# Analyzing Web Robots and Their Impact on Caching

Virgílio Almeida[†] Daniel Menascé[‡] Rudolf Riedi [§]
Flávia Peligrinelli[†] Rodrigo Fonseca[†] Wagner Meira Jr.[†]

[†] Dept. of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG 31270 Brazil
{virgilio,flavia,rfonseca,meira}@dcc.ufmg.br

[‡] Dept. of Computer Science
George Mason University
Fairfax, VA 22030 USA
menasce@cs.gmu.edu

[§] Dept. of Electrical and Comp. Engineering
Rice University
Houston TX 77251 USA
riedi@rice.edu

## ABSTRACT

Understanding the nature and characteristics of Web robots is an essential step to analyze their impact on caching. Using a multi-layer hierarchical workload model, this paper presents a characterization of the workload generated by autonomous agents and robots. This characterization focuses on the statistical properties of the arrival process and on the robot behavior graph model. A set of criteria is proposed for identifying robots in real logs. We then identify and characterize robots from real logs applying a multi-layered approach. Using a stack distance_based analytical model for the interaction between robots and Web site caching, we assess the impact of robots' requests on Web caches. Our analyses point out that robots cause a significant increase in the miss ratio of a server-side cache. Robots have a referencing pattern that completely disrupts locality assumptions. These results indicate not only the need for a better understanding of the behavior of robots, but also the need of Web caching policies that treat robots' requests differently than human generated requests.

## 1. INTRODUCTION

Robots play a particularly central role in information economy. Robots automatically search the internet for information, goods and services on behalf of customers. Indeed, directories and search engines are among the most popular sites of the Internet. At the same time, with the dawn of e-business and time-sensitive information, such as news and financial data, came along a steep growth of dynamic documents on the web. Thus, search engines require exhaustive crawling work to maintain and update their indices to the increasingly time-sensitive web content. Currently, publicly indexed documents exceed one billion in numbers [1]. In addition to the general-purpose crawlers, an ever growing number of focused *crawlers* selectively seek out documents that are relevant to a specific pre-defined set of subjects [2]. To obtain information about a product or service requested by a customer (e.g., price, expected delivery time, etc.) might require to query hundreds of sites within seconds. For example, www.shopper.com claims to compare 1,000,000 prices on 100,000 specific products.

The growing popularity of crawlers, shopbots, and other robots on the web, demands for an understanding of their behavior and their impact on the infrastructure of the Internet, in particular on the performance of Web caching. Towards this end, we analyze three different types of logs from actual web sites: an online bookstore, servers for the 1998 FIFA World Cup, and the site of the Computer Science Department at UC Berkeley.

The goals of this paper are twofold. We first aim at identifying, characterizing and eventually distinguishing two major categories of robots, namely "Crawlers" and "ShopBots" solely based on observations at the server. These two classes produce quite distinctively different streams of requests. A typical Crawler will request a site's Home Page, wait for the response, parse it and determine the links present in the page. It then waits for a predetermined amount of time (possibly zero), and sequentially issues requests for each link found, repeating the process for each page received. On the other hand, PriceBots issue requests triggered by human action on a remote site, for example the search for a book by author in a price comparison site. One should expect the arrival process of requests and the popularity of objects requested to differ substantially for the two classes, which is what we set out to show. Second, it is impossible to ignore the impact of web robots on Web caching. Using a stack distance_based model, we analyze the interaction between robots and server-side caches. Based on actual logs we compute cache hit ratios for different types of robots and analyze their impact on the cache behavior.

There are very few studies on Web robots available. Most concentrate on defining architectures and implementations for crawlers and shopbots. In [8], the authors survey the state-of-the-art of Web robots and discuss robot crawling, a technique for building indices for search engines. In [3], the authors examine the problem of identifying navigational patterns of Web robot sessions using standard classification techniques but do not cover features and statistical characterization of robot accesses. Reference [4] searches for invariants in e-business workloads. The authors studied the workload of two actual e-business sites and found the presence of robots in the workload. They also estimate that robot requests correspond to roughly 16% of the total workload of the sites analyzed.

Section two shows our hierarchical approach to characterizing robot workloads and summarizes the main features of the three actual logs used in this study. Section three models in detail the robot requests found in the bookstore logs. Section four assesses the impact of robots' requests on web caches. Finally, section five presents concluding remarks.

## 2. ROBOT WORKLOAD MODELING

This section describes the approach used for characterizing robot workloads. The resulting characterization not only provides information about the behavior of the various classes of robots but also allows for the determination of the resource demands imposed by them. Since not all robots identify themselves to the server, the task of characterizing robot workload is a combination of two highly intervowen sub-tasks: identifying robots for analysis, and an analysis which allows to characterize and thus identify robots. Identification may result from clues as diverse as session length, patterns, and others, as we are about to elaborate on.

