Improving Web Usability Through Visualization

Predictive Web usage visualizations can help analysts uncover traffic patterns and usability problems.

Ed H. Chi
Palo Alto Research Center

Despite major efforts in Web design research, many sites remain difficult to use. In a recent study of e-commerce sites, for example, users successfully performed only 56 percent of their intended tasks.[1] Forrester Research has reported that 65 percent of all online shopping trips end in failure,[2] and that 40 percent of all visitors choose not to return to a site because of design problems.[3] Clearly, usability can make or break a Web site.

One reason for poor Web site usability is researchers' lack of understanding about how users surf for information. The field of Web analytics, which is currently a billion-dollar-a-year industry, arose to address such usability issues. Web analytics attempts to increase usability and visitor retention through examination of server log data. A variety of tools extract metrics from these files, such as the number of unique users, page visits, and session lengths. While some of these tools have evolved into products, administrators often must sift through large amounts of simple statistics that provide little insight into user behavior and traffic patterns. Moreover, analysis efforts are complicated by the fact that a site's usage, as well as its content and structure, change greatly over time.

In the past decade, some researchers have explored visualization methods to help understand usage data and identify major traffic patterns. In this article, I describe some of the techniques we in the User Interface Research Group at PARC have developed to visualize content changes, linkage structures, site usage, and so on. I explore how these techniques can be used to identify specific usability problems at large Web sites. As part of our work, we have created a predictive visualization model called Information Scent, in which simulated users help uncover patterns and deficiencies in information accessibility. We demonstrate these techniques using a prototype system called ScentViz to illustrate how Web usage analysis can be enhanced using visualization methods.

 http://computer.org/internet/ 1089-7801/02/$17.00 ©2002 IEEE

Challenges in Analyzing Web Usage

Recent advances, from the data mining movement to new statistical techniques in natural language processing and information visualization, have made it easier to manage Web data analysis. Because they help combat complexity in understanding user behavior, these techniques will remain important elements in every Web analyst's toolbox. Nonetheless, visualizing Web usage data to improve usability presents several challenges.

Data Mining

Effective visualization requires sophisticated data analysis and data mining techniques to distill the raw data. Analysis algorithms and open-source software such as Analog (www.analog.cx) can provide descriptive statistics on Web traffic, such as the frequency and deviation of accesses for a given document or area of a site. This simplistic approach is undesirable because it paints a fairly inaccurate picture of site usage. As one usability guru remarked, "Hits is what idiots use to track success." Sophisticated data mining requires four data sets:

- a copy of the content of each page on the site
- the linkage topology structure between the pages
- the usage logs from each server
- user self-reported demographic or ratings data

Missing any of these data sets can make it impossible to apply a given analysis technique. Moreover, the data sets are usually large and difficult to process. Researchers have therefore devised methods to construct sessions from server logs.[4] These user paths tell us how the users navigate through the site. Cleaning up the logs and dividing them into sessions to create user paths is difficult, especially when multiple distributed servers are servicing a single site. This nontrivial step still generates more data than can be easily analyzed without visualization techniques.

We can gain insights into the user paths and profiles by applying various data mining algorithms. Clustering algorithms, for example, can classify users into categories such as investors, buyers, product seekers, or article readers.[5] We can use association rule data mining to discover interesting access patterns, such as that people who download scanner drivers always download the scanning software update as well. By applying decision-tree algorithms on the fly, we can create optimized paths dynamically. Based on the path taken, we can use estimation algorithms to guess unknown values, such as demographic attributes.

