

Using Interactive Visualizations of WWW Log Data to Characterize Access Patterns and Inform Site Design

Harry Hochheiser, Ben Shneiderman*

Human-Computer Interaction Lab, Department of Computer Science

*Institute for Systems Research and Institute for Advanced Computer Studies,

University of Maryland, College Park, MD 20742

{hsh,ben}@cs.umd.edu

ABSTRACT

HTTP server log files provide Web site operators with substantial detail regarding the visitors to their sites. Interest in interpreting this data has spawned an active market for software packages that summarize and analyze this data, providing histograms, pie graphs, and other charts summarizing usage patterns. While useful, these summaries obscure useful information and restrict users to passive interpretation of static displays.

Interactive visualizations can be used to provide users with greater abilities to interpret and explore web log data. By combining two-dimensional displays of thousands of individual access requests, color and size coding for additional attributes, and facilities for zooming and filtering, these visualizations provide capabilities for examining data that exceed those of traditional web log analysis tools. We introduce a series of interactive visualizations that can be used to explore server data across various dimensions. Possible uses of these visualizations are discussed, and difficulties of data collection, presentation, and interpretation are explored.

Keywords

World Wide Web, Web Site Design, Log File Analysis, Information Visualization, Human-Computer Interaction

1. INTRODUCTION

For WWW information providers, understanding of user visit patterns is essential for effective design of sites involving online communities, government services, digital libraries, and electronic commerce. Such understanding helps resolve issues such as depth vs. breadth of tree structures, incidental learning patterns, utility of graphics in promoting exploration, and motivation for abandoned shopping baskets.

WWW server activity logs provide a rich set of data that track the usage of a site. As a result, monitoring of site activity through analysis and summary of server log files has become a commonplace activity. In addition to several research projects on the topic, there are over 50 commercial and freeware products supporting analysis of log files currently available (Uppsala University, IT Support, 1999). Unfortunately, these products tend to provide static displays of subsets of the log data, in a manner that can obscure patterns and other useful information.

Interactive visualizations of log data can provide a richer and more informative means of understanding site usage. This paper describes the use of Spotfire (Spotfire, 2000) to generate a variety of interactive visualizations of log data, ranging from aggregate views of all web site hits in a time interval to close-ups that approximate the path of a user through a site. We begin with a discussion of currently available solutions and research efforts, followed by examples of the visualizations created in Spotfire. Additional examples illustrate the use of multiple, coordinated visualizations to increase the expressive power of the visualizations. Difficulties of data collection, presentation, and interpretation are discussed, along with suggestions for future improvements.

2. CURRENT EFFORTS

Log analysis efforts can be divided into two categories: products and research projects.

2.1 Products

Early log analysis tools such as *wwwstat* (Fielding, 1998), *analog* (Turner, 2000), and *Wusage* (Boutell, 2000) provide static reports based on parsed log files. While features differ, these tools generally provide a range of reports aggregating requests by date, time of day or day of week, originating regions, user agents, page, or other criteria. Tables, histograms, pie charts, and other visual aids are generally provided to aid in data interpretation. In addition to aggregate data, some packages provide user level information, such as "example visits" or "document trails", which describe paths that have been taken through the site. Most products provide numerous reports that can be easily generated, along with configuration options that can be used to fine-tune the selection of reports generated. Output is generally delivered on one or more web pages for ease of retrieval, and automated scripts can be used to generate new reports at regular intervals.

More recently, advanced commercial products have augmented this basic approach with additional expressive power. For example, *Hit List* (Accrue, 2000) and *WebTrends' Enterprise Reporting Server* (WebTrends, 2000) add reports detailing entry and exit pages, number of single page visits, and most common authenticated users, along with technical detail describing response delay and volume of content delivered (in bytes/hour).

Further innovation in web log analysis has been fueled by the need for information that can be used to drive web-based businesses. Retailers and other businesses that depend upon web traffic (and, frequently, conversion of that traffic to sales) need to understand how their site traffic translates to income. *HitList* (Accrue, 2000) and other products address these concerns by supporting the integration of external databases with web log analysis, providing site operators with the ability to link web visits to sales and customers.

Although these packages offer a wide range of functionality and analytic power, they are often constrained by the use of static, low-dimensional reports. As these graphs and charts lack interactive facilities, they must be examined sequentially. Reports generally focus on aggregations, with minimal (if any) support for direct examination of records relating to individual page requests. As a result, users are unable to dynamically explore the data. Furthermore, existing products often fail to integrate analysis output with available information regarding site topology.

2.2 Research Efforts

Since the early *WebViz* effort (Pitkow & Bharat, 1994), various projects have revisited the issue of log display and visualization. *Disk Trees* and *Time Tubes* (Chi, Pitkow, Mackinlay, Pirolli, Gossweiler, & Card, 1998) provide three-dimensional visualizations of web "ecologies", displaying the evolution of a web site over time, using attributes such as display line color or thickness to encode multi-dimensional information. Other efforts, such as *Palantir* (Papadakakis, Markatos, & Papathanasiou, 1998) and *Chitra* (Abrams, Williams, Abdulla, Patel, Ribler & Fox, 1995) examined the use of log analysis for specific goals, such as understanding of patterns in geographic origin of requests or caching performance. However, these tools lack facilities for general-purpose, interactive exploration of log data. In a different domain, *SeeLog* (Eick, Nelson, & Schmidt, 1994) combines dense displays with interactive filtering and information visualization techniques to support browsing of log files generated by telephone switching software.

Characterization and modeling of web-site access patterns has been an active area of research (Tauscher & Greenberg, 1996; Pitkow 1996; Cooley, Mobasher, & Srivastava, 1999; Pirolli, Pitkow, & Rao, 1996). While these efforts often rely upon web log analysis, their focus is generally on modeling and data mining. Visualization tools developed by NIST's *WebMetrics* project (WebMetrics Development Team, 2000) take a slightly different approach, using visualization of paths through web sites to support usability experiments (Cugini & Scholtz, 1999).

Other efforts have begun to address the electronic-commerce applications of log analysis. For example, micro-conversion rates, measuring the rates at which users view product information, add products to shopping carts, and make purchases, have been suggested as a tool for understanding the effectiveness of site design (Gomory, Hoch, Lee, Podlaseck, & Schonberg, 1999).

3. STARFIELD VISUALIZATIONS

Starfield visualization tools (Ahlberg & Shneiderman, 1994) such as Spotfire (Spotfire, 2000) combine simultaneous display of large numbers of individual data points with a tightly-coupled interface that provides facilities for zooming, filtering, and dynamic querying. By using these facilities to examine the content of web server logs, we can gain an understanding of human factors issues related to visitation patterns.

Interactive visualizations of visits to the web site of the Human-Computer Interaction Lab (HCIL, <http://www.cs.umd.edu/hcil>) were generated from the logs of the University of Maryland's Computer Science department (<http://www.cs.umd.edu/>). In an attempt to generate meaningful page request data, these logs were processed to remove any accesses that either came from machines with the cs.umd.edu domain or referenced pages outside the ``hail" subdirectory. Requests for non-HTML objects (images, applets, etc.) were also eliminated, in order to avoid generating multiple data points for any single page request. This process can be viewed as a simplified version of the pre-processing performed by WebMiner (Cooley, et al., 1999) and similar systems.

During this processing, each entry was also assigned to a category, based on a simple pattern match that assigns pages to categories based on URLs. Furthermore, client host names were parsed to allow categorization by top and second-level Internet domain names, and attempts were made to identify host names for accesses from visits that were logged only by IP number. In addition to identifying the requesting host, timestamp, URL, and Category, the resulting visualization file includes HTTP Status, number of bytes delivered, HTTP-referer [sic], and User-Agent for each hit. The available data fields are summarized in Table 1.

After they are pre-processed, the log records can be visualized using Spotfire. A visualization is constructed by selecting two of the data dimensions to serve as the x and y coordinates on the starfield plot. Additional codings can be selected by using the size and color of the points to represent other dimensions in the data. Zooming sliders can be used to restrict the values presented in the x and y coordinates. A control panel in the upper-right-hand pane contains scrollbars, alphasliders, checkboxes, and other controls in the right-hand control panel that can be used to filter points based on values of the other dimensions. Finally, clicking on a data point leads to display of complete data regarding that point. This data is displayed in the lower-right-hand "details on demand" pane. An example visualization is given in Figure 1.

