USING ACCESS INFORMATION IN THE DYNAMIC VISUALISATION OF WEB SITES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE,
FACULTY OF SCIENCE
AT THE UNIVERSITY OF CAPE TOWN
IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Bryan Wong
November 2004

Supervised by
Gary Marsden

Abstract

Log file analysis provides a cost-effective means to determine web site usage. However, current methods of displaying log analysis results tend to be limited in that they either contain no reference to a web site's structure, or else they portray this structure as a standard graph or tree. This dissertation presents a visual representation of web server log information, which addresses these limitations by incorporating log file data into a visualisation of a web site's layout.

The devised visualisation utilizes properties unique to web sites in order to create a compromise between the clutter-prone network graph and the information incomplete tree representations that have traditionally been used to depict web sites. As such, the visualisation emphasises typical web site features such as the home page, sub-sites and navigation bars. This approach permitted the introduction of the concept of implying the presence of links without explicitly rendering them. This notion has many implications, not least of which is the reduction of cluttering. The visualisation combined several other techniques to address the issues of structure and data representation, data exploration, scalability and context maintenance.

Assessment of the visualisation consisted of a heuristic evaluation by an expert from the web site usage industry, a test to determine the intuitiveness of the representation, and a series of user experiments. Results of the assessment were generally promising although a few areas of concern, such as the difficulty experienced by users in navigating the visualisation with a trackball, were identified. These issues should not prove to be too difficult to overcome however. The visualisation could thus be said to have successfully met the aim of developing a representation of web site usage information that incorporates site structure and treats web sites as unique entities, thereby taking advantage of their particular characteristics. It is hoped such a visualisation will be of benefit to web site designers and administrators in analysing and ultimately improving their web sites.

Acknowledgments

There are many people who provided me with aid and encouragement throughout the course of completing this dissertation. In particular, I wish to thank the following: my supervisor, Assoc Prof Gary Marsden, for his guidance, motivation, inspiration and most all patience; my parents and my sister Andrea, for all their love and support; and finally the entire "gang" in the CVC lab for the friendships and many happy memories which I take away with me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This research concerns the visualisation of the usage patterns of a web site. In particular, focus is given to visualising web site access information from the perspective of a web site's structure as well as designing a visualisation that specifically takes advantage of properties unique to web sites. After an investigation was conducted to identify and explore factors that influence the visualisation of web site traffic, a visual representation, or *metaphor*, was developed. The metaphor was then evaluated to determine how effectively it addresses those factors, and thus how successfully it achieves its goal of promoting better understanding of the manner in which users navigate a web site.

This chapter motivates the need for web site evaluation and introduces the topic of visualising web site access information in Sections 1.1 and 1.2 respectively. Section 1.3 then presents the aims of this dissertation while Section 1.4 discusses the approach adopted in conducting this research. The chapter is then concluded by Section 1.5, which provides an outline of the rest of the dissertation.

1.1 Analysing Web Site Usage

The advent of the World Wide Web (WWW) has resulted in radical changes to the way in which information is exchanged. With this new mode of communication presenting opportunities to reach an ever increasing audience, new web sites, for both private and commercial use, are continuously being created. This is true to such an extent that most businesses now consider it to be imperative to maintain a web "presence". However, to date there exists no set method or formula for producing a web site that effectively utilises the full potential of the medium. Although there is literature available that discusses web site design, such as [37], following the guidelines laid out in these publications is still no guarantee of creating a successful web site. As a result, creating a lucrative web site has become an exercise in personal judgement and experimentation.

As long as this situation remains, web site designers have little option but to continually redesign and analyse their site until they are satisfied that it is a success. There are many criteria for defining success. One good way to gauge it is by examining the visiting patterns of users browsing the site, and then comparing the results to the expected or ideal patterns that the site designer anticipated.

With the escalating number of web sites competing for attention, possessing a web site that fails to address its intended audience is disastrous. This is particularly vital for commercial sites such as those that offer services or sell products on-line. To these sites, those users who are unable to navigate the site to their satisfaction represent a loss of revenue, as they are transformed from potential customers to dissatisfied, confused browsers. In addition, often the only residual information about users who have left a site is the trace of the pages they accessed. Thus the traces of users who have an incorrect idea of how the site should be explored distort the statistics about which pages are popular or correlated. This can lead to inaccurate conclusions being drawn on product or service interest. A poorly designed web site displaying characteristics such as being misleading, unattractive or difficult to navigate, is not serving its purpose and results only in a further loss of potential clients. Thus, investigating user access patterns becomes an essential task if one is to gain any benefit from one's web site.

Investigating access patterns has led to the emergence of a new field of research, namely *web site usage analysis*. Motivated by the desire of organisations to determine whether the investment they made in setting up their web page is providing satisfactory returns, web site usage analysis offers an idea of the extent of a site's usefulness. If an analysis of a particular site indicates that the users are not able to retrieve the information they require, then that site is not achieving its purpose and needs to be redesigned.

Aside from providing an indication of whether the site design is effective or not, web site usage analysis also aids in the understanding of how people browse a web site. Taking advantage of this knowledge may someday lead to the formulation of well-defined guidelines which will aid designers in creating better, more effective web sites.

1.2 Visualising Web Site Traffic

In order to improve a web site its current usage must first be evaluated. The question now arises as to what form this evaluation should assume and what tools will be required.

1.2.1 Evaluating Web Sites

Although the technique of monitoring visitor patterns or traces was mentioned in the previous section, this is by no means the only, or even the best, manner of determining a web site's usability.

Preece et al. define evaluation as being

"*concerned with gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment or work context*" [41].

Thus, if this definition is taken into account, to correctly evaluate a web site firstly requires the assembly of a collection of test subjects that is representative of that site's targeted audience. Next, the expected activities of such a group need to be identified. Then, once the test participants have been placed in a suitable environment, i.e., one that mirrors their typical working situation, their behaviour and actions while partaking in these activities should be monitored. Additional feedback in the form of subjects' opinions and comments may also be obtained. Finally, the results can then be examined and conclusions drawn. Such an approach was used by Eighmey [14] during his experiments on users' responses to different commercial web sites.

However, although this may be the ideal solution in that it is most likely to generate results of greater accuracy, its cost can be prohibitive. The resources required to establish a suitable testing environment, as well as to gather a sample of the population and to ensure they are suitable test subjects, prevents most parties from being able to evaluate their web sites in this manner. As a result, most individuals and organisations make use of a method known as *log file analysis*, which is more economically viable, in terms of both time and capital.

1.2.2 Visualising Web Server Log Statistics

Every web site is hosted by a server. This server automatically records activities that have taken place on the site in the form of a log file. Thus any traces left by users who visited the site are contained in this file. Detailed analysis of the log file can lead to valuable information concerning visitors' paths through the site, and hence the site's effectiveness.

Discerning users' particulars from a log file's raw format, however, is not always practical. This is due to the difficulty in identifying relevant details from the clutter of the log file format. Also, considering that log files for highly active sites can grow to several gigabytes per day, the volume of data can be overwhelming. One solution to this problem lies in the area of *information visualisation* [20][42]. Through the use of graphical representation, visualisation exploits the human visual system to provide insights into large and potentially confusing amounts of data. Most of the web site log analysis tools available today visualise log data in the form of two-dimensional tables, histograms and pie charts. These representations are useful and are certainly adequate for providing an approximate idea of a site's usefulness.

However, Mulvenna et al. [33] point out that there are three facets of a web site that play a role in it's ability to provide the intended service to its users. These are the actual content of the site, the layout of the individual pages, and finally the structure of the web site itself. The structure of a

web site is determined by the presence of links between the various pages the site consists of. The structure therefore restricts navigation through the site to paths predefined by the site's designer and thus determines the ease with which users can access relevant pages. If a site designer's perception of users' needs is inaccurate, then the designer's idea of the manner in which the site should be navigated is likely to differ to the users'. This can have a negative impact on the users' satisfaction and thereby harm the site's effectiveness. A web site's structure is thus an important consideration when determining its usage.

By only using simple graphics such as tables, pie charts and histograms, most log analysis products therefore do not provide site designers with all the information they require in order to correctly analyse and improve their sites. Bar charts and tables by their nature offer particular types of insights, which though useful, are limited, especially in terms of incorporating site structure. This is significant as intimate knowledge of the structure of a site is vital in correctly analysing log files [17]. In addition to their general lack of interactivity, these log analysis tools also provide poor support for miscellaneous data exploration, such as viewing the number of times a particular web page was accessed. Furthermore, there is no support for identifying visiting patterns to the site as a whole.

Therefore, in order for web site designers to benefit fully from the information available via web server log statistics, improved visualisations which include site structure need to be developed. As web sites are a distinctive type of media these visualisations should be developed with the unique characteristics of web sites in mind.

1.2.3 Other Uses of Visualising Web Site Usage

The main purpose of a visualisation of a web site's usage would be to aid designers in assessing their site's effectiveness. By gaining an improved understanding of how users navigated the site, and thereby obtaining an indication of their interests and any difficulties they encountered, designers would be able to improve the design to increase the site's profitability. This would involve making alterations to the site and then collecting data about the new design before visualising the results. If necessary, further modifications might be made and the cycle would repeat. This process is illustrated in Figure 1.

Visualisations of site usage could prove useful in other applications as well. One such application concerns the creation of adaptive web sites that are personalised for each individual user. This is an active area of research consisting of contributions such as those by Spiliopoulou [47] [48] and Buchner et al. [4], who are seeking to enable web sites to address the particular requirements of all, or any, of its users. These efforts typically make use of a software application known as a *web miner* which extracts sequences of page accesses and attempts to use these sequences to determine

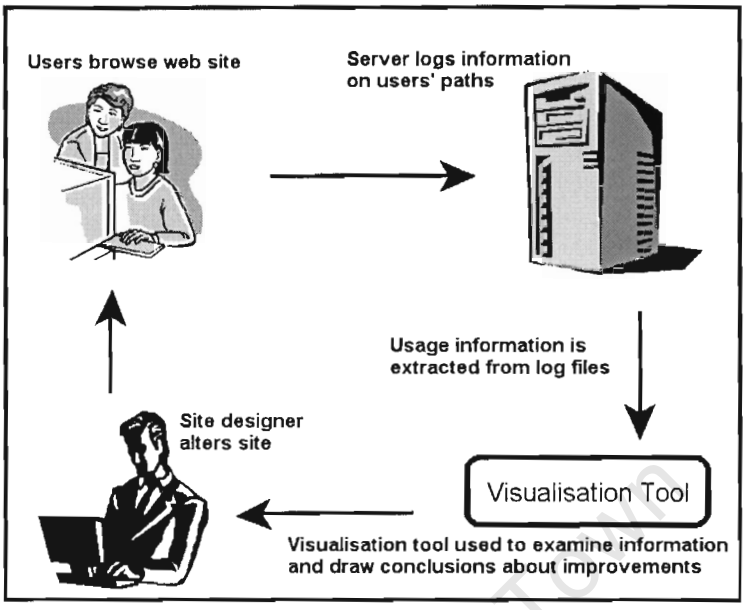


Figure 1: **The cycle of designing, analysing and improving a web site.**

the paths users took through a site (see Figure 2). By aiding in the discovery of navigation patterns as well as enhancing understanding of a web miner's results, web site usage visualisation tools perform an important role in the endeavor to realise the automatic personalisation of web sites.

Another application that could benefit from an effective web site visualisation is the provision of additional navigational aids to a web browser's history list and back and forward buttons. As web sites tend to be nonlinear in structure it is not uncommon for users to lose track of how the page they are currently viewing is integrated with pages they have already seen and with the rest of the site. Conklin described users as suffering from disorientation, which is the tendency to lose one's sense of location and direction in a nonlinear environment, and as becoming confused due to cognitive overload, which refers to the additional effort and concentration necessary to maintain several tasks or trails at one time [12]. The problem with web sites is that as a user only views one page at a time she/he has no physical context on which to base her/his current "location". To compound this, the user may have arrived at the present page by a number of different means such as a link within the site, an external link or else a link created from a text search. Landow referred to this as the "rhetoric of arrival" [26].

The use of a visualisation of a web site that incorporates the site's structure can alleviate these problem by presenting the user with a graphical representation of the entire site. Making such a visualisation easily accessible from the site itself in place of traditional, limited site maps would

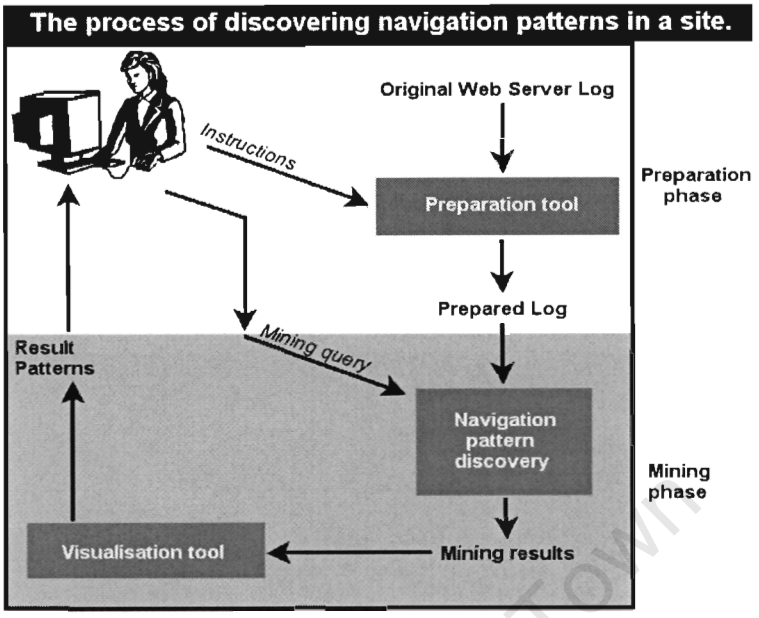


Figure 2: **Discovering navigation patterns in web sites.** This figure, which was reproduced from a similar image in [47], illustrates how visualisation aids the process of identifying user browsing patterns.

provide users with an improved overview of the site layout and thus would aid them in reaching the pages of interest to them.

1.3 Aims

The main objective of this research is to explore methods of effectively visualising the usage of a web site. The purpose of such a visualisation is to aid web site designers in understanding how their sites are being navigated and thus how they can improve the site's design. As was indicated in the previous section, for any visualisation of a site's usage to be complete, it must incorporate the underlying structure. In addition, as web sites possess several unique properties, an effective visualisation should exploit these traits.

As well as displaying the characteristics mentioned above, our aim is also to develop a visualisation that addresses certain factors satisfactorily. Investigating information visualisation literature [5][49][50][51], as well as past work in this area resulted in the identification of several key issues that have an impact on the effectiveness of a web site usage visualisation. These issues, which form the factors that we wish to address, can be stated as follows:

- *Structure Representation* – This involves the choice of arrangement chosen to depict the structure of a web site. The chosen configuration should accurately portray the site's structure without confusing the user.

- *Data Representation* – This concerns the manner in which information (such as page accesses) are encoded in the visualisation. Data should be represented in such a way that their values are easy to ascertain and that interesting patterns are readily apparent.