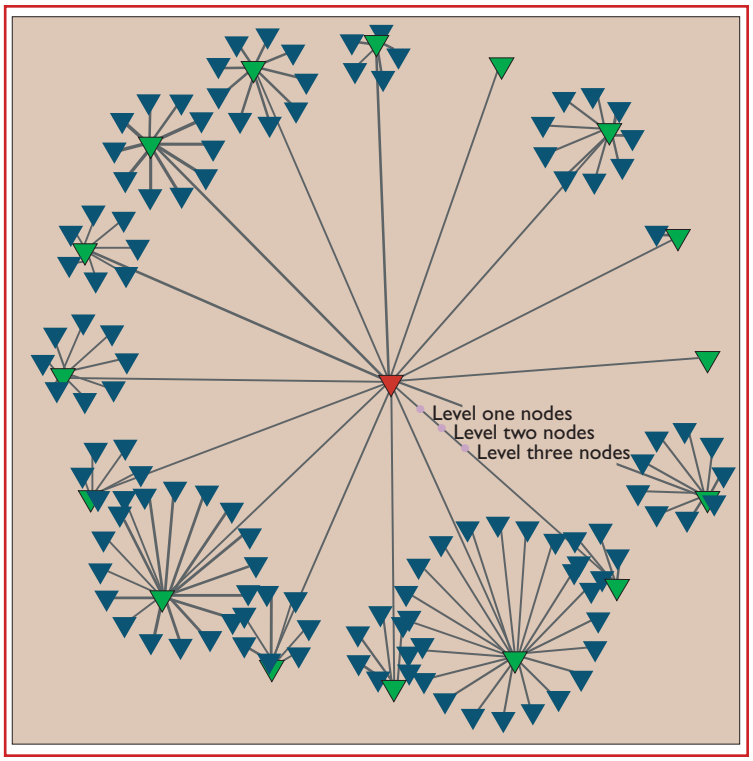

Figure 1. Top view of a site structure visualization. The red glyph in the center of the graph depicts the root node of the site. The green glyphs represent pages one click away from the root; blue glyphs represent pages two clicks away.

Once we apply these data analysis techniques, we need to make sense of the results. Unfortunately, data analysis and data mining algorithms often produce extensive reports, which can be difficult to distill into action items. Visualization can give a global picture of data analysis results.

Web Structure Visualization

An entire field of research is devoted to the general problem of visualizing graph structures, and one branch of researchers has taken up the challenge of visualizing the structure of the Web. Fortunately, Web sites have certain qualities — hierarchical structures, for example — that make visualizing them slightly easier than visualizing a general graph.

The first-level goal of early Web visualizations was to help users navigate more effectively by visually representing the document structure. Figure 1 shows one of the earliest visualizations, based on the *cone tree* visualization[6] of a medium-size site (about 100 nodes).[7] Andrews described a landscape approach for small Web sites,[8] and Munzner used a 3D hyperbolic browser to map fairly large sites.[9] Inxight Software's (www.inx-

ight.com) SiteLens used the hyperbolic tree technique[10] to visualize Web site structure, while NicheWorks[11] used an angular layout similar to disk trees.[12]

Although these early Web site visualization systems could faithfully represent Web site structure, they were never widely used, probably because sites started providing site maps that worked just as well as automated visualization solutions.

More recent Web structure visualization efforts focus on visualizing the global network topology of the entire Web. With the goal of "understanding the organization," these later efforts visualize the connectivity of major traffic servers around the world. The rationale is that we should be able to reason with the structure to pinpoint interesting artifacts of the server organization. Understanding the structure of large servers should let us tune Web performance, including server delay and connectivity. CAIDA, for example, (www.caida.org) attempts to map the global Internet infrastructure. For a list of other Web visualization efforts, see CyberGeography (www.cybergeography.org).

Web Evolution

Web analysts realized early that understanding the users' behavior patterns was more important than understanding the Web's structure. However, the constant evolution of content and site usage compounds the challenge of Web analytics and our ability to generate good visual representations that include both user paths and site structure. Indeed, extracting trends from the time series of large-scale Web usage data is difficult.

Most past Web usage visualization efforts worked with very few paths or focused on sites that were too small to be useful by today's standards. Early work visualized small-scale structures along with descriptive statistics such as hit frequency, visitor domains, and browser usage. In probably the first work in this area, Pitkow and Bharat produced a Web site visualization in which server logs could be played back as animations with interactive filters of the logs based on host names, access times, and URLs.[13] As each server access was played back, link edges lit up in red and faded to blue. Although the techniques clearly do not scale to sites with more than about 100 pages, this first attempt at Web usage visualization successfully presented an organized view of the visitor activities in log files.