Client Host	Client's Internet host name: "cs.umd.edu"
TLD	Top-level Internet host name: "edu"
SLD	Second-level Internet name: "umd.edu"
Timestamp	Date and time of Client's request: ``980822 17:05:03" indicating August 22, 1998 at 5:05:03 PM EST
URL	Uniform Resource Locator: the name of the file that was requested
Category	Classification within the web site. Possibilities include projects within the group, such as "Visible Human", "Pad+", or "Lifelines"
HTTP Status	The web server's response to a request. Values include "OK", "Unauthorized", "Not Found", and other values specified in the HTTP specification (Fielding, Gettys, Mogul, Frystyk, & Berners-Lee, 1997).
Bytes	The size of the resource delivered, in bytes
HTTP-Referer	The URL that the user's browser was on before making the current request. When present, identifies the page that links to the requested page
User Agent	A description of the specific client software used to make a request (e.g., "Mozilla/4.0 (compatible; MSIE 4.01; MSN 2.5; Windows 98)"). Can be used to identify user's operating system and browser. Also useful for identifying WWW robots - automated web traversing programs. Example robots include "ArchitextSpider" and "Slurp/2.0 (slurp@inktomi.com; http://www.inktomi.com/slurp.html)"

Table 1: *Visualization Data Fields*: Each data point in the pre-processed log files has values (possibly null) for each of these 10 attributes. In a visualization, two of these attributes will be used as the spatial (x & y) dimensions of the starfield, with other attributes being optionally represented by size and color of the points, with dynamic query sliders, checkboxes, and other controls used to filter out data points based on values of the other attributes. Figure 1 presents a

sample visualization, with time on the x-axis, URL on the y-axis, HTTP-Status represented by color, and the topmost slider adjusted to limit points displayed to one specific value of Client Host.

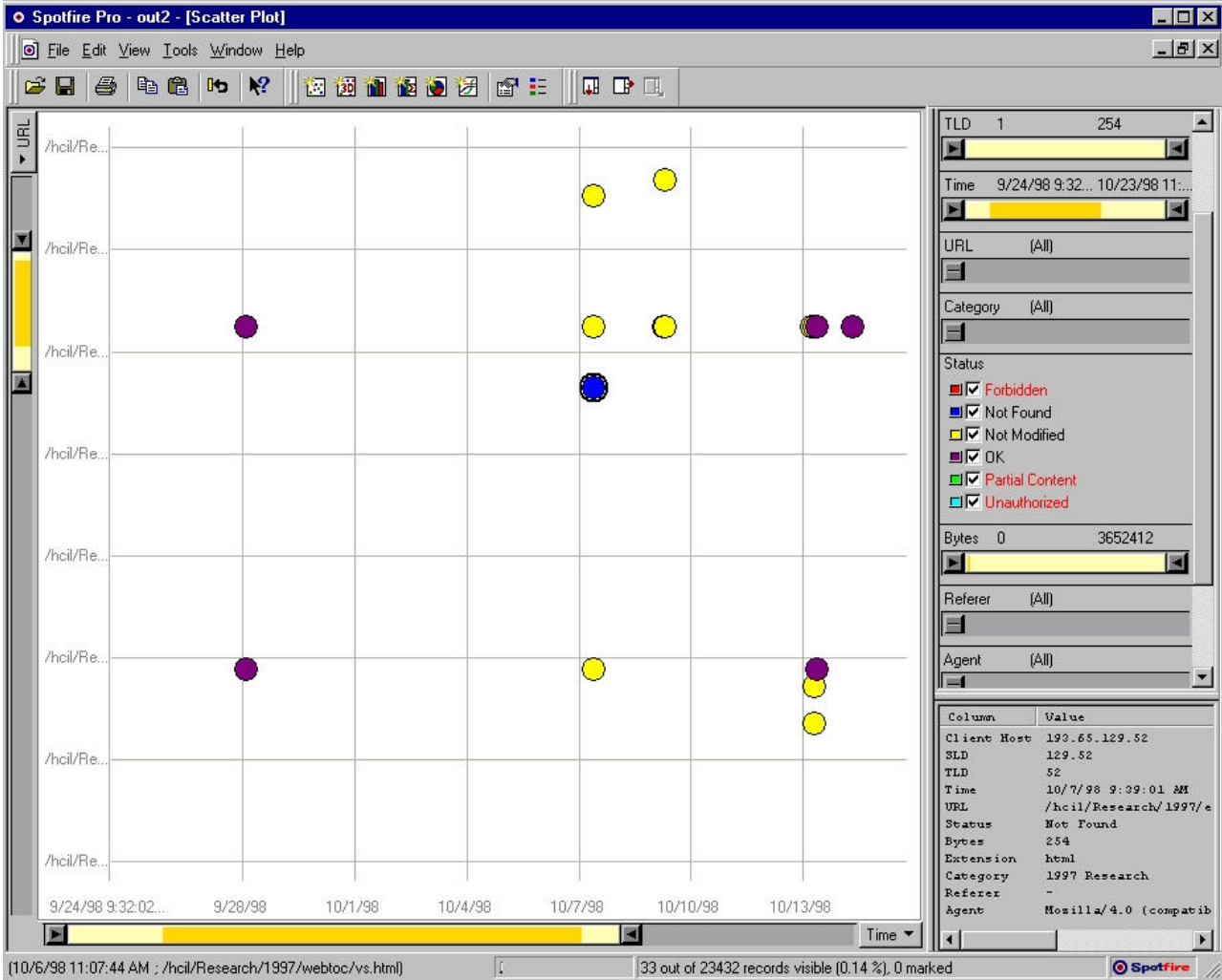


Figure 1: *Interactive Visualizations in Spotfire*: A Spotfire visualization, with the URL requested on the y-axis, the time of request on the x-axis, and the HTTP-Status is represented by the color of the data points. Zooming sliders on the axes have been adjusted to narrow the time values to approximately 10 days on the x-axis, and URLs (on the y-axis) have been restricted to a limited alphabetic range. The topmost slider in the right-hand control panel has been adjusted to limit requests to a single host. The resulting display shows requests made by one or more users from that host machine to URLs in the limited alphabetic range during the restricted time range. The other sliders in the control panel might be used to further restrict the points displayed by eliminating undesired values of SLD, TLD, Time, Category, or other attributes (see Table 1). From this display, it can be inferred that the user(s) of the given machine visited several pages on the site during this time, with at least one visit on each of four days. Furthermore, repeated revisitation of a single page is indicated by the presence of several points on the same horizontal line. The lighter color points correspond to the HTTP-status code “not modified”, indicating that the page requested has not changed since the data specified by the browser, and the client’s cached copy can be used.

For a two-month period covering late August to late October 1998, the resulting data set consisted of over 33,000 data points. This data was used to generate several visualizations, some of which required additional processing.

3.1 Time vs. URL, Macro View

Accesses were plotted with time on the x-axis and URL (alphabetically) on the y-axis. Secondary codings include size coding for document size and color-coding for HTTP response code. This “all at once” overview provides a high-level view of major usage patterns of web site visits (Figure 2), including:

1. HTTP errors: color-coding of HTTP status responses allows for quick visual scanning to identify document requests that caused errors (Region 1)
2. Correlated references: short vertical groupings indicating pages that had similar URLs (due to prefix similarity) and references that were close together in time (Region 2)
3. Bandwidth usage: frequency of hits to larger files (Region 3)
4. Usage frequency: Dense rows of points indicate frequently requested pages (Region 4)
5. Weekly usage: vertical “lanes” of lower hit density correspond to weekends (Region 5)

