

Visualizing the Evolution of Web Ecologies

Ed H. Chi[†], James Pitkow, Jock Mackinlay, Peter Pirolli, Rich Gossweiler, Stuart K. Card

Xerox Palo Alto Research Center

3333 Coyote Hill Road

Palo Alto, CA 94304 USA

{echi,pitkow,mackinla,pirolli,rcg,card}@parc.xerox.com

([†] current address: echi@cs.umn.edu, University of Minnesota, Computer Science Department,
200 Union St. SE Room 4-192, Minneapolis, MN 55455)

ABSTRACT

Several visualizations have emerged which attempt to visualize all or part of the World Wide Web. Those visualizations, however, fail to present the dynamically changing ecology of users and documents on the Web. We present new techniques for **Web Ecology and Evolution Visualization** (WEEV). Disk Trees represent a discrete time slice of the Web ecology. A collection of Disk Trees forms a Time Tube, representing the evolution of the Web over longer periods of time. These visualizations are intended to aid authors and webmasters with the production and organization of content, assist Web surfers making sense of information, and help researchers understand the Web.

Keywords

World Wide Web, visualization, log file analysis, temporal analysis, information ecologies, hypertext, documents.

INTRODUCTION

Currently, producers, organizers, and users of content on the World Wide Web must use technologies that are spartan in providing cues that would make their lives easier. For comparison, the librarian or experienced user of a physical library can usually tell, with some ease, the general layout of content, figure out which books are of greatest interest to the community (e.g., by the wear on the book spines), and figure out what's timely or deadwood (e.g., by looking at the circulation check-out stamps inside the book covers). The organizers and authors of such content may use such cues as data for different purposes, to structure content in more appealing and useful ways.

It is important to note that these cues are not just about the content: they are about the relationships among content (e.g., the books themselves), their topology (e.g., layout in physical and topical space), and use (e.g., circulation). To use a biological metaphor, such a complex set of relations form an *ecology* among users and their information

environment, and its change through time is a form of *evolution*. Here we present visualization techniques that provide cues that enhance the understanding of relationships among Web content, topology, and usage through time. These techniques are examples of Web Ecology and Evolution Visualization (WEEV). Such techniques should aid users in finding content, keeping track of it, making sense of it, and understanding its timeliness. The techniques should also help authors, sponsors, and site administrators understand how content is being used and the effectiveness of their designs.

Visualizations of entire information spaces as large as the World Wide Web have been pursued by various researchers [7,8,17]. However, most systems have yet to deal with the larger problem of how these information spaces evolve over time. The problem is more difficult because: 1) the ecology of these information spaces changes rapidly, and 2) the size of these information spaces grows explosively within the time span that we are interested in examining.

In this paper, we introduce a new visualization called the **Time Tube**. The Time Tube organizes and visualizes the evolution of Web sites. For the purpose of this paper we will explore the **application of Time Tubes to the Xerox WWW site** (www.xerox.com), a medium sized *time-extended document ecology* (7,500 items across a 30 day period). Time Tubes are composed of one or more *Disk Trees*, which **represent the hyperlink structure** of a Web site. Figures 1 and 2 show the Time Tube and Disk Trees, respectively.

These new visualizations are designed to **assist users** in the difficult task of **understanding the complex relationships and interactions among the production and consumption of information within Web sites**. If we divide tasks into three levels:

Level 1: (Local) Finding some specific content

Level 2: (Comparison) Comparing information at two places

Level 3: (Global) Discovering a trend or pattern



then most Web interaction methods focus on Level 1 tasks. One of the goals of WEEV techniques is to address tasks at **Levels 2 and 3**.

Different communities of Web users are likely to share at least some common questions and tasks. For instance, consider these questions about the usage and structure of the Web (this list is by no means exhaustive):

Usage

- **How has information been accessed?** How frequently? How recently? What is the most/least popular information?
- **Where do people enter the Web site? Where do people spend time on the site?** How long do they spend in the site?
- **How do they travel within** the site? What are the most popular paths? What is the road least traveled?
- **Who** are the people accessing the site? From which geographic location? From what high level domain (.com, .edu, etc.)?

Structural

- What information has been added? Modified? Deleted? Remained the same, but moved?

Another set of common questions might arise concerning the relations among usage and structure. For example, Web site administrators and content providers are often curious about the following:

- How is new information accessed? When does it become popular? When does it die?
- How does the addition of information change the navigation patterns within the site? Can people still navigate to desired information?
- Would a site benefit from a redesign? Was the redesign of a site effective? What about minor redesigns and restructuring?
- Do people look for deleted information? Could unused information be deleted or relegated to cheaper storage?

