



## ***Virtual evidence: analyze the footsteps of your users***

**Win Shih**

Denison Memorial Library, University of Colorado at Denver and Health Sciences Center, USA

### **Introduction**

Inside the networked academy of today, the library's Website is the portal to ever-burgeoning electronic resources, support and services. Understanding the nature and characteristics of how your site is utilized becomes crucial for ensuring and improving the quality of services your prized patrons receive upon virtual demand. Monitoring Website usage and server workload not only assures 100% uptime performance, but concomitantly assists in improving Website design, systems performance, and resource utilization, in conjunction with future hardware and infrastructure capacity planning.

The intensifying imperative of mega-powerful, Web-driven search engines such as Google and Yahoo, combined with specialized "Internet Search Assistants," has contributed to an exponential increase in the population, as well as variety, of software agents, commonly known as Internet robots or simply "bots." These agents are software programs that traverse the Internet in an automated, methodical manner, gathering or "harvesting" the content of each Website they visit, including those of academic libraries. Bots tend to consume a considerable amount of system resources and network bandwidth from the sites they frequent, at the expense of regular site users and usage. Additionally, they generate concerns regarding unauthorized content collection, privacy, not to mention a host of other security-centric matters. A few cases in hand, American Airlines and eBay won cases against companies using robots to mine the pricing information of their sites (Compart 2003, Freeman 2002, Graham 2000).

This paper reports a study on identifying autonomous software agents and their impact on a library's Website based on Web access logs. Furthermore, it characterizes behavior of the major software agents.

### **Guess Who Is (under)Mining Your Website?**

Consider these scenarios –

- The public service personnel at your library begin receiving calls and complaints from patrons that they are being denied access to important resources at your Website. Upon further testing, you discover to your own dismay that you are likewise receiving an "access denied" message when you try to use the resources in question.
- You receive an E-mail from a vendor informing you that your institution's access to its resources has been revoked due to excessive downloading activities from a specific campus IP address. The vendor further requests that you conduct an investigation immediately and that you provide a

detailed explanation of what happened, before they restore access to their resources at your institution.

- Your staff members report that they are not able to log onto your integrated library system in order to process materials. The message they receive indicates that the system has reached “maximum user licenses allowed.” After further examination of access logs, you discover unusually high accessing activities from a specific range of off-campus IP addresses.
- You receive an E-mailed alert from a vendor indicating they’ve identified an open proxy server on your campus. This server will relay requests from off-campus machines without proper authentication (hence referred to as an “open proxy”) to access the vendor’s resources. Such unauthorized access to licensed and copy-righted contents, as stipulated in the vendor’s alert, represents a violation of contract and thus your vendor has suspended printing and downloading of privileges to the “offending server” until the problem is rectified.
- Your collection development librarian receives a warning from one of your major database vendors, indicating that your library’s Website has posted a training user name and password to access their databases. Since this page is open to the general public for viewing and hence violates licensing agreements, your vendor has subsequently deactivated the training account.

These incidents, though ostensibly fictitious, are unfortunately all-too-real and have been reported by a number of University libraries. Service disruption and inconvenience to library users, the costs incurred in resolving these problems, and the damage caused to your library and your institution’s image overall can be potentially enormous. No matter how severe these problems might be, they can be often attributed to Web robots. When applied benevolently, Web crawlers can be invaluable tools for indexing and organizing Internet content. It’s nearly a given that Web search engines have become an indispensable desktop reference aid in our daily “cyber lives.” However, without proactively defending or protecting your server, Web crawlers can be anything but benevolent applets – sometimes in fact they can be so incursive and disruptive that they overload your system, reducing throughput to a true “crawl,” all at the expense of legitimate users and right-minded resource utilization.

### **Crawlers, Robots, Spiders and Spambots**

Like all software agents or “bots”, Web crawlers are “soft robots” which will do precisely what they are programmed to do: Collecting whatever is accessible from your Web server, be it public, private, password-protected, even “hidden” information – everything is potentially ripe fodder to a “bot.” If sections of your Website contain confidential or sensitive data but are not adequately password-protected, Web crawlers will make your data publicly available to others. This is known as “Information Leakage.”

Web crawlers may also be deployed by spammers, known as “spambots,” to harvest e-mail addresses, online resumes, business intelligence and product prices. “EmailSiphon” and “Cherry Picker” are just two examples of spambots, searching for and collecting e-mail addresses as targets for spamming. “YahooSeeker” is another type of search agent, sometimes called a “shopbot” or a “pricebot,” that collects pricing and product information from e-business sites for the Yahoo Shopping service (<http://shopping.yahoo.com/>).

The behavior of any given Web crawler will impact the performance of your Website for good or ill. The typical Web crawler is designed to “mine” a Website as quickly as possible and may use multiple connections to read data from your Web server. A crawler’s actions can resemble a “blitzkrieg” of online Ariadnes swarming about your Website for an extended time (known as “rapid fire”). This intensive activity from multiple crawlers not only potentially overloads the Web server, consuming vast quantities of user licenses, but also can flood the network with spurious information requests, possibly clogging your network connection dangerously, bringing bandwidth to a standstill. As a result, your regular users experience a slowdown in response time or may even be denied accessing vital resources altogether. Finally, Web crawlers can skew usage statistics of your library’s electronic resources, which libraries depend upon for making crucial collection-development decisions. One library system administrator reported an incident involving “Googlebot,” which tied up at least 20 of their integrated library system’s logins for four days while simultaneously indexing their integrated library system<sup>1</sup>.

There are techniques and standards which may either reduce the impact of or prevent crawlers from “tromping through” certain pages on your server. The “Robot Exclusion Standard” allows a site administrator to put a text file (“robots.txt”) on the Web server instructing robots not to index certain parts of your site<sup>2</sup>. This helps; however, not every Website administrator actively implements such measures and not every Web robot abides by “the rules.” Another prophylactic measure therefore is to block known robots like known viruses or known Trojan horses using their associated IP addresses. Yet here again, it is problematic to detect every known bot<sup>3</sup> and passively block it. Worse, some spambots disguise themselves as benign programs to deflect detection altogether.

