Effectively Exploiting Server Log Information for Large Scale Web Sites

B. Wong G. Marsden

Collaborative Visual Computing Laboratory,
Department of Computer Science,
University of Cape Town, Rondebosch 7701, South Africa.

Abstract

With the continuing growth of the Internet, an increasing number of organisations are incorporating the Web into their business activities. The appeal of a site to users in terms of both attractiveness and usability determines whether it will improve profits or cause on–line failure. It is thus vital that web site designers have access to tools that will aid them in evaluating site usage so that they can identify problem areas and alter them accordingly. At present, the most popular tools utilised in this evaluation make use of a technique called log file analysis, a process by which server log files are parsed to extract information about visitors to a site. However, when visualising this information, current tools have either neglected site structure or else failed to utilise properties unique to web sites. We address both these issues by developing a visualisation of server log data that incorporates site structure and caters specifically for web sites by exploiting their unique characteristics.

Keywords: *World Wide Web, Log File Analysis, Information Visualisation*
Computing Review Categories: *H.5.4*

1 Introduction

The role that the World Wide Web (WWW) plays in day-to-day activities continues to grow rapidly as both individuals and organisations realise the importance of maintaining an effective “web presence”. Whether one’s purpose is to simply publish information or else to sell some service online, it is vital that one’s web site competes favourably with the increasing number of new sites being created. The audience targeted by web sites can afford to be both discerning and demanding as they possess an abundance of choice. As such, any web site that fails to properly address their needs, or meet their expectations, is never visited again resulting in the site’s failure.

Thus, in order to remain competitive in the online world, constant evaluation and improvement of a web site is required. The only economically obtainable information about a visitor available to a site designer is the trace of the pages they accessed. Web site evaluation therefore mostly takes the form of a technique known as *log file analysis*. This approach involves the parsing of web server access logs to glean information concerning the usage, or browsing patterns and behaviour of visitors, of a particular site.

Unfortunately, garnering useful information by viewing the log files in their raw format is rather impractical. This is due to both the unstructured character of the unaltered log file format, as well as the sheer size to which log files of modern sites are inclined to grow. To alleviate this problem, web designers and researchers turned to the field of *information visualisation*, which exploits the human visual system by making use of graphical representations to provide insights into vast amounts of data [8]. As such, a large market for visualisation tools that assisted in web site usage evaluations arose. This led to the appearance of a plethora of commercial products, such as those reviewed in [9]. Most of these tools analysed log files and visualised the results using simple graphics such as tables, pie charts and histograms. While such graphs offer useful information, by their very nature they provide specific types of insights, which may be restrictive. In particular, they suffer from limitations with regards to imparting information about the actual *structure* of a web site. This is significant since a visitor’s path, and hence behaviour, could be said to be determined by the site’s structure, as their navigation is confined to predefined routes by the existence or absence of links between the various pages comprising the site. In addition, intimate knowledge of the structure of a site is vital in correctly analysing log files [1].

There have been a few projects developed that do incorporate site structure in their visualisations of site usage, such as [7][4][3]. However, these efforts tend to approach the problem by treating a web site as a normal tree or directed graph, thus failing to take advantage of properties unique to web sites.

This paper describes our ongoing research regarding the visualisation of web site server log information. We aim to create a tool that offers improved insights into users’ needs than those provided by current log file analysis products, by visualising the underlying structure of the site in addition to its log statistics. In developing our visual representation, or *metaphor*, we intend to take advantage of those attributes that are unique to web sites.

The remainder of the paper is outlined as follows. Section 2 provides some background information about log file analysis. Section 3 then describes those factors that affect the effectiveness of a web site usage visualisation. This is followed by Section 4, which describes common features which are unique to web sites and which we therefore wish to utilise in our visualisation. A description of our resulting metaphor is then provided in Section 5. After this, a discussion regarding our metaphor is presented in section 6. Finally, Section 7 concludes the paper with a description of possible future work.

2 Log File Analysis

Log file analysis is at present the most widely used means of determining web site usage (there were over 50 commercial and freeware products available in 1999 [2]). This popularity is likely due to the economy and ease with which log information concerning users' browsing patterns can be obtained.

Each time a user visits a site, a connection is established between the web server on which the site resides and the client browser of the user. Whenever the user wishes to view a page from that site, which they indicate by selecting the appropriate link, their browser sends a request for that page to the server. The server then records this request in a text file, known as a log file, along with various information about the client that made the request and whether the request was successfully granted.

Log files contain a rich set of data that when compiled and combined in various manners can provide statistics describing the usage of a site. Statistics that one can derive for certain from log files include:

- the number of requests made, which are commonly referred to as hits. This may be ordered by HTTP status codes (eg. successful,not found, etc), by type of file (eg. HTML document, JPG image) or by domain suffix, which are derived from IP addresses,
- the distinct IP addresses served and the number of requests each made,
- the number of requests for specific files or directories,
- the number and size of files successfully served,
- the URL's of the referring pages from which a user came,
- the browser type and version making the requests, and
- the totals and averages for a specific time period.

