Understanding Patterns of User Visits to Web Sites: Interactive Starfield Visualizations of WWW Log Data

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 starfield 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 starfield visualizations, which 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, 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 common-place 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 [2]. 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 [17] 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. 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 [20], analog [5], HitList [11], and wusage [19] 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 displays* : 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 BazaarSuite [6] 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 [16], various projects have revisited the issue of log display and visualization. Disk Trees and Time Tubes [8] 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 [13] and Chitra [1] 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 [18] [15] [9] [14]. 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 [3] such as Spotfire [17] 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.

Interactive visualizations of visits to the web site of the Human-Computer Interaction Lab (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

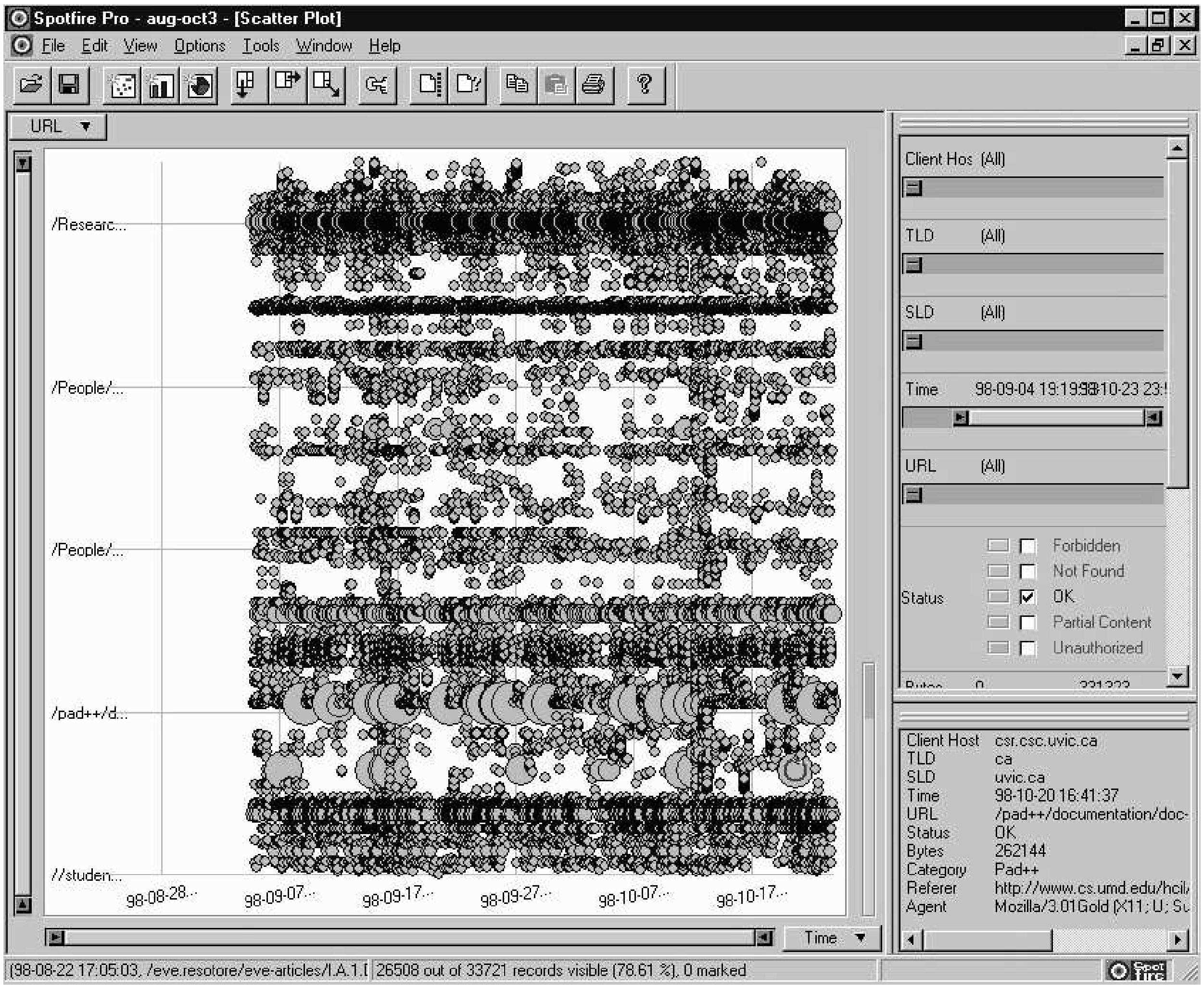

