Analyzing Robot Behavior in E-Business Sites

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

1. INTRODUCTION

The population of robots, i.e., crawlers, shopbots, pricebots, and autonomous software agents interacting with Web servers and e-business sites, will increase significantly in the future [2, 3, 5]. 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 News sites 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 [2].

In addition to the general-purpose crawlers, an ever growing number of focused *crawlers* selectively seek out information relevant to a specific pre-defined set of subjects [3]. These are part of a broader class of agents that perform resource discovery and retrieval functions, such as address collectors, off-line browsers, site maintenance agents that 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 generally 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*. As a consequence, crawlers are placing an ever growing demand on site resources and on the whole internet infrastructure.

Another class of robots is that of agents associated with meta-search engines and price comparison sites. They automatically search the internet for goods and/or services on behalf of customers, and play a particularly central role in E-business and information economy in general. To acquire 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 role of agents will naturally evolve from information providers into decision-makers. Currently, it is common to see two types of robots being used by e-business sites. Economically-motivated agents called *pricebots* are used to set prices with the goal of maximizing profits for companies. As the degree of autonomy and responsibility of agents increase over time, it is expected that transactions among economic software agents will be a significant part of the information economy. We collectively call them *ShopBots*. ShopBots are employed, for instance, by sites which search for prices of items in several e-tailers and present the findings summarized in a single page to the user.

The goals of this paper are twofold. First, we aim at identifying, characterizing and eventually distinguishing two major categories of robots, namely "Crawlers" and "ShopBots" solely based on observations at the server. Workload characterization may be accomplished at many levels: user level, application level, protocol level, and network level [4]. We looked at the robot workload at three levels, represented by request layer (HTTP protocol level), function layer (application level) and session layer (user level). This hierarchy may be used to capture changes in the behavior of robots and map the effects of these changes to the lower layers of the model, which can be used to provide input information to performance models.

The two classes of robots produce quite distinctively different stream 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, ShopBots 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 E-business sites and information provider sites. Robots consume computational resources at the site as well as valuable bandwidth. We develop analytic models for the interaction between robots and e-business sites [1]. Based on actual logs, we derive performance models of a typical online bookstore through which we assess the impact of robot activities in several what-if scenarios.

2. ANALYZING ROBOTS

The growing use 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 and the performance of servers. 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. We identify and characterize robots from the real logs applying our multi-layered approach [4]. Thereby, we concentrate on the bookstore log not only 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 which are not directly generated by browsers of human users. Through our analysis we identified both ShopBots and Crawlers.

As an example of how the behavior of the two types of robots varies, we analyzed 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, and is fit by a lognormal distribution. 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 IAT around their mean. Variability might be caused by network transfer delay – which is known to be sharply peaked log-normal [6] – and servicing delay. ShopBots, on the other hand, exhibit a distribution that is closely fit by an exponential distribution, suggesting the human driven request generating process. We also found differences at the function and session level, which are presented in an extended version of this paper [7].

Robots use E-business site resources such as CPU time, storage subsystems, and networks. We characterize the service demands placed by robots on processors and disks with the goal in mind to determine the portion of *utilization* of these resources that can be attributed to ShopBots and Crawlers. To this end, we construct a model simulating an online bookstore. We studied the utilization of the bottleneck device—the disk at the database server—attributed to ShopBots and Crawlers, at different time scales. We noticed that the bursts of activity by these robots, though present at all time scales, decrease as the time scale becomes coarser. For example, for a time scale of 14,400 sec., i.e., for data averaged in bins of 14,400 seconds, the utilization peaks range from 15 to 28%, while the highest peaks at time scale 1,800 sec lie between 40 and 65% and for 60 sec between 75% and 124%. This indicates that the site configuration specified in the model would be unable to accommodate such a high arrival rate of requests from Crawlers and ShopBots. For instance, in the 60-sec time scale, the requests generated by the Crawlers resulted in utilization peaks from 95 to 120%, being the main cause of the observed overload.

3. CONCLUSIONS

Very few studies have been published regarding the behavior of robots in the Web. We used a hierarchical approach for workload characterization of requests generated by robots. Using several criteria, we were able to show 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 analysis of the robot request arrival process was carried out at different time scales. Using information derived from the log of a real online bookstore, we also developed a performance model of an online bookstore servicing the robot-generated workload. The performance model provides service demands that, in conjunction with laws from operational analysis, are used to assess the impact of robot workloads on the consumption of the site resources.

In summary, we verified that i) robots can consume a significant amount of system resources, ii)Crawlers consume more resources than ShopBots, and iii) utilization peaks increase in intensity as we consider finer time scales. This indicates that averaging the effect of robots over large time periods does not accurately reflect their behavior in short time scales. At these fine time scales, robots can steal precious resources from other, in most cases, more important requests. Also, as the size of the database increases, we see more instances in which the site would need significantly more capacity just to cope with requests originating from robots.

4. REFERENCES

[1] D. A. Menascé and V. A. F. Almeida, *Capacity Planning for Web Performance: metrics, models and methods*, Prentice Hall, Upper Saddle River, NJ, 1998.

[2] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles and M. Gori, “Focused Crawling Using Context Graphs,” *Proc. 26th VLDB Conference*, Egypt 2000.

[3] S. Chakrabarti, M. Berg, and B. Dom, “Focused Crawling: a new approach to topic-specific web resource discovery,” Proc. 8th WWW Conference, Canada, 1999.

[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 Conference in Electronic Commerce*, Minneapolis, 2000.

[5] F. Cheong, Internet Agents: spiders, wanderers, brokers, and bots, New Riders, 1995.

[6] 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.

[7] D. A. Menascé, V. Almeida, R. Fonseca, R. Riedi, F. Ribeiro, and W. Meira Jr., “Analyzing Robot Behavior in E-Business Sites” e-SPEED Technical Report 003/2001, February, 2001.