### **Spider Spotting**

Acknowledging the potency, however, of the “cyber arachnids” on your Web server is the mark of a very wise Web server administrator. Web servers can be configured to record every visit or request for file(s) from your Website vis-a-vis server log files. They are a rich source for tracking how a Website is used and accessed. A typical log entry captures up to 21 pieces of data, including the date and time of the request,

---

<sup>1</sup> From a message posted on the Innopac List (<http://www.innopacusers.org/list/archives/>), May 11, 2001.

<sup>2</sup> For details, see: <http://www.robotstxt.org/wc/exclusion.html>

<sup>3</sup> As of August 2005, there are 298 “known robots” listed on the Web Robot Website (<http://www.robotstxt.org/wc/active/html/index.html>). Another list, maintained by Mike Shor, Assistant Professor of Economics at Vanderbilt University, includes 625 robots (<http://www2.owen.vanderbilt.edu/mike.shor/diversions/analog/>)

the user's IP address, the URL (universal resource locator) of the file/page requested, the protocol used by the request, and the browser or user agent employed by the requesting computer. A sample log entry is shown below:

```
2005-05-04 02:27:23 66.249.64.52 -127.0.0.1 80 GET /interlibrary/index.html - 200
Googlebot/2.1+(+http://www.google.com/bot.html)
```

Table 1 provides a synopsis of this sample log entry, piece by piece, with descriptions of each component of the entry. Software Web crawlers, often referred to as “spiders,” just like their analogous, human counterparts, leave traces of their “footprints” in the server access log files. The user agent field for instance, which is intentionally bold-faced in the sample log entry, is the area for spotting a Web crawler’s I.D. Other fields in the log entry will reveal additional information about the crawler’s activity, such as when it visited and what page(s) got “harvested.” By regularly analyzing your Web server’s log files, a system administrator is empowered to fathom the extent to which and the impact of what any given Web crawler did to your library’s Website.

Table 1. A description of the Extended Log File Format entries used in this study.

Date	Time	Client IP	User name	Server IP	Server Port	Method	URI Stem	URI Query	HTTP Status	User Agent
2005-05-04	02:27:23	66.249.64.52	-	127.0.0.1	80	GET	/interlibrary/index.html	-	200	<b>Googlebot/2.1+(+http://www.google.com/bot.html)</b>

### Literature Review on Web Access Log Analyses

As we migrate ever more swiftly and thus closer to a working realization of the digital library, not only libraries but their vendor-side partners in digitization are paying greater attention to the Web access log in order to monitor resource usage, as well as to study user behavior and improve Website design. There are several attractions which make Web log analyses increasingly powerful and by extension, tremendously popular. Unobtrusive and automatic logging, for example, renders data collection virtually effortless. A few log analysis software programs, known as Web log analyzers, freely available from download, are easy to use and relatively straightforward to install and set up. Web log analyzers can be configured to generate reports automatically with little or no staff involvement required. Reports can be viewed as Web pages and may provide impressive-looking, professional-appearing, colorful graphs and tables, permitting ease of readability and vaunting seamless distribution of descriptive statistics.

There is a growing contingent of library researchers embracing Web log analyses. Several authors have advocated the use of Web access logs to learn more about patron behavior and characteristics, as well as how library resources are regularly used (Bauer 2000, Breeding 2002, Guenther 2000, 2001, Schuyler 2001). Other scholars provide step-by-step instructions on selecting and implementing log analysis tools and systems (Bertot and McClure 1997, Coombs 2005). Cohen (2003a, 2003b), in her two-part article, proposes a double-tiered model, one for library administrators and another for Website managers, in the process of properly reporting Web usage statistics within the academic library milieu. Mariner (2002) provides guidance

for Webmasters in gleaning clues from error log files, agent information, and top entry pages from Web logs to enhance overall Website usability and design.

A voluminous portion of library literature in this area reports findings based on logs derived from individual library Websites. Various authors detail how they are using Web log analysis to learn more about average patron behavior and resource utilization of their sites. The type of library Website here investigated includes: Health Sciences Libraries (Rozic-Hristovski et al. 2002, Stabin and Owen 1997); Government Publication Libraries (Xue 2004); Ready Reference (Mudrock 2002, Silet 1999); Academic Libraries (Coombs 2005, Ren et al. 2000); the National Science Digital Library at Cornell University (Pan, 2003); and a university project site (Thelwall 2001). The authors of these studies further proceed to address how their research outcomes assisted them in improving or enhancing their Website designs. They then proceed to discuss, in considerable depth and across broad latitudes of cognition and subject interpenetration, issues and limitations for implementing Web access logs and interpreting their results meaningfully, consistently, and carefully.

Take a multi-library study for example. Hightower et al. (1998) endeavored to benchmark Web usage statistics by comparing log files from 14 science and engineering academic libraries. They encountered several challenging and seemingly-incompatible findings, including differences in log file formats, site architectural discrepancies, target audience differences, and Website-design philosophical disparities.

In yet another study on a major scale, Covey (2002) interviewed librarians from 24 Digital Library Federation member libraries on how they assess, use and govern the usability of their own online collections and services. She discovered wide divergence in why and how these libraries conduct their Web log analyses, as well as how the results are interpreted. Further she offers and subsequently provides details on the issues, problems and challenges of such practices as reported by the subjects.

In another innovative experiment, Huntington et al. (2004) change the physical, virtual placement of Web links on a consumer health Website to see if altering this lone factor affects usage statistics derived from the Web access logs. Interestingly or perhaps logically enough, they found that the more prominent or visible the link is on a Web page, the greater the number of visits that page received, at least – according to the access logs.

