based on resource request patterns of sessions allocated to a training set.

Let $\boldsymbol{S} = (x^1, x^2, \ldots, x^a)$ be the resource request pattern of a new session containing a requests. The log-probability that a DTMC will generate $\boldsymbol{S}$ is given by the formula:

$$\log \boldsymbol{Pr}(\boldsymbol{S}|\mathbf{s}, \mathbf{P}) = \log \mathbf{s}_{x^1} + \sum_{i=2}^{a} \log p_{x^{i-1},x^i} \qquad (1)$$

For each session class (robot or human) a separate DTMC model is trained. Let $\mathbb{R} = (\mathbf{s}_r, \mathbf{P}_r)$ be the DTMC trained with robot sessions and $\mathbb{H} = (\mathbf{s}_h, \mathbf{P}_h)$ the DTMC trained with human sessions. A class of a new session $\boldsymbol{S}$ is determined using (1) according to the following rule: if $\log \boldsymbol{Pr}(\boldsymbol{S}|\mathbb{R}) > \log \boldsymbol{Pr}(\boldsymbol{S}|\mathbb{H})$, $\boldsymbol{S}$ is classified as a robot and as a human otherwise.

Session Labeling

In order to evaluate the efficiency of a classification approach, sessions have to be labeled as driven by a bot or a human. We labeled the sessions using the following heuristic approach.

A base for session labeling was an online database of user agent fields and IP addresses known to correspond to various kinds of robots and Web browsers. Since a database used to session labeling by Doran and Gokhale (2016) has not been available anymore, we used the *Udger* database (Udger 2017). It contained 2,832 known user agent fields and 996,657 known IP addresses of Web robots, as well as 843 user agent fields of known browsers. Besides, we used an older, freely available database of *User agents* (User-agents 2014). In our basic labeling procedure (*labeling procedure 1*) a session of a given Web client was marked as a robot when:

- the client's user agent was found in the *Udger* database and the corresponding class was "crawler", "validator", "e-mail client", "library", "multimedia player", or "offline browser" or
- the client's user agent was found in the *User-agents* database and it was labeled as robot or
- the client's user agent contained a keyword suggesting a robot or
- the client's IP address was found in the *Udger* database and the corresponding class was "crawler", "known attack source - mail", "fake crawler", "known attack source - http", or "known attack source - ssh" or
- the file "robots.txt" was requested in session.

Otherwise, the session was marked as a human when the client's user agent was found in the *Udger* database and the corresponding class was "browser" or "mobile browser". Sessions that were not marked as a robot or a human, were labeled as unknown and were excluded from the experimental analysis.

Since in reality many robots are impersonators retaining legitimate user agent fields (Zeifman 2017), we decided to broaden our heuristics and to additionally label some sessions as robots based on some session features that are untypical for humans. In the broader labeling procedure (*labeling procedure 2*) a session was additionally marked as a robot ("probable bot") when:

- the image to page ratio in session was zero or
- all page requests in session had empty referrer fields or
- all responses in sessions were 4xx or
- all requests were of type HEAD.

Applying the two labeling procedures resulted in two scenarios in our experimental analysis, referred to as scenario 1 and scenario 2 for the labeling procedure 1 and 2, respectively.

Experimental Setup

Log file parsing, data preprocessing, and session reconstruction was done for each of the three online stores. For each store session labeling was performed using the labeling procedures 1 and 2, leading to two session datasets, used in the corresponding two experimental scenarios. The experimental analysis was conducted separately for each store and for each scenario. To evaluate the efficiency of resource request patterns-based classification for each scenario, a five-fold cross-validation was applied. A given session dataset was partitioned into five subsets so that each subset contained the same number of bot sessions and the same number of the human ones. In each round of cross-validation another subset was a testing set whereas the remaining four subsets aggregated made a training set. Finally, classification results (i.e., performance measures) were averaged over the five rounds.

