Characterizing Crawler Behavior from Web Server Access Logs

Marios Dikaiakos[1], Athena Stassopoulou[2], and Loizos Papageorgiou[1]

[1] Department of Computer Science
University of Cyprus
PO Box 20537
Nicosia, Cyprus
{mdd, loipap}@ucy.ac.cy
http://www.cs.ucy.ac.cy/mdd
[2] Dept. of Computer Science
Intercollege
P.O. Box 24005
Nicosia, Cyprus
stassopoulou@cytanet.com.cy

Abstract. In this paper, we present a study of crawler behavior based on Web-server access logs. To this end, we use logs from five different academic sites in three countries. Based on these logs, we analyze the activity of different crawlers that belong to five Search Engines: Google, AltaVista, Inktomi, FastSearch and CiteSeer. We compare crawler behavior to the characteristics of the general World-Wide Web traffic, and to general characterization studies based on Web-server access logs. We analyze crawler requests to derive insights into the behavior and strategy of crawlers. Our results and observations provide useful insights into crawler behavior and serve as basis of our ongoing work on the automatic detection of WWW robots.

1 Introduction

Log analysis and World-Wide Web characterization have been the target of intensive research in recent years [7, 6, 2, 3]. Web characterization results have provided significant insights into Web usage, performance analysis, infrastructure design, etc. Machine learning and data mining techniques have been applied to process logs in order to mine user profiles, communities of pages, patterns of use, and guide the improvement of Web design, etc. [8]. So far, however, most studies have focused on general Web traffic. Very little emphasis has been put on the characterization of automated Web clients like robots or crawlers, and the contribution thereof to WWW workloads. *Web crawlers* are programs that traverse the hypertext structure of the Web starting from a "seed" list of hyperdocuments and recursively retrieving all documents accessible from that list [1]. The number and the variety of active robots operating on the Internet increases continuously, resulting to a noticeable impact on WWW traffic and Web-server activity.

Log Acronym	CS-UCY	CC-UCY	ICS-FORTH	SL-NTUA	CSE-TOR
Log Origin	CS, UCY	CC, UCY	ICS, FORTH	SOFTLAB, NTUA	CSE, U. of Toronto
Country Code	CY	CY	GR	GR	CA
Log Duration (days)	176	114	45	58	42
Starting Date	11/9/01	15/1/02	11/3/02	1/1/02	13/2/02
Ending Date	6/3/02	9/3/02	25/4/02	27/2/02	27/3/02
Log Size (MB)	184.8	150.7	81.8	525.9	243.7
Total Requests	1,767,101	1,467,266	786,300	2,724,074	2,565,214
Distinct URL's Requested	69,918	47,751	58,225	102,088	48,229
Avg Requests/Day	10,040.35	12,850.46	17,473.33	46,966.8	61,076.52
Bytes Transferred (MB)	23,618.53	18,946.92	8,021.12	745,387.42	36,234.55
Avg Bytes/Day (MB)	134.20	166.20	178.25	12,851.51	862.73

Table 1. Summary of access log characteristics.

In this paper, we seek to characterize the activity of Web crawlers, gain an insight into their behavior, and identify common characteristics of different crawlers. Investigating and understanding crawler activity is important as it enables researchers to:(i) estimate the impact of robots on the workload and performance of Web servers; (ii) investigate the contribution of crawlers in WWW traffic; (iii) discover and compare the strategies employed by different crawlers to reap resources from the Web and (iv) model the activity of robots to produce synthetic crawler workloads for simulation studies. Finally, characterization is the basis for the automatic detection of robots. For the purposes of our study, we concentrate on the characterization of five different crawlers, four of which belong to well-known search engines: *Google* (http://www.google.com), *AltaVista* (http://www.altavista.com), *Inktomi* (http://www.inktomi.com), and *FastSearch* (http://www.fastsearch.com). The fifth crawler belongs to *CiteSeer*, also known as "ResearchIndex," the Digital Library and Citation Index of NEC Research Institute (http://www.citeseer.nj.nec.com).