Our starting point is the multi-layer hierarchical model proposed by Menascé et. al. [4]. The model proposes a three-layer characterization strategy (see Fig. 1) to facilitate the understanding of how business-level decisions impact the resource level. The business strategy of robots is quite simple: gather information from the e-business site for tasks such as indexing or price/product comparison. The session layer focuses on the session length and on the overall behavior of each robot in terms of functions (e.g., search) invoked per session. The function layer specifies features associated with each function, such as their parameters, inter-arrival times, and the number of embedded files (e.g.,images) requested when a function is invoked. The request layer characterizes the individual HTTP requests in terms of an analysis of their arrival process on multiple time scales. Finally, the resource level characterizes all components of the computational platform of the e-commerce server.

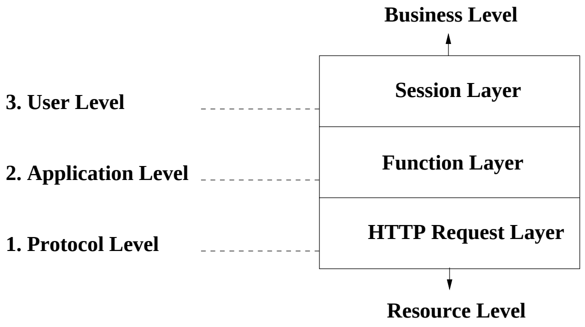

**Figure 1: A hierarchical workload model.**

Robots behave quite differently from humans; they invoke a smaller variety of functions and exhibit different repetition patterns and search strategies. The major simplification is that we do not expect robots to present a buying behavior, that is, they usually do not select products for later acquisition ("add to cart") and, in consequence, never purchase. However, we expect that the evolution of agent technology will allow such tasks to be performed automatically in the future. On the other hand, the process of robot characterization demands not only the determination of visiting patterns and arrival process, but also the nature of the parameters requested by robots. For instance, a crawler visits each object served by the Web site just once, producing an access pattern that is quite different from human users.

Inspired by [3, 4] we introduce now several criteria, based on the aforementioned hierarchical model, which allow us to identify robots in real logs. We group these characteristic criteria according to the three layers of the hierarchical model.

### 2.1 The Robot Criteria

#### *2.1.0.1 Session Layer*

The session-layer criteria for identifying robots are *session length* and *function*. Session length is defined as the number of functions invoked during a session. It is intuitive that thousands of requests within a few hours is highly atypical for a single human user, making session length an obvious criterion for identifying robots. But robots and humans differ also in terms of the type and variability of functions they invoke. For instance, some robots perform only searches and others visit systematically all the pages of a site, whereas humans typically follow some procedure of "narrowing the search and potentially "buying'.

#### *2.1.0.2 Function Layer*

The criteria used at the function layer are the execution of *human-like functions* and the occurrence of *embedded files* in the request stream. The "human-like function criterion' discards sessions in which products are bought or added to a shopping cart as being non-robotic. This criterion might fall short of being accurate in the future as robots start to buy on behalf of human customers. Robots usually do not request embedded files, since their goal is to obtain the textual information resulting from the requests.

#### *2.1.0.3 Request Layer*

At the request layer the criteria are *self-identification* and request *inter-arrival times*. Some robots identify themselves by requesting the `robots.txt` file, which specifies the actions that may be performed by robots. This particular file request is a reliable identifier of robots since the file is hardly of any interest to a human. Unfortunately, not all robots check this file, though highly recommended by *net etiquette*. The request arrival process generated by human users typically exhibits an exponential distribution for inter-arrival times, while robots produce more periodic arrivals and, thus, quasi constant inter-arrival times.

## 2.2 The Logs

The above criteria are applied to three actual logs. Obtaining HTTP logs from actual e-commerce sites can be a challenge since these logs may contain information that is quite revealing about the nature and degree of success of the business. We are grateful to an online bookstore which was kind enough to provide us with a sanitized version of their HTTP logs. Due to a non-disclosure agreement we are unable to name the company or to provide information on sale-related matters. The other two logs used in this paper are publicly available: logs from the UC Berkeley CS department Web server [11] and logs from the servers for the 1998 FIFA World Cup [10].

The characteristics from these logs are presented in Table 1. Notice that the number of functions issued by robots is largest for the bookstore. Because the site is a bookstore we found some ShopBots in the log, justifying the larger number of functions associated with robots. Notably, the logs of the bookstore and the World Cup servers contain more images than the Berkeley log due to the nature of the sites. Bookstores hope to improve sales by displaying images of book covers. The World Cup site used pictures of soccer players and games to attract visitors to the site. The Berkeley site, on the other hand, is an institutional site and does not exhibit a large amount of image files.

Let us now discuss the application of the robot identification criteria to the three logs. Table 2 summarizes the results obtained for the bookstore log. Each column header labels a specific criterion and each line corresponds to a single robot. The id numbers are extracted from the logs and are used throughout this section to uniquely identify each robot. For each position in the table, except for column 2, a "•" sign indicates that the robot was identified by that specific criterion.

## 2.3 Applying the Criteria

The criterion "session length" (session layer) simply says that if a session has a length (in number of requests to functions) larger than a threshold, it is a robot with high probability. Figure 2 plots the probability that a given session has more than $x$ requests. Notice a sharp change in behavior at about 500 function-requests after which a heavy tail appears. We conjecture that this change is due to the increased fraction of robot sessions as the session length increases. Consequently, we used this value as the "session length threshold": sessions longer than 500 requests to functions are considered to be originated by robots.

