Footprints: History-Rich Web Browsing

Alan Wexelblat, Pattie Maes
MIT Media Lab, Software Agents Group
E15-305, 20 Ames St., Cambridge, MA 02139
wex, pattie@media.mit.edu

Introduction

Browsing is hard. Web pages present a plethora of options and there are often no clear guidelines to help a person find her way among the many possibilities. Particular goals may be addressed, but it is nearly impossible to get an overall sense of the options and possibilities that might be found.

One way to help people deal with this problem is to make available the work of large numbers of people who have tried solving the problem beforehand. This past work can provide hints or clues to help the problem-solver. In essence, we allow new users to cooperate passively with past users. While the record of any one attempt at solving the problem is unlikely to be helpful, by looking at the collected patterns of many users we can see useful patterns emerge. This solution is often available in physical spaces because physical spaces and artifacts provide means of recording what has been done. For example, the campus at the University of Michigan has walkways which do not seem to follow any regular pattern, but which usually provide pedestrians with efficient routes between buildings. This is because the paths were not predesigned, but were first made on grass by people actually using the space. Designers observed the results of this behavior and put in walkways where people had actually trod. The designers cooperated passively with the users and the result is a much better design. In addition, subsequent users have the advantage of the prior work. No particular person’s set of paths is likely to be optimal for all subsequent users, but the collected solution is useful.

Footprints is a prototype system created to help people browse complex web sites by visualizing the paths taken by users who have been to the site before. These paths are shown as a graph of linked document nodes, with the links color-coded to visualize the frequency of use of the different paths.

The goal of this visualization is to give users a sense of the kind of experience they can have from a given web site before they enter. The resulting display turns out to be useful both to end users and to information designers, who know what information is being presented but lack ways of understanding how users are navigating the information.

This paper describes the Footprints prototype system and some of the issues in implementing tools for this kind of problem.

Footprints

The prototype consists of two main pieces: a front end and a back end. The back end runs in a batch mode, usually once per night, processing the web server logs. It produces a data file which is read in by the front end and which is used to create the Footprints site map. The front end is a Java applet which handles the display and interaction. The data flow in the system is shown in Figure 1. The back end is a pair of C++ programs which can be run on any computer by an administrator with access to the web server log files. The front end is run by any user visiting the Footprints-enhanced web site with a Java-capable browser such as Netscape Navigator or Microsoft Internet Explorer.
The back end uses web logs in the standard referrer format. An example of a log file entry in this format, with explanations of the various fields, is shown in Figure 2. Data from the web logs is cleaned — failed requests, image requests and such are removed — and validated to ensure that all URLs actually point to retrievable documents. The document references are linked into paths, and the paths are combined into a map. Most web servers simply append new information to the end of their log files, and it is up to the human administrator to determine when to clean out old data; likewise, Footprints will operate on whatever log data it is given; the site administrator must determine what data is too old to be useful. Footprints can combine previously-generated map data with new web log entries, or it can generate a fresh map.

As shown in Figure 2, we do not keep track of individual users' identities. While it is possible to say that a user followed a given path through the site, it is not possible in this system to say which user followed a given path. Additionally, the map display does not show individual data. The map, an example of which can be seen in Figure 3, contains only aggregate data for all users to the site. We believe that these steps are important to protect users' privacy. Without such privacy protections being designed into the system, users will become unwilling to be collaborative participants, even passive ones.

Footprints presents the path data map in a window separate from the main display. Document icons represent web pages and the directed links between pages are color-coded to show levels of activity: blue (cool) links are the least traveled, red (hot) links are the most used, and purple shades represent intermediate levels of use. Because the amount of data shown in the map can vary widely — for example, one map might represent a day's worth of use, another might show an entire month — the links only show normalized use levels rather than absolute numbers of traversals. The map of Figure 3 shows about five days' worth of use of the Software Agents' group web site.

The applet displaying the map can be used to direct navigation. Users can single-click on any node to reveal the title of the corresponding page. Single-clicking again hides the title, which helps prevent the display from becoming too cluttered, since titles of web pages are often quite long. A double-click brings the corresponding page up in the browser window; the most-recently selected document is colored black on the map as a sort of "you are here" reminder.

