Paper 226-28

Usage Statistics for Your Web Site: Leveraging the Flexibility of SAS® and WebHound™

Lynn Rohrs, Columbia University, New York, NY
Carol Markowitz, Columbia University, New York, NY

ABSTRACT

Within our university, we have a variety of web sites for which we needed to obtain usage data, from the general (number of unique visitors to our site) to the specific (how many visitors viewing particular course descriptions went on to register for those courses).

Using WebHound, we were quickly able to see answers to many of our general questions through the out-of-the-box setup and reports. But the real payoff came when we utilized the underlying power and flexibility of SAS to create multiple "webmarts" and accompanying reports specific to several groups within our institution.

In this paper, we will explain our reasons for choosing WebHound over competing products, describe the experience of using WebHound for both novice and experienced SAS users, and detail some of the customizations we have applied in order to meet our specific needs. No previous knowledge of WebHound or SAS is required to understand this paper.

AUTHORS

The two authors of this paper are the WebHound developers for their respective sites and have worked closely with one another in this process. Carol Markowitz works for Columbia's Digital Knowledge Ventures (DKV) and Lynn Rohrs for Columbia's Academic Information Systems (AcIS).

BACKGROUND: COLUMBIA UNIVERSITY'S WEB ANALYSIS CHALLENGES

In the mid-1990's, Columbia University had been providing information on the web for a year or two. Web managers and site developers wanted to understand how their sites were being used. After consideration, Academic Information Systems (AcIS), the central academic computing department on campus, chose a public domain program called Getstats. We edited this program to meet our needs.

While the output was helpful, it was evident early on that there was a plethora of information to be garnered from these web logs that Getstats (and other similar software) could not give us. Cross-tabular processing and subsetting web logs are useful techniques for answering many questions about usage but are not easy to obtain in these primarily one-dimensional programs.

Within this context, we began writing our own homegrown web analysis tools using SAS. The functionality of SAS enabled us to group web log records based on factors such as virtual host, domain, directories within URL, and referring records, over any time period we defined. We were also able to customize our analyses to include and eliminate log records as needed (for images, staff hits, etc.).

Also around this time, others at Columbia were working on several early web publishing projects. In addition to basic reporting, the developers of these sites were interested in detailed studies that brought to life the concepts of "sessions" and "repeat visitors." They also wanted detailed information, such as whether certain online books invited more repeat visitors than others. SAS was an excellent tool for this task.

While we were able to provide many of these analyses, the cost of doing so was steep. It took extensive staff programming time, a limited resource. Web logs can be time-consuming to work with due not only to their large size and varying record lengths but also to the laborious process of identifying what we considered to be true "hits" (e.g. counting only pages and not images associated with pages). These difficulties meant that producing these specialized two- and three-dimensional reports on an ongoing basis was not possible for us.

By the end of 2001, we had at our disposal the weekly reports Getstats generated plus the occasional in-depth reports on single part of our web site for a certain snapshot of time.

DKV APPROACHES ACIS ABOUT WEBHOUND

Digital Knowledge Ventures (DKV) is a division of Columbia University that develops online seminars, certificate programs, and discipline-oriented knowledge centers. In late 2001, DKV launched Columbia Interactive (http://ci.columbia.edu), a site with two functions. The first purpose of this site is to organize campus resources, such as class websites, learning tools, and online events, and allow users to access these resources through searching and browsing. The second function is to offer over 100 "e-seminars," online classes that bring together knowledge from Columbia professors and a host of interactive teaching tools, to an audience both on and off campus.

To determine the success of Columbia Interactive, DKV wanted to get a baseline on usage of the site and all of its areas and then track usage over time to see if it increased. We also wanted to make changes and improvements and run promotions and then see if these initiatives led to increased interest. Specifically, we were interested in the following:

- Total visitors
- Unique visitors
- Total page views
- Average session duration
- Top 10 course description pages (i.e., descriptions of our e-seminars)
- Top 10 referring URLs
- Top directories accessed

While the web log records had course numbers embedded in them, we knew we would need to combine the web logs with information from a growing external database of course names in order to make the reports of course description pages easily readable. DKV therefore needed a tool that would easily combine the logs with information from our database and give us these statistics. (See Figure 1 for an example of a WebHound report that incorporates outside information).

