Paper 136-26

WebHound: Your Best Friend for Tracking Web Traffic

Dean Duncan, School of Social Work, University of North Carolina – Chapel Hill, NC
Frank Lieble, SAS, Orlando, FL
Carol Martell, Highway Safety Research Center, University of North Carolina – Chapel Hill, NC
Sally Muller School of Social Work, University of North Carolina – Chapel Hill, NC

The University of North Carolina at Chapel Hill has two state-chartered organizations engaged in e-campus and e-business initiatives. The Work First Group (WFG), in the School of Social Work, provides a web site that presents dynamic statistics and analyses about the welfare reform program, "Work First," in North Carolina. The Highway Safety Research Center (HSRC) develops project-specific sites for contracting agencies. These sites are as diverse as the projects themselves, ranging from a site designed to deliver specific information for college students, to sites created to increase public awareness. This paper presents two case studies on lessons learned from the experiences of implementing SAS WebHound^{TM} at WFG and HSRC. We describe how the study of web traffic logs allows both organizations to understand how their users are using their web sites.

Introduction

The Internet has revolutionized how we access and retrieve information. It has touched all industries including commercial, government, and academia. The need to easily provide accurate information in a timely manner is more important today than it ever has been. The need to find solutions to meet this level of service is crucial in the success of these organizations.

SAS WebHound^{TM} was used to analyze web logs at two University of North Carolina – Chapel Hill sites, the School of Social Work (SSW) and the Highway Safety Research Center (HSRC). We will discuss the need for web traffic analysis, the SAS WebHound solution, what issues it is addressing, and finally present two case studies documenting discoveries found from analyzing web logs using the SAS WebHound solution.

Why Web Traffic Analysis?

Web traffic analysis is crucial for any organization that has a web site. For example, commercial corporations or the private sector want to increase revenue and

profitability, while government agencies, universities and colleges or the public sector want to provide services and information. Even though theses two sectors have different goals they do have one goal in common. They must be able provide the highest level of service to their users. This can be achieved by:

- Providing the most relevant information to the people visiting the web site.
- Improving communication via the intranet and extranet.
- Optimizing the buying process to maximize revenue.
- Promoting the information that is the most popular.
- Creating a convenient experience by organizing information more effectively on the site.
- Improving user satisfaction by reducing the number of clicks it takes to access information.

Web traffic analysis is the study of web user usage patterns. This will allow an organization to determine:

- Where are my web users coming from?
- What path do they take before finding the information they want?

- What path do they take before leaving?
- Which pages are the most popular?
- Which pages are the least popular?
- How many clicks did it take to find the information?

But before these questions can be answered a process and/or application is needed to capture, store, manage, analyze, and report web usage data.

SAS WebHound

SAS WebHound is a solution that incorporates SAS technologies to allow web analysts to understand web traffic usage within their organization's web site. With WebHound they can identify who is browsing their web site, what are they looking at, and how often they visit. WebHound also provides ease of use and flexibility by:

- Processing large volumes of web log data (scalability).
- Integrating web data with other data sources on-line and off-line.
- Providing dynamic reporting through any browser by providing accurate information about how users interact with the web site.
- Tracking visitor usage trends across time of day, weeks, months, and even years.

WebHound will allow an organization to measure and improve the effectiveness of their web site, which will provide the highest of service to their users.

WebHound Case Studies

The following are two case studies on lessons learned from the experiences of implementing SAS WebHound. The first case study, the School of Social Work, have never performed web traffic analysis at their site. The second, the Highway Safety Research Center, have been using other web analysis tools for web traffic analysis for the past year. Each case study will introduce their organization, describe their web environment, discuss their e-problem and

solution implementation, and close with a conclusion.

Case Study 1:
School of Social Work – University of North Carolina at Chapel Hill

The North Carolina University - Chapel Hill (UNC-CH) School of Social Work (SSW) has 45 full-time faculty and more than 300 graduate level students. The School ranked 4th in the nation, according to the 2000 U.S. News and World Report survey of social work at public universities and 7th overall of 140 graduate schools of social work. At UNC-CH, the School is ranked third in externally funded projects. The School's mission is to provide an interface between the physically and economically disenfranchised people of North Carolina and the government, non-profit, and private foundations of North Carolina. The School's site (http://ssw.unc.edu) hosts 12 individual programs: such as the Work First project. The Work First's Web site (http://ssw.unc.edu/workfirst) provides longitudinal (i.e. historic) statistics on more than 335,000 families and 795,000 individuals who have received assistance through welfare (i.e. Work First) since January 1995. This SAS/IntrNet web site was developed by Dr. Dean Duncan and his staff in 1998 to provide performance measures to county Departments of Social Services (DSS) to administer the Work First program. Also it enables DSS staff, social workers, researchers, policymakers, and the public to access analyses regarding welfare reform.

