Analysis of Web logs: Challenges and Findings

Maria Carla Calzarossa and Luisa Massari

Dipartimento di Informatica e Sistemistica
Università di Pavia
I-27100 Pavia, Italy
{mcc,massari}@unipv.it
http://peg.unipv.it

Abstract. Web logs are an important source of information to describe and understand the traffic of the servers and its characteristics. The analysis of these logs is rather challenging because of the large volume of data and the complex relationships hidden in these data. Our investigation focuses on the analysis of the logs of two Web servers and identifies the main characteristics of their workload and the navigation profiles of crawlers and human users visiting the sites. The classification of these visitors has shown some interesting similarities and differences in term of traffic intensity and its temporal distribution. In general, crawlers tend to re-visit the sites rather often, even though they seldom send bursts of requests to reduce their impact on the servers resources. The other clients are also characterized by periodic patterns that can be effectively represented by few principal components.

1 Introduction

The Web has become a phenomenon of growing social, economic and cultural importance and an essential component of the modern society that attracts million of users and accesses daily. On the Web, users distribute and share information and knowledge, conduct businesses, communicate, socialize and develop relationships. To discover, locate and retrieve the huge amount of information published on the Web, crawling has emerged as a key enabling technology [17].

Many applications and services rely on crawling. For example, to facilitate navigation and provide users with up-to-date information, search engines periodically crawl Web sites to index, group and cache Web content. Other applications crawl the Web for different legitimate or malicious purposes: to maintain a site, discover Web services, collect email addresses and personal information, extract business intelligence, exploit vulnerabilities.

Crawling employs programs, known as Web crawlers or robots, that automatically access and download Web pages without continuous involvement of human users. These programs are expected to comply with the Robot Exclusion Protocol [11], a standard that allows Web site administrators to specify, in the `robots.txt` file, the rules of operation of the crawlers. Nevertheless, some of them ignore the file and the rules, thus leading to potential performance problems as well as to privacy and security concerns [24]. It is then important to

identify the presence of both ethical and malicious crawlers as they might have a considerable impact on the infrastructures, thus hindering normal user accesses and causing damages and even economic losses.

Web access logs represent an important source of information to describe and understand Web server traffic and usage patterns as well as users behavior. Logs provide useful inputs to a large variety of engineering activities, ranging from the improvement of the site structure and organization, to the provisioning of personalized content, the development of recommendation systems, the selection of prefetching and caching policies, the formulation of content distribution and replication strategies. Moreover, the multiplicity of statistics, metrics and diagrams about the visitor traffic derived by the tools specialized in the analysis of the content of Web logs, can be used for commercial purposes, to develop, for example, customized marketing strategies or new business models or to attract advertisements.

In this paper we study Web servers access logs with the objective of modeling the access patterns of the visitors and identify typical navigation profiles as well as clients trying to compromise the servers by issuing malicious requests. The outcomes of this analysis could be used to develop proactive policies aimed at enhancing server availability, security and performance as well as to define the input of load generators used, for example, for benchmarking experiments.

Our study is experimental, that is, based on the analysis of the logs collected during more than one year on two Web servers. The choice of these servers is motivated by their characteristics, such as, potential users and traffic, that make them particularly suitable to assess our methodological approach. One server hosts an academic site mainly used by students and researchers of Computer Science [19], whereas the other hosts the European mirror of the SPEC (Standard Performance Evaluation Corporation) Web site [21] whose content is of interest for the entire community of IT specialists.

The paper is organized as follows. Section 2 briefly discusses the state of the art in the area of the analysis of Web workload. The main characteristics of the two Web servers considered in our study and the results of the preliminary exploratory analysis of their traffic are presented in Section 3. The methodological approach applied for the identification of the navigation profiles and its outcomes are described in Section 4. Finally, Section 5 concludes the paper by pointing out the major findings and challenges encountered in the analysis of Web logs.

2 Related work

Logs have been used as the basis of many studies since the early days of the Web (see e.g., [2, 3, 8, 14, 15, 20, 25]). Most of these studies focused on the characteristics of the workload being processed by the servers and used the information extracted from their access logs to describe the properties of the workload in terms of various metrics, such as, document types and popularity, file size distribution, concentration of references, inter-reference time. In particular, Arlitt and Williamson presented in [2] ten invariants, i.e., characteristics common to

all the sites that are likely to persist over time. A more recent paper [25] has actually shown that, even though the Web traffic has dramatically increased in ten years, the same invariants can accurately capture its properties.