Top 25 Course Description Pages and Percent of Total

Rank	Requested File	Course Name	Page Views	Percent of Total
1	/ciseminars/1401_detail.html	Environmental Sustainability	85	4.9%
2	/ciseminars/1002_detail.html	The Future of English	61	3.5%
3	/ciseminars/0740_detail.html	The Origins of the First World War	58	3.4%
4	/ciseminars/0301_detail.html	Mathematics of Finance	54	3.1%
5	/ciseminars/1323_detail.html	America and the Muslim World - E-Seminar 3	54	3.1%
6	/ciseminars/0751_detail.html	Slavery and Emancipation - E-Seminar 1	46	2.7%
7	/ciseminars/1001_detail.html	The Shakespearean Sonnet and the Modern Voice	45	2.6%
8	/ciseminars/1332s_detail.html	Israeli and Palestinian Nationalism - Two E-Seminars	43	2.5%
9	/ciseminars/0511_detail.html	Developing Your Classroom Management Plan	40	2.3%
10	/ciseminars/0210_detail.html	Shoenberg and Modernism	39	2.3%

Figure 1: A Custom Report Using External Data

Prior to purchasing WebHound, we compared it to other web usage tracking products. We preferred WebHound for several reasons. First, we liked the drillable and visualization reports, which are standard in WebHound. Second, we already had an existing and extensive relationship with SAS. Third, we knew that there were specific questions we wanted answered. We were aware that because WebHound is a SAS application, we would be able to incorporate database information and customize it.

However, DKV had neither the hardware nor the systems and SAS expertise to run WebHound. We approached AcIS to see if the two entities could use WebHound as a team. AcIS was already hosting the Columbia Interactive web site and housing the web log records. They were also providing usage statistics on any web sites using Columbia's central servers. It seemed a natural merger of resources.

ACIS CONSIDERS DKV'S PROPOSAL

When DKV approached AcIS, we took a look at WebHound to see what it had to offer. While it had certain limitations (the current version's reports can only be viewed under MSIE on Windows, for instance), it was clear it could give us everything we got from Getstats and more. Our main reasons for choosing to team up with DKV and use WebHound were:

- Many more out-of-the-box reports than other products
- Multiple time periods automatically available (hourly, daily, weekly, quarterly, yearly)
- Ability to edit source files
- Drillable reports
- Visualization reports

The fact that WebHound is an open-source SAS product meant we had all of the power and functionality of SAS available underneath and could take advantage of the knowledge of SAS that we have spent years developing. We would be able to obtain the information we had to program ourselves before, and we would also be able to view this information over time without additional programming.

The built-in drillable and visualization reports give a multitude of information. Drillable reports (see Figure 2) are reports that are created "on the fly" by doing live hits on the underlying SAS WebHound datasets. The values of the top-level variable are charted with counts on the first page of the report. You can then "drill down" within the variable value, essentially creating a giant cross-tabular matrix going down as many levels within as many variables as you want. Since drillable reports are easily created in WebHound simply by specifying which variables you want to use, there is tremendous flexibility. Also, because the tables are created only upon drilling, you do not need to dedicate disk space for each of the infinite possible combinations.

Top Entry Points per Week -> Referring Domains

Drill Path: 08SEP2002

Requested File	Session Starts: Sum	Session Starts: Percent of Sum
/acis/eds/data_tools/ftp-man/ftp_com.html	985	1
/acis/ets/Graesse/contents.html	576	1
/cgi-bin/cul	1,179	1
/cu/libraries/	510	0
/cu/libraries/indexes/clio.html	483	0
/cu/libraries/indexes/lexis-nexis-uni.html	681	1
/cu/libraries/inside/	3,976	4
/cu/libraries/reserves/butler/	459	0
/cu/lweb/	24,998	24
/cu/lweb/eguides/amerihist/error.404.html	1,755	2
/cu/lweb/eresources/clio.html	1,748	2
/cu/lweb/eresources/databases/	585	1
/cu/lweb/eresources/ejournals/	957	1
/cu/lweb/indiv/africa/cuvl/	1,008	1
/cu/lweb/indiv/africa/cuvl/maps.html	757	1
/cu/lweb/indiv/business/	1,749	2
/cu/lweb/indiv/chemistry/	472	0
/cu/lweb/indiv/dsc/nyc.html	1,175	1
/cu/lweb/indiv/dsc/wtc.html	3,791	4
/cu/lweb/indiv/eastasian/	416	0
/cu/lweb/indiv/mideast/cuvlm/	786	1
/cu/lweb/indiv/mideast/cuvlm/women.html	576	1
/cu/lweb/indiv/oral/sept11.html	1,611	2
/cu/lweb/indiv/southasia/cuvl/	717	1
/cu/lweb/news/files/2001-10-30.wtc_archives.html	568	1
^ALL OTHERS NOT IN THE TOP 25	52,010	50
TOTAL	104,770	100

Figure 2: A Drillable Report

Figure 2, above, shows the second page of a drillable report on "Entry Points by Referring Domains." The first page gives a list of dates. Drilling down on a date, September 8 in this case, leads to the data above. If you click on any of the requested files, you see a list of all of the referring domains to that particular entry point. In this way, you can see not only how people are coming into your site and from where, but the two metrics combined.