Web Environment

SSW's web site (Figure 1) consists of three web servers: two of which are HP Netserver 5 servers running Windows NT 4.0 Service Pack 4. Respectively, installed with 180 MHz and 200 MHz processors, attached to 12 and 32 GB RAID drives. Both servers have 128 MB of RAM. The third server is a HP Netserver LX-Pro running Windows 2000. Installed with a 600MHz processor, 640 MB of RAM, 2 20 GB RAID drives. Two servers collect IIS web logs and one collect Netscape web logs.

The SSW web sites are distributed across all three machines, but each individual site is only on one machine. SSW decided that, rather than include all of their web logs in one WebHound report, for each web server they would essentially run a separate WebHound application. Thus, each web server has it's own URL for the WebHound reports. This case study will examine web usage logs collected for six weeks from January 2001 to February 2001.
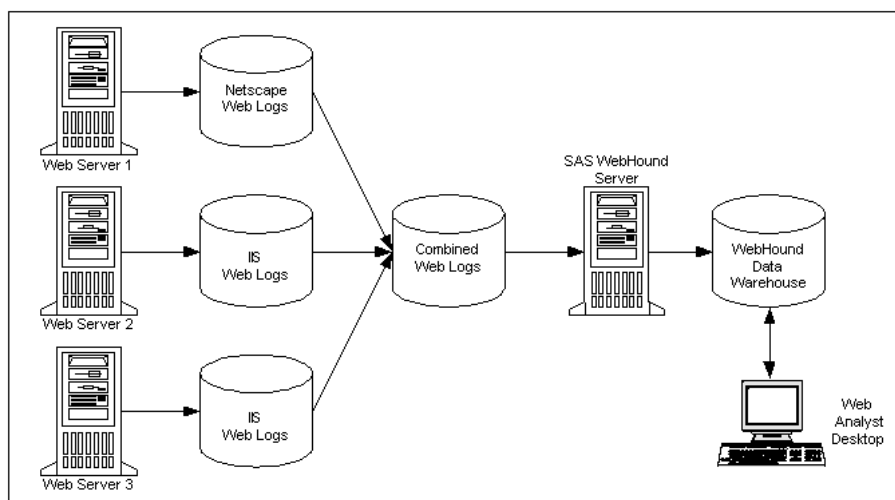
Figure 1

e-Problem

In 1999, the SSW faculty realized that their School's web site was the first place people went for information about the School, and so the dean formed a committee to look at the site. Their consensus was that the web site could make a significant contribution toward achieving the School's mission, provided:

- There are clear goals and valid measures for tracking progress.
- Faculty and students have easy access to computers, software, and the Internet.
- Faculty know how to use the technology effectively.

A member of the committee and director of the Computing Information and Technology Unit, Laura Zimmerman, Ph.D. said, "We want to implement a look and content that appeals to our students and faculty as well as potential students, researchers, faculty from other campuses, and the media." The School encourages faculty to use the web site to promote their research and courses. And because the site is often the first entry point for alumni and constituents, SSW administrators, such as the Director of Alumni Relations, increasingly are using the web site.

Just when the School's faculty was voicing concerns that students and researchers who visited the SSW site could not easily find what they needed, Duncan was considering ways his group could determine if their Work First web site was effective.

Duncan and his staff decided that in order to improve their Internet presence it would be necessary to identify the purpose of their site and then identify valid measures for tracking their progress. They identified the mission of the Work First web site as a vehicle for easy access to statistics regarding Work First recipients for DSS staff, social workers, researchers, policy makers, program leaders, and the public.

The measures Duncan and his staff identified for tracking progress:

- Can visitors, who know what they are searching for, find it?

3

- If visitors know where to go to find the information, is it still too difficult to navigate?
- Where are visitors coming from and going to and which links to the site are the most popular?
- Are pages that we promote actually the pages that are visited the most?
- Are the visitors who come to the site, the visitors that we expect?
- Are our web pages being developed in a cost-effective manner?
- Is the impact of our web pages on the School's computer minimal?