Log file analysis does suffer from a number of weaknesses, though. For example, certain data, such as individuals' identities, is not logged or is inherently incomplete. In addition, the increasing use of caching imposes several difficulties. The introduction of cookies did overcome many of these problems, although their use presented other concerns such as privacy and security issues. However, since it is possible to make reasonable estimates of missing data by using heuristics, log file analysis remains a useful technique even without the use of cookies.

3 Factors Affecting Web Site Usage Visualisation

Before we designed our visual metaphor, we first needed to decide what exactly makes an effective web site usage visualisation. In other words, we needed to identify those issues that a site usage visualisation must adequately address in order to be useful to users. This would include, at least, the following factors:

- *Structure* – This involves the choice of arrangement chosen to depict the structure of a web site. The user should be able to clearly perceive the site structure in a manner that is consistent with their knowledge of the site.
- *Variable Representation* – This concerns the manner in which variables (such as page accesses) are encoded in the visualisation. Users should be able to readily ascertain the values of different variables. They should also be able to identify interesting patterns in the data.
- *Maintaining Context* – With the size of large web sites, users are unable to survey an entire site in great detail at the same instant. Instead, they view only subsections of a site in great detail. It is important however that in doing so, they have the means of keeping track of the relative position of the area of the site they are currently viewing in context with the overall site.
- *Scalability* – This refers to how well the visualisation scales with size. When viewing large sites, users are likely to suffer from information overload if they are presented with too much information at one time. This needs to be avoided. In addition, users should be able to perceive as much of the site at one time as possible. Therefore *cluttering*, whereby certain items of the visualisation obscure others, needs to be kept to a minimum.

Now that we have outlined those factors that are critical to a site usage visualisation, we must look at those features unique to web sites that we wish to exploit, if we are succeed in our aim of take advantage of those attributes. These features are discussed in the next section.

4 Interesting Web Site Features

Every organisation strives to create sites that are interesting and original. However, in order not to confuse or alienate visitors, sites often contain common features that are familiar to web surfers (Jakob Nielsen advocates adopting the same approach as others due to the “Law of the Web User Experience” [5]). From a visualisation perspective, these features are of interest as they can be taken advantage of to create improved metaphors that cater specifically for web sites. Examples of exploitable features include:

- *Organisational Homepage* – Sites are devised with the intention that visitors will enter the site via the homepage. The majority of site designers therefore use this initial page to organise the contents of the rest of the site. The links leading off the homepage then serve to divide the site into various sections of interest such as products, people, etc.

- *Global Navigation Menu* – To aid users in navigating the site, a large number of sites include a navigation bar, or menu, that contains global links. These bars are often present on the majority of pages comprising the site (normally situated on the left edge of a page [6]) allowing users to access the major links from any page.

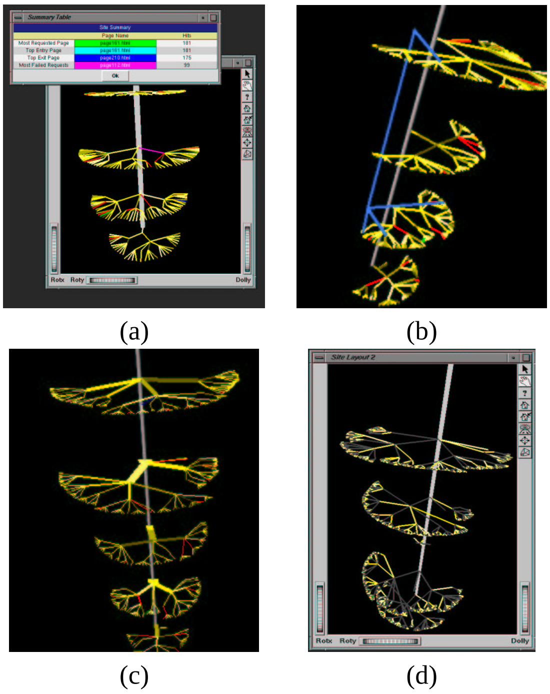
- *Self-Contained Web Sites* – It is not uncommon for several smaller web sites to be contained within a larger one. Examples include sites such as personal sites enclosed within an institution’s web site and individual sports sites within a general sports news web site.

Once those features we wished to exploit were identified, we could commence on developing our visualisation. The resulting metaphor is described in the following section.

