TECHNICAL RESEARCH REPORT

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

by Harry Hochheiser, Ben Shneiderman

T.R. 99-70

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. Sample visualizations of server data from two web sites are presented. Coordinated, snap-together visualizations (STVs) of log data are introduced as a means of gaining additional expressive power. Possible uses of these visualizations are discussed, and difficulties of data collection, presentation, and interpretation are explored.

Keywords
World Wide Web, Log File Analysis, Information Visualization

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, 1999) 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 snap-together visualizations (STVs) (North & Shneiderman, 1999) 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

Products, such as wwwstat (Fielding, 1998), analog (Turner, 1999), HitList (Accrue, 1999), and Wusage (Boutell, 1998) parse log files in order to produce aggregate reports, such as ``transfers by request date'', ``transfers by URL/archive'', most popular pages, visits by time of day or day of week, originating regions, user agent, or other criteria. While these packages focus on aggregate statistics, some provide user level information, such as ``example visits'' or ``document trails'', which describe paths that have been taken through the site. Output reports are generally presented as tables, histograms, or pie charts.

While these packages differ in the specific reports available, they generally share several characteristics:

- *Static display*: Reports are generally presented on an HTML page, without interactive facilities.
- *Low-dimensionality of reports}*: Results are presented in a series of low-dimensional reports, which must be scanned sequentially.
- *Lack of low-level details*: Reports focus on aggregations, with minimal (if any) support for direct examination of records relating to individual page requests.
- *Relative lack of flexibility* : while most tools include some configuration functionality to allow for selection of reports to be generated, customization facilities are often quite limited.
- *Lack of integration of knowledge of site layout* : reports are presented in terms of visits by URL, without any further information regarding the site layout.

More recently, products such as BazaarAnalyzer (Aquas, 1999) have augmented these facilities with additional reports tracking entry and exits sites and visual displays of user paths through a site.

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.

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. This paper presents a series of interactive visualizations that might be used to augment these data models.

3. STARFIELD VISUALIZATIONS

Starfield visualization tools (Ahlberg & Shneiderman, 1994) such as Spotfire (Spotfire, 1999) combine simultaneous display of large numbers of individual data points with a tightly-coupled interface that provides facilities for zooming, filtering, and dynamic querying (Figure 1). 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.

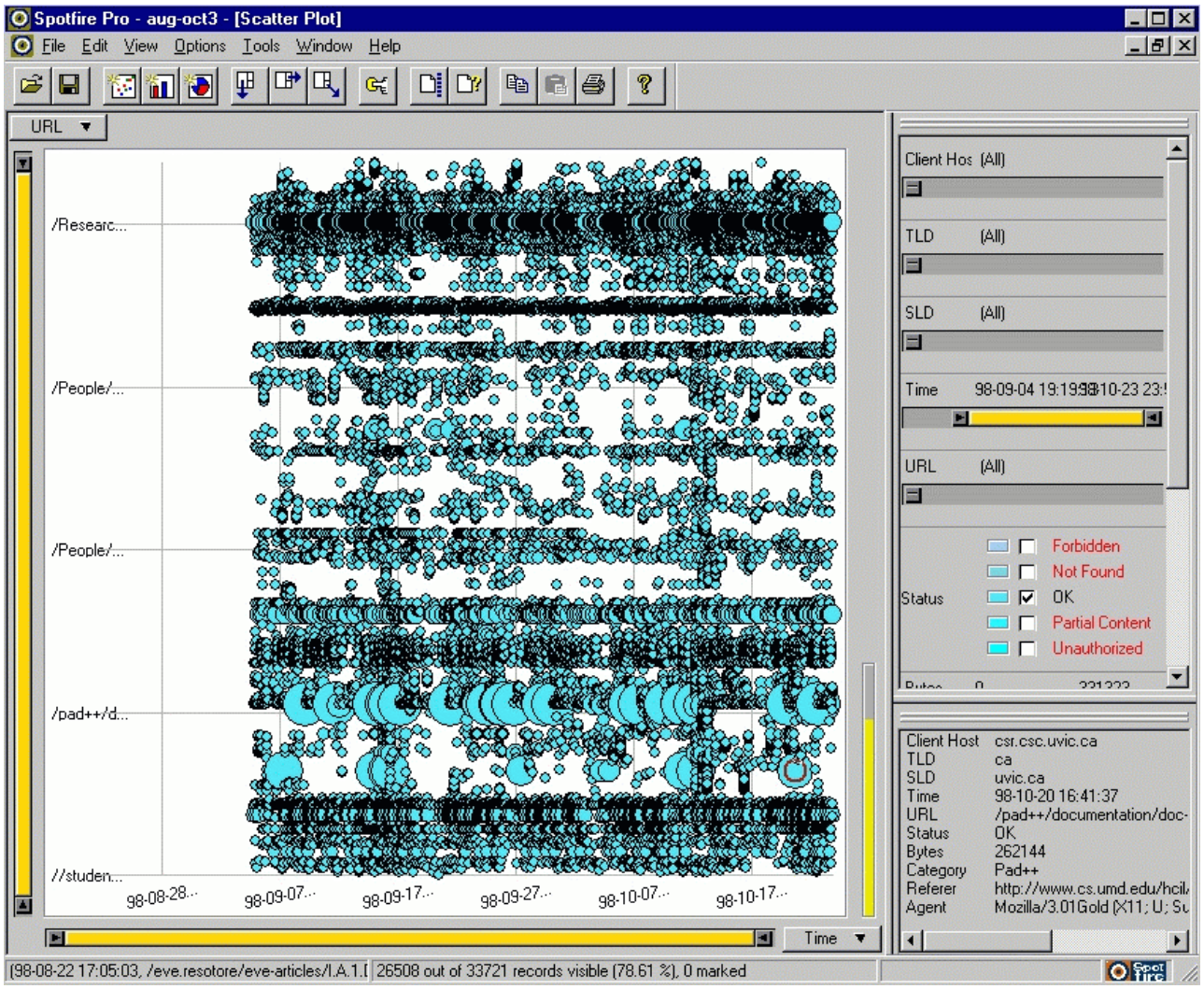

Figure 1: *Interactive Visualizations in Spotfire*: A Spotfire visualization, with the URL requested on the y-axis and the time of request on the x-axis. Checkboxes on the right-hand side have been selected to display only data points with ``OK'' status, and a slider labeled ``Time'' has been adjusted to eliminate points corresponding to requests that occurred before September 4, 1998. Detailed information about a selected point is displayed in the lower right-hand corner.

