MODELING A SESSION-BASED BOTS' ARRIVAL PROCESS AT A WEB SERVER

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

Daria Wotzka
Faculty of Electrical Engineering, Automatic Control and Informatics
Opole University of Technology
ul. Prószkowska 76
45-758 Opole, Poland
E-mail: d.wotzka@po.opole.pl

KEYWORDS
Web traffic, Web workload, Web server, log file, user session, Web bot, Internet robot, analysis and modeling, regression analysis

ABSTRACT

The paper deals with the problem of modeling key features of the Web traffic generated by Internet bots, observed at the input of a Web server. Based on real log data of an online store, a set of bot sessions was prepared and analyzed. Three session features connected with bots' arrival process at the server were analyzed: session interarrival time, request interarrival time, and the number of requests in session. Distributional models for these bot session features were developed using regression analysis and validated through graphical comparisons of histograms for the empirical data and simulated values. As a result, interarrival times of bot sessions and interarrival times of requests in bot sessions were modeled by a Weibull and a Pareto distribution, respectively, and the numbers of requests in session were modeled by a function being a combination of a sigmoid and exponential distributions. The aim of our analysis was to develop a model of a session-based bot arrival process on a Web server which may be then implemented in a bot traffic generator integrated with a Web server simulator.

INTRODUCTION

The continuous growth of the Internet community and development of Web-based technologies have been accompanied by a persistent growth and proliferation of Internet robots. According to (Geroimenko 2004), an *Internet robot*, also called *Web bot*, *Web agent*, *software agent*, or *intelligent agent*, is "a software tool that carries out a task on behalf of a user or computer, typically relatively autonomously". Search engine crawlers, shopbots, link archivers, and other autonomous software agents are constantly crawling the Web, making it possible to provide Web users with a variety of fast, up-to-date, and reliable information. There are also harmful bots, however, which hack computer systems or user accounts, steal Web content, carry out DoS (Denial of Service) attacks, or perform other malicious actions.

The increasing activity of bots has motivated research into bot traffic characterization and analysis. In particular, many studies have focused on the detection of various types of malware and hacking bots activities, both in computer networks (Kołaczek and Juszczyszyn 2012; Skrzewski 2014) and on end hosts (Liu et al. 2008; Stevanovic et al. 2011; Skrzewski 2016).

The analysis of bot traffic has been typically performed using data recorded in Web server access logs (Doran and Gokhale 2010; Doran et al. 2013; Suchacka 2014). As regards the characterization of various kinds of bots, much attention has been paid to Web crawlers (Calzarossa and Massari 2013; Dikaiakos et al. 2005). Some studies have tried to classify different types of bots, including text crawlers, link checkers and icon crawlers (Lee et al. 2009) based on the finding that various types of bots exhibit different traffic characteristics.

A variety of statistical and machine learning techniques have been applied to detect bot traffic from Web server log data (Saputra et al. 2013; Stassopoulou and Dikaiakos 2009; Stevanovic et al. 2011; Suchacka and Sobków 2015).

Although many studies have addressed the issue of modeling Web traffic features for the overall traffic incoming to the server, relatively few such studies have been dedicated to bot traffic. Therefore, the arrival process of Web clients on a Web server has been well characterized and modeled but this is not the case for bots' arrivals. What is more, the character and properties of bot traffic is subject to more rapid changes than the traffic resulting from human users' visits on the Web, mainly due to the dynamic development of Web technologies relying on bots' indexing and monitoring activities – Web analytics, Internet marketing, price and product comparison Web services, etc.

As regards distributional models describing bots' arrival process on a Web server, some traffic features have been analyzed and modeled for known bots (Doran and Gokhale 2010) and for specific bots, including crawlers and shopbots (Almeida et al. 2001; Calzarossa and Massari 2013). In (Doran and Gokhale 2010) interarrival times of known bots' sessions were shown to be heavy-tailed and follow a hybrid lognormal-Pareto distribution. On the other hand, request interarrival times in known bots' sessions were not heavy-tailed and

Proceedings 31st European Conference on Modelling and Simulation ©ECMS Zita Zoltay Paprika, Péter Horák, Kata Váradi, Péter Tamás Zwierczyk, Ágnes Vidovics-Dancs, János Péter Rádics (Editors)
ISBN: 978-0-9932440-4-9/ ISBN: 978-0-9932440-5-6 (CD)

such distributions as lognormal, Weibull or Pareto did not fit the analyzed data set. Request interarrival times were modeled by a lognormal distribution for crawler sessions (Almeida et al. 2001; Calzarossa and Massari 2013) and with an exponential distribution for shopbot sessions (Almeida et al. 2001).