While the WEEV system we developed was designed to answer questions such as those above, here we restrict ourselves to only a subset, and use the Xerox Web site in our examples. Before describing the Time Tube system and demonstrating how the Time Tube and Disk Trees help answer these questions we describe the underlying mapping of data to visualizations.

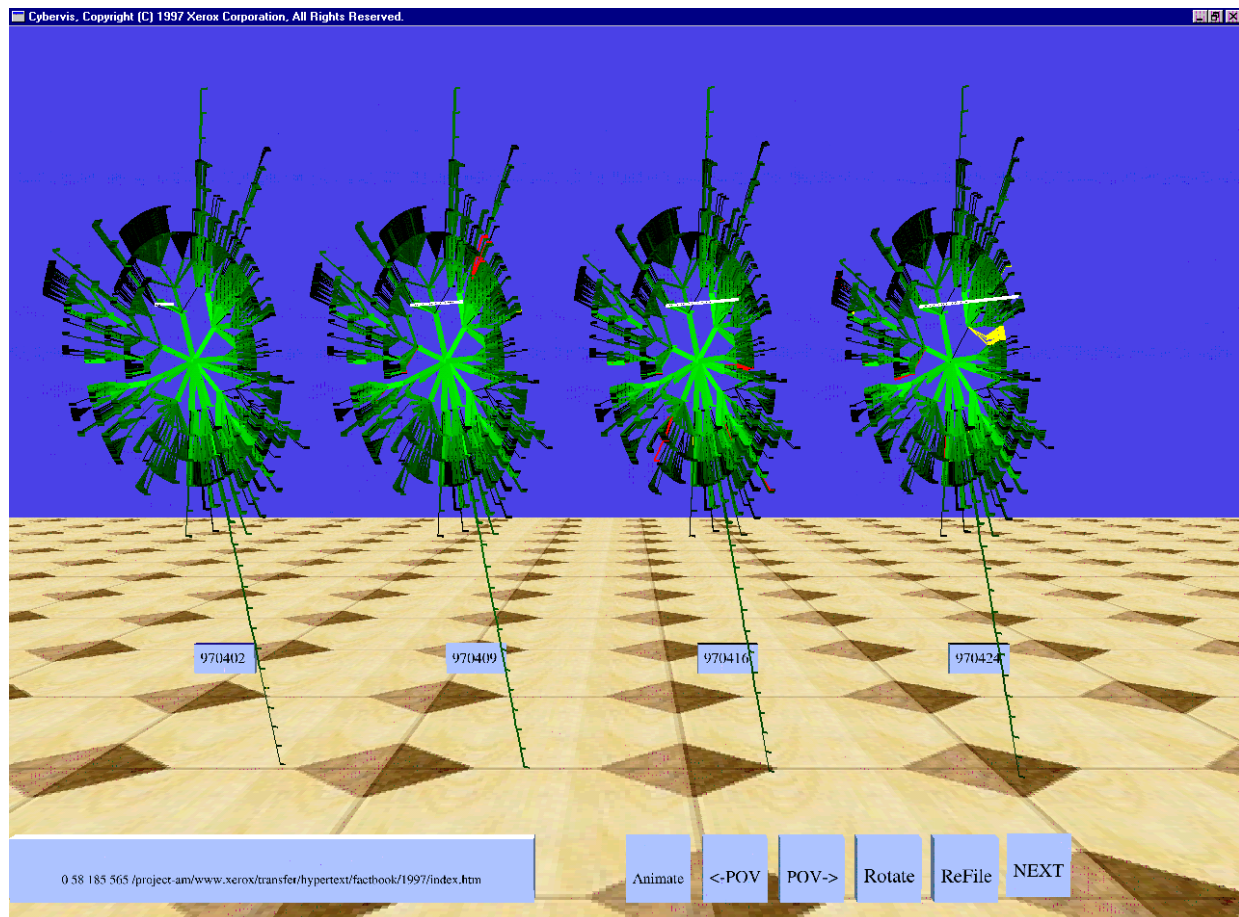


Figure 1: Time Tube shows the evolution of a time-extended document ecology (www.xerox.com during April 1997). This Time Tube is composed of four Disk Trees, **each representing one week of April**.