Other studies focused on the analysis of Web logs with the objective of identifying the users behavior. Graph-based models were proposed to represent the navigation profiles of the customers of e-commerce sites [16].

As Web crawlers are responsible of a large fraction of the Web traffic, several authors addressed specifically their attention to the identification and characterization of this type of traffic (see, e.g., [1, 4, 13, 22, 23]. Some of these studies analyzed its overall characteristics, whereas others took into account some more specific aspects. For example, Dikaiakos et al. [4] characterized and compared the behavior of the crawlers of five popular search engines by analyzing access logs collected on various academic Web servers. The study introduced a set of metrics that provide a qualitative description of the behavior of these crawlers. Lee et al.[13] analyzed a very large number of transactions recorded by a commercial server over a 24 hours period to investigate the characteristics of various Web robots. Metrics associated with HTTP traffic features and resource types were then used for the classification of the robots.

In [23] Tai and Kumar introduced the concept of Web robot sessions and used some access features derived from the Web server logs for their identification and classification. Sessions considered in the framework of search engines were studied in [6] where a multidimensional approach was applied to Web search logs to derive a systematic classification of users as humans or robots.

On the contrary, the classification of Web robots presented in [5] took into account the influence exerted by the goals and the functions performed by robots on their navigational patterns. Mouse clicks streams were used in [18] to infer whether the traffic source is a human or a robot.

Our study complements these studies because of the perspective adopted for the analysis of Web logs. More specifically, starting from the access patterns of the individual visitors, we apply various types of statistical techniques to highlight differences, similarities and peculiarities in the behavior of Web crawlers and human users.

3 Data sets

The data sets used in our investigation are represented by the logs collected on two Web servers, hosting an academic site at the University of Pavia in Italy and the European SPEC mirror site, respectively.

Both servers record the details of the HTTP transactions being processed according to the Extended Log File Format [7]. In particular, a transaction is described by the IP address of the client that issued the HTTP request, the timestamp of the transaction, the method and resource requested, the status code of the server response, the number of bytes transmitted by the server, the referrer of the previous site visited by the client, and the user agent, that is, the browser used by the client to issue the request.

As a first step, we performed an exploratory analysis of the Web logs to derive from these large volumes of information some preliminary insights in the characteristics of the workload of these Web servers. Note that the information stored in the two logs files accounts for about 50Mbytes and 970Mbytes, respectively.

Table 1 summarizes the main characteristics of the transactions processed by the two servers. As already pointed out, the sites considered in our study

Table 1. Main characteristics of the servers workload.

	Academic server	SPEC server
Measurement interval	14 months	12 months
Total number of transactions	239,081	5,098,621
Total number of 2xx transactions	144,081	3,977,929
Total number of 4xx transactions	27,197	143,863
Total GBytes transmitted	8	129
Number of clients	7,940	19,135
Number of one-time clients	1,034	3,364

differ in terms of potential users, hence, their traffic intensity is quite different. In a period of approximately 14 months, since April 2009, the academic server processed some 560 HTTP transactions per day and transmitted 18.5Mbytes of data. The SPEC server, with its 14,000 transactions per day, was much busier. In 12 months, it processed more than five million transactions and transmitted 129Gbytes of data in total.

The transactions with status code 2xx refer to the requests of the client successfully received, understood and accepted by the server, whereas the 4xx status code refers to bad requests due to client errors. As can be seen, the large majority, i.e., 78%, of the transactions processed by the SPEC server was successful and bad transactions were very few: their fraction did not reach the 3%. It is also worth noting that 1,556 requests could not be processed because of temporary server errors. On the contrary, for the academic server about 60% of the requests were successful but a good fraction of requests, i.e., 11.4%, was bad. Most of these requests were to non-existing resources, e.g., various types of PHP scripts mainly developed to exploit server vulnerabilities.

Another clear indicator of the different behavior of the two servers is represented by their hourly traffic. As Fig. 1 shows, the traffic over the 24 hours of the academic server follows a typical diurnal pattern with its highest peak at noon, whereas for the SPEC mirror the traffic is basically flat with transactions evenly distributed and no significant difference between day and night. As we will explain in more details later on, this is mainly due to the very strong presence of crawlers that are responsible for the majority of the traffic of this server.