The literature review has shown that the main goal of previous analyses of a bot arrival process on a Web server has been to characterize and detect Web bots to cope with the consequences of their visits. In contrast, the motivation for our study was the need for differentiation between bots and human users in synthetic Web traffic being generated during simulation experiments testing the performance of a Web server system under various overload conditions.

Based on real log data of a Web store site we analyze and model key features of Web bot traffic in order to create a simulation model of bot arrivals on a Web server. The model is intended to be implemented in a bot traffic generator and used in simulation experiments to generate a stream of HTTP requests emulating the real bot traffic. An assumption underlying the planned simulation model is the ability to emulate many independent clients (including bots) interacting with the server. Thus, the implementation of such a model will make it possible to monitor and treat the emulated sessions in the server system independently from one another – which is not possible when using aggregated Web traffic models or reproducing real Web traces.

RESEARCH METHODOLOGY

Data Collection and Preprocessing

Source data of the Web traffic are Web server access log files. Basic log data describing each HTTP request coming to the server includes:

- information on the Web client sending the request (an IP address, a user agent string, a version of HTTP protocol, a referrer),
- information on the requested server resource (an URI of the resource, an HTTP method);
- information on request's processing at the server (a timestamp indicating request's arrival time, an HTTP status code, and a size of the file transferred to the client).

Processing and analysis of the requests' data using a computer program allows one to describe Web clients' sessions in an observation window. A *session* may be defined as a sequence of HTTP requests sent by a Web client during a single visit to the website (in some cases a session may contain only one request). A client is identified based on two data combined: the IP address and the user agent string. To identify multiple visits of the same client a minimum value of a time gap between any two consecutive requests of the client is defined: in the literature this value was established as 30 minutes (Catledge and Pitkow 1995; Bomhardt et al. 2005; Stevanovic et al. 2011).

Preparation of a Dataset of Web Bot Sessions

In the set of all sessions only the ones accomplished by Web bots were used for the analysis. Identifying bot sessions is not a trivial task. Only some bots may be recognized by requesting the file "robots.txt" in their sessions or by a bot name declared in the user agent string.

Furthermore, we assume that a Web client is a bot if the mean time per page in session is less than 0.5 second or by identifying some atypical behavioral patterns (Suchacka 2014): the empty referrer of the first request in session, all page requests or even all HTTP requests in session with empty referrers, combined with such observations as all requests containing the HTTP method *HEAD* (instead of the most popular method *GET*) or all requests with the erroneous HTTP status code and the image-to-page ratio in session equal to zero.

Moreover, each session containing only one request may be attributed to a Web bot as well. Even if a human user visits only one page of the online store website, their client – Web browser – generates and sends to the server multiple HTTP requests (a hit for a page description file and the following hits for embedded objects, which in the case of an e-commerce site are typically image files).

To avoid truncation of reconstructed bot sessions on the edges of the observation window, the target *bot sessions' assembling phase* was extended by two additional phases: an *initial phase* and a *final phase* (Fig. 1). Bot sessions which started in the *initial* phase are not used for the analysis. A *final phase* was introduced to allow time for completion of sessions initiated in the previous phase (sessions initiated in the *final phase* are not taken into consideration).

Developing Distributional Models for Bot Sessions' Features Using Regression Analysis

Three features of bot sessions, connected with the character of bots' arrival process on the server, were modeled in our study:

- session interarrival time,
- request interarrival time (in session),
- the number of requests in session.

These features are analyzed while maintaining the integrity of sessions. The way of measuring session interarrival times and request interarrival times is illustrated in Fig. 2.

Before the analysis a dataset of all bot sessions is divided into N subsets corresponding to consecutive days (because in the case of huge datasets there might be a problem with data processing). Each subset contains samples from sessions which were initiated on the corresponding day (although duration of some sessions might extend to the following day). Thus, for each session feature there are N subsets of data samples and each subset is analyzed separately.