Performance Measures

Three standard measures were adopted to evaluate the classification efficiency: recall, precision, and F1. *Recall* is the number of correctly classified robot sessions divided by the number of all robot sessions – thus, it reflects the percentage of correctly classified bot sessions. *Precision* is the number of correctly classified robot sessions divided by the number of all sessions classified (correctly or incorrectly) as robots – thus, this measure reflects the scale of wrong classification of human sessions as robots. Finally, F1 summarizes precision and recall into a single value as their harmonic mean – thus, it reflects the overall quality of the classifier. Moreover, we visualized matrices of transition probabilities between the resource types in sessions of both classes (without the cross-validation) to illustrate and assess the strength of differences in robot and human resource request patterns.

RESULTS AND DISCUSSION

Dataset Description

In the experimental study we used access log data for three various e-commerce websites from different domains over the Internet:

1) *Auto* – the online store offering car parts and accessories,
2) *Books* – the site of a publishing house with its own online bookstore, offering books, films, and multimedia,
3) *Elderly* – the online store offering products and services for elderly people.

The data covered different time spans in 2015 and 2016. The intensity of session arrivals was very differentiated across the websites: 148, 921, and 2,154 sessions per day had arrived on average at *Auto*, *Books*, and *Elderly* websites, respectively (Tab. 1). Although these three websites may be not representative of all e-commerce sites, they differ in many aspects – especially in resources request patterns.

Tab. 2 and Tab. 3 summarize the number of robot, human, and unknown sessions determined by using the labeling procedure 1 (without "probable bots") and 2 (with "probable bots"), respectively. One can notice that

regardless of the labeling procedure used, the proportion of robot traffic is very differentiated across the websites.

Table 1: Basic Information on the E-Commerce Datasets Used in the Experiments

	Auto	***Books***	***Elderly***
Time of data collection	Oct 1-Dec 27, 2015	May 1-31, 2016	Jan 1-17, 2016
No. of sessions	13,012	28,565	36,618
Avg. no. of sessions per day	147.9	921.4	2,154.0

When we consider only known bots, identified with a user agent field or an IP address (Tab. 2), bot sessions constitute only 7% of all sessions on *Elderly* site compared to 22% on *Books* site and as much as 71% on *Auto* site. However, when bots are additionally identified based on some specific session features (Tab. 3), the share of their sessions grows a bit for *Auto* (by 9%) and for *Books* (by 10%) and increases really drastically for *Elderly* (from 7 to as much as 86%, i.e. by 79%). This increase is due only to a small extent to sessions in which all requests are of type HEAD or all responses are 4xx – the main reason is a huge number of sessions in which all page requests have empty referrers and sessions with zero image to page ratio.

Table 2: Number of Human, Robot, and Unknown Sessions Identified by the Labeling Procedure 1 (Scenario 1 – without "Probable Bots")

	Auto	***Books***	***Elderly***
Robot	9,256 (**71.1%**)	6,432 (**22.5%**)	2,671 (**7.3%**)
Human	3,644 (28.0%)	21,339 (74.7%)	33,643 (91.9%)
Unknown	112	3	5

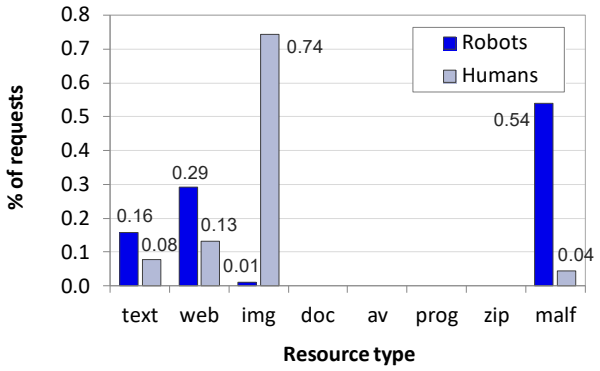
Table 3: Number of Human, Robot, and Unknown Sessions Identified by the Labeling Procedure 2 (Scenario 2 – with "Probable Bots")

	Auto	***Books***	***Elderly***
Robot	10,399 (**79.9%**)	9,086 (**31.8%**)	31,487 (**86.0%**)
Human	2,613 (20.1%)	19,476 (68.2%)	5,126 (14.0%)
Unknown	0	3	5