We employ and analyze access logs from five different academic sites in three countries: (a) The University of Cyprus; one set from the departmental Web server of the Computer Science Department (log acronym *CS-UCY*) and one set from the main University Web server (*CC-UCY*); (b) The Institute of Computer Science, Foundation of Research and Technology, Hellas in Greece (*ICS-FORTH*); (c) The Software Engineering Laboratory server at the National Technical University of Athens, Greece (*SL-NTUA*); (d) The departmental server of Computer Science and Engineering at the University of Toronto, Canada (*CSE-TOR*). These logs were given to us under a non-disclosure agreement in order to protect the privacy of end-users accessing the respective sites. The logs capture a period spanning from the fall of 2001 to the winter of 2002; log durations range from 42 to 176 days. Overall, these logs capture a total of 9,3 million HTTP requests for 326,211 distinct URL's, and a total of 812 GB of transferred data. In Table 1 we present an overview of the log-suite employed in this paper. To process the access-logs, we designed and developed ALAN (A Log ANalyzer). ALAN is a library written in JAVA that provides classes and methods for pre-processing and filtering access-logs, identifying the IP addresses of known crawlers. ALAN produces output compatible to Matlab.

Log Acronym	CS-UCY	CC-UCY	ICS-FORTH	SL-NTUA	CSE-TOR
% of total requests	10.32 %	9.48 %	12.67 %	4.02 %	10.18 %
% of bytes	5.72 %	5.11 %	4.33 %	0.08 %	5.84 %

Table 2. Contribution of selected crawlers to Web-server activity.

Response Codes	2xx	304	3xx (except 304)	4xx
All clients	72.26%	16.47 %	1.98%	9.29%
Google	41.86%	42.21%	3.31%	16%
Inktomi	33.73%	33.14%	7.4%	25.7%
AltaVista	80.18%	0%	0.53%	19.28%
Fastsearch	52.58%	33.17%	2.25%	11.99%
CiteSeer	59.59%	1.09%	4.21%	35.10%

Table 3. Percentage of HTTP responses to selected crawlers over all logs.

The characterization of our logs is given in Section 2 and a summary of our conlucions in Section 3. An extended presentation of our analysis is given in [5].

2 Characterizing Crawler Behavior

Considering all crawlers and logs, we end up with a total of 792,285 crawler-induced requests generating a traffic of 5GB of data; individual crawlers generate HTTP traffic of 1.1 to 33.52 MB/day on each Web server examined. Collectively, the activity of the five crawlers represents the 8.51% of the total number of requests included in our logs and the 0.65% of the bytes transferred. The impact of the five crawlers on each individual Web server is presented in Table 2. From this table we can see that in four out of the five sites, the five crawlers are responsible for the 9.48-12.67% of the total incoming requests and for the 4.33-5.84% of outgoing traffic, which represent a sizable proportion of the overall HTTP activity in these particular servers. Crawler contribution to the outgoing traffic in the fifth site (SL-NTUA) is negligible; this is because the SL-NTUA server hosts very large and very popular multimedia files, which are of no interest to crawlers.

2.1 HTTP-traffic Characteristics

An analysis of general HTTP traffic captured by our logs discovers trends similar to those published in the literature [2, 7]. These trends change, however, if we focus on traffic stemming from the five crawlers of our study. Almost 100% of all HTTP requests are `GET`'s. The percentage of different response codes for the five crawlers examined, averaged over all five logs of our log-suite, are presented in Table 3: crawlers implementing caching, such as Google, Inktomi, and FastSearch, issue cache-validation commands at a rate much higher than the rate observed in the general population of WWW clients: 42.21%, 33.14% and 33.17%

versus 16.47%. From Table 3 we can observe also that responses to crawler requests exhibit a proportion of `4xx` error codes higher than the observed rate for all clients. Most of the error codes are due to unavailable resources "`404 Not found`." The higher rate of `4xx` codes can be explained by the fact that human users are able to recognize, memorize and avoid erroneous links, unavailable resources, temporarily unavailable sites etc. It is the (rational) behavior and choices of those users that determine the all-clients characterization. A crawler should try to minimize the number of HTTP requests that lead to `4xx` replies, as these represent a mere overhead in its operation.