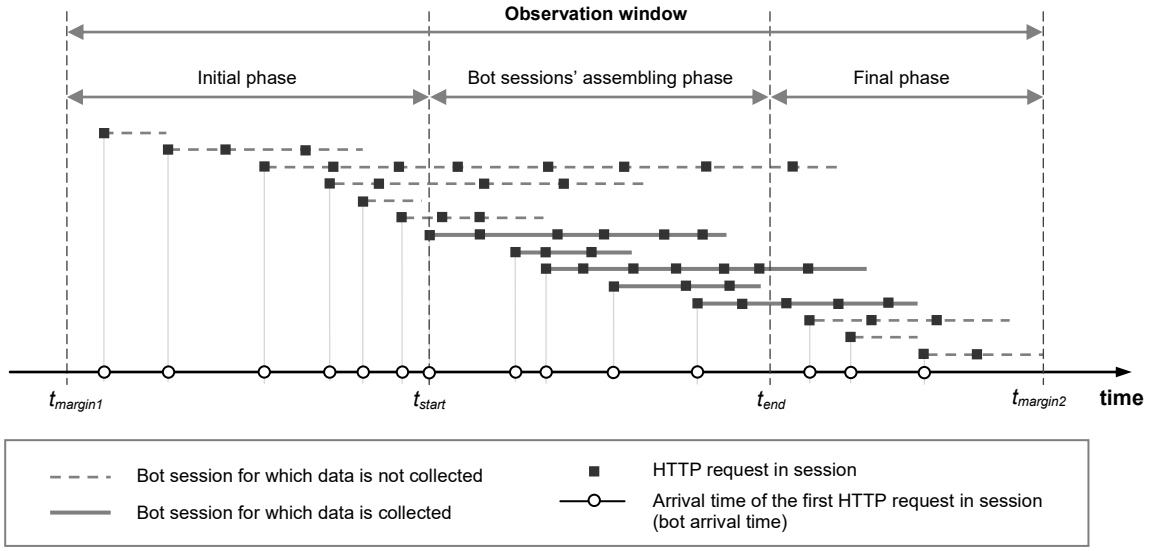

Figure 1: The Way of Collecting Bot Sessions' Data to Be Analyzed

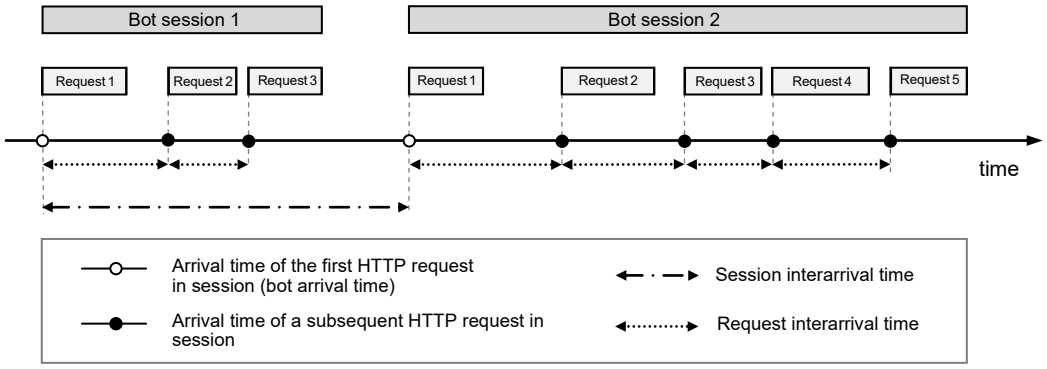

Figure 2: The Way of Measuring Interarrival Times

A distributional model for each session feature for a given subset is developed using statistical analysis software in the following way. First, for empirical data contained in the subset a histogram is built and based on its shape some candidate regression functions are chosen. The suitability of the functions is then analyzed, resulting in the selection of the function $f(x)$ characterized by the highest correlation of the model with the empirical data.

Values of parameters of a regression function are estimated by applying the least squares method and optimized by applying the Nelder-Mead simplex method. The optimization criterion is minimization of the residual norm, given by the formula:

$$\delta = \|\hat{\mathbf{y}} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)}, \qquad (1)$$

where:

n – the number of samples in a given subset,

y_i – the ith sample, $i = 1, 2, \ldots, n$,

$\hat{y}_i$ – the ith estimate of the regression function.

The suitability of a model to the empirical data is measured with the correlation coefficient, R^2, given by the formula:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \right)^2, \qquad (2)$$

where:

$\bar{y}$ – the average over y_i,

$\bar{\hat{y}}$ – the average over $\hat{y}_i$.