- *Scalability* – This refers to how well the visualisation scales with size. As it is possible to encounter web sites consisting of several thousand pages, the visualisation should be able to gracefully handle sites of reasonable size.

- *Context Maintenance* – With the size of large web sites it is unlikely that one would be able to survey an entire site in great detail in the same view at the same instant. Instead, users examine subsections of a site, which are usually depicted in some type of magnified display. This issue therefore deals with a user being able to keep track of the relative position of the subsection, or area, of the site they are currently viewing with regards to the layout of the rest of the site.

- *Data Exploration* – This relates to the mechanism of navigating around the visualisation in order to examine various parts of the site, as well as to the technique/s used for allowing users to obtain more details about items of interest. Navigation in the visualisation should be consistent and easy to perform.

To be effective, any visualisation system developed should meet the above criteria.

1.4 Methodology

The approach utilised for this dissertation was as follows:

The first concern to be dealt with was the identification of the various aspects that would affect a web site usage visualisation. These were defined in the previous section. This task was necessary in order to provide a framework by which related efforts in this area could be evaluated, as well as to provide a guide concerning the development of a new visualisation.

Once this was accomplished, research was carried out to locate previous or related work dealing with visualising web site usage. In addition to those projects that visualised web site usage specifically, other works that contained more generalised web site visualisations were also examined. These projects were then evaluated according to the predetermined factors. The results of this

evaluation was then used to guide the development of a metaphor of a web site that was suitable for displaying web site usage.

A metaphor was then devised, with an effort being made to design specifically with a web site's properties in mind. In order to assess the new metaphor's feasibility, a prototype visualisation system was developed. The prototype generated a random web site with every execution. In this manner, the metaphor's suitability with regards to sites of varying size and layout structure could be evaluated. As a result, simulated, instead of real, usage data was utilised. Upon examination, the initial metaphor was deemed to be of limited success in terms of meeting the previously stated goals and was therefore discarded. A second metaphor was thus devised and another prototype developed. With initial experiences of the new metaphor proving to be promising, it was formally adopted so that progress could continue. Throughout this period of prototyping, feedback was obtained from several experts, whose input aided in the decision to abandon the initial metaphor as well as to proceed with the second.

The second metaphor was then further tested through the use of a user experiment. Users were asked to identify the meaning of certain sections of the metaphor using only their intuition (they were not provided with any prior information other than that they were viewing a web site usage visualisation).

Following the results of the intuitiveness test, certain modifications were made to the metaphor, after which a final system was implemented. The aim of this system was to prove the concept, rather than to provide the full functionality that would be present if it were designed for a commercial market. Pilot and full user experiments were then carried out.

Finally the results of the experiments were analysed and conclusions drawn on the effectiveness of the developed metaphor.

1.5 Outline of this Dissertation

Chapter 2 provides some background information relevant to visualising web site access, or usage, information. In particular, it defines what comprises a web site and describes common properties web sites display. In addition, web server log files and log file analysis are discussed in greater detail, with a view as to the type of data that can be reliably extracted from log files.

Chapter 3 concerns previous and related work in this area. Brief descriptions of these are presented according to the manner in which they address the factors of data representation, scalability, context maintenance, structure and data exploration.

Chapter 4 describes how the two metaphors were developed and discusses their anticipated strengths and weaknesses with regards to the factors that works in the previous chapter were examined by.

Chapter 5 details the development of the prototypes developed to evaluate the feasibility of the devised metaphors. This chapter also includes an account of the user experiment performed in order to determine the intuitiveness of the metaphor that was finally selected.

Chapter 6 outlines the modifications to the metaphor that resulted from initial evaluation of the prototype. In addition, this chapter relates certain aspects of the implementation of the final system, such as the manner in which real data was collected.

Chapter 7 presents the experiments carried out to evaluate the completed system. The aims and execution of the experiments are described, after which the results are discussed.

Chapter 8 concludes this thesis with a summary of the results obtained, some concluding remarks and suggestions for future work.

Appendix A contains the questionnaire that was used in the user evaluation experiments,

Chapter 2

Background

In order to develop an effective metaphor, a clear understanding of the available data, and how that data is attained, is required. If that data has an associated underlying structure, a closer inspection of that structure (which in this case pertains to web sites) is also necessary.

This chapter provides some background information about web sites and their properties in Section 2.1. As the data that is to be visualised was obtained from server logs by employing log file analysis, web site log files, the information available in them and the manner in which that information is extracted is discussed in Section 2.2. While being relatively inexpensive and convenient, log analysis does possess several limitations however. As a result other approaches to determining web site usage are being sought, some of which are mentioned in Section 2.3. The decision to use log analysis for the purposes of this research is then justified in Section 2.4. Finally, Section 2.5 concludes the chapter with a summary of its main points.

2.1 Web Sites

The development of the Internet has had a major impact on the publication of information. The extent to which one can potentially reach people in greatly varying geographical locations has made the World Wide Web a highly popular medium for communication. In addition to its scope, the Internet also offers new opportunities as web sites possess unique properties not shared by more traditional forms of publication.

Web sites incorporate the advantages of text, audio and video media as they are an amalgamation of those forums. As a result, web sites can become quite complex structures and a single site can be comprised of pages that contain text, images, audio files, video clips, animations, multiple links, some combination, or even none, of the above. However, there are certain attributes that are common to most sites. These include:

- *Base URL* – All sites have an address or Universal Resource Locator (URL) which points to the location of the initial or starting page of the site.

- *Links* – Web pages on most sites will have at least some links which provide access to other pages. These links, which are uni-directional, are created by using tags containing HREF attributes such as **A**, which is used to indicate single text or image links and **MAP**, which is used to create multiple links on a single image. Links are able to have four types of destinations:

  1. a location in the same page or document
  2. a document on the same web site (relative URL)
  3. a document on another web site (absolute URL)
  4. a program that results in a web page (CGI script).

  In addition, links may also be coded into non-HTML formats using Javascript, VBscript, ShockWave, Flash or Java applets.

- *Images* – Although a large number of text-only sites do exist, as the World Wide Web is very much a visual medium most sites will contain some form of graphics. These may either be purely decorative in function, as in background images, or else may perform some navigational or informative role.

Web sites may also contain additional features such as complex animations, streaming video and audio files or Java applets depending on the web site's purpose and the site designer's discernment.

The individual pages comprising a web site may themselves be categorised according to how they are utilised by visitors. Possible categories include:

- *Entry* pages – the first page a user visits when entering a site.

- *Exit* pages – the last page viewed before the user leaves the site.

- *Entry/exit* – the first and last page visited.

Note that a page may move from one category to another, since this distinction is determined entirely by visitor browsing patterns.

Although each organisation strives to create sites that are interesting and original, certain trends do appear. These trends may result from past experiences and research and are perceived to improve the site's usability, or else may be influenced by currently favoured technologies (e.g. frames, dynamic html,etc). Of more interest from a visualisation perspective, such trends may provide common features which can be utilised in order to create improved metaphors. Examples of exploitable features include:

- *Organisational Homepage* – While various pages within a site may be accessed directly from an external link, sites are devised with the intention that most visitors will enter the site via the homepage. As such, most homepages are designed to reflect this. The majority of site designers 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. An approximate analogy would be a table of contents which partitions a book into various chapters. However, as a web site is not a linear construct, the function the homepage performs becomes more vital.

- *Global Navigation Menu* – Navigating a web site is not always an intuitive procedure. This problem is exacerbated by the unreliability of the "back" button employed by web browsers to return the user to the previous page that was viewed (although this is often due to the site design). To aid users, a large amount of sites contain a navigation bar or menu that contains global links. These bars are often present on the majority of pages composing the site allowing users to access the major links from any page.

- *Self-Contained Web Sites* – This feature occurs more often in fairly large web sites or else those belonging to organisations composed of multiple departments. In these cases it is not uncommon for several smaller web sites to be contained within the main site. Examples include sites such as personal sites enclosed within an institution's web site, faculty sites contained inside a university site and individual sports sites within a general sports news web site.

Having examined web sites in closer detail, the next concern is the actual data to be visualised, as well as the means for acquiring that data. The following section therefore discusses web server logs and log file analysis, the process by which the data for this research was obtained.

2.2 Log File Analysis

At present the most common method of measuring web site usage is to analyse web server log files. Log files are large text files generated by web servers. They contain records of any activity that took place between a web server and users browsers during a particular time period. Log file analysis consists of parsing these files to extract useful information and then summarising it in reports.

2.2.1 Web Server Activity Logs

When a user visits a web site, a connection is established between the web server on which the site resides and the client browser of the user. Each communication between the browser and server,

Figure 3: **Sample of a log file segment.** This segment is taken from a log file generated by an Apache server recording requests for items from the Collaborative Visual Computing Laboratory web site.

such as the requests for a page, then results in an entry being added to the server's log, recording the transaction. The data stored in a log file varies depending on the type of server being used and the log file formats that it supports.

The two most widely used formats are the *Common Log Format* and the *Extended Log Format.*

The Common Log Format utilises four different log files to track user information, namely the access log (sometimes known as the transfer log), which logs data about the request being made, the agent log, which collects details about users' browsers and operating systems, the referrer log, which contains information about the site that referred the user to the current site, and the error log, which stores data about failed requests.

The Extended Log Format tracks the access, agent, referrer and error log into one log file and supports additional directives that provide meta data about the log file, such as the version and start and end date.

Figure 3 shows an extract from an log file generated by an Apache server, which contains requests for pages and images from the Collaborative Visual Computing Laboratory web site. Extracting information from this log file is then a straightforward process which requires simple parsing of the file.

2.2.2 Information that can be Inferred from Log Files

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 (commonly referred to as hits),
- the number of requests by type of file, such as HTML documents, JPG images, etc.,
- the distinct IP addresses served and the number of requests each made,
- the number of requests by domain suffix (derived from IP addresses),
- the number of requests for specific files or directories,
- the number of requests by HTTP status codes (such as successful, failed, redirected, informational),
- the number and size of files successfully served,
- the URLs 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.

Take for example, the final line in the log file in Figure 3, which reads

access.log.20000801.gz:cable.cs.uct.ac.za – – [31/Jul/2000:09:55:25 +0200] "GET /Research/CVC/ HTTP/1.1" 200 8475 "http://www.cs.uct.ac.za/Research/" "Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt).

Examining this line yields the following information:

- *the domain name* – namely cable.cs.uct.ac.za,
- *the time and date the request was made* – i.e., 31st July 2000 at 9:55 am,
- *the page being requested* – in this case the CVC homepage (http://www.cs.uct.ac.za/Research/CVC),
- *the version of the protocol being used* – which was HTTP 1.1,
- *the result of the request* – which, as the status code was 200, was that the transfer was successful,

- *the volume being transferred* – namely 8475 bytes,
- *the referring site* – which was the Research page of the Computer Science department site (http://www.cs.uct.ac.za/Research/),
- *the browser that made the request* – which was Microsoft Internet Explorer 5.0, and
- *the user's operating system* – namely Windows 98.

2.2.3 Disadvantages of Log Files

While log file analysis provides some indication of determining site usage, it does suffer from several major flaws [15][17][39]. These are summarised as follows:

Caching

Since the advent and increasing use of caching, log files may no longer be able to accurately report the correct amount of activity for a web site. This is due to the fact that all requests for a page that has been stored in cache are not recorded in the servers log, as no request is made to the actual server.

Incomplete Data

Another drawback of log files is that they contain data about files transferred from client to server and not information about people visiting the web site. This means that certain usage data is not logged, while other data that is logged is inherently incomplete. Data not captured in log files includes items such as individuals' identities, sites which users visited after leaving a particular site, or any qualitative data such as user motivation for viewing a site and reactions to site content. Inherently incomplete data includes the number of requests and all other statistics based on that figure. This information is incomplete due to local and regional caching.

Incorrect Assumptions

In addition, many commercial log analysis products employ complex heuristics in order to make educated guesses about information that is excluded from log files. However, not all of the inferences drawn in such a manner are sound. Unsound inferences include the concept that "user sessions" can be isolated and counted. Many log analysis products calculate user sessions by tracing requests received from a particular IP address until a sufficient period of inactivity suggests that the session ended. This calculation is based on two unsound assumptions; first, that a host corresponds to an individual (indeed, uniquely identifying visitors from log files is a considerable problem), and second,

that the individual would not pause (to perform some other task) while within a site. As such, many statistics provided by log analysis products, which are based on user sessions, are also unreliable. These include average page per views, average length of session, average length of a page view, top entry and exit pages, single use pages and top paths through a site.

Researchers such as Pirolli et al [38] have suggested more complicated heuristics to uniquely identify users, such as examining successive requests with regards to site topology to determine if the subsequent page is reachable from the first. If not, this would indicate multiple users as would requests for pages that have already been visited (if it were a single user no requests for pages already stored in cache would be made). However, Pitkow (a co-author on [38]) himself stated that these measures have not been shown to reliably identify users with any degree of accuracy [39].

2.2.4 Cookies

Some of the difficulties experienced when analysing log files outlined in Section 2.2.3 can be overcome through the use of *cookies*. With log analysers having to deal with the same proxy being utilised by multiple users as well as dynamic IP addresses, many log analysis packages eagerly embrace the use of cookies to isolate individual users.

Cookies are small samples of textual information that a web server stores on a client's (user's) computer. Several specifications exist that control both the cookie file maintained by the user's browser and the cookie string that is passed between servers and clients.

Cookies are transferred in the header of an HTML file and contain up to six variables, two of which are obligatory and four of which are optional. The mandatory variables are the cookie name, which could be any unique name provided by a specific site, and the cookie value, which could either a variable name gathered from within a form or else a unique identifier to be used by a database. It is this second variable that is used to uniquely identify a user by log analysers. The four remaining optional variables are the expiration date, the valid path and the valid domain of the cookie as well as whether the web site issuing the cookie can only be used under a secure connection.

Unfortunately for web site usage evaluators, cookies are not universally accepted by the user community. Many users are hesitant to access sites that install cookies due to privacy concerns as they do not wish their virtual movements to be known. Some also refuse to accept cookies for security considerations as cookies have been exploited by malicious programmers to spread viruses.

2.2.5 Proposals for Improving Usage Statistics Gathering

New proposals are being put forward to make usage data obtained without the use of cookies more reliable. Examples include *hit-metering* [31] and *user sampling* [39]:

- Hit-metering proposes the implementation of a new HTTP header, called "Meter". This header would enable proxy-caches to report usage and referral information to the original server. Additional extensions could also permit the originating server to limit the number of times a proxy-cache returns a document before requesting a fresh copy.
- User sampling involves continuous sampling of a random set of users. The users being sampled are identified using cookies. Once a user has been identified, caching is defeated for all subsequent requests by that user during the sampling period.

2.3 Other Approaches

Due to the shortcomings of log file analysis other approaches to determine web site usage have also been investigated. However, each of these has its own benefits and weaknesses.

2.3.1 Qualitative Methods

Different approaches include qualitative methods of data collection. These range from guest books and feedback forms to user surveys and focus groups. The advantage of these types of data collection methods is that they include information such as user opinions on site content, navigation and "look-and-feel". In addition they can provide indications as to user satisfaction and motivations. A disadvantage of these methods is that they are costly and time-consuming. In addition, many users are unwilling to participate in surveys or to fill out forms.