Figure 1: *Interactive Visualizations in Spotfire:* A Spotfire visualization, with the URL requested on the y-axis and the time of request on the x-axis. Checkboxes on the right-hand side have been selected to display only data points with "OK" status, and a slider labeled "Time" has been adjusted to 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.

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 [9] 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, and User-Agent for each hit. The available data fields are summarized in Table 1.

For a two-month period covering late August to late October 1998, this data set consisted of over 33,000 data points. This data set 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.

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).

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

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 [10]

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

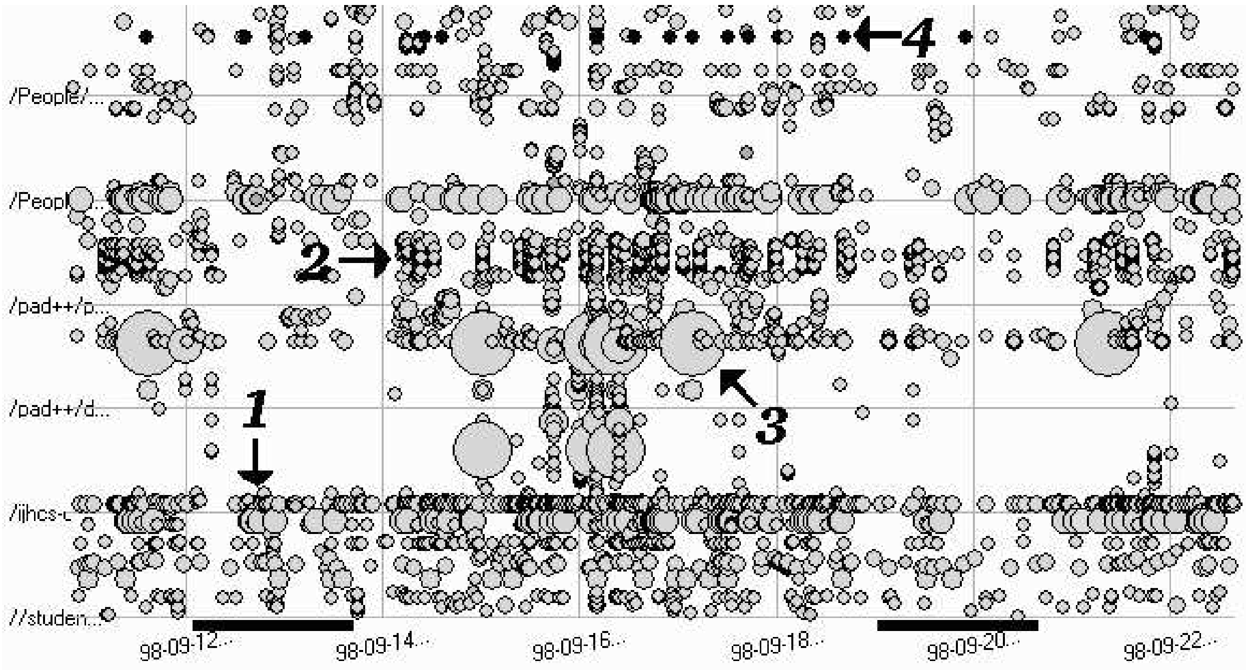

Figure 2: *Time vs. URL, Macro View:* Two weeks of accesses to a subset of the HCIL pages. The requested URL is on the y-axis, with the date and time on the x-axis. The dark lines on the x-axis correspond to weekends. Each circle represents a request for a single page. The size of the circle indicates 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 3.1: 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.