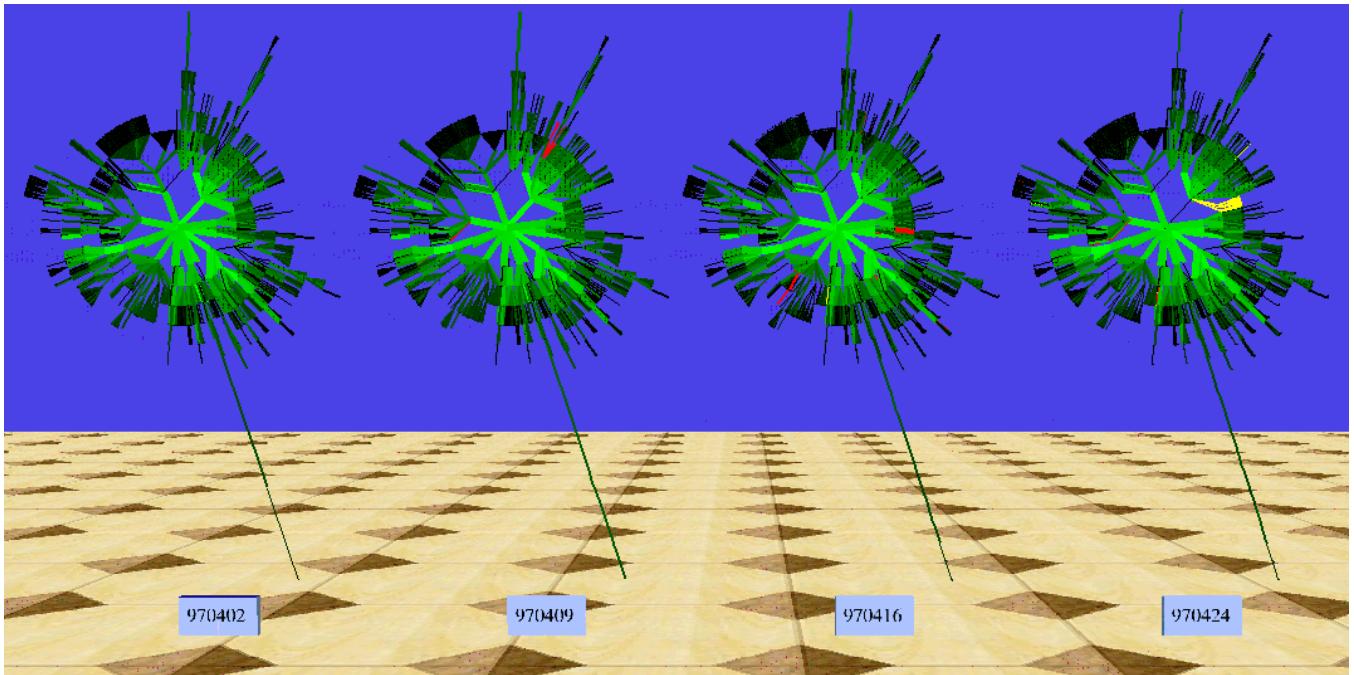
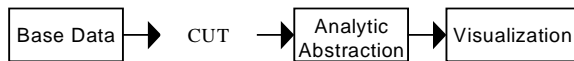


Figure 2: Each of the four **Disk Tree** shows a discrete time slice of a Web ecology (www.xerox.com, April, 1997).

FROM RAW DATA TO USEFUL ABSTRACTIONS

We can think of information visualization as involving a set of mappings from some Base Data set, in some idiosyncratic representation through possibly intermediary representations to some Analytic Abstraction appropriate for insight or communication and then into a Visualization representation suitable for visual analysis and communication with the user.



Current attention within the Web measurement community centers around the number of times items are requested (hits) and hyperlink topology. In the case of the Web, our previous work [13] has shown that it is much more useful to organize an intermediate representation CUT based on **Content (C), Usage (U) and Topology (T)**. In addition to the explicit linkages among documents, the intermediate CUT representation can be used to create implicit relationships with a much more richly connected space.

Previous work has tended to focus on CUT elements in isolation. Early work on information retrieval of document collections, for example, focused on Content, usually represented as document vectors [16]. Later work, e.g., [4] considered Topology, especially in the form of explicit hypertext links. But we have argued [13] that a richer set of analysis can be preformed by combining Content, Topology, and Usage.

We further extend these analyses by examining changes through time. Such data more directly address deeper and

more difficult questions like "Was newly added information well consumed?" and "What faddish trends occur within our site?" Here we concentrate on Topology and Usage computed over time, deriving these from available data and mapping these into an Analytic Abstraction that can be mapped into an information visualization.

RELATED SOLUTIONS

Although the Web has grown rapidly, one can easily argue that the state-of-the-art tools for the analysis of WWW logs have not. Many of the commercially available analysis programs perform the same functionality as the earliest descriptive statistics reporting programs like Kevin Hughes' *getsites*, released during the summer of 1993 [9] and the earliest access log visualization programs like WebViz and WebSpace [14,7].