Later visualizations incorporated additional derived statistics, such as user sessions and user paths. In addition to our 1997 work on visualizing usage, WebPath visualized a handful of user access patterns at a Web site,[14] and VisVIP let administrators visualize user paths through a Web site — but for only one path at a time.[15] While these two systems offered a good start, neither was designed to handle a collection of user paths from an entire Web site's log. Moreover, none of the systems uses a predictive model of Web usage.

Both our ScentViz system and Astra's SiteManager (www.merc-int.com) visualize entire Web site usage logs and site structure, and let users drill down on user path statistics. SiteManager extracts patterns showing how users navigate throughout a site, allowing analysts to compare navigation patterns between different parts of a site. Site Manager is not designed to discover patterns across time, however, because it is hard to compare usage logs side by side in the application.

Accrue Insight (www.accrue.com) has also made significant contributions to analyzing entire usage logs. With its focus on business-oriented metrics such as customer retention and conversion rates, however, this software is not primarily designed to improve Web site usability. Instead, the system's simple visualizations use charts and tables to show advertising campaign results, bi-graphs to show user entrance and exit points, and so on.

Although little has been written on Anemone, the system is important because it introduces a way to grow the layout for a site's structure "genetically."[16] The system adds new pages to the structure as they appear, and nodes shrink and wither away as pages receive fewer visitors. Anemone shows usage evolution through animation, whereas our system shows usage evolution statically. Because Anemone does not incorporate a predictive model of usage as ours does, truly detecting outliers requires more effort because analysts must thoroughly understand the descriptive statistics.

A Visualization-Based Approach

Our approach to improving site usability through Web usage visualizations includes four parts:

- visualizing site structure with usage-based layout algorithms,
- visualizing site evolution through time tube and visual subtraction techniques,
- visualizing significant traffic routes by first extracting important paths, and
- incorporating a predictive model of user paths.

Usage-Based Layout

While various techniques for laying out graphs exist, they cannot handle the size and complex structure of a typical enterprise site. To improve our ability to visualize such structures, we can exploit aspects of the underlying data in the layout.

In a departure from traditional graph layout methods, which rely exclusively on structural relationships, usage-based layout (UBL) lets us reduce the Web graph to a hierarchy derived from usage data.[17] UBL uses *link induction* and *priority-based traversal* to determine hierarchical relationships that provide a sense of direction for user flow. As a side benefit, UBL provides visualizations of Web paths with fewer path crossings because the user paths are given preferential treatment during graph layout decisions.

Link induction. Link induction identifies user paths between pages that are not explicitly hyperlinked. The transaction log might record a traversal path of A → B → C → L, for example, even though there is no hyperlink between documents C and L. This could arise from the use of Back buttons, search results, dynamic pages, and so forth. The actual path might have been A → B → C → B → L if B was a search result page. Whatever the cause, these induced links indicate information that is related but not directly hyperlinked. The greater the induced link's frequency, the more likely that two documents are related. By mining usage logs, we can thus create visualizations that show usage linkages more accurately than those based solely on structural hyperlinks. Link induction enhances our understanding of user behavior better than merely visualizing the site's structure.

Priority-based traversal. To determine the hierarchical relationships between documents, we conduct a priority-based traversal of the site based on usage data. We first examine the hyperlink structure and the induced links to determine the root node's children, which we insert into a priority queue from which we extract and expand the most frequently accessed page. We then insert that page's children into the priority queue and pick out the next most popular page, expand to its children, and so on. This priority scheme should thus ensure that the visualization roughly represents the most popular routes through the data.