Solution Implementation

In January 2001, SAS consulting staff installed and configured the WebHound solution at the School of Social Work. A Windows NT AT process was created to archive the web log files collected by Expanded Log Format (ELF) each night. They were then copied from the three NT web servers to a fourth NT server on which SAS WebHound was installed (Figure 1). After WebHound processes the logs, reports are generated back to one of the three web servers where they can then be accessed via any web browser. The reports produced by WebHound at this point are all static. Dynamic reports are also produced using SAS/IntrNet features of WebHound. These applications are submitted to a broker, which resides on one of the three web servers. During the implementation, decisions had to be made as to where to put various pieces of the process and also how to surface the resulting reports.

As soon as the reports became available, we began investigating which of the reports provided measures for tracking progress toward achieving web site goals.

Duncan discovered that the WebHound report, "Browsers By Version," (Figure 2), could provide him with another source of information regarding his client's use of IT. Duncan can use this information in several ways, including testing new releases of the web site by using the same browsers and browser versions that visitors use. This was salient because last year Duncan got a call

from the NC Division of Social Services that a staff member was having trouble accessing the Work First site. After investigating, Duncan discovered the problem was the version of the staff's browser.
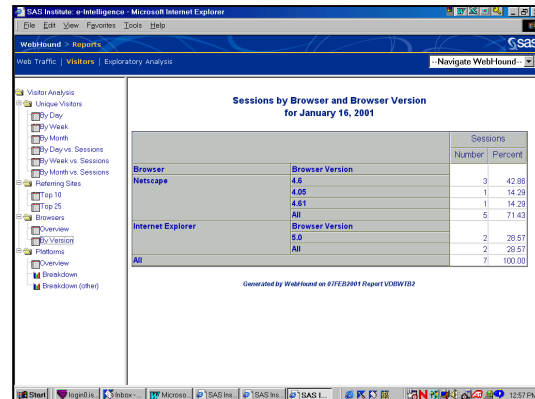

Figure 2

As the Principle Investigator of the Work First project, Duncan is also interested in the information provided by the "Sessions -- Day/Hour Contour" report (Figure 3). From this report, Duncan can determine the times of the day when usage peaks. He can examine those days when his staff have issued announcements or given presentations about the site and determine the impact of these announcements and presentations.
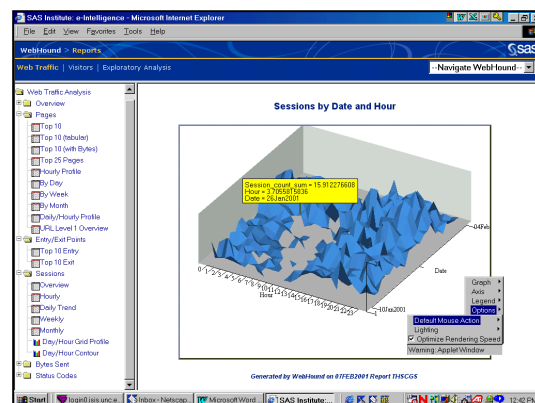

Figure 3

In justifying resource allocation to funding agencies, it is helpful to be able to substantiate projections with statistics such as the actual number of web visitors to the site (Figure 4).
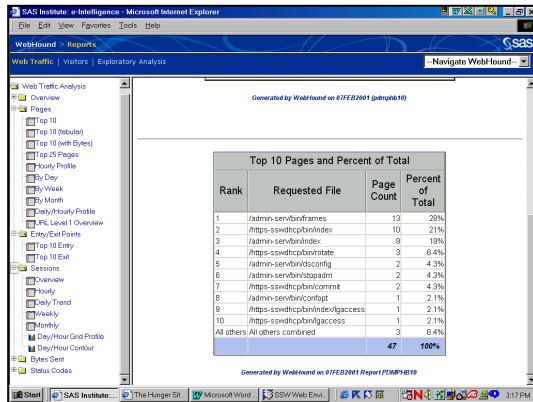
4