2.3.2 Using Information Scent

Attempts concerned more with predicting web site usage rather than displaying current usage information have also been proposed. One such attempt makes use of information scent, which is the "imperfect, subjective perception of the value, cost or access path of information sources obtained from proximal cues, such as web links, or icons representing the content sources" [10]. To our knowledge, this method is restricted to academic use and has yet to be adopted outside the research community.

2.3.3 Human Browsing

Another approach involves companies, such as Vividence [54], that employ people to browse web sites. These organisations will gather and provide user feedback from their employees concerning a site, for a certain fee. While such an approach can provide one with different types of information than log file analysis, it does also face certain drawbacks. For instance, people who are browsing

a Web site as part of their job may have very different interests and motivations to those who are browsing either for leisure, business or information gathering.

2.3.4 Software Agents

A new emerging approach is to employ software agents as surrogate users to traverse a web site and determine usage information. Systems such as WebCriteria SiteProfile [55] use a browsing agent to traverse a web site using a modified GOMS model [6] and record download times and other data. The problem with these types of systems is that they are limited to metrics such as load times and amounts of content versus hyperlink structure. Any system attempting to provide more information would have to show that their software agent had similar browsing patterns to a human.

2.4 Use of Log File Analysis

The decision was made to obtain data using log file analysis for several reasons. Firstly, it is presently the most common method of determining web site usage as the number of log analysis products available can attest. Secondly, it is an economically and chronologically inexpensive process, requiring no specialised equipment or highly trained personnel. In addition, as web servers automatically keep logs it is convenient as data is readily available. Finally, although, many of the statistics log analysis provides without the use of cookies are estimates at best, these still offer some indication of site usage as long as it is kept in mind that they are only estimates.

2.5 Summary

This chapter provided some background information relevant to this research.

Web sites and the components of which they are comprised were examined. These include a base URL, links connecting the various pages that make up the site and optional text, images, animations, audio and video files according to the designers discretion. In addition many sites possess an organisational homepage, a global navigation bar and the inclusion of smaller, self-contained web sites.

Log file analysis was then discussed. It consists of parsing text files called log files, which are created by web servers to store interactions between servers and client browsers. Analysing these files can provide information concerning requests for individual pages, such as the date and time of request, as well as information about the user for whom the request was made, such as browser type.

However, log file analysis is prone to certain errors resulting from caching, incomplete data and incorrect assumptions. The use of cookies can reduce these errors but they introduce their own issues such as privacy and security concerns.

Other approaches to measuring web site usage include qualitative methods, information scent, human browsing and software agents. Each of these has its own strengths and weaknesses and it remains to be seen if any of them will replace log file analysis as the currently most widely used method of determining site usage.

In spite of its failings log file analysis was chosen as the method for obtaining data for this research as it is economical as well as convenient.

Chapter 3

Previous Work

As the role that the World Wide Web performed in day-to-day activities increased, so the importance of understanding web sites and peoples' interaction with them became more apparent. However, attaining this understanding requires the analysis of vast quantities of data, as popular sites may receive many thousands of visits per day. With information visualisation proving an important tool in extracting useful information from large data sets, numerous research efforts were developed that address the visualisation of web sites.

This chapter investigates several of these previous works.

It begins with Section 3.1, which lists the purposes that related research projects attempt to address. This is followed by Section 3.2, which provides a recapitulation of the factors identified to have an effect on visualising web site usage, which were introduced in Section 1.3 in Chapter 1. Section 3.3 then examines several classes into which previous works can be classified, and investigates the manner in which each approached the factors mentioned previously. Finally, the chapter is concluded by Section 3.4, which summarises the main points presented.

3.1 Purposes of Past Web Site Visualisations

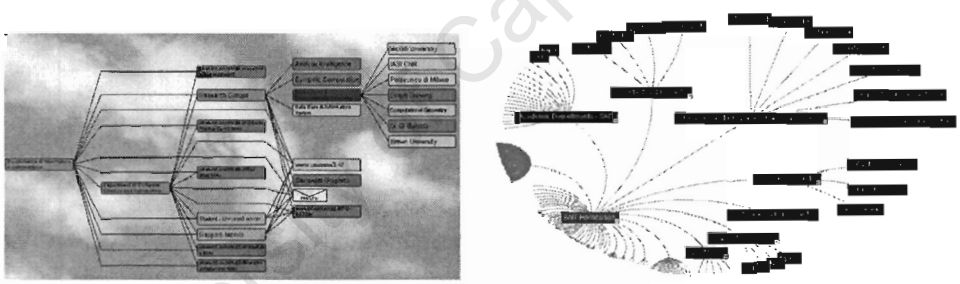
The topic of web site visualisation has been approached from several different perspectives, the purposes of which vary as greatly as the resulting metaphors. Although when examining previous works it is important to bear their purpose in mind, it is more the manner in which these efforts chose to represent web sites that is of concern to this dissertation. Thus, in the discussion to follow, past projects will be considered in relation to the visual representations, or metaphors, they utilise, rather than according to their function.

However, as the objectives of a project has a direct influence on the resulting metaphor, a brief look at the varying purposes of past efforts is first provided.

While a number of web site visualisations exist, most projects, whether commercial or academic in nature, can be generally grouped into one of three categories according to their intended function. These are: *web site navigation visualisations*, *web query visualisations* and *web site usage visualisations*.

3.1.1 Web Site Navigation Visualisations

Of the three, web site navigation visualisations are probably the most numerous. The nonlinear qualities of many web sites, in addition to the fact that only one page may be viewed at a time, has led to users experiencing problems (see Section 1.2.3) while navigating certain sites. Visitors to complex or sizable sites may often become "lost" in that they lose track of the location of the page they are currently viewing in relation to the rest of the site. The simple history list and back and forward buttons, which many browsers employ, often add to the predicament, as they are more suited to a linear construct. Activating them can thus lead to pages that are different to those that the users expected. As a result, instead of providing help, these aids can cause users to become more confused.



(a) (b)

Figure 4: **Examples of Web Site Navigation Visualisations.** (a) Ptolomaeus, the web cartographer by Battista et al. (b) Inxight Software's Tree Studio.

To solve these difficulties, various attempts have been made to create visualisations of web sites that aid users in effectively navigating those sites. These range from simple site maps located on the web sites themselves, such as those produced by Dynamic Diagrams [13], and applications designed to augment currently existing navigational support (browsers' history lists), such as Ptolomaeus [3] (Figure 4a), to more complicated stand-alone products, like Inxight Software's Tree Studio [23] (Figure 4b).

Regardless of their approach, these efforts all seek to improve users' understanding of a web site by familiarising them with its structure and improving their conceptual overview of the entire

site.

3.1.2 Web Query Visualisations

The World Wide Web has evolved into an important information resource, containing information concerning a wide variety of topics. Indeed, it is rare to encounter a subject upon which there does not exist at least one web page. To date, however, successfully utilising this source of information has not been an issue of whether the required information is available, but rather of whether that information is locatable. Performing a search using a search engine that retrieves only the necessary information is a task that is far from simple.

As a result, several researchers have developed systems designed to make the searching process less vulnerable to inaccurate or undesired results. These efforts consist of supplementing search engines by visualising the search and/or the results returned. As pages containing related topics are often located in close proximity to each other, most of these visualisations include some form of visualising the site structure, in order to illustrate interconnected pages on the required subject. An examples of a project that falls into this category is WebQuery by Carriere and Kazman [8].

3.1.3 Web Site Usage Visualisation

Relatively little research has been conducted in the field of visualising web site usage. Some exceptions include the early effort of Pitkow and Bharat [40], the contribution by Hochheiser and Shneiderman [20] (Figure 5a) and the various works produced by the Xerox PARC group [38],[9],[10].

There are however, a plethora of commercial products which determine web site usage from log file analysis and then visualise the results in the form of tables, histograms and pie charts. These include WebTrends Log Analyzer [56] (Figure 5b), NetTracker Enterprise [43] and Funnel Web Professional [1], to name but a few (others can be seen in an article run in PC Magazine [52]).

Aside from these, more complex products exist that include some form of site structure visualisation as well as the log analysis data. Examples of these more advanced offerings are those such as Mercury Interactive's Astra SiteManager [28] (Figure 5c) and Microsoft's Site Server [30] (Figure 5d), which incorporates visualisation research performed at Xerox PARC that was later licensed as a software component to OEM developers.

All the tools in this category attempt to aid in the analysis and understanding of the manner in which users "use" or navigate web sites. Feedback is thus obtained on whether current sites are fulfilling their aims, as well as on possible avenues of improvements.

Before proceeding to examine the metaphors that have been developed to address the problems stated above, a recapitulation of the factors by which they will be evaluated is required. This is presented in the following section.

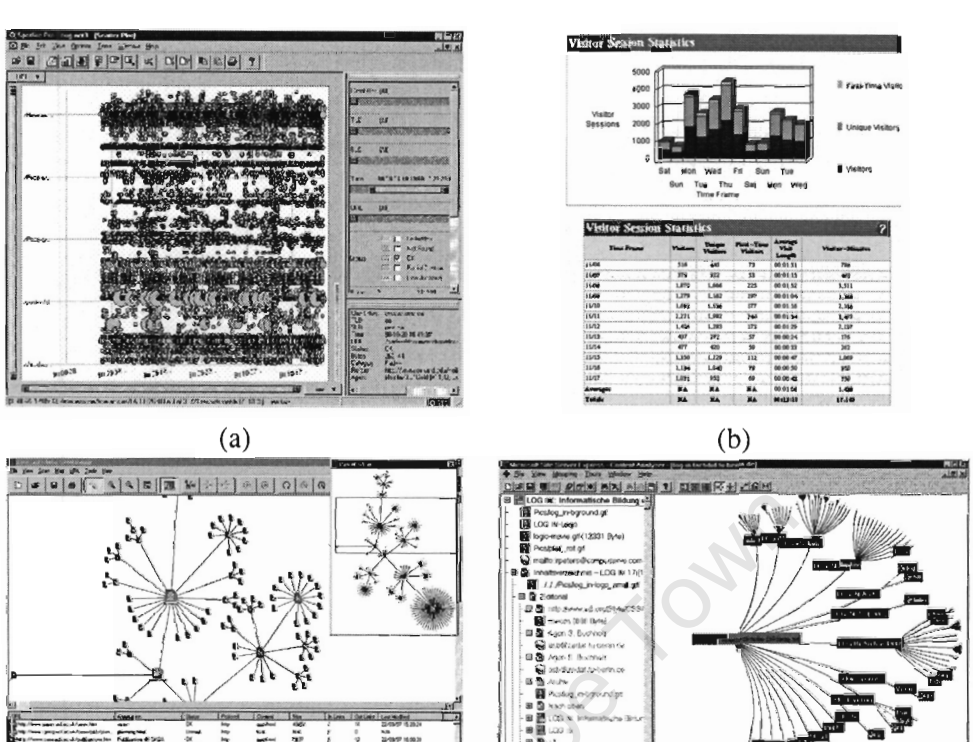


Figure 5: **Examples of Web Site Usage Analysis Tools.** (a) Hochheiser and Schneiderman's Spotfire tool. (b) WebTrends Log Analyzer. (c) Mercury Interactive's Astra SiteManager. (d) Microsoft's Site Server.

3.2 Factors Affecting Web Site Visualisations

Irrespective of the purpose for which a web site visualisation is devised, there are certain factors that which it must successfully address in order to be effective. These factors were identified through an investigation of information visualisation literature as well as previous attempts at visualising web sites. Common issues faced by past works as well as approaches to visualisation such as Schneiderman's visualising information mantra of "overview, zoom and filter then details on demand" [46] can be summarised into the following five factors:

1. *Structure Representation*,
2. *Data Representation*,
3. *Scalability*,

4. *Context Maintenance*, and

5. *Data Exploration.*

Attention will now be given to each of these in turn.

3.2.1 Structure Representation

Of the five factors identified, the manner in which the structure of a site is represented has the greatest impact on the appearance of a metaphor. As such, the choice of how to represent site structure is a vital one. This importance is compounded by the fact that the chosen representation affects the other four factors to a large extent as well.

The most important requirement of the selected representation is that the user is able to correctly identify and interpret it. If users are unable to form a clear mental conception of the site's structure from the visual display they are viewing, then it is unlikely that they will make effective use of the visualisation.

A number of other concerns exist. One of these is the *type* of structure being visualised. Visualisation developers can decide to either represent the *link* structure of a site or else its *file* structure. A web site's link structure refers to the logical organisation of the site as determined by the presence of links between pages, while the file structure depends on the folder organisation of the actual HTML files that comprise the site. The type of structure being displayed has an influence on the representation utilised as certain metaphors may be more suited to one structure type than the other (for example, a hierarchical tree representation may be the most appropriate for visualising file structure). Most often, the choice of structure type used is dictated by the purpose of the visualising system. It is possible for a system to visualise both structure types, although utilising a single metaphor to do so may lead to a reduction of effectiveness due to necessary compromises. A better option may be to switch between multiple metaphors instead.

Another concern to be addressed regards the decision of implementing a two dimensional metaphor or else a three dimensional one. Two dimensional representations are generally simpler and easier to use than their more complex three dimensional counterparts. However, an extra dimension permits designers to include additional information than would otherwise be possible.

All these concerns need to be carefully considered so that an effective structure representation can be chosen.

3.2.2 Data Representation

The ultimate goal of any visualisation is to graphically represent data in a meaningful manner. Thus, the issue of how to display information is one that requires careful consideration.

The nature of the data to be visualised clearly plays a large role in this consideration. Web site visualisations usually contain either site usage data, which includes items such as server loads, page accesses, error statistics and browser details, or else query data, which comprises the results of a site search. Navigation data, such as the location of pages or history lists of links already visited, are normally encoded in the structure representation itself. The different data categories present designers with different issues, which need to take into account when creating a metaphor.

In addition, complications arise in that the options of displaying data are restricted by the entities to which they pertain. With regards to web sites, data could be associated with web pages, links, servers, visitors or entire sites. As these entities vary greatly in character, so their representation is likely to differ significantly as well. However, all these entities will generally be represented by some form of geometric object, and there is a limit to the number of methods by which information can be encoded in a geometric body. These include altering the item's colour, size, transparency, shape and texture. Deciding on which of these will be the most effective can prove to be a challenging matter.

Whatever the chosen means of representing data, disparities between individual data values should be readily discernible. In addition, any trends or patterns present in the data should be highlighted and made apparent.

3.2.3 Scalability

As the World Wide Web expands, so the factor of scalability gains in importance. With the increasing number of services available online, the size of modern web sites is continually rising. Thus, it is possible to encounter sites that consist of several thousand pages. This introduces several potential problems in terms of web site visualisation.

The value of a visualisation which is restricted to depicting sites of small or intermediate size is severely limited. Therefore, a requirement exists that visualisations are able to gracefully handle sites of large scale.

When sites reach a certain size, several concerns materialise. The first of these is *cluttering*. Cluttering occurs when certain elements of a visualisation obscure other elements. This is mostly the result of attempting to display too much content for the available screen estate.

The second concern is that of presenting the user with too much information at the same time. The user therefore suffers from *cognitive overload*, whereby s/he is unable to mentally acknowledge and process all the information currently being displayed, which results in confusion.