Another team of authors examined usage statistics of a specific online collection. Research conducted by the Institute for the Future for Stanford University (2002) analyzed access logs of fourteen medical and life sciences journals published by Highwire Press. What they uncovered were distinctive user behavioral patterns: downloading the PDF version of the article is the major goal of the user majority; up to 60% of users are redirected from PubMed and up to 25% of users are referred by academic institution Websites. In studying the usage pattern of chemists at Cornell University, Davis and Solla (2003) and Davis (2004) used the transaction logs of 29 scholarly journals from the American Chemical Society (ACS). Using IP addresses

as surrogates for individual users, they discovered that there is a quadratic relationship<sup>4</sup> between the number and types of referrals, as well as between the number of journals consulted and the number of articles downloaded. These relationships fit the Inverse-Square Law, also known as Lotka's Law<sup>5</sup>. They also suggest that the size of the user population can be estimated by the number of downloads per journal. SciFinder Scholar and PubMed are two top sources referring users to ACS journals.

Another study by Stemper and Jaguszewski (2003) compares the usage statistics supplied by four E-journal publishers and statistics from Web access logs. They found that there is a strong similarity between these two sets of data.

Two studies examine the usage of SinceDirect, a portal for Elsevier journals. Ke et al. (2002) review the usage logs provided by the Taiwan-based ScienceDirect E-journal system and report various usage patterns, including PDF accounting for 30% of all result-format requests, the most commonly downloaded journals and the most popular subject areas. On a smaller scale, Taha (2004), reports that the usage pattern of ScienceDirect at United Arab Emirates University identically matches the research focus and graduate programs of that institution.

Finally, in yet-another collection-usage study, Jones et al. (2000) review the search queries and strategies derived from the usage log of the Computer Science Technical Reports Collection, the largest collection of the New Zealand Digital Library. From the medical field, two papers report the usage pattern based on access log analysis of the Family Practice Handbook online collection at University of Iowa (D'Alessandro et al. 1998, Graber et al. 1998).

A few general characteristics of user behavior come out from these collection studies. First, these studies report that a vast number of the users only visit the sites less than twice during the study period. On the other hand, there are a small number of heavy, repeat users (Davis and Solla 2003, Jones 2000, and Ke 2002). They also notice that users prefer PDF versus HTML for full text viewing and downloading (Davis and Solla 2003, Institute for the Future 2002, Ke 2002, and Taha 2004).

Another group of authors used Web access log analysis as a way to monitor the "readership" of online publications. Marek and Valauskas (2002) conducted a citation analysis based on the server logs of the online journal, *First Monday* (<http://www.firstmonday.dk/>). Based on its three years of log files, the authors were able to identify the most "read" or "classic" articles, which continue to be downloaded heavily long after their original publication. In another Web log analysis, Zhang (1999) found that the electronic journal,

---

<sup>4</sup> A quadratic relationship between two variables involves the power 2 in the equation. The simplest equation for a quadratic is  $y = x^2$ .

<sup>5</sup> Lotka's law describes the frequency of publication by authors in a given field. It states that the number of authors making  $n$  contributions is about  $1 / n^a$  of those making one contribution, where  $a$  is often nearly 2 (see: <http://www.wikipedia.org>).

*Review of Information Science* (<http://www.inf-wiss.uni-konstanz.de/RIS/>), has a worldwide, but fixed, readership. However, the log shows that the publication's online interactive communication forum is less popular. In another study, Nicholas et al. (1999) discuss in detail – various measurement and methodological problems based on their experience studying *The Times* Web server log.

### **Limitations of Utilizing the Web Access Log**

Several authors question just how truthful and reliable Web access logging really can be (Cohen 2003a 2003b, Dowling 2001, Fieber 1999, Haigh and Megarity 1998, Pan 2003, Van der Geest 1999, Yu and Apps 2000). Log files are originally set up to monitor network traffic, system performance, server workloads, as well as to detect unusual activities, so that server administrators may deploy necessary actions. The logs are not intended to serve as a substitute for how visitors use your site (Bauer 2000, Goldberg 1995). In addition, log files only record IP addresses of the user workstation. They do not reveal if an IP address is used by a single person, a group of users, or users within a proxy or firewalled environment. Log file analysis will not generate real user demographics, the purpose of user visits, or why users choose certain resources (Tarr 2001). Other means of investigation, such as the Association of Research Library's "MIMES" for Libraries (<http://www.arl.org/stats/newmeas/mines.html>), a transaction-based online survey that randomly samples online patrons, may provide an entirely different picture of usage statistics (Franklin 2004).

Only a few library authors have reported behavior of Web robots and their impact upon their research. Both Pan (2003) and Xu (2004) acknowledge the existence of Web crawlers in their analyses of Web statistics. Pan's statistics show that 11.6% of total requests of the National Science Digital Library site at Cornell University are from Google's crawler. In Xue's study, visits by Web Crawler amount to 5.9%, still a significant impact, of total visits. Cohen (2003a) further points out that Web crawler visits can cause a library to overcount usage statistics and therefore ought to rightfully be removed from the raw data. However, she also indicates that identifying some of the less known or disguised Web crawlers can be difficult.