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 ``hcil'' 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.

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*

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:

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

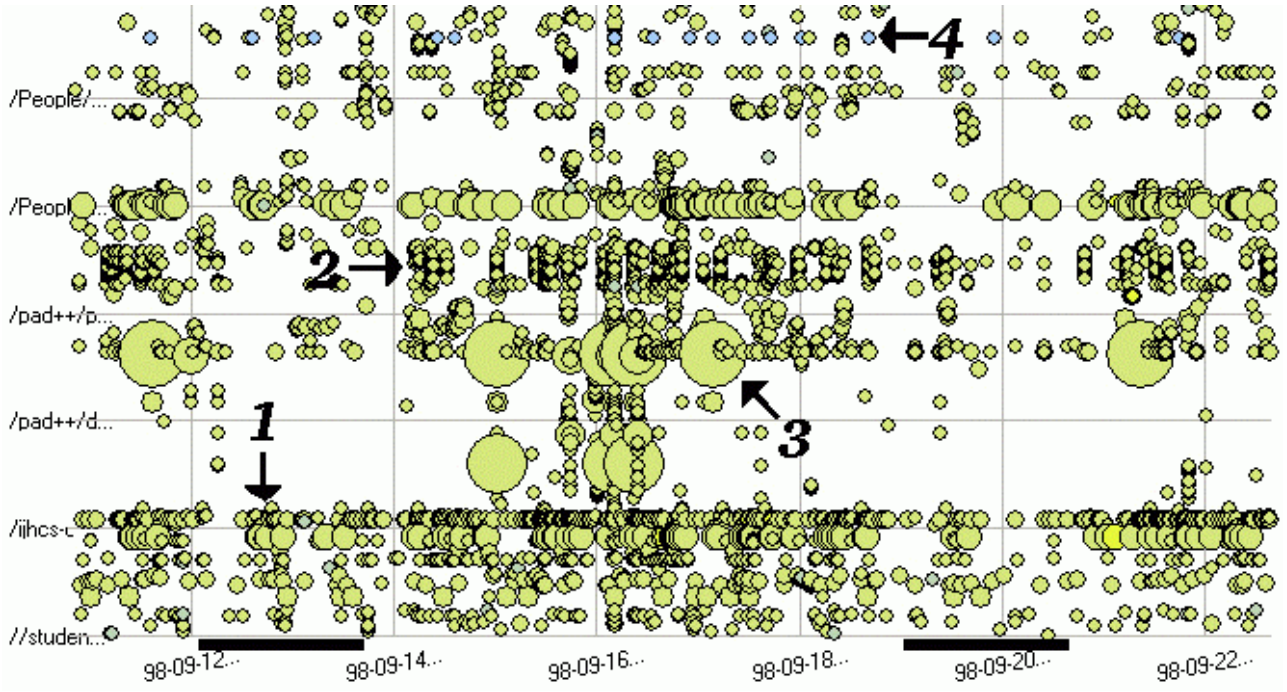

Figure 2: *Time vs. URL, Macro View*: Two weeks of accesses to a subset of the HCIL pages. The requested URL is on the y-axis, with the date and time on the x-axis. The dark lines on the x-axis correspond to weekends. Each circle represents a request for a single page. The size of the circle indicates 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. Numbered arrows point to examples of interesting patterns that can be seen in the visualization: 1) 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. The gap to the left of the arrow shows a slight dip in access frequency, due to the weekend of September 12-13, 1998. 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) Color coding for HTTP status codes allows for quick identification of errors: the straight line of error code here indicates a non-existent URL that is frequently requested – perhaps from an outdated link on an external page.

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 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).

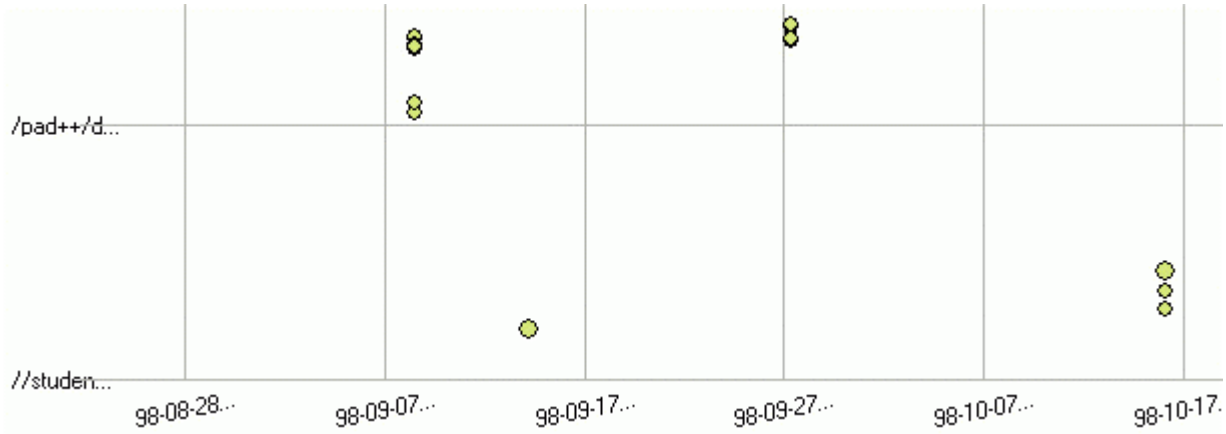

Figure 3: *Time vs. URL, Micro View*: A series of requests from a single client. Over the course of five weeks, this client made several series of requests to the HCIL web site: 4 pages on September 8, one on September 14, 3 on September 27, and 4 (of which three are shown) on October 16. URLs are alphabetized 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 the x-axis and fully-qualified-domain-name (or IP number, if the complete domain name is unavailable) on the y-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.

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 4). This information may be useful for site operators interested in knowing when an automated agent is visiting their site.