It is important to remember that — unlike many web-site visualizations — the map does not represent all the possible paths within a site, or even all the possible links a user could follow from any given page. Rather, the
map shows what people actually did in the represented site over the sample time. These two representations are often quite different. The Footprints map is usually simpler and less organized than the total site map. Nodes which were never visited are not shown at all in Footprints, and we expect browsing users to move more randomly through the site than any preplanned layout would suggest. Additionally, users may jump between pages which are not explicitly linked in the site simply by typing in a URL or calling one up from the browser software’s bookmark or history lists.

Creating a use map is an interesting display problem. Graph layout is well-studied (see [3] for a comprehensive review) but there are no good general solutions for laying out arbitrary graphs on a 2D surface. If the graphs are planar and/or highly connected there are good solutions; however, usage graphs tend to be both highly non-planar and highly disconnected. In addition, if we are making a map which can be used by newcomers to a web site, we have the problem of not knowing in advance who our users are likely to be. Without the ability to characterize the user, it is hard to say that one display method is better than any other.

For these reasons, we decided to create a self-organizing display. The Java applet which shows the map does not come up in a predetermined configuration. Instead, the graph is displayed in an initial random configuration, which sorts itself out into a more or less acceptable display via a modified Boltzmann algorithm. The Boltzmann algorithm [1] is usually used to model ideal gasses and works by computing attraction and repulsion forces among objects. In our applet, these objects are the document icons, which are attracted to other documents to which they have a link and repelled by documents to which there is no link. The trick to using a Boltzmann algorithm for graph layout is to balance the positive and negative forces so that nodes are drawn together and repelled in just the right ways so that they do not over-expand nor collapse, and so that the system does not oscillate.

For simplicity, we model our forces linearly, as shown in Figure 4, with the positive (attracting) force increasing with increasing distance and dropping off to zero when the nodes are "close enough." This ensures that the system does not collapse to a 'black hole.' Conversely, the negative force is strongest when nodes are close together and drops off as distance increases. This ensures that the graph stops expanding after a while.

If the algorithm does not produce an acceptable layout, or if the user wishes to view the graph differently, the nodes and clusters of nodes can be grabbed and moved. The layout algorithm then adapts to the user's input and adjusts the display to take the new configuration into account. A node that has been clicked on by the user (i.e. is displaying its title) is assumed to be of interest and is not moved by the algorithm. This allows users to control the focus and
separation of key nodes. Once the user is done viewing the nodes, she can click on them again to hide their titles and the graph relaxation algorithm will move them in accordance with the new configuration.

![Diagram of Boltzmann Algorithm forces](image)

**Figure 4 — Boltzmann Algorithm forces**

The resulting display is one which our informal test users have found to be quite successful. In particular, the display allows naive users to get a sense for what kinds of experience a web site can offer — a measure of the dynamic use of the information space. This makes a good complement to traditional measures such as hit counts. Figure 5 shows a portion of a Footprints map with three different kinds of browsing experiences highlighted. One — circled on the left-hand side of the map — is a traditional linear narrative experience where nodes are viewed in sequence: show me one thing, then show me the next thing, then show me another, possibly with brief diversions to get more detail.

Another experience — covering the middle of the map — is more like a storefront: richly interconnected and with no obvious order to the nodes, though some nodes clearly show stronger relationships to other nodes in the cluster. This is similar to a store where you walk in and are immediately confronted with a variety of departments and options with no obvious preferred path or sequence among them.

![Footprints Map Detail](image)

**Figure 5 — Footprints Map Detail**

The third experience — in the upper-right part of the map — is called "central message" because it shows a pattern wherein a user enters a cluster of documents and, after taking a look at a node one hop from the center, returns to the central node (probably by use of the browser's "Back" button). Like the other experiences shown here, this can be either good or bad, depending on the user's goals and the site designer's goals. If the central page is a contextualizing message which it is important to see periodically, then this kind of pattern can be quite good. This is why lecturers often put copies of their "Outline" slides at strategic points in the lecture, to remind listeners where they are and
what is to come. In other cases, this kind of display can reveal user confusion or indecision, indicating that it is probably not worth the user’s time to visit that cluster of pages.

The goal of providing the map in Footprints is not to rate the kind of experience which is shown, but rather to allow people to understand what the information space has to offer. This understanding should facilitate browsing.