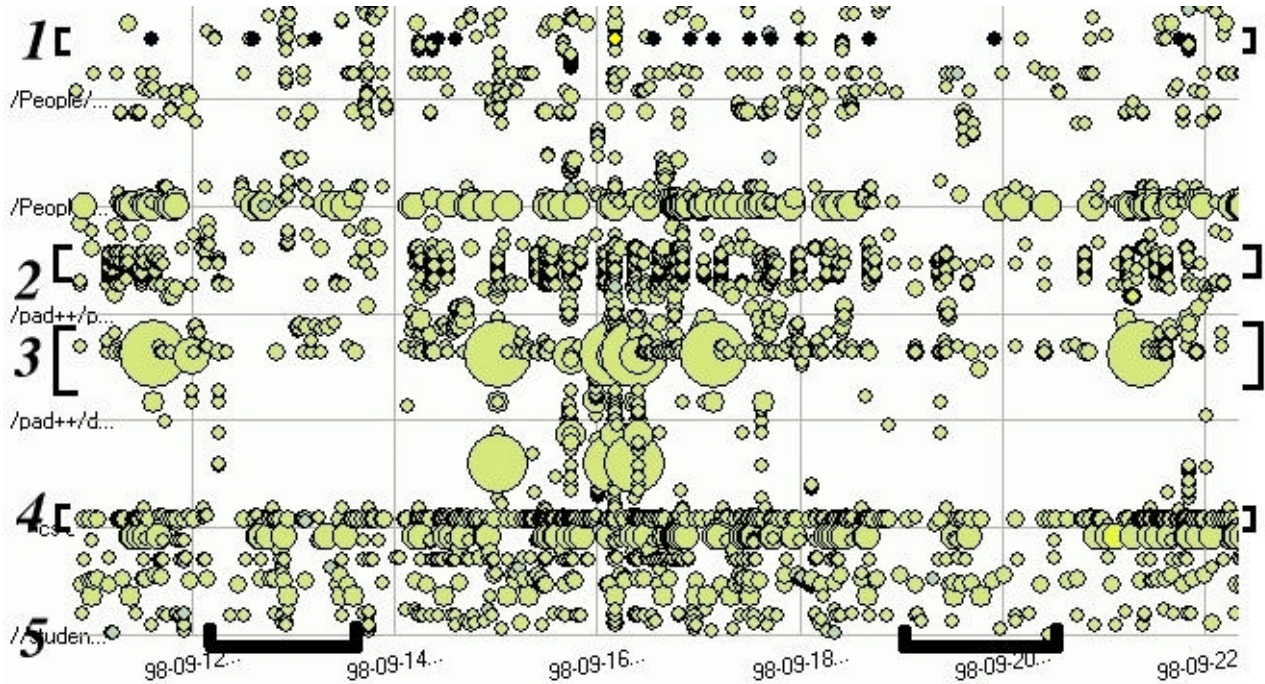


Figure 2: *Time vs. URL, Macro View*: Two weeks of accesses to a subset of the HCIL pages. The requested URL is on the y-axis, with the date and time on the x-axis. The dark lines on the x-axis correspond to weekends. Each circle represents a request for a single page. The size of the circle indicates the number of bytes delivered for the given request. Color is used to indicate the HTTP status response, with the majority of points being “OK”, indicating a successful request. Labeled regions point to examples of interesting patterns that can be seen in the visualization: 1) Color coding for HTTP status codes allows for quick identification of errors: the straight line of darker points indicates an error response corresponding to a non-existent URL that is frequently requested – perhaps from an outdated link on an external page. 2) Groups of access points clumped together vertically indicate pages that both have similar URLs and were accessed at points close together in time, possibly indicating user sequences of requests that form user sessions. 3) Large circles indicate large files. Frequent accesses to such files might cause concerns regarding bandwidth allocation. 4) The group home page, “/index.html”, shows a steady stream of visits, as indicated by the horizontal line of access points that spans the entire graph. 5) Lower frequency of requests on weekends is indicated by vertical lanes of lower request density.

By displaying all of these usage patterns in one screen, the visualization gives a compact overview of site activity. Due to their qualitative nature, these observations are more useful for identification of potential areas of interest than for direct comparison. However, Spotfire's zooming and dynamic query facilities can be used to quickly narrow in on interesting subsets of the data.

Replacing URL with category on the y-axis groups points into horizontal bands, based on the semantic category assigned during pre-processing. While potentially hiding the information carried in the distinct URLs, the discrete categories provide a more orderly display that can simplify investigations of relative usage of different parts of the site. Specifically, category usage information may provide insights into the topics and areas that were of interest to users, as opposed to simply identifying the pages that were accessed. This information might be useful for designers interested in focusing maintenance efforts on the most highly used portions of a site, or for researchers testing hypotheses about site design.

3.2 Time vs. URL, Micro View

Zoom and filter techniques can be used to modify the time vs. URL visualization to display lower-level usage patterns, such as per-host visits. By restricting the above visualization to display hits from particular clients, we can examine patterns of repeated visits over extended periods of time, in order to identify host machines that may have repeatedly returned to the site over the course of several weeks. Zooming in to display smaller time slices provides a potential visualization of the events in a given visit (Figure 3).

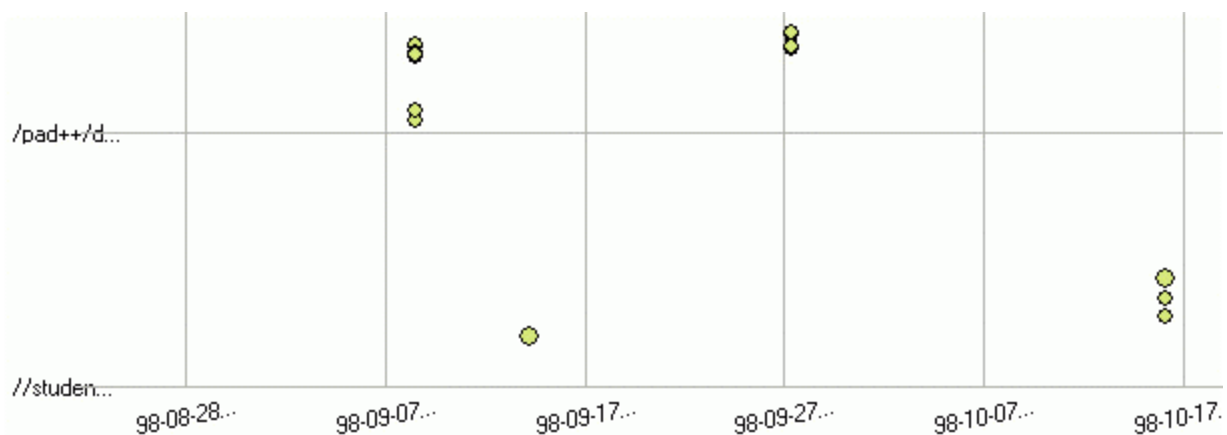


Figure 3: *Time vs. URL, Micro View*: A series of requests from a single client. Over the course of five weeks, this client made several series of requests to the HCIL web site: 4 pages on September 8, one on September 14, 3 on September 27, and 4 (of which three are shown) on October 16. URLs are alphabetized on the y-axis, so closely-packed points in a vertical line are accesses occurring on a single day involving files with similar file names. Each of these request clusters may constitute a visit to the site.}

Of course, these visualizations must be interpreted carefully: hits from hostnames that indicate proxy hosts or dynamically-assigned hostnames (for ISP dialups) are less likely to indicate single visits from a small group of individuals.

Use of this visualization to examine patterns found for multiple hosts can also reveal some interesting patterns. For this data set, this visualization clearly indicated that the vast majority of individual hosts had recorded only one request to the site.

3.3 Time vs. Hostname

Examination of trends in accesses by hostname can provide insights into the patterns of visitors into the web site. By plotting time on one axis and fully-qualified-domain-name (or IP number, if the complete domain name is unavailable) on the other axis and maintaining the size and color codings used previously, we can see trends in requests from different hosts.

As with the “time vs. URL” visualization (Section 3.1), this display may illustrate usage patterns that would not be obvious in output from traditional log analysis tools. For example, horizontal lines indicate sites that have been visited repeatedly by a given host, perhaps over a period of days or weeks. Particularly strong trends in the horizontal - a given host visiting the site repeatedly and regularly over an extended period of time - may indicate a visit from an automated web agent, or classes of visitors coming from a proxy or cache server (Figure 4).

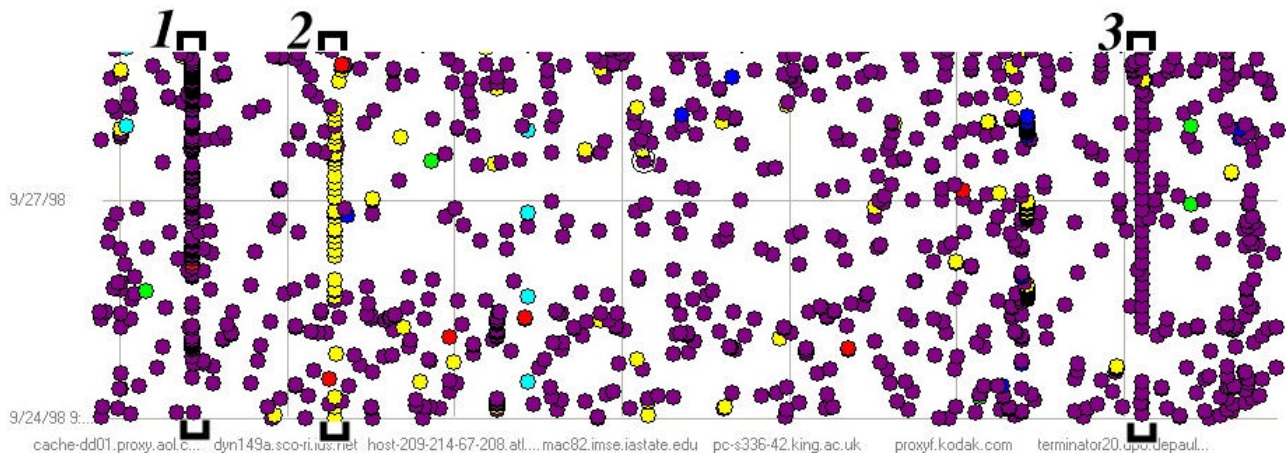


Figure 4: *Time vs. Hostname*: Time of access is displayed on the y-axis, client hostname on the x-axis, and color is used to display the HTTP-Status. Strong vertical lines – as exemplified by the highlighted sections labeled 1, 2, and 3 – indicate hosts that have visited repeatedly over the time period displayed. Patterns as strong as these are usually caused by web crawlers: examination of detail points in regions 1 and 2 reveals – in both cases – access to a wide range of pages by hosts with hosts names such as “crawler.foo.com”, strongly indicating a crawler visit. Pattern 3, however, is a series of repeated requests to a single page by a single machine. Although the reasons for such repeated requests are not clear, this pattern is potentially interesting and would have been hard to identify with traditional tools.

