Temporal analysis of crawling activities of commercial Web robots

Maria Carla Calzarossa and Luisa Massari

Abstract Web robots periodically crawl Web sites to download their content, thus producing potential bandwidth overload and performance degradation. To cope with their presence, it is then important to understand and predict their behavior. The analysis of the properties of the traffic generated by some commercial robots has shown that their access patterns vary: some tend to revisit the pages rather often and employ many cooperating clients, whereas others crawl the site very thoroughly and extensively following regular temporal patterns. Crawling activities are usually intermixed with inactivity periods whose duration is easily predicted.

Key words: Web robots; Crawling; Temporal patterns.

1 Introduction

Web robots are agents that traverse the Web and access and download Web pages without any significant human involvement [7]. These agents are a fundamental component of many applications and services, e.g., search engines, link checkers, Web services discovery. Nevertheless, some robots open up privacy and security issues as well as performance issues [11]. Although robots employed by major search engines tend to behave, the highly dynamic nature of Web content requires some sort of aggressive crawling policies. For example, to provide up-to-date content, commercial robots frequently revisit Web sites, thus draining server resources and causing potential bandwidth overload and overall performance degradation of the sites. Hence, to avoid

M. Calzarossa · L. Massari
Dipartimento di Ingegneria Industriale e dell'Informazione
Università di Pavia, Via Ferrata 1, I-27100 Pavia, Italy
e-mail: {mcc,massari}@unipv.it

damages and economic losses, it is important to identify Web robots and understand their behavior and their impact on the workload of a Web site.

This paper focuses on the analysis of some commercial robots with the objective of characterizing their behavior and their access patterns. The outcomes of this temporal analysis could be very useful for Web site administrators to estimate and predict the traffic due to robots and develop regulation policies aimed at improving site availability and performance.

Our study relies on the Web access logs collected on the European mirror of the SPEC (Standard Performance Evaluation Corporation) Web site. The choice of this site is motivated by its content, i.e., the performance results of standardized benchmarks of the newest generation high-performance computers. This content makes the site very relevant to the entire community of IT specialists and even more to search engines.

The paper is organized as follows. Section 2 briefly presents the state of the art in the area of the Web robot identification and characterization. The methodological approach applied in our study is presented in Section 3. The results of the exploratory analysis and of the temporal patterns followed by Web robots are discussed in Sections 4 and 5, respectively. Finally, Section 6 summarizes the major findings of this study.

2 Related work

The identification and characterization of the traffic generated by Web robots have been addressed in several papers (see e.g., [1, 2, 4, 5, 6, 9, 10]). Some studies analyze the overall properties of the traffic, whereas others take into account more specific aspects. A detailed survey of the existing robot detection techniques is presented in [3], where authors classify the techniques into four categories, discuss strengths and weaknesses of the underlying detection philosophy and suggest new strategies that try to overcome current limitations.

As commercial robots employed by search engines produce a large fraction of the overall traffic experienced by Web sites, some papers specifically focused on the characterization of this type of traffic. Dikaiakos et al. [2] compare the behavior of the crawlers of five popular search engines by analyzing access logs collected on various academic Web servers and introduce a set of metrics for their qualitative description. Similarly, Lee et al. [5] investigate the characteristics of some popular Web robots by analyzing a very large number of transactions recorded by a commercial server over a 24 hours period. Metrics associated with HTTP traffic features and resource types are then used for the classification of the robots. The analysis, though very detailed, fails to investigate the temporal or periodic patterns of the traffic.

In this study we complement previously published results in that we focus on the temporal behavior of commercial robots. More specifically, our

investigation extensively analyzes the temporal properties and patterns of the transactions belonging to Web robots with the aim of identifying models able to represent and predict their behavior.

3 Methodological approach

Our study relies on the log collected on the European mirror of the SPEC Web site [8] for one year, from April 2009. The information recorded in the log according to the Extended Log File Format, refer to the main characteristics of each HTTP transaction processed by the site, such as, IP address of the client that issued the HTTP request, timestamp of the transaction, method and resource requested, user agent used by the client to issue the request.