Experience
Our experience with the prototype has been quite positive. We have not done any formal user testing and cannot make claims about improved browsing efficiency. However, our users’ feedback has been quite positive, with many requesting copies of the software to run on their own sites or suggesting features and enhancements they would like to see in the next version.

We first applied the system to our own web site because we believe that people who do not use their own systems are probably not serious about their research. In doing so, we were both pleased and surprised to find a bug in our web site. This “bug” consisted of a popular page (our research description page, highlighted black in Figure 5) to which many users went but from which no users departed. This was a surprise, since that page contains a large number of links to individual research projects done by members of our group.

In looking at the research description page itself, we quickly realized our mistake — the links to other projects were buried well down the page and the top of the page did not make clear that links could be found lower down. Before using Footprints, this all-too-common mistake had been completely invisible to us. Since the individual research project pages were getting hits from other sources, we had no way to know that the research overview page was not doing its job.

We also found an unexpected class of users among our initial testers: information designers. This class of users knows what information is being presented by the space, but lacks tools for understanding how that information presentation is perceived by outsiders, particularly those new to the space. Simple hit counts can give information designers a beginning but they do not help the designer understand why her presentation is or is not effective. A Footprints-style map permits designers to do source and sink analyses, which can be extremely useful for detecting effective or ineffective information designs.

For example, Figure 6 shows the same display as Figure 5 but analyzed from an information designer's point of view. Again, three kinds of experience are highlighted. The first — in the lower-center of the map — is the common error described above, where the popular page serves as a sink when in fact it should be a gateway to a whole host of other pages. The second is the in-out pattern in the upper right corner. As noted above, this can be either good or bad. If the central page is a corporate home page and the adjoining pages are the lead pages for various product departments, then this configuration can be quite troubling.

The third grouping — in the center of the right-hand side of the map — was called by one of our users the "Advertiser's nightmare." It was so named because it shows a cluster of external pages which have led users to a page at our server. These external pages include search engines like Yahoo and sites with advertisements for and links to the central page. In this case, the advertiser has done her job: the page is publicized and the advertisements are drawing people to the site. However the page itself is serving as an information sink — people are not being drawn farther into the site, an alarming sight for an advertiser. In another situation, this configuration could be much more pleasing; for example, if the surrounding pages were product descriptions and the central page was the place where customers entered their credit-card numbers to place orders. In either case, it is the way in which Footprints shows the use of the space which allows these kinds of determinations to be made.
Related Work
The major inspiration for this project has been the idea of using history information to help people deal with information spaces. A similar idea was promoted by Hill and Hollan [4] in their Bellcore project on history-rich digital objects, from which we took this paper’s title. Their writings give excellent rationale for the notion of objects containing and revealing their use histories. Hill and Hollan make strong analogies between computer files and physical books such as parts catalogs, pointing out the rich history records borne by the latter which are absent in interaction with computer-based versions of the same information.

The Bellcore group led by Hill and Hollan [5] built a series of prototypes (Edit Wear, Read Wear, Email Wear, Source Code Wear, and Vita Service) to demonstrate the advantages of using objects in the context of their past interaction histories. These prototypes were aimed at changing the way people did various document processing tasks, such as collaborative creation of papers, source code editing, email processing, and so on. None of their tools were explicitly aimed at the problem of information navigation in complex spaces, but the relationship of their work to Footprints is quite clear.

An interesting ongoing project is called Aqui [2], from IBM. Aqui presents a front end to a database populated by links which Aqui users have added in a sort of competitive environment. The basic thesis of the project is that no reasonable page can contain all the links to relevant pages. Even assuming that a page's author could somehow keep up with the rapidly growing Web, the sheer number of possible links would overwhelm the information on the page.

Aqui also points out that certain kinds of links are made for certain purposes, based on the underlying theories and biases of the person doing the linking. For example, a site propounding a particular point of view would be likely to be linked to other sites with a similar or sympathetic point of view. However, the author of such a site would have little or no incentive to link his site to pages which put forth an opposing point of view. The database maintains links suggested by users and keeps rankings of which suggested links are the most popular; unfortunately, because the link database is on a separate site, its data is not available while the user is browsing.