Changing the view to display second-level domains (e.g., umd.edu) or top-level-domains (e.g., .edu) provides information regarding the organization or locality of the originating host. Filtering and zooming to specify specific hostnames can be used to provide another version of the usage patterns from individual hosts described under the “time vs. URL, micro view” visualization (Section 3.2).

Unfortunately, the high frequency of hosts that do not have resolvable hostnames results in a large proportion of the hits being classified by IP number only. Furthermore, some of the hostnames that were found in the log either came from proxies (proxy.host.com), or were obviously associated with dialup PPP lines (ppp.dialup.isp.net). In the data set used to generate these visualizations, approximately 2500 hits (roughly 7%) involved hosts with names containing “proxy” or “dialup”, and approximately 6200 (roughly 18%) were identified solely by IP number. While these percentages are not necessarily typical, these difficulties clearly present challenges for any analysis system that hopes to extract useful information from hostname information in log files.

3.4 Client Host vs. URL

Visualization of client hostname (x-axis) vs. requested URL (y-axis) can illustrate trends in access patterns for individual Internet hosts. In this display, each vertical lane corresponds to requests from a single host: examination of these lanes can provide insights into the files requested by different hosts.

This display might also be used to identify URL request patterns that are shared by multiple hosts. Specifically, multiple parallel vertical lanes that have data points (hits) in the same vertical positions indicate groups of clients that visited similar pages. Unfortunately, the alphabetic ordering of client hosts and URLs might make such patterns difficult to identify.

The visualization might also be used to identify visits from web robots. Vertical lines that extend throughout large portions of the URL space show time periods when many pages on the site were hit by a single host in a short time period, indicating a possible robot visit (Figure 5). This information may be useful for site operators interested in knowing when an automated agent is visiting their site.

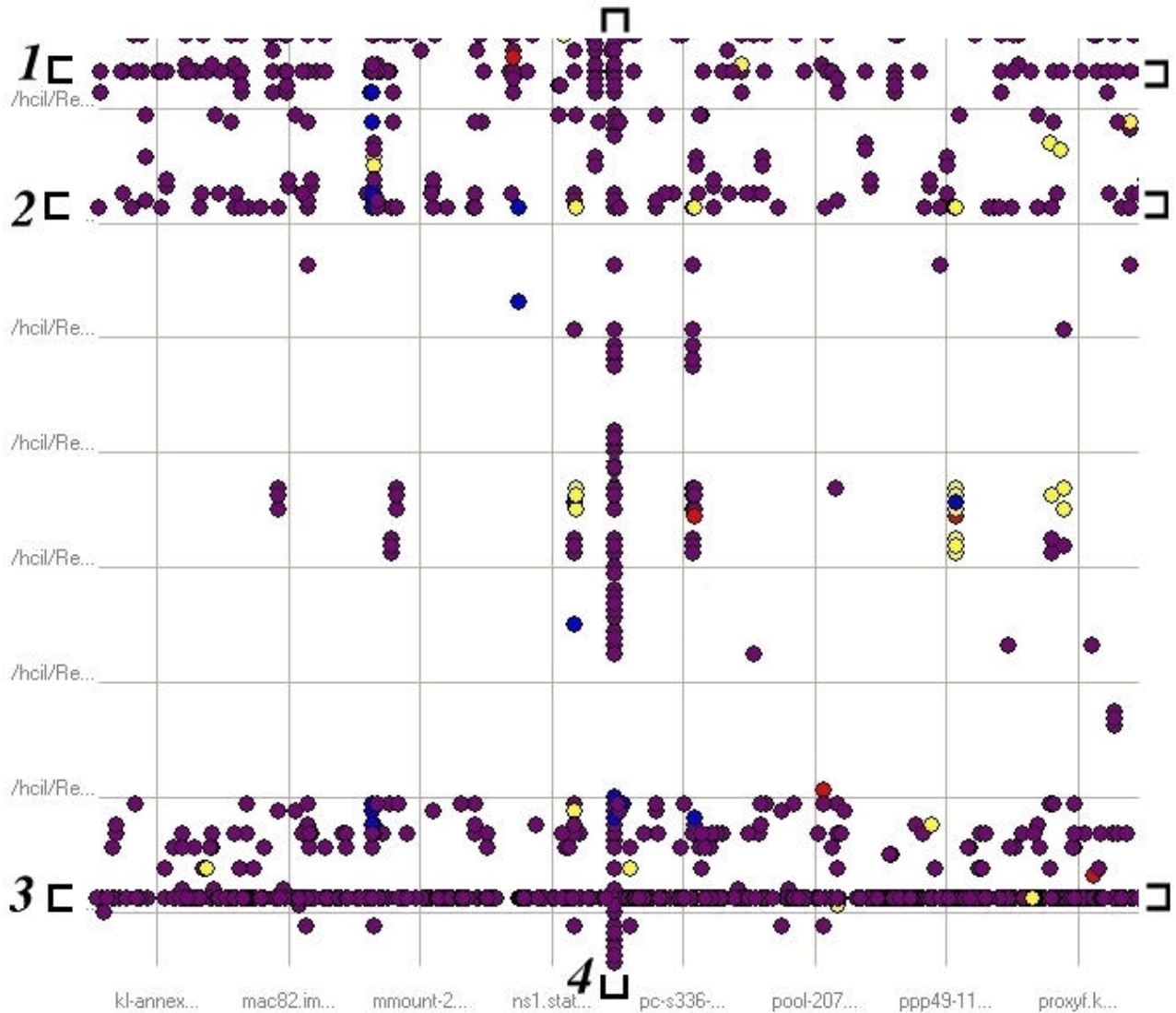


Figure 5: *Client Host vs. URL*: Requested URL on the y-axis and client hosts on the x-axis, with borders around patterns added for emphasis. This visualization depicts patterns of page accesses by hostname, providing insight into which pages are requested by which hosts. The rows pointed to by regions 1, 2, and 3 correspond to URLs that are accessed by a significant fraction of the hosts that visit the site, with the relative “coverage” indicated by the density of the row: almost all of the hosts visit the URL corresponding to region 3, while somewhat fewer visit the URLs corresponding to regions 1 and 2. Since region 3 points to the URL for the lab home page, it is not surprising to see that it was visited by most of the hosts that visited the site. In the HCIL data, very few of the URLs had dense access patterns similar to those indicated by the arrows. Vertical slices in this diagram indicate the pages visited by a single host, with densely packed vertical lines corresponding to hosts that visit large portions of the site. The host corresponding to region 4 gives an example. Closer examination confirmed that this host was a web crawler.

Of course, the difficulties with unidentified or uninformative hostnames (described above) apply to this visualization as well.

3.5 Index Page Link Requests

Researchers and web site designers may be interested in using data regarding hits to links on a site's home page as a means of evaluating the effectiveness of the site's design. One way to perform this assessment would be to track the frequency of user visits to URLs that are referenced from the home page. In order to visualize this data, we reprocessed

the visualization files, calculating the total number of hits per day per linked URL for each of the 35 links found on the HCIL home page. As part of this processing, each URL that was linked from the index page was assigned a number (links on the home page to off-site resources were ignored). Numbers were assigned in descending order, starting with -1 for the top link on the home page, thus guaranteeing that a link's position in the visualization will correspond to its position in the home page.

This revised data was then displayed in a visualization, with date of access on the x-axis, rank on the y-axis, color coding for the URL, and size coding for the number of hits on each day, with larger points indicating more hits. This provides a visualization with a series of horizontal lines, each tracking accesses to a given link on the HCIL home page.