In what follows, we focus on the analysis of the visitors identified by means of the IP addresses specified in the logs. Although these addresses do not univocally identify individuals because of the dynamic assignment of addresses and of their

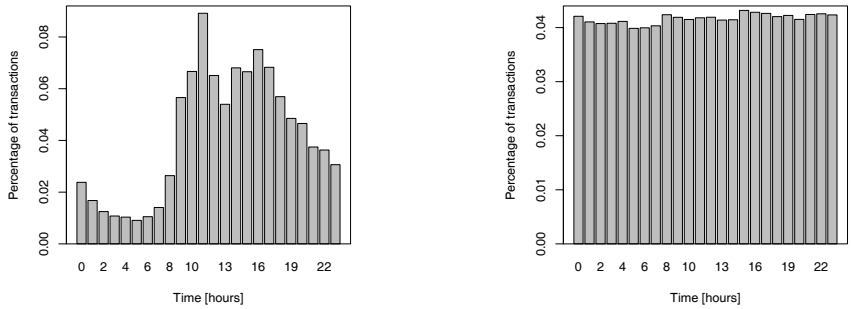


Fig. 1. Percentage of transactions processed by the academic server (a) and by the SPEC server (b) over the 24 hours.

management within organizations, they seemed appropriate for the purposes of our work.

The behavior of the clients in terms of total number of requests and total number of bytes transmitted by the servers during our measurement interval are shown in Fig. 2 and in Fig. 3, respectively. Note that the y axes of the plots are in log scale. The clients of the academic server issued on average 30 requests each, even though one of them issued as many as 11,953 requests. Moreover, three-quarter of the clients sent at most 20 requests, and 13% one request only, that is, they are the so-called "one-timers". The clients of the SPEC server exhibit a rather different behavior: one client is responsible of the 7.2% of the total traffic of this server and some 30 clients account for half of the traffic. Moreover, three-quarter of the clients issued 27 requests at most.

In terms of bytes, the number of bytes downloaded by three-quarter of the clients of both servers did not exceed 375Kbytes, nevertheless, few clients downloaded most of the bytes transmitted by the servers. For example, one client downloaded from the SPEC server as many as 9.2Gbytes. It is worth noting that about 5% of the clients of the academic server did not download any byte because their HTTP requests either used a HEAD method or specified an "`If-modified-since`" header and the corresponding pages were not transmitted as they were not modified by the server since their latest download. These requests represent about 4% of the workload of this server. In summary, on average clients of the academic server downloaded 1Mbytes of data each, compared to 6.8Mbytes of the SPEC clients.

Before studying the navigational profiles of the clients, we did some preprocessing of the log files to identify "well-known" crawlers and assess their impact on the overall traffic of the servers. More precisely, we recognized clients

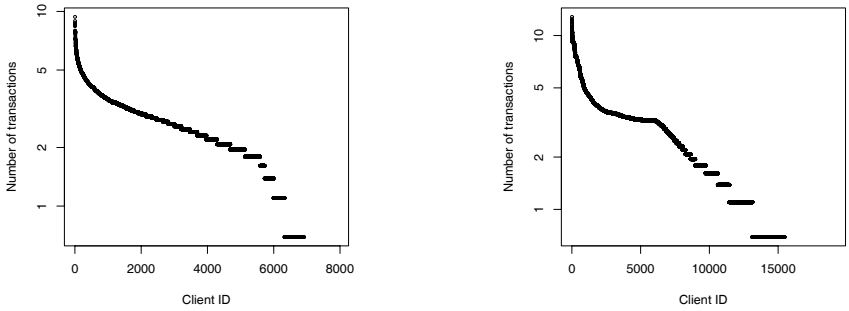


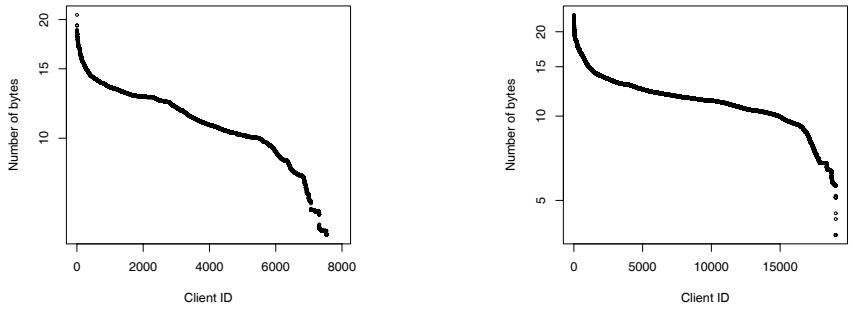
Fig. 2. Total number of transactions per client of the academic server (a) and the SPEC server (b).

as being crawlers, either because they accessed the `robots.txt` file or because of some explicit information in the user agent field.