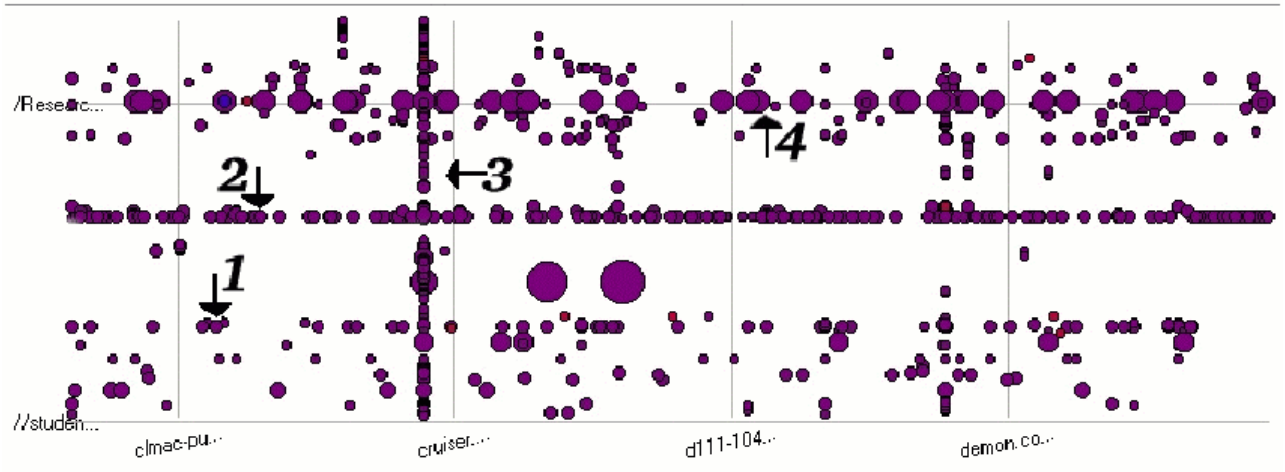

Figure 4: *Client Host vs. URL*: Client hosts on the X-axis, and requested URL on the y-axis. Vertical slices display files visited by each host, while horizontal slices indicate the patterns of requests for a given web page. This display indicates that the HCIL index page (arrow labeled 1) is visited by many of the hosts that come to the site, but the percentage of visitors that visit the page for the lab's Visible Human project (2) or technical reports (4) appears to be higher. The vertical line indicated by arrow 3 is a visit from a web robot.

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 index page (Figure 5). 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 it's 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 5: *HCIL Homepage*: A section of the HCIL homepage, with links for the BLC project and the technical reports indicated by arrows labeled 1 and 2, respectively.

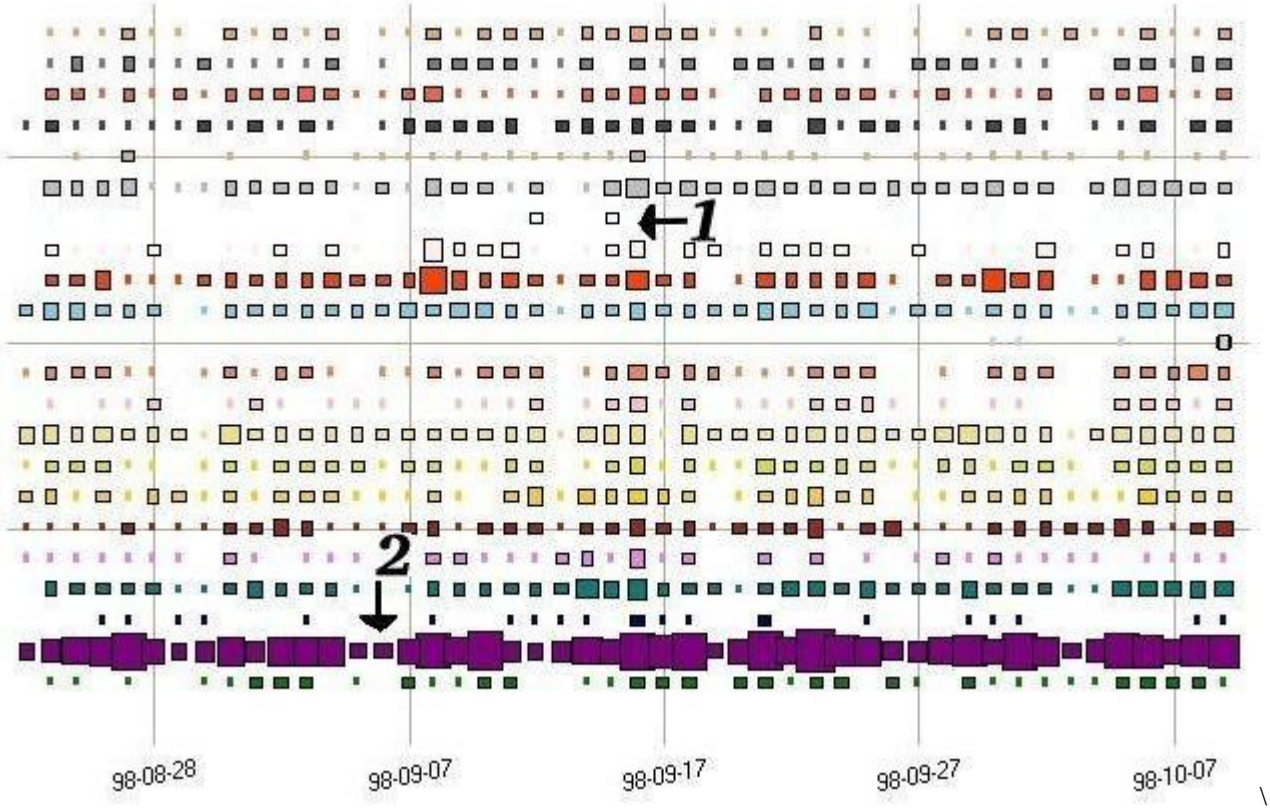

Figure 6: *Index Page Link Requests*: Requests for pages that have links on the group index page. Each row corresponds to a link on the index page. The vertical position of each row in the visualization corresponds to the vertical position of

the link on the index page (Figure 5), 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. From this visualization, we can see that some links placed fairly high on the page are not referenced very frequently (arrow 1, indicating links to the pages for HCIL's Baltimore Learning Community project), while frequently requested links such as HCIL's technical report page are placed further down the page (arrow 2). This information might be used to redesign the index page, by identifying frequently requested links that might be moved to more prominent positions.}

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).