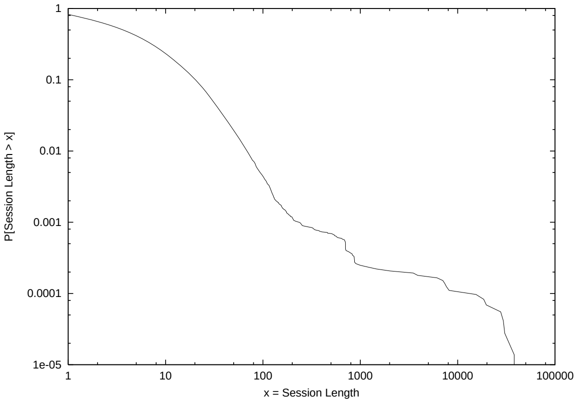

**Figure 2: Tail of the distribution of session lengths (1 - CDF) for all sessions in the bookstore log.**

This criterion is very effective, and all the sessions we classified finally as robotic satisfy it. False positives, on the other hand, mainly mistake proxy servers for robots.

The criterion "function" (session layer) regards the frequency of requests for each of the store's functions. The letters "C" and "S" in the table indicate a behavior that would be expected for Crawlers (C) or for ShopBots (S). Some robots, previously identified as crawlers by manual inspection of their sessions had at least 65% of their functions corresponding to "browse" or "info" requests, while manually identified ShopBots exhibit 95% of "search" requests. We used these numbers as criteria for automatically classifying other sessions. This procedure requires a semantic analysis of the frequency distribution of functions, and it cannot be used alone to determine the *type* of robot because these numbers may not always be valid, even though they showed to be effective in determining whether or not a session *is* robotic.

The "human-likely function" criterion (function layer) says that a session is *not* associated with a robot if a function from a well-defined pool of typically human functions is present. For instance, in the case of the bookstore, a session with any request to the function "pay" is considered non-robotic. It is important to note that this is a negative criterion: it does not imply that a given session exhibits human like behavior, but rather that it contains non-robotic functions. This criterion, unfortunately, did not classify any of the sessions in Table 2 as a robot.

The "embedded files" criterion says that sessions with no or very few requests to embedded files are considered to be initiated by robots. This criterion may fail in the presence of some client or proxy cache configurations, because cached images might not be requested. Nevertheless, even cached images generally result in "if modified since" requests, which appear in the log.

| Source | Bookstore | Berkeley | WorldCup Site |
|---|---|---|---|
| Interval | 01-15 Aug 1999 | 01-30 June 2000 | 23 May 1998 |
| Number of requests | 3,630,964 | 3,643,208 | 2,225,475 |
| Percent of images | 74% | 44% | 84% |
| Number of functions | 955,818 | 2,038,249 | 340,719 |
| % of robot's functions | 33.51% | 16.53% | 6.46% |
| Number of sessions | 130,314 | 371,242 | 33,995 |
| Avg. robot's session length | 2,409.60 | 1,324.93 | 1,398.16 |

**Table 1: Characteristics of the Log Files**

| Robot Id | Session Length | Function | Human-Likely Function | Embedded Files | Self Identification | IAT Distribution |
|---|---|---|---|---|---|---|
| 2 | ● | S | | ● | | |
| 6 | ● | S | | ● | | |
| 8 | ● | S | | ● | | |
| 25 | ● | S | | ● | | |
| 104 | ● | S | | ● | | |
| 3784 | ● | C | | ● | | ● |
| 0 | ● | C | | | | ● |
| 45282 | ● | C | | ● | ● | ● |
| 584 | ● | C | | ● | ● | ● |
| 47277 | ● | C | | ● | | ● |

**Table 2: Criteria used for the identification of the ten most important robots in the bookstore log**

The "self identification" criterion (request layer) assumes that only robots request the `robots.txt` file before accessing a site, since it describes the policies for robot access to the various site resources, such as which can be indexed and/or fetched.

The last criterion concerns the inter-arrival time distribution. Poisson processes generate exponentially distributed inter-arrival times (IAT). The request generation process for some automated robots is almost periodic with quasi-deterministic IAT. In fact, as we shall see, the IAT distribution for some robots was found to be very close to a log-normal distribution which concentrates around one value. This criterion is positive if the IAT distribution is not exponential, possibly indicating a request generation process not driven by humans.

The result of applying the above criteria to the sessions present in the log showed that no single criterion is always effective but a consensus among two or more of the presented criteria generally suffices for identifying a session as generated by a robot.

## 3. ROBOT WORKLOAD CHARACTERIZATION

In this section we identify and characterize robots from real logs applying our multi-layered approach (see section 2). Thereby, we concentrate on the bookstore log because of its significant robot workload, but also because we identified a more diversified mix of robots than on the other logs.

The bookstore log shows requests that are not directly generated by browsers of human users. Through our analysis, we identified search engines' crawler agents, which are part of a broader class of agents that perform resource discovery and retrieval functions. In this class we can also find email address collectors, off-line browsers, site maintenance agents which probe the site at regular intervals to check whether it is alive, as well as database dumpers which, in the case of a bookstore, perform extensive ISBN searches for price comparison or retrieve information on books.