2.2 Resource Referencing Behavior

Classification of requested URL resources by file type For most Web sites, the resources that receive the overwhelming majority of requests are text (`text/plain`, `text/html`) and image files (`image/jpeg`, `image/gif`, etc.) [7, 2]. The remaining content types constitute a relatively small portion of requested URL resources (postscript and PDF, audio and video, scripts, applets). However, the mixture of content types requested may vary dramatically from site to site according to the site's design and the profile of its user base. In our log-suite, over 90% of all requests in four out of the five logs target text or image resources.

The situation is different if we focus on requests arising from the five crawlers studied. Text-file requests represent the 71.67-97.22% of total requests, whereas requests for image resources are practically non-existent. Finally, crawlers such as Google and NEC's CiteSeer, which try to index other than textual documents where available (e.g., postscript, pdf, compressed files), pursue the retrieval of the corresponding URL resources more aggressively than the general population of WWW clients.

Distinct requests When studying the patterns of URL requests arriving at a particular Web server, it is interesting to estimate the percentage of *separate* (distinct) resources requested over the *total* number of requested resources [2]. Taking into account requests from all clients captured in our log suite, gives a small percentage of distinct requests between 1.88-7.4%. This observation agrees with prior Web characterization studies (e.g., [2]). Nevertheless, it changes drastically if we focus on requests arriving from IP addresses that belong to individual crawlers: percentages increase by an order of magnitude and up to 100%. This is because the percentage of distinct over total requests coming from a crawler depends on the (typically limited) number of visits this particular crawler pays to the Web site at hand, within the time frame captured by the access log under study. For instance, in the period captured by the ICS-FORTH log, CiteSeer visits ICS's Web site only once and the corresponding percentage is 100%.

Resource popularity and Concentration of Requests *Popularity* of a URL resource is measured as the proportion of requests accessing the resource over the total number of requests reaching its Web site. A large number of Web

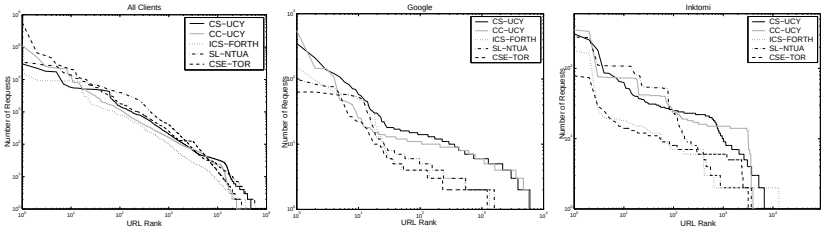

Fig. 1. Resource popularity (All clients, Google, Inktomi)

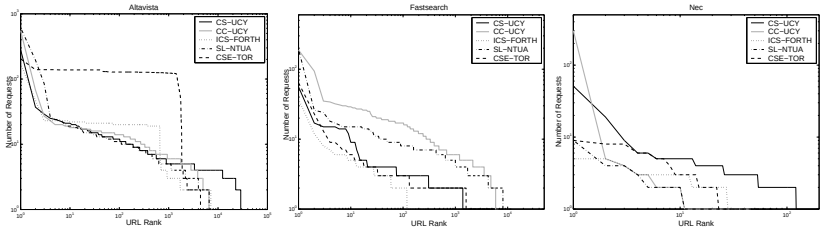

Fig. 2. Resource popularity (AltaVista, FastSearch, NEC's CiteSeer)

characterization studies showed that resource popularity follows a *Zipf*-like distribution [7]. To test if a distribution is Zipf-like, we produce a log-log plot of the number of requests for each resource versus the resource's popularity rank. Resources are placed on the horizontal axis in decreasing order of rank; if the distribution is Zipf-like, the graph should appear linear with a slope of -0.5 to -1. Figures 1 and 2 present popularity plots for the distinct resources in our logs versus the rank of these resources.

We calculate the popularity of Web resources based on *crawler requests only* and plot our findings in the diagrams of Figures 1 and 2. Statistical observations for the case of crawler-logs are harder as the total number of requests issued by each crawler during our period of observation is small, i.e., one to two orders of magnitude smaller than the total number of requests coming from all clients. These "crawler" plots, however, provide some insights into crawler referencing patterns: many of the popularity diagrams display a step-wise shape with URL resources clustered into smaller subsets of equal popularity. In other words, the frequency of visits of a crawler on a particular Web site varies for different subsets of resources within the site.

