An extensive study of Web robots traffic

Maria Carla Calzarossa
Dip. Ingegneria Industriale e Informazione
Università di Pavia
via Ferrata 1
I-27100 Pavia, Italy
mcc@unipv.it

Luisa Massari
Dip. Ingegneria Industriale e Informazione
Università di Pavia
via Ferrata 1
I-27100 Pavia, Italy
massari@unipv.it

Daniele Tessera
Dip. Matematica e Fisica
Università Cattolica Sacro Cuore
via Musei 41
I-25121 Brescia, Italy
daniele.tessera@unicatt.it

ABSTRACT

The traffic produced by the periodic crawling activities of Web robots often represents a good fraction of the overall websites traffic, thus causing some non-negligible effects on their performance. Our study focuses on the traffic generated on the SPEC website by many different Web robots, including, among the others, the robots employed by some popular search engines. This extensive investigation shows that the behavior and browsing patterns of the robots vary significantly in terms of requests, resources and clients involved in their crawling activities. Some robots tend to concentrate their requests in short periods of time and follow some sorts of deterministic patterns characterized by multiple peaks. The requests of other robots exhibit a time dependent behavior and repeated patterns with some periodicity. A frequency domain methodology is applied for modeling the traffic represented as a time series. The models, consisting of trigonometric polynomials and Auto Regressive Moving Average components, accurately capture the behavior of the traffic and can be used as a basis for forecasting.

1. INTRODUCTION

Web robots are agents that crawl the Web on behalf of various systems and services with the main objective of automatically discovering and harvesting pages [12]. For this purpose, robots systematically browse entire Web domains and fetch pages at various levels of the site hierarchies. The pages downloaded by the commercial robots deployed by search engines are then indexed and presented to the users as a result of their queries.

To keep their content and search indices current, robots continuously recrawl websites at rates that vary according to some custom algorithms. Hence, the traffic due to crawling activities represents a good fraction of the overall traffic of websites and produces a non-negligible impact on their performance. Moreover, the deployment of multiple instances of robots operating in parallel often overloads Web servers and makes this traffic appearing as a sort of distributed denial of service attack. Crawling activities could also hide other types of attacks perpetrated by malicious robots with unethical purposes aimed, for example, at exploiting website vulnerabilities, collecting email addresses and personal information, discovering business intelligence [16].

Effective regulation policies are then of paramount importance. These policies have to identify and predict the patterns of requests originating from Web robots and assess their overall impact on the websites as well as their security and privacy implications.

In this paper, we focus on the crawling activities of a large set of robots deployed by different organizations and analyze their behavior with the objectives of discovering similarities and differences in their browsing patterns and developing models able to capture and reproduce these patterns. Our study relies on the logs collected on the website of the Standard Performance Evaluation Corporation (SPEC), a nonprofit corporation formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers [13].

The paper is organized as follows. After a survey of the work in the field of Web robot characterization, presented in Section 2, Section 3 summarizes the main properties of the traffic of the Web robots considered in our study. The behavior of the robots in terms of resources requested and clients employed for their crawling activities are described in Sections 4 and 5, respectively. The models of the traffic produced by these robots are then presented in Section 6. Finally, Section 7 summarizes the paper with some concluding remarks.

2. RELATED WORK

Several papers focus on the characterization of the crawling activities of Web robots and analyze various aspects of their behavior. In [15] the ethicality of Web robots is assessed against the rules specified by website administrators in the `robots.txt` file. The study shows that commercial robots are typically very ethical, nevertheless, many of them constantly violate some of the rules. The impact of Web robots on the overall load of a dynamic website is investigated in [8]. Through a quantitative measurement study, authors analyze the behavior and access patterns of the robots and outline the differences from human users. These results are then

used to devise caching policies aimed at mitigating the overload imposed by robots. The rate of behavioral changes expressed in terms of switching factors is considered in [10] as a key indicator to characterize the degree of uniformity in the request patterns and detect sessions initiated by Web robots. Markov models are applied in [7] to represent the intrinsic behavior and the contrasts in the resource request patterns of robots and human users.

Some papers specifically study commercial Web robots deployed, for example, by search engines. A comparison of the main characteristics of the robots of five search engines is presented in [6]. Metrics associated with workload features and resource types are proposed in [11] and used in an empirical study aimed at assessing the behavior of several robots. The classification of the sessions associated with three popular commercial Web robots, namely, Googlebot, MSNBot and YahooBot, is used in [14] to study the differences in their browsing styles. In particular, the dissimilarity between MSNBot and GoogleBot is rather high, whereas the crawling behavior of YahooBot is in between the other two and shares some characteristics with them.