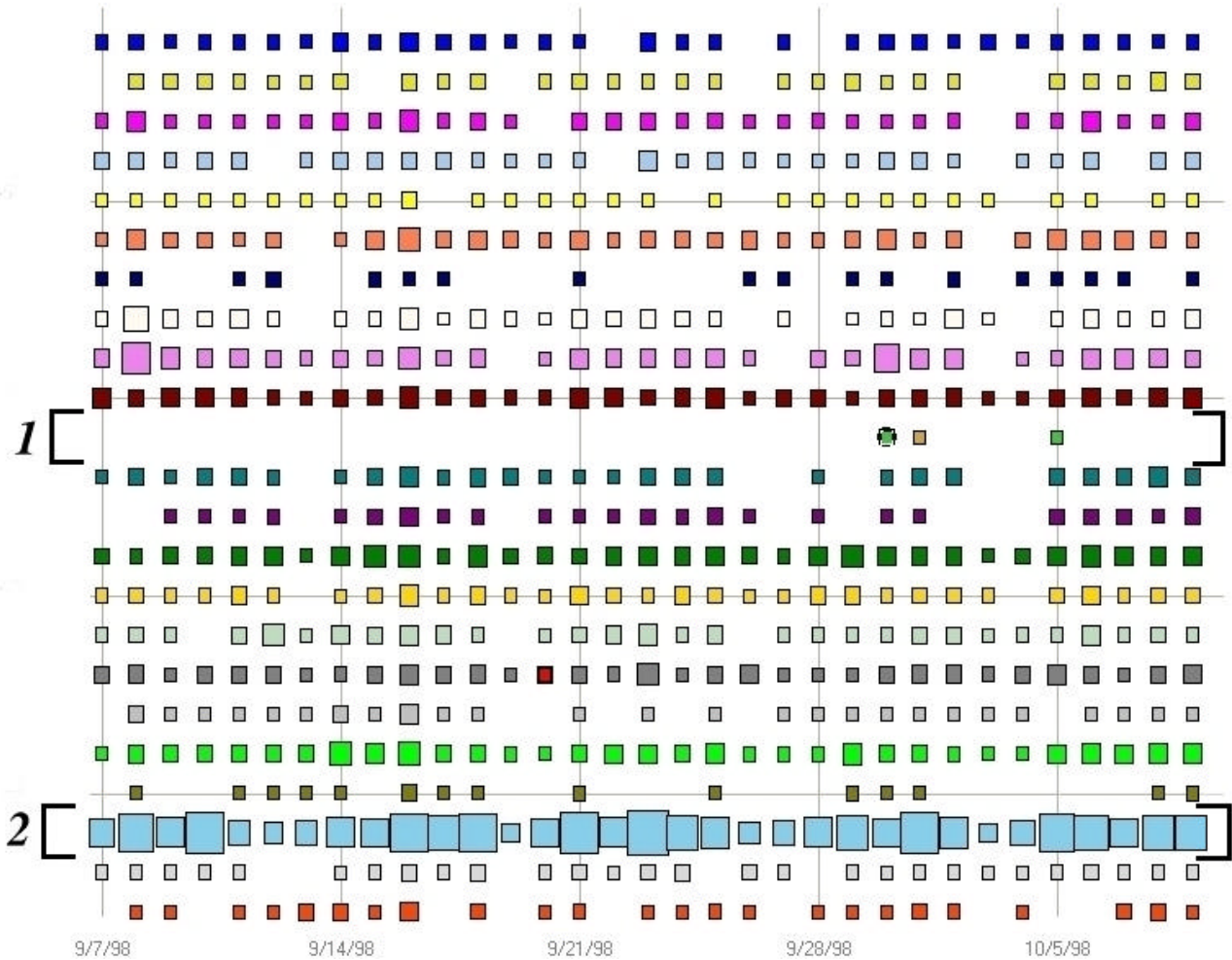


Figure 6: *Index Page Link Requests*: Requests for pages that have links on the group index page. Each row corresponds to a link on the index page. The vertical position of each row in the visualization corresponds to the vertical position of the link on the index page, with links at the top of the page found at the top of the visualization. Date of access is plotted on the x-axis, and the points are scaled to indicate the relative number of requests on each day - larger points indicating more frequent accesses. The regions bracketed on the left-hand side display two patterns of interest. Region 1 shows the requests for a link to one of the HCIL project pages. This page is not referenced until September 30, as indicated by the empty space in the left side of the row. However, the link is placed fairly close to the top of the page. Region 2 corresponds to the link to HCIL's frequently accessed technical report page, which is placed further down the page. These observations might be used to redesign the home page: for example, this information might be used to justify moving the link to the technical report page to a more prominent position.

This visualization can be used to track frequency and regularity of user visits to the home page links. However, as references to pages linked from the home page do not necessarily involve selections from that page, this display can be somewhat misleading. Specifically, in situations where site visitors might arrive at these pages by selecting links from some page other than the home page, or by typing a link directly into their browsers, this summary might be very inaccurate. This was the case in the current data set, as this visualization helped confirm our suspicions that many of the user visits to HCIL pages were coming from external links.

This one-screen display of the relative frequency of use of the various links can provide valuable insights to designers and webmasters interested in improving page performance. For example, rarely-used links towards the top of a page might be occupying space that would be better allocated to more popular resources (Figure 6) (Nielsen, 1999). Alternatively, high-interest items found at the end of a long page might show lower levels of access, perhaps reflecting users' unwillingness to scroll to the end of longer pages.

3.6 Referrer vs. Time

Many web site operators are interested in understanding their site's position in the web universe. While search engines may provide facilities for searching for links to a given URL, such searches do not provide any information about the actual use of these links. Fortunately, many web logs contain the HTTP-referer field, which indicates the URL that a browser was viewing before a given page request was made, thus indicating the page that led to the request. Log files containing HTTP-referer fields can be used to derive visualizations that might provide some valuable insights into the use of internal and external links. By plotting time on the x-axis, referrer URL on the y-axis, along with color coding for HTTP status and size coding for size of resource requested, we can generate a visualization that displays trends in referring URLs that lead users to the site (Figure 7).