With this pre-processing, we classified as crawlers about 16% of the clients of the SPEC server, namely, 3,108, and 12% of the clients of the academic server, namely, 974. It is interesting to outline that in terms of traffic, while crawlers account for about 15% of the traffic of the academic server, the situation is completely different on the other server, where crawlers are responsible for the vast majority of its traffic, namely, for about 4.8 million requests, out of approximately five million, and of 122Gbytes of data, out of 129Gbytes. The crawlers of three major search engines, i.e., Google, Microsoft and Yahoo, emerged as the top crawlers on both servers as they generated about 80% of their traffic. Table 2 presents the main characteristics of the traffic produced by these three top crawlers on the SPEC server.

Table 2. Main characteristics of the traffic of the three top crawlers of the SPEC server.

	Google	Microsoft	Yahoo
Total number of transactions	1,429,954	2,147,582	238,202
Total number of 2xx transactions	1,156,072	1,434,838	227,640
Total number of 4xx transactions	36,952	34,742	1,150
Total GBytes transmitted	48	46	6
Number of clients	535	958	268



(a) (b)

Fig. 3. Total number of bytes transmitted per client by the academic server (a) and the SPEC server (b).

From now on, we investigate separately the behavior of clients identified as crawlers and of the remaining clients, that might include human users as well as crawlers that did not identify themselves mainly because of their malicious intentions.

4 Navigation profiles

The methodological approach followed for the analysis and characterization of the navigation profiles of the visitors of the two Web servers is based on the selection of the parameters that describe their behavior and the application of various statistical techniques to uncover differences and similarities among profiles.

The parameters used to describe the navigation profiles of the individual clients were related to the traffic generated by the clients and their temporal distribution. More specifically, the inter-reference time, that is, the time elapsed between two consecutive requests of a given client, is a good metric to describe the profiles in terms of traffic intensity.

Figure 4 shows the details of the distributions of the inter-reference times measured on the academic server for all the requests of the crawlers and of the other clients. The average inter-reference time of crawlers is much larger than for the other clients; the 90-th percentile of the distribution is about 310,000 seconds, that is, more than 86 hours, compared to 22 seconds for the other clients.

This investigation has shown that, whenever the inter-reference time was larger than 240 seconds for the crawlers and 120 seconds for the other clients, a navigation session was basically over, that is, the client will start a new session. A session is then defined as the sequence of requests issued by a client and

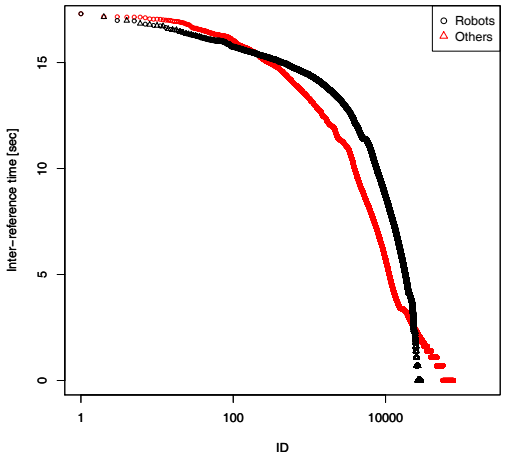


Fig. 4. Inter-reference times for the clients identified as crawlers and for the remaining clients of the academic server.

characterized by inter-reference times smaller than the selected thresholds. As a consequence, the navigation profile of a client can be described in terms of number of sessions and their duration, number of requests per session and inter-session time, that is, the time between two consecutive sessions of a given client. Table 3 presents the average characteristics of the navigation profiles of the crawlers and of the other clients of the SPEC server in terms of these parameters. Crawlers sessions were bigger in terms of average duration and average number of

Table 3. Main characteristics of the navigation profiles of the SPEC clients.