For each of the three session features values of R^2 are calculated for all N subsets of samples (N days) and then the mean value of the coefficient for a potential model is determined. The optimal model for each feature is selected based on the highest mean value of R^2.

RESULTS OF MODELING A BOTS' ARRIVAL PROCESS AT A WEB SERVER

Data used for the analysis was log data of a Web server hosting a middle-sized online bookstore. The store

software was osCommerce, a PHP-based electronic commerce platform.

A timespan of the used log data corresponded to the observation window ranging from 31st March 2014, 7:00 pm to 1st May 2014, 5:00 am with a 5-hour *initial phase* and a 5-hour *final phase*. Thus, the *bot sessions' assembling phase* covered the period from midnight preceding 1st April 2014 to midnight on 30th April 2014, i.e., 30 days ($N = 30$).

As a result of log data processing and analysis, a set of bot sessions was prepared. Basic information about the traffic on the server for the analyzed log data are given in Tab. 1. The entire set contained 42,782 Web bot sessions, in which 1,148,863 HTTP requests were sent to the server in total. Basic statistics for the three bot sessions' features are presented in Tab. 2.

Most active bots in terms of the number of sent requests were search engine crawlers (Googlebot, Yahoo! Slurp, MJ12bot, Bingbot, Nekstbot), SEO spybots (AhrefsBot and BLEXBot), online advertising bot Google AdsBot, e-commerce bots (ShopWiki, WillyBot, DotBot), and the Facebook bot called FacebookExternalHit.

Table 1: Basic Statistics for the Server Log Data (Timespan: 1st – 30th April 2014)

	All sessions	Bot sessions
Number of sessions	51,121	42,782 (83.7%)
Number of requests	3,200,228	1,148,863 (35.9%)
MB transferred	25,123.4	10,782.5 (42.9%)

Table 2: Statistics for Bot Sessions' Features

Statistics	**Session interarrival time [s]**	**Request interarrival time [s]**	**Number of requests in session**
Minimum	0	0	1
Maximum	6,178	1,799	47,624
Mean	61.0	28.2	26.9
Median	36	0	2
Std. dev.	83.3	142.0	242.4

As a result of the regression analysis of bot session features, three distributional models were best fitted to the corresponding empirical data: a Weibull model for the session interarrival time, a Pareto model for the request interarrival time and a model based on sigmoid and exponential type functions for the number of requests in session.

Session Interarrival Time

As regards the session interarrival time (*SIT*), the mean, equal to 61 seconds, is much higher than the median, equal to 36 seconds – in fact, almost 66% of samples are below the mean. It suggests that a distribution of the session interarrival time is right-skewed. This result confirms findings reported in previous studies, e.g. in (Doran and Gokhale 2010).

The distribution of the bot session interarrival time is modeled by a Weibull function given by the formula:

$$f_{SIT}(x) = A\frac{k}{\lambda}\left(\frac{x+\mu}{\lambda}\right)^{k-1} e^{-((x+\mu)/\lambda)^k}, \qquad (3)$$

where:

A – the amplitude parameter,
λ – the scale parameter, $\lambda > 0$,
k – the shape parameter, $k > 0$,
μ – the location parameter.

The independent variable x is the value of the time interval in seconds, $x \in \mathbb{N}$.

The shape of the regression curve differs slightly between the 30 subsets (measurement days) so the parameters of (3) are different for each day. Fig. 3 presents a histogram of session interarrival times along with the probability density function (PDF) of the estimated distribution for one sample day (27th April). To improve the clarity of the graph only 400 first bars of the histogram are shown in the figure in a semi-log scale.

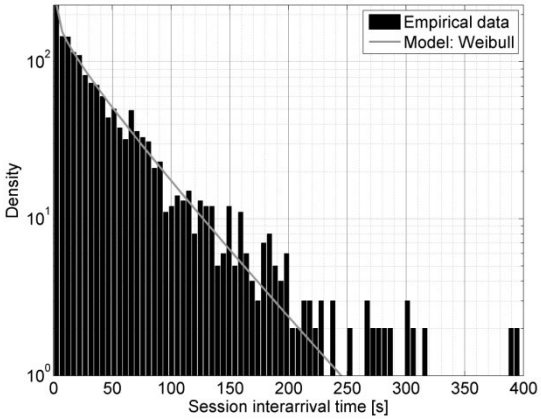

Figure 3: Histogram of the Empirical Session Interarrival Times and PDF of the Estimated Distribution (a Semi-Log Plot)

The correlation coefficient, R^2, varies from 0.77 to 0.99 and its average over all 30 datasets is 0.87. This shows a good fit of the distribution model to the empirical data.