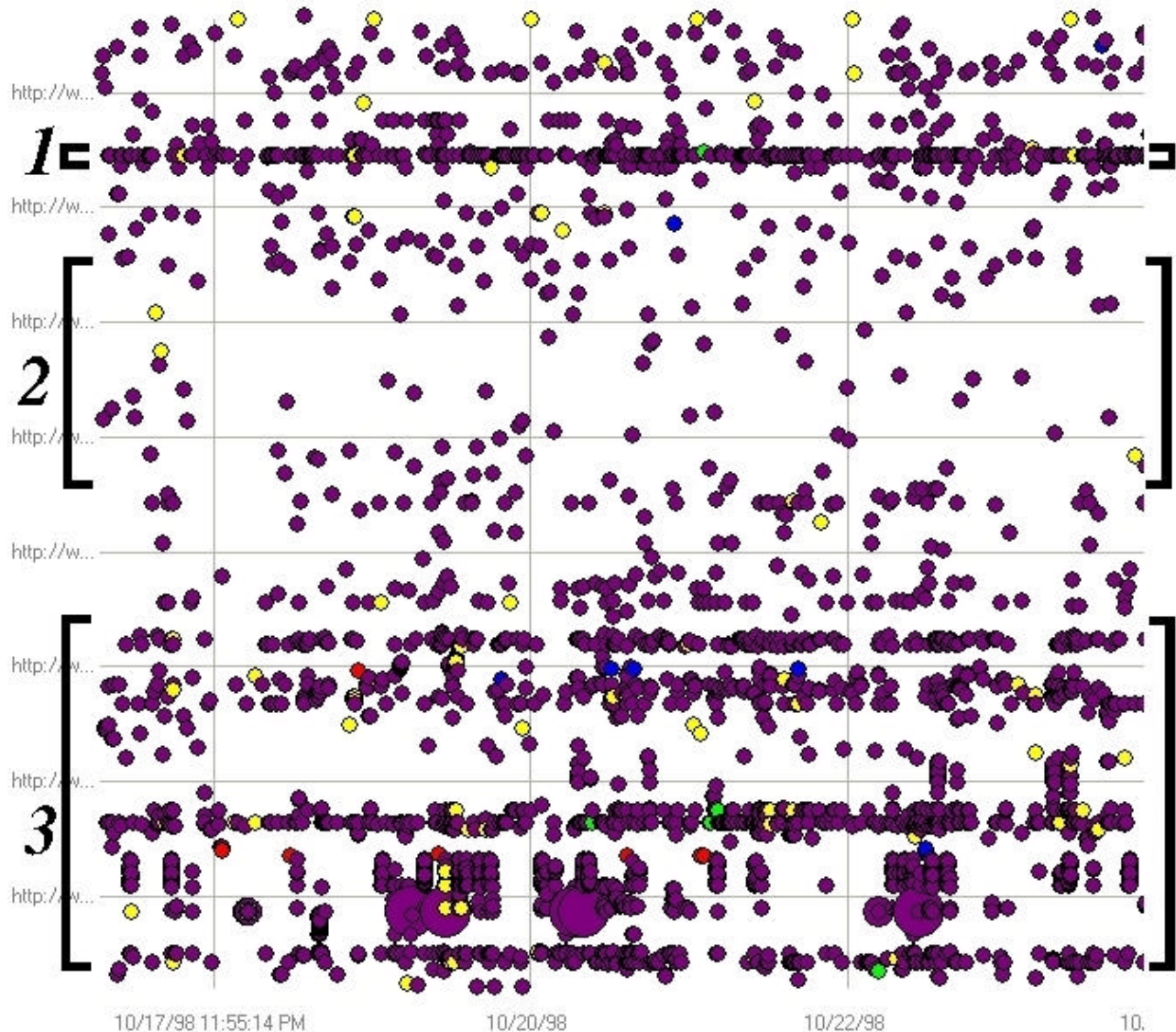


Figure 7: *Referrer vs. Time*: The URL of the referring page is given on the y-axis, and the request date and time is on the x-axis. The line marked by bracket 1 indicates a URL at the National Library of Medicine that consistently refers users to the pages for the HCIL's Visible Human project. The area indicated by the bracket 2 is a band of referrer URLs corresponding to requests that came from a search engine. The repeated and frequent occurrence of such requests indicates that users are regularly using search engines to find pages within the HCIL web site. Bracket 3 is a band of referrer URLs that are within the HCIL web site, indicating visitors who went from one page to another within the site. As this class makes up a significant portion of the data points, we can conclude that many of the page visitors to the HCIL site move from one page to another within the site.

For example, dense horizontal bands indicate referrer URLs that are continually and regularly leading people to the site. Of these URLs, external sites are likely to be the most interesting, but internal referrers may provide interesting clues as to which links on the site are being used. Furthermore, changes in the referrer profiles over time may indicate the addition or deletion of new links to the site.

Examination of the range of referrers is also instructive. Search engines often return responses to queries as dynamically generated content with similarly dynamic URLs. As a result, visits that originated with search engines have distinct referrers, leading to horizontal bands in the visualization. Each of these bands indicates a class of visits from a single search engines. Furthermore, search terms are often encoded in the URLs of search results, so

examination of individual referrer URLs for these search engine referrers may provide some insights into the search keywords that are leading visitors to the site.

3.7 Referrer vs. URL

Further insight into paths that users take to reach various pages can be gained by plotting the HTTP-referrer (x-axis) vs. the URL being retrieved (y-axis), while maintaining the size and color codings used above for HTTP status and resource size, respectively. While this visualization may provide interesting insights, the presence of a large number of intra-site and search engine referrers may lead to possibilities for misinterpretation. If these potential confounds are properly accounted for, several interesting patterns may be observed:

- *Pages accessed from a variety of external referrers:* Horizontal bars correspond to pages that are referenced from multiple sources - either external or internal. These bars may be used to gauge the relative external visibility of different web pages, in a manner that identifies the links that actually bring users to the site (as opposed to links that may exist but are never visited).
- *Frequent referrers:* Vertical lines (or bands) indicate URLs (or groups of URLs) that may reference multiple pages on the site. In the case of external referrers, these patterns may be used to identify WWW resources with a strong affinity to the material on a given site.
- *Non-link references:* The referrer field is only recorded for HTTP requests that originate when a user clicks on a link found in a web page. Examination of the entries that do not have referrer values may provide insights into the prevalence of users who are reaching the site in question by manually providing a URL to their browser. This may be used to gain some understanding of the extent to which knowledge about the site is propagating via non-WWW mechanisms.
- *Problem Links:* As described above, color-coding based on HTTP status can be used to quickly identify requests that corresponded to problem responses. In particular, referrer/URL combinations that result in the "not found" response can be quickly identified, and this information might be used to locate external pages that may include links to one or more references on the site that do not exist. This information might be used to determine when appropriate redirection may prove useful, or to identify web site operators who might be asked to update their pages.

The use of this visualization for the HCIL web site provided an example of the problems of artifacts in the data that present potential pitfalls in the use of these techniques. Specifically, we observed strong patterns in the visualization, in the form of multiple data points that seemed to form two distinct lines of non-zero slope, cutting across large sections of the URL space (Figure 8).

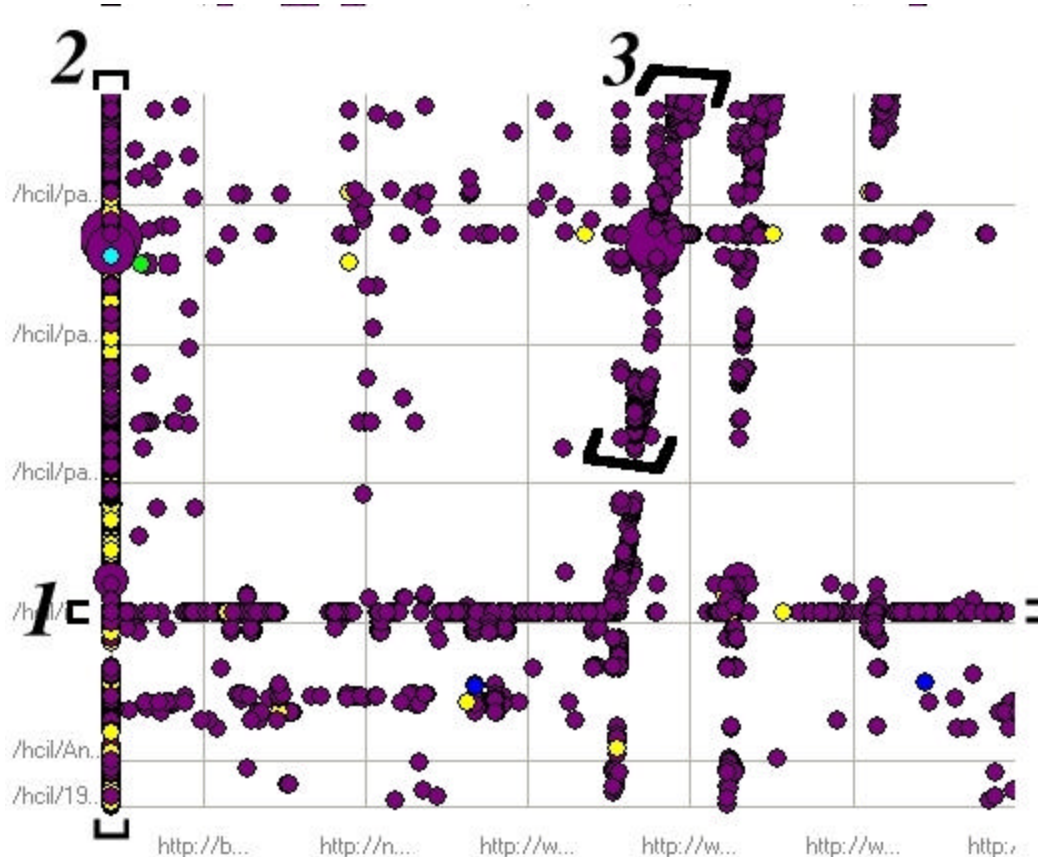


Figure 8: *Referrer vs. URL*: URL on the y-axis, referrer on the x-axis. This visualization illustrates patterns of references from one page to another, with green borders added to emphasize patterns. The row marked by region 1 is a page on the site that has numerous associated referrers. Specifically, this row corresponds to the HCIL home page, which is the target of many external links. The points in this row reveal that external links are often used to refer users to the HCIL home page. Region 2 contains “non-link” references: requests that did not involve named referrers. These requests may have been for URLs that were manually typed into the browser, or they may have originated from browsers that did not report the referring URL correctly. The diagonal pattern of points marked as region 3 shows requests involving a page on the HCIL site that refers to another page on the site that is found in the same directory in the file-system hierarchy, as described in Section 3.7.

While these lines present a striking visual image, the phenomenon being observed is actually quite simple. Like many other web sites, the HCIL pages are arranged hierarchically on a Unix file system, where pages for a given interest area - such as a research project or user home pages - are stored in a single directory. As a result, a page in one of these areas is likely to contain links that refer to other pages in that area: a user's home page might contain links to her CV, and vice-versa. Since the URLs differ only slightly, page requests that move between these pages will generate tight clusters in the visualization.

Furthermore, the presence of areas on a web site with common prefix (i.e., “/Research/1997” and “/Research/1998”) will lead to a juxtaposition of these clusters, thus forming easily visible lines. While this display may provide the impression of a strong pattern of usage and references, the understanding of usage patterns that is gained is actually quite small. Further clarification of the data, either through elimination of intra-site referrers, or through aggregation of referrers by URL domain (as opposed to complete URL path) may eliminate the potential problems caused by this sort of display.