Popularity studies for WWW access show that URL requests are highly concentrated around a small set of resources. The *concentration* of requests can be expressed by sorting the list of distinct URL resources requested into decreasing order of rank, and then plotting the cumulative frequency of requests versus the *fraction* of the total URL resources requested. Previous Web characterization studies have shown that resource-popularity is highly concentrated [7]. Our logs

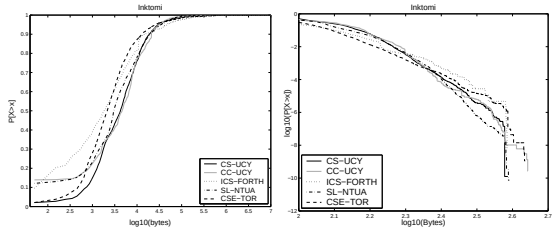

Fig. 3. Size distribution for successful responses: cumulative histogram and heavy tail of Google.

Log acronym	CS-UCY	CC-UCY	ICS-FORTH	SL-NTUA	CSE-TOR
All clients	13.69	13.22	10.45	280.20	14.46
Google	30.34	13.45	19.83	7.17	38.67
Inktomi	4.31	3.31	1	6.91	3.12
AltaVista	3.02	9.24	2.85	5.12	6.62
FastSearch	1.67	6.30	3.83	6.45	28.89
CiteSeer	32.37	7.49	51.75	17.12	115.82

Table 4. Average size of HTTP responses (in KB).

exhibit a high concentration of references on a small subset of unique resources: 10% of separate (distinct) URL resources attract a 75-90% of all requests. Focusing on crawler-induced requests, however, it becomes apparent that crawler-references are *not* highly concentrated around a small set of URL resources. For instance, in the vast majority of cases, 50% of the most popular resources attract between 60-80% of all crawler-induced requests. This behavior is expected since crawlers typically try to reach as many resources as possible when visiting a particular Web site.

2.3 Size Distributions

Several Web characterization studies have explored the size distribution of URL resources and HTTP messages and showed that average resource sizes are relatively small with an average size of 4 to 8 KB for HTML and 14 KB for images, and a wide variability [4, 7]. Resource-size distribution is typically captured by a hybrid model that describes the body of the distribution with a *lognormal* and the tail with a *heavy tailed* (*Pareto*) distribution [4, 7]. Analyzing our log-suite reveals that the mean size of an HTTP response across all logs and for all clients is 91.53 KB. The average response size for all clients is very high if compared to observations of other studies because of the very large files downloaded from the SL-NTUA server; omitting the SL-NTUA logs results to a drop of the mean size to 13.5 KB.

If we concentrate on crawler traffic only, we get a mean HTTP-response size of 7.03 KB. Evidently, there is a high variability in actual response sizes to the same crawler across different logs, as it can be seen from Table 4. Further evidence

for the response-size variability can be derived if we compare the mean to the median values of HTTP responses. For instance, looking at the ICS-FORTH logs, the mean transfer size of HTTP responses that carry HTML resources is 5.53 KB, whereas the median is 0.19 KB. The respective values for image resources are 6.43 KB and 1.19 KB respectively. Similar observations hold for other types of URL resources and other logs. From Table 4 we can also notice that different crawlers exhibit widely different average response sizes. This is attributed to the fact that, in contrast to Inktomi and AltaVista, Google and CiteSeer download postscript, pdf, and image resources, which have larger average sizes.

In our study, a significant portion of the HTTP traffic corresponds to messages carrying no content and having a very small size. For instance, over 40% of Google messages have `3xx` and `4xx` codes. Therefore, it is interesting to study the size distribution of successful messages with a `200 OK` code; this will provide insights on the size and type of content *downloaded* by users and crawlers. In Figure 3 we present diagrams with the body and the tail distribution of the sizes of successful responses to Inktomi (similar plots for the remaining crawlers are given in [5]). From these diagrams we observe that high variability is also present in the sizes of successful HTTP responses.