The third concern is that of speed considerations. As the size of the site being visualised increases, so the number of graphics primitives required to render its representation increases as well. While modern graphics accelerator cards are able to render vast numbers of primitives at interactive

rates, they still possess limits. Thus, once a site requires a certain number of objects to be drawn, the rate at which the visualisation display is refreshed will be reduced. Once this rate is reduced to a critical level, the user will perceive the delay and may have difficulties in effectively interacting with the system. Therefore, the choice of which graphics objects to utilise in representing the site becomes vital, as certain objects may be more efficient to render using graphics hardware than others.

These considerations all need to be addressed in order to ensure that a visualisation is not rendered unusable due to attempting to visualise a large and complex site.

3.2.4 Context Maintenance

With the size of many modern web sites, displaying the entire site in great detail in a single view is impractical. Instead, visualisers rely on the concept of "focus and context". This notion involves the display of highly detailed depiction of a subsection of the visualisation (i.e., the focus), together with a low-level detail view that provides the perspective location of the high detail view, in terms of the entire visualisation (i.e., the context). Without such a context view, users could lose track of the position of the subsection that they are currently viewing, thereby becoming "disorientated". What would follow, would be a loss of productive time while the user attempts to reorient him/herself.

While there is no argument concerning the need for a focus and context mechanism, the best method to implement such a mechanism could provide much debate – several different possibilities exist. Among others, these include the use of two separate windows, the use of a "fish–eye" algorithm and the use of alternative geometric spaces.

3.2.5 Data Exploration

The final issue to be discussed is that of data exploration.

A visualisation is only of use if users are able to explore and manipulate it effectively. If details of the information being displayed are not readily accessible, it may lead to frustration as well as a reluctance to continue utilising the visualisation. Thus, interaction with the data needs to be implemented in such a manner that it is intuitive, consistent and "user-friendly".

Data exploration consists of two components, namely the manner in which the overall visualisation is manipulated and thereby "navigated", and the method by which further details concerning a particular item are obtained (this is sometimes referred to as "drilling down"). While the former is generally of vital importance regardless of context, the latter may depend on the application type. Certain applications contain less information than others, and so may be able to display all relevant details at a single level without requiring further user interaction. For those projects that do necessitate drilling down to acquire additional information, both the actual mechanics of how the

drill-down process is performed, as well as the manner in which the further information is presented, contribute to the *usability* of the end systems.

Unfortunately, the issue of data exploration is complicated by the fact that the effectiveness of particular interaction techniques is often subjective. Interaction methods that one user is able to utilise efficiently, may daunt another user. Selecting data exploration methods that most, if not all, users find beneficial is a nontrivial task.

Now that the factors which affect any web site visualisation have been outlined, a discussion of previous attempts at representing web sites follows. The decisions taken by each, with regards to the factors described above, will be examined for any benefit they may provide to the development of a new metaphor.

3.3 Past Metaphors Used

With the number of people searching for effective visualisations of web sites, there are bound to be certain metaphors that occur frequently, regardless of the manner of tool in which they appear or that tool's purpose. Although certain disparities may exist within each metaphor group depending on the requirements of the systems that utilise them, in general, examples of a particular metaphor will display similar behaviour and share like appearances. Such metaphors include *cyclic graphs*, *hierarchical trees*, *cone trees*, *radial views* and *hyperbolic trees*. These metaphors, which are encountered in many disparate applications ranging from site navigation aids to web query tools, will now be each described in turn.

3.3.1 Cyclic Graphs

The majority of past web site visualisations depict sites as either cyclic graphs or else hierarchical trees. Of the two, cyclic graph metaphors are encountered far less frequently, even though they are more accurate representations of web sites.

Cyclic graphs are comprised of a network of nodes, usually denoted by circles or rectangles, which are connected by lines that signify links. In terms of web site visualisation, the nodes commonly represent individual web pages, while the lines portray the links between those pages. Cycles may occur in these graphs, as it is not uncommon for web sites to contain paths that lead back on themselves.

Given the current popularity of tree representations of web sites, most of the projects that depict web sites as graphs tend to be early efforts, such as WebViz [40] and AOLpress' now defunct mini-web representation [25], which is shown on the left side of Figure 6.

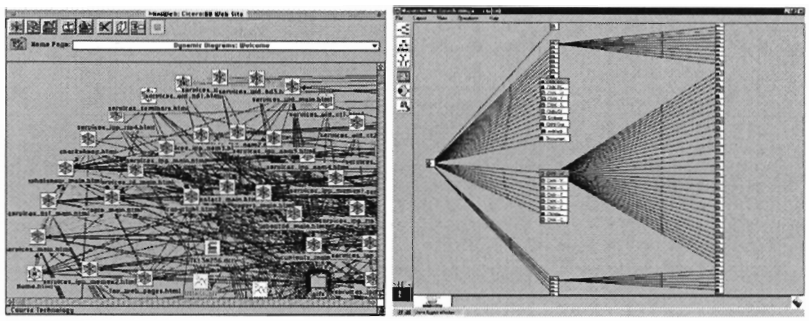


Figure 6: **Cyclic graphs versus hierarchical trees, the two main metaphors used to depict web sites.** In the picture on the left is a cyclic graph representation of the Dynamic Diagrams web site (*www.dynamicdiagrams.com*), which was created using AOLpress' mini-web tool. The image on the right shows a hierarchical tree representation of CNN's web site (*www.cnn.com*) created using IBM's Mapuccino tool.

Structure

Both the appearance and the interactivity of graph representations depend a great deal on the layout algorithm used. Determining an appropriate positioning of the nodes and edges of a graph is a complex task, one that has generated an entire field of research (refer to [18] for a brief treatise). Aside from those constraints imposed by aesthetic rules, such as nodes and edges being evenly distributed, edges having the same length, edges being straight lines, etc., devising graph layout algorithms is further complicated by the requirement that the resulting graphs possess certain properties. One such property is *planarity*, which refers to whether it is possible to draw a graph on a plane without any edges intersecting.

Numerous approaches have been taken to solve these problems. The ensuing algorithms, many of which have been applied to web sites, can be categorised either according to the type of layout they produce, or else according to the methodology on which they are based. An example of the former would be a grid layout, such as the hexagonal grid layout in NicheWorks [57], that place nodes at regular positions (usually with integer coordinates). Layouts that are defined by the methodology utilised include non-deterministic approaches, such as simulated annealing algorithms and force-based models.

The different classes of layout algorithms produce graphs that may vary greatly in appearance. While each possesses desirable traits, each class also generates its own difficulties. However, one problem shared by all is the scale of the graph. Few systems claim to effectively handle graphs that contain many thousands of nodes and edges, NicheWorks and H3Viewer [34] being the exceptions. Unfortunately, as already discussed, encountering web sites of this size is becoming increasingly

common. Layout algorithms that generate good layouts for several hundred nodes may no longer be suitable for graphs that are several orders of magnitude larger. Cluttering becomes a major issue, and occlusions may become severe enough to cause navigation and queries about particular nodes to be impossible.

There have been attempts to alleviate the problems posed by scale through the use of three dimensional layouts. The HyperSpace system [58], developed at the University of Birmingham, takes this approach. However, on its own, the addition of an extra dimension can only assuage the predicament to a certain extent. Ultimately, the limits imposed by available screen estate renders the display of all the nodes and edges of large scale graphs impractical.

Aside from 3D, the use of non–Euclidean geometry has also been proposed as a means of handling large graphs, although these techniques are more often applied to trees. As a result, these will be discussed in Section 3.3.5.

Data Representation

WebViz, a tool that visualises the results of web server log file analysis, encodes frequency and recency of access information in the width and colour of links and the borders of their associated nodes. The nodes themselves are represented by text boxes that contain the names of the related pages. NicheWorks includes the ability to map data to additional node and edge attributes such as shape, size and line style.

The problem with embedding information in the node icons, or representations, is that such an approach is limited to relatively small sites with regards to displaying overall patterns in the data. Once a certain limit is reached, discerning the individual nodes can become problematic due to the effect of cluttering, as well as to the necessarily reduced size of the icons in order to fit the entire site into view. A side effect of this is that node icons very rarely resemble the appearance of the pages they represent. Altering node attributes then, is best suited to the mapping of information that will only be required for drill-down operations.

On the other hand, the encoding of information in the edges (or links) is not affected as greatly by the scale of the visualised site, as these are generally easier to perceive at greater distances.

An interesting use of node colour is that implemented in the HyperSpace system, which is a navigational and search aid tool. The system renders nodes in greyscale according to their distance from the viewer in the 3D representation. Nodes that are light are therefore closer to the user whereas those in the distance fade into the dark background.

In addition to the information mapped onto the actual structure representation, supplementary data may be presented in tables or standard graphical user interface widgets such as text boxes.

Scalability

Unlike early systems, most tools that make use of cyclic graph representations do not attempt to portray entire web sites. Instead, they utilise some means of reducing the number of nodes and edges required to be shown. Common means of achieving this include the use of some form of *clustering*, whereby parts of the graph are abstracted or "summed" and replaced by a single representation (for example the "supernodes" in HyperSpace), and the use of user–defined filters, such as those utilised in NicheWorks.

The use of filters introduces the danger of suppressing data to such an extent that the user loses the context of the filtered results. This becomes possible when a system no longer renders any representation of those sections that are filtered out at all.

Context Maintenance

Aside from those projects which make use of non-Euclidean geometry or fish–eye views, most graph–based systems do not make any provision for context maintenance. This is understandable in that when many of these systems were devised, web sites consisted of far fewer pages. There are, of course, methods of maintaining context that can be applied across a range of metaphors, such as the use of fish–eye distortions and secondary overview windows. These will be discussed when examples of systems that implement them are encountered.

Data Exploration

Systems that provide navigational support contain standard zoom and pan abilities. Drill–down operations are typically performed either by mouse–over gestures, node selection or else by typing the identity of the node into a text field on the interface. The latter is probably the least effective, as it requires the user to break focus from the visualisation to look at the surrounding interface instead.

3.3.2 Hierarchical Trees

As already mentioned, web sites are most commonly displayed in the form of hierarchical trees. Trees have received a lot of attention in graph literature and thus there exist numerous different tree layout algorithms, many of which have been to used to represent web sites. This subsection discusses only systems which make use of a classical tree layout. Relevant variations and adaptations will be dealt with later.

The traditional tree contains an initial node called the *root* node. This node generally possesses several subnodes, which are said to be *children* of the root node, and which may or may not contain children of their own. If a node does not have any children it is referred to as a *leaf* node. Trees

are created by adding nodes to the root node and subsequently adding further nodes to these nodes. In this manner, a *hierarchy* consisting of several levels is set up. A classical layout algorithm will place children nodes "below" their common ancestor.

With regards to web sites, the home page is usually associated with the root node of a tree. Any pages which are directly accessible from the home page then form the children of the root node. Links leading from these pages correspond to the nodes of the next level of the hierarchical tree and so on.

Most of the early "site maps" that many web sites provided in order to aid the user in navigating the site, such as the *Nature Neuroscience* site [16], were hierarchical trees. An example of a tool that uses this representation is IBM's *Mapuccino* product [21], a screenshot of which is shown on the right side of Figure 6.

Structure

There are a few advantages to representing a web site as a tree rather than as a cyclic graph. For the same site, the number of links required to be rendered is much less, as any links that are not from parent to child pages are ignored. This results in more screen estate being available, which permits sites of greater size to be represented. Another advantage of tree representations is that links do not cross over each other, potentially making them easier to explore and understand. Finally, it can be argued that many users think of the structure of a site conceptually as a tree rather than a cyclic graph (it would be interesting to determine the influence of "site map trees" employed by many sites on this perception). Users might thus find a tree representation to be more familiar and may therefore be more comfortable interacting with one.

Unfortunately, tree metaphors do possess some weaknesses as well. The most serious of these is the same attribute that makes them scale better than graph representations, namely, the lack of "non–tree" links (i.e., those links that do not follow the hierarchy). By not depicting these links, tree representations discard possibly valuable information about a site's structure. Many tree maps completely ignore non–tree links, which may have an adverse effect on how a site is navigated. Most systems that do address this problem, approach it by displaying non–tree links "on demand", or in other words, they will display non-tree links associated with a particular page when that page has been highlighted. This is only a partial solution however, as users are still presented with a poor idea of how the various pages comprising a site are interlinked beyond the links from parent to children. As a result, their understanding of the site layout suffers, as even if they intellectually understand that there exist links in the site that are not currently being displayed, their subconscious impression of the accessibility of certain pages may be limited to hierarchical paths.

Data Representation

Classical tree representations are generally used only as web site navigational aids or else as complementary displays for systems using other metaphors. Consequently, the only data they are generally used for displaying is the layout of the site. Naturally, they could be modified to display other information, although this would depend on their actual implementation. For example, some systems, such as Spotfire [20], use windows file directory tree type browsers for representing trees. In these cases, mapping data to link attributes would be ineffective, as the lines representing the links are typically very short and therefore difficult to discern. Altering the page icons on the other hand may be a more promising route.

Scalability

Ordinary tree representations generally do not scale well beyond a few hundred nodes, which is inadequate for visualising modern web sites. Implementing a tree as a file directory type view, allows one to represent size of much greater size, as the majority of the tree will be hidden at most times. Such an approach does mean however, that more interaction is required in order to obtain an idea of the site's entire layout, especially if the site is relatively large and consists of many depths.

Context Maintenance and Data Exploration

With a tree representation such as those used by site maps, the entire site is normally in view. Thus, there are no requirements for the provision of context maintenance or navigation features. Tree implementations that hide sections of the tree are usually manipulated by mouse clicks on either nodes or else "activation points" (e.g. small boxes containing positive and negative signs), which then show or hide the relevant subsections of the tree. If the site is particularly deep, there is the risk of expanding so much of the site that the entire site no longer fits into view. In such a case, context can be easily lost.

3.3.3 Cone Trees

Devised by Robertson et al in 1991 [42], cone trees are hierarchical trees that are displayed using an alternative layout.

The original cone tree was a three dimensional construct that was viewed from the side. The root node of the hierarchy was located at the highest vertical point of the cone tree and formed the apex of a cone, with its children placed evenly along the cone's base. Each of the root's children then formed the apices of cones located beneath the root cone and that contained their own children as the bases, thus making up the next layer and so on. To ensure that cones didn't overlap and

thereby obscure each other, the radius of each cone was calculated to be proportional to the amount of space it requires. This space is determined by the total number of nodes contained in all paths leading from the nodes making up the cone itself. Cones belonging to the same layer were all of equal height. In addition, all the cones were rendered with transparent shading so that while they could readily be perceived, the cones located behind them were not obstructed.

Cone trees can also be converted to two dimensions by viewing the cone tree from the top. In this form they are sometimes referred to as "dandelion" or "balloon" representations instead of cone trees.

