Millennium Web Logs

Web Wanderers slowing down your system? LCC1; .NET CLR 2.0.50727; Media Center PC 5.0; InfoPath.2; OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.5.30729; .NET CLR 3.0.30618") 3029

GAELIC Show ‘n Tell Workshop
held at the National Library, Pretoria
15 July 2009

Anette Lessing
Library Services, University of Pretoria
Agenda

• Why analyze logs?
• Logs from Millennium server
  – Setup and retrieving logs
  – Tools to analyze logs
• Robots, Crawlers and Spiders
  – Control by Systems Administrator
  – Robot.txt file on server
Why analyze logs?

• Support web development
  – Pages most used
  – What scopes are patrons searching
  – What browsers are used by clients
  – Where traffic is coming from

• Control over who is accessing your WebPAC
  – Robots (Web wanderers, Crawlers or Spiders used by search engines to index web content)
Set-up and retrieving Millennium logs

- Web server logs available as from Release 2007
  - Must run Millennium Rel. 2007 behind Apache WebServer
    (Innovative does not provide a log analyzer)

Setup to access logs

- Login Manager
- Web Master
- Activate Web Server Logs
Retrieving Millennium logs

In Web Master Mode - Web Server Logs
Retrieving Millennium logs

Download files and save to your PC
What the logs look like

Use a log analyzer to interpret eg.

Webalizer (only analyze one log at a time)

or

Web Expert Lite

"Web Log Expert Lite is the most straightforward, and will accept wildcards to analyze all the logs in a directory o you can have it by month....")

From IUG Listserv
Or go the Google Analytics way

- Reports than can be retrieved
  - Traffic sources
  - Visitors
  - Map overlay
  - Site usage
  - What browsers were used
  - And more (can set various filters)
Google Analytics set-up

Copy code to web page
Google Analytics Reports
(Trial log for restricted period)
Google Analytics Reports
(Trial log for restricted period)

6,556 visits came from 61 countries/territories

<table>
<thead>
<tr>
<th>Visits</th>
<th>Country/Territory</th>
<th>Page/Visit</th>
<th>Avg. Time on Site</th>
<th>Per-Visit \# Pages</th>
<th>Per-Visit Site Avg. Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,556</td>
<td>South Africa</td>
<td>1.44</td>
<td>00:02:06</td>
<td>20</td>
<td>1.14 (0.00%)</td>
</tr>
<tr>
<td></td>
<td>Norway</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Swaziland</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>India</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sweden</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>United Kingdom</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>China</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Egypt</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Iran</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Robots

- Known as Web Wanderers, Crawlers or Spiders
- Used by search engines to index web content
- Crawlers try to follow every link embedded in catalog pages
- Used by spammers to scan for email addresses
- Increase load on OPAC and slow system down
List non-local access attempts allowed

Limit NETWORK access

Are there many postings for the SAME IP address in very short time span?

<table>
<thead>
<tr>
<th>DATE</th>
<th>TIME</th>
<th>REMOTE IP ADDRESS</th>
<th>SERVICE NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>03-07-09</td>
<td>08:49:03</td>
<td>38.99.13.119</td>
</tr>
<tr>
<td>0002</td>
<td>03-07-09</td>
<td>07:26:30</td>
<td>38.99.44.101</td>
</tr>
<tr>
<td>0003</td>
<td>03-07-09</td>
<td>07:26:50</td>
<td>38.99.44.104</td>
</tr>
<tr>
<td>0004</td>
<td>03-07-09</td>
<td>07:25:50</td>
<td>38.99.44.101</td>
</tr>
<tr>
<td>0005</td>
<td>03-07-09</td>
<td>07:25:50</td>
<td>38.99.44.104</td>
</tr>
<tr>
<td>0006</td>
<td>03-07-09</td>
<td>07:24:30</td>
<td>38.99.44.101</td>
</tr>
<tr>
<td>0007</td>
<td>03-07-09</td>
<td>07:24:50</td>
<td>38.99.44.104</td>
</tr>
<tr>
<td>0008</td>
<td>03-07-09</td>
<td>07:23:03</td>
<td>72.30.87.98</td>
</tr>
<tr>
<td>0009</td>
<td>03-07-09</td>
<td>07:21:57</td>
<td>38.99.44.102</td>
</tr>
<tr>
<td>0010</td>
<td>03-07-09</td>
<td>07:21:12</td>
<td>38.99.44.104</td>
</tr>
<tr>
<td>0011</td>
<td>03-07-09</td>
<td>07:21:09</td>
<td>38.99.44.101</td>
</tr>
<tr>
<td>0012</td>
<td>03-07-09</td>
<td>07:19:57</td>
<td>38.99.44.102</td>
</tr>
<tr>
<td>0013</td>
<td>03-07-09</td>
<td>07:16:51</td>
<td>38.99.44.105</td>
</tr>
<tr>
<td>0014</td>
<td>03-07-09</td>
<td>07:16:46</td>
<td>38.99.13.117</td>
</tr>
<tr>
<td>0015</td>
<td>03-07-09</td>
<td>07:16:13</td>
<td>38.99.44.102</td>
</tr>
<tr>
<td>0016</td>
<td>03-07-09</td>
<td>07:15:52</td>
<td>38.99.44.103</td>
</tr>
</tbody>
</table>
Look up the host info

IP address lookup.
Find IP address by host domain name.

Tell me about this host
Hostname or IP address:
e.g. company.com or 12.34.56.78

The 38.99.44.104 report:
IP Address: 38.99.44.104
Hostname: crawl.13.XMXXXXXXXX

More info about this host
Country: United States
IP address owner: PSINet, Inc
1015 31st St NW
Washington
DC 20007
US

What is a Hostname?
As it is rather complicated to memorize IP addresses, most people use Domain Names instead. Domain name looks like www.company.com. The resolution between domain names and IP addresses is handled by the Domain Name Servers (DNS).

Since IP numbers are usually assigned to internet service providers within country-based blocks, an IP address can often be used to identify the country where the
Restrict Access on Innopac

Limit NETWORK access

Added entry for IP will block access immediately.
Robots.txt file on server

- `http://<your server/>/robots.txt`
- Allow legitimate search engines to index main page of catalog (set by III)
Specific segment for Google Scholar

# For the WebBridge Google Scholar Extension. Allows googlebot_IA to crawl
# /screens
User-agent: Googlebot-IA
Disallow: /acquire
Disallow: /airpac
Disallow: /airwkst
Disallow: /articles
Disallow: /availlim
Disallow: /bookill
Disallow: /bookit
Disallow: /circhistlim
Disallow: /circpix
Disallow: /cisti_order
Disallow: /clearhist
Disallow: /documents
Disallow: /donate
Disallow: /extlang
Disallow: /feeds
Disallow: /ftlist
Disallow: /goto
Disallow: /iii
Disallow: /ill
Disallow: /illframe

Allow crawling of both / and /screens

From: Mark Welge, Robots, Crawlers and spiders...Oh My! Automated Searches and your WebPAC, IUG Anaheim, 17-20 May 2009
References

• (Alan Dyck, WebPAC product manager Web Access Log Analysis IUG Anaheim, 17-20 May, 2009)