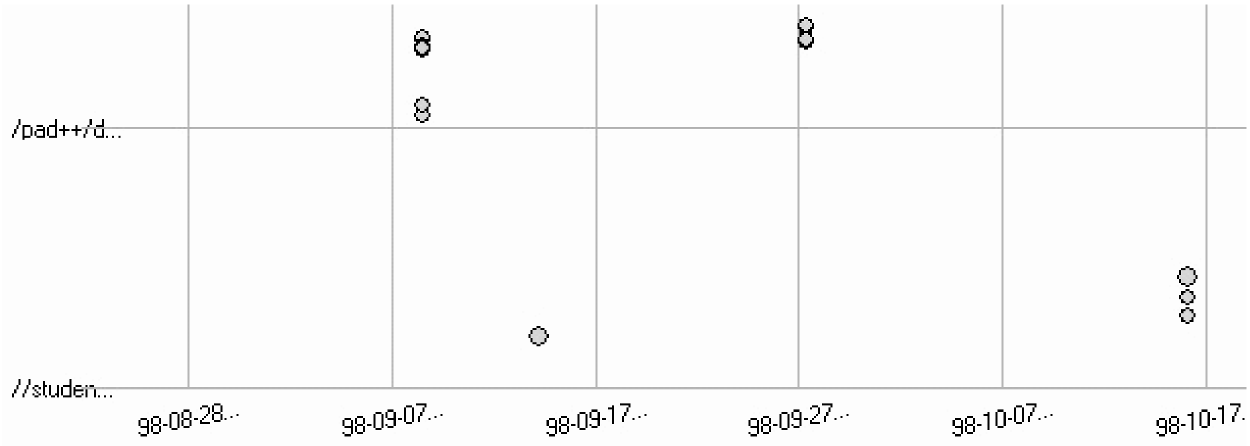

Figure 3: *Time vs. URL, Micro View:* A series of requests from a single client. Over the course of five weeks, this client made several series of requests to the HCIL web site: 4 pages on September 8, one on September 14, 3 on September 27, and 4 (of which three are shown) on October 16. URLs are alphabetized 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.

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 on the y-axis (or IP number, if the complete domain name is unavailable) 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 their site is being visited by an automated agent.

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. 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.

This visualization can be used to track frequency and regularity of user visits to the home page links. Of course, this display does not address the relative use of the underlying pages: in fact, this visualization confirmed our suspicions that many of the user visits to HCIL pages were coming from external links. Unfortunately, this visualization is potentially misleading, as references to pages linked from the home page do not necessarily involve selection from the home page: site visitors might arrive at these pages by selecting links from some other page, or by typing a link directly into their browser.

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 5). 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 Referer vs. Time

Many web site operators are interested in understanding their site's position in the web universe. While search engines such as Altavista [4] 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,