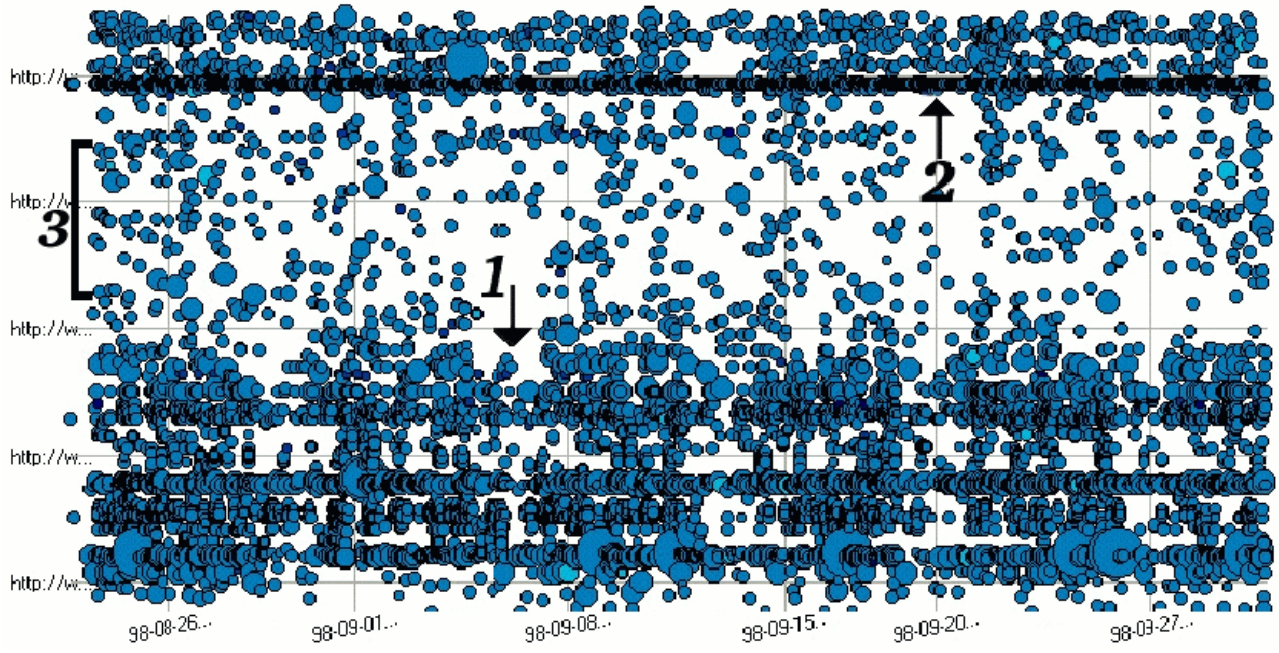

Figure 7: *Referrer vs. Time*: The URL of the referring page is given on the y-axis, and the request date and time is on the x-axis. Referrer patterns seem to be fairly regular across the 5 weeks of data displayed. The points shown below arrow 1 are those that have referring URLs that are inside the HCIL web site, indicating visitors who went from one

page to another within the site. As expected, this class makes up a significant portion of the data points. The line marked by arrow 2 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 vertical brace on the left (labeled 3) indicates a band of referrer URLs corresponding to a search engine.

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-referer (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).

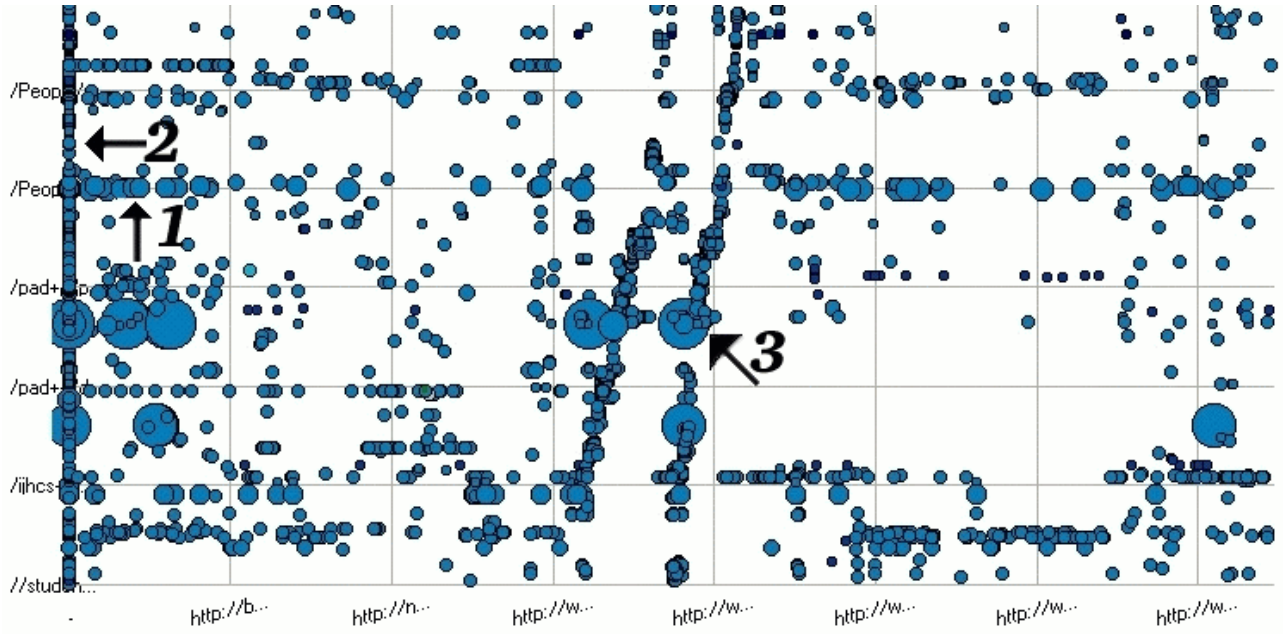

Figure 8: *Referrer vs. URL*: URL on the y-axis, referrer on the x-axis. From this visualization, we can see that some URLs have numerous associated referrers (see arrow 1), indicating either multiple links to the URL or frequent searches that lead users to the URL. Arrow 2 points to the ``non-link'' references: requests that did not involve named referrers. Arrow 3 illustrates the patterns formed when users follow paths through the site, 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. Possibilities include:

- *User Agent*: 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 "mapping"*: plots containing category identifiers vs. URL illustrate the layout of the site, in terms of categories occupied by various URLs.

4. ICP LOG DATA

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.

As one might expect, the ICP's site is heavily focused on photographic images, in the form of JPEG files. As these files are a crucial part of the site's content, inclusion of requests for image files might help illustrate viewing patterns of the various photographic resources on the site. Towards that end, the pre-processing of the raw log data was modified to include all GIF or JPEG files in the visualization input data. Unfortunately, this additional power does not come without a cost: as images are generally requested through HTML tags embedded in web pages, requests for images do not necessarily correspond to user actions. As a result, the one-to-one correspondence between user actions and visualization data points does not hold for this data set.