The visualization reports (see Figure 3) provide a picture of the mapping of your web site with color-coded values of usage and the ability to drill down the map to a particular corner of your web site and see the details. The two standard visualization reports include the number of hits to each level of web site and an actual mapping of the paths people take through your site from entry point to exit point. Figure 3 shows the latter.

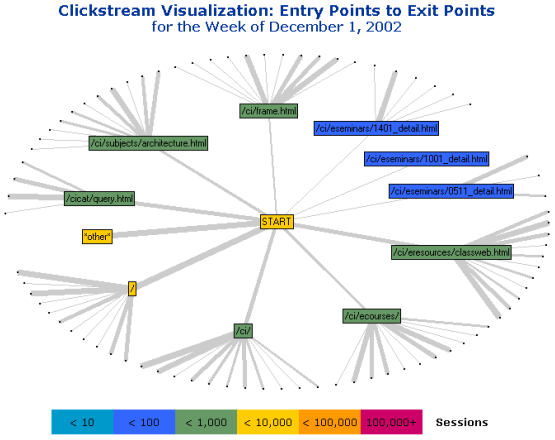

Figure 3: A Visualization Report

The thickness of the lines represents the relative counts of those pages. If you were to double-click on one of the nodes, you would see a new view in which that node is the new center and the pages visited from there are the new nodes.

USING WEBHOUND

Once we decided to purchase WebHound, we needed to learn how to install and to customize it for use at our site. Prior to purchasing WebHound, Carol had never used SAS; Lynn was a longtime SAS user. Below are our impressions.

A NOVICE SAS USER'S EXPERIENCE OF WEBHOUND (BY CAROL)

When we installed WebHound and began using it in February 2002, I had no previous knowledge of SAS. For me, the learning curve was incredibly high. To help me through the initial learning phase, I had three means of support: representatives from SAS, who made themselves available in person and via email; Lynn, an experienced SAS user; and previous experience with computers and computer programming. This last advantage meant that I knew I could do what I wanted to accomplish, even if I needed to develop the skills.

I found the setup of a new webmart to be fairly straightforward. Running reports, once someone showed me how to do it, was similarly easy. The hard part for me was creating custom reports. To create a custom report, one must take the following steps:

1. Create any necessary variables and/or formats.
2. Create or customize the summary table to be used for the report.
3. Copy an existing report.
4. Copy and rename the prep and presentation source files for the new report.
5. Modify those source files.
6. Make sure the new report "points to" the new prep and presentation files.

Some of these steps are WebHound-specific – and fairly straightforward once I did them once or twice – but steps 1 and 5 require knowledge of SAS. At first, I had no idea how to navigate the code and find the places that needed modification. Eventually, after looking at several source files, and plenty of trial and error, I began to understand the code and be able to manipulate it. Lynn's help in this process was essential, as was the ability to contact SAS for especially sticky problems. In fact, there is now a dedicated tech support person for WebHound making the support process smooth and efficient.

With our WebHound license, we purchased two weeks of on-site support from SAS. Decoupling those two weeks worked very well for us. The first week was devoted to setup and training. The second week, which we scheduled for three months later, was a time for us to ask the questions we had amassed over those months. We focused on the more complex questions and training that were more easily explained face-to-face. I was able to learn enough in that time to know what questions to ask in the second session. Once that second week ended, I felt that I had a solid grounding and could handle most tasks alone.

In any case, I now feel that if I can quantify a question, I can figure out how to find the answer using WebHound. For instance, I wanted to create a drillable report that showed pages requested, by username. Our customer service people would use this report so that they could easily determine the source of a problem as it arose for each individual. It was easy to create such a report using WebHound, once I got accustomed to the steps.

My impression is that once I got over the initial hump of learning some SAS code, I had greater control of WebHound. Using SAS, I can manipulate my data in any way that I can conceive, and I can use the SAS log to debug my code quickly and easily.

A SAS EXPERT'S EXPERIENCE OF WEBHOUND (BY LYNN)

Before we purchased and began using WebHound, I had many years of experience using SAS and had even used it to run many analyses on web logs. So I had a basic idea of what the code behind WebHound could entail.