Disk Tree Visualization

Figure 2 illustrates the *disk tree* visualization technique for about one week's worth of usage data for the 7,588 documents at the Xerox site.[17] The center of the disk tree is the root page, and each line represents a link to another page in the site. Yellow lines represent links to deleted content, and red lines represent links to added content. The brighter the green links, and the wider the lines, the more frequently the link was traversed. Links one hop away are on the first concentric ring; two hops away are on the second concentric ring; and so on. For each node on each concentric ring, the basic disk tree algorithm allocates angular space proportional to the number of children of that node. The system displays detailed usage information on the side for each node the mouse passes over. The user can thus probe the structure to learn where each area of the site appears on the disk tree.

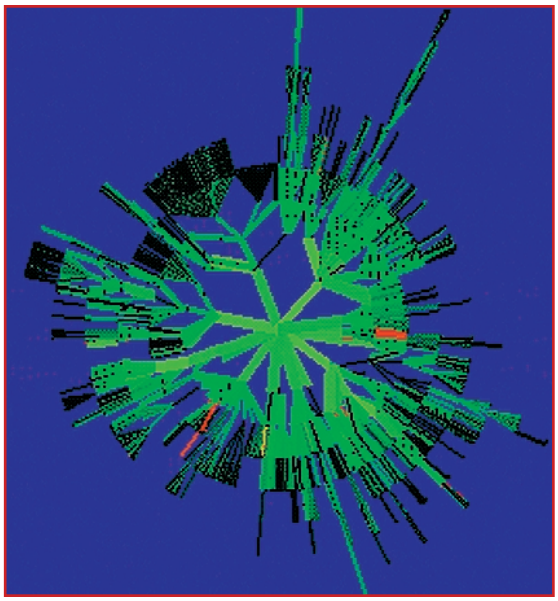

Figure 2. Disk tree technique. This visualization represents a week's worth of usage data for the 7,588 documents at www.xerox.com. The center of each disk is the root page of the Web site. Yellow lines indicate deleted content and red lines show added content.

Because the disk tree is based on UBL, a page node's distance from the root node is based on the most popular path to it. A page that is only two hops from the root node at its closest point, for example, might typically take a user along a five-hop path to reach it. Notice that UBL yields disk tree layouts that provide a sense of direction for user flow: traffic generally starts in the high-traffic parts of a site (shown in the center) and moves outward.

Visualizing Usage Evolution

The rapid pace of evolution at most complex sites makes it both more difficult and more rewarding to analyze the usage data. To solve real usability prob-

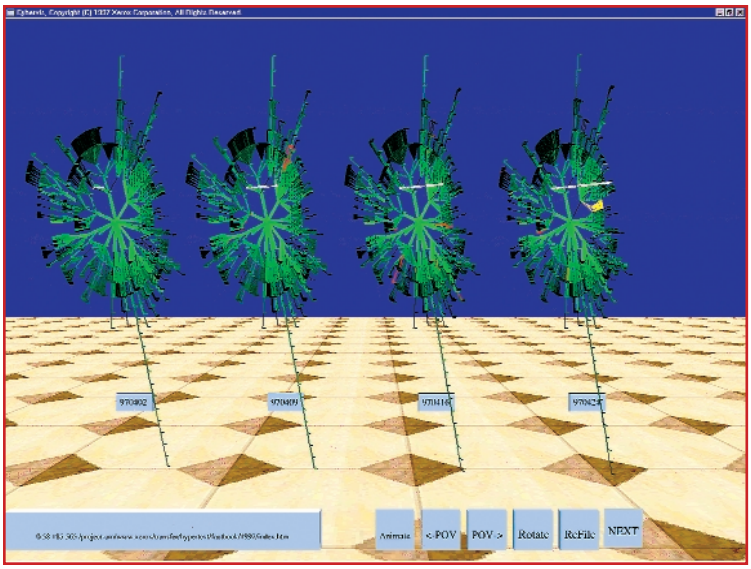

Figure 3. Time tube visualization. Document access trends are shown as slices of disk trees. Each slice here represents one week's worth of content, usage, and structure changes at the Xerox Web site.