The mix of requests for HTML pages, GIF images, and JPEG images can be displayed through the use of color coding for document file name extensions. As in previous examples, data points were size coded according to the number of bytes requested. This information might be particularly valuable if the ICP site had unusually large image files that were frequently accessed.

The resulting visualization contains roughly 25,000 requests spanning a period of 58 hours during January, 1999. An overview of the requests to the ICP site during this period is given in Figure 9. Like the ``Time vs. URL, Macro View" display of the HCIL data (Section 3.1), this visualization shows request times on the x-axis and URLs (ordered alphabetically) on the y-axis. This display clearly shows an overall usage pattern that is very different from that of the HCIL web site (Figure 2).

Relative to the HCIL data, the ICP site appears to have a significantly higher percentage of requests that form vertical lines. As these lines indicate requests that are close together in time, these patterns might provide indications of clustered requests from individual visitors as they browse through the site. In fact, closer examination of individual requests and requests by individual hosts revealed a clear pattern: many visitors to the ICP site seem to visit sequences of several pages. This is qualitatively different from observed patterns for the HCIL data, which appeared to have a higher percentage of users who visited only one or two pages.

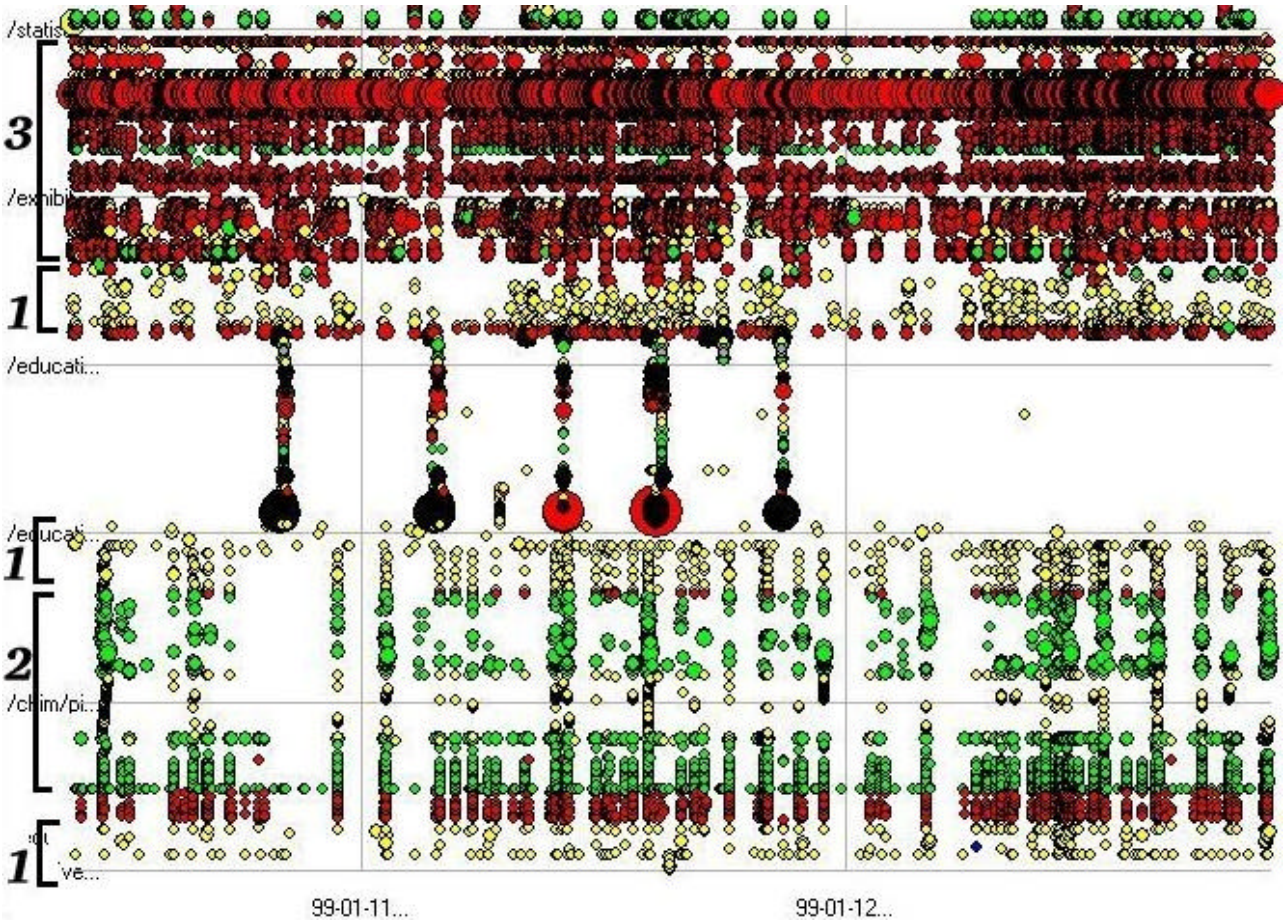

Figure 9: *ICP Data, Time vs. URL*: An overview of the ICP log data, with request time displayed on the x-axis and URL on the y-axis. File types are color-coded, as indicated by the vertical braces on the left: 1) HTML pages, 2) JPEG images, 3) GIF images, with requests of each type distributed throughout the display due to alphabetical ordering of URLs. Requests that form vertical lines indicate user visits consisting of requests for multiple pages within the site.

The difference in access patterns between the HCIL data and the ICP data might be explained in terms of the site content. Where the HCIL site is composed of large numbers of individual project pages with relatively few cross-links, the ICP site is composed of several thematic areas with links that encourage sequential browsing of pages. Clearly defined vertical lines in the visualization indicate visits from users who viewed a sequence of related pages in an exhibit or biography (Figure 10).