Analyses

Essentially, modern day log file analysis programs provide spreadsheet-like descriptive summaries of Web site activity. A typical package [12,10] reports the number of times an object was requested, simple characterizations of the machines requesting objects from the site, entry and exit points, and temporal summaries of the requests on a per hour basis. Most packages also create charts that summarize the descriptive data. It is important to note though that while these systems are able to read several months worth of WWW access data, **reports are usually only produced on a per-day basis.**

Visualizations

Recently, several analysis packages have emerged that provide visualizations of an entire site [3,1]. While Accure Insight™ primarily focuses on the real-time monitoring of

Web sites, the Astra SiteManager™ (ASM) concentrates on the structure and usage of a site. Since ASM provides the closest functionality to Time Tubes, we briefly review the functionality of ASM.

ASM provides the analyst with a split screen display, with one part dedicated to visualizing the site and the other towards textual descriptions of the site. The topology of a site is gathered by walking the structure of a site via a spider. Pages and links are displayed using a 2-D cone-tree layout scheme similar to [6]. In this scheme, WWW objects are laid out along a circle with the space occupied by each object being proportional to the number of children it contains. The layout is recursive in the sense that the same circular layout is used for the children of an object. Users are able to interactively zoom-in and zoom-out of the visualization and collapse and expand the children of a page.

With ASM, the analyst is able to create daily maps of a Web site based upon the crawls of the site and the site's usage. The system enables user-specified filtering based upon several parameters (e.g., media type, origin of link, etc.) and colors usage according to user-specified thresholds (e.g., blue for low usage, pink for more active usage, and red for high usage items). Additionally, support is available to help understand the structural changes to a site. This is accomplished by color-coding additions, deletions, and changes that occur between two maps and displaying the result as a single graph.

Although ASM provides many useful functions, it does not easily answer questions about evolving Web ecologies. First and foremost, the system has difficulty visualizing a large number of documents, such as those within the Xerox site¹. A considerable amount of screen real estate is consumed at the expense of displaying the link structure among documents. This effect is further amplified as pages with a large number of children are allocated a circle with a large radius (The large radius is necessary to separate all the children).

Another limitation is that the system only permits the visualization of a site at a single point in time, making it very difficult to make comparisons across times. Additionally, no mechanism is provided that allows *differences* in usage to be identified as well. Without the basic ability to identify usage and structure differences, one can not expect to readily understand how these dimensions change across time.

BASE DATA AND ANALYTIC ABSTRACTIONS

For our exploration of the evolution of a Web ecology, we examined data for the Xerox Web site for a one-month

period starting April, 1997. Content providers for the Xerox site contributed 889 new items (11%), bringing the site to a total of 7,588 items by the end of April. For each day, additions, modifications, and deletions of content to the site were monitored and the corresponding structural changes recorded. Additionally, the daily usage logs from the site were collected. The combined data set was chosen to be large enough to be a significant challenge to visualize without being so large as to incur extreme computational penalties during the exploration phase.

We derive the Base Data from the link information extracted from WWW pages and from the usage log information collected by Web servers. WWW servers are instrumented to record relevant transactional information in the Common Log Format (CLF) [11]. For each item requested, the Web server logs the following information:

Server Log (CLF format)

```
{<Requesting host IP address or DNS name>,
  <The remote logname of the user.>,
  <User authenticated name>,
  <Time to nearest sec>,
  <name of object requested>,
  <HTTP version>,
  <HTTP method invoked>,
  <Bytes transferred>,
  <HTTP status code returned>}
```

While additional transaction information is now available since the advent of CLF, e.g., user agent, referrer, cookie, etc., our efforts concentrate only on the information contained in CLF.

The topology information is stored in a custom hyperlink database, where a complete snapshot of the topology is captured daily. The format used to store the topology is as follows:

Topology (custom)

```
Day1: {URL1: {<link> <link> ...}
      {URL2: {<link> <link> ...} ...}
Day2: {URL1: {<link> <link> ...}}
```

From the Base Data representation, we construct a CUT representation.

Topology

```
URL1: {day1: <link> <link> ...}
      {day2: <link> <link> ...} ...
URL2: {day1: <link> <link> ...}
      {day2: <link> <link> ...} ...
...
```

Usage

```
URL1: {day1: <freq>, day2: <freq>, ...}
URL2: {day1: <freq>, day2: <freq>, ...}
```

¹ Unfortunately we are not able to include a screen shot of the ASM's presentation of the Xerox Site for comparison due to a bug in the ASM's Web crawling spider that caused it to infinitely recurse the Xerox Site.