Request Interarrival Time

Request interarrival times (*RIT*) are much more differentiated than session interarrival times, with the mean of 28.2 seconds and the standard deviation of 142 seconds. It is worth noting that the median is zero – in reality, above 73% of samples are equal to zero. This is a consequence of the fact that inter-request times in session are usually very short, even of the order of milliseconds (especially between hits for embedded objects) and on the other hand, request timestamps are registered in Web server log with an accuracy of one second.

Fig. 4 presents a histogram of request interarrival times and PDF of the estimated distribution for one day (27th

April). To capture data in a heavy-tail of the distribution, *y*-axis is plotted on a logarithmic scale.
Due to the very large number of zero intervals the empirical distribution is extremely heavy-tailed, similarly to distributions fitted in previous studies, e.g. in (Calzarossa and Massari 2013). We model this distribution by a Pareto function given by the formula:

$$f_{RIT}(x) = A\frac{k(x-\mu)^k}{(x-\mu)^{k+1}}, \quad (4)$$

where:
- A – the amplitude parameter,
- k – the shape parameter, $k > 0$,
- μ – the location parameter.

The independent variable x is the value of the time interval in seconds, $x \in \mathbb{N}$.

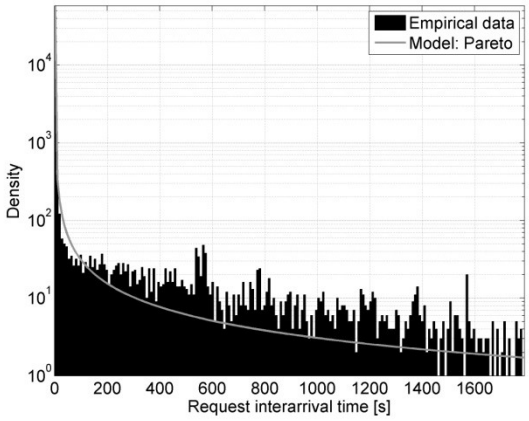

Figure 4: Histogram of the Empirical Request Interarrival Times and PDF of the Estimated Distribution (a Semi-Log Plot)

The correlation coefficient varies from 0.96 to 0.99 and its average over all 30 datasets is 0.99. So the correlation is even higher than that for the session interarrival time model.

Number of Requests in Session

Statistics for the numbers of HTTP requests in sessions (*NR*) generated by Web bots, included in Tab. 1, show that the minimum is one and the maximum is 47,624. In fact, as much as 41% of bot sessions contained only one request. The mean is 26.9 requests per session but the median is only two requests.
Fig. 5 shows a histogram of the numbers of requests in session (150 first bars only) and the corresponding PDF graph of the estimated distribution for one day (27th April). *Y*-axis is plotted on a logarithmic scale. One can notice in Fig. 5 that although the histogram is evidently right-skewed, there are some peaks, especially a very significant peak with a maximum value at about 120 requests per session (such a high peak is present in all 30 analyzed subsets).
We propose a model of the number of requests in bot session based on a regression function which is a combination of one sigmoid and two exponential type functions and is expressed by the formula:

$$f_{NR}(x) = e^{-\left(\frac{(x-\mu_1)}{c}\right)^2}\frac{A}{1+ke^{-b(x-\mu_2)}} + Be^{-\left(\frac{(x-\mu_3)}{d}\right)^2}, \quad (5)$$

where:
- k – the scale parameter of the sigmoid function, $k > 0$,
- b – the shape parameter of the sigmoid function,
- c, d – scale parameters of the first and the second exponential functions, respectively,
- A – the amplitude parameter of the sigmoid function,
- B – the amplitude parameter of the second exponential function,
- μ_1, μ_2, μ_3 – location parameters of the first exponential, sigmoid, and the second exponential functions, respectively.

The independent variable x is the number of requests in session, $x \in \mathbb{N}_+$.