The methodological approach adopted for the analysis of the Web log consists of various steps. In particular, after the identification of the transactions belonging to commercial Web robots, exploratory statistical techniques are applied to discover and highlight the main characteristics of these transactions and select the parameters that describe their temporal behavior. The investigation has then to focus on the distributions of these parameters to find out their properties, e.g., correlations, time dependence, and identify their temporal patterns. In particular, the analysis of the times between consecutive transactions of a given robot is the basis for the identification of its sessions, that is, the set of transactions issued within a given time interval. Sessions are typically intermixed with inactivity periods during which a robot does not issue any transaction. Temporal patterns are then studied as sequences of activity and inactivity periods. Through the application of numerical fitting techniques, models that predict these patterns, in terms, for example, of times elapsed between consecutive transactions and between consecutive sessions, are identified. These models can then be used to predict crawling activities and estimate the impact of the traffic of Web robots on the overall workload of a Web site.

4 Exploratory analysis

A preliminary processing of the log has shown that the site was visited by more than 19,000 different clients and approximately 18% of them visited the site only once during the entire year. In terms of traffic, the site served some five million HTTP transactions and generated about 130 Gbytes of data. Most of the transactions used the GET method and were successful, that is, the status code of the server responses was 2XX, whereas the number of bad transactions with status code 4XX was almost negligible. Moreover, most of

the referrer fields were empty, thus denoting navigation profiles characterized by a very limited degree of interactivity.

From a more detailed analysis of the IP addresses of the clients coupled with their corresponding user agents, we discovered that well-known commercial Web robots generated the majority of the traffic. In particular, as shown in Table 1, we ascribe about two million transactions to `msnbot`, that is, the robot employed by the Microsoft search engine – also known as `bingbot` after Microsoft officially introduced the Bing search engine – and about 1.5 million to `Googlebot`, the robot employed by the Google search engine. Apart

Table 1 Characteristics of the traffic generated by commercial Web robots.

	# transactions	# clients	Data volume [GB]
Microsoft	1,975,232	1,653	39.59
Google	1,424,054	533	47.85
Dotnet	313,861	2	8.03
Yahoo	237,328	262	6.01
Scirus	165,037	1	5.98
Baidu	7,978	327	0.14
Exalead	4,390	1	0.10

from these robots, most of the others are definitely less active. For example, this is the case of `Baiduspider` and `Exabot`, the robots of the Chinese search engine Baidu and of the French search engine Exalead, respectively, that are responsible for less than 8,000 transactions each. The number of different clients involved in the crawling activities varies significantly from organization to organization. For example, Scirus, the engine dedicated to search scientific information on the open Internet, employs one client only, whereas Microsoft relies on a very large number of clients often operating in parallel and cooperating in a crawling session. In terms of data volume, Google contributes for more than 40% of the data transmitted by the site, whereas Microsoft is responsible for less than 35% although its number of transactions is about 39% larger.

From the analysis of the resources requested by the Web robots considered in our study we discovered some interesting findings. More specifically, robots spread their requests over some 300,000 different resources whose popularity, as shown in Figure 1, varies significantly and does not follow any Zipf distribution. The `robots.txt` file, i.e., the file used by Web site administrators to specify the rules of operation of robots, and the index file of the site were retrieved more than 36,000 and 9,500 times, respectively, whereas most of the other resources were retrieved much fewer times, namely, once or twice for one third of the resources and up to 20 times for about three quarter. This is in line with the behavior expected by robots whose main goal is to index and keep fresh the entire content of a site. We have also discovered that resource popularity is independent of their size.

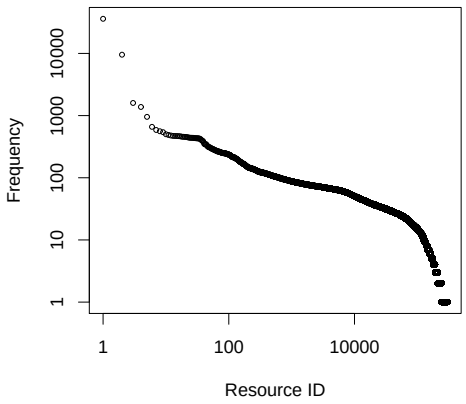

Fig. 1 Log-log scale plot of the popularity of the resources.