	Crawlers	Others
Number of sessions per client	462.44	2.62
Number of transactions per session	29.43	13.68
Session duration [sec]	467.09	22.40
Inter-session time [sec]	18,697.00	468,279.00
Number of one-transaction sessions	869,678.00	14,953.00
Number of one-session clients	485.00	8,752.00

transactions. Nevertheless, these sessions were characterized by a large variability across clients. The standard deviations of these parameters were an order of magnitude larger than the corresponding averages. From these results, it appears

that crawlers that identify themselves tend to behave and do not send their requests in bursts to reduce their impact on the server resources.

In terms of re-visit patterns, crawlers re-visit the site very often: on average every 31 minutes and with requests distributed across many sessions. This is mainly related to the use of some sort of distributed crawling policies to speedup the process. In details, the majority of the crawlers (i.e., 88%) re-visited the site at least three times, whereas this was the case of very few of the other clients. It is also interesting to outline that after a session with a large number of transactions, crawlers were likely to re-visit the site very soon, that is, their inter-session times were small. For example, clients identified as Google crawlers sent as many as 6,470 requests each, distributed across as many as 870 sessions and spanning over a time period of more than four months. On the other hand, we have discovered that the number of sessions characterized by one transaction was not negligible for both crawlers and the other clients, namely, 69% and 44% of the total number of sessions, respectively, and about one third of the clients characterized by one session were the so-called "one-timers".

Another metric used to describe the navigation profiles was related to distribution of the requests across months, days of the month and hours of the day. In particular, for each client we counted the number of requests issued in each of these time periods. We then obtained a tuple of N parameters, N being equal to 69 for the academic server and 67 for the SPEC server, because its logs contained the measurements of 12 months instead of 14.

These parameters allows us to identify clients with similar patterns and discover periodic patterns, i.e., clients visiting the site regularly, for example, in the first day of the month at noon or in last day of May and August.

To make this large number of parameters more manageable, we applied various multivariate statistical techniques in combination [9, 10]. The Principal Component Analysis was used to linearly transform these correlated parameters into a much smaller set of uncorrelated parameters, the principal components. The Correspondence Analysis was used to visually display the clients and the parameters used for their description. Finally, the application of hierarchical clustering techniques allowed us to discover groups of clients with homogeneous behavior.

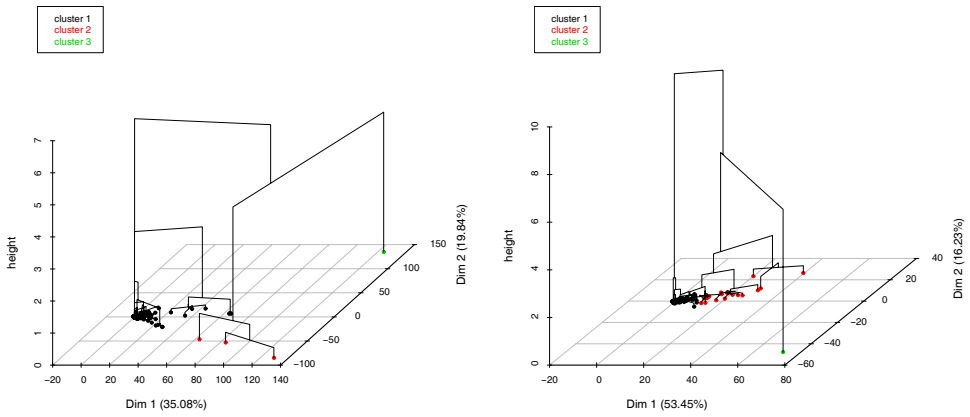
The rest of this section is dedicated to present the classification of the hourly patterns of the crawlers and of the other clients of the academic server, each described in terms of number of requests issued in the various hours of the day, i.e., 24 parameters. Moreover, to take into account the distribution of the requests of each client across months and days of the month, we used two additional parameters, that is, the total number of months and the total number of days during which the client sent at least one request. Note that for this analysis we used the FactoMineR package [12].

The application of the Principal Component Analysis to both sets of clients described by these 26 parameters has shown that few principal components could summarize very well the variability in the original data. More specifically, the first two principal components computed for the other clients accounted for 55% of their variance, whereas in the case of crawlers the principal components

could capture their behavior even better. The first two principal components accounted for approximately the 70% of their variance and four principal components covered 80% of the variance. The weights associated with the first principal component are about equal. This means that each of the parameters is equally represented in the linear composition, i.e., this component represents crawlers that do not differentiate their traffic among the various hours of the day. On the contrary, the second principal component represents the contrast between day traffic and night traffic.