Outside library literature, there are more studies with focus on the impact of Web crawlers on Web server performance, network traffic, and Web usage statistics. In an analysis of more than 1.1 billion access requests to CERNET (China Education and Research Network, <http://www.edu.cn>) during a six month period, Ye et al. (2004) found that none of the top five Web crawlers identified in their study paid attention to the workload of the Website itself. Reportage came through instead that Web crawlers paid 15 times more visits during the site's busy hours than those of quiet times. In another study, Dikaiakos et al. (2005, 2003) analyzed Web access logs of five academic institutions' Websites for periods varying from 42 days to 176 days. Their results indicate that the activities of the top five crawlers represented 8.52% of total requests for all five sites combined. However, they only accounted for 0.65% data bytes transferred. In another study, Tan and Kumar (2002) reviewed the traffic of the Web server at the University of Minnesota Computer Science Department during a one-month window. They reported that, on average, Web crawler accounts for only 5% of the total session. However, they also discovered that Web crawlers (of all varieties) may account

for as many as 85% of the total number of HTML pages requested during “busy times.” In still another study comparing the behavior of different types of robots at an E-commerce site, Almeida et al. (2001) reported that robots can consume a significant amount of system resources. However, Web crawlers utilize more resources than shopbots and the larger the site the greater the utilization. In another paper of similar study reported by Menasce et al. (2000), like-minded researchers found that at least 16% of the requests of two E-commerce sites are generated by robots.

### Collection of Data

This study employs the Web access logs recorded on the Web server at University of Colorado Health Sciences Center Library (<http://denison.uchsc.edu/>). The data covers a four-month period, from March 27 through July 26, 2005. Table 2 provides a summary of the characteristics of the raw data. “Analog” (<http://www.analog.cx/>), a popular and free software program which analyzes and summarizes Web access log files, was used to process 122 daily log files.

Table 2. A summary of access log characteristics (the raw data).

Log period	3/27/05 – 7/26/05
Log duration (days)	122
Log size (MB)	762
Total <i>Requests</i> (total number of files downloaded, including graphics)	4,820,709
Total requests for <i>Pages</i>	701,261
Average daily <i>Requests</i>	39,514
Average Web crawler daily <i>Requests</i>	1,346
Average daily requests for <i>Pages</i>	5,748
Average Web crawler daily requests for <i>Pages</i>	1,144

### Web Usage Statistics

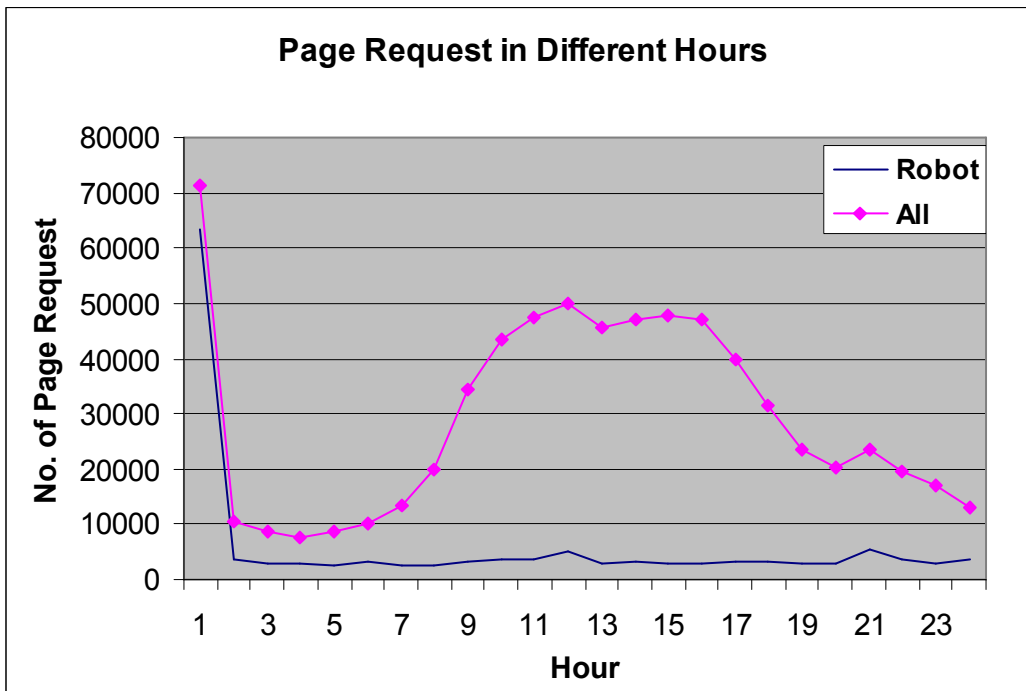
In gauging Web site activities, two commonly-employed indicators or factors are *Request* and *Page*. For example, each file sent to the end user by the Web server is counted as a *Request*, or sometimes a *Hit*. By comparison, a *Page*, sometimes called a *Page View*, represents each time a visitor viewed a Web page of the site. It is important to understand the difference between *Request* count and *Page* count and how they are generated when assessing Web site usage reports. A typical Web page is composed of a collection of the essential HTML code file, as well as images, pictures, and other type of files embedded within the page. If a user requests a Web page that contains an HTML text page with five GIF-formatted image files contained therein, the Web server will open a total of six connections (five for the image files, and one for the HTML file) and record six *Request* entries in the Web access log file. However, the server will count this specific access as only one “*Page request*.” As a result, in reviewing a Web analysis, the number of total *Requests* is always larger than that of total *Page* requests. Relying on the *Request* count runs the risk of artificially

inflating actual usage. In reviewing Website usage statistics, the number of *Page* requests means more than *Total Requests*, unless one is interested only in the number of images files that are requested by users. As Table 2 shows, during the study period, more than 4.8 million *Requests* were recorded, but merely about 0.7 million *Pages* were viewed.

### Crawler Activities

On average, Web crawlers daily paid 1,144 *Page* visits and 1,346 *Requests* during the period of investigation. Figure 1 shows the distribution of *Page* requests during different hours of the day. To our surprise, the line chart shows a spike of heavy usage at midnight. After further examining the log entries, we learned that the University’s Information Technology Services ran an enterprise search engine software package, called “Ultraseek,” early every morning to index all Web servers on campus. The information collected by Ultraseek crawlers is used to construct the University Website’s search engine.