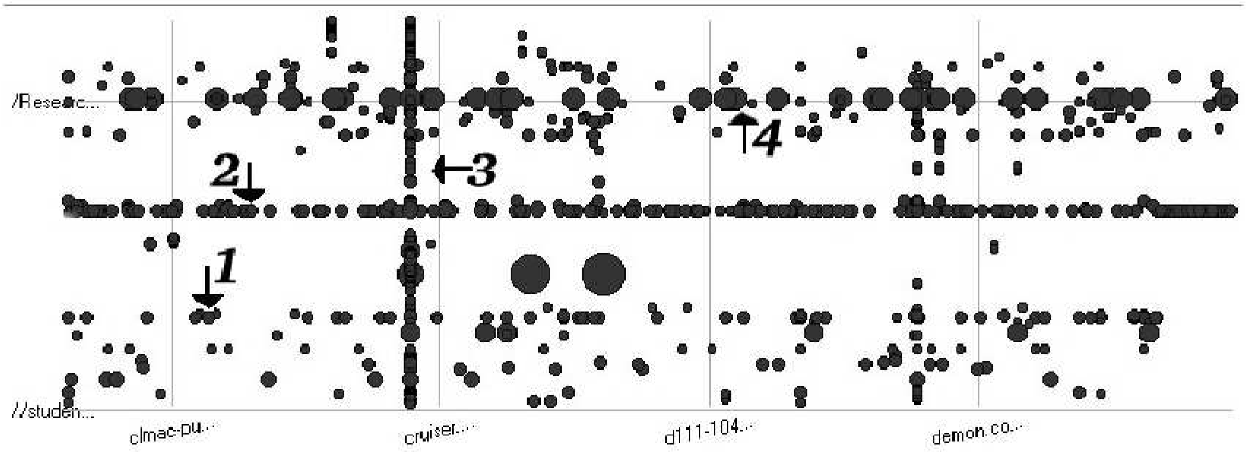

Figure 4: *Client Host vs. URL*: Client hosts on the X-axis, and requested URL on the y-axis. Vertical slices display files visited by each host, while horizontal slices indicate the patterns of requests for a given web page. This display indicates that the HCIL index page (arrow labeled 1) is visited 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.

referer 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 6).

For example, dense horizontal bands indicate referer 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 referers may provide interesting clues as to which links on the site are being used. Furthermore, changes in the referer profiles over time may indicate the addition or deletion of new links to the site.

Examination of the range of referers 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 referers, leading to horizontal bands in the visualization, indicating classes of visits from different search engines. Furthermore, search terms are often encoded in the URLs of search results, so examination of individual referer URLs for these search engine referers may provide some insights into the search keywords that are leading visitors to the site.

3.7 Referer 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 referers 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 referers*: Horizontal bars correspond to pages that are referenced from multiple sources - either external or internal. These bars may be used to