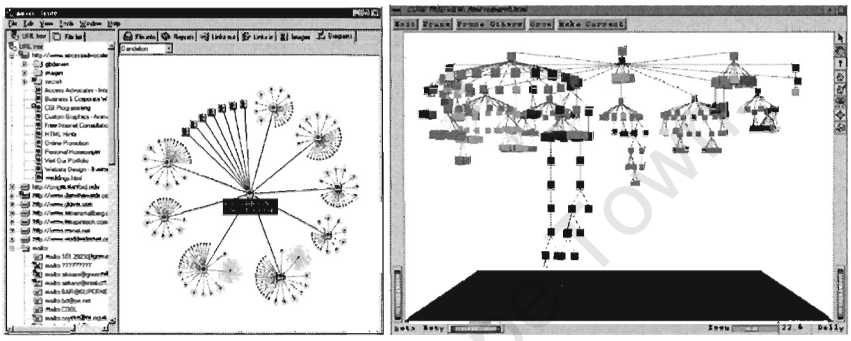


Figure 7: **Cone Tree Metaphor.** The image on the left displays a screenshot from IXACTA's Ixsite tool [24], which makes use of a 2D cone tree representation of a web site. The picture on the right was taken from Mukherjea and Foley's Navigational View Builder tool [32], which utilises several different metaphors, this one in particular being a 3D cone tree.

Since their development, the use of cone trees to illustrate site structure has become quite widespread. Although the top–down, two dimensional view (2D) is seen more frequently, such as in tools like Mercury Interactive's Astra SiteManager [28], IXACTA's Ixsite [24], Merzcom's Netscope [29] and the Visual Web product [53], the three dimensional (3D) variation is also occasionally encountered, such as in Mukerjea and Foley's Navigational View Builder [32]. One unusual example of a cone tree was that utilised in Apple's Hotsauce tool [2]. Hotsauce utilised a top–down cone tree without the links being portrayed. Pages were thus represented by floating labels which were colour–coded according to the level of the tree they were located in. A large amount of overlap was experienced with higher levels obscuring lower ones. This was alleviated by "flying" down the tree so that the top levels disappeared from view. Hotsauce is no longer supported by Apple.

An example of a 2D cone tree is illustrated on the left–hand side of Figure 7, while a 3D representation is provided on the right.

Structure

As cone trees are merely hierarchical trees laid out in an alternative manner, they possess some of the same strengths and weaknesses as traditional hierarchical tree layouts. They are able to effectively display larger sites than graph metaphors and are conceptually simpler, but leave out structural information in the form on non–tree links.

It may be argued that as the 3D cone tree is more aesthetically similar to a classical 2D hierarchical layout, it may reflect the intrinsic hierarchy of the data more apparently than the 2D version. Whether this is desirable in terms of web sites, considering that web sites are not always hierarchical in nature, is debatable.

Data Representation

Again, both node and link attributes can and have been utilised to encode web site data. Astra SiteManager renders links in one of three colours (blue, pink or red) depending on how many hits the associated page received. The use of three distinct colours has the advantage of removing any ambiguities the user may have concerning which range the number of hits for a page falls into. However, the particular choice of colours appears to have been made arbitrarily. As such, users have the additional overhead of recalling which range was indicated by what colour. A key is provided, but looking up this key requires the user to break their focus from the displayed data. To alleviate this problem, perhaps a colour scheme based on the rainbow colours might be used. However, an improvement is not guaranteed, as not all users may be familiar enough with such a scheme to be able to offhandedly rank colours accordingly.

Scalability

The original cone tree is a 3D structure, which allows it to make better use of screen estate than traditional 2D tree layouts. As such, larger sites can be displayed at the cost of using a more complicated depiction. However, a limit will be eventually reached when the cone tree becomes unmanageable. In such cases, the possibility of expanding and contracting (hiding) subtrees does exist, as with traditional trees.

Context Maintenance

3D cone trees that do not fit into a single view suffer more than their 2D counterparts, in that the 2D versions can make use of standard 2D context maintenance approaches, whereas context maintenance is more complicated for 3D cases. One standard 2D method is to use an additional "overview" window that portrays the entire site, while indicating the position and range of the

primary "focus" window. Astra SiteManager adopts this practice whereby a rectangular outline is used to indicate the "zoomed region" displayed in the other window. However, the system's approach breaks down in that it is possible for a user to zoom in to such an extent that the zoom outline rectangle becomes too small to perceive. The user thus loses all context and must zoom out until the outline is visible once more.

The original 3D cone tree aided in maintaining context by presenting a single, unalterable view and rotating the necessary cones until the desired sections are in front.

Data Exploration

As already stated, it is not possible to alter the viewing angle of the original cone tree in order to investigate different sections of the tree. Instead, selecting a node has the effect of rotating the various cones comprising the tree, so that the selected node and each node in the path from the selected node to the root are brought to the front. The rotations of each substructure are done in parallel and are animated to aid the user in tracking the transformations. Although this system aids in context maintenance, it does require the user to wait while all the rotations and animations take place.

3D cone trees that do not follow the original's interaction policy of a single viewing angle, permit users to examine the structure from various perspectives by rotating and possibly panning the tree.

2D cone trees display similar browsing features to those of 2D graph and tree displays.

3.3.4 Radial Views

Radial views are hierarchical trees whose nodes are positioned in a radial manner. The root of the tree is located in the centre of a circle or disk. Successive levels of the tree are then mapped to expanding concentric rings around the root node, with subtrees occupying non–overlapping segments of the circle. Examples of tools which describe web sites as bullseye views include IXACTA's Ixsite [24], InContext System's WebAnalyzer [22], CLEARWeb [11] and IBM's Mapuccino [21]. Some of these systems utilise a radial placing of nodes but only display the links between nodes on demand. Mapuccino initially uses a radial setup but allows the user to move the pages around thereby rearranging the layout.

A radial view with displayed links features in the image on the left in figure 8,while the right shows a radial view where links are hidden.

Two interesting variations on the radial theme were developed by researchers at Xerox PARC. The first of these, called a *disk tree*, was developed by Chi et al. [9]. The disk tree is similar to a traditional radial layouts, except that Chi et al. stacked several disk trees on top of each other. Each

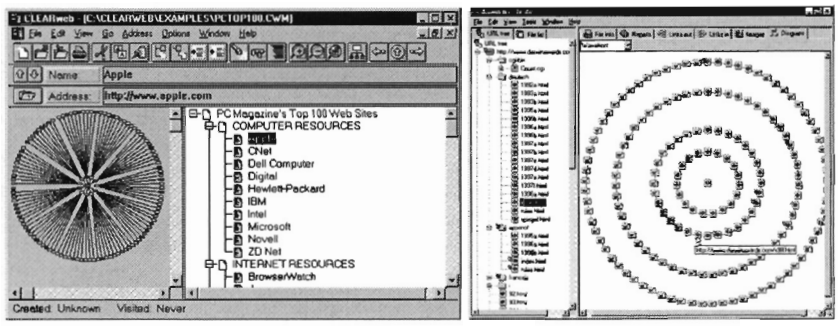


Figure 8: **Conventional Radial Views.** On the left is an image taken from the CLEAR-Web tool [11]. This image shows a bullseye view of a web site that also illustrates the links down the site hierarchy, unlike the right–hand picture, which shows IXACTA Ixsite tool [24].

disk represents a "time slice" created at different periods to form a *time tube*. This time tube thus enables one to view how a web site changed or evolved over time. The second metaphor, which was devised to minimise the inter–crossing of links [10], took a single disk tree and mapped it onto a 3D parabola or dome. This metaphor is known as a *dome tree*. The dome tree has a segment cut out so that one is able to view into the dome.

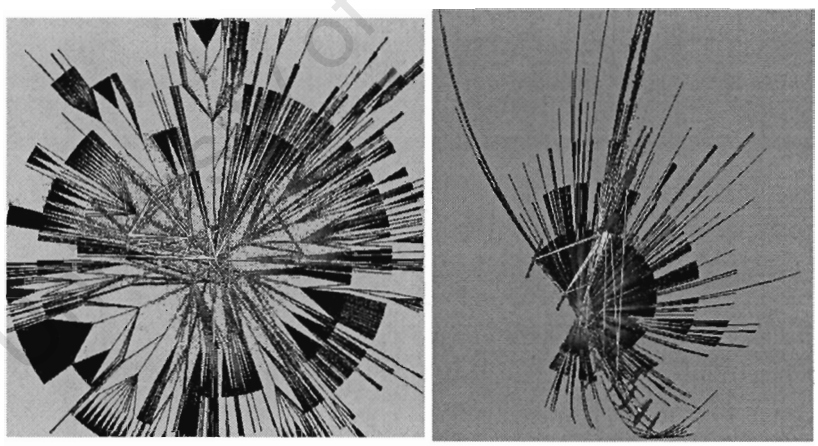


Figure 9: **Chi et al.'s [9][10] Disk Tree and Dome Tree Metaphors.** A disk tree that represents a snapshot in time of a web site, multiples of which can be stacked on top of each other and combined to form a time tube, is shown on the left. On the right is a dome tree, which is a single disk tree that has been mapped to a 3D parabola with a segment missing.

A disk tree is shown in the left–hand picture of Figure 9, whereas the right–hand image depicts

a dome tree.

Structure

As has already been mentioned, radial views present web sites as hierarchical trees. However, because of the altered layout, it is less obvious where the root node and hence the home page of a site is located, as compared to a classical arrangement. Users may thus explore the visualisation in a less hierarchical manner. While this effect may result in new insights into the data, it may also prove to be disconcerting for users who cannot conceptually orient themselves without firm knowledge of the position of the home page.

Two other consequences of this layout are also apparent. Firstly, it is more difficult to ascertain whether nodes occupying separate segments of the circle occur at the same depth as each other. Secondly, if a subtree is present that consists of only single branches it will be represented as a single straight line. Thus, if the node icons are sufficiently small it becomes problematic determining how many nodes and levels that line contains.

Data Representation

Some systems that use radial views dispense with displaying node icons. Instead, they rely on the understanding that at the end of each line representing a link, there exists the node, or page, to which that link leads (node icons may be explicitly displayed for highlighted nodes, however). Accordingly, these systems encode information pertaining to a node in the actual link representation leading to that node.

Chi et al. double encode page access frequency using the colour and size of the links in their dome and disk tree representations. The links are shaded from bright green to black. This is an improvement over the colour scheme used by Astra SiteManager, as the progression from bright green to dark green to black suggests an intuitive scaling system. In addition, patterns in the data, such as areas of high usage, are easy to perceive as light areas against dark surroundings.

One problem with this approach of encoding usage information about a page in a link to that page, is that an impression is created that all the accesses or hits that the page in question received came from that link. However, this may not be the case. For example, say there exists a page to which there are two links; one from its parent and one non–tree link from some page other than its parent. If most of the traffic to the page proceeded along the non-tree link, it would not be apparent, as this link would not be explicitly displayed in a tree representation. Instead, the data could be misinterpreted in that it would seem that the link from the parent was responsible for all the page accesses, as the total page hits would be encoded in the parent link, regardless of where they came from. This may result in a negative impact on the usefulness and accuracy of the site

usage visualisation. Unfortunately, this is a problem shared by all tree representations, due to their limitations in showing non–tree links.

Scalability

By laying out branches concentrically around a central point, radial algorithms are able to utilise screen estate more effectively than classical tree layouts. Again, this means that sites of greater size can be depicted.

Context Maintenance and Data Exploration

Although in the systems listed above, there is no specific support for context maintenance (entire sites are depicted instead), as they all utilise 2D representations, with the exception of the dome tree, there is no reason why a zoom and overview window system cannot be used. Standard 2D data exploration techniques may also be applied.

3.3.5 Hyperbolic Trees

Hyperbolic trees represent another specialised form of hierarchical trees. Unlike conventional hierarchical trees however, hyperbolic trees are laid out using an alternative, non–Euclidean geometry, namely hyperbolic geometry.

Hyperbolic geometry can be distinguished from Euclidean geometry by observing the behaviour of parallel lines: in Euclidean space, for any given point there is exactly one line passing through it that is parallel to a given line, whereas in hyperbolic space there are many. In addition, in Euclidean space, the area of a circle expands linearly with respect to its radius, while in hyperbolic space the area grows exponentially (Munzner provides a more detailed account of hyperbolic space in [36]). Hyperbolic space also causes a visual effect similar to that obtained by a fish–eye camera lens, such as that used by Sarker and Brown [44], which expands a section of interest while displaying the remainder of the structure with successively less detail. These properties all prove to be useful in visualising very large trees, and hence their use in visualising web sites.

Silicon Graphic's Site Manager, which was based on research conducted by Munzner and Burchard [36][35], provides an example of web sites displayed as hierarchical trees in 3D hyperbolic space. However, the 2D form of hyperbolic tree, which is attained by uniformly laying out a hierarchy on an imaginary hyperbolic plane and then mapping the plane to the Euclidean space of a circle, is more common, as is evidenced by its use in products like Microsoft's Site Server [30] and Inxight Software's Tree Studio [23].

Figure 10 shows examples of a 2D and a 3D hyperbolic tree on the left and right–hand sides respectively.

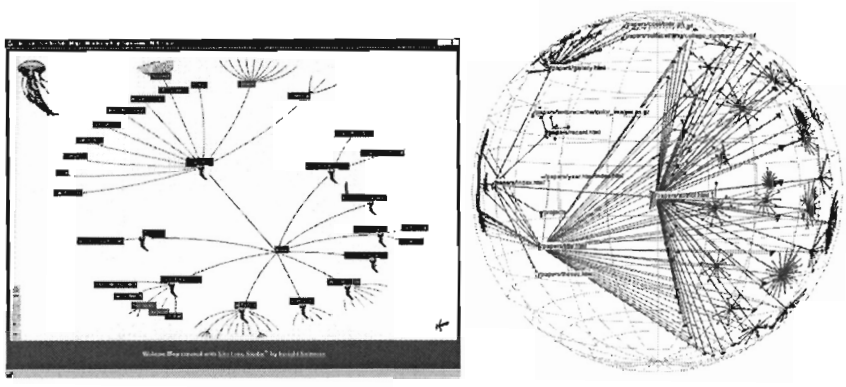


Figure 10: **Hyperbolic Trees.** The picture on the left shows a web site depicted as a 2D hyperbolic tree that was created using Inxight Software's Site Lens Studio [23]. The right–hand image illustrates a hierarchical tree laid out in 3D hyperbolic space that resulted from research conducted by Munzner [35].

Structure

Although laid out in an alternative geometric space, web sites that are represented by hyperbolic techniques tend to be displayed as trees. In fact, many tree layout algorithms including classical and cone tree variations can be reapplied in a hyperbolic setting. Hyperbolic trees thus share many characteristics with their conventional Euclidean counterparts.

Data Representation

One side effect of the distortion caused by hyperbolic space is that the size of the link and node representations are constantly altered depending on the current point of interest. Thus, a line representing a given link may be long at one moment and hardly distinguishable the next. The human visual system generally gives emphasis to larger objects, and given that node and link attributes are commonly used to encode data, the effect of the same object changing size could have an influence on interpretation of the data. This influence could be nullified somewhat by extended exploration. Many systems, such as Microsoft's Site Server, that use hyperbolic views, map data to node rather than link attributes as a result.

Scalability

Due the fish–eye lens effect they have, hyperbolic trees are well suited to depict large sites. For very large sites, structure hiding may still occur whereby branches not currently of interest are not

displayed, but even in such cases more of the structure is in view than would be possible with Euclidean representations.