These agents are the actual robots, as they are entirely automated and have a request generation process that is not human-driven, but rather the consequence of a computer program. In this study we collectively call robots of this class *Crawlers*.

Another class of robots is that of agents associated with meta search engines and price comparison sites. We call them *ShopBots*. ShopBots are employed, e.g., by sites which search for prices of items in several e-tailers and present the findings summarized in a single page to the user. What is seen from the perspective of the bookstore is a long stream of requests coming from a single IP address, which can account for a significant portion of the store's load.

Proxy servers are a third source of requests that do not come directly from user browsers, but rather appear to the store as a long sequence of requests coming from a single IP address. By maintaining state information on the user, however, the store can identify the different user sessions present in the stream. We did not consider this type of request generating process, as it is essentially analogous to the direct human interaction.

## 3.1 Request Layer Characterization

We first analyze the distributions of the inter-arrival times (IAT) for each individual robot, and then show our results for the arrival rate sampled at multiple time scales.

### 3.1.1 *Distribution of Inter-arrival Times*

As mentioned, requests triggered by human interaction have a high potential to possess characteristics of a Poisson process, i.e., exponentially distributed IATs. In fact, over 1/2- to 1-hour periods, session arrivals on Internet links have been shown to be consistent with a homogeneous Poisson process (e.g., see [6] for FTP and TELNET sessions, and [7] for Web sessions). Systematic crawlers, on the other hand, should be expected to show more regular, periodic behavior, i.e., IAT distributions which cluster heavily around an average value.

Figure 3 displays the probability distribution function for the interarrival times (IAT) for Crawlers and ShopBots. The distribution for Crawlers exhibits a well defined peak, which varies from one robot to the other.

Crawlers, dumping a database or following links systematically and without human interaction will show a fairly periodic pacing of requests, i.e., a strong clustering of the IAT around their mean. Variability might be caused by network transfer delays—known to be sharply peaked log-normal [9]—and servicing delays. Indeed, we were able to successfully fit, by minimizing the sum of the squared residuals, sharply peaked log-normal distributions for the IAT of crawlers. The probability density function (pdf) for the log-normal distribution is given by $p(x) = 1/Sx\sqrt{2\pi}e^{-(\ln x - M)^2/(2S^2)}$, where $M$ and $S$ are the mean and standard deviation of $ln(x)$, respectively. The parameters of the distribution, as well as the mean and variance thereof can be seen in Table 3.

Shopbots, on the other hand, show IAT distributions that are fairly well approximated by exponential distributions with pdf given by $p(x) = \lambda e^{-\lambda x}$ for $x > 0$, as can be seen on the plot at the bottom of Fig. 3. A slight mismatch can again be explained by accounting for the convolution with transfer delay (for an exponential, a convolution with a sharply peaked log-normal distribution has little effect). Again, the parameters for the fitted distributions for the Shopbots can be seen in the left half portion of Table 3.

### 3.1.2 *Multi-scale Time Analysis*

In this section we study the arrival rate of robot requests on different time scales, i.e., sampled in time intervals of varying length. Figures 4 and 5 show the results for sampling intervals of 1 minute, 30 minutes, and 4 hours. A visual inspection of the plots reveals completely different behavior for Crawlers and Shopbots, at least on intermediate to large time scales.

Crawlers turn on sporadically, send requests at a high steady rate through the duration of their activity and produce clearly visible bursts of activity, not necessarily following any daily patterns. Human-triggered agents such as the Shopbots produce request arrivals that show characteristics similar to the overall requests (all clients). Here, trends emerge which show in terms of long sequences of increase or decrease of volume, particularly pointed at intermediate time scales of the order of minutes. This reflects the overall pattern of the human work-cycle.

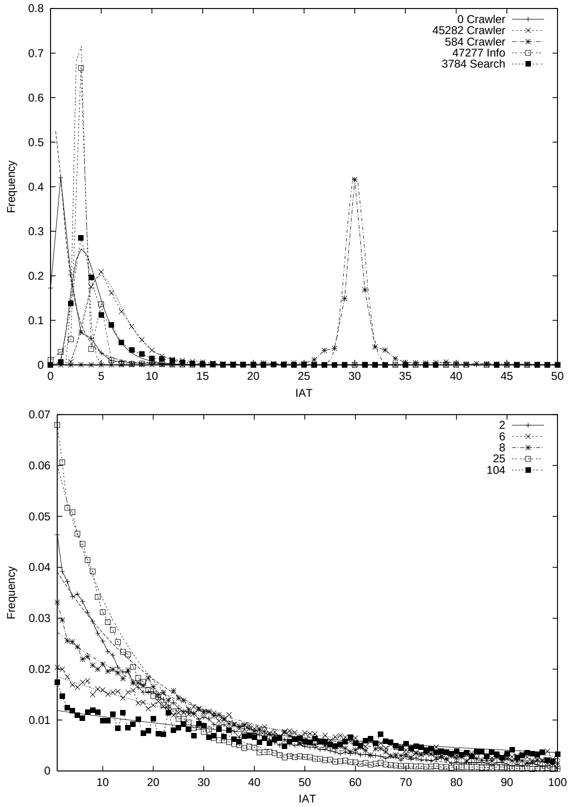

**Figure 3: Probability density functions of the IAT for Crawlers (top) and ShopBots (bottom). Plotted next to each curve is the fitted log-normal or exponential theoretical curve, which show underlying machine and human triggered request generation processes, respectively.**

It is also interesting to indicate the effectiveness of the multi-scale analysis to point out short bursts of activity, seen only on finer times scales. For example, in the case of Crawlers the arrival rate averaged over periods of 1 minute peaks at close to 2 requests per second (Fig. 4). These peaks are averaged out when larger sampling intervals are considered since they are of short duration. This can be invaluable when planning to handle bursts.