In previous studies [3],[4], we have shown that the temporal behavior of commercial robots is characterized by some regular access patterns in terms of sessions, number of transactions and times between consecutive transactions within each session. Crawling activities are often intermixed with inactivity periods, whose duration follows some specific patterns. In this study, we analyze the characteristics of a large set of commercial Web robots. Even though our analysis focuses on the logs collected on one website only, to the best of our knowledge, no other studies have ever compared the behavior of such a large number of robots.

3. CHARACTERISTICS OF THE TRAFFIC

The traffic considered in our analysis refers to the HTTP transactions, that is, the HTTP requests issued by the clients and the corresponding responses generated by the Web server, collected on the SPEC website in 2011 during an observation interval of 12 consecutive weeks.

To investigate the characteristics of the traffic produced by Web robots, we first identify and extract from the log files storing transactions, those that can be associated with robots. For this purpose, we apply a syntactic analysis of the strings denoting the user agents, that is, the applications used by the clients to issue the requests. More specifically, we check these strings against some common keywords, such as, bot, crawler, spider, and some lists of user agent strings published on the Web (see, e.g., [17]). As an additional criterion, we check the presence of a URL or of an email address as most robots include in their user agent strings either URLs pointing to some explanatory pages or reference email addresses. The results of the syntactic analysis are then coupled with the reverse DNS lookup of the IP addresses of the clients generating the HTTP requests. This step is very important as it allows us to validate and ensure the authenticity of the robots. Indeed, user agent strings can be easily forged and the clients used by robots continuously change and become rapidly stale.

We further group the transactions classified according to the previous criteria as a function of the type of robots, e.g., text crawlers, image collectors, mobile content collectors, and the organization operating them. We discover a large number of different robots: some associated with major search engines, e.g., Yahoo!, Google, others associated with specific search engines, including, for example, a shopping search engine, i.e., ShopWiki, an engineering search engine, i.e., GlobalSpec, and image search engines, e.g., Picsearch.

Table 3 summarizes the characteristics of the most active robots identified in our logs. The table clearly shows that their crawling activities and strategies vary a lot. Some organizations employ different robots, each specialized for crawling specific types of content. For example, this is the case of Yandex, the company operating the most popular Russian search engine, and Google that employ specific robots for collecting images and mobile content. Moreover, even though the content of the SPEC website, that is, standardized benchmarks, is highly relevant to IT specialists and usually very highly ranked in search queries results, we observe that robots tend to crawl the website to a very different extent, with as many as 870,000 transactions or as little as 8,650.

A significant portion of the traffic, that is, slightly more than one million transactions, has been generated by `wget` and `libwww-perl` scripts. Despite their number, we do not include these transactions in the following analysis because these open source scripts can be used by anybody for any legitimate or malicious purpose. Hence, their access patterns change from time to time depending on the specific goals of the crawling sessions. Similarly, we will not consider the transactions associated with image and mobile content robots since the SPEC website stores a limited number of images and its pages are not customized to mobile devices. Hence, our analysis will be focused on some 2.1 million transactions associated with the 17 remaining robots. These transactions generate in total about 122Gbytes of traffic, i.e., 1.5Gbytes per day.

As shown in Table 3, Web robots usually identify themselves with a limited number of different user agent strings, whereas during their periodic crawling activities, they retrieve resources to a different extent. While some robots focus on a small pool of resources, others involve a much larger number of resources. It is worth emphasizing that the tendency to revisit the same resource varies across robots and does not depend on how often robots crawl the website. For example, even though Voilabot and Sosospider crawl a very small number of different resources, namely, 2,832 and 695, respectively, their average revisit ratio, that is, the ratio between number of transactions and number of resources, is quite large, namely, about 20. On the contrary, the Ocelli bot does not revisit any of the 80,001 resources. The independence between the revisit ratio and the number of transactions of each robot is confirmed by the corresponding correlation index that is equal to -0.12.