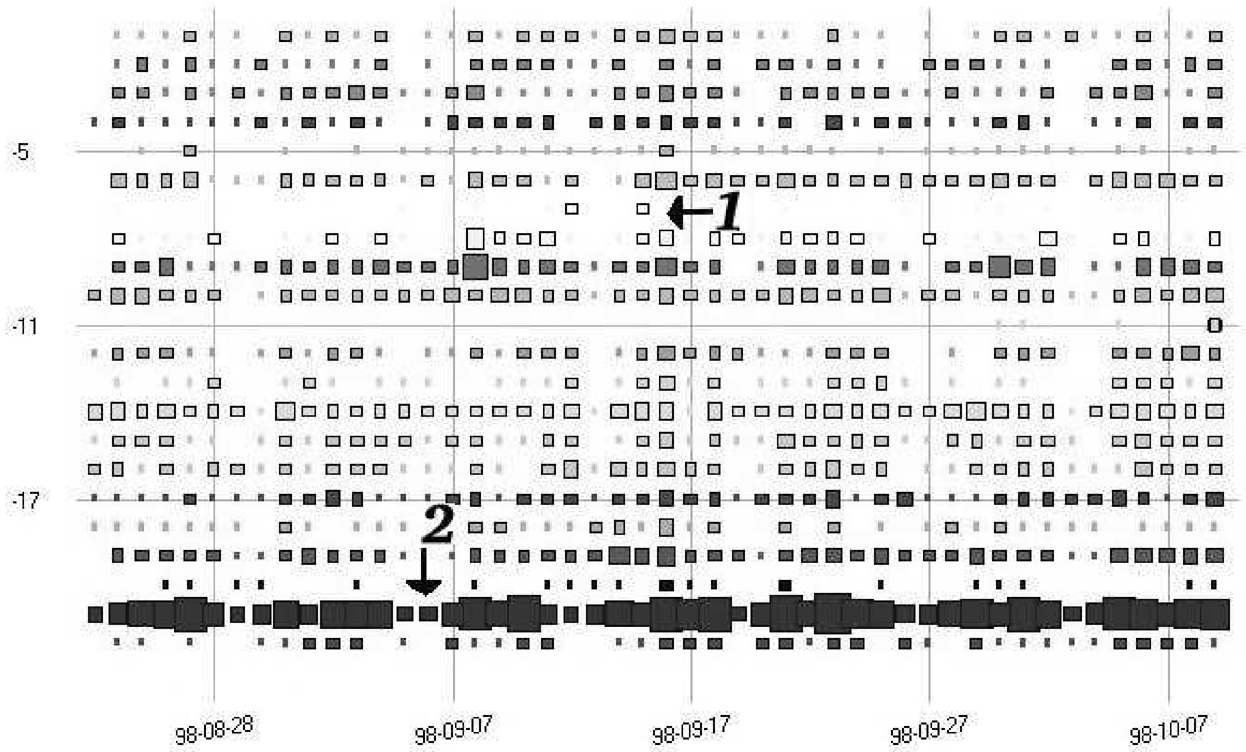

Figure 5: *Index Page Link Requests*: Requests for pages that have links on the group index page. Each row corresponds to a link on the index page. The vertical position of each row in the visualization corresponds to the vertical position of the link on the index page, with links at the top of the page 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. 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.

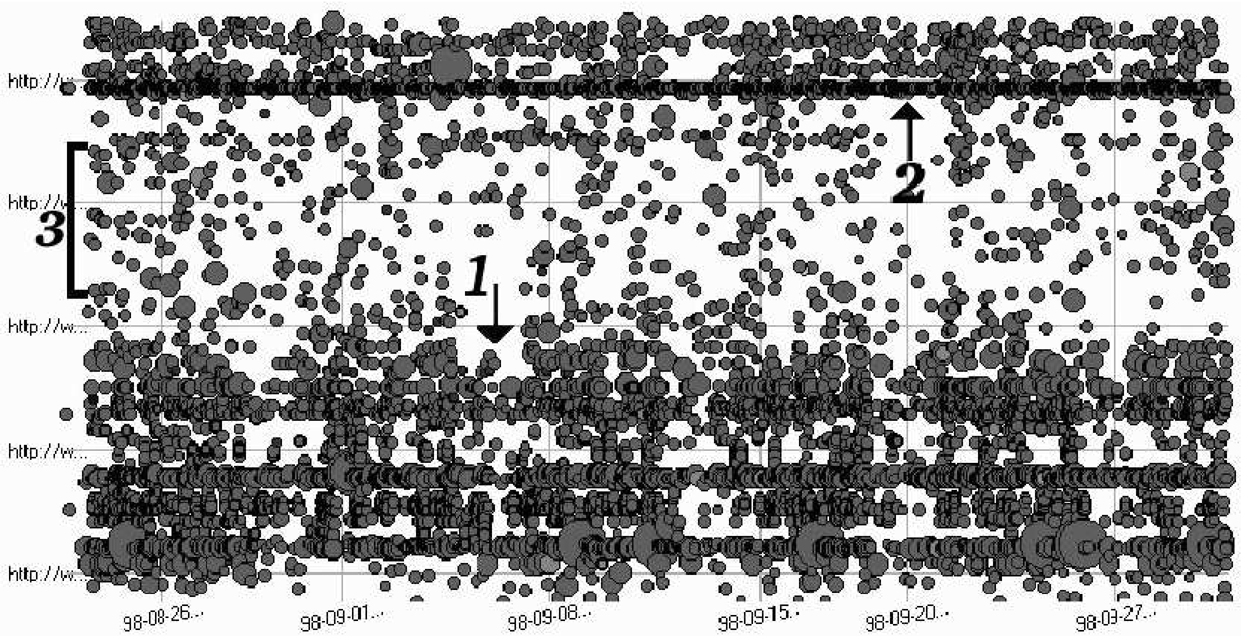

Figure 6: *Referer vs. Time*: The URL of the referring page is given on the y-axis, and the request date and time is on the x-axis. Referer patterns seem to be fairly regular across the 5 weeks of data displayed. The points shown below arrow 1 are those that have referring URLs that are inside the HCIL 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 referer URLs corresponding to a search engine.

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 referers*: Vertical lines (or bands) indicate URLs (or groups of URLs) that may reference multiple pages on the site. In the case of external referers, these patterns may be used to identify WWW resources with a strong affinity to the material on a given site.