...

Finally, to support the visualization, we build an Analytic Abstraction by mapping the CUT structures into trees aggregated over time forming combined trees over all weeks, where each node has the following attributes:

{ <freq over 1 week>, <age of node>, <status- (added, modified, deleted)> <relation to other nodes> }

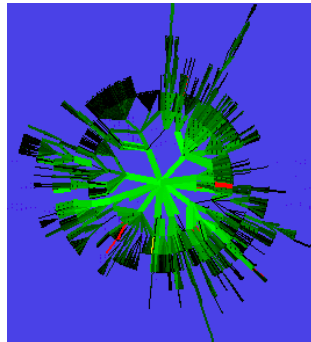
WEB ECOLOGY AND EVOLUTION VISUALIZATION

In analyzing the linkage structure of the Web, we are often interested in the shortest number of hops from one document to another. Breadth-first traversal transforms the Web graph into a tree by placing a node as closely to the root node as possible. After obtaining this tree, we then visualize the structure using the Disk Tree technique.

Instead of having a different layout for each Disk Tree, we use a combined layout for all trees. We do this by taking into account of all of the documents that ever existed in the entire time range of the time tube, then computing a single Disk Tree layout that is then used across all of the Disk Tree slices. This produces a layout that remains consistent across Disk Trees.

Disk Tree Visualization

The Disk Tree uses a circular layout to visualizing the hierarchy. Each successive circle denotes levels in the tree. The layout algorithm runs in two passes. In the first pass, the algorithm traverses the entire hierarchy using post-order traversal. At each node, the algorithm calculates the number of leaf nodes in that subtree. So we know the total number of leaves in this tree. We then calculate the amount of angular space each leaf node should be allocated (360/Total # of Leaves). In the second pass, the algorithm traverses the hierarchy using breadth-first traversal. At each node, it allocates the amount of angular space for that node by looking to see how many leaf nodes are rooted at that subtree. In this manner, each leaf node is guaranteed a fixed amount of angular space. The result is shown in the figure below. Each concentric circle is a new level in the tree.



We can gain increased understanding of visualizations by paying attention to the choices made in mapping data into visual presentations [5]. In our use of Disk Trees to visualize the Web site, the computed hierarchy is mapped (see Figure 2):

Analytic Abstraction	Visualization
Treelinks	→ a line mark in X and Y
Page access frequency	→ line size and brightness

Page lifecycle stage → color (new page = red; continued page = green; deleted page = yellow)

Because of the large number of nodes and because of its importance, page access frequency is double-coded. This has the effect of making the diagram visually simpler and allows the diagram to separate out three hierarchical levels of attention: (1) Documents added or deleted in a week standout highlighted in yellow or red. (2) Frequently visited parts of the Web site (thick bright green) form a noticeable foreground. Finally, (3) low frequency areas of the Web site seem to recede into the background.

The Disk Tree has several advantages: First, the structure of the tree is visualized compactly, with the pattern easily recognizable. Second, when viewed straight on or at slight angles, there are no occlusion problems since the entire layout lays on a 2D plane. Third, unlike Cone Trees, since it is a 2D technique, we can utilize the third dimension for other information, such as time, or a 3D glyph at each node. Lastly, the circularity is aesthetically pleasing to the eye.

The visualization itself actually validates the choice of our transformation algorithm—breadth first traversal. Notice that the high traffic areas are concentrated near the root node. This means that the algorithm places easy to reach nodes starting from the root node. As the document gets farther and farther away from the root node, the document has a lesser possibility of being accessed.

Time Tube Visualizations

Using Disk Trees, we can utilize the third dimension to represent time. In the Time Tube visualization, we layout multiple Disk Trees along a spatial axis. By using a spatial axis to represent time, we see the information space-time in a single visualization, thus facilitating easy sense making on the entire information space-time. Essentially in Figure 2 we are reusing the X dimension, once as part of the layout of the Disk Tree and again to represent the passage of time.

Treelinks	→ a line mark in X and Y
Time	→ an area on X

The mapping of multiple variables is mitigated if the Disk Trees are turned as in Figure 1. By making the Disk Trees be 2D in a 3D world, we get additional flexibility in the mapping at the cost of perspective distortions and lower readability.