The rate at which a robot crawls a site depends on many factors, including, among the others, the type of content and how often it is modified as well as the crawling policy adopted. Our investigation has outlined the thorough and extensive crawling activities performed by the Google robots: over one year they downloaded at least once almost all the resources of the site. On the contrary, the Microsoft robots reached a much smaller set of resources, each characterized by a higher revisit rate, on average about 12 visits per resource, compared to five of the Google robots.

5 Temporal behavior

To study the temporal properties of the overall traffic generated by Web robots, we have first analyzed the interarrival times, that is, the time elapsed between two consecutive transactions of any robot. As shown in Fig. 2, the distribution is positively skewed with most of these transactions arriving very close to each other. Note that even though the plot does not span the entire range of variability of the interarrival times, whose maximum value is 1,818 seconds, it covers more than 99.99% of the observations. Indeed, the median and the third quartile of the distribution correspond to 5 and 10 seconds, respectively and 99.9-th percentile to 74 seconds. The application of numerical fitting techniques has shown that a lognormal distribution best fits the observed data as most of the mass is concentrated on small values. The location and shape parameters identified by fitting are equal to 1.74 and 0.92, respectively.

From the analysis of the behavior of the individual robots we have discovered that some robots visit the site rather regularly with transactions spread over the entire year, whereas others visit the site only sporadically, with transactions concentrated in some periods of the year. These different visit patterns are reflected in the interarrival times computed for the indi-

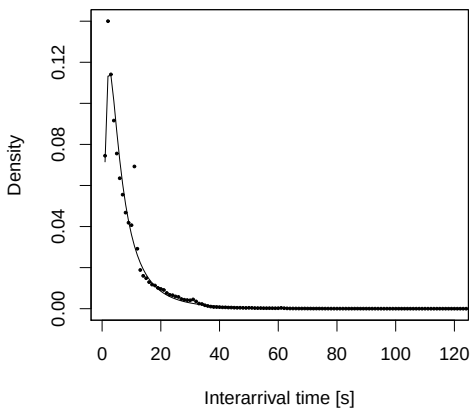

Fig. 2 Interarrival times of the transactions of all robots: measured (dotted curve) and fitted model (continuous curve).

vidual robots. As can be seen in Table 2, for 95% of the transactions of all robots, but Baidu and Yahoo, the interarrival times are rather small and do not exceed two minutes. Nevertheless, their maximum values are much larger, especially in the case of robots that sporadically visit the site. For example, the robot of Scirus is characterized by several inactivity periods of more than a week. On the contrary, the robots of Microsoft are almost always active: their maximum period of inactivity is about two hours.

Table 2 Statistics of the interarrival times, in seconds, of the individual robots.

					Percentiles			
	Mean	St. Dev.	Max	Skewness	25-th	50-th	75-th	95-th
Microsoft	15.8	28.1	7,420	31.0	3	9	29	56
Google	21.9	58.3	23,645	96.2	4	11	34	70
Dotnet	98.9	5,913.4	1,824,148	227.8	10	10	11	23
Yahoo	131.4	362.9	73,878	97.8	19	71	170	439
Scirus	154.6	10,063.8	2,512,479	169.7	60	60	61	119
Baidu	3,908.6	4,034.9	76,241	2.6	444	3,115	6,935	10,012
Exalead	3,031.3	29,592.2	956,147	15.2	10	10	10	22

To investigate the time dependence of the interarrival times of the individual robots, we computed the autocorrelation function with various lags. The transactions of some robots, e.g., Scirus, Yahoo, are not characterized by any time dependence, whereas this is not the case of other robots, e.g., Baidu, Google, Microsoft. Moreover, the slow decrease of the autocorrelation functions computed for these robots suggests the presence of long-term correlations or self-similar behavior that imply heavy-tailed distributions. This conclusion is confirmed by the Hurst parameter whose estimates are always larger than 0.6.

To increase the accuracy of the predictive models, we subdivided the transactions of each robot into sessions, that is, sequences of transactions whose interarrival times are below a certain threshold. In particular, our analysis

has suggested a 300 seconds threshold. The effect of this subdivision is a general decrease of the variability of the interarrival times within a session. This is especially true for the robots that tend to visit the site either periodically or sporadically. For example, the crawling activities of Dotnet robots resulted in some 7,000 sessions of about 45 transactions whose interarrival time is almost constant. The time between two consecutive sessions, that is, the intersession time, is at most one hour for about 90% of the sessions of all robots.