## 3.2 Function Layer Characterization

In this section we consider, at the function level, the pattern of object references of different robots. Here the term "object" is employed in a broad sense, representing the information targeted by a request, such as the query string for a search request, and the name of the category for a browsing request. For example, given a Search request, its parameter is the query string, and given a Book Info re-

| ShopBots | | | | Crawlers | | | | |
|---|---|---|---|---|---|---|---|---|
| ID | $\lambda$ | mean | $\sigma$ | ID | M | S | mean | $\sigma$ |
| 2 | 0.0407 | 24.57 | 24.57 | 3784 | 0.448 | 1.3280 | 3.78 | 8.31 |
| 6 | 0.0188 | 53.19 | 53.19 | 0 | 0.230 | 0.9120 | 1.90 | 2.17 |
| 8 | 0.0280 | 35.71 | 35.71 | 45282 | 1.733 | 0.3750 | 6.06 | 2.35 |
| 25 | 0.0639 | 15.64 | 15.64 | 584 | 3.403 | 0.0031 | 30.05 | 0.09 |
| 104 | 0.0120 | 83.33 | 83.33 | 47277 | 1.057 | 0.1764 | 2.92 | 0.51 |

**Table 3: Parameters for the fitted distribution of IAT's**

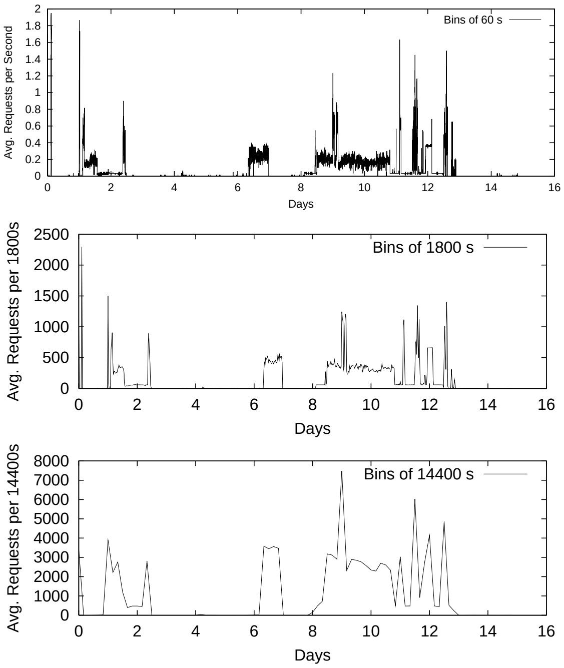

**Figure 4: Arrival process for the Crawlers combined, considering bins of 1 minute, 30 minutes, and 4 hours, showing well defined, short, and irregular bursts of requests**

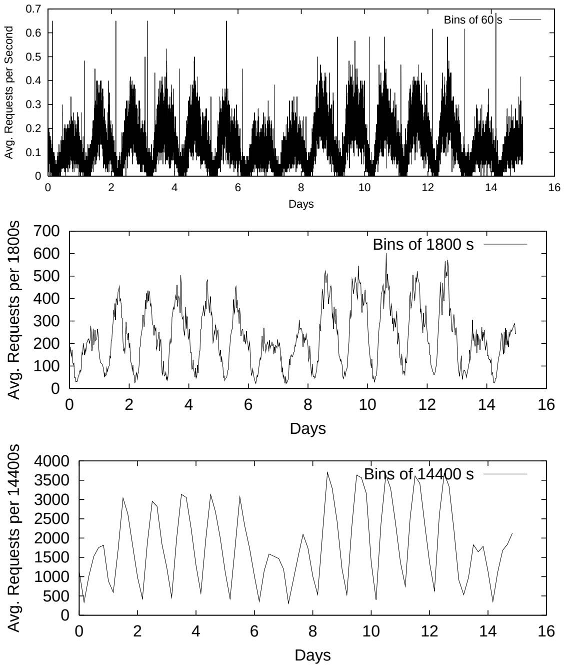

**Figure 5: Arrival process for the ShopBots combined, considering bins of 1 minute, 30 minutes, and 4 hours, showing a well defined pattern that follows the days of the week at the coarser scale, a very human pattern**

quest, its parameter is the book itself. Intuitively, crawlers should request several items approximately the same number of times. Shopbots, on the other hand, should follow an object popularity similar to that of the human users that trigger the activation of the Shopbots, as discussed in more detail in what follows.