In addition, although Web robots are expected to comply with the rules of operation specified by the website administrator in the `robots.txt` file [9], from our data it appears that not all robots bother about requesting this file before crawling the website, thus confirming the findings presented

Botname	Organization	Transactions	Resources	Clients	User-agents	robots.txt
wget	Unknown	870,900	6,273	2	2	0
Slurp	Yahoo!	848,255	190,560	82	10	1,627
Googlebot	Google	430,799	263,801	264	4	1,067
YandexBot	Yandex	182,787	88,582	29	8	256
libwww-perl	Unknown	179,230	25,964	1,270	37	0
Bingbot	Microsoft	85,660	33,595	485	8	5,035
Ocelli	GlobalSpec	80,001	80,001	1	1	1
Baiduspider	Baidu	76,900	51,609	382	8	668
MJ12bot	Majestic-12	82,469	10,084	291	2	5,010
DotBot	Dotnetdotcom	74,253	37,397	2	1	532
Ezoom	Not specified	61,932	34,139	2	1	1,548
VoilaBot	Orange France	56,275	2,832	23	1	263
Slurp China	Yahoo!	41,641	4,602	598	3	731
Psbot	Picsearch	35,060	33,901	6	1	718
YandexImages	Yandex	30,274	8,691	8	4	145
Exabot	Exalead	22,955	20,305	4	3	241
Google-Image	Google	21,643	21,004	31	1	9
Sogou spider	Sogou	13,310	2,297	47	1	98
Sosospider	Tencent	12,326	695	91	2	300
NaverBot	NHN	12,519	4855	51	2	429
ShopWiki	ShopWiki	12,018	7,877	5	1	156
archive.org_bot	Internet Archive	10,672	9,996	10	2	55
Googlebot-Mobile	Google	8,650	2,298	40	2	14

Table 1: Characteristics of the traffic generated by the most active Web robots identified in our logs.

in [8] about its use in the identification of Web robots. The requests of the robots.txt file do not depend on the number of transactions and the number of resources retrieved by a each robot, while there is a moderate correlation with the number of clients used to crawl the site. The corresponding index is equal to 0.54.

4. RESOURCE POPULARITY

A more detailed analysis of the resources retrieved by the robots shows that the transactions involve some 380,000 resources. About 173,000 of these resources refer to HTML files and about 133,000 are almost evenly distributed among files with ps, pdf, cfg and txt extensions. The diagram of Figure 1 plots the popularity of the resources. Apart from two resources, i.e., the robots.txt file and the home page of the website, retrieved some 18,000 and 15,500 times, respectively, on average the remaining resources are retrieved 5.5 times each and the resources retrieved more than 10 times represent a small fraction, namely, about 12%, of the total number of resources. Nevertheless, these resources account for about 43% of the overall number of transactions.

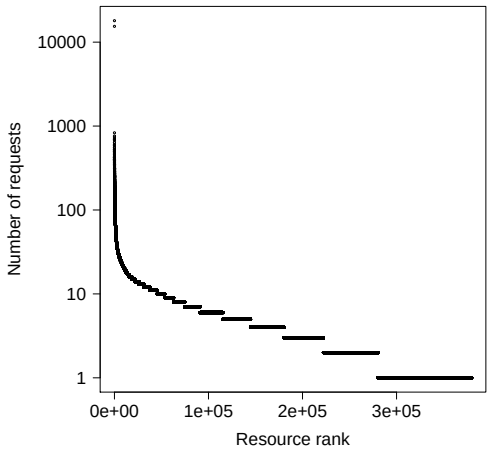


Figure 1: Popularity of the resources retrieved by the Web robots.

It is also interesting to study the popularity of a resource under a different perspective, that is, how many robots show an "interest" in the resource and retrieve it. While some 147,000 resources are retrieved by one robot only and very few resources, namely, 10, by all 17 robots, we identify 60% of the resources, that is, about 226,000, retrieved by at least two and at most six robots. The box plot of Figure 2 summarizes the distribution of the requests per robot to the resources, as a function of the number of robots requesting the resources. In the figure, we observe that, for the resources retrieved by all 17 robots, the distribution is characterized by a maximum equal to 1,059 requests per robot, corresponding to the robots.txt file, and much lower values of the median and the inter-quartile range, that are equal to 30 and 20 requests, respectively. In general, to an increase in the number of robots corresponds an increase in the requests per robots. This means that popular resources represent the favorite target of a large number of robots although these resources represent a small fraction of the resources they request. Furthermore, we notice that whenever resources are retrieved by a limited number of robots, the distribution of requests per robot is very spread. For example, the requests per robot to resources retrieved by at least two and at most six robots span a large range, with resources retrieved once