3.8 Other Visualizations

Several other possible visualizations may provide further understanding of site access patterns. Plotting user-agent vs. time, URL, or domain, may prove useful for understanding the software used to access a given web site. This

information might be useful for web site designers interested in deciding which HTML features to use. Site “map” plots containing category identifiers vs. URL would illustrate the layout of the site, in terms of categories occupied by various URLs, providing assistance with ongoing web site maintenance.

Examination of additional data sets can illustrate the generality of the visualization techniques described above. Towards that end, we applied these techniques to log data from the web server of the International Center of Photography (ICP, <http://www.icp.org>). As the web presence for a photographic museum, this site includes thematic content relating to museum exhibits, along with educational material and general information about the museum. Unlike the HCIL web site, the ICP site is composed of several thematic areas with links that encourage sequential browsing of pages. The success of this design was demonstrated by time vs. URL visualizations (similar to those in Figures 2 and 3) that showed clearly defined vertical lines, indicating visits from users who viewed sequences of related pages.

4. DISCUSSION

All of the data trends discussed above might be included - in some form - in the output of a traditional web log analysis tool. However, interactive starfield visualizations offer several advantages (Ahlberg & Shneiderman, 1994) in understanding user visits, including:

- *Rich display of multiple-dimensional data, allowing discovery of multiple trends.* Many of the visualizations described above can potentially reveal several usage patterns in the data. For example, the “Time vs. URL” visualization (Figure 2) illustrates trends including relative request frequency for URLs, changes in request frequencies by day of week, HTTP errors, and potential bottleneck bandwidths caused by frequent requests for large files, all in a single screen.
- *Simultaneous display of large numbers of individual data points.* While traditional analysis tools display bar charts or tables containing dozens of data points, Spotfire can present thousands of data points, each representing an individual request, on a single screen. The visualizations presented in this paper involve display of roughly 25,000 individual points. When combined with advances in hardware and software, appropriate use of aggregations in coordinated visualizations should support significantly larger data sets.
- *Filter and zoom for access to detail.* In generation of aggregate summaries, traditional tools obscure most information about individual events. The visualizations described above allow analysts to move seamlessly from viewing the roughly 25,000 hits in the overview visualizations covering several weeks (Figure 2) to several individual accesses from a single user (Figure 4).
- *Goal-neutral, interactive output.* Existing log-analysis tools provide reports and output that are limited in flexibility and tied directly to the problem domain. As a result, the analyst's ability to expand the range of questions being asked, or to simply “explore” the data, is limited. The lack of domain knowledge in a tool such as Spotfire is in many ways an advantage, as it may avoid over-constraining analysts in their efforts to find meaningful patterns.

These facilities combine to provide an environment that may prove useful for generating hypotheses about web usage patterns that would be difficult to make with traditional tools. For example, the combination of the Time vs. URL and Front Page Visit visualizations was used to identify pages that were entered “through the side door” - pages that had user visits from links that originated outside of the local site. This provides another perspective on the notion of “entry points” (Pirolli, et al., 1996; Cooley, et al., 1999).

Visualizations helped illustrate data artifacts that might have been obscured by the output of traditional packages. For example, some projects described on the HCIL web page have all of their information on a given web page, while others use multiple pages. Using traditional tools, it might appear as if the former projects had more user visits, because these hits would be focused on a small number of pages, instead of being distributed across a larger set. The categorization of web pages as described above helps avoid this problem, and could easily be added to traditional tools. However, the interactive visualization provides analysts with the ability to quickly switch between the categorized and non-categorized views, thus presenting a means of visually identifying a trend that might be obscured in the static layout of a traditional tool.

Effective use of log visualizations will depend upon selection of the appropriate level of granularity. Many of the visualizations described above presented each page access as a distinct point in the starfield visualization. This use of individual points instead of aggregate summaries is a double-edged sword: while visualizations eliminate the data loss that is inherent in summaries, they also mask some of the more basic –and very useful - information provided by traditional tools.

Visualizations involving multiple coordinated displays offer another solution to the problem of selecting the correct granularity. By presenting two or more tightly coupled views at varying levels of granularity, coordinated visualizations provide users with both overview and detail. We have used Snap-Together Visualizations (STV, <http://www.cs.umd.edu/hcil/snap>) (North & Shneiderman, 2000) to explore the possibility of using coordinated displays of web log data. Using STV, aggregate views of log data such as index page link requests (Figure 6) can be displayed alongside a view containing individual data points. Selection of a point for a given URL, day combination in the first view leads to the display of the time and requesting host for each of the hits that are found in the aggregate, thus providing the analyst with the ability to “drill-down” into finer levels of detail.

The use of a general-purpose architecture such as STV might also be useful for generating additional contextual information or other details appropriate for interpreting the log information. For example, STV can be used to create views that coordinate display of log data with external site maps and display of individual web pages, providing context that may aid interpretation of the results. Alternatively, integration with analytic tools might be used to develop quantitative reports to accompany the qualitative visualizations. For example, data points displayed in a starfield might be linked to a spreadsheet containing reports of request aggregates and statistical analyses. Such combinations would combine the support for exploration and investigation inherent in interactive visualizations with quantitative detail comparable to the output of traditional analysis tools.

More generally, interactive visualizations of web log data might work best as complements to, rather than substitutes for, traditional analysis tools. The wide range of reports that can be automatically generated with these tools provides site operators with large amounts of easily understood information which may be very useful for many tasks. A hybrid system might contain these reports with facilities for interactively drilling-down into the aggregate summaries, thus providing a smooth transition between traditional analysis tools and interactive visualizations. This design would have the advantage of providing detailed data on demand and hiding complexity from users who do not need the interactive visualization techniques.

Interactive visualizations and traditional log analysis tools can both be used to gain an understanding of usage patterns that can be used to guide web site design or research. For maximal benefit, this analysis will be done in the context of a clear understanding of the goals of a site: usage patterns from an academic site are likely to be very different from those of an online supermarket. By providing direct access to data from large number of user visits, interactive visualizations provide web site operators with the ability to answer questions such as “which links are being used?”, “when are people visiting the site”, “where are visitors coming from?”, and others. Answers to such questions can be valuable inputs to the process of site and page design.

5. FUTURE WORK

Additional insights may be gained from visualizations covering a longer time range. By extending the above visualizations to cover longer time periods - perhaps 6 months or one year, we might gain an understanding of seasonal usage trends, the impact of site redesign, or other factors that might be missed in a smaller time sample. Unfortunately, such expanded visualizations might exceed the capabilities of the visualization tool: appropriate tools should handle rapid updates for data sets of 10,000 to 100,000 points. For larger data sets, display technologies may not be able to adequately handle the hundreds of thousands or millions of data points that might be involved in visualizing usage patterns for larger sites. Appropriate use of aggregation and coordinated visualizations might be particularly helpful for management of larger data sets.

The utility of web log visualizations is also limited by the available data that can be manipulated, and by the types of manipulations that can be done. Inclusion of additional data, along with tools to manage that data, may increase the expressive power of these visualizations.

Specifically, visualizations that combine web log data with other data may help users place data in the appropriate contexts. The most basic external data sources include additional log files, tracking errors, cookies, or other web server output. Visualizations that combine web log data with site "maps" might improve the utility of visualizations that approximate user sessions. For sites aimed at accomplishing specific goals, data relevant to those goals might provide further utility. For example, visualizations of log data for electronic commerce sites might be enhanced through inclusion of relevant marketing data (Büchner & Mulvenna, 1998).

Further improvements might be made through the addition of data modeling tools to the visualization environment. Potentially useful additions to the visualization environment include:

- *Improved aggregation facilities:* facilities for generating "on-the-fly" aggregations of data may prove useful for identifying trends. Fully general aggregation facilities could be used to generate aggregations that would go beyond those provided by traditional tools.
- *Generalized handling of hierarchical data:* Log data has several attributes that are hierarchical in structure: URL file names, timestamps, and client host names. Facilities to easily move through views at different levels of the hierarchy, in combination with improved aggregation tools, would simplify the process of building models. For example, users would be able to move from display of all hits in a given month, to aggregate counts by hour, day, or week.
- *Increased support for coordinated visualizations:* Development of coordinated visualizations using STV currently involves manual creation of appropriate SQL queries and is limited to a small set of "snappable" visualization tools. Appropriately designed tools could support the use of coordinated visualizations while expanding the range of visualization tools that could be used.
- *Integration with site data:* Understanding of user behavior through analysis of user paths might be facilitated by tools that would incorporate display of site topology alongside log visualizations. Such tools might provide analysts with the ability to "step through" paths followed by site visitors. Alternatively, log data might be used to generate annotated versions of web pages, indicating the percentage of visitors to a page that followed each link on the page.
- *Integration with external data sources:* Web site operators may have any number of additional databases that could be coordinated with web log data, including lists of registered users, purchase data, and other relevant sources. Just as products like HitList (Accrue, 2000) are providing tools for integrating these data sources into static log analyses, interactive visualizations might provide similar access to external data sources.
- *Integration with traditional tools:* As described above, a system that combined the summaries generated by traditional analysis tools with the access to individual request data provided by the Spotfire visualizations described in this paper might provide the best combination of ease of access and expressive power.