2.4 Temporal Behavior

Distribution of inter-arrival times Earlier studies have shown that general HTTP traffic is bursty and highly variable, and inter-arrival times of HTTP requests are heavy tailed [7, 4]. To investigate the inter-arrival-time distribution of crawler requests, we process our logs to measure and extract all time-intervals between successive HTTP requests issued by a particular crawler. A crawler employs multiple fetchers to crawl a site and different fetchers may reside on different IP addresses. Therefore, we take into consideration requests coming from all IP addresses identified with that crawler and study the statistical characteristics of the union of all corresponding time-intervals.

In Figure 4, we present logarithmic diagrams of the empirical density of inter-arrival times of HTTP requests from Google, AltaVista and Inktomi on CS-UCY. From these diagrams we can see that the time between subsequent HTTP requests is highly non-uniform and heavy-tailed. The observed distributions reflect the presence of multiple underlying distributions representing the behavior of fetcher-processes residing at different IP hosts of the same crawler. This effect is more pronounced for Google and Inktomi, which use a very high number of different IP hosts, than for AltaVista. Furthermore, the inter-arrival-time distribution of requests coming from an individual IP host is the combination of two underlying distributions: the first distribution represents the inter-arrival times of HTTP requests generated by the fetcher-process(es) of this IP host within one "crawling-session." The second represents the times between subsequent crawling-sessions. Shorter inter-arrival times are observed within a crawling-session whereas longer intervals correspond to periods of "silence," or crawler inactivity.

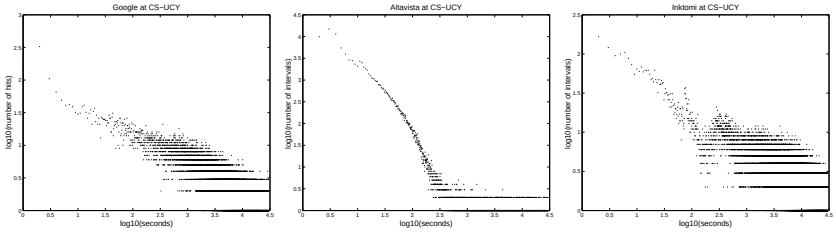

Fig. 4. Distribution of inter-arrival times for Google, AltaVista and Inktomi on CS-UCY.

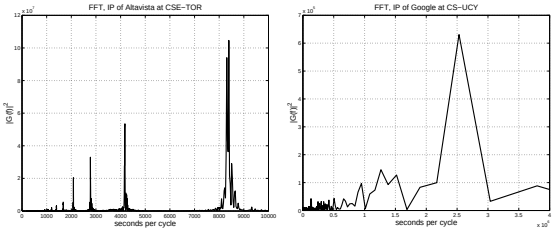

Fig. 5. Power spectral density of an AltaVista and Google IP address hitting CSE-TOR.

Crawler Periodicity An interesting point that arises when investigating a crawler's activity is whether they exhibit a periodic behavior. By plotting the time activity (i.e. the active and inactive periods of time) of a crawler's processes that issue requests to a server, we have observed that several of them seem to exhibit, at least partially, a periodic pattern. We investigated further this observation and verified the periodicity for several IP addresses belonging to crawlers and estimated their time cycles.

For this task, we use the Fast Fourier Transform (FFT). The FFT maps a function in the time field to a, complex in general, function in the frequency field. The idea is that by observing peaks of magnitude in the frequency field we can easily conclude that time activity has periodicity. The frequency coordinate of each possible peak is inversely proportional to the time cycle of the periodicity. Since we are not interested in the phase of the frequency plot, we will illustrate the spectral density function that is the square of the magnitude of the FFT. Before implementing the FFT, we pre-process the requests issued from a certain IP address belonging to a crawler. Time is assumed to be sliced; we use a 10 sec time interval (granularity). Ideally, the granularity should be as small as possible, but we tried to keep the number of resulting points relatively small for a faster FFT computation. We count the requests issued from the IP of interest in each time interval. Because our focus at this stage is on the presence of some periodic action, we assign the value of one to the intervals that have at least one

hit and the value of zero to the ones with zero hits. Consequently, we produce an ON-OFF signal that represents the crawler's time activity for the selected granularity. This signal is passed as input to the FFT function. The resulting diagrams reveal periodic behavior for several crawlers' IP addresses and in some cases this phenomenon is rather intense.