Context Maintenance

Hyperbolic layouts were designed largely with the "focus+context" distortion in mind. Allowing the user to focus on some detail without losing the context, promotes greater data exploration and aids the user's understanding of the site structure. The use of a single view to achieve this, has the added advantage of permitting users to concentrate on areas of interest without being required to divide their attention between two separate windows in order to maintain context.

However, if structure hiding is implemented, context information can be lost. In Inxight's Tree Studio, which utilises a 2D hyperbolic view, viewing nodes that are very deep in a large tree results in the ancestor nodes higher up in the tree being concealed. It is thus no longer possible to identify which part of the site is currently in focus, without moving the focus back to where earlier nodes are brought into view.

Data Exploration

Navigating a hyperbolic tree involves the translation of the focal point of the visualisation. Shifting this point could be achieved in a number of various ways, such as causing a selected node to become the focus or else allowing the user to use a mouse to "grab" and move the point freely. Because of the nature of the hyperbolic layout, drill-down operations are generally limited to nodes within the current focus. However, as users are typically interested in acquiring further details about nodes they are currently focusing on, this does not present any difficulties.

Difficulties may be encountered with very large scale sites, however. When a certain size limit is reached, a zoom feature might be necessary to observe minute details, as the distortion caused by the hyperbolic effect may not expand the area of interest enough. However, the addition of a zoom ability would defeat the initial purpose of the focus+context distortion that hyperbolic trees were designed for.

3.3.6 Other Metaphors

Although the metaphors already described are those that are seen most frequently, there do exist other, less popular means of representing web sites.

One example of these is the *treemap* conceived by Schneiderman [45]. This approach makes use of a 2D space filling algorithm in which each node is denoted by a rectangle, the area of which is proportional to some node attribute.

Another metaphor is the perspective wall developed by Mackinlay et al. [27]. This metaphor requires data that is linearly organised along at least one dimension, such as date or time. The different walls then show data items which correspond to these different dates or times. The user focuses on one wall at a time with the others disappearing into the distance on either side.

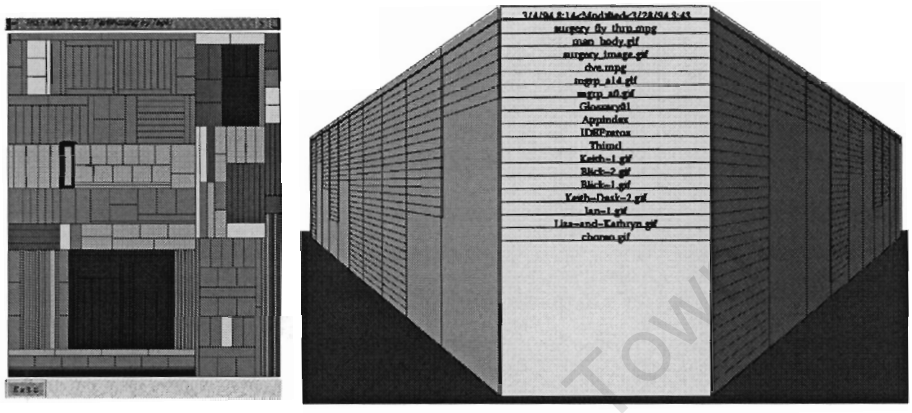


Figure 11: **The Treemap and Perspective Wall Metaphors.** The left image shows the Navigational View Builder's treemap view of a hierarchy of research pages [32]. The node colours represent author types. The right–hand image, also taken from [32], depicts a perspective wall view of a linear arrangement of files based on the last modification time. The different walls show files which were last modified in different time frames.

Both a tree map and a perspective wall representation can be found in Mukherjea and Foley's Navigational View Builder tool [32] and are shown on the left and right–hand side of Figure 11 respectively.

These metaphors will not be considered further as they were judged to be ill-suited to the task of representing web site structure. This is due to the difficulty of showing links, both hierarchical and non-tree in their current incarnations.

Having examined previous works in the field of web site visualisations, one common trait has become apparent. The problem has always been approached by treating web sites as normal trees or graphs and no attempt has been made to exploit any site features that are unique. Therefore, there exists a motivation to design and develop a metaphor that specifically caters for web sites. Such a metaphor is proposed in the next chapter.

3.4 Summary

In this chapter, previous efforts in the field of web site visualisation were examined.

Attempts have been made to visualise web sites for many purposes including aiding web site navigation, improving web search results and promoting greater understanding of web site usage.

Regardless of their purpose, web site visualisations are required to address the following factors: *structure representation*, or the manner in which the site structure is depicted; *data representation*, which is how information is encoded; *scalability*, which refers to how well a system handles large scale sites; *context maintenance*, which deals with whether contextual information is retained while focusing on details; and lastly, *data exploration*, which regards the techniques used for navigation and exploration.

Metaphors created to represent web sites can be grouped into certain classifications according to their appearance and behaviour. These include *cyclic graphs* representations as well as the various tree metaphors such as *classical hierarchical trees*, *cone trees*, *radial views* and *hyperbolic trees*.

These various classes of metaphors generally approach structure representation by depicting web sites as either cyclic graphs or hierarchical trees. Data representation is achieved by encoding information in node and link attributes including size, shape, colour and pattern. Scalability is addressed by attempting to efficiently utilise screen estate, as well as by reducing the amount of content required to be displayed. Context is maintained through either the use of an additional overview window or else by the use of focus+context techniques in a single window. Data exploration is typically dealt with by providing zoom, pan and rotation features and drill–down operations are mostly performed by mouse selection of items of interest.

Previous works have displayed a trend to regard web sites as normal trees or graphs. Breaking this trend presents an opportunity for creating a web site metaphor that is more specialised and effective.

Chapter 4

Metaphor Development

Devising a web site metaphor that effectively addresses all the key factors of structure and data representation, scalability, context maintenance and data exploration is a nontrivial task. Indeed, it is questionable whether it is possible to create a "perfect" representation that is versatile, simple to use and quick to understand. Instead, designing a metaphor usually involves certain compromises and tradeoffs, as was evident from previous efforts discussed in the preceding chapter.

Much can be learned from previous attempts to balance the factors mentioned above. By adapting past proposals, so that they are more suitable for visualising web site usage in particular, and combining them with novel ideas and the approach of treating web sites as special entities rather than as typical graphs, a metaphor that is more effective than those currently in use may be developed.

This chapter describes an attempt made to create such a metaphor.

Section 4.1 describes the design decisions taken while developing an initial metaphor. Section 4.2 then provides a summary description of the resulting metaphor. This is followed by Section 4.3, which discusses the initial experiences and evaluations of the initial metaphor. After serious flaws were found in the initial metaphor, a second metaphor was developed, the design of which is outlined in Section 4.4. Section 4.5 then gives a brief description of the final metaphor, after which Section 4.6 discusses the initial experiences of the final metaphor. Finally, Section 4.7 provides a summary of the chapter.

4.1 Initial Metaphor Design

This section describes the initial attempt made to develop a web site metaphor for the purpose of visualising site usage. At the time of its conception, attention was focused primarily on the early goals of displaying the site layout, as well as presenting usage statistics that may have been found

in typical log analysis tools.

The development of the metaphor will be discussed according to the factors affecting web site visualisation that were outlined in Section 3.2 of Chapter 3. The manner in which each factor was approached will be treated separately in the relevant sections below.

4.1.1 Structure

The first decision to be made involved whether to represent a web site as a hierarchical tree or as a cyclic graph. Both representations have their advantages and disadvantages. For instance, tree metaphors generally scale better and may bear greater similarities to users' existing conceptions of the site layout but exclude vital information concerning a site's link topology. On the other hand, graph metaphors embody a more accurate representation and contain additional details about page connectivity. They are, however, prone to problems such as cluttering for large scale sites.

Bearing these considerations in mind, it was ultimately decided to utilise a tree representation, due to both the need to deal with increasingly large web sites as well as to the perception that trees provide greater familiarity. Accordingly, site structure is exhibited in the initial metaphor through the use of a hierarchical windows directory tree, such as those typically found in popular file system browsers. The directory tree is easily adapted by simply replacing file and folder icons with page icons and the associated page names or URL's. The resulting 2D structure representation is shown in Figure 12.

Employing such a device has several benefits. For example, as a result of their widespread application, especially in terms of depicting file structures, these directory trees are well understood by most computer literate people. In addition, directory trees are able to represent reasonably large scale hierarchies, as due to their mechanism of expanding and contracting individual branches, only partial sections of a tree need be shown at any particular instant.

Adopting such an approach does present certain difficulties, however. In particular, a means of displaying detailed information as well as a method of handling non-tree links must be found.

In terms of presenting more information, the size of the page and link representations in a directory tree precludes the possibility of illustrating numerous details concurrently. Instead, it is likely that only a single item of information, or variable, will be displayable at a time. As a result, the simultaneous portrayal of multiple variables needs to be accommodated separately. One viable option for achieving this is to include several details pertaining to a web page in a table, which is then presented whenever that particular page's icon is selected. However, a drawback of switching directly from a directory tree view to a table is that it is possible that a loss of contextual information will occur. A more favourable alternative might be to use an intermediate representation and then to only pop up the table following an extended drill-down operation. In this manner, a user could

obtain a slightly more detailed view of a subsection of the site, decide which page appears to be the most promising and then view the table for that page. Comparisons between multiple data items could also be performed. To this end, the initial metaphor contains an additional, intermediate view of a portion of the site structure. This view consists of two alternative layouts, namely a flat classical tree layout and a 3D cone tree layout (these can be seen in Figure 13a and Figure 13b respectively). Regardless of which layout is currently activated, the selected page forms the root node of the tree. Links from pages that seemingly lead nowhere are simply indicators that the associated page is not a leaf node, i.e., it contains a sub–tree of descendants that are not in view. Of the two views, the classical tree presents a layout that is simpler, requires less manipulation and makes direct comparisons between two pages within the view easier. Conversely, the cone tree layout requires increased interaction by means of necessary rotations in order to locate suitable viewing points but makes more effective use of the available screen estate. Both layouts provide a view of a fragment of the site topology while simultaneously depicting several usage statistics.

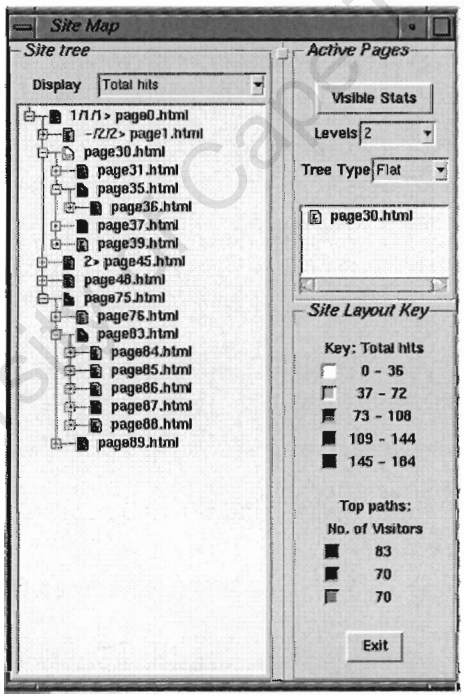
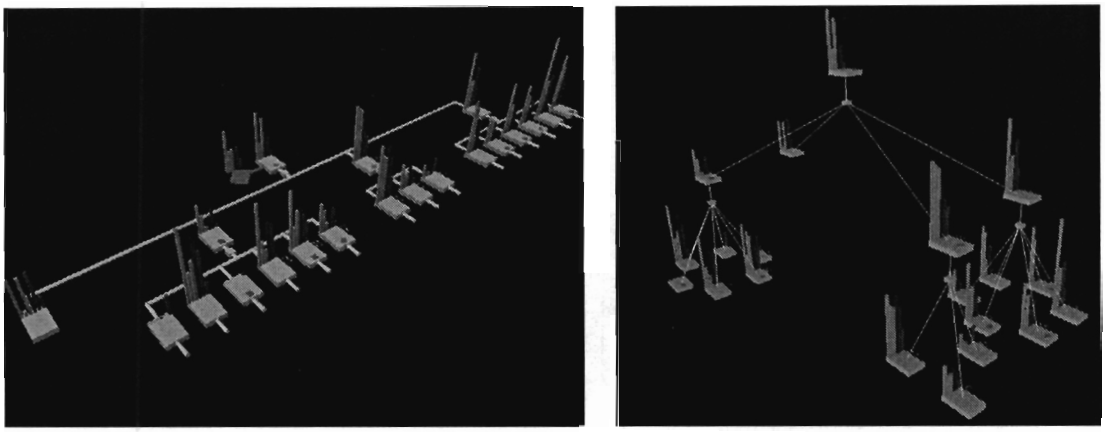
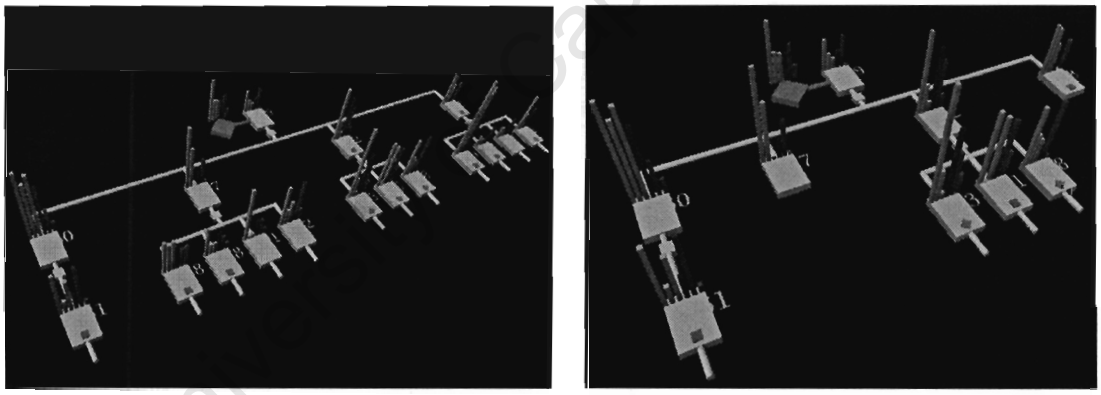


Figure 12: **Directory Tree View.** This view provides an overview of the entire site using a windows directory type browser. Page icons are colour coded on a grey scale according to the amount of traffic that page received. Page icons can be dragged and dropped with the mouse into the box labeled "Active Pages" (on the top right) in order to access the intermediate and more detailed view of that page.