The large space of possible visualizations of log data presents a challenge for effective use of these tools: further exploration of these possibilities might lead to identification of an "optimal" set of visualizations. Data mining techniques such as cluster detection, outlier detection, and correlation analyses might also be used to identify regions of potential interest. Ideally, these and other techniques for identifying interesting data subsets or views would provide necessary understanding with minimal effort.

Some of the suggestions provided above may be possible within current versions of Spotfire. Specifically, histograms and pie charts might be used to provide some of the aggregate data currently provided by traditional tools. However, Spotfire's generality may be a liability for user's interested in this specific domain. A special-purpose starfield visualization tool specifically tailored for web log analysis might use a restricted set of functionality to provide users with easier access to commonly-requested visualizations and aggregations. This tool might use knowledge of the information domain to provide additional desirable features, such as support for easily moving between different levels of hierarchical aggregation by hostname and /or request date. Another area for enhancement might be in the identification of user paths: research models aimed at identifying user paths (Pirolli, et al., 1996; Cooley, et al., 1999; Pitkow, 1997) might be used to augment the paths of individual users through a site.

Visualization of web log data might also prove useful for real-time monitoring, as opposed to retrospective analysis. By feeding the real-time request data directly into the visualization tool, this monitoring tool would provide network engineers with a real-time view of the ongoing activity of a web server. This information might be valuable for identifying network problems and ensuring availability.

Despite the efforts of several research projects (Pitkow 1996; Cooley, et al., 1999), modeling of web usage remains an inexact science (Monticino, 1998). Interactive visualizations of web log data may be useful complements to static reports generated by current tools and session models currently being developed. These visualizations might also work well alongside data mining efforts aimed at understanding of customer records and other non-web data.

Finally, no matter how rich or accurate the log data, answers to many questions may require coordinated observations or interviews with users. For example, a long visit to many pages on a site may indicate satisfaction and interest in the contents, or confusion and frustration due to an unsuccessful search for information. While visualizations of the log data may expose patterns that provide some insights into the user's experience, the characterizations of user behaviors provided by these patterns will be at best indirect, and may require interviews for clarification.

ACKNOWLEDGEMENTS

This research was supported by a grant from IBM's University Partnership Program. Thanks to Anne Rose for help with generation of the visualizations, Edward Earle from for his help with the ICP logs, and Chris North for his assistance with Snap-Together Visualizations. We wish to thank JASIS editor Don Kraft for inviting us to expand and update our conference paper from ASIS 1999. The anonymous reviewers provided several helpful comments.

BIBLIOGRAPHY

- Abrams, M., Williams, S., Abdulla, G., Patel, S. Ribler, R., & Fox, E. (1995) Multimedia traffic analysis using CHITRA95. Proceedings of the third Annual Conference on Multimedia (ACM Multimedia '95), 267-276.
- Accrue (2000). HitList overview [Online] Available at <http://www.acrue.com/products/hitlist.html>. (Accessed June 16, 2000).
- Ahlberg, C., & Shneiderman, B. (1994) Visual information seeking: tight coupling of dynamic query filters with starfield displays. Conference Proceedings on Human Factors in Computing Systems. (ACM CHI '94), 313-317.
- Boutell, T. (2000) Wusage Home Page [Online] Available at <http://www.boutell.com/wusage/> (Accessed June 16, 2000).
- Büchner, A. & Mulvenna, M. D. (1998) Discovering internet marketing intelligence through online analytical web usage mining. ACM SIGMOD 27(4), December 1998, 54-61.
- Chi, E., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., & Card., S. (1998). Visualizing the evolution of web ecologies. Conference Proceedings on Human Factors in Computing Systems (ACM CHI '98), 400-407.
- Chimera, R., & Shneiderman, B. (1994) An exploratory evaluation of three interfaces for browsing large hierarchical tables of contents. ACM Transactions on Information Systems 12(4) October 1994, 383-406.
- Cooley, R. Mobasher, B., & Srivastava, J.(1999). Data preparation for mining world wide web browsing patterns. Journal of Knowledge and Information Systems 1(1).
- Cugini, J. & Scholtz, J. (1999). VISVIP: 3D Visualization of Paths through Web Sites. Proceedings of the International Workshop on Web-Based Information Visualization (WebVis '99), in conjunction with DEXA '99 Tenth International Workshop on Database and Expert Systems Applications, 259-263.
- Eick, S., Nelson, M., & Schmidt, J. (1994). Graphical Analysis of Computer Log Files. Communications of the ACM 37(12), December, 1994, 50-56.
- Fielding R. (1998). Wwwstat: httpd logfile analysis software [Online] Available at <http://www.ics.uci.edu/pub/websoft/wwwstat/> (Accessed June 16, 2000).
- Fielding, R., Gettys, J. Mogul, J. Frystyk, H. & Berners-Lee, T. (1997) RFC 2068: Hypertext transfer protocol – http/1.1. [Online] Available at <http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2068.txt> (Accessed June 16, 2000).
- Gomory S., Hoch, R., Lee, J., Podlaseck, M., & Schonberg, E. (1999) Analysis and Visualization of Metrics for Online Merchandising. Conference Proceedings: WebKDD '99 Workshop on Web Usage Analysis and User Profiling. [online] Available at <http://www.acm.org/sigkdd/proceedings/webkdd99/papers/paper8-jylee.ps> (Accessed June 29, 2000).
- Monticino, M. (1998) Web-analysis: stripping away the hype. IEEE Computer 31(12), December, 1998, 130-132.

- Nielsen, J. (1999) Good content bubbles to the top (Alertbox Oct. 1999). [Online] Available at <http://www.useit.com/alertbox/991017.html> (Accessed June 16, 2000).
- North, C. & Shneiderman, B (2000). Snap-together visualization: A user interface for coordinating visualization for relational schemata. Conference Proceedings of ACM Advanced Visual Interfaces 2000.
- Papadakakis, N., Markatos, E. P., & Papathanasiou A.E. (1998) Palantir: A Visualization Tool for the world wide web. Proceedings INET '98 Conference [Online] Available at http://www.isoc.org/inet98/proceedings/1e/1e_1.htm (Accessed June 29, 2000).
- Pirolli, P., Pitkow, J., & Rao, (1996). R. Silk from a sow's ear: extracting usable structures from the Web. Conference Proceedings on Human Factors in Computing Systems (ACM CHI' 96), pp. 118-125.
- Pitkow, J. (1996) In search of reliable usage data on the WWW. Technical Report 97-13, Georgia Tech, College of Computing, Graphics, Visualization, and Usability Center [online] Available at <ftp://ftp.gvu.gatech.edu/pub/gvu/tr/1997/97-13.pdf> (Accessed June 16, 2000).
- Pitkow, J. & Bharat, K.(1994) Webviz: A tool for world wide web access log analysis. Proceedings of First International Conference on the World Wide Web.[Online] Available at <http://www1.cern.ch/PapersWWW94/pitkow-webvis.ps> (Accessed June 16, 2000).
- Shneiderman, B., Shafer, P., Simon, R., & Weldon, L. (1986) Display strategies for program browsing: concepts and an experiment. IEEE Software 3 (3), March 1986, 7-15.
- Spotfire. (1999). Spotfire [Online] Available at <http://www.spotfire.com> (Accessed June 16, 2000).
- Tauscher, L, & Greenberg, S. (1986) Revisitation patterns in world wide web navigation. Conference Proceedings on Human Factors in Computing Systems (ACM CHI '97), 399-406.
- Turner, S. (2000). Analog: WWW logfile analysis [Online] Available at: <http://www.statslab.cam.ac.uk/~sret1/analog/> (Accessed June 16, 2000).
- Uppsala University, IT Support (1999) Access log analyzers [Online] Available at: <http://www.uu.se/Software/Analyzers/Access-analyzers.html> (Accessed June 16, 2000).
- WebMetrics Development Team (2000). WebMetrics Tools [Online] Available at: <http://zing.ncsl.nist.gov/webmet/> (Accessed June 27, 2000).
- WebTrends (2000). Enterprise Reporting Server, Web Trends Corporation, Portland, Oregon [Online]. Available at <http://www.webtrends.com/products/ers/Erlog.htm> (Accessed June 16, 2000).