Figure 4

Another measure Duncan selected to track was the impact of the Work First site on the School's web server. Two of the School's system managers, Andy Broughton, Ph.D. and Manuel Garcia, were consulted. The system managers agreed that the reports they needed most were those that would allow them to identify and track "trends" in system usage. Only by identifying trends can they make informed decisions about the need for system expansion. They found that the "Sessions By Month," report (Figure 5), provides the kind and level of information that they are looking for: a summary of total number of web sessions over months. At the time of this writing, only six weeks of data were available. However, as new months are added, the system managers will be able to detect patterns of usage and use this information for capacity planning.
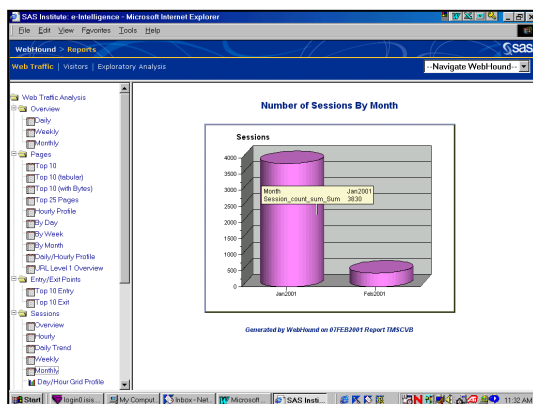

Figure 5

The system managers also found the report, "Sessions Day/Hour Grid Profile," helpful for identifying patterns of usage (figure 6). The report provides information regarding the total number of sessions for any particular

hour on any particular day, for as many weeks as were specified in the WebHound configuration. With this information, the system managers can correlate NT performance measures on memory, disk, processor, and network usage with these measures of total number of sessions for specific date/time combinations. Note that the report allows for "rubber banding" a piece of the report may be selected for further analysis.
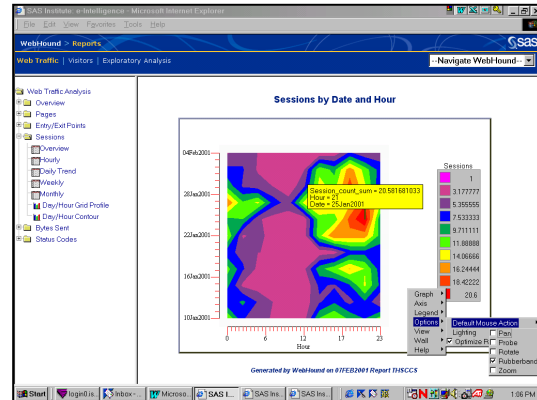

Figure 6

Conclusion

Using results "straight out of the box", the web server system managers, project director, and IT director were each able to immediately find reports appropriate to their areas of interest. The system managers have a finger on the pulse of system parameters. The project director can see what users respond to his announcements, he has documentation to justify funding, and he knows the range of browser versions for future development parameters. The IT director can learn which areas of the sites are most popular and plan future development to use the most appealing approaches. This is just the tip of the iceberg. They will next explore: 1) removing in-house traffic, 2) using the TreeView Applet to examine pathing, and 3) examining the times spent on individual pages. WebHound has already provided leads that have allowed SSW and the Work First Group to address a multitude of problems, issues, and opportunities. Now they need only follow these leads to obtain the information that they are seeking.

5

Case Study 2:
Highway Safety Research Center – University of North Carolina at Chapel Hill

The UNC Highway Safety Research Center (HSRC) conducts interdisciplinary research aimed at reducing deaths, injuries and related societal costs of roadway crashes in North Carolina and the nation. Our research addresses crashes involving motor vehicles, bicyclists and pedestrians, and takes into account the various human, vehicular, roadway and environmental components of these risks. HSRC strives to translate developed knowledge into practical interventions that can be applied at local, state, national and international levels. While public service announcements, posters and printed documents are still important ways of sharing life-saving transportation safety messages, the web offers an explosion in capabilities for marketing our research and outreach projects.