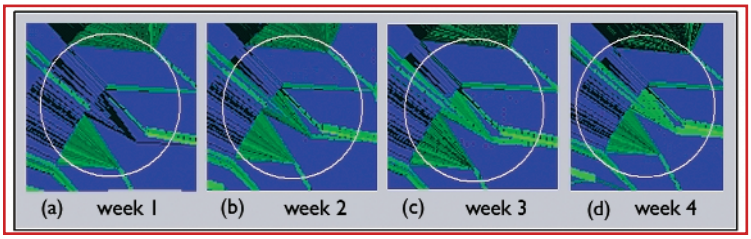

Figure 4. Visualizing traffic increase. The triangular structure in the center of the circle became greener from the first week (a) to the second (b), and from the third week (c) to the fourth (d), meaning traffic was increasing to the corresponding pages.

lems at such sites, analysts need tools for discovering new relationships or correlations in Web traffic data. Issues such as how recently added content affects user traffic can provide important insights into various trends in user interests. We can improve nearly every aspect of the user experience on a Web site by understanding the users' goals and traffic composition. Webmasters, marketers, and content producers can use this information to better tailor sites to user needs by presenting highly relevant materials as well as targeted promotions and advertisements.

Identifying Traffic Patterns

Figure 3 shows a visualization for a month's worth of server log data along with the changes in the Web site's structure and content. The *time tube* technique represents time on the horizontal axis and uses multiple disk tree slices to represent specific time intervals. In Figure 3, for example, each disk tree slice presents the usage patterns for a week's worth of data. The time tube technique allowed us to see trends in traffic pattern changes as well as changes in the document collection at the site.[17]

Figures 4a through 4d, which are zoomed in to a particular area of each slice, show a noticeable traffic increase for one collection of pages in the disk tree. The area marked gets greener from week 1 (a) to week 4 (d), indicating an increase in usage. This increase during the second week is due to a site modification that helped promote the pages. After locating interesting traffic patterns using these techniques, we started looking for ways to support this task more directly.

Visualizing Usage Changes Directly

Figure 5 illustrates the visual *usage pattern subtraction* method we employed to depict traffic changes directly. It shows the difference between weekly usage patterns. That is, column one is the result of subtracting week one from week two; column two shows week two minus week three, and column three is week three minus week four. Row 1 shows only the decrease in traffic (blue links), and row 2 shows only the increase (red links).

Even with high traffic and evolving content, usage, and structure, we can create visualizations that show trends for a site. Using pattern subtraction, we found an area (marked with a yellow oval) that showed increases in user interest in the information related to two software packages, which are indicated by red links in the figure. A second area (marked with a yellow rectangle) showed traffic increases during the second week, but not the third or fourth week. Without visual subtraction, these trends are not easy to see. By illustrating what information is most popular, this technique gives Webmasters a better sense of how to direct content development efforts to improve a site's usability.

Identifying Significant User Paths

Direct visualizations of user sessions can help developers identify specific bottlenecks in a site. The *longest repeating subsequences* (LRS) technique can model and extract significant surfing paths from server logs.[18] An LRS is, roughly, a subsequence of a user path that has been repeated more than once in other user paths. We try to find the longest of these subsequences in the logs. The LRS technique combines several heuristics and pattern-matching techniques to identify the more informative or prevalent paths and store them efficiently. It reduces the complexity of the path model required for representing the raw data, while maintaining an accurate profile of future usage patterns.

Pitkow and Pirolli showed that this algorithm compresses and extracts the top 10 percent of all paths and retains 90 percent of the predictive power of the full data set.[18]

To track well-traveled user paths we modified the disk tree technique to develop the dome tree visualization. Figure 6 shows the dome tree visualization for the user paths related to the TextBridge 98 product page (marked by the white arrow) at the Xerox site. Root pages still appear in the center of the visualization and aggregate user traffic is still shown in shades of green, but dome tree overlays this structure with yellow lines that represent significant surfing paths extracted using the LRS method. The numerous yellow paths show that information related to the document is spread across multiple areas of the site. This suggests that a redesign might bring more cohesion to the site. I have outlined one interesting well-worn path in orange on the left. This serial pattern corresponds to the tutorial pages related to the TextBridge homepage, which is marked by the white arrow.