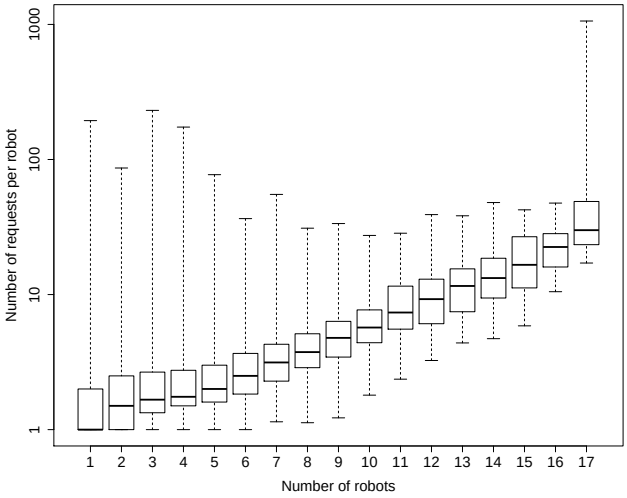


Figure 2: Box plot of the requests per robot to the resources as a function of the number of robots.

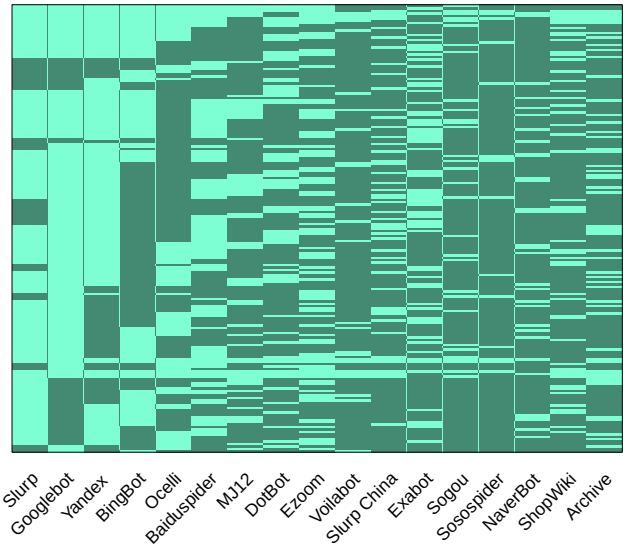


Figure 3: Combinations of the robots with respect to the retrieved resources.

and as many as 231 times, while the range is much narrower, i.e., with resources retrieved from 10 up to 28 times, whenever 15 or 16 robots are involved.

To further explore this new concept of popularity, we analyze the combinations of the robots with respect to the resources they retrieve, that is, which robots retrieve which resources. In total, out of the 380,000 resources, we identify 4,056 combinations of the 17 robots considered in our study, that is, very few with respect to the total number of possible combinations. This means that it is very likely for the resources to be requested by the same pool of robots. Of course, the number of resources per combination varies a lot. On average a combination groups 93 resources, even though few combinations, namely 118, involving one or two robots only, group 258,423 resources. The diagram of Figure 3 shows a snapshot of the combinations identified from our data. Light green and dark green denote the presence and the lack of a given robot in the combination, respectively. In this example, for the sake of clarity, each of the 185 combinations shown in the diagram involves the same number of resources, namely, four. In total, these combinations account for 740 resources that have been retrieved in total approximately 30,000 times. In the figure, we easily notice a large light green area on the left hand side of the diagram corresponding to the robots whose crawling activities involve larger numbers of resources.

5. CLIENT USAGE PATTERNS

The organizations operating the various robots tend to employ for their crawling activities a variable number of clients (see Table 3). Moreover, as shown in Figure 4, the clients are used to a very different extent. A very small fraction of clients is responsible of a significant portion of the requests: about 1.5 million requests, that is, 71%, are issued by ten clients belonging to different organizations, including among the others, Yahoo!, Google and Yandex.