Web Environment

HSRC is, at the time of this writing, bringing online a Sun Enterprise 250 with 1GB of RAM, one 400MHz UltraSparc-II processor and three 18 GB drives to consolidate web and application server needs, which had been distributed across in-house and UNC campus servers (Figure 7). Other in-house Sun servers are an Ultra 5, an Ultra 1 and a SparcStation 5. This case study will examine logs collected in December of 2000 from a Sun Ultra 5, serving at the time as the Center's Apache web server. Virtual sites are not configured to log separately, so a single log file contains all entries. The log is archived out twice a day. DNS lookup has already been performed for these logs.
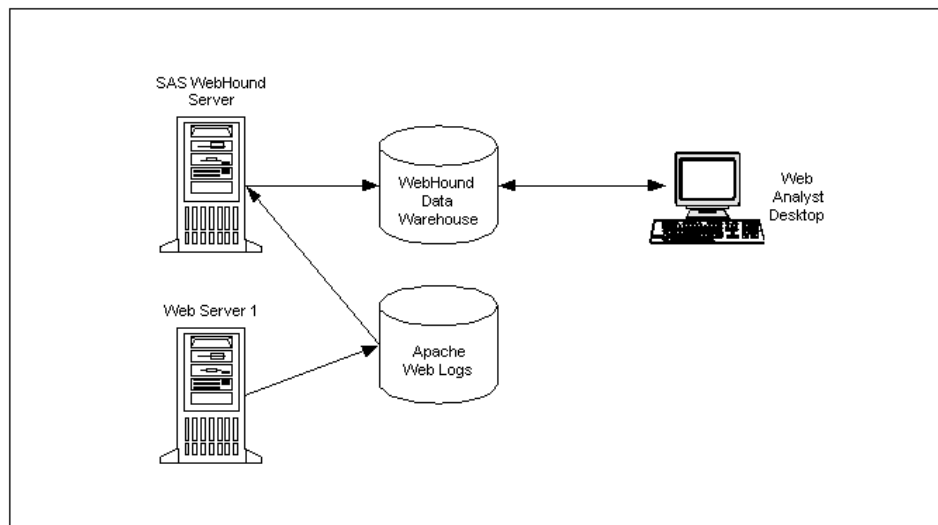

Figure7

e-Problem

Web traffic statistics have been implemented at HSRC using other software for more than a year. These numbers have allowed us to examine traffic volumes, but these static reports cannot answer all our questions. Each new web project has design strategies influenced by lessons learned in earlier projects. The web development team is extremely talented but small. Efficiency, not a new concept, must be applied to this relatively new environment.

Solution Implementation

Archived log files were transferred via FTP from the Ultra 5 to the E 250, which will serve both to process the logs and to serve the results to the web. After the initial installation and test run of WebHound, we immediately wanted to incorporate the virtual site name as a variable, and to

eliminate all in-house traffic. Since the logs were collected using the default Common Log Format (CLF), the virtual site name was not part of the logging information. Also, since virtual hosts were not configured to log separately, the virtual host could not be determined from the source file.

Our site name workaround was to employ operating system utilities to parse out the log files. We "grep'd" the logs by writing each virtual site to a separate file. The first level of each resulting filename was assigned the value we wished to see in the reports to represent the virtual site. For example, this command:

cat logfiles | grep www.hsrc.unc.edu > hsrc.log

creates a file containing every log entry involving the main HSRC virtual site. In order to use the filename to assign a value to the variable SITENAME, we added code to a catalog in the USERMODS library. USERMODS is employed to house override code for a WebHound data store. The variable SITENAME already exists in WebHound, so we were populating an existing variable with values rather than adding a new variable.

USERMODS was also the appropriate place to delete log entries from in-house users. WebHound already has a facility for ignoring traffic from of a range of IP addresses. We could not use the facility, however. These logs no longer had IP addresses since DNS lookup had already been performed. The following code, placed in entry USER_ASSIGNMENTS_AFTER_INPUT in the USERMODS.WBETL catalog performs both customizations:

```
if index(client_id,'.hsrc.unc.edu')>0 then delete;
 length _sitetmp $200;
 _sitetmp = scan(File_Name, -1, "/");
 _sitetmp = scan(_sitetmp, 1, ".");
 sitename = _sitetmp;
```

Our next task was to make SITENAME available in all the Exploratory Analysis groups. The SAS table WBCROSS contains all the class variables that will be used to create the MDDBs for the Exploratory Analyses. We found SITENAME in two groups, and added entries for the remaining groups (Figure 9)

	table_prefix	VariableName
62	referrers	Date
63	referrers	Week
64	referrers	Month
65	browsers	Browser_Version
66	browsers	Browser
67	browsers	Platform
68	browsers	Date
69	browsers	Week
70	browsers	Month
85	pagesviewed	sitename
86	urldirectory	sitename
87	visitors	sitename
88	client	sitename
89	referrers	sitename
90	statuscodes	sitename
91	browsers	sitename

Figure 9

We could have stopped at this point and had SITENAME available in all the interactive table groups. We wanted, however, to change some of the hierarchical groupings available in those tables. Again, we used override code in the USERMODS library.