- *Non-link references*: The referer 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 referer 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, referer/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 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 referers, or through aggregation of referers 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:

- *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.

- *Totals by Time Period*: With the exception of the "front page visits", the above visualizations do not summarize information in terms of hits by day, week, or month. Visualizations based on these aggregates (which can be easily generated through re-processing of the log files) might be used to identify further patterns.

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

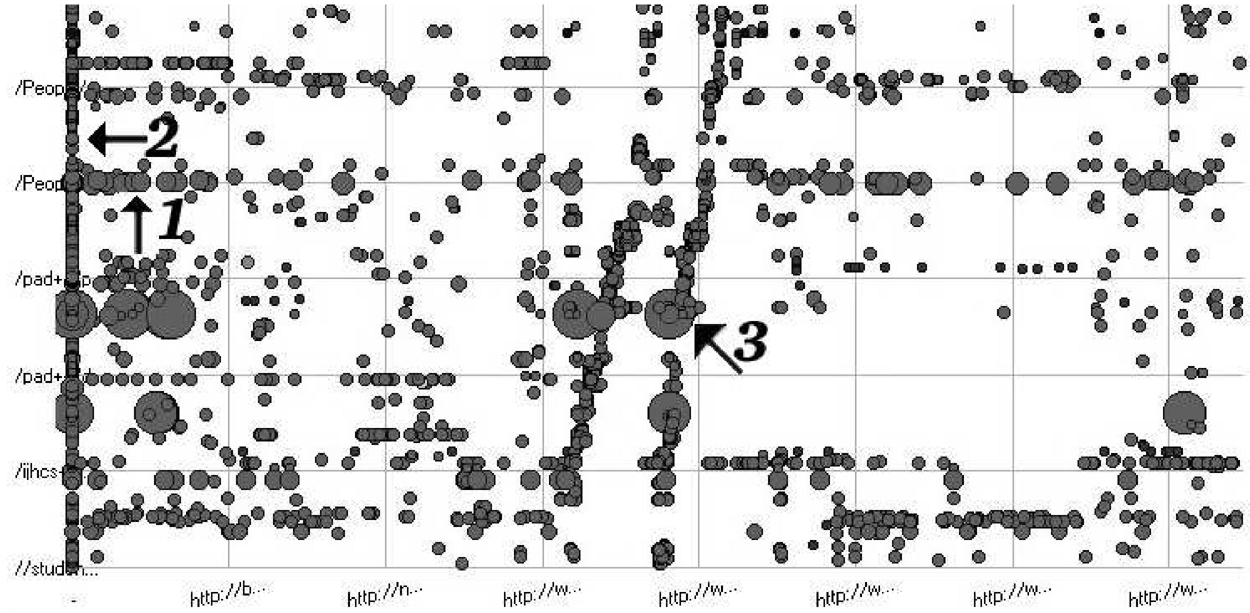

Figure 7: *Referer vs. URL:* URL on the y-axis, referer on the x-axis. From this visualization, we can see that some URLs have numerous associated referers (see arrow #1), indicating either multiple links to the URL or frequent searches that lead users to the URL. Arrow #2 points to the "non-link" references: requests that did not involve named referers. Arrow #3 illustrate the patterns formed when users follow paths through the site, as described in Section 3.7.

4 Discussion

All of the data reported above might be included - in some form - in the output of a traditional web log analysis tool. However, interactive starfield visualizations offer several advantages [3] in understanding user visits, including:

- *Rich display of multiple-dimensional data, allowing discovery of multiple trends.* The Time vs. URL visualization described above can potentially reveal several usage patterns in the data.
- *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 over thousands of data points, representing every site user visit for a month-long period, on a single screen.
- *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 thousands of hits from a period of several weeks to individual accesses from an hour-long visit by a single user.
- *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" [14] [9].

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.

With the exception of "Front Page Visits" 3.5, the visualizations described above presented each 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. For example, the "time vs. URL" visualizations provide access to each individual data point, but identification of aggregate trends may not be immediately available. Interactive visualizations might work best as complements to traditional analysis tools. Specifically, visualizations might be most useful for identifying questions to be asked, while traditional log analysis tools might provide the answers to those questions.