Figure 1. *Page* Requests During Different Hours

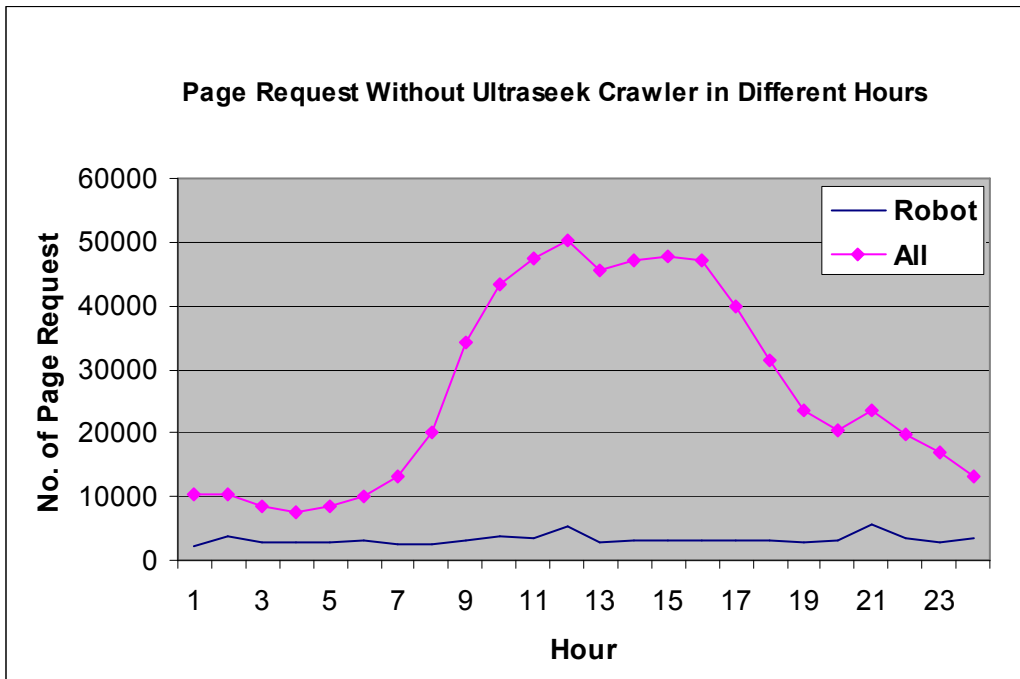


To better understand the actual Web crawler impact outside the context of our unique, localized situation, the entries from the University’s Ultraseek crawlers were removed from our log analysis. Figure 2 shows the revised distribution of *Page* requests by “hours of the day.”

As one can see from the line chart, our Web site traffic starts to increase after 9 a.m. and reaches its peak around noon time. The usage sustains for another three hours before gradually dropping off at 4 p.m. On the other hand, the activities attributed to Web crawlers display a steady and stable flow regardless of the time of day. Our findings indicate that Web crawlers pay little attention to the activity status of our Website,

nor do they pay heed to the workload of our Web server or network traffic. As a result, Web crawler activities during busiest periods for our Website put an additional burden upon an already overloaded system and network. Furthermore, they compete with our regular users for system resources and are likely to impair services overall. Not only can our users not fully access our resources, Web crawlers likewise are unable to harvest their content in the most efficient manner. A similar result was reported by Ye et al. (2004).

Figure 2. Page Requests During Different Hours Excluding Ultraseek Crawler Activities



### Crawler Impact

As Table 3 shows, Web crawlers account for less than 3.5% of all *Requests* received by the Web server during the 122-day period. However, when we focus on the number of Web pages viewed by Web crawlers, the percentage jumps to almost 20% of total Web *Page* requests. This prodigious percentage difference between *Requests* and *Page* requests is due to the very nature of browsing behavior by Web crawlers. Search engines or Website harvesters do not usually index image, PDF, or other non-HTML files, and thus their “crawlers” only target HTML files most of the time.

If we exclude the University’s Web crawler activities from total *Requests*, the activities of the rest of Web crawlers reduces to 2.41% of total *Requests*. *Page* requests by remaining Web crawlers similarly drop to 12.26% of total *Page* requests.

Table 3. Contribution of Web crawlers to Web server activity

	All	Web Crawlers	Percentage
Total <i>Requests</i> (total number of files downloaded, including graphics)	4,820,709	164,175	3.41%
Total <i>Requests</i> (total number of files downloaded, including graphics), excluding Ultraseek crawler activities	4,758,355	101,821	2.41%
Total requests for <i>Pages</i> (HTML pages)	701,261	139,534	19.90%
Total requests for <i>Pages</i> (HTML pages), excluding Ultraseek crawler activities	640,236	101,821	12.26%

### Resource-type

The difference in browsing behavior between Web crawlers and regular Website visitors can be further observed by comparing the type of files requested by each group, as shown in Table 4. More than 90% of the requests by Web crawlers are for text files, which consist of .html, .htm, and directories. These types of files match the definition of a “Page visit.” Whereas text files represent less than 16% of regular human user visits. On the other hand, image files (.gif, .jpg, and .pdf files), which are practicably negligible to a Web crawler, account for more than 65% of regular user requests. Our findings match those reported by Dikaiakos et al. (2005, 2003) and Tan and Kumar (2002).