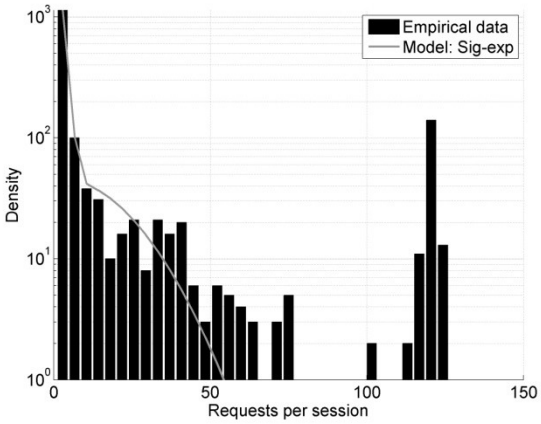

Figure 5: Histogram of the Empirical Numbers of HTTP Requests in Session and PDF of the Estimated Distribution for a Sample Day (a Semi-Log Plot)

This model is very well fitted to the empirical data as evidenced by the high correlation: R^2 ranges from 0.88 to 0.99 with the average equal to 0.97.

Summary of the Model

Since for each traffic feature each of the 30 datasets (measurement days) was analyzed and modeled separately, shapes of the regression curves differ slightly between the datasets. Thus, the parameters of the model for each feature are different for each day. Comparison of mean values of the correlation coefficient over all three features determined for each day showed that the highest overall correlation was achieved for the 27th April dataset (mean R^2 = 0.97). Therefore, parameters of the distribution functions in our model are given just for that day. The proposed model, including three features of Web bot arrival process, is summarized in Table 3.
The suitability of fit of the three distributional models to the empirical data is illustrated in Fig. 6. The best average fit was achieved for the request interarrival time

model (R^2 equal to 0.99). The lowest average value of R^2, equal to 0.87, was achieved for the session interarrival time model.

Table 3: Model of a Web Bots' Arrival Process

Traffic feature	Distribution	Parameters
Session interarrival time [s]	Weibull (3)	A = 3,176.4 λ = 43.43 k = 0.86 μ = 0
Request interarrival time [s]	Pareto (4)	A = 1228.6 k = 2.50 μ = 1.43
Number of requests in session	Sig-exp (5)	k = -0.0003 b = 64801.55 c = 4.32 d = 27 A = 4988.86 B = 46.1 μ_1 = -2.45 μ_2 = 2.9 μ_3 = 1.48

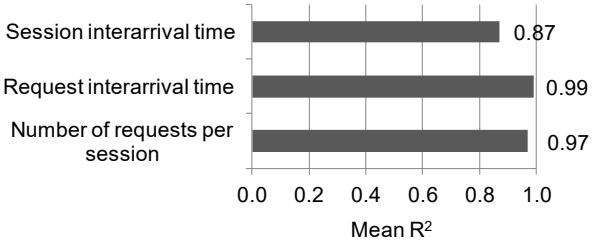

Figure 6: Suitability of Fit of the Model to the Empirical Data

VALIDATION OF THE MODEL

To check if modeled data will have a similar character to the real data, the proposed model was implemented in a C++ computer program. Numbers generated according to the model functions were recorded and then their distributions were compared with those of the corresponding empirical data (recorded in subsets for the 27th April). Results are shown in Fig. 7-9.

A visual inspection of the real data and the simulated results allows one to observe a very high degree of similarity between empirical and model data (note that the data is presented in a semi-log scale). One can notice some discrepancies, however. The Pareto model of the request interarrival time (Fig. 8), for which the highest mean R^2 was achieved, seems to produce too many extremely low values in the range of about 30 to 100 seconds and too few high values, exceeding 150 seconds. A big advantage of this model, however, is its ability to capture values in the distribution tail.

In the case of two other kinds of bot data, inter-session interarrival times generated according to the Weibull model (Fig. 7) and numbers of requests per session generated according to the sigmoid-exponential model (Fig. 9), one can notice that bodies of the distributions are very well fitted but the distribution tails are not well reflected.

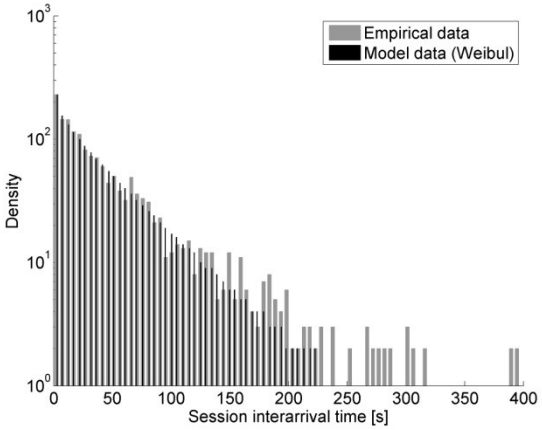

Figure 7: Histograms of the Session Interarrival Times for the Weibull Model Data and the Real Data (a Semi-Log Plot)