From the analysis of the usage patterns of the clients, we notice that all robots, but Ocelli and DotBot, have been crawling the SPEC website for the entire observation interval. Nevertheless, their activities are characterized by considerable differences. For example, VoilaBot employs 23 clients, each operating between eight and ten days, whereas NaverBot relies on 51 clients, used for no more than 10 hours each. On the contrary, one client of the 29 clients employed by YandexBot is used for the entire observation interval, whereas the others participate rather sporadically

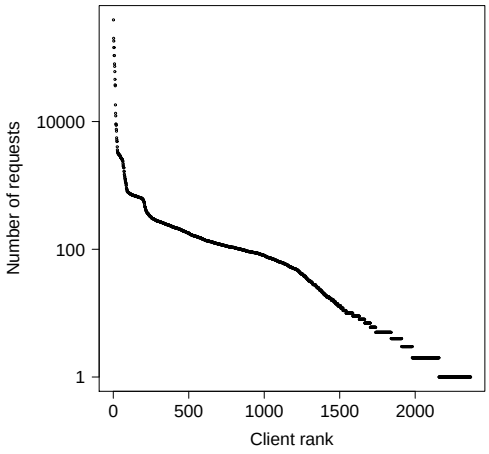


Figure 4: Popularity of the clients used by the Web robots to issue their requests.

in the crawling activities.

The usage patterns of the clients employed by the various organizations have been further investigated by analyzing the extent of the overlap of their crawling activities. Figure 5 shows the patterns of the clients employed by Googlebot and Baiduspider during our observation interval. As can be seen, patterns are characterized by significant differences. On one side, we observe organizations, such as, Baidu, that base their crawling activities on a large number of cooperating clients that operate in parallel. On the other side, other robots, such as, Googlebot, rely on much fewer concurrent clients. Moreover, there is a tendency to systematically change the clients, thus making robot identification more challenging. This is more evident for Googlebot, whose clients are mostly active in non overlapping intervals (see Fig. 5 (a)). For example, the six most active clients, responsible of about 91% of the requests, i.e., 390,000, perform their crawling activities in completely disjoint time intervals of different durations. In the case of Baiduspider, we notice some 100 new clients appearing in the last week of our observation interval (see Fig. 5 (b)) and the clients previously used slowly disappearing, that is, a new set of clients replace the old ones in the crawling activities.

6. MODELS OF THE TRAFFIC

As already pointed out, the traffic produced by Web robots represents a good fraction of the overall traffic of the websites. It is then important to model this type of traffic as to be able to predict its impact on the website performance.

A snapshot of the traffic generated on the SPEC website by the 17 Web robots considered in our investigation over the first four weeks of our observation interval is displayed in Figure 6. Note that the requests shown in the diagram

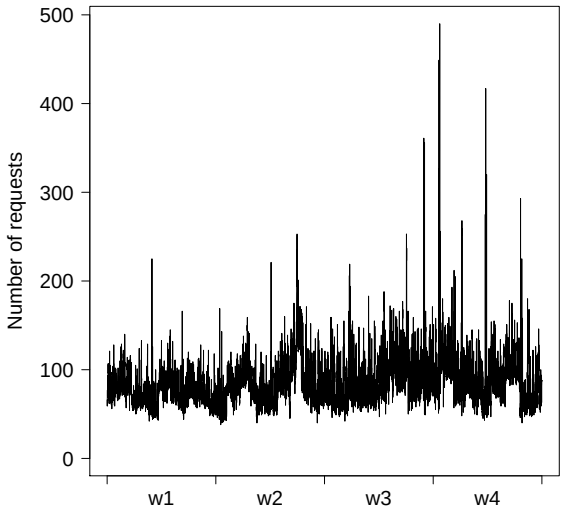


Figure 6: Snapshot of the traffic generated by the 17 Web robots.

are aggregated into five minutes intervals. The figure shows that the traffic is not evenly distributed over the weeks and exhibits large fluctuations. On average, the website receives 87 requests per interval, with some intervals characterized by high peaks with more than 200 requests. Moreover, we can recognize some sort of repeated patterns, even though we cannot identify any clear periodicity.

The idea is then to represent the traffic by means of a time series and build simple models which capture its behavior and provide an adequate basis for forecasting [1], [2]. We apply a frequency domain methodology for time series modeling as it provides very powerful tools for understanding the dynamic behavior of the data without any well defined periodicity. More specifically, tools, such as, the autocorrelation function and the periodogram, allow us to diagnose the properties of the time series. In our analysis, both tools confirm the presence of repeated patterns. Moreover, the periodogram highlights spikes in the frequency spectrum, that lead us to consider the traffic as a multiple seasonal process with cycles of different lengths. In our data, these spikes correspond to periods ranging from four hours up to four weeks.