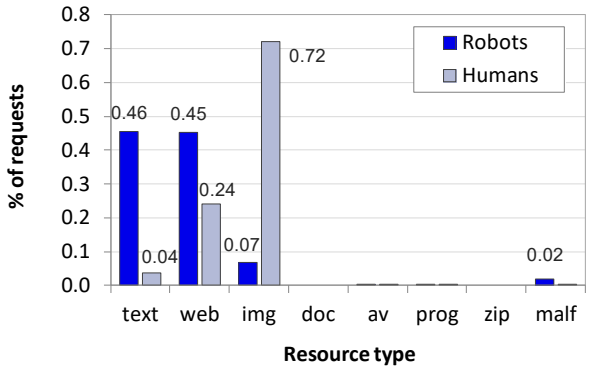
Resource Request Distribution

Fig. 1 presents distributions of resource request types on the analyzed websites for robots and humans. Only distributions for scenario 2 are shown due to the space limit. The corresponding results for scenario 1 lead to the same conclusions. The most clear differences were in bot sessions (1) for *Books*: shares of text and web requests 0.68 and 0.19 instead of 0.46 and 0.45, respectively, and (2) for *Elderly*: shares of web and malformed requests 0.86 and 0.08 instead of 0.79 and 0.15, respectively.

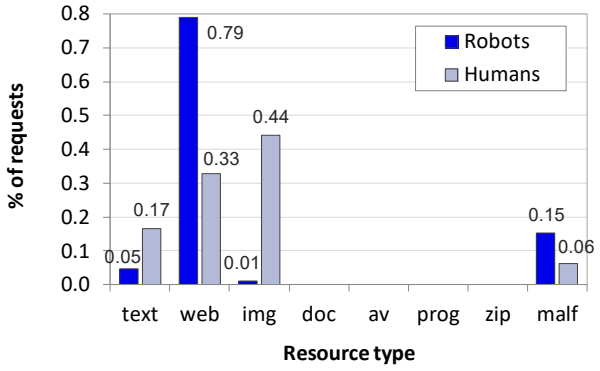
The main observation from Fig. 1 is that in human sessions similar resource types are requested across the websites (there are similar shares of requests of type img, web, text, and malformed) whereas for robot sessions access patterns differ across the websites (bot traffic on *Auto* site is dominated by malformed request, on *Elderly* site – by web request, and on *Books* site – by both text and web requests). A common feature of bot sessions is a very small share of image requests. Another observation is that on the analyzed e-commerce sites requests regarding resource type doc, av, prog, and compressed are extremely rare. This observation is consistent with results obtained in (Doran and Gokhale 2016) for the e-commerce site (this case of their experiments is referred to in our paper as "Ref_EC").



(a) *Auto*



(b) *Books*

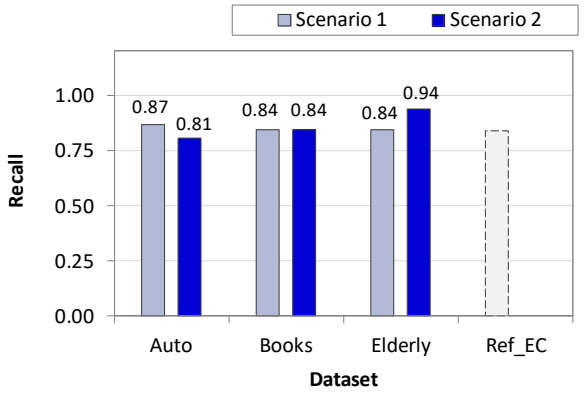


(c) *Elderly*

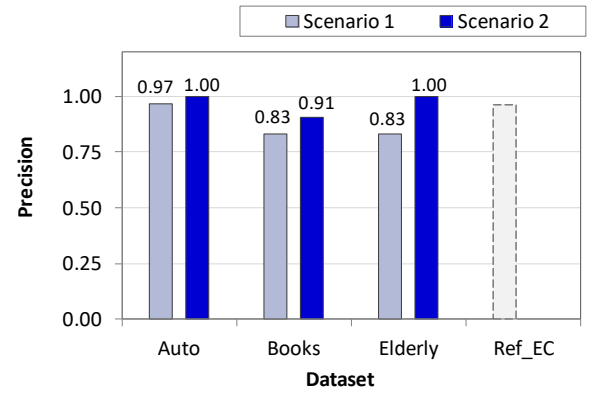
Figure 1: Distribution of Resource Requests for the Three E-Commerce Websites Broken Down by Session Class (Scenario 2).