Figure 10 shows the original code creating the Top Pages Exploratory Analysis.

```
%let mddbrc=;
%Create_MDDB_from_Summary_DS (
   summary_ds   = summary.pagesviewed_1_2,
   mddb         = mddbs.pagesviewed,
   mddb_label   = Webhound Pages Viewed MDDB,
   hierarchy1   = ,
   drillhier1   = Requested File->Status Code: Re
   drillhier2   = Requested File->Referrer Domain
   timehier1    = Week->Date: week date ,
   filtervars   = Status_Code File_Type ,
   repository   = &repository,
   nuniques     = ,
   nunique_labels = ,
   detail_dataset = detail.weblog_detail_1,
   _rc          = mddbrc
   );
%put NOTE: Return code from Create_MDDB_from_sum
```

Figure 10

In our situation, file structures and names are often duplicated across sites, so without SITENAME, the requested file is ambiguous. Consequently, we modified the hierarchical definitions as seen in Figure 11. A new hierarchy replaces drillhier2 and drillhier3 is the same as the original drillhier2 with SITENAME inserted as the first level.

7

```
%let mddbrc=;
%Create_MDDB_from_Summary_DS (
    summary_ds  = summary.pagesviewed_1_2,
    mddb        = mddbs.pagesviewed,
    mddb_label  = Webhound Pages Viewed MDDB,
    hierarchy1  = ,
    drillhier1  = Requested File->Status Code: Re
    drillhier2  = Site Name->Referrer Domain: Sit
    drillhier3  = Site Name->Requested File->Refe
    timehier1   = Week->Date: week date ,
    filtervars  = Status_Code File_Type ,
    repository  = &repository,
    nuniques    = ,
    nunique_labels = ,
    detail_dataset = detail.weblog_detail_1,
    _rc         = mddbrc
    );
%put NOTE: Return code from Create_MDDB_from_sum
```
Figure 11

To first examine the enhancement introduced by adding SITENAME to the WBCROSS table, we examine Top File Types. Suppose we'd like to know how the heavy use of graphics affects our traffic. The initial report is seen in Figure 12.

Figure 12

What we see is file type by week (graphic is truncated here). The boxes between the title and the table indicate that the 'down' variable is hierarchical: 'File Type->Status Code (hier)'. The other boxes indicate that the top 5 file types are displayed and that the 'across' variable is also hierarchical: 'Week->Date'. Clicking on the date causes the table to refresh with each day of that week appearing separately across the table. The hierarchy of the 'down' variable may be explored in two ways. Clicking on the arrow to the right of a file type causes the second level of the hierarchy to appear (Figure 13), whereas clicking on the file type itself drills down to a table of only that file type (Figure 14)
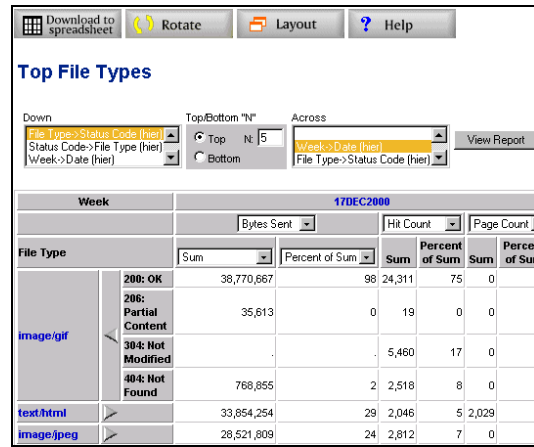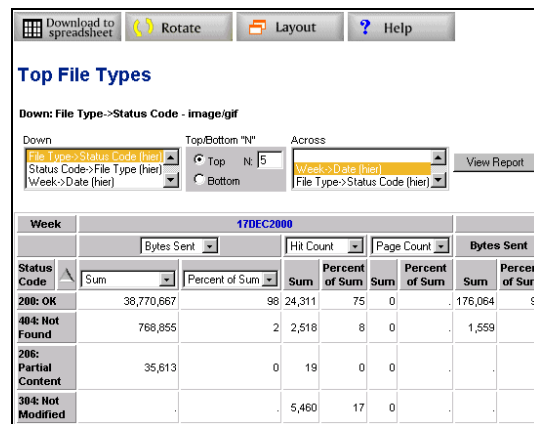