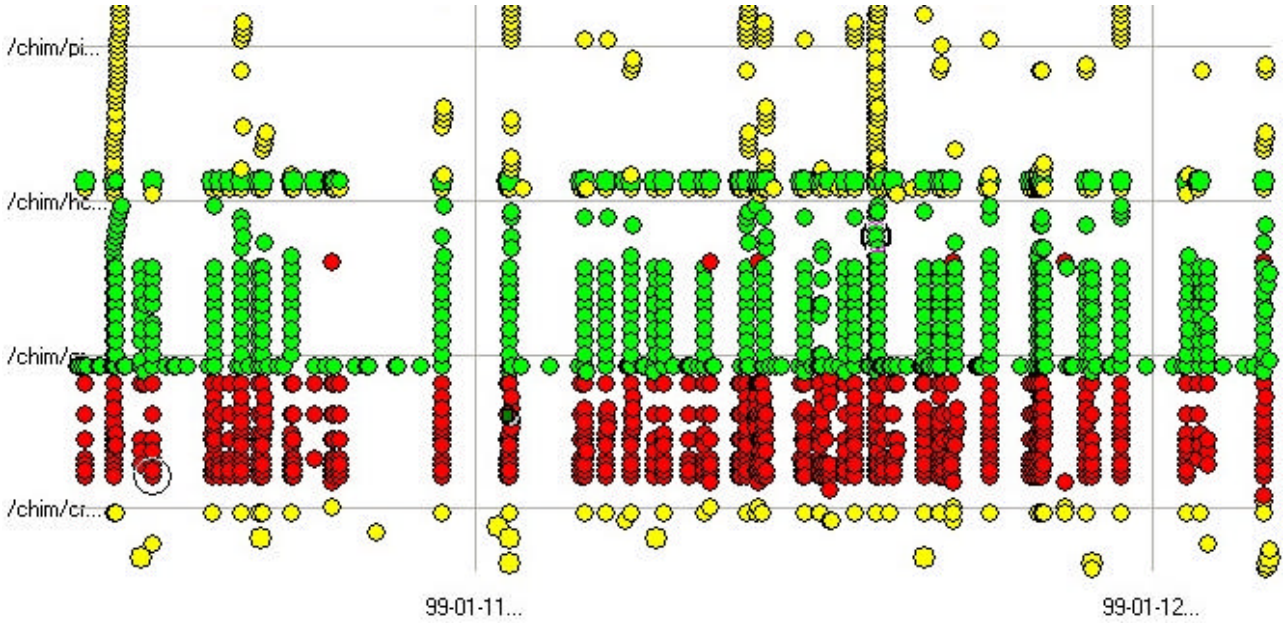

Figure 10: *Visits to Thematic Sections of a Site*: Examination of visits to URLs for the ICP's pages on David Seymour (http://www.icp.org/chim) reveal strong vertical lines, indicating users who accessed multiple pages from that section of the ICP site.

The use of color to coding for the file extension provides potentially interesting information regarding the relative frequency of request for image files as compared to HTML files. For the current data set, examination of these details led to the observation that requests for GIF images were more frequent than requests for JPEG files. Although this initially seemed surprising for a site so heavily focused on photographic images (which are almost exclusively stored as JPEGs), we quickly realized that the high frequency of requests for GIF files was due to the site's frequent use of GIF images for other visual effects. This frequency, taken together with the presence of the vertical lines indicating extended visits from individual users, might be used to conclude that visitors were viewing the JPEG photo files "in-context" - as they are found embedded in pages on the HTML site. This might provide reassurance to operators of sites such as ICP's, who might be concerned about external pages that might link to the ICP images but not to the ICP pages. Of course, further exploration of referrer data might be necessary to identify instances where users followed links on external pages to retrieve images from ICP pages.

5. MULTIPLE COORDINATED VISUALIZATIONS

Our work with these visualizations led to the identification of two areas of support that would significantly increase the utility of the starfield visualizations of log data. This section introduces these issues, and describes the use of multiple coordinated visualizations to provide additional expressive power.

5.1 Inter-Visualization Visualizations

The visualizations described above present a variety of perspectives on the log data. While some patterns can be inferred from examination of a single display, identification of other trends might require examination of several visualizations. For example, investigation of the index page link request visualization (Section 3.5), identified a link that was not heavily used, despite the fact that the corresponding area on the site had a high level of traffic. We hypothesized that user's were arriving through an external link to another page in that section of the site. Although examination of the referrer vs. time and referrer vs. URL visualizations eventually verified this, manipulations of the two visualizations were cumbersome and uncoordinated.

5.2 Aggregations

The visualizations described thus far (with the exception of Index Page Links) generally present each request as an individual point, leading to densely populated displays that can be used – in combination with Spotfire's dynamic query

tools - to infer patterns. This approach is fundamentally limited, as it does not account for aggregate counts that many site operators find useful. Visualizations of total number of hits by URLs, or hits counted by time of day, increase the expressive power of the visualizations. The Index Page Link Requests visualization described in Section 3.5 provides one example of an aggregate view, but the ad-hoc reprocessing of the data required to generate this display does not scale to other interesting visualizations.

A more general approach to visualizing aggregate data would be to harness the power of a relational database system. SQL queries could express interesting aggregations without requiring special-purpose programming. Results of these queries could then be loaded into Spotfire for visualization.

Ideally, views based on aggregations would support moving between different levels of detail. For example, an initial display in one window might present requests aggregated by day of the week. Selection of a point in that visualization might display individual hits for that URL and day of week broken down by request time, along with a third window that would display the URL in question.

5.3 Snap-Together Visualizations of Log Data

Coordinated, tightly coupled displays have been shown useful in a number of domains (Chimera & Shneiderman, 1994, Shneiderman, Shafer, Simon, & Weldon, 1986, North & Shneiderman, 1999). Application of these techniques to log visualizations could address both the "inter-visualization visualization" and the aggregation problems. Specifically, views that presented tightly coupled displays of the log data would support the use of multiple displays for inferring usage patterns. Since the displays could be at different levels of granularity, we can build coordinated visualizations between aggregated and disaggregated data.

We used Snap-Together Visualization (North & Shneiderman, 1999) to explore the applications of coordinated visualizations to web log visualization. STV is an architecture that allows users to connect visualization tools such that actions of selections, navigation, and querying are coordinated. Furthermore, STV supports several visualization tools (including Spotfire), raising the possibility of coordinating starfield visualizations with tables, outline views, and web browsers. As STV uses Microsoft's ODBC tools for database connectivity, data preparation is straightforward: we import the data into Microsoft Access, and design appropriate database queries, using SQL or Access' visual query editor.

A simple example of the use of STV's to coordinate web log visualization involves coordination between an outline view of site URLs, a browser window displaying a page from the site, and a Spotfire window displaying requests to a URL by time (x-axis) and hostname (y-axis) (Figure 11). When the user selects a URL in the outline view, data for that page is displayed in the Spotfire window, while the page itself is loaded into the browser window. The scrollable overview of the available URLs provided by the outline view can be useful for understanding the range of available URLs, and the browser display is useful for relating the requests in the log data to the actual site content. Tight coupling between the three windows insures that all displays are constantly coordinated. Although this display only includes one view of the log data, the overview and contextual information provided by the other windows may simplify the process of understanding patterns in the data.