Classification Results

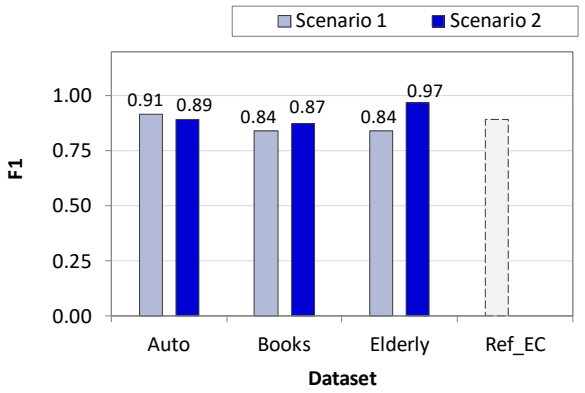
The experiments resulted in very high values of all three performance measures for both scenarios across all the datasets (Fig. 2). Recall is in the range of 0.84 to 0.87 for scenario 1 and 0.81 to 0.94 for scenario 2. This result is comparable with recall for "Ref_EC" (Fig. 2a).



(a) Recall



(b) Precision



(c) F1 measure

Figure 2: Classification Performance Compared to the Performance for the E-Commerce Dataset ("Ref_EC") Reported in (Doran and Gokhale 2016)

The tested approach is especially effective in terms of precision for scenario 2 (precision is in the range of 0.83 to 0.97 for scenario 1 and 0.91 to 1 for scenario 2). The high precision corresponds to the low false positive rate – this means that very few humans are wrongly classified as bots. This finding is also consistent with the results of the original experimental study reported for the e-commerce site ("Ref_EC").

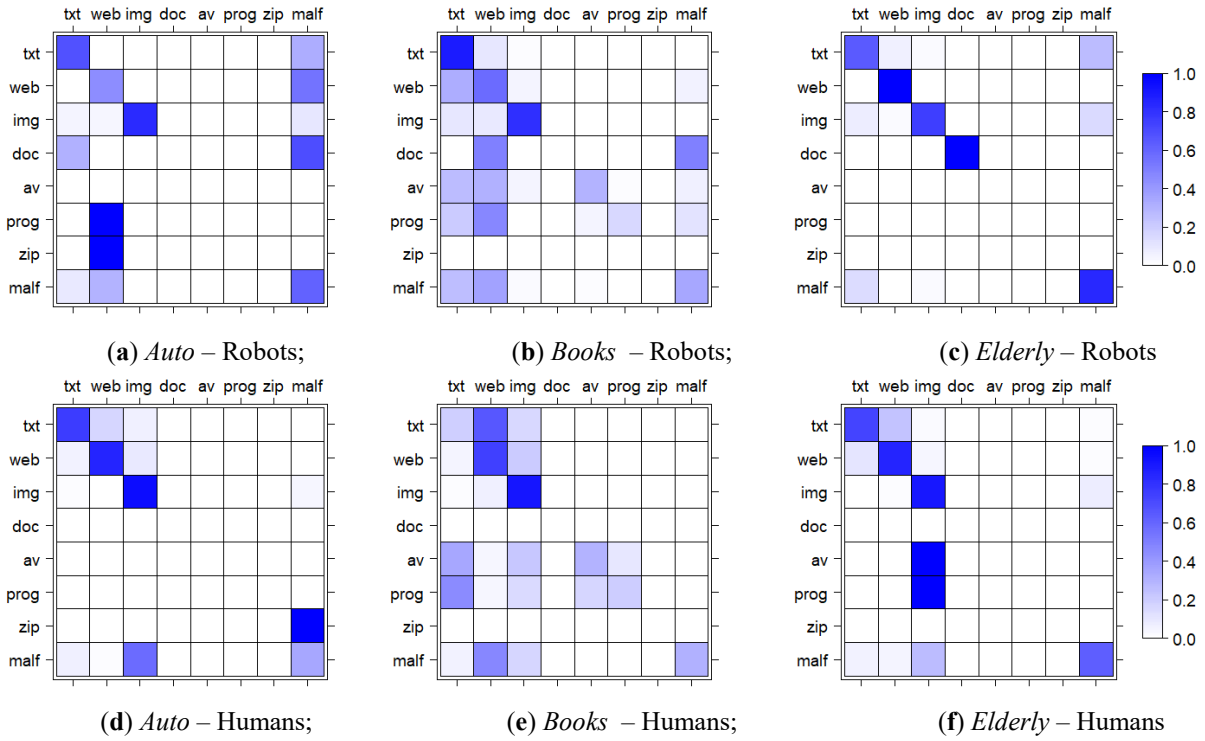
The resulting recall values do not allow us to clearly state that one of our labeling procedures is better than the other regarding classification capabilities of the tested approach. However, the resulting precision values for all three datasets demonstrate an evident superiority of labeling procedure 2. The overall quality of the classifier (F1 measure) also tends to favor the broader session labeling procedure.