Figure 13

Figure 14

The Layout button provides complete control over the report. Every variable in WBCROSS for the group as well as the hierarchies defined in MDDB creation are available. Filter variables allow the end user browsing to subset based on filter variable values (Figure 15).
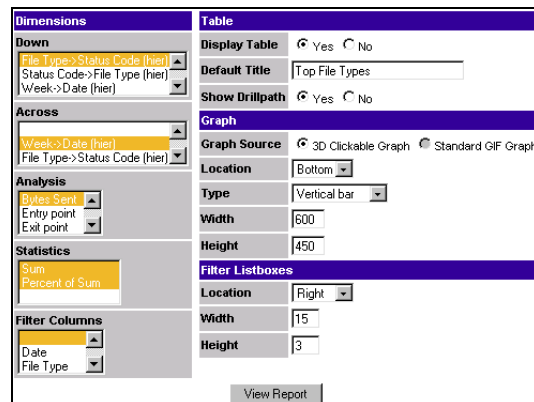

Figure 15

Changing the configuration to reveal bytes sent for file type by site generates Figure 16.

Figure 16

Changing 'Sum' in the select box in the table to 'Percent of Sum' yields Figure 17. Seeing that 89% of the bytes sent for the 2outof3 site are gif or jpg files, which represent 58% of our traffic (Figure 18) would make it seem a byte hog were it not revealed in Figure 19 that 2outof3 represents only 5% of the traffic.

Figure 17

Figure 18

Figure 19

To illustrate the success of the hierarchy modification, we see in Figure 20 that SITENAME appears as a first level in two of the available hierarchies for the Top Pages reports.

Figure 20

Conclusion

SAS WebHound Exploratory Analysis tools give us access to a frontier of information that was not provided in our other web analysis tools. Our job is to ask questions. When the answers raise more questions, it is easy to dig around for those answers as well. We have only begun to explore the information available using this e-tool. There are two areas we will pursue but have not addressed in this paper: customization of static reports to automate virtual site-specific information and implementation of the TreeView Applet, which visually presents click stream data.

Summary

The Internet has transformed how corporations, government, and academia conduct day-to-day business. The common goal of these organizations is to provide the web user with the highest level of service

9

possible. To achieve this web traffic analysis must be performed. SAS WebHound provides the ability to capture, store, manage, analyze, and report web usage data.

The School of School of Social Work (SSW) and the Highway Safety Research Center (HSRC) had the need to analyze web logs to improve the level of service for users visiting their web sites. They both installed SAS WebHound on their web site servers and started to analyze six weeks of web log data. SSW, who never has analyzed their web site, is now able to access web usage information they were not able to access before. This information will allow them to predict user impact on hardware resources and justify resource allocation to funding agencies. HSRC, who has analyzed their web site with other web analysis tools, discovered SAS WebHound exploratory analysis tools could provide additional information about their web site and user patterns. They now have the flexibility to determine how user traffic will affect their web site. These discoveries have provided new information for both SSW and HSRC to improve the level of service for their web users.

Acknowledgements

The authors thank the following people for their contribution to this paper:

Andy Broughton, Ph.D., Manual Garcia, Harvey Hou, Michael Ingraham, Andy Parks, Stephen Schultz, Christian Valiulis, Jia Xu, and Laura Zimmerman, Ph.D.

Contact Information

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

Dr. Dean Duncan
UNC School of Social Work
301 Pittsboro Street, CB# 3550
Chapel Hill, NC 27599-3550
919-962-7897
Dean_Duncan@unc.edu

Frank Lieble
SAS Institute Inc.
Orlando Regional Office
1035 Greenwood Blvd., Suite 465
Lake Mary, FL 32746
407-804-1995 x237
Frank.Lieble@sas.com

Carol Martell
UNC Highway Safety Research Center
730 Airport Road, CB# 3430
Chapel Hill, NC 27599-3430
919-962-8713
Carol_Martell@unc.edu

Sally Muller
UNC School of Social Work
301 Pittsboro Street, CB# 3550
Chapel Hill, NC 27599-3550
919-843-7798
sally@email.unc.com