Despite other studies (see, e.g., [5]), to capture this behavior, we do not resort to the standard time series decomposition in seasonal, trend and irregular components; instead, we apply a direct approach based on the results of the spectral analysis. In particular, we apply a linear regression using a trigonometric polynomial with frequencies corresponding to the largest spectral power densities. Figure 7 shows the trigonometric polynomial with the 15 frequencies identified by the goodness of fit analysis.

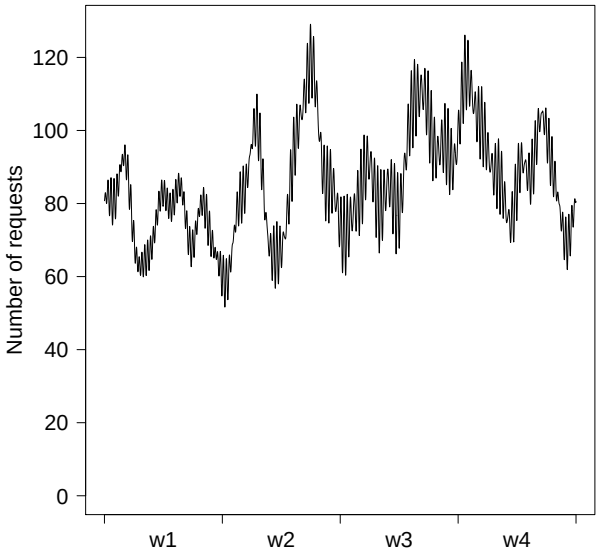
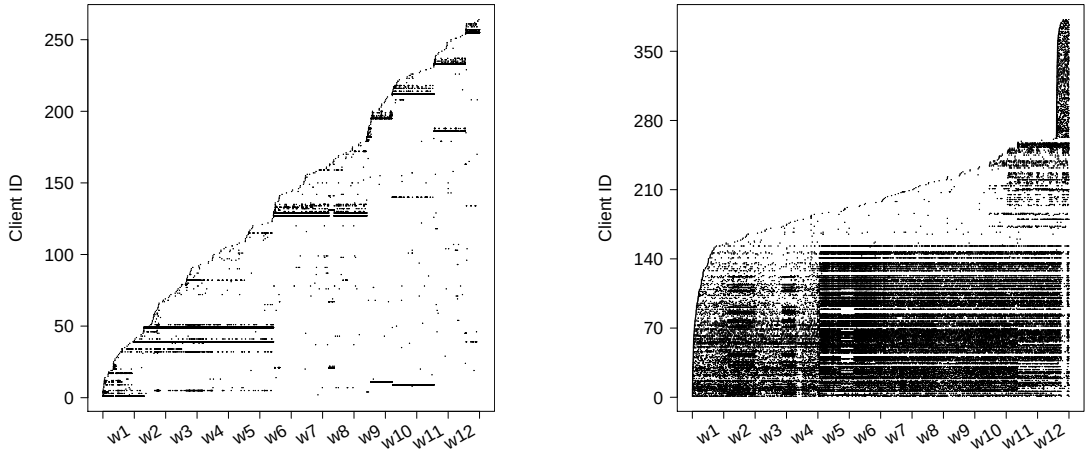


Figure 7: Trigonometric polynomial obtained from the time series analysis of the traffic shown in Figure 6.

The irregular component of the time series, that is, the residuals not accounted for by the trigonometric polynomial, is then estimated using Auto Regressive Moving Average mod-



(a) (b)

Figure 5: Usage patterns of the clients employed by Googlebot (a) and Baiduspider (b) for the crawling activities performed during the observation interval of 12 weeks.

els. The choice of these models is driven by the short term autocorrelations characterizing the residuals. More specifically, we identify an ARMA $(2, 1)$ with autoregressive coefficients equal to 1.0944 and -0.2045, and moving average coefficient equal to -0.6695. The overall traffic and its final model, obtained applying an additive approach, are displayed in Figure 8. For legibility, the diagram refers to an

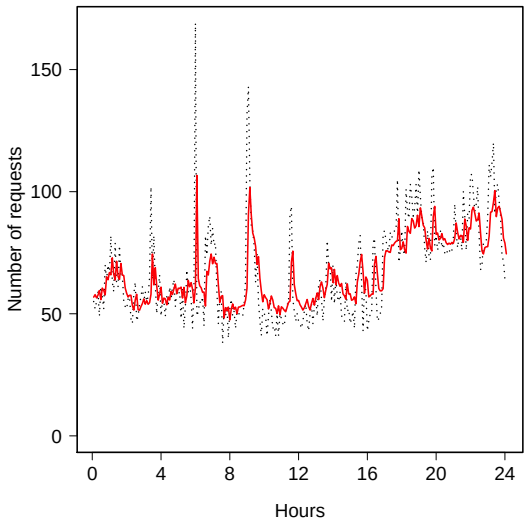


Figure 8: Overall traffic generated by the 17 Web robots (dotted pattern) and corresponding model (solid red pattern) over a 24 hours time interval.

interval of one day only.