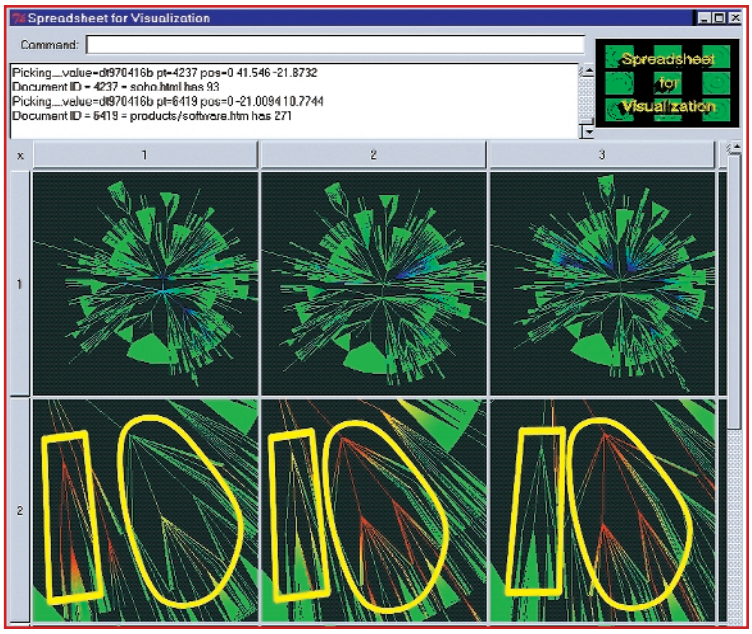

Figure 5. Visual usage pattern subtraction. Each column in this visualization shows changes in usage patterns between weeks, starting with the difference between weeks 1 and 2. Row 1 shows only decreases in usage (blue), and row 2 shows only increases (red).

Predicting Paths by Information Scent

The previous examples showed how we might improve Web site usability by visualizing usage analytics, but a tool that could directly locate problem areas in a site would be more useful. Discussions with Web usage analysts suggest two key questions:

- How do users locate a given page?
- Do users follow the expected navigation paths to the page?

A predictive model of how users search for information at a site would enable us to look for data anomalies and locate potential problem areas. Good visualizations are usually coupled with good analysis models, and Web usage visualizations seem to work best when coupled with good Web analytic algorithms.

One class of models developed in information foraging theory[19] addresses how users navigate through content by following *information scent* — the collection of cues, such as graphics and text snippets surrounding links, available to aid navigational decisions.[20] Researchers have used this notion to develop detailed computational models that can predict users' moment-by-moment behaviors in document browsing. We have integrated the information scent model into the ScentViz visualization system for predicting user surfing patterns.[19]

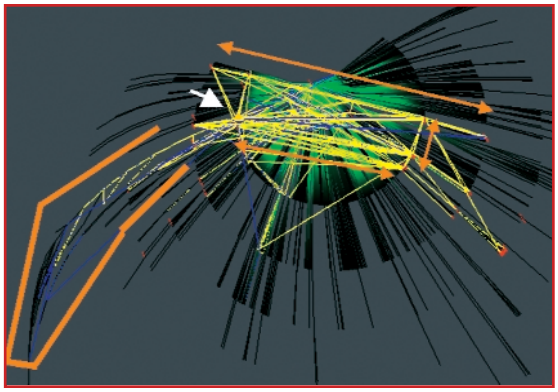

Figure 6. Dome tree visualization. Yellow lines indicate well-traveled user paths involving the document marked by the white arrow. The orange lines indicate major traffic routes.

Assessing Navigation Design

Our approach assumes that users have some information goal, and that a site's information scent determines their surfing patterns. The predictive model employs a simulation in which an arbitrary number of agents traverse the links and content of a Web site. Keyword strings represent the agents' information goals, and the model assesses the information scent by comparing the keywords against the content of each visited page. ScentViz incorporates the information scent prediction algorithms.