For the other clients, the first principal component mainly describes the traffic sent during business hours, i.e., between 8am and 5pm, whereas the second component represents the difference between day and night traffic and traffic sent across few months and days.

We then applied hierarchical clustering techniques to the data of both sets of clients represented in the principal components space. The partitions in three clusters obtained for both sets are shown in Fig. 5. Each plot represents the



(a) (b)

Fig. 5. Clusters obtained for the other clients (a) and for the crawlers (b) of the academic server.

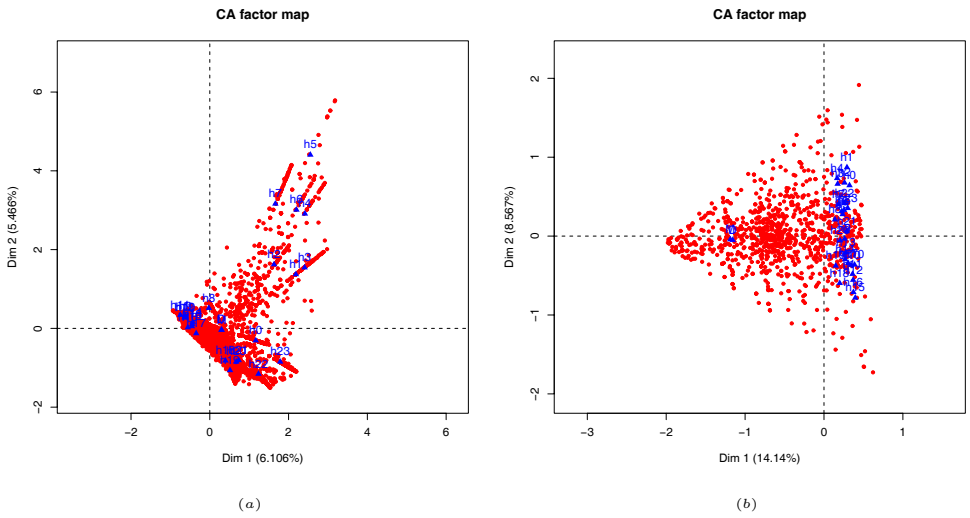
projection of the data in the space of the first two principal components. Thus, these diagrams highlight the structure of the data and their similarities and differences. Clients close to each other in this space were similar to each other in their original data.

From the figure, we can notice that the behavior of the other clients is quite homogeneous: two of the three clusters consist of one and three clients, respectively, while the remaining clients belong to one big tight group. It is worth

noting that the client belonging to cluster 3 is characterized by requests concentrated in one month and in two days. A further analysis of this client has shown that it was probably a crawler that did not identify itself as being such: all its HTTP requests used the HEAD method and specified the root of the site (i.e., /) as resource. On the contrary, the three clients of cluster 2 are actually human users, whose behavior is fully explained by the first principal component: all their requests were issued during the working hours only.

For the crawlers, clustering identified one very persistent crawler (belonging to cluster 3) and two other groups of crawlers with requests distributed over the 24 hours, sometimes in the day and other times at night. It is interesting to outline that many of the Google crawlers were grouped in cluster 2.

All these conclusions are also supported by the diagrams of Fig. 6 obtained by applying the Correspondence Analysis. This geometric representations display the relationships between clients and parameters. and point out their mutual influence and ease to identify visually clustered observations. As can be seen, while the diagram obtained for the crawlers does not show any specific association between them and the parameters used for their description, there is a stronger association between some of the other clients and the parameters describing night traffic. Note that not to clutter the presentation, we did not plot the identifiers of the clients, represented by red dots.



(a) (b)

Fig. 6. Correspondence Analysis maps of the other clients (a) and of the crawlers (b) of the academic server.

5 Conclusions

The analysis of Web logs is very useful to discover interesting properties in the traffic of the servers and in the behavior of their clients. This analysis becomes rather challenging especially for very busy servers that receive many requests by many different clients. Our study focused on the characterization of the access patterns and navigation profiles of the clients of two Web servers. The traffic of one of the servers was heavily dominated by some very persistent crawlers. Nevertheless, on both servers we have noticed that, despite the intensity of their traffic, clients that identify themselves as being crawlers tend to behave and avoid sending bursts of requests. The application of various statistical techniques in combination has allowed us to highlight similarities and differences among the navigation profiles of the clients. The other clients, that is, human users and unidentified crawlers, are characterized by a rather homogeneous behavior. Nevertheless, while some clients return to the sites periodically to download the pages available on the servers, some others visit the site with the only intention of discovering vulnerabilities and compromising the servers.