Aqui is that it attempts to characterize links by asking users for suggestions as to what kind of user would find the link useful. Aqui supports a limited set of stereotype users such as "business person" or "knowledge worker." Using these stereotypes and the user's self-description, the system attempts to rank the suggested links so that the
most useful suggestions are presented first. Of course, this suffers from the usual limits of stereotypes: no characterization is ever complete and users' interests shift, often rapidly.

The situation on the Web today puts almost all the power in the hand of the information designers. They control the look, feel, and content of the information. Readers, though, are not passive consumers of information. One of the effects of Aqui is to give some measure of power to the user of the information, a goal it shares with Footprints.

Future Work
Given that Footprints is a prototype system built to show that history visualization can be useful for information browsing, we feel we have accomplished our initial purpose. The research project of which Footprints is a part is continuing, and we are pursuing improvements in several directions.

Display
Although the display produced by the Java front end has a number of advantages, we do not claim that it is in any way the optimal or ultimate display technique for this information. Rather, it is a first attempt to show the information in a way that does not interfere with or change the display of the actual web pages themselves. An obvious alternative to this "off-board" display would be to embed the use information in the web pages themselves, perhaps by changing the appearance of the links in a recognizable way or by inserting information next to the link on the page.

The problem with "on-board" displays of this type is that they cannot be done in a way which is generally applicable. Display properties such as font size and font color may be used by the page designer to convey information already. If the Footprints system was to attempt to use these display properties, it might end up destroying the information it is trying to augment. Likewise, the insertion of additional text or icons into a page risks destroying the alignment or spacing on which the information display may depend. Finally, many cross-document links are not immediately visible on the display: they are hidden in imagemaps or Java applets, for example.

Given these constraints, an off-board display seems to be the only generally practical solution, unless you know in advance that the information designer will not make use of certain visual properties and HTML techniques. At this time we are investigating other options for off-board displays. It is likely the next version of the system will support at least two different interfaces, one tailored for newcomers to the site and one tailored for information providers who know the structure and content of the site.

Showing paths
To some extent, the "storefront" type experience described above is due to the fact that the current Footprints prototype shows only aggregate data — the activities of many users. For any given user, it is possible to extract a path (or set of paths) followed through the site. If a reliable way can be found to classify these paths, it may be more helpful for a new user to look at a relevant set of paths rather than at aggregate data for the site as a whole. We are exploring ways to extract the paths followed by individual users and clusters of users, without violating the privacy of users.

Interaction capabilities
Even though we strongly distinguish browsing from search (see [6] for discussion), search techniques can also be useful for people who are browsing. Additionally, document titles are currently not highlighted in the Footprints demo. Our test users found titles to be quite helpful, so we plan to make the page titles more available and more centrally a part of the interaction mechanisms.

Finally, we anticipate extending Footprints' capabilities from purely passive collaborations to ones in which users may be free to take an active part. It is one thing to notice that a page in a book has a number of notes in the
margin; reading the content of those notes is an entirely different level of collaboration. We would like to make it possible for people to annotate their footprints with information on what was good and bad, and what purpose following a given path served. These annotations would become part of the public record, available for later users to see and take advantage of, as well as adding to. This will help Footprints move from being a passive cooperation environment to a more active one.

Summary
We built a prototype system called Footprints in an effort to find a way to help people browse complex information spaces. This browsing is an instance of a problem for which few good solutions exist and little expertise is available to help new users.

Footprints makes records of past user activity available to new users, in the form of an active, Java-based, graph visualization. This visualization is adaptive and interactive, allowing users to customize the graph view. It also serves as an interaction tool for viewing the actual web pages as the user browses the site. This tool was informally found to be useful for new users and, unexpectedly, for a class of experienced users: the information designers familiar with the site.

Both classes of users benefited from Footprints’ ability to provide a visualization which is complementary to the usual quantitative measure such as hit counts. Simply knowing how often a page is visited is often enough. In order to get a sense for the kind of experience a user is having, it is important to know the sequence in which things are seen, what pages encouraged further browsing, what pages were discouraging, and so on.

Although the prototype is primitive and clearly can be improved a number of ways, it has still proved useful, including helping us find a glaring error in our own web site. Test users have consistently shown interest, which has helped us validate our innate belief that we are on the right track.

References