Figure 7 shows two predictive visualizations for

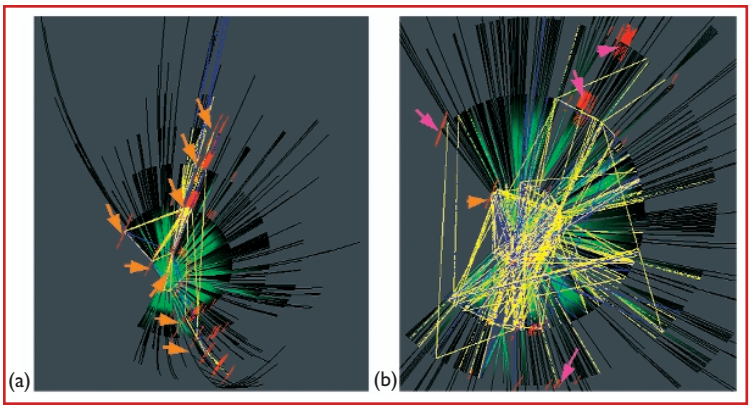

Figure 7. ScentViz information scent predictions. Red bars show predicted user paths. (a) For the Pagis home page, there is a good match between predicted and actual usage—indicated by orange arrows. (b) For the products home page, poor matches between predicted and actual paths indicate a potential usability problem.

users searching the Xerox Web site for information on Pagis, a scanning product. In Figure 7a, the information scent algorithm predictions (represented by red bars) match actual user surfing paths (represented by yellow paths), indicating a well-designed navigation scheme for the Pagis home page.

Potential Navigation Problems

Figure 7b shows the predicted and actual paths for users starting from the products.html page looking for information on Pagis. The fact that the observed user paths (yellow lines) do not match the predicted destinations (red bars) signals a navigation problem.

Because the agents in our information scent simulations read every word on each page to decide which link to follow, they actually represent idealized users. Thus, any deviation from the predictive simulation represents a situation that requires the designer's attention. On careful examination, we discovered that while information related to Pagis was available near the products page, the "scent" was buried in layers of graphics and text that made the information cues hard for real users to see. While simulated users could read and locate these links, real users were confused. This case study showed that the products page provides inadequate access to Pagis information that users might search for. Our predictive model lets us pinpoint potential problem areas to help focus our efforts for improving a site's usability.

Future Work

As Herb Simon once said, "a wealth of information creates a poverty of attention."[20] Visualizations, along with Web analytics, can help us understand the complex relationship between information producers and consumers as the Web continues to grow. Anecdotal evidence suggests that our ScentViz visualization system can help analysts assimilate larger amounts of usage data and identify Web usability problems more quickly than with traditional log analysis reports. Although it has proven useful on one large-scale corporate Web site, however, the system is essentially untested on a larger scale.

Our next challenge is to distribute the tools to Webmasters at a variety of site types. For this strategy to succeed, the visualization techniques must be intuitive and easy to use. We have found informally that, while our technique is not immediately intuitive, users begin to discover traffic patterns with it after a relatively short demonstration. We are thus planning a systematic deployment and technology transfer to explore the long-term benefits of Web usage visualization on site development and usability improvement. Coupled with ways to predict user traffic using information scent, usage visualization could transform our ability to directly understand Web usability.

Acknowledgments

This article describes results from several years of collaboration with Peter Pirolli, Jim Pitkow, and the User Interface Research Group at PARC. I thank them for comments and suggestions. This work was supported in part by U.S. Office of Naval Research grant No. N00014-96-C-0097, awarded to Peter Pirolli and Stuart Card.