Consider a crawler that indexes a site or downloads the entire book database of an online bookstore. During its session, the crawler would typically visit each page once (or a constant number of times). Let the popularity of an object be defined as the number of times it is requested in a session, divided by the total number of requests in the session. For a crawler session, the popularity of all objects should be very similar.

The popularity of objects referenced by human beings, on the other hand, is far from constant. In fact, it has been shown in several domains that the distribution of the popularity among objects of a referenced set is highly concentrated. If we rank the objects by the number of references, i.e., the most referenced objects first, Zipf's law states that the popularity is inversely proportional to the rank. Thus, in a doubly logarithmic plot of the popularity versus the rank Zipf's law appears as a straight line with negative slope given by a parameter $\alpha$. The larger the absolute value of the slope, the more skewed the distribution is. This behavior should also determine the reference pattern of ShopBots, since they re-issue search requests that originate from humans.

Figure 6 plots the rank versus the popularity for objects referenced by a typical Crawler and by a typical ShopBot, in a log-log scale, for the robots with identifications 0 and 25, respectively. We can see that they differ considerably. In the top plot, the Crawler exhibits very large quasi-horizontal regions, indicating near uniform referencing patterns for large groups of objects. In fact, for one of the crawlers we found that more than 90% of the objects were referenced only once. In the bottom plot, the ShopBot conforms very closely to a Zipf-like distribution. By using a linear least-squares fit we found a good fit for $\alpha = 0.49$. The popularity distribution for the objects referenced in the sessions which were classified as human-generated sessions exhibits a similar behavior, i.e., a Zipf-like distribution with $\alpha = 0.62$, indicating a more skewed distribution.

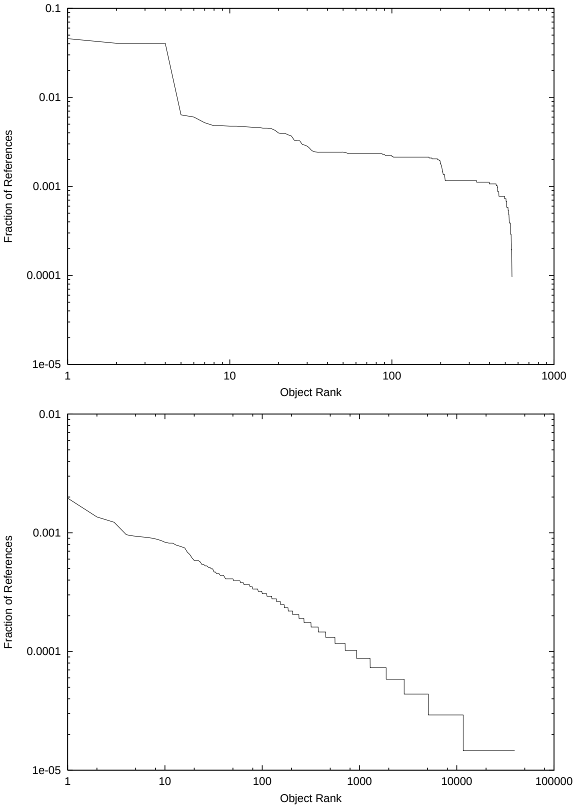

**Figure 6: Rank vs. popularity graph for a typical Crawler (top) and a typical ShopBot (bottom).**

## 3.3 Session Layer Characterization

In this layer,we target the session lengths of the robots and the semantics derived from the access pattern to functions as represented by the Customer Behavior Model Graph (CBMG) [5]. This is a state transition graph used to describe the behavior of groups of customers who exhibit similar navigational patterns. The graph has one node for each possible state (e.g., home page, browse, search, select, add, and pay) and transitions between these states. A probability is assigned to each transition. Different types of users may be characterized by different CBMGs in terms of the transition probabilities. From the CBMG one can derive the session length and the average number of visits per session to each state, and the semantics of the access pattern.

We represent the CBMGs as square matrices, where an entry $i, j$ represents the percent probability of going to state $j$ from state $i$. We removed the states that had no visits.

| | Entry | Exit | Home | Search |
|---|---|---|---|---|
| Entry | 0 | 0 | 0.118 | 99.882 |
| Home | 0 | 0 | 0 | 100 |
| Search | 0 | 0.003 | 0 | 99.997 |

**Table 4: ShopBot CBMG.**

| Function | Function Distribution(%) | # of Visits |
|---|---|---|
| home | 0.12 | 36 |
| other | 0.05 | 14 |
| search | 99.83 | 30391 |

**Table 5: ShopBots Function Distribution and Number of Visits**

Table 4 shows the averaged CBMG for the ShopBots. It shows a simple structure: the first state is almost always "search', and the robot leaves the "search" state with a very low probability. This is confirmed by Table 5, which shows the distribution of visits to each state per session, derived from the CBMG. The behavior indicates that the ShopBots perform (almost) exclusively searches, and that usually they have long sessions.

Table 6 shows the averaged CBMG for the different crawlers. The first observation is the much broader pool of states that are visited. This is only natural, since some of the crawlers tend to visit the whole site. This involves spanning various states, with the exception of the "Humanoid Functions". Most of the visits are to "browse" and "view" (information on specific items). Visits to the search state are rare, and they are usually associated with the initial search page that is reached by the crawler (see Table 7).