The analysis of the traffic of the individual Web robots outlines interesting differences in their behavior. We identify two broad categories of robots: robots whose requests are mostly concentrated in short periods of time and characterized by some deterministic patterns, and robots whose requests are characterized by a time dependent behavior and some repeated patterns. Figure 9 shows the traffic generated by robots of the two categories, namely, NaverBot and YandexBot. The diagrams plot the traffic generated by these robots over the first four weeks of our observation interval. As can be seen, the traffic of NaverBot does not follow any specific pattern. We can identify several peaks of requests intermixed with intervals without any request. In the case of YandexBot, the traffic is characterized some periodicity and repeated patterns.

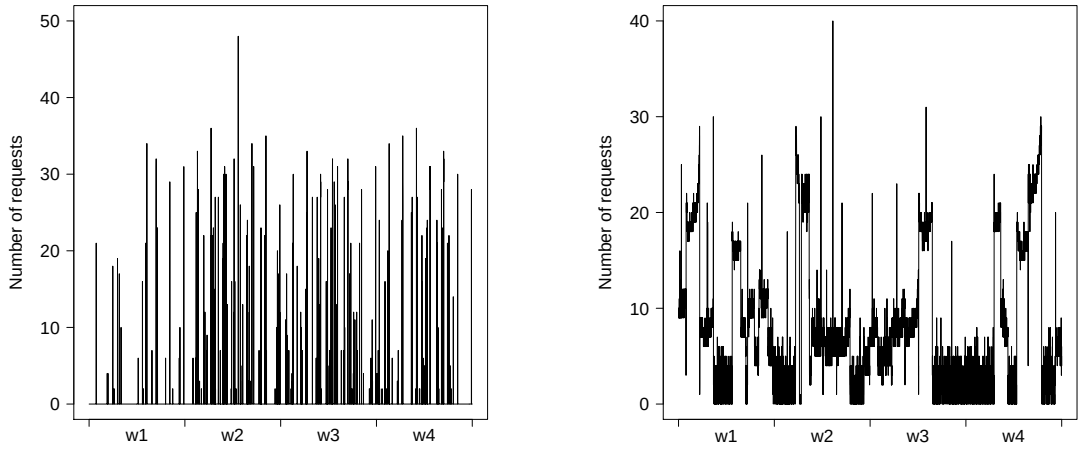
To model this type of traffic, we apply the same methodological approach used for the overall traffic. Figure 10 shows the model of the traffic of YandexBot. For legibility, the plot refers to a 24 hours time interval. The identified model, consisting of a trigonometric polynomial with 11 frequencies and an ARMA $(1, 2)$ model, with the autoregressive coefficient equal to 0.9825 and the moving average coefficients equal to -0.4803 and -0.054, accurately captures the experimental data.

7. CONCLUSIONS

The extensive investigation presented in this paper has focused on the crawling activities of a large set of commercial Web robots with the objectives of outlining their browsing patterns and developing models to capture and reproduce their traffic.

The study is based on the analysis of the logs collected on the SPEC website during an observation interval of 12 consecutive weeks. A syntactic analysis of the strings denoting the user agents, together with the reverse DNS lookup of the IP addresses of the clients generating the HTTP requests is applied to identify 17 popular Web robots, issuing some 2.1 million transactions.

The analysis of the traffic in terms of clients employed in



(a) (b)

Figure 9: Snapshot of the traffic generated by NaverBot (a) and YandexBot (b) over a four weeks time interval.