Table 4. Comparison of file type requests between Web crawlers and regular users

File Type	% of Web Crawler Requests	% of Non-Crawler Requests
.html (Hypertext Markup Language)	73.89%	8.12%
.htm (Hypertext Markup Language)	9.07%	0.96%
(directories)	7.53%	6.85%
.txt (Plain text)	5.07%	0.21%
.gif (GIF graphics)	1.64%	42.24%
.jpg (JPEG graphics)	1.07%	23.23%
.pdf (Adobe Portable Document Format)	0.53%	0.24%
.css (Cascading Style Sheets)	0.14%	11.45%
.js (JavaScript code)	0.07%	5.50%

### Top Crawlers

Web crawlers can be identified by reviewing the user agent field of each log entry. Table 5 lists the top ten crawlers and their frequency of visits. After excluding Ultraseek (our campus search engine crawler) we were not surprised to see that the top three crawlers are from the three major commercial Web search engine vendors: Google, Yahoo, and MSN Search. The combined crawling activities from these three

commercial search engines account for 43% of all crawler activity (after excluding crawler activity from the University's Ultraseek bot). We were quite amazed to see "YahooSeeker," a "shopbot," listed among the top crawlers, because our site is not an E-commerce site.

Table 5. Top Ten Web Crawlers by User Agent

Rank	Crawler	Requests	Pages	Web Crawler Site	Crawler Type
1	UCHSC Ultraseek	62,354	61,025	<a href="http://www.verity.com/products/ultraseek/">http://www.verity.com/products/ultraseek/</a>	Enterprise Search Engine
2	Googlebot	13,997	13,503	<a href="http://www.google.com/bot.html">http://www.google.com/bot.html</a>	Search Engine
3	Yahoo Slurp	13,445	10,002	<a href="http://help.yahoo.com/help/us/ysearch/slurp">http://help.yahoo.com/help/us/ysearch/slurp</a>	Search Engine
4	MSNBot	11,923	10,639	<a href="http://search.msn.com/msnbot.htm">http://search.msn.com/msnbot.htm</a>	Search Engine
5	LookSmart	6,845	6,788	<a href="http://www.WISEnutbot.com">http://www.WISEnutbot.com</a>	Search Engine
6	Nutch (Linux-based Open Source free search engine)	5,765	5,454	<a href="http://lucene.apache.org/nutch/bot.html">http://lucene.apache.org/nutch/bot.html</a>	Search Engine
7	FAST Enterprise Crawler	5,512	5,434	<a href="http://www.fastsearch.com">http://www.fastsearch.com</a>	Unknown
8	YahooSeeker	3,898	3,798	<a href="http://help.yahoo.com/help/us/shop/merchant/">http://help.yahoo.com/help/us/shop/merchant/</a>	Shopbot
9	Aipbot	2,737	2,606	<a href="http://www.aipbot.com">http://www.aipbot.com</a>	Unknown
10	FAST MetaWeb Crawler	2,243	2,220	<a href="http://www.fastsearch.com">http://www.fastsearch.com</a>	Search Engine

### HTTP Status

The HTTP status codes represent the outcomes of user requests. The 304 status code is of specific interest. It indicates that file caching is used by the requesting client, either the user's browser or a Web crawler. The client checks with the Web server to see if its cached version of the requested Web page has been modified. If no change has occurred since the previous visit, there is no need to download and index the file again and thus, caching reduces or eliminates superfluous loading from the Web server, and unnecessary network traffic is therefore obviated. This technique furthermore improves response time experienced by end users (Arlitt and Williamson 1997, Nicholas et al. 1999, Pan 2003, Tarr 2001).

Table 6 illustrates how Web crawlers employ this strategy to even greater advantage than the regular user. This discovery corresponds to similar discoveries made through other studies (Arlitt and Jin 2000, Dikaiakos 2005, 2003).

Table 6. Percentage of HTTP responses to Web crawlers and all requests.

HTTP Status	Web Crawler	All
2xx	41.72%	45.05%
<b>304 (Not Modified)</b>	<b>47.60%</b>	<b>42.36%</b>
3xx (except 304)	4.90%	4.30%
4xx	5.78%	8.16%
5xx	0.00%	0.13%
Total	100.00%	100.00%

### Conclusions and Implications for Libraries

This paper has explored in-depth the extent of and impact by Web crawlers on the Website of the University of Colorado Health Sciences Center Library. Analyzing our Web access logs, we have concomitantly been able to learn more about their behavior. Our results strongly suggest that Web crawlers pay frequent visits, yet give little attention to the activity levels of our Website. Crawlers possess distinctive behavior as compared to that of regular users. They focus on harvesting text-based information, ignoring image files. Our results also identify the most active top ten crawlers. The top three crawlers are from the major commercial search engines, Google, Yahoo, and MSN Search. We've also observed that Web crawlers employ extensive caching techniques to verify if webpage content has been modified since a prior visit, thus reducing or virtually eliminating unnecessary "re-crawling" and by extension, re-indexing.

There are several limitations of our study. First, in data collection the task of isolating Web crawlers from regular and legitimate users is a difficult one. The proliferation of Web crawlers is complicated by some robots, such as spambots, disguised as regular browsers and intentionally not providing identification. In this study, we were not able to identify robots camouflaged as regular users and thus in all probability, underreported actual crawler behavior. Moreover, our primary dataset is derived from only one Web server and, therefore, our results are preliminary, limited by sample size, and cannot be generalized to all library Websites or Web servers. Our quarterly data set likewise poses a potential shortfall due to lack of comprehensiveness.

Future studies should further identify the intensity and duration of Web crawler visits, the amount of information in bytes requested by crawlers, and finally – more extensively cross-compare crawler behavior with the behavior of regular users. Furthermore, future studies should include Web access logs from multiple Web servers to eliminate uniqueness of a single Web site, as well as longer time coverage (longitudinal studies) to eliminate possible seasonal effects.

"The use of robots comes at a price," stated Koster (1995) when he proposed a voluntary robot exclusion standard eventually adopted by robot writers and Website administrators. There is a delicate tradeoff

between complete accessibility versus total exclusion from your Website. Armed with a sharper, clearer picture of the behavior patterns of Web crawlers attracted to our site at UCDHSC, we have implemented a combination of measures to adjust security of our network and servers. For our Web server, we specify in the “robots.txt” file exactly which portions of our server are off-limits from crawlers. Moving expired Web pages and files to password-protected folders or unloading them from the server altogether effectively prevents our server from becoming clogged with needless or junk files. For our Integrated Library System, we exclude IP addresses from known Web crawlers so we can maximize the usage of limited user licenses. Regularly monitoring access logs and server performance is the best practice for all library Website system administrators.