(a) (b)

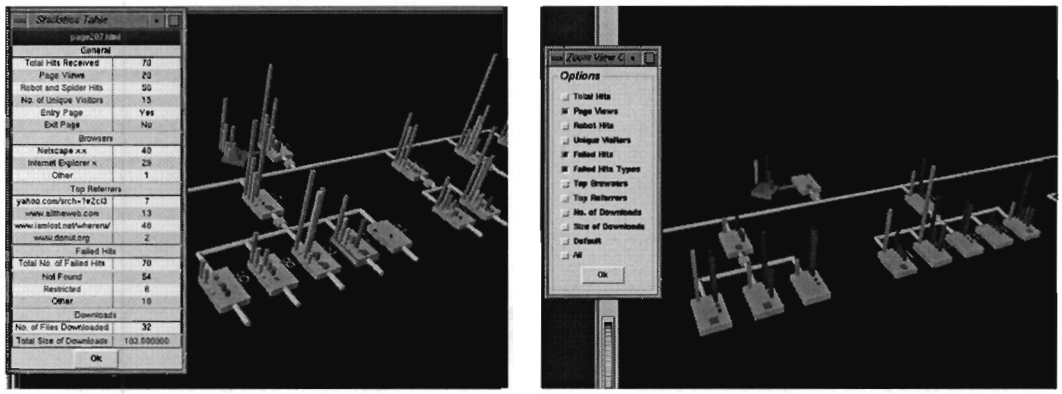
Figure 13: **Intermediate Tree View.** The intermediate view provides a more detailed view of a page and up to 3 levels of the hierarchy below that page. Page icons have bars present on them, representing various usage statistics. This view can be toggled from a flat hierarchical tree view (a) or a 3D cone tree view (b).



(a) (b)

Figure 14: **Focusing the Intermediate Tree View.** The amount of pages displayed in the intermediate view can be reduced by hiding or "pruning" pages or branches. This is performed by clicking on the small square below a page icon which will cause that pages below that page to be hidden. (a) An intermediate view with no pruning. (b) The same view with a branch pruned.

Having decided on a method for displaying detailed views of subsections of a site, attention was turned to solving the second issue resulting from the use of a directory tree, namely how to portray non-tree links. This task presented a greater challenge, as even though non-tree links would,



(a) (b)

Figure 15: **User Interaction.** (a) In order to view detailed information concerning a particular page, the user enters pick mode by pressing escape and then clicking on the page of choice with the mouse. This will cause a table containing usage statistics for the chosen page (which will have a red outline) to appear. (b) The statistics which appear on the page icons in the intermediate view can be selected and changed by the user.

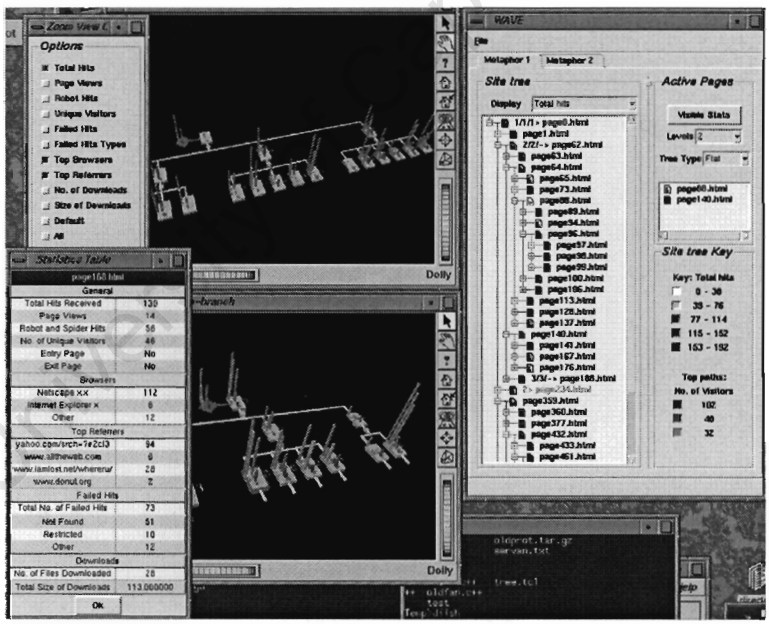


Figure 16: **Entire System View.** This figure shows the directory tree overview, two intermediate views and a table

by necessity, only be shown on demand, directory trees were designed solely for the purpose of displaying hierarchical structures. Thus, a difficulty exists with regards to the overlaying of link representations on a hierarchy without causing clutter. One plausible suggestion for resolving this dilemma would be as follows: when a page is selected, all the branches containing pages that are reachable from that page (i.e., those pages to which the original page has a non-tree link *to*), would be expanded and the respective page icons altered to reflect this (for instance, by altering the border colour of the icon). For example, if a page is selected that has a non-tree link to a page in another branch, that other branch would be expanded and the page that is the destination of the non-tree link would have a different border colour (say blue, if the normal border colours are black). Non-tree links *from* the current page of interest (i.e., links where the current page is the destination page) could be handled in an identical fashion, with appropriate changes to the border colours of the page icons representing the origins of the non-tree links. If the situation arises that, due to user interference, a branch containing a page with a non-tree link is then contracted, an arrow could be drawn from the selected page to the lowest visible level of that branch. The indication would then be that the branch being pointed to includes pages that are linked to the currently selected page.

4.1.2 Data Representation

Depicting the site structure as a directory tree limits the choices available concerning data representation. As the lines representing links are comparatively small and therefore difficult to clearly perceive, the remaining entities suitable for encoding data are the page icons and page names. Of the alterable attributes belonging to these icons and names, size was considered unsuitable for representing data, as aside from the fact that the size of the page icons and names are linked (representing the same page with a large name but a small icon could prove to be distracting), varying the size attribute could provide users with a distorted idea of the site layout. This effect would be accentuated if a user either switches to a different variable (recall that it is likely that only a single variable will be viewed at a time) or else decides to invert the scale by placing emphasis on, say, low data values as opposed to high. The same section of the site could thus be enlarged, and thereby made more noticeable during one viewing, only to be reduced in both scale and importance in another. Although transferring emphasis from one segment of a site to another to indicate varying data items and values may be desirable, doing so at the cost of potentially creating the impression that a topological alteration has occurred, is not.

If size is discounted, the remaining attributes that may be suitable for embedding information are shape, colour, transparency and possibly texture, although the use of texture and colour is generally mutually exclusive. Of these, variations in colour is probably both visible and distinguishable from the greatest distance. Since this is a valuable characteristic, especially in terms of discovering

patterns in the data, colour was chosen as the property to represent data. Accordingly, the page icons in the directory tree are colour-coded to reflect the values of a variable, which by default is the number of page hits or accesses. The scale used is a grey scale ranging from white for pages with a high number of hits, to black for pages with few or no hits. This grey scale was employed (as opposed to a scale consisting of different hues such as a rainbow scale) because it provides a more intuitive value indication [49].

Given the importance of the number of page accesses as an indication of site usage, it would not be unreasonable to double encode page hits in both the page icons and their names. However, the number of page hits was considered evident enough in the colour of the page icons to render this redundancy unnecessary. Instead, the page names are used to provide information about popular user paths though the site. If a page belongs to one of the top three paths then its page name is colour coded accordingly, with brighter shades indicating more frequently followed trails (see Figure 12). In addition, the page name is preceded by a number that denotes that page's position in the path. In the case of a page belonging to multiple paths, the name is shaded according to the most used path and is preceded by multiple numbers. The first number refers to the location of this page in the most popular path, the second number to its position in the next most used path and so on. Glancing at the page names will thus provide the user with an idea of the paths taken by the majority of visitors to the web site. Insights about these paths is important as understanding the manner in which users browse a web site allows designers to improve the site design.

The intermediate view was designed to present the user with more information about a subsection of the site than was available using the directory tree. To facilitate this, pages in this view are represented by rectangular blocks. Present on each block are a number of pillars, each of which corresponds to a particular usage statistic (Figure 15b). The heights of these pillars or bars are altered to signify the values of the associated variables. Theoretically, the pillars could have been projected flat onto the surface of the blocks thereby resulting in a 2D view instead of a 3D one. However, adopting such an approach would have greatly limited the number of variables which can be depicted. The arrangement of the pillars is such that the major statistics are positioned in a row along the top of the block with derived variables placed in a column below the appropriate base statistic. For example, if a pillar in the top row represents the total number of failed hits, then the number of failed hits due to *file not found* errors and *access denied* errors would be represented by two pillars in a column below the top pillar. In this manner, statistics with the highest values, and therefore the tallest pillars will be at the top, or back, of the block with shorter pillars in the front. The pillars that are present at any stage is determined by user-defined options. In addition, to the pillars, the page blocks also contain a small icon indicating the page type. There exist four different possibilities: a pink diamond representing an entry page, a blue square denoting an exit page, both of the above icons showing an entry-exit page and no icon at all, which symbolizes an ordinary page

(refer to Chapter 2, Section 2.1 for a description on page types). Finally, attached to the root page of the layout is a green diamond. This diamond supports pillars whose heights indicate the average values of the corresponding variables for the entire sub-tree.

4.1.3 Scalability

When dealing with large scale web sites, depicting the entire site at the same instant is not a viable option. Instead, the amount of information shown needs to be made manageable either through the use of *hiding*, whereby sections of the site are not displayed, or else *clustering*, which involves replacing a cluster of nodes by a "super-node" representation. The directory tree chosen to represent the site structure makes use of the former method in that the majority of the site is hidden, except for the top layer of the hierarchy, which corresponds to those pages that are reachable from the home page. Pages belonging to the remainder of the site are revealed when the appropriate branches are expanded. When a sufficient number of branches are expanded so that the tree no longer fits inside the window, vertical and horizontal scrollbars enable the user to move the view in order to contain the desired subsection. In this manner, fairly large hierarchies can be accommodated. Furthermore, since each page is allocated its own space and location, displaying strict hierarchies will never suffer from cluttering.

With regards to the intermediate view, in order to remain manageable, the size of the sub-tree displayed is required to be limited. This is accomplished by restricting the user-defined number of levels of the sub-tree to portray to a maximum of three levels. However, if the sub-tree being viewed is very broad, an unwieldy visualisation could still result. In such cases, provision has been made for hiding sections of the display by "pruning" unwanted branches. Pruning is performed by clicking on the small squares present on each link. Doing so will result in the sub-tree leading from that link to be hidden (Figure 14).

4.1.4 Context Maintenance

Context maintenance is generally achieved by one of two means, namely using a single view that incorporates a focus and context technique, or by using a multiple view system consisting of a zoom and an overview window. Since the intermediate view is presented in a separate window, context will be maintained using the latter option. The border of the currently selected page icon is shaded red, thereby indicating the relevant page's relative location in the site. Red was selected as the border colour as its appearance against the grey-scaled icon interior is prominent from a distance.

4.1.5 Data Exploration

Data exploration inside the windows directory tree should be self-evident to any user with previous experience of such a device. Indeed, this well-defined behaviour is one of the reasons a directory tree representation was selected. Data exploration thus consists of expanding and contracting desired branches of the site until a page which provokes further investigation is encountered. Once this has occurred, the page in question needs to be rendered with further detail in the intermediate view. This switch to a new view could have been accomplished by having the user simply select the page of interest by clicking or double clicking with the cursor placed over the page icon or name. However, it was decided to provide increased feedback of the pages being examined by placing copies of the name and icon in a separate text box. In order to activate an intermediate view of a page, a user must therefore drag and drop either that page's icon or name from the directory tree into the text box. This text box is known as the "active area" and a page can be said to have been made "active" if it was selected and placed in this box (Figure 12). Once the page has been activated, a separate window will appear in which a sub-tree with the active page as the root will be rendered. The layout utilised will depend on the option selected by the user. Deselecting a page in the active box will have the same result as closing the window containing the intermediate view of that page. Exploration within an intermediate view is achieved with a combination of mouse buttons and movements. Available interactions include rotation (left button and mouse move), pan (middle button and mouse move) and zoom (left and middle buttons and mouse up and down). Up to a total of three pages can be activated at any one time for comparison, after which it was deemed that the screen became too cluttered with multiple windows.

The intermediate view provides only an approximate estimate of usage statistics values via vertical bars. Should users desire actual figures, they are required to access the table view. The simplest and most intuitive method of achieving this would be to present the table for a page when that page is selected with the mouse. Thus, in order to bring up a table of data figures users must press the escape key to enter selection mode (as opposed to manipulation mode), position the mouse pointer over the desired page and then click the left button. The resulting table appears in a new window (Figure 15a). Any of the windows can be moved and placed according to the users discretion.

Now that the design approach to each factor has been discussed, a summary description of the metaphor is provided in the next section.

4.2 Resulting Metaphor – Directory Tree Metaphor

The metaphor resulting from the initial manner in which each factor was addressed can be summarised as follows:

The metaphor incorporates a tree–based approach that consists of two separate views. The first of these contains a windows style directory tree that provides the user with an impression of the overall site layout. Based on usage information that is embedded in the colour intensity of the page icons, the user decides on a subsection of the site for which s/he requires further details. The selected sub–tree is then viewed in the second display, which provides a greater level of detail. This second view is accessed via dragging and dropping the page icon of a node into a box on the interface representing the "active" pages list. The sub–tree, whose depth is dependent on a user–specified preference, headed by the active page is then rendered as a 3D object in a second window. The object has the appearance of a hierarchical tree that is laid out as either a flat 2D tree or else a 3D cone tree. Nodes that do not belong to an area of interest can be pruned in order to maximise screen estate. Various usage statistics can then be discerned by examining the height of the bars extending from each rectangular page icon. The decision of which statistics to present, is again, user defined. Averages for the statistics visible for the displayed sub–tree are encoded as height variable bars present on a green diamond that is attached to the sub–tree root node. Finally, if the user wishes to view all the available information concerning a particular page, s/he is able to select that page's rectangular icon in order to prompt the appearance of a table containing that page's details.

4.3 Initial Experiences and Outcomes of Initial Metaphor

Once the initial design of the directory tree metaphor was completed, a decision was required as to whether to adopt this metaphor and develop it further or to abandon it in favour of a more promising one. To this end, a prototype system utilising the metaphor was built, so that an evaluation concerning the metaphor's effectiveness and potential could be made. This process, as well as the outcome will be described in the sections to follow.

4.3.1 Prototype Implementation

Before investing in developing a full web site usage visualisation, the metaphor was first employed in a prototype system. The system, which was implemented using an inhouse development kit linking Open Inventor and Tcl/Tk called DIISH, ran under Unix on an SGI O2.

Once implementation was complete, the prototype was assessed using a test web site and simulated usage data. In order to evaluate the effectiveness of the metaphor in representing web sites of varying scales and arrangements, the test web site was randomly generated each time the system was run. After the site was created, the structural informational for each page, including items such as parent and children pages as well as incoming and outgoing links to the rest of the site, was stored in a data structure. A base value was then generated for the total hits for each page. Although

this value was mostly random, it was constrained by factors like the page's position and type (for example, it makes no sense for a page to receive a higher number of hits than its parent if it is not an entry page and there are no links to it aside from its parent). Once the total hits were obtained, several other usage statistics were generated, using the total hits value as a basis.

Although the resemblance of data obtained in such a manner to real usage data for actual sites is difficult to establish, visualising the simulated data does at least provide an estimate of the metaphors effectiveness.

4.3.2 Informal Evaluation

Aside from the informal practice of running the system multiple times and viewing the results obtained for web sites of different sizes and layouts, feedback was also obtained through the means of a heuristic evaluation by an expert from the web site usage industry as well as a demonstration to a human computer interaction expert.

The outcome of these processes was that several flaws in the metaphor were exposed. These weaknesses were deemed serious enough that it was decided to discard the directory tree metaphor, as it failed to achieve our stated objectives. Further evaluation was thus not required. The metaphor's identified drawbacks are discussed in further detail in the next section.