This conclusion is confirmed by the visual inspection of transition probabilities between the resource types in sessions of both classes for scenario 1 (Fig. 3) and 2 (Fig 4). Transition matrices, computed for all sessions of a given class, are presented as level plots. The higher transition probability between two request types in the matrix **P** is, the darker color the corresponding cell has on the level plot. The most significant changes between scenario 1 and 2 are marked in Fig. 4 with black ovals.

A general look at the figures and the comparison of the plots in a vertical dimension (*Auto* vs. *Books* vs. *Elderly*) allows one to observe that each website has its own specific resource request patterns. Similarly, the comparison of the plots in a horizontal dimension (robots vs. humans) indicate clear distinctions between resource request patterns for both session classes at the same website. *Books* website is characterized by much larger diversity in the types of resource requested than other two websites – it should be emphasized, however, that the total numbers of requests for types doc, av, prog, and compressed are very low (cf. Fig. 1b).

There are some common features of robot and human patterns across the three websites. First, robots tend to issue requests of malformed type after requests of other types much more often than humans (there are more shaded cells in the rightmost columns of their matrices). This distinction is even clearer for scenario 2 (Fig. 4) than scenario 1 (Fig. 3). Second, both robot and human sessions reveal a high transition probability from type web to web and from type img to img. Third, for robot sessions transitions from type web to img almost do not happen, in contrast to human sessions. Also this feature is more evident for scenario 2.