This work has a number of possible extensions and improvements. For example, it will be necessary to develop more reliable criteria to identify and classify clients that visit the sites for malicious purposes. Scrapers are an example of these clients in that they send legitimate requests with the objective of automatically creating copies of Web sites to be used for malicious purposes, such as, phishing. Moreover, starting from these results we plan to develop proactive policies aimed at improving the security of Web sites.

References

1. Almeida, V., Menascé, D., Riedi, R., Peligrinelli, F., Fonseca, R., Meira, W., Jr.: Analyzing Web robots and their impact on caching. In: Proc. of the Sixth Web Caching and Content Delivery Workshop (2001)
2. Arlitt, M.F., Williamson, C.L.: Web server workload characterization: the search for invariants. In: Proc. of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. pp. 126–137 (1996)
3. Crovella, M., Bestavros, A.: Self-similarity in World Wide Web traffic: evidence and possible causes. IEEE/ACM Trans. on Networking 5(6), 835–846 (1997)
4. Dikaiakos, M.D., Stassopoulou, A., Papageorgiou, L.: An investigation of web crawler behavior: characterization and metrics. Computer Communications 28(8), 880–897 (2005)
5. Doran, D., Gokhale, S.: Discovering new trends in web robot traffic through functional classification. In: Proc. of the International Symposium on Network Computing and Applications. pp. 275 –278. IEEE Computer Society (2008)
6. Duskin, O., Feitelson, D.G.: Distinguishing humans from robots in web search logs: preliminary results using query rates and intervals. In: Proc. of the Workshop on Web Search Click Data. pp. 15–19. ACM (2009)
7. Hallam-Baker, P.M., Behlendorf, B.: Extended Log File Format. W3C Working Draft WD-logfile-960323 (1996)

8. Iyengar, A.K., Squillante, M.S., Zhang, L.: Analysis and characterization of large-scale Web server access patterns and performance. World Wide Web 2(1-2), 85–100 (1999)
9. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Data Analysis – Sixth Edition. Pearson Prentice Hall (2007)
10. Jolliffe, I.T.: Principal Component Analysis – Second Edition. Springer (2002)
11. Koster, M.: A method for Web Robots control. Network Working Group - Internet Draft (1996)
12. Lê, S., Josse, J., Husson, F.: FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software 25(1), 1–18 (2008)
13. 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)
14. Mahanti, A., Williamson, C., Wu, L.: Workload characterization of a large systems conference Web server. In: Proc. of the Seventh Annual Communication Networks and Services Research Conference. pp. 55–64. IEEE Computer Society (2009)
15. Menascé, D.A., Almeida, V.A.F., Riedi, R., Ribeiro, F., Fonseca, R., Meira, Jr., W.: A hierarchical and multiscale approach to analyze E-business workloads. Performance Evaluation 54(1), 33–57 (2003)
16. Menascé, D.A., Almeida, V.: Capacity Planning for Web Services: metrics, models, and methods. Prentice Hall (2001)
17. Olston, C., Najork, M.: Web Crawling. Journal of Foundations and Trends in Information Retrieval 4(3), 175–246 (2010)
18. Park, K., Pai, V.S., Lee, K.W., Calo, S.: Securing web service by automatic robot detection. In: Proc. of USENIX '06. pp. 23–23. USENIX Association (2006)
19. Performance Evaluation Group Web site – University of Pavia: http://peg.unipv.it
20. Pitkow, J.E.: Summary of WWW characterizations. World Wide Web 2(1-2), 3–13 (1999)
21. SPEC Web site – European mirror: http://spec.unipv.it
22. Stassopoulou, A., Dikaiakos, M.D.: Web robot detection: A probabilistic reasoning approach. Computer Networks 53(3), 265–278 (2009)
23. Tan, P.N., Kumar, V.: Discovery of Web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery 6(1), 9–35 (2002)
24. 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)
25. Williams, A., Arlitt, M., Williamson, C., Barker, K.: Web workload characterization: Ten years later. In: Xueyan Tang, J.X., Chanson, S.T. (eds.) Web Content Delivery, pp. 3–21. Springer (2005)