Slices in the information space-time of Figure 1 are actually not laid out parallel to each other. Each slice is rotated so that it occupies the same amount of screen area as other slices. Because of perspective effects, if each slice were parallel to each other, then slices in the center would occupy smaller amounts of space than slices on the side. Also, we would see the front side of the slices that are on the left side of the viewing frustum, and the backside of the

slices that are on the right side of the viewing frustum. By carefully monitoring the viewing degree of interest, we can also emphasize certain slices, and de-emphasize others to get a focus+context effect.

Another interesting variant of the Time Tube is to stack the Disk Trees in the Time Tube and then fly through the tube, or similarly, to play the Disk Trees one after another in time order so as to create an animation of change. That is, instead of mapping time into space, we just map it into time.

Time → time.

This method is more compact, hence the Disk Trees can be larger, and it engages the motion detection capabilities of the human perceptual system. The detection of change and the interpretations of series of changes are enhanced at the cost of the ability to do comparisons between different points in time.

Interaction Model

The system allows users to interact with the visualization in various ways:

- By clicking a button, the system rotates all of the slices so that we are **viewing it head-on**, as shown in Figure 2.
- Clicking on a slice brings that slice **to the center focus**, thus allowing us to view that week's worth of data in more detail. That slice is also drawn on a transparent circular background, so the slices in the Time Tube are still visible.
- Using a "Flick-Up" gesture, the slice goes back into the Time Tube.
- Using a "Flick-Down" gesture, the slice becomes the floor (at a slight angle).
- Right-clicking a node zooms to the local area of that node to **show more detail**. Hitting the home key sends the user back the global view. This enables a "Drill-Down" operation that is of favorite among analysts. The rapid exploration of local patterns is of great interest to them.
- We can also use the mouse to poke around in the slices. When a mouse is over a node, **that node is highlighted in all of the slices**. In addition, a small information area shows the details on that node. This interaction is like brushing your finger through the Time Tube, seeing the detail of the point of interest.
- While poking around with the mouse, you can also instruct the program to notify the Netscape browser to bring up that particular page. Making this into a novel web-surfing tool!
- When the mouse is activated on a particular node, the 1-hop links are also shown using blue lines.

- Another button changes the point of view so that we are viewing straight down the Time Tube.
- We can also see an animation of each successive slice shown head-on. This maps the time dimension of the data into a time dimension in the visualization.

REAL-WORLD ANALYSIS SCENARIOS

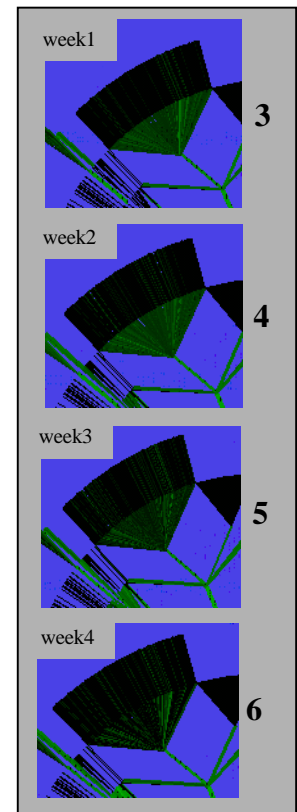
Given the ability to visualize usage patterns on web sites, analysts can now answer some interesting questions. We started this paper by motivating a number of questions that site administrators, content producers, and users may have about the behavior of a specific Web site. Given the space limitations, we will focus on the following questions with respect to the Xerox Web site, and answer them in order:

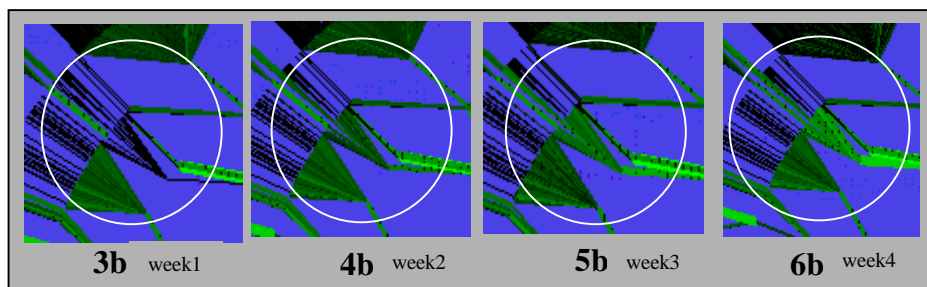
1. What devolved into deadwood? When did it? Was there a correlation with a restructuring of the Web site?
2. What evolved into a popular page? When did it? Was there a correlation with a restructuring of the Web site?
3. How was usage affected by items added over time?
4. How was usage affected by items deleted over time?