Once we purchased WebHound, SAS came in for an initial week to assist us in getting it started. With their help, we got the software installed and the full set of standard reports up and running. We wanted to supplement them by customizing WebHound to answer additional questions. For example, I created several dummy variables that would indicate:

- A hit from a campus-based machine
- A staff hit
- The "only" hit in a session (many on-campus users have Columbia University as their default home page)

Then, I generated reports using these variables including:

- Number of visitors from on-campus/off-campus hosts
- Top pages broken down by on-campus/off-campus hosts (see Figure 4)
- Filtered drillable reports using the "staff" and "only session record" variables, which allow you to view the counts of pages and directory hits including or eliminating either or both of these

The Ten Most Popular Pages by User Location
for the Week of September 29, 2002

Top Ten Pages from CU Campus						
Rank for CU Campus Users	Page Viewed	Page Views*	Percent of Total	Rank for Off-Campus Users	Page Views*	Percent of Total
1	/cu/web/	41,012	22%	1	11,073	5.7%
2	/cu/web/eresources/clio.html	16,019	8.8%	3	4,827	2.5%
3	/cgi-bin/cul	10,670	5.9%	2	5,137	2.6%
4	/cu/libraries/inside/	7,355	4.0%	215	101	.05%
5	/cu/web/eresources/databases/	7,095	3.9%	5	3,213	1.6%
6	/cu/web/eresources/ejournals/	4,918	2.7%	6	2,638	1.4%
7	/cu/web/indiv/business/	4,839	2.7%	24	673	.35%
8	/cu/web/eresources/databases/alphabetical_list.html	2,583	1.4%	16	935	.48%
9	/cu/libraries/reserves/butler/	2,459	1.3%	348	65	.03%
10	/cu/web/indiv/eastasian/	1,986	1.1%	129	161	.08%
All others	All others	83,439	46%	All others	165,935	85%
		182,375	*100%*		*194,758*	*100%*

Figure 4: A Report with a Custom Variable

The SAS code associated with creating variables and editing reports was standard. The biggest challenge was learning how and where to do it within the WebHound environment. While maneuvering through WebHound can be cumbersome, having the understanding of how SAS works enabled me to take many shortcuts in the process. In addition, my experience with SAS allowed me to believe that anything we wanted to do with the web logs, we could do with WebHound.

In my estimation, someone who did not know SAS could use WebHound and derive substantial benefit from the out-of-the-box reports, but it might be difficult to create reports relevant to their individual web site without access to somebody who knows SAS. While we have not used it at our site, it is very possible that using Enterprise Guide could help with this process.

If you already know SAS, however, or have other similar programming experience that would speed up your learning curve, WebHound is an ideal environment within which to run web log analysis. The body of reports that comes with WebHound is large and diverse so that much of what you want may already exist. Additionally, you can build on the large foundation SAS has created and begin doing the more complicated analyses you have always wanted or have struggled to obtain.

FUTURE USES FOR WEBHOUND

Once we had WebHound up and running for two sites, leveraging it to run on additional sites was an easy task. We are currently in the process of expanding its usage to provide site usage statistics for EPIC (the Electronic Publishing Initiative at Columbia), for the Libraries web site, for our online course management and delivery system, and for several others.

Web developers for each site have been impressed with the breadth of standard reports and the information it gives them. They have also been able to define requests for additional specifications. We have been able to define each webmart differently and have added much functionality such as running reports by semester, merging with multiple databases to identify affiliations (e.g. faculty, student, etc.) for internal reporting, and discarding the developer staff hits for each site.

Accomplishing all of this growth has taken a large investment in staff time – mainly during the development stages – but the payback in terms of the information has been tremendous. Furthermore, once you have learned how to develop a single webmart, adding new reports or creating a new webmart is relatively easy and not very time-consuming. It is at this point that you begin to see the true value of WebHound.

CONCLUSION

We feel that the decision to purchase WebHound was a good one. Once we installed and learned how to use the product, it exceeded our expectations for analyzing our web logs. We obtained much of the information we had originally sought by using the reports out of the box. In addition, once we had learned the ins and outs, we were inspired to use WebHound for more web sites than we had originally intended. We found that the startup time for each new webmart dropped dramatically with greater experience, so the return increased.

If you are an experienced SAS user, it is primarily a matter of learning how to use SAS in a new context. If you are a beginner, you will have a fairly high learning curve, but once you gain basic SAS programming skills, you will also be able to create custom reports. At Columbia University, we found that we could accomplish our initial objectives and answer additional questions as they arose. We also see many ways to leverage WebHound for other sites and departments within the University. The development we have done so far is just the beginning – and we are looking forward to finding the many additional uses for WebHound.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Lynn Rohrs
Columbia University Academic Information Systems
612 West 115th Street, Room 818
New York, NY 10025
Work Phone: (212) 854-5482
Fax: (212) 662-6442
Email: lynn@columbia.edu
Web: http://www.columbia.edu

Carol Markowitz
Columbia University Digital Knowledge Ventures
514 West 113th Street
New York, NY 10025
Work Phone: (212) 854-7838
Fax: (212) 854-7555
Email: carol@dkv.columbia.edu
Web: http://dkv.columbia.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.