5 Metaphor Description

Our current metaphor (Figure 1) is a three dimensional structure that consists of a vertical column with flat branches forming “fans” at different heights coinciding with cubes placed along the column. The column represents the global navigation bar with the cubes along it representing the pages that the links on the global navigation bar lead to. The fans then correspond to pages accessible from those navigation bar pages with lines portraying the links to the pages in question. Each page is understood to link to any of the navigation bar pages, unless the link leading to that page is portrayed by a dashed line instead of the normal solid one. Any outstanding links that are not displayed are then shown on request by lines proceeding straight to the target page if the two pages are only one level apart, or else flow to the back of the structure and from there to their destination (Figure 1b).

Whenever a self contained site is encountered, instead of being portrayed as a normal page it is indicated by a smaller vertical column representing its own nav bar. By definition, all pages on the fans of this mini-site are no



Summary Table

Site Summary		
	Page Name	Hits
Most Requested Page	page161.html	181
Top Entry Page	page161.html	181
Top Exit Page	page210.html	175
Most Failed Requests	page112.html	99

(a) (b)

(c) (d)

Figure 1: **Metaphor Screenshots** (a) Table of important site statistics and corresponding link colours. (b) Links not already implied or displayed are shown passing through behind the fans. (c) An option is provided that alters the line width according to the variable value of the connecting page and all the pages accessible from it. (d) Users can filter out pages containing values outside the user-defined range. These pages are then displayed in grey.

longer connected to the global nav bar of the main site as the fans are joined to their own nav column.

Data pertaining to a particular page is encoded in the line corresponding to the link to that page. The intensity of this link indicates the value of the current variable (default is page hits) for the associated page. The pages themselves are represented by varying icons, depending on the distance from the camera to the page. These range from no icons at all (furthest distance) to different shape (cylinder,cube and cone) icons representing different page types (medium distance) to different shape icons with numbers indicating the range of the values of other variables (e.g. failed hits) for a page (closest distance). In relation to this, when the camera is far out, branches of links are approximated by a polygon, the intensity of which indicates the average value of the current variable for pages comprising that branch (Figure 2).

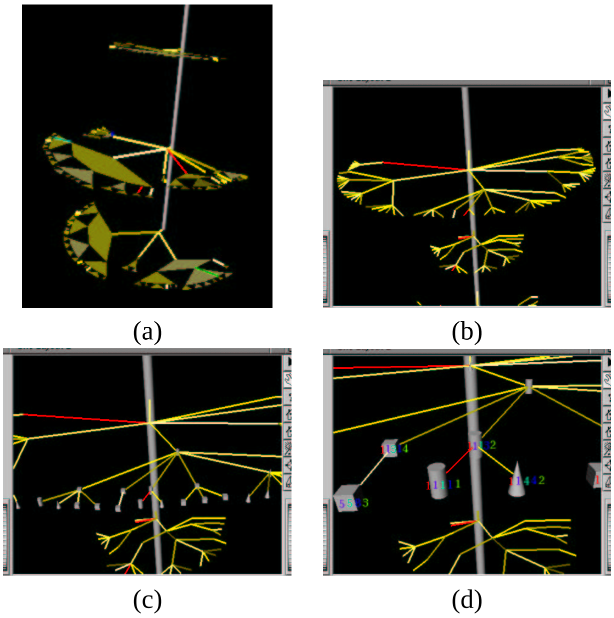

(a)
(b)
(c)
(d)

Figure 2: **Level of Detail** (a) When the camera is the furthest individual branches are approximated by polygons. (b) Moving in closer results in the polygons resolving into individual lines representing links. (c) Zooming even closer causes icons for the individual pages appearing. (d) At the closest level, numbers indicating the range of various statistics for individual pages are shown.

Certain important site statistics, such as the most popular entry and exit pages, are presented in a table (Figure 1a). The links leading to those pages which are included in the table are coloured to match the corresponding cells of the table.

There is also a line-drawing and a filtering option. The user is able to alter the line type so that lines or branches possess varying width, depending on the value of the variables of the pages belonging to that branch and the branches following from it (Figure 1c). The filtering tool allows the user to specify a value range of interest and greys out all those pages which fall outside this range (Figure 1d).

Finally, when navigating, the user is able to zoom, pan and rotate in one window while observing the entire site in a second window, which indicates the zoomed section using a “zoom box”.

6 Discussion of Metaphor

We will now discuss our metaphor according to the factors outlined in section 3.

6.1 Structure

The vertical bar and fan structure used has numerous benefits. First of all the user is always certain of which links lead to which pages, as implying links instead of displaying them avoids a lot of cluttering and has the effect of preventing links crossing each other. Cluttering is further reduced by the fact that the fans are only half circles, so that links that have been neither depicted nor implied can be shown by lines leading behind the semi-circles instead of by lines passing through the fans.

The user is also able to differentiate between pages belonging to a “subsite” included in the main site and pages belonging only to the main site, due to the 3D nature of the structure.