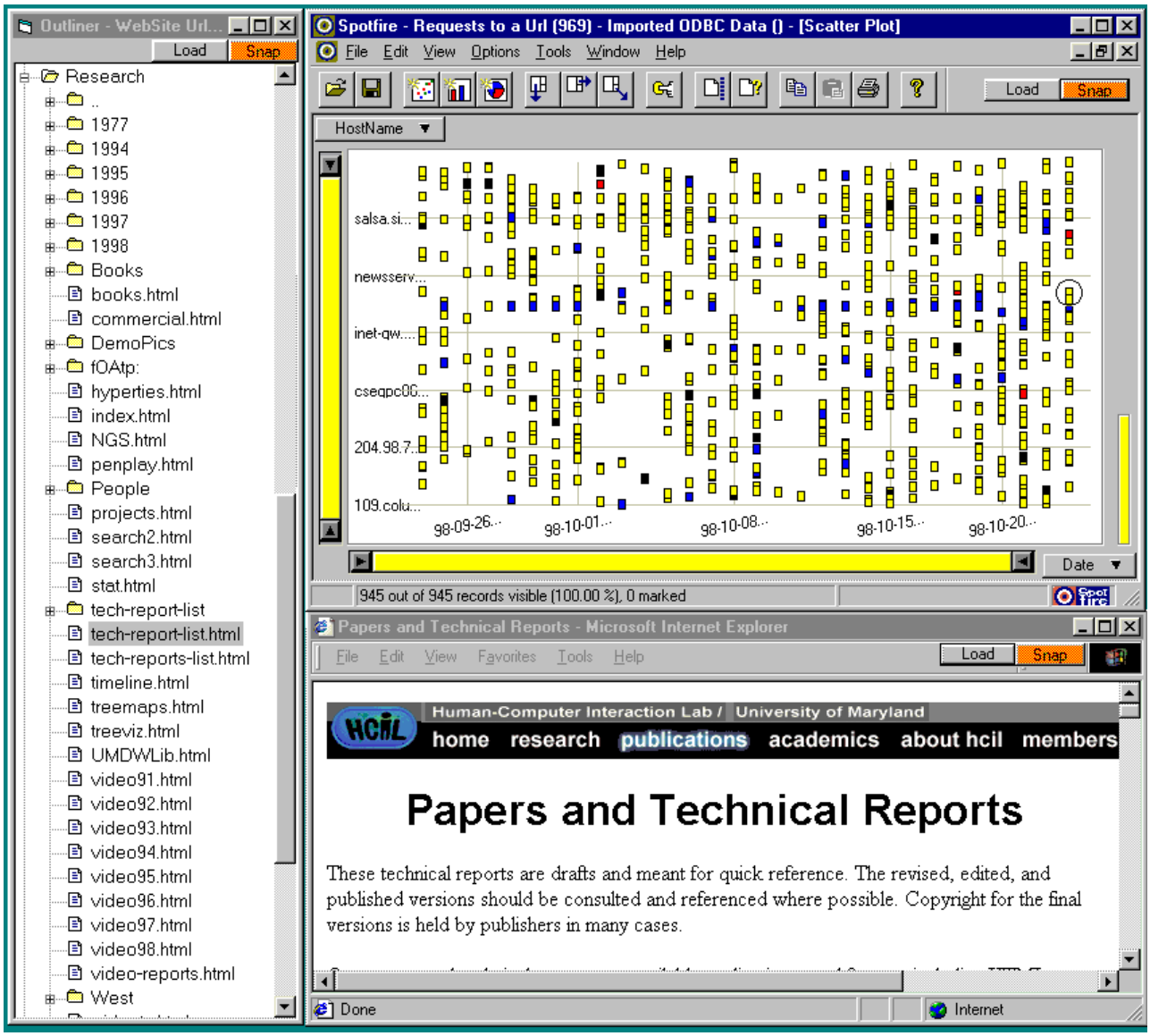

Figure 12: *Simple STV Example*: This display shows three coordinated windows. The outline window on the left-hand side provides a hierarchical view of URLs on the site. The web browser window in the lower right corner displays a selected web page. Finally, the Spotfire display plots requests for a given URL, with time on the x-axis and hostname on the y-axis. Selection of a URL in the outline window leads to updates of both the Spotfire and web browser displays.

Coordinated visualizations based on aggregations of data points can provide additional expressive power and flexibility. By snapping a view of aggregations to a second window containing individual data points, users can move quickly between overview and detail analysis. Selection of an aggregate in the first visualization will lead to the display of the component requests in the second display, thus allowing users to ``drill-down'' to finer levels of detail.

An example of this technique is shown in Figure 12. The aggregate display shows totals of the number of hits to a given URL (y-axis) on a given date (x-axis). Size coding displays the number of hits, so the larger circles indicate higher number of hits for the given URL on the given day. This visualization might be used to determine which pages are accessed most frequently, or how usage varies across dates (or days of the week). The components of each aggregation can be displayed in a second visualization, which might present time on the x-axis and hostname of the requesting computer on the y-axis, presenting each request for a given URL on a given day. Coordination between the

visualizations provides for easy navigation between the views: when an aggregate point in the first visualization is selected, appropriate data for the components of that aggregation is displayed in the second visualization window.