The CMBG is very effective in identifying the type of a robot, but it is worth noting that it should be used in conjunction with the other criteria, such as the arrival process characterization to correctly classify the robots. Indeed, we found a Crawler (actually a database dumper) that performed only searches.

| Function | Function Distribution(%) | # of Visits |
|---|---|---|
| view | 38.34 | 6221 |
| browse | 36.53 | 5926 |
| aux | 13.61 | 2208 |
| home | 4.41 | 716 |
| search | 2.56 | 415 |
| acc | 2.24 | 364 |
| add | 2.09 | 339 |
| other | 0.21 | 34 |
| robot | 0.01 | 1 |

**Table 7: Crawler Function Distribution and Visits**

| | Entry | Exit | Home | Browse | Search | View | Add | Acc | Robo | Aux |
|---|---|---|---|---|---|---|---|---|---|---|
| Entry | 0 | 0 | 33.333 | 33.333 | 0 | 33.333 | 0 | 0 | 0 | 0 |
| Home | 0 | 0.0196 | 16.606 | 24.29 | 0.031 | 8.794 | 30.332 | 4.061 | 0.016 | 15.844 |
| Browse | 0 | 0.006 | 8.088 | 68.817 | 0.098 | 2.783 | 0.053 | 0.002 | 0.006 | 20.146 |
| Search | 0 | 0 | 1.426 | 54.028 | 16.125 | 25.716 | 0 | 0 | 0 | 2.705 |
| View | 0 | 0.0020 | 1.240 | 4.182 | 0.783 | 92.601 | 0.124 | 0.015 | 0 | 1.052 |
| Add | 0 | 0 | 6.979 | 0 | 0 | 0 | 11.419 | 81.119 | 0 | 0.483 |
| Acc | 0 | 0 | 2.292 | 0.040 | 0.040 | 0 | 3.023 | 13.485 | 0 | 81.121 |
| Robo | 0 | 0 | 50.000 | 0 | 0 | 0 | 0 | 0 | 0 | 50.000 |
| Aux | 0 | 0 | 1.832 | 26.017 | 32.916 | 13.273 | 0.060 | 0 | 0 | 25.903 |

**Table 6: Crawler CBMG.**

# 4. IMPACT OF ROBOTS ON WEB CACHES: A FIRST CUT

In this section we assess the impact that robots have on Web caching. Our discussion applies mostly to server-side caches or to other types of caches that are part of the server cluster.

We concentrate our analysis on the bookstore log. We start by partitioning the log using the criteria of Sec. 2 into three categories: crawlers, shopbots, and others (i.e., requests generated directly by human beings). This procedure yields three new logs, which we compare with the original complete log, leaving us with a total of four data sets. In order to compare the effectiveness of caching, we assume a server-side cache system which is able to cache results of dynamic requests, specifically search and info requests. From the four logs we derive the corresponding request streams that the cache would be subject to, by isolating the objects referenced in the requests to search and info functions.

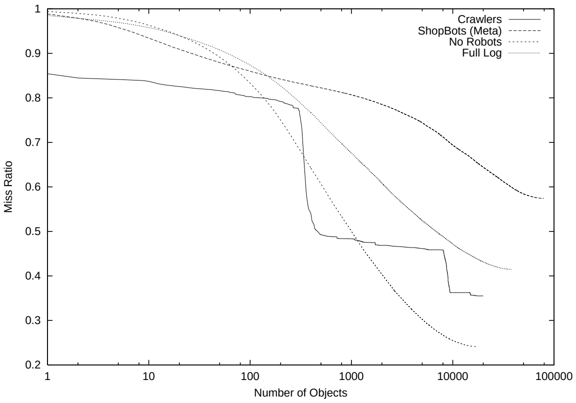

**Figure 7: Miss ratio as a function of cache size (in number of objects)**

Figure 7 plots the expected miss ratio on a cache subject to the four streams discussed above, as a function of cache size, measured as the number of objects it can hold. We consider all objects to be equally sized. This model, albeit simplistic, can give us insight on the behavior of the cache for the different streams. We use the marginal distribution of Least Recently Used (LRU) stack distances to determine the cache miss ratio [12, 14] under the LRU policy. If $D$ is the random variable corresponding to the stack distances and $F_D$ is the cumulative distribution function of $D$, then the miss ratio $m\ (x)$ for a cache of size $x$ is given by $P[D > x] = 1 - F_D(x) = m\ (x)$.

We start by noting the significant difference in the asymptotical miss ratio of the four streams, i.e., if the cache had infinite size. The asymptotical miss ratio is due to the mandatory miss caused by the first reference to each object, and thus relates to the number of different objects in the reference stream. The 'No Robots' curve has the lowest miss ratio, lower than 25%, almost half that of the 'Full' log. It is also the one that decreases the most as we increase the cache size. The 'ShopBots' curve shows an asymptotical miss ratio close to 57%. The log we analyzed is from a specialized bookstore that has books from restricted domains. We conjecture that the human users that access the site know this and thus submit queries that are also restricted to a narrower domain. The ShopBots, on the other hand, submit queries to this and to other bookstores on behalf of users that are not aware of the domain of these backend bookstores. This behavior would tend to increase the number of different objects requested. The 'ShopBots' curve is also much less sensitive to increases in cache size. This is very much in tune with the observation that the inclination of the popularity versus rank curve is much lower for the Shopbots than it is for human users, indicating a less skewed distribution.