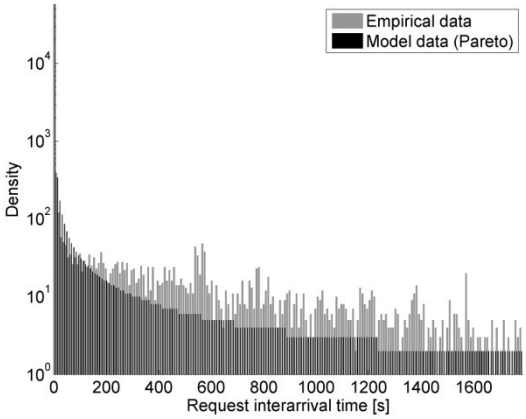

Figure 8: Histograms of the Request Interarrival Times for the Pareto Model Data and the Real Data (a Semi-Log Plot)

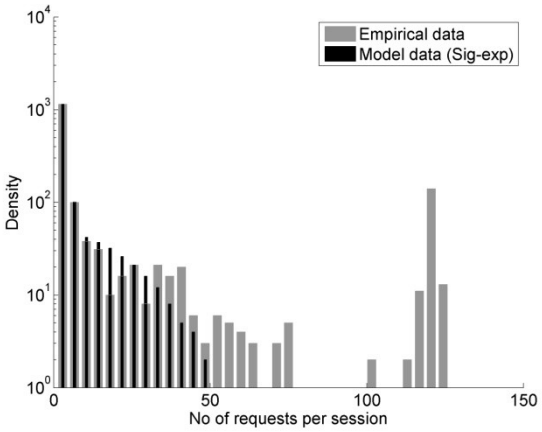

Figure 9: Histograms of the Numbers of Requests in Session for the Sigmoid-Exponential Model Data and the Real Data (a Semi-Log Plot)

A possible reason for these discordances may be the fact that no outliers have been eliminated before the analyses and the presence of the outliers might have negatively affected the resulting distributions of the modeled data. Besides, in the case of the numbers of requests per session the accurate modeling of a distribution peak with a maximum value at about 120 requests in session would probably require a combination of the bigger number of functions.

CONCLUSIONS AND FUTURE WORK

A result of the study presented in the paper is a mathematical model of a session-based bots' arrival process at an e-commerce Web server. The advantage of the model is that it is completely session-based and was developed based on real log data obtained from an online retailer. The model may be implemented in a discrete-event simulator, where the Web traffic is generated by many independent traffic sources (including Web bots). Thus, a feedback-based interaction of the server with the clients may be simulated.

A comparison of histograms for the real data and simulated results shows that the estimates generated according to the proposed model are characterized by the similar distributions as real data. However, further research is required to improve the quality of the model, to increase the value of the correlation coefficient and to capture data in heavy tails of the distributions.

Although this is a preliminary study, its results proved the efficiency of the proposed approach and provided some conclusions for our future work. The first issue to improve is connected with the coarse time resolution of log files, which causes most inter-request times in session to be equal to zero. Thus, it seems reasonable to eliminate zero intervals from the set of the analyzed request interarrival times by converting them to values of the order of milliseconds. Similarly, due to the high number of bot sessions containing only one request, it would be worth separating one-request sessions from the longer ones and to analyze the longer sessions separately. Another step which could improve the suitability of fit of the three distributional models to the empirical data might be the elimination of outliers.

The developed models describing a bots' arrival process on a Web server may be used in a traffic generator at the input of a Web server simulator. Such a simulation tool will make it possible to test Web service degradation under various load levels. We leave these issue to our future work.

ACKNOWLEDGEMENT

Grażyna Suchacka is a member of the consortium ICT COST Action IC1406 "High-Performance Modeling and Simulation for Big Data Applications" (cHiPSet).

REFERENCES

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

Bomhardt C.; W. Gaul; and L. Schmidt-Thieme. 2005. "Web Robot Detection - Preprocessing Web Log Files for Robot Detection". In *New Developments in Classification and Data Analysis*. *Studies in Classification, Data Analysis, and Knowledge Organization*, H.-H. Bock et al. (Eds.). Springer, Berlin-Heidelberg, 113-124.