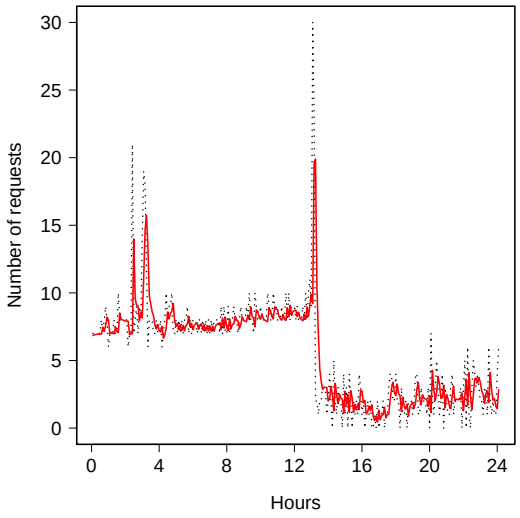


Figure 10: Traffic generated by YandexBot (dotted pattern) and corresponding model (solid red pattern) over a 24 hours time interval.

the crawling activities and resources requested by the Web robots has outlined some interesting findings. While some robots focus on a small pool of resources, others involve a much larger number of resources. Moreover, the tendency to revisit the same resource varies across robots and does not depend on how often robots crawl the website.

Popular resources are the favorite target of a large number of robots although these resources represent a small fraction of the resources they request. Furthermore, it is very likely for the resources to be requested by the same pool of robots.

The analysis of the usage patterns of the clients has shown that some Web robots employ a large number of cooperating clients for their crawling activities, while others rely on much fewer concurrent clients, thus producing different impacts on the performance of the website. In addition, we have detected the tendency of the organizations operating the robots to systematically change the clients and this makes robot identification rather challenging.

By looking at the traffic generated by the robots, we have discovered some large fluctuations and a time dependent behavior with no clear periodicity. Time series analysis is applied to model the overall traffic as well as the traffic of the individual Web robots. The identified models, based on trigonometric polynomials and ARMA components, represent a good basis for forecasting of Web robots traffic.

Acknowledgment

Authors are very thankful to Alessandro Vito for his valuable support in the analysis of the client usage patterns.

8. REFERENCES

[1] G. E. P. Box and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control*. Holden-Day, 1976.

[2] D. R. Brillinger. *Time series: data analysis and theory*, volume 36 of *Classics in Applied Mathematics*. SIAM, 2001.

[3] M. Calzarossa and L. Massari. Analysis of Web logs: Challenges and findings. In K. Hummel, H. Hlavacs, and W. Gansterer, editors, *Performance Evaluation of Computer and Communication Systems - Milestones and Future Challenges*, volume 6821 of *Lecture Notes in Computer Science*, pages 227–239. Springer, 2011.

[4] M. Calzarossa and L. Massari. Temporal analysis of crawling activities of commercial Web robots. In E. Gelenbe and R. Lent, editors, *Computer and Information Sciences III*, Lecture Notes in Electrical Engineering, pages 429–436. Springer, 2012.

[5] M. Calzarossa and D. Tessera. Time series analysis of

the dynamics of news websites. In *Proc. PDCAT 2012*, pages 529–533. IEEE Computer Society Press, 2012.

[6] M. Dikaiakos, A. Stassopoulou, and L. Papageorgiou. An investigation of web crawler behavior: characterization and metrics. *Computer Communications*, 28(8):880–897, 2005.

[7] D. Doran and S. Gokhale. Detecting Web Robots Using Resource Request Patterns. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, pages 7–12, 2012.

[8] A. Koehl and H. Wang. Surviving a search engine overload. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 171–180. ACM, 2012.

[9] M. Koster. A Method for Web Robots Control. Network Working Group - Internet Draft, 1996.

[10] S. Kwon, M. Oh, D. Kim, J. Lee, Y.-G. Kim, and S. Cha. Web Robot Detection based on Monotonous Behavior. In *Proc. International Conference on Information Science and Industrial Applications*, pages 43–48, 2012.

[11] J. Lee, S. Cha, D. Lee, and D. Lee. Classification of web robots: An empirical study based on over one billion requests. *Computers & Security*, 28(8):795–802, 2009.

[12] C. Olston and M. Najork. Web Crawling. *Journal of Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.

[13] SPEC corporate website. http://www.spec.org.

[14] D. Stevanovic, N. Vlajic, and A. An. Detection of malicious and non-malicious website visitors using unsupervised neural network learning. *Appl. Soft Comput.*, 13(1):698–708, 2013.

[15] Y. Sun, I. G. Councill, and C. Giles. The Ethicality of Web Crawlers. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 668–675. IEEE Computer Society, 2010.

[16] M. Thelwall and D. Stuart. Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, 57(13):1771–1779, 2006.

[17] User-agent-string.info. http://user-agent-string.info/list-of-ua/bots, Last visited: May 16, 2013.