Needless to say perhaps, but nonetheless important to reiterate, the challenge of the library Website administrator is unquestionably a daunting one — balancing the behavior of “benign bots” with that of your regular library user base, while all the while keeping out the “bad guys” – spambots, viruses, worms, or “poisonous cyber spiders” or any Web crawler whose intent is malign – harvesting your library Web server’s invaluable content sans authorization, getting in without a permission or a “clearance.” The very best practice we have learned, through trial, error, tons of hard work simply but complicatedly enough: “follow the middle way” to the best of your ability. Not all bots are bad, just as not all human users with a valid password are beyond reproach.

Managing your library’s Web server is as much art as it science. Studying those access logs assiduously, learning everything you can about which bots and crawlers are safe or even useful to you and your system, keeping your “robots.txt” file up-to-date, knowing your users – not just the bots but your human user-base as well.

## References

- Almeida, V., D. Menasce, and R. Riedi. 2001. Analyzing robot behavior in e-business sites. *ACM SIGMETRICS Performance Evaluation Review* 29(1): 338-339.
- Arlitt, M. and C. Williamson. 1997. Internet Web servers: workload characterization and performance implication. *IEEE/ACM Transactions on Networking* 5(5): 631-645.
- Arlitt, M. and T. Jin. 2000. A workload characterization study of the 1998 World Cup Web site. URL: <http://www.comsoc.org/ni/private/2000/may/pdf/Arlitt.pdf> [viewed August 19, 2005]
- Bauer, K. 2000. Who goes there? Measuring library web site usage. *Online* 24(1): 25-31. URL: <http://www.infotoday.com/online/OL2000/bauer1.html> [viewed August 19, 2005]
- Bertot, J. and C. McClure. 1997. Web usage statistics: measurement issues and analytical techniques. *Government Information Quarterly* 14(4): 373-395.
- Borghuis, M. 1997. User feedback from electronic subscriptions: the possibilities of logfile analysis. *Library Acquisitions: Practice & Theory* 21(3): 373-380.

- Breeding, M. 2002. Monitoring the use of your Web site. *Information Today* 19(11): 40-41.
- Cohen, L. 2003a. A two-tiered model for analyzing library Website usage statistics, part 1: Web server logs. *Portal: Libraries and the Academy* 3(2): 315-326.
- Cohen, L. 2003b. A two-tiered model for analyzing library Website usage statistics, part 2: log file analysis. *Portal: Libraries and the Academy* 3(3): 517-526.
- Compart, A. 2003. Injunction bars FareChase from AA.com. *Travel Weekly* 62(1): 4.
- Coombs, K. 2005. Using Web server logs to track users through the electronic forest. *Computers in Libraries* 25(1): 16-20.
- Covey, D. 2002. Usage and usability assessment: library practices and concerns. *Digital Library Federation and Council on Library and Information Resources*. Washington, D.C.: Digital Library Federation, Council on Library and Information Resources. 93 p. URL: <http://www.clir.org/pubs/reports/pub105/pub105.pdf> [viewed August 19, 2005]
- D'Alessandro, M., D. D'Alessandro, J. Galvin, and W. Erkonen. 1998. Evaluating overall usage of a digital health sciences library. *Bulletin of Medical Library Association* 86(4): 602-609.
- Davis, P. 2004. Information-seeking behavior of chemists: a transaction log analysis of referral URLs. *Journal of the American Society for Information Science and Technology* 55(4): 326-332.
- Davis, P and L. Solla. 2003. An IP-level analysis of usage statistics for electronic journals in chemistry: making inferences about user behavior. *Journal of American Society for Information Science and Technology* 54(11): 1062-1068.
- Dikaiakos, M., A. Stassopoulous, and L. Papageorgiou. 2005. An investigation of Web crawler behavior: characterization and metrics. *Computer Communications* 28(8): 808-897.
- Dikaiakos, M., A. Stassopoulous, and L. Papageorgiou. 2003. Characterizing crawler behavior from Web server access logs. *Lecture Notes in Computer Science* 2738: 369-378.
- Dowling, T. 2001. Lies, damned lies, and Web logs. URL: <http://www.libraryjournal.com/article/CA106218.html> [viewed August 19, 2005]
- Fieber, J. 1999. Browser caching and Web log analysis. *ASIS Midyear Conference May 25, Pasadena, CA*. URL: <http://ella.slis.indiana.edu/~jfieber/papers/bcwla/bcwla.html> [viewed August 19, 2005]
- Franklin, B. and T. Plum. 2004. Library usage patterns in the electronic information environment. *Information Research* 9(4) URL: <http://informationr.net/ir/9-4/paper187.html> [viewed August 19, 2005]
- Freeman, E. 2002. Software robots and trespass to chattels: eBay v. Bidder's Edge. *Information Systems Security* 10(6): 6-9.
- Goldberg, J. 1995. Why Web usage statistics are (worse than) meaningless. URL: <http://goldmark.org/netrants/webstats> [viewed August 19, 2005]
- Graber, M., D. D'Alessandro, M. D'Alessandro, G. Bergus, B. Levy, and S. Ostrem. 1998. Usage analysis of a primary care medical resource on the Internet. *Computers in Biology and Medicine* 28(5): 581-588.
- Graham, L. 2000. Keep your bots to yourself. *IEEE Software* 17(6): 106-107.
- Guenther, K. 2000. "Evidence-based Web redesigns". *Online* 24(5): 67-72.
- Guenther, K. 2001. "Know they remote users". *Computers in Libraries* 21(4): 52-54.