Deadwood

In our previous work [15], we motivated the notion of the time-course desirability of information. That is, information is subject to various patterns of faddishness across time. For example, while some pages remain perennially popular, the desirability of other information decreases over time, i.e., becomes deadwood. Content providers need to be made aware of the utility of their material. Popular information needs to remain accessible whereas information that has lost its appeal can be removed, relegated to cheaper storage devices, given a face-lift, or integrated into other content.

The wide fan in Figure 3a corresponds to a set of product safety information. During the first week, the partial green color of the area reveals moderate usage of the information. As time progresses (Figures 4a, 5a, 6a), the area becomes darker and darker revealing that **the material is becoming less and less popular**. While Time Tube does not inform us in this case **why** the material lost its popularity, it does **clearly identify the change** in usage and enables further exploration.





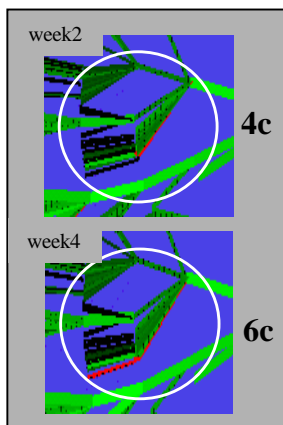
The Rise of Information

Fortunately, not all material becomes deadwood. Each year, Xerox publishes a Fact Book that contains all relevant product, service, and financial information for the current year. For 1997, the Fact Book was published online prior to the month of April but received very little attention from users of the site. This is shown in Figure 3b, where the black coloring of the pages in the Disk Tree reveals that the pages existed but were not receiving much traffic. Due to a modification during the second week within the site that helped promote attention to the 1997 Fact Book (Figure 4b), the area becomes noticeably greener indicating increased traffic to the area. In each successive week (Figure 5b and 6b), the Fact Book continues to become more and more popular and the corresponding area in the Disk Trees becomes greener and greener. In this case, Time Tubes enabled the quick verification that a minor redesign was able to successfully influence the traffic to an area of the site.

Addition of Items

The addition of new material into a document ecology competes with other items for attention. Therefore, one would like to know how the addition of new information affects global usage across time. It is possible that the new material is never used, or it may become wildly popular. Without an understanding of the evolution of the topology and the usage patterns, this determination cannot be easily performed.

Using the Xerox Web site Time Tube visualization in Figure 4c, we see the addition of a press release announcing a new family of Xerox products. In Figure 4, we see that the node is colored black in the first week, indicating the page does not yet exist. In Figure 4c, the red line depicts the introduction of the press release into the Web site. By brushing this area in Figure 4c, we are able to drill down and discover that usage in the third week skyrocketed from 1 access to 871 accesses. In week four (Figure 6c), additional red lines indicate

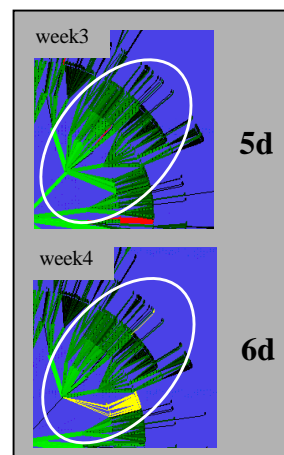


that more related press releases were added. From this example, we are able to understand the impact of a series of important press releases on the Xerox Web site. The addition of this material positively affected the usage, probably indicating a well-received product line.

Deletion of Items

Equally interesting as the incorporation of new material into a Web ecology is the deletion of material from that ecology. The deletion of material not only affects the content within the site, but potentially the navigational structures that enable users to get from one portion of the Web site to another as well. Clearly, the haphazard deletion of material can negatively impact the usability of a site. Both end-users as well as administrators can profit from being made aware of deletions.

The area displayed in Figure 5d corresponds to Xerox print drivers information. The green color and the width of the lines indicate that the material is well-used. The root of the branch serves as the navigational hub to pages in this area, and has a direction from the Xerox home page. During the fourth week, a different navigation path replaced the direct link from the Xerox home page to the main driver page. The Disk Tree in Figure 6d shows the link from the home page to the driver page as black, indicating no usage of that link. Despite this dramatic change in topology, usage of the driver area remained strong (notice the subtree remains green). The yellow lines in Figure 6d indicate that a subset of the driver information was deleted in the fourth week. From this example, we can see the deletion of material and navigation redesign did not negatively affect the overall usage of driver information.