The 'Crawlers' curve is very different from the others, with very pronounced steps. It presents regions in which the miss ratio remains almost the same as the cache expands, to fall abruptly after threshold cache sizes. This is a consequence of the non-human request generating process, which can be also seen in Figure 6 that shows the popularity distribution as a function of rank. The crawlers have a 'round-robin' like referencing pattern, requesting each object approximately the same number of times. Some crawlers have a list of items which they request sequentially, while others crawl the site in breadth-first order. In the database literature, it is well known that repeatedly reading a relation sequentially can render an LRU cache useless, unless the cache can hold the entire relation [13]. We interpret each step in the curve as the size of the set of pages requested by each different crawler, but this requires further investigation. Another interesting observation from the 'Crawler' curve is that with a cache of size 1 we can verify a hit ratio of almost 15%, while at the same size the other curves exhibit hit ratios close to 0%. We inspected the logs and found out that this is due to one particular crawler that always requests the same object twice within a very short interval. Even on a cache of size 1 the second references always generate a hit.

The main conclusion drawn from the above is that the presence of robots causes a significant increase in the miss ratio of a server side cache. Crawlers have a referencing pattern that completely disrupts locality assumptions, while ShopBots show a reference stream with less reference locality. Further, the arrival process of both Crawlers and ShopBots poses a significant additional load on the servers, which have to handle the costs associated with both a higher miss ratio and the requests submitted by robots. This suggests that robots should be treated differently than humans by the cache and by the server.

## 5. CONCLUSIONS

Very few studies have been published regarding the behavior of robots on the Web and we are not aware of any studies that focus on the impact of robots on cache performance. This paper used a hierarchical approach for workload characterization of requests generated by robots. Using several criteria, the paper shows the presence of different types of robots in logs from actual web sites. The characterization was done at the session, function, and request levels. Statistical analyses of the robot request arrival process was carried out at different time scales. Using information derived from the log of a real online bookstore, the paper then discusses the impact of robots on the performance of Web caches. From the reference locality perspective, the presence of robots causes a significant increase in the miss ratio of a server side cache, resulting in higher costs for request responses, which affect already overloaded servers by the robot requests themselves. These results suggest the need for strategies of service differentiation between robots and human users, not only guaranteeing the response time of the later, but also the efficacy of cache mechanisms employed by the server.

## 6. REFERENCES

[1] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles and M. Gori, "Focused Crawling Using Context Graphs," *Proc. 26th VLDB Conf.*, Egypt 2000.

[2] S. Chakrabarti, M. Berg, and B. Dom, "Focused Crawling: a new approach to topic-specific web resource discovery," Proc. 8th Int. World-Wide Web Conf., Canada, 1999.

[3] P. Tan and V. Kumar, "Modeling of Web Robot Navigational Patterns," Proc. ACM WebKDD Workshop, 2000.

[4] D. A. Menascé, V. Almeida, R. Fonseca, R. Riedi, F. Ribeiro, and W. Meira Jr., "In Search of Invariants for E-Business Workloads," *Proc. 2000 ACM Conf. in Electronic Commerce*, Minneapolis, 2000.

[5] V. A. F. Almeida, D. A. Menascé, R. Fonseca, and M. Mendes, "Business-oriented Resource Management Policies for E-commerce Servers," *Performance Evaluation: An International Journal*, Vol. 42 (2000), pp. 223–239.

[6] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking* **3**, pp. 226–244, 1995.

[7] A. Feldmann, A. C. Gilbert. W. Willinger and T. G. Kurtz, "The changing nature of network traffic: Scaling phenomena," *Computer Communication Review* **28**, No. 2, pp. 5–29, April 1998.

[8] O. Heinomen, K. Hatonen, and M. Klemettinen, "WWW Robots and search engines," Seminar on Mobile Code, Dept. of Computer Science, Helsinki University of Technology, Finland, 1996.

[9] W. Matthews and L. Cottrell, "Internet End-to-End Performance Monitoring for the High-Energy Nuclear and Particle Physics Community," *Passive and Active Measurement Workshop (PAM 2000)*, Hamilton, 2000.

[10] Soccer World Cup Server log http://ita.ee.lbl.gov/html/traces.html.

[11] Berkeley CS Department Web Server log http://www.cs.berkeley.edu/logs/http.

[12] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing reference locality in the www. In *Proc. 4th Int. Conf. on Parallel and Distributed Information Systems*, pages 92–103, December 1996.

[13] B. T. Jonsson, M. J. Franklin, and D. Srivastava. Interaction of query evaluation and buffer management for information retrieval. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 118–129, June 1998.

[14] T. Kelly and D. Reeves. Optimal Web cache sizing: scalable methods for exact solutions. *Computer Communications*, Vol. 24, pages 163–173, February 2001.