Calzarossa, M.C. and L. Massari. 2013. "Temporal Analysis of Crawling Activities of Commercial Web Robots." *Computer and Information Sciences* III, 429-436.

Catledge, L.D. and J.E. Pitkow. 1995. "Characterizing Browsing Strategies in the World-Wide Web." *Computer Networks and ISDN Systems* 27, No.6, 1065-1073.

Dikaiakos, M.D.; A. Stassopoulou; and L. Papageorgiou. 2005. "An Investigation of Web Crawler Behavior: Characterization and Metrics." *Computer Communications* 28, No.8, 880-897.

Doran, D. and S.S. Gokhale. 2010. "Searching for Heavy Tails in Web Robot Traffic". In *Proceedings of the 7th International Conference on the Quantitative Evaluation of Systems* QEST'10 (Williamsburg, Virginia, USA, Sep.15-18). IEEE, Piscataway, N.J., 282-291.

Doran, D.; K. Morillo; and S.S. Gokhale. 2013. "A Comparison of Web Robot and Human Requests". In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* ASONAM'13 (Niagara, ON, Canada, Aug.25-29). IEEE, Piscataway, N.J., 1374-1380.

Geroimenko, V. 2004. *Dictionary of XML Technologies and the Semantic Web*. Springer-Verlag, London, UK.

Kołaczek, G. and K. Juszczyszyn. 2012. "Traffic Pattern Analysis for Distributed Anomaly Detection". In *Proceedings of the 9th International Conference on Parallel Processing and Applied Mathematics* PPAM'12 (Toruń, Poland, Sep.11-14, 2011), *Lecture Notes in Computer Science* 7204, R. Wyrzykowski; J. Dongarra; K. Karczewski and J. Waśniewski (Eds.). Springer, Berlin-Heidelberg, 648-657.

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.

Liu, L.; S. Chen; G. Yan; and Z. Zhang. 2008. "BotTracer: Execution-Based Bot-Like Malware Detection". In *Proceedings of the 11th International Conference on Information Security* (Taipei, Taiwan, Sep.15-18, 2008), *Lecture Notes in Computer Science* 5222. Springer, Berlin-Heidelberg, 97-113.

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.

Skrzewski, M. 2014. "System Network Activity Monitoring for Malware Threats Detection". In *Proceedings of the International Conference Computer Networks* (Lwówek Śląski, Poland, Jun.23-27), *Communications in Computer and Information Science* 431, A. Kwiecień; P. Gaj and P. Stera (Eds.). Springer, Berlin-Heidelberg, 138-146.

Skrzewski M. 2016. “About the Efficiency of Malware Monitoring via Server-Side Honeypots”. In *Proceedings of the International Conference Computer Networks* CN’16 (Lwówek Śląski, Poland, Jun.14-17), *Communications in Computer and Information Science* 608, P. Gaj; A. Kwiecień and P. Stera (Eds.). Springer, Cham, 132-140.

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

Stevanovic, D.; N. Vlajic; and A. An. 2011. “Unsupervised Clustering of Web Sessions to Detect Malicious and Non-Malicious Website Users.” *Procedia Computer Science* 5, 123-131.

Suchacka, G. 2014. “Analysis of Aggregated Bot and Human Traffic on E-Commerce Site.” In *Proceedings of Federated Conference on Computer Science and Information Systems* FedCSIS'14 (Warsaw, Poland, Sep.7-10), Annals of Computer Science and Information Systems (ACSIS), Vol. 2. IEEE, Piscataway, N.J., 1123-1130.

Suchacka, G. and M. Sobków. 2015. “Detection of Internet Robots using a Bayesian Approach”. In *Proceedings of the 2nd IEEE International Conference on Cybernetics* CYBCONF'15 (Gdynia, Poland, Jun.24-26). IEEE, Piscataway, N.J., 365-370.

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 Web Service with special regard to electronic commerce. Her e-mail address is: gsuchacka@uni.opole.pl.

DARIA WOTZKA received the M.Sc. degree in Computer Science from the Technische Universität Berlin, Germany and the Ph.D. degree in Electrical Engineering from the Opole University of Technology, Poland. She is a lecturer and research fellow at the Opole University of Technology, Poland. Her research interests include digital data processing, modeling and simulation of multiphysical phenomena occurring in electric power devices. Her e-mail address is: d.wotzka@po.opole.pl.