Futhermore, users are able to correlate their knowledge of a web site with the visualisation of its usage information. This is due to the emphasis placed on the global navigation bar, and the splitting of the site into sections or fans according to the pages comprising the navigation bar, which ties in to our aim of accentuating and utilising web site features. The navigation bar is present on most pages and is likely to be the most used means by which users travel to various parts of the site. The links present on the navigation bar divide the site into various sections of interest such as people and products etc. Separating these areas in the visualisation makes sense as users would not necessarily want to determine the usage of the people section relative to the unrelated product area.

6.2 Variable Representation

We propose that there are several advantages to the manner in which we chose to portray variables. Encoding page information in the link leading to that page, instead of in the page icon, enables the user to determine this information without requiring to zoom too much, as the user can discern the link line at distances where they can no longer view the page icon.

Users also do not have to contend with the additional overhead of remembering which colour represents a higher value, which they would have experienced if we had utilised hue instead of intensity to indicate variable values. In addition, by activating the option of varying the line width the user can identify important sections of the site as s/he is then able to view information pertaining to, not just the page connected by a particular line, but to all the pages accessible from that page as well. Thus the user is able to view information on *areas* of a site as opposed to only individual *pages*.

Finally, the user experiences no ambiguity about the relative values of a variable for different pages, as by looking at the number representing that variable on the page icons they will immediately know whether one page's data falls into a greater range than another's. This would not be the case if they had to differentiate between some scaling geometrical body such as say, bars, as the distortions (both perceived and real) due to relative distances would complicate such comparisons.

6.3 Maintaining Context

The user is able to keep track of the location of the subsection of the site currently of interest relative to the rest of the site due to our use of two windows. When the user zooms in and out, s/he can follow the camera's movements due to the smooth animation of the result of their manipulation. A user is thereby able to maintain context of where and how far they have moved.

6.4 Scalability

By using a vertical column for the navigation bar pages, we are effectively removing one level of the hierarchy. Considering that the last level of a hierarchy generally contains the most pages we save a lot of screen estate. In addition, by splitting the site into various fans we reduce the number of pages shown in each fan. Thus, if there are five pages leading from the home page, we have five fans in which to split up the pages comprising the site.

When viewing a particularly large site, users might have difficulties in determining the variable values of a branch, as the individual lines showing the links comprising the branch all "blur" together. However, users are given an indication of these values by the polygon approximation of a branch. They are thus allowed to efficiently choose those branches of interest to them.

Finally, users can ignore all the areas that are not of interest to them by filtering them out with the filter tool. The amount of information overload they suffer is thus reduced as all unnecessary information is greyed out.

7 Conclusions and Future Work

At present our metaphor lacks any provision for time. Instead it represents a particular snapshot in time of a web site. As web sites are highly evolving and constantly changing entities it would be desirable to view any alterations and the manner in which they affected site usage. One possible method of providing this would be through the use of animation. A collection of data taken from a specific period could then be stepped through while the changes in the data are animated from one value to the next.

Of more immediate concern to us is a proper evaluation of the current metaphor. We plan to perform this evaluation by designing a set of user experiments and then carrying them out with an assortment of users in a controlled environment. Their performance in the given tasks should then provide us with a clearer idea of whether we succeeded in our aims of creating an effective site usage visualisation.

Log file analysis performs a vital role in modern online business activities. By utilising effective web site usage visualisations site designers gain an improved understanding of how their sites are being navigated and thus obtain a clearer insight into the manner in which people browse sites in general. This knowledge empowers designers to create better, more efficient web sites, which ultimately benefits all who conduct their business through the World Wide Web.

References

[1] S Haigh and J Megarity. Measuring web site usage: Log file analysis. http://www.cranfield.ac.uk/docs/stats, 1998. Information Technology Services, National Library of Canada.

[2] H Hochheiser and B Shneiderman. Using interactive visualizations of www log data to characterize access patterns and inform site design. ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/pdf/99-30.pdf.

[3] Mercury Interactive. Astra sitemanager. http://www-svca.mercuryinteractive.com. Version 2.0.

[4] Microsoft. Microsoft site server 3.0. http://www.microsoft.com/siteserver/.

[5] J Nielsen. Do interface standards stifle design creativity? http://www.useit.com/alertbox/990822.html, August 1999.

[6] J Nielsen. Jakob nielsen's alertbox. http://www.useit.com/alertbox/991114.html, November 1999.

[7] J Pitkow and K Bharat. 'Webviz: A tool for world-wide web access log visualization'. Proceedings of the First International World-Wide Web Conference, Geneva, Switzerland, (May 1994).

[8] G Robertson, J Mackinlay, and S Card. 'Cone trees: Animated 3d visualizations of hierarchical information'. Proceedings of the Conference on Human Factors in Computing Systems CHI'91, pp. 189–194, (1991).

[9] Various. Web site analysis tools review. PC Magazine, May 2000.