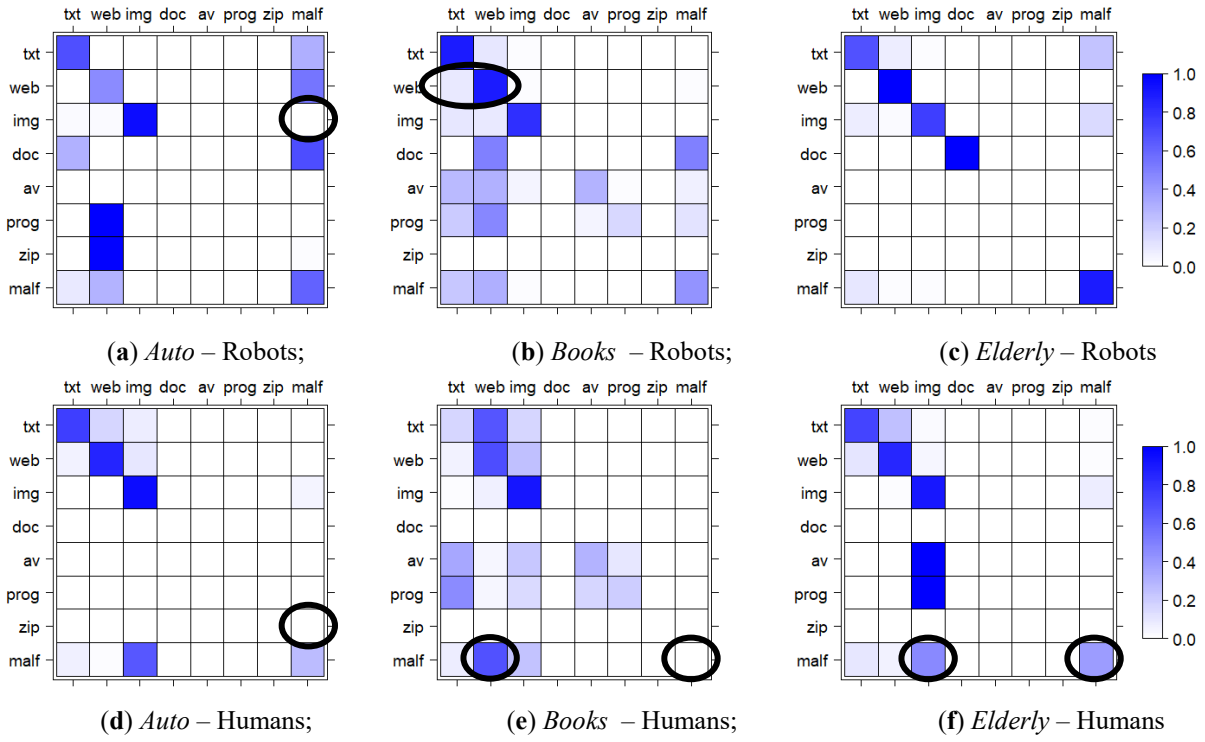
Generally, the level plots show that distinctions between resource request patterns of bots and humans are more significant when labeling procedure 2 is applied to identify bots than for procedure 1. In particular, broadening a subset of bot sessions led to the clarification of transition matrices for humans. For example, for human sessions it turned out that there are not transitions from compressed to malformed types on *Auto* website (Fig. 4d), there are not transitions from malformed to malformed types on *Books* website (Fig. 4e), and there are significantly less transitions from malformed to malformed on *Elderly* website (Fig. 4f).



(**a**) *Auto* – Robots; (**b**) *Books* – Robots; (**c**) *Elderly* – Robots

(**d**) *Auto* – Humans; (**e**) *Books* – Humans; (**f**) *Elderly* – Humans

Figure 3: Transition Probabilities Between Request Types in Session for the Three E-Commerce Websites for the Scenario 1: For Robots (Top) and Humans (Bottom).



(**a**) *Auto* – Robots; (**b**) *Books* – Robots; (**c**) *Elderly* – Robots

(**d**) *Auto* – Humans; (**e**) *Books* – Humans; (**f**) *Elderly* – Humans

Figure 4: Transition Probabilities Between Request Types in Session for the Three E-Commerce Websites for the Scenario 2: For Robots (Top) and Humans (Bottom).

CONCLUSIONS

In the study discussed in this paper we experimentally evaluated the efficiency of the session classification approach known from the literature, which uses differences in resource request patterns of bots and humans. We used real actual log data from three e-commerce sites from different domains of the Web. We proposed and implemented two heuristic procedures of labeling bots' and humans' sessions: (1) the labeling procedure based solely on known user agent fields and IP addresses and (2) the procedure which additionally uses intuition regarding bot session features untypical for human sessions. The broader session labeling procedure allowed us to label almost all the sessions in the analyzed e-commerce datasets, in contrast to other labeling approaches reported in the literature. Thus, we were able to verify the efficiency of the classification approach for almost all the sessions on the analyzed websites.

Moreover, differences between resource request patterns of bots and humans are more clear when the labeling procedure 2 is used. The general conclusion is that for the purpose of labeling sessions as bots or humans it is worth including session features untypical for humans, like the image to page ratio equal to zero, all page requests with empty referrers, all responses erroneous, and all requests of type HEAD.

Our analysis showed that bot and human sessions reconstructed from HTTP log data reveal significant differences in resource request patterns. Our findings confirmed high performance of the tested classification approach across multiple e-commerce websites from multiple domains, especially in terms of precision and F1 measure. Furthermore, our results indirectly confirmed that in reality many bots are camouflaged and use legitimate names of known Web browsers in their user agent string fields.