- Jenkins, C. 1997. User studies: electronic journals and user response to new modes of information delivery. *Library Acquisitions: Practice & Theory* 21(3): 355-363.
- Haigh, S. and J. Megarity. 1998. Measuring Web site usage: log file analysis. *Network Notes* 57. URL: [www.collectionscanada.ca/9/1/p1-256-e.html](http://www.collectionscanada.ca/9/1/p1-256-e.html) [viewed August 19, 2005]
- Hightower, C., J. Sih, and A. Tighman. 1998. Recommendations for benchmarking Web site usage among academic libraries. *College & Research Libraries* 59(1): 61-79.
- Huntington, P., D. Nicholas, and D. Warren. 2004. Digital visibility and its impact upon online usage: case study of a health Web site. *Libri* 54(4): 211-220.
- Institute for the Future. 2002. E-journal user study: report of Web log data mining. URL: <http://ejust.stanford.edu/logdat.html> [viewed August 19, 2005]
- Jones, S., S. Cunningham, R. McNab, and S. Boddie. 2000. A transaction log analysis of a digital library. *International Journal on Digital Libraries* 3(2): 152-169.
- Ke, H., R. Kwakkelaar, and Y. Tai. 2002. Exploring behavior of E-journal users in science and technology: transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library & Information Science Research* 24(3): 265-291.
- Koster, M. 1995. Robots in the web: threat or treat? URL: [www.robotstxt.org/wc/threat-or-treat.html](http://www.robotstxt.org/wc/threat-or-treat.html) [viewed: August 19, 2005]
- Marek, K. and E. Valauskas. 2002. Web logs as indices of electronic journal use: tools for identifying a "classic" article. *Libri* 52(4): 220-230.
- Mariner, V. 2002. Logging usability. URL: <http://www.libraryjournal.com/article/CA190393.html> [viewed: August 19, 2005]
- Menasce, D., V. Alemeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. Meira. 2000. In search of invariants for e-business workloads. *Proceedings of the 2<sup>nd</sup> ACM Conference on Electronic Commerce, Minneapolis, MN*. p. 56-65.
- Mudrock, T. 2002. Revising ready reference sites: listening to users through server statistics and query logs. *Reference & User Services Quarterly* 42(2): 155-163.
- Nicholas, D., P. Huntington, P. Williams, N. Lievesley, T. Dobrowolski, and R. Withey. 1999. Developing and testing methods to determine the use of Web sites: case study newspapers. *Aslib Proceedings* 51(5): 144-154.
- Pan, B. 2003. Capturing users' behavior in the National Science Digital Library (NSDL). URL: <http://dlist.sir.arizona.edu/848/01/nsdl-user-report.pdf> [viewed August 19, 2005]
- Ren, J., M. Zemon, and K. Rodriguez. 2000. A model for monitoring use of your library's Web site. *College & Undergraduate Libraries* 7(1): 1-9.
- Rozic-Hristovski, A., D. Hristovski, and L. Todorovski. 2002. Users' information-seeking behavior on a medical library Website. *Journal of Medical Library Association* 90(2): 210-217.
- Schuyler, M. 2001, Cutting-edge statistics. *Computers in Libraries* 21(3): 51-3.
- Silet, S. 1999. Anatomy of the Internet reference resources Web page: a UVA Library experiment. *Virginia Libraries* 45(3): 6-10. URL: [http://scholar.lib.vt.edu/ejournals/VALib/v45\\_n3/v45\\_n3.pdf](http://scholar.lib.vt.edu/ejournals/VALib/v45_n3/v45_n3.pdf) [viewed August 19, 2005]

- Stabin, T. and I. Owen. 1997. Gathering usage statistics at an environmental health library Web site. *Computers in Libraries* 17(3): 30-37.
- Stemper, J. and J. Jaguszewski. 2003. Usage statistics for electronic journals: an analysis of local and vendor counts. *Collection Management* 28(4): 3-22.
- Taha, A. 2004. Wired research: transaction log analysis of e-journal databases to access the research activities and trends in UAE University. *Twelveth Nordic Conference on Information and Documentation, September 1-3, 2004, Aalborg, Denmark*. URL: <http://www2.db.dk/NIOD/taha.pdf> [viewed August 19, 2005]
- Tan, P. and V. Kumar. 2002. Discovery of Web robots sessions based on their navigational patterns. *Data Mining and Knowledge Discovery* 6(1): 9-35.
- Tarr, B. 2001. Looking for numbers with meaning: using server logs to generate Web site usage statistics at the University of Illinois Chemistry Library. *Science & Technology Libraries* 21(1-2): 139-152.
- Thelwall, M. 2001. Web log file analysis: backlinks and queries. *Aslib Proceedings* 53(6): 217-223.
- Underwood, Lee. The Inner workings of robots, spiders, and web crawlers. URL: <http://www.winplanet.com/article/2551-.htm> [viewed August 19, 2005]
- Van der Geest, T. 1999. Evaluating a Web site with server data. *Document Design* 1(2): 131-134.
- Xue, S. 2004. Web usage statistics and Web site evaluation: a case study of a government publications library Web site. *Online Information Review* 28(3): 180-190.
- Ye, S., G. Lu, and X. Li. 2004. Workload-aware Web crawling and server workload detection. *Network Research Workshop, 18<sup>th</sup> Asian Pacific Advanced Network Meeting (APAN 2004), July 2004*. URL: <http://www.cn.apan.net/cairns/NRW/28-Ye%20Shaozhi.pdf> [viewed August 19, 2005]
- Yu, L. and A. Apps. 2000. Studying e-journal user behavior using log files: the experience of SuperJournal. *Library & Information Science Research* 22(3): 311-338.
- Zhang, Z. 1999. Evaluating electronic journals services and monitoring their usage by means of WWW server log file analysis. *Vine* 111: 37-42.