Website Traffic Analysis Tools: Connecting User Data to Page Design and Content

Timothy D. Hirzel, Andrew R. Volpe and Jeffrey D. Enos
Division of Engineering and Applied Science
Harvard University
Cambridge, MA 02138
{hirzel,volpe,enos}@fas.harvard.edu

17 May 2000

Abstract

Existing web statistics visualization tools do not allow the user to analyze the effectiveness of page layout and how it affects the way in which visitors navigate through a site. We present the OSM tool for viewing link hit statistics in a per-page context, and the Parallel Page Viewer for visualizing user navigation over an entire site.

1 Introduction

Current tools available for understanding web site traffic are fairly limited in their scope. One of the most widely used tools is Webtrends' Log Analyzer [1], whose approach is similar to most competing tools such as Sawmill [2], Netstats2000 [3], and HitList Pro [4]. Using standard Apache or IIS site logs, Log Analyzer creates a multitude of pie charts, bar and line graphs, and tables of figures detailing HTTP requests broken down by large number of different metrics.

These statistics visualization tools share a common "server-centric" approach: the information gathered and presented is based on the point of view of the server rather than the user. They present what pages are being requested and from where the request was initiated, but they don't give a sense of how people are using the site. It is this information with which a web site designer or any person, technical or not, involved with the production of a web site is concerned. Important questions for such analysts and users surround the layout of a page, the navigation of the site, and how users are moving around the site.

Our goal is to create a package that will allow web site producers to explore these questions and to understand how their site is functioning from a user's perspective, that is, a "user-centric" approach that allows a designer to tailer a site to the needs of potential users. Currently these questions are difficult to answer and web site producers must conduct time-consuming user studies to address them. We seek to turn the everyday use of a site into a user study by keeping detailed information on how each user navigates through the site. Then we will present this information in an intuitive manner that allows a site designer to explore site traffic with the user-centric approach.

2 Statistics Overlay

The Overlaid Statistics Mechanism (OSM) tool aims to provide knowledge of user experience by examining each link in the context of the original page. As a result, the tool can assist in the evaluation of effective placement, style, and general aesthetic of the links in a web page. By looking at hit statistics on a per-link basis the tool can also attack problems that conventional server hit statistics can not. For instance, suppose that on a single page there are two links which reference the same page. If two users click on these different links yet arrive at the same page, conventional server statistics treat the hits equally. However, by examining the number of hits received from each link, one may be able to decide the effectiveness of one link over another, an ability lost using only server statistics.

To present per-link statistics to the analyst, the OSM tool overlays values directly onto a re-rendered version of the original web page. A rectangular box containing the statistic for a given link is placed over the lower-right corner of the link itself. The length of this rectangle depends upon the value of the displayed statistic to provide the analyst with a preattentive notion of link quality and usage without reading the statistic values themselves. In order to view the statistics of links with comparable values, the OSM tool divides the links into four "buckets" of equal size, each containing links of similar statistical value. Any subset of these buckets can be viewed by clicking on the checkboxes on the left

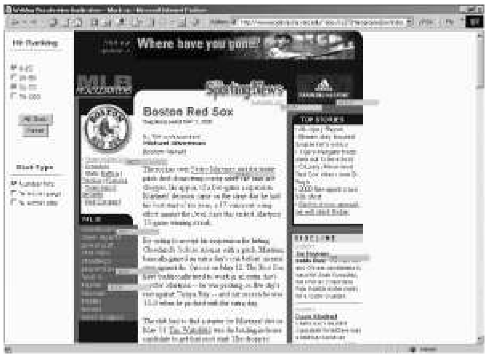

Figure 1: The OSM Tool

side of the interface. For instance, the analyst may just wish to view statistics for the quarter of the links which have the most hits on the page.

In addition to different "buckets", the analyst may select on the left any subset of types of statistics to display over each link. The types of statistics available are the number of hits for the link, the percentage hits the link received relative to the surrounding page, and the percentage hits the link received relative to the entire site. Current HTML rendering engines made the implementation of this feature impossible as of the writing of this paper. In order to place statistics over links, the OSM tool requires the location of each link on the page prior to HTML rendering. The tool's functionality calls for the use of layers, for which there are only two options for page placement provided by the CSS standard: absolute and relative positioning. Absolute positioning requires the pixel coordinates of each link on the displayed page. Using this type of positioning is useless, however, seeing as how different browsers types, font styles, and window sizes affect the location of the pixel coordinates of links, especially if they are composed solely of text. Relative positioning, however, allows the

This is text shifted using relative positioning.

Figure 2: Faulty relative positioning

programmer to position a layer with offsets relative to where it would be placed on the page if no offsets were given. However, under current HTML rendering engines, such layers are not "floated" over the page as expected. Rather, space is reserved for the layer in the final rendering, but its contents are simply moved according to the specified offset.