As part of future work we will address the problem of online Web bot detection and investigate the efficiency of resource request patterns in recognizing bot sessions in real time. In terms of the practical application it would be worth striving for elimination of false positives and increase in recall.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Centre (NCN) in Poland under Grant No. DEC-2017/01/X/ST6/01070.

REFERENCES

Almeida, V.; D. Menascé; R. Riedi; F. Peligrinelli; R. Fonseca; and W. Meira Jr. 2001. "Analyzing Robot Behavior in E-Business Sites". In *Proc. of ACM SIGMETRICS* (Cambridge, Massachusetts, USA, Jun.16-20). ACM, New York, NY, USA, 338-339.

Balla, A; A. Stassopoulou; and M.D. Dikaiakos. 2011. Real-Time Web Crawler Detection. In *Proc. of the 18th* ICT'11 (Ayia Napa, Cyprus, May 8-11). IEEE, Piscataway, N.J.

Doran, D. and S.S. Gokhale. 2011. "Web Robot Detection Techniques: Overview and Limitations." *Data Mining and Knowledge Discovery* 22, No. 1-2, 183–210.

Doran, D. and S.S. Gokhale. 2016. "An Integrated Method for Real Time and Offline Web Robot Detection." *Expert Systems* 33, No. 6, 592-606.

Doran, D.; K. Morillo; and S.S. Gokhale. 2013. "A Comparison of Web Robot and Human Requests". In *Proc. of the IEEE/ACM* ASONAM'13 (Niagara, ON, Canada, Aug.25-29). IEEE, Piscataway, N.J., 1374-1380.

Lee, J.; S. Cha; D. Lee and H. Lee. 2009. "Classification of Web Robots: An Empirical Study Based on Over One Billion Requests." *Computers & Security* 28, No.8, 795-802.

Saputra, C.H.; E. Adi; and S. Revina. 2013. "Comparison of Classification Algorithms to Tell Bots and Humans Apart." *Journal of Next Generation Information Technology* 4, No.7, 23-32.

Stassopoulou, A. and M.D. Dikaiakos. 2009. "Web Robot Detection: A Probabilistic Reasoning Approach." *Computer Networks* 53, No.3, 265-278.

Stevanovic, D.; A. An.; and N. Vlajic. 2011. "Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers." *Foundations of Intelligent Systems*. ISMIS 2011, LNCS 6804.

Suchacka, G. 2014. "Analysis of Aggregated Bot and Human Traffic on E-Commerce Site." In *Proc. of* FedCSIS'14 (Warsaw, Poland, Sep.7-10), ACSIS, Vol. 2. IEEE, Piscataway, N.J., 1123-1130.

Suchacka, G. and M. Sobków. 2015. "Detection of Internet Robots using a Bayesian Approach". In *Proc. of* CYBCONF'15 (Gdynia, Poland, Jun.24-26). IEEE, Piscataway, N.J., 365-370.

Udger. 2017. https://udger.com (access: September 4, 2017).

User-agents. 2014. http://www.user-agents.org (access: September 4, 2017).

Zeifman, I. 2017. "Bot Traffic Report 2016". Imperva Incapsula, https://www.incapsula.com/blog/bot-traffic-report-2016.html.

AUTHOR BIOGRAPHIES

GRAŻYNA SUCHACKA received the M.Sc. degrees in Computer Science and in Management, as well as the Ph.D. degree in Computer Science from Wroclaw University of Technology, Poland. Now she is an assistant professor in the Institute of Mathematics and Informatics at Opole University, Poland. Her research interests include analysis and modeling of Web traffic, Web mining, and Quality of Service with special regard to electronic commerce support and Web bot detection. Her e-mail address is: `gsuchacka@uni.opole.pl`.

IGOR MOTYKA is a student of Computer Science at the Faculty of Mathematics, Physics and Computer Science at Opole University, Poland. He is passionate about object-oriented programming, especially in C# and Java. His current area of interest is primarily development of WPF applications and Android software. His e-mail address is: `igor_motyka@mail.com`.