Table 3 summarizes the main characteristics of the sessions of the various robots. As can be seen, these characteristics vary from robot to robot, especially in terms of number of transactions per session and session duration. On the contrary, the average number of clients employed in a session is usually small, with the exception of Microsoft whose crawling activities rely on a pool of about 28 clients operating in parallel. It is also worth noting that some of the sessions of the most active robots span days. Finally, let us remark that about 0.35% of the transactions issued by robots belong to sessions consisting of one transaction only, that is, their interarrival times are larger than the fixed threshold.

Table 3 Average characteristics of robot sessions. Times are in seconds.

	Duration	Intersession	# transactions	# clients	# sessions
Microsoft	26,443.28	521.03	1,463.13	28.3	1,350
Google	10,845.72	637.75	308.97	1.3	4,609
Dotnet	773.19	3,951.84	45.10	1	6,958
Yahoo	804.85	529.02	9.09	1.1	26,088
Scirus	7,282.34	7,481.30	83.90	1	1,967
Baidu	140.62	4,831.72	1.24	1.24	6,413
Exalead	280.31	70,154.23	23.22	1	189

To obtain a more accurate prediction of the crawling activities of the robots, we have analyzed the intersession times by applying fitting techniques and we have discovered that these times can be modeled by a Pareto distribution, thus denoting a heavy tail.

6 Conclusions

Web robots are responsible of large fractions of the traffic received by the sites. To cope with their presence, it is then important to understand and predict their behavior. In this paper, we have analyzed the access log of the European SPEC Web site, to investigate the properties of the traffic generated by some commercial Web robots and outline the similarities and differences existing among them. More specifically, we have discovered that some robots employ in their crawling activities a large number of clients working

in parallel, whereas others rely on one client only. Moreover, not all robots tend to revisit the resources: a good fraction of the resources was retrieved only once or twice during the entire year. The analysis of temporal behavior of the robots has shown that most robots are characterized by regular access patterns in terms of sessions, number of transactions and times between consecutive transactions within a session. Crawling activities are intermixed with inactivity periods, whose duration follows some specific patterns. We can conclude that, despite these differences, the behavior of individual robots is well defined and, once a robot is identified, it is possible to predict its visit patterns.

As a future work, we plan to assess whether the characteristics of the traffic of commercial robots hold across sites. Based on the identification of the robots, we will then develop regulation policies that predict their behavior.

References

1. Calzarossa, M., Massari, L.: Analysis of Web logs: Challenges and Findings. In: K. Hummel, H. Hlavacs, W. Gansterer (eds.) Performance Evaluation of Computer and Communication Systems - Milestones and Future Challenges, *Lecture Notes in Computer Science*, vol. 6821, pp. 227–239. Springer (2011)
2. Dikaiakos, M., Stassopoulou, A., Papageorgiou, L.: An investigation of web crawler behavior: characterization and metrics. Computer Communications **28**(8), 880–897 (2005)
3. Doran, D., Gokhale, S.: Web robot detection techniques: overview and limitations. Data Mining and Knowledge Discovery **22**, 183–210 (2011)
4. Kwon, S., Kim, Y., Cha, S.: Web robot detection based on pattern-matching technique. Journal of Information Science **38**(2), 118–126 (2012)
5. Lee, J., Cha, S., Lee, D., Lee, H.: Classification of web robots: An empirical study based on over one billion requests. Computers & Security **28**(8), 795–802 (2009)
6. Lourenco, A., Belo, O.: Catching Web Crawlers in the Act. In: Proc. International Conference on Web Engineering, pp. 265–272 (2006)
7. Olston, C., Najork, M.: Web Crawling. Journal of Foundations and Trends in Information Retrieval **4**(3), 175–246 (2010)
8. SPEC Web site – European mirror: http://spec.unipv.it
9. Stassopoulou, A., Dikaiakos, M.: Web robot detection: A probabilistic reasoning approach. Computer Networks **53**(3), 265–278 (2009)
10. Tan, P., Kumar, V.: Discovery of Web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery **6**(1), 9–35 (2002)
11. Thelwall, M., Stuart, D.: 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)