References

1. J. Nielsen, "Did Poor Usability Kill E-Commerce?" *Alertbox*, 19 Aug. 2001; online at http://www.useit.com/alertbox/20010819.html.
2. R. Souza et al., "The Best Of Retail Site Design," white paper, Forrester Research, Cambridge, Mass., Oct. 2000.
3. H. Manning, J.C. McCarthy, and R.K. Souza, "Why Most Web Sites Fail," white paper, Forrester Research, Cambridge, Mass., Sept. 1998.
4. J.E. Pitkow, "Summary of WWW Characterizations," *Web J.*, vol.2, nos. 1-2, 1998, pp. 3-13.
5. J. Heer and E.H. Chi, "Identification of Web User Traffic Composition Using Multi-Modal Clustering and Information Scent," *Proc. Workshop on Web Mining, SIAM Conf. Data Mining*, SIAM Press, Philadelphia, Apr. 2001, pp. 51-58.
6. G.G. Robertson, S.K. Card, and J.D. Mackinlay, "Information Visualization Using 3D Interactive Animation," *Comm. ACM*, vol. 36, no. 4, 1993, pp. 57-71.
7. E.H. Chi, "WebSpace Visualizations," *Proc. 2nd Int'l WWW Conf.*, World Wide Web Consortium (W3C), Oct. 1994; http://www.geom.umn.edu/software/weboogl/webspace/webspace.html.
8. K. Andrews, "Visualizing Cyberspace: Information Visual-

ization in the Harmony Internet Browser," *Proc. IEEE Symp. Information Visualization* (InfoVis 95), IEEE CS Press, New York, 1995.
9. T. Munzner, "Drawing Large Graphs with H3Viewer and Site Manager," *Proc. Graph Drawing 98*, Springer-Verlag, New York, Aug. 1998, pp. 384-393.
10. J. Lamping and R. Rao, "Laying Out and Visualizing Large Trees Using a Hyperbolic Space," *Proc. 7th Symp. User Interface Software and Technology* (UIST 94), ACM Press, New York, 1994.
11. G.J. Wills, "Nicheworks – Interactive Visualization of Very Large Graphs," *Proc. Graph Drawing 97*, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1997.
12. E. Chi et al., "Visualizing the Evolution of Web Ecologies," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, New York, 1998.
13. J.E. Pitkow and K. Bharat, "WebViz: A Tool for World Wide Web Access Log Visualization," *Proc. 1st Int'l WWW Conf.*, W3C, 1994; available at http://www1.cern.ch/WWW94/PrelimProcs.html.
14. E. Frécon and G. Smith, "WebPath: A Three-Dimensional Web History," *IEEE Symp. Information Visualization* (InfoVis 98), IEEE Press, Piscataway N.J.,1998.
15. J. Cugini and J. Scholtz, "VisVIP: 3D Visualization of Paths Through Web Sites," *Proc. Int'l Workshop on Web-Based Information Visualization (WebVis 99)*., IEEE Press, Piscataway, N.J., 1999, pp. 259-263; available at http://www.itl.nist.gov/iaui/vvrg/cugini/webmet/visvip/webvis-paper.html.
16. B. Fry, "Mapping How People use a Website," *Mappa Mundi*, June 2001; online at http://mappa.mundi.net/maps/maps_022/.
17. E. Chi et al., "Visualizing the Evolution of Web Ecologies," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, New York, 1998, pp. 400-407, 644-645.
18. J.E. Pitkow and P. Pirolli, "Mining Longest Repeated Subsequences to Predict World Wide Web Surfing," *Proc. 2nd Usenix Symp. Internet Technologies and Systems*, Usenix Assoc., Berkeley, Calif., 1999; available at http://www.parc.xerox.com/istl/projects/uir/projects/1999-08-usenix-lrs-final.pdf
19. E. Chi, P. Pirolli, and J. Pitkow, "The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, New York, 2000, pp. 161-168, 581-582.
20. H.R. Varian, "The Information Economy," *Scientific American*, Sep. 1995, pp. 200-201.

Ed H. Chi is a research scientist in Palo Alto Research Center's User Interface Research Group. He received a PhD in computer science from the University of Minnesota. His area of expertise is software systems for computer-human interaction and 2D/3D user interfaces. Chi is a member of the IEEE, the IEEE Computer Society, the ACM, and ACM SIGCHI.

Readers can contact the author at echi@parc.com.