GENERAL DISCUSSION

We have presented the Disk Tree and Time Tube as examples of Web Ecology and Evolution Visualization (WEEV) techniques. These techniques are designed to provide users with visual cues about patterns and trends in the relationships among content, topology, and use over time. There remain many limitations to the current system that need to be addressed. For instance, there is no simple method for presenting abstracts or summaries of node content—the current system just presents URLs. Despite these limitations, our scenarios illustrate that the

techniques reveal information that is not readily available otherwise.

The scenarios showed how it is possible to see the birth and death of information structures on the Web environment. Moreover, the scenarios showed how activity in the information environment often begets activity in the user community, and vice versa. Each adapts to the other, since the survival of information depends on the interest it receives, and lives of users are simplified by getting the best information at the lowest effort. The WEEV techniques presented here are aimed at enhancing our understanding of the patterns of dependency and adaptation among documents and use.

While it would be nice to have developed a system that is able to answer all of the possible questions dealing with the interactions between usage and structure of Web sites, we readily recognize several limitations of Time Tubes.

Since we used a global layout (laying out every node that ever existed during the time period), the documents and links did not shift from disk to disk. This made temporal comparisons much easier, but masked the movement of documents (for example, moving from /xps/driver.html to /driver.html). This movement is not directly apparent in this visualization because

1. The document always exists, and therefore has one representation in the global layout.
2. May continue to have similar usage patterns even after being moved, so its color brightness does not change.

As potential future work, we can imagine creating visualizations that show before-and-after views and how documents shifted around.

CONCLUSION

Given the infancy of the Web, it is not surprising that the interactions and relationships within Web ecologies are not very well understood. The visualization system presented in this paper pushes the capabilities of current Web analysis programs in the amount of data it is able to handle as well as making the evolutionary patterns of Web ecologies more apparent. As the World Wide Web continues to grow in the number of users and the number of documents made accessible, the problem of understanding the correlations between the producers of the information, the characteristics of the information, and the users of the information will most likely remain. While the Time Tube was able to address several real world analysis scenarios that are not possible with other systems, we look forward to expanding the capabilities of the Time Tube, and as a result, our understanding of Web ecologies and other time-dependent document ecologies.

ACKNOWLEDGMENTS

This work was supported in part by Office of Naval Research contract N00014-96-C-0097 to Peter Pirolli and Stuart Card.

REFERENCES

1. Accrue Insight™. <http://www.acrue.com> (1997).
2. Andrews, K. Visualizing Cyberspace: Information Visualization in the Harmony Internet Browser. In: *Proceedings of Information Visualization '95*. (1995).
3. Astra SiteManager™. <http://www.merc-int.com> (1997).
4. Botafogo, R.A., Rivlin, E. and Schneiderman, B. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142-180 (1992).
5. Card, S. K. and Mackinlay, J. The structure of the information visualization design space. InfoVis '97. IEEE Symposium on Information Visualization. (1997).
6. Carrière, J. and Kazman, R. Interacting with huge hierarchies: Beyond cone trees. In: *Proceedings of Information Visualization '95*. (1995).
7. Chi, E. H. WebSpace Visualization. <http://www.geom.umn.edu/docs/webboogl/>. The Geometry Center, University of Minnesota. (Also appeared in First WWW Conference, Chicago, 1994)
8. Hendley, R.J., Drew, N.S., Wood, A.M., and Beale, R. Narcissus: Visualizing information. In: *Proceedings of Information Visualization '95*. (1995).
9. Hughes, K. *getsites* (software no longer supported), private communication (1997).
10. Interse Market Focus™. <http://www.backoffice.microsoft.com> (1997)
11. Luotonen, A. The common logfile format. *CERN httpd user manual*. (1995)
12. net.Genesis. net.Analysis™. <http://www.netgen.com> (1997)
13. Pirolli, P., Pitkow, J. and Rao, R. Silk from a sow's ear: Extracting usable structures from the web. In: *Conference on Human Factors in Computing Systems (CHI '96)*, Vancouver, Canada (April 1996).
14. Pitkow, J. and Bharat, K. WebViz: A tool for World Wide Web access log visualization. In: *Proceedings of the First International World Wide Web Conference*. Geneva, Switzerland (May 1994).
15. Pitkow, J. and Pirolli, P. Life, Death, and Lawfulness on the Electronic Frontier. In: *Conference on Human Factors in Computing Systems (CHI '97)*. p. 383-390. Atlanta, Georgia. (April 1997).
16. Salton, G. *The SMART Retrieval System—Experiment in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, (1971).
17. Wise, J., J. Thomas, K. Pennock, D. Lantrip, M. Pottier, Anne Schur, and Vern Crow. Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents. In: *Proceedings of Information Visualization '95*. (1995)