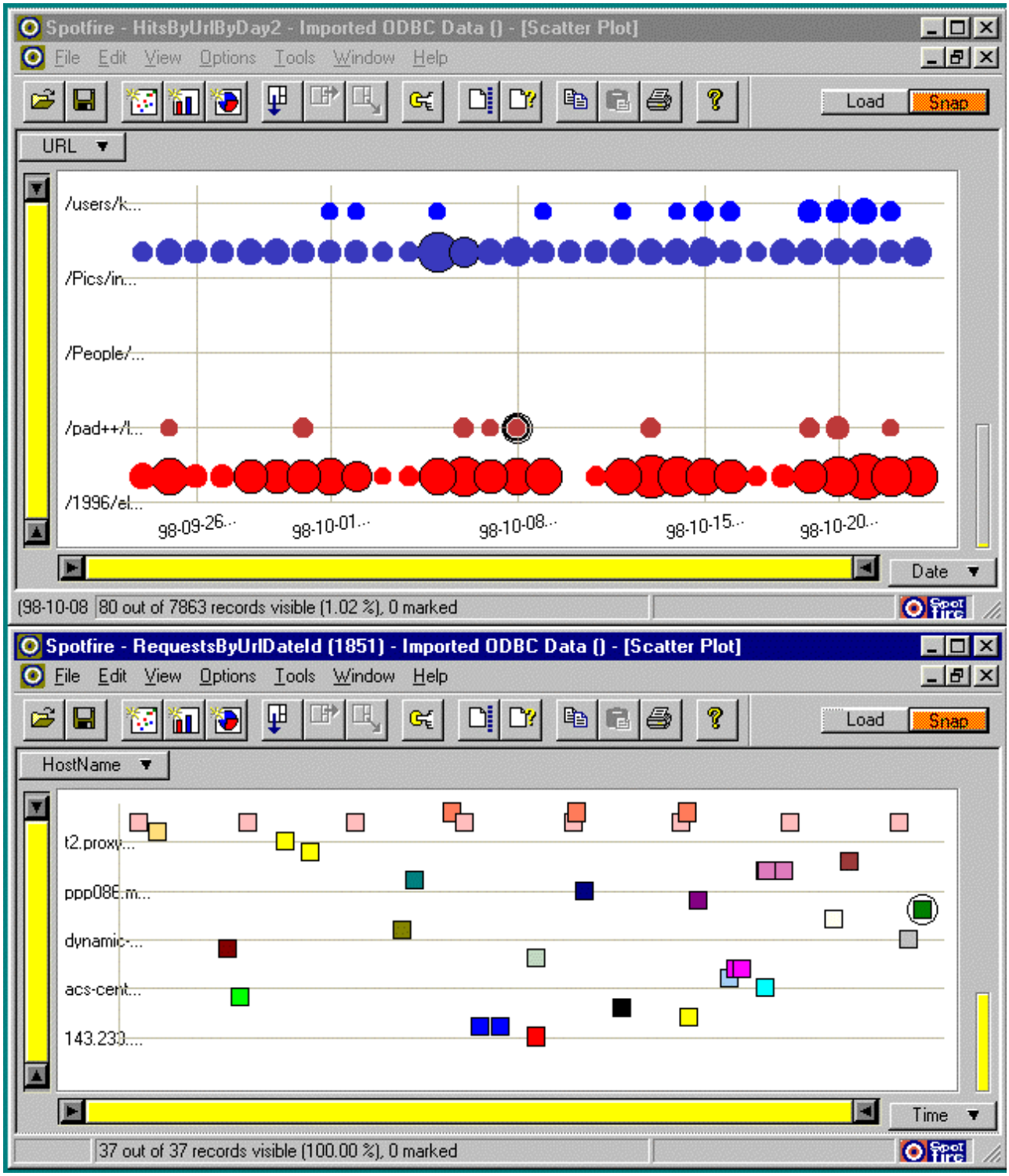

Figure 12: *Coordinated Visualizations and Aggregation*: Two coordinated displays support visualization of aggregated and disaggregated data points. The upper window displays URLs on the x-axis and times on the y-axis, with points size-coded to display the number of requests for each URL on each day, and filters set to remove data points containing fewer than 37 requests. Selection of a data point in this view leads to display of the individual request records in the

second visualization window, with requesting hostname on the y-axis and time of request on the x-axis. Currently, a data point for the ``/pad++" URL on October 8 is selected in the upper window, and the 37 requests to that URL for that date are displayed below.

This scenario is just one application of coordinated visualizations to web log data. As STV provides a general-purpose, database-driven, platform for visualization coordination, log data might be visualized alongside other relevant organizational data. For example, operators of e-commerce sites might construct coordinated visualizations that relate web log access patterns to customer purchase records. Furthermore, the flexibility provided by STV's use of a relational database provides the possibility of visualizing the results of arbitrary aggregations through SQL queries, side-by-side with "snapped" visualizations providing context and drill-down capabilities.

6. 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 3).

- *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; Coolet, 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 information provided by traditional tools. In some cases, interactive visualizations might work best as complements to traditional analysis 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. The use of a general-purpose architecture such as STV might also be useful for generating quantitative reports to accompany the qualitative visualizations. For example, data points displayed in a starfield might be linked to a spreadsheet containing reports of aggregate sizes 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.

Ideally, web log analysis will lead to an understanding of usage patterns that can be used to guide web site design or research, in order to effectively realize the goals of the site. 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.

7. 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: the performance of the current version of Spotfire (3.2) degrades noticeably on data sets containing more than twenty thousand points. While improved software and processing hardware should help, 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. Spotfire is primarily a data visualization tool: facilities for data modeling are limited. 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.

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.

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.

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) (pp.267-276).

Accrue (1999). HitList overview [Online] Available at http://www.accrue.com/products/hitlist.html. (Accessed November 11, 1999).

Aquas (1999). Aquas home page [Online] Available at http://www.bazaarsuite.com (Accessed November 16, 1999).

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) (pp. 313-317).

Boutell, T. (1998) Wusage Home Page [Online] Available at http://www.boutell.com/wusage/ (Accessed November 11, 1999).

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) (pp. 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).

Fielding R. (1998). Wwwstat: httpd logfile analysis software [Online] Available at http://www.ics.uci.edu/pub/websoft/wwwstat/ (Accessed November 11, 1999).

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 November 11, 1999).

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 November 16, 1999).

North, C. & Shneiderman, B (1999). Snap-together visualization: coordinating multiple views to explore information. University of Maryland, Department of Computer Science Technical Report CS-TR-4075.

Papadakakis, N., Markatos, E. P., & Papathanasiou A.E. (1998) Palantir: A Visualization Tool for the world wide web. Proceedings INET '98 Conference.

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. In search of reliable usage data on the WWW. (1996)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 November 16, 1999).

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 November 16,1999).

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 November 16, 1999).

Tauscher, L, & Greenberg, S. (1986) Revisitation patterns in world wide web navigation. Conference Proceedings on Human Factors in Computing Systems (ACM CHI '97), (pp. 399-406).

Turner, S. (1999). Analog: WWW logfile analysis [Online] Available at: http://www.statslab.cam.ac.uk/\verb+~+sret1/analog/ (Accessed November 16, 1999).

Uppsala University, IT Support (1999) Access log analyzers [Online] Available at: http://www.uu.se/Software/Analyzers/Access-analyzers.html (Accessed November 11, 1999).