In Figure 5, we present two examples. We plot the power spectral density function with respect to the inverse frequency (time) of the activity of an IP belonging to Altavista and hitting CSE-TOR, and of a Google IP hitting CS-UCY. FFT specifies the main periods observed on that signal. We observe a periodic behavior which corresponds to the peak of around 8400 sec in the left diagram of Figure 5, and a dominating period of about 2.5×10^6 seconds in the right diagram.

3 Conclusions

Our analysis produced a number of insights regarding crawler traffic and crawling characteristics. In particular: (a) Crawler-induced HTTP messages carry *GET* requests at a much higher percentage than the general population of Web clients. Crawlers that implement caching and employ conditional *GET*'s, receive *304* replies at a rate significantly higher than "average" Web clients. Therefore, caching at the crawler-side can reduce significantly the crawler-induced traffic on the World-Wide Web. (b) Crawler requests result to a percentage of HTTP replies carrying error codes (with *4xx* numbers) at a rate higher than what is observed for the general Web-client population. Therefore, there is room for improvement in crawler design, so that crawlers avoid following broken or erroneous links. (c) As expected, a crawler seeks text and HTML resources at a rate much higher than the general population of Web clients. Crawler interest for images is minimal. Crawlers that belong to Search Engines which index non-textual formats (postscript, PDF, etc.), however, fetch this type of resources more aggressively than the general Web-client population. (d) In contrast to observations for "concentration" of HTTP requests, crawler-induced requests are not highly concentrated to a small subset of Web-site resources. Furthermore, crawler visits at a Web site do not result to "Zipf-like" popularity plots. It seems that crawlers classify resources into subsets; each subset is being visited by a crawler with a different frequency. (f) HTTP replies to crawler requests exhibit a high variability in size; the same remark holds for successful responses to crawler requests that carry Web resources back to a crawler. The size of these messages can be modeled as a heavy-tailed, Pareto distribution. Average and median size of crawler-induced HTTP responses are much smaller than those for the general Web-client population. (g) Inter-arrival times of crawler requests are highly variable and exhibit characteristics of heavy-tailed distribution. Periodicity properties can be investigated with Fourier transforms, which help us identify the time-periods of crawler visits upon a Web site.

Acknowledgments The authors wish to thank professors A. Bilas of the University of Toronto, V. Markatos of the University of Crete, M. Skordalakis of the National Technical University of Athens, and the University of Cyprus Computer Center for providing access to their Web server logs. This work was supported in part by the Research Promotion Foundation of Cyprus, under the PENEK 23/2000 project, and by the Planning Bureau of the Republic of Cyprus, through the WebC-MINE grant for Scientific Collaboration between Cyprus and Greece.

References

1. A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, 1(1):2–43, 2001.
2. M. Arlitt and T. Jin. Workload Characterization of the 1998 World Cup Web Site. Technical Report HPL-1999-35R1, Hewlett-Packard Laboratories, September 1999.
3. P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns: Characteristics and caching implications. *World Wide Web (special issue on Characterization and Performance Evaluation)*, 1999.
4. M. Crovella. Performance Characteristics of the World-Wide Web. In C. L. G. Haring and M. Reiser, editors, *Performance Evaluation: Origins and Directions*, pages 219–233. Springer, 1999.
5. M. Dikaiakos, A. Stassopoulou, and L. Papageorgiou. Characterizing Crawler Behavior from Web Server Access Logs. Technical Report TR-2002-4, Department of Computer Science, University of Cyprus, November 2002.
6. A. Feldmann. Characteristics of TCP Connection Arrivals. In K. Park and W. Willinger, editors, *Self-Similar Network Traffic and Performance Evaluation*. John Wiley, 2000.
7. B. Krishnamurthy and J. Rexford. *Web Protocols and Practice*. Addison-Wesley, 2001.
8. G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C. Spyropoulos. Clustering the Users of Large Web Sites into Communities. In *Proceedings of the International Conference on Maching Learning (ICML)*, pages 719–726, 2000.