3 Parallel Page Viewer

Another area where a user-centric analysis can take place is on the scale of an entire web site. We wanted a way to view how users moved through the site. Currently, web site maps focus upon on the design of a site and how a user could move between pages, not upon the manner in which this navigation takes place. Advancements in the field of site mapping have been more in how to display complex graphs than in ways to map without nodes and edges.

Inspired by the success of Parallel Coordinates [5] we decided to approach the problem with a similar technique. In the Parallel Page Viewer, vertical lines represent pages in the site, horizontal lines represent links between pages, and horizontal line height represents the amount of use of that link. A closed circle is placed at the endpoint of each horizontal line, and a second open circle indicates the origin of the link. With this technique, PPL can display the most fundamental information in the most fundamental way. Before dealing with color, line thickness, or shapes, the analyst can view a map of site use.

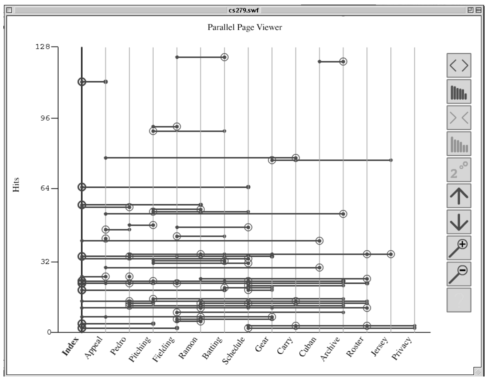

Figure 3: The PPL Tool

Several enhancements to the visualization tool complement this parallel coordinate skeleton. Colors, for instance are used to allow the analyst to preattentively filter different types of information. The horizontal lines representing links are colored green, and vertical page lines are grayed to draw attention to the links, the important data. If interested in a particular page, the analyst may click on that page's corresponding vetical line. In doing so, the appropriate page line and all lines representing links out of that page are highlighted in blue to separate them from the other data in the display. Red and orange markers appear to show total values of links pointing into and leaving a page. These markers may be turned on and off.

One of the main purposes of this visualization tool is to provide the analyst with the capability to explore data. As a result, the graphs which the tool produces are highly interactive. The pages in the display may be sorted and reordered by, for instance, the total number of links in or out of a page. Vertical page lines may be relocated by dragging the actual lines to different positions on the visualization itself. Such flexibility facilitates the discovery of trends, as well as allows the analyst to create custom visualizations for presetation purposes.

Implementation of the Parallel Page Viewer was accomplished with a combination of Macromedia Flash v4.0 and Java servlets. The servlets produced text files that consisted of variables in CGI form which could be loaded into Flash. Once the variables were loaded into Flash, through using a form of Lingo special to Flash, the graphs could be dynamically generated and updated. We chose Flash over all other GUI creation options because it is by far the easiest and best way to create an interface today. The author can easily create a fully working prototype which can distributed to test users on-line. Significant design changes can then be incorporated into the software in a matter of minutes and hours, instead of days and weeks. Because flash is a based on graphical objects called symbols, it is easy to change the appearance or performance of any single entity in order to change the entire application. Complex algorithms and tasks better left to old fashioned programming can be performed in the language of one's choice and handed to Flash to present to the user.

4 Conclusions

As websites continue to grow exponentially both in numbers of users and number of pages, it has become clear that pie charts and line graphs are not enough to understand how a site is being used. The OSM/PPV system combines two tools which provide a better sense of a site's use; both on the page and site levels. OSM gives a clear picture of the flow of users through a particular page, and the PPV gives an overview of the flow of users around the site as a whole. Together the two tools allow site producers to turn the everyday use of their site into an on-going user study.

5 Future Work

Currently the two tools are complementary, but function completely independently; ideally OSM and PPV would be joined together into one integrated system. If a user clicks on a page name in the PPV, the OSM should be brought up displaying stats for that individual page. And as the user navigates to different pages in the OSM, the currently selected page in the PPV window should change.

Our system of tracking users according to links required the creation of a unique ID for each link, and the insertion of a query string into each link to record each click through. For our logging system to work on a real site, this process would need to be automated. PPV currently can present approximately 40 pages at any one time before the screen becomes too cluttered. Most websites contain significantly more than 40 pages, so PPV needs a method to deal with displaying larger sites. One approach we have considered is to build up the pages shown by links from currently shown pages. The user would select a page to begin with, and when this start page was clicked on, each page linked to from it would be brought up (perhaps with a threshold so that seldom utilized links would be ignored). Another technique would be to break up the site into sections and display each section as a page. Links between sections would be displayed in the normal fashion; in this way the site producer can get a sense of how a very large site is functioning on a macro level.

References

[1] http://www.webtrends.com/.

[2] http://www.flowerfire.com/sawmill/.

[3] http://www.netstats2000.com/.

[4] http://www.accrue.com/products/hitlist.html.

[5] A. Inselberg. Multidimensional detective. In *Proceedings of EEEI Symposium on Information Visualization*, 1997.