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.

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: 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.

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 appropriate 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 [7].

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 somewhat 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.
- *"Inter-visualization" visualization:* While investigating the index page link request visualization, we noticed that one area of the web site had a high level of traffic, despite the fact that the corresponding link on the index page was not heavily used. We hypothesized that user's were arriving through an external link to another page in that section of the site. Examination of the referer vs. time and referer vs. URL visualizations verified that this was indeed the case. Tools that would allow for coordination of data between different visualizations would increase the likelihood of identifying trends of this sort.

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. This reduced set would ideally provide necessary understanding with minimal effort.

Despite the efforts of several research projects [15] [9], modeling of web usage remains an inexact science [12]. Interactive visualizations of web log data may be useful complements to static reports generated by current tools and session models currently being developed.

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 direct observations for further clarification.

Acknowledgments

This research was supported by a grant from IBM. Thanks to Anne Rose for help with generation of the visualizations.

References

[1] Abrams, M., Williams, S., Abdulla, G., Patel, S., Ribler, R., and Fox, E. Multimedia traffic analysis using chitra95. In *ACM Multimedia* (1995). http://ei.cs.vt.edu/~succeed/95multimediaAWAFPR/95multimediaAWAFPR.html.

[2] Access Log Analyzers. http://www.uu.se/Software/Analyzers/Access-analyzers.html.

[3] Ahlberg, C., and Shneiderman, B. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *ACM CHI Conference on Human Factors in Computing Systems* (1994), pp. 313–317.

[4] Altavista. http://www.altavista.digital.com.

[5] Analog. http://www.statslab.cam.ac.uk/~sret1/analog/.

[6] BazaarSuite. http://www.bazaarsuite.com.

[7] Büchner, A. G., and Mulvenna, M. D. Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record 27*, 4 (December 1998), 54–61.

[8] Chi, E., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., and Card, S. Visualizing the evolution of web ecologies. In *ACM CHI Conference on Human Factors in Computing Systems* (1998), pp. 400–407.

[9] Cooley, R., Mobasher, B., and Srivastava, J. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems 1*, 1 (1999). http://maya.cs.depaul.edu/~mobasher/papers/webminer-kais.ps.

[10] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., and Berners-Lee, T. Rfc 2068 – hypertext transfer protocol – http/1.1. http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2068.txt.

[11] Hit List. http://www.marketwave.com.

[12] Monticino, M. Web-analysis: Stripping away the hype. *IEEE Computer 31*, 12 (December 1998), 130–132.

[13] Papadakakis, N., Markatos, E. P., and Papathanasiou, A. E. Palantir: A visualization tool for the world wide web. In *INET'98 Proceedings* (1998). http://www.csi.forth.gr/~papathan/papers/INET98_Palantir/.

[14] Pirolli, P., Pitkow, J., and Rao, R. Silk from a sow's ear: Extracting usable structures from the web. In *ACM CHI Conference on Human Factors in Computing Systems* (1996), pp. 118–125.

[15] Pitkow, J. In search of reliable usage data on the www. Tech. rep., Georgia Tech. College of Computing, Graphics, Visualization, and Usability Center, 1996. ftp://ftp.gvu.gatech.edu/pub/gvu/tr/1997/97-13.pdf.

[16] PITKOW, J., AND BHARAT, K. Webviz: A tool for world-wide web access log analysis. In *First International Conference on the World-Wide Web* (1994). http://www1.cern.ch/PapersWWW94/pitkow-webvis.ps.

[17] SPOTFIRE. http://www.spotfire.com.

[18] TAUSCHER, L., AND GREENBERG, S. Revisitation patterns in world-wide-web navigation. In *ACM CHI Conference on Human Factors in Computing Systems* (1997), pp. 399–406.

[19] WUSAGE. http://www.boutell.com/wusage/.

[20] WWWSTAT. http://www.ics.uci.edu/pub/websoft/wwwstat/.