4.3.3 Weaknesses of the Directory Tree Metaphor

The weaknesses of the directory tree metaphor will, once more, be discussed in terms of the factors that a web site usage visualisation must address in order to be effective. Although the devised metaphor satisfactorily dealt with some of these factors, its approach to others cannot be deemed successful. The factors where the metaphor's weaknesses were exposed are discussed below.

Structure

The main weakness of the proposed metaphor is one that is induced by the utilisation of a windows style directory tree. While such a structure may be sufficient for visualising a file system, where one is more concerned about the full path of the current file being accessed, it provides a poor indication of the overall system layout. This is due to the fact that, by default, nodes that are not children of the root node, are hidden from view unless the user explicitly expands the sub–tree containing those nodes. Thus, while using such a device for visualising web sites is well suited to identifying the path from the home page to the currently selected page, it does not promote a good understanding of the site's overall structure.

Another problem is that of displaying non–tree links on a 2D representation without causing cluttering and confusion due to inter–crossing links. Complicating the matter further is the strong

possibility that the origins or destinations of certain links are not even in view as they belong to sections of the tree which have yet to be expanded. The mechanism of depicting non–tree links that was proposed in Section 4.1.1 may actually cause more difficulties than it solves. Having been responsible for expanding and contracting desired sections of the site, the user should have an accurate mental impression of the context of the visible portions of a site. However, if s/he then selects a page which has non–tree links to areas of the site that s/he has not exposed, those areas will be automatically expanded to show the non–tree links. The user will thus suddenly be confronted with multiple pages that have not been seen before. In addition, the revealed segments of the site would have increased, possibly even to an extent that the sections the user was familiar with are no longer in view but off–screen. These side effects are all likely to confuse the user, at least until they have readjusted to the alterations. This lack of an effective method of displaying non–tree links severely hampers the metaphor's handling the vital factor of displaying site structure.

Scalability

Although a windows style directory tree is theoretically not restricted by the scale of the site being visualised, in practice the opposite is true. Thus, even though the scrolling mechanism does not limit the size of a displayable site to the available height of the screen, scrolling up and down to locate items of interest becomes too taxing once the site reaches a certain size. Expanding and contracting multiple branches also becomes unmanageable for sites consisting of upwards of several hundred nodes. Visualising sites consisting of several thousands of nodes is therefore out of the question, which severely limits the metaphors usefulness.

Data Exploration

The directory tree window was also designed to be the main tool for exploring the usage data of a site, as it is only when a user locates a potentially interesting section of the site by examining the page icon intensities that s/he will drill–down for further details. However, continuously expanding and contracting nodes and then scrolling up and down to expose new sections of the site proved to be a cumbersome and time–consuming task. In addition, as the visible sections of the site are limited to those sub–trees which are currently expanded, discerning patterns and trends in the data is also problematic. Thus the manner in which data exploration is handled is also inadequate.

4.3.4 Outcome

All things considered, even though the directory tree metaphor adequately deals with certain factors like data representation and context maintenance, the metaphor's addressing of factors such as structure representation, data exploration and scalability are not acceptably effective. Since attempting

to improve the metaphor's handling of these factors would require comprehensive alterations, it was decided that a new metaphor, better suited to our aims, was required. The design process of this new metaphor is outlined in the next section.

4.4 Final Metaphor Design

With the previous metaphor proving to be insufficient for our needs, a new representation of web sites was needed. This section describes the design of this new metaphor.

As with the description of the initial metaphor design, the development of the second metaphor will be discussed according to the factors affecting web site visualisation that were outlined in Section 3.2 of Chapter 3.

4.4.1 Structure

As with the initial metaphor devised, the first design decision to be made involves the choice between utilising a tree or a network graph representation. Past efforts, in adopting one or the other, have had to make a trade-off between the cluttered accuracy of a network graph and the more discernible yet less correct representation of a tree. Ideally, the best representation of site structure would be a compromise that contains the strengths of both.

There are two approaches to achieving such a compromise: use a network graph as a base representation and devise a method of displaying less structural information on it, or else use a tree as a basis and show more. Either approach is non-trivial, as is evident in the current lack of existence of such a hybrid representation. This lack is possibly due to the popular conception that a web site is either a graph or a tree. However, as pointed out in Section 2.1, web sites do possess unique characteristics that orthodox trees and graphs do not. Among other features, these include an organisational home page, which is intended to be the entry point into the site for the majority of users, and a navigation bar, which contains links that are accessible from most of the web site. Utilising these features, a second metaphor was designed.

A tree representation was chosen for a basis, based on the same reasoning for designing the initial metaphor as a tree, namely greater familiarity and the perception that users conceptualise a web site as a tree rather than a graph. Having chosen a tree representation, the next step was to decide how to modify the tree in order to portray further structural information. Actually depicting increased information on the tree itself (say by rendering non-tree links) would not represent an improvement, as the increased clutter would negate the main advantage of utilising a tree in the first place. However, due to the manner in which typical web sites are structured, such as their use of navigation bars, there exists the possibility of providing extra structural information through the

implication of links without having to physically represent them.

The devised metaphor is as follows. It consists of a vertical column to which several lines are attached. From each line horizontal "fans" of branches extend (Figure 17a). The vertical column represents a global navigation bar and the lines along it are the links to pages that form that bar (referred to as *global* links). The navigation bar is automatically derived from the site topography as described in Section 6.4.2. The home page is represented by a cube placed at the apex of the column. Each fan of branches then represents a sub-tree consisting of those links and pages accessible from one of the global links (this sub-tree is only approximated by a fan when the viewed from a distance and is represented by lines and node icons at closer views). The metaphor is thus a modified tree, the difference being that it is understood that every page contains links to the pages comprising the navigation bar (column). These links between the pages of the web site and the navigation bar pages are therefore *implied* and thus do not have to be explicitly illustrated. If a page is not linked to the navigation bar (i.e., that page does not contain links to the navigation bar on it), then the link to that page is represented by a stippled rather than a solid line. Any remaining links that are neither implied nor already depicted are then shown on request for individual pages, as is commonly found in many existing tree visualisations of sites.

Another common feature of web sites that was mentioned in Chapter 2 is that they often contain sub-sites. Whenever a sub-site is encountered, it is illustrated by the presence of its own navigation column (Figure 18a). By definition, all pages on the fans of this sub-site are no longer connected to the global navigation bar of the main site but are instead joined to their own navigation column.

In order to clearly display the site structure, each fan is restricted to a semicircle, which limits the extent to which one part of a single fan obscures another part of the same fan. Positioning the columns representing the navigation bars perpendicularly to the fans does introduce the requirement of the use of a third dimension. As a result the metaphor is more complex and potentially less usable than a 2D representation would be. The metaphor could conceivably be "flattened" into a 2D representation, provided that a suitable layout algorithm could be found so that each fan was allocated a sufficient amount of screen estate. The loss of a dimension however, would require that navigation bars be indicated by some other means, which would result in one less attribute being available to depict other information. In addition, less space would be available for the rendering of non-tree links that cannot be implied. As such, it was felt that the utilisation of the third dimension was justified.

4.4.2 Data Representation

Abandoning the windows directory tree representation used in the initial metaphor had the consequence of the size of the node icons being decreased. In fact, when viewed from a distance the node

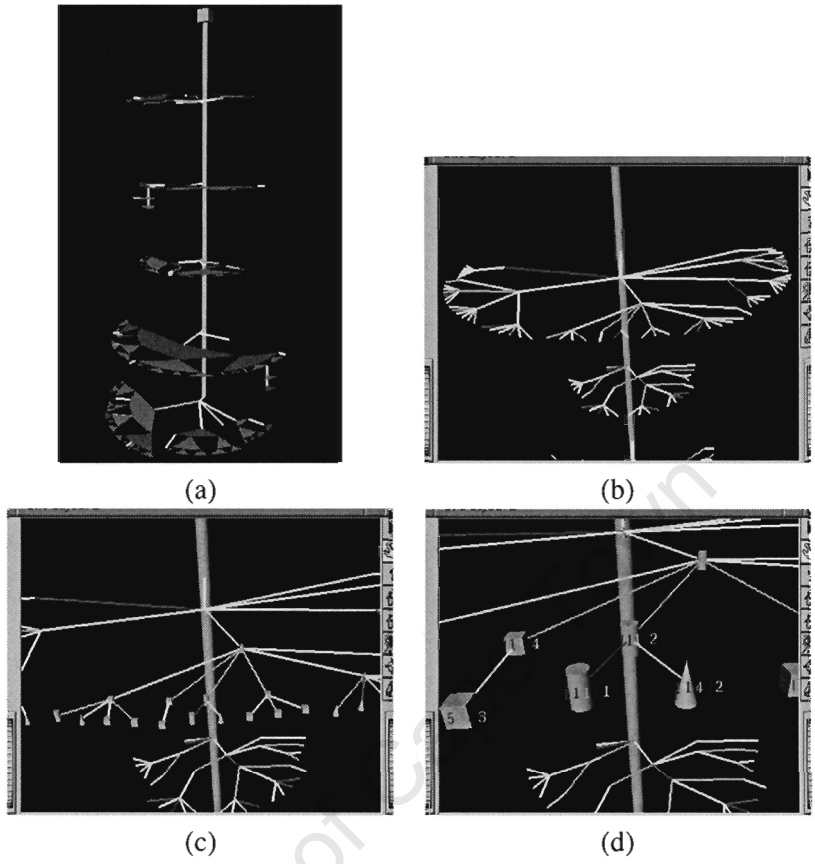
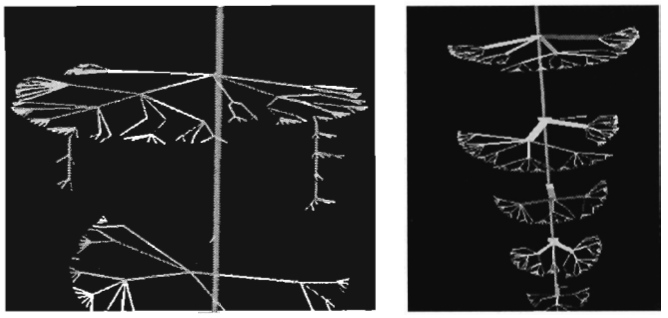


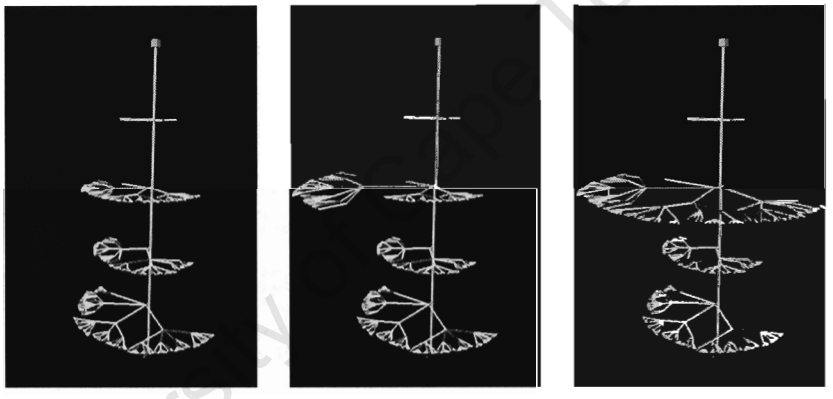
Figure 17: **Level of Detail.** (a) The metaphor consists of a vertical column with a cube, representing the home page, at the apex. Horizontal "fans" extend from the column represent branches of pages. The lines attached to the column represent links to pages forming the global navigation bar. At a distance, the branches are approximated by "wedges", whose intensity represents the average usage of the pages in that branch. (b) At closer inspection, individual lines representing links are discernible. Line intensity indicates the usage of the associated page. (c) Moving closer still results in the individual page icons representing different page types appearing. (d) Finally at the highest level of detail, numbers indicating value ranges for various usage statistics for each page appear on the relevant page icon

icons of the new metaphor could not be distinguished at all. The lines indicating the links between pages however, are discernible from a much greater distance. As a result, data pertaining to a particular page is encoded in the line corresponding to the link to that page. All links are shown in yellow, with the usage for the associated pages depicted using a mechanism similar to [10] and [9], namely by varying the intensity of the links according to the number of times that page was viewed. Links to pages which received no traffic, however, are indicated by red links, so as to emphasise



(a) (b)

Figure 18: **Structure Indication.** (a) Sub-sites are identifiable by their own vertical columns or navigation bars. (Two are visible in this screenshot.) (b) The width of a line indicates the total usage of the associated page and every page in the subtree of which that page is the root.



(a) (b) (c)

Figure 19: **Zoom Options.** Users have the option of enlarging either an entire fan or else an individual branch of interest so that it may be more easily examined (a) The normal view. (b) Enlarging a single branch. (c) Enlarging an entire fan.

pages that are potential problems.

The line width is utilised in a manner akin to that used by [19] to indicate areas of complexity in general trees, whereby the thicker the line, the greater the size of the sub–tree at the end of that line (Figure 18b). However, this approach has been modified so that the line width adjusts depending on the usage of the sub-tree (made up of all the children pages and branches) leading from the link represented by that line. Although line width has been used by previous systems, such as [40] and [10] ([10] double encodes page usage in both the intensity and thickness of a line), to represent

usage information, an important distinction is that by using line width to represent usage for entire sub-trees, the user is able to obtain information about *areas* of a site as opposed to individual *pages*. Thus, if two sub-trees of a site that consisted of approximately the same number of pages were compared, it would be apparent which sub-tree represented an area of greater interest by observing which line corresponding to the link to the respective root nodes of the trees was thicker. In order to acquire the same information from a visualisation utilising line width to portray the usage of a single page, one would have to inspect the lines to each of the nodes comprising the two sub-trees.

An essential aspect of approaching data representation is the ease with which patterns may be discerned in the data being visualised. While making use of line width and intensity certainly contribute to the pattern detection potential of the metaphor, there is still the danger of the user suffering from cognitive overload when faced with a large visualisation consisting of thousands of lines. To this end, when the visualisation is viewed from a certain distance, the individual branches of the site are no longer represented by lines but are instead denoted by polygonal wedges. The intensity of these wedges then indicate the average usage of the pages comprising the branch that they represent. In this manner, general areas of varying usage can be more readily perceived from a greater viewpoint, allowing users to view more of the site and gain an impression of overall usage. Once, the viewpoint has moved closer, the wedges are reconstituted into individual lines representing the links to the branches' pages (Figure 17b).

A disadvantage of portraying information on the links to pages rather than on the pages themselves is that only a single variable can be shown. While this should not prove to be all that problematic, given that in terms of web site usage visualisation users are likely to be only interested in a single variable at a time, some manner is required to show additional details. By necessity, further details are provided on a per page basis, and are depicted on the individual page representations. These page representations consist of varying icons, depending on the page's type:

- Spheres, representing entry pages.
- Cylinders, representing exit pages.
- Cones, representing entry/exit pages.
- Cubes, representing normal pages.

In order not to cause too much cluttering, page icons are only rendered when the viewpoint is sufficiently close so that the icons are distinguishable and so the number of pages in view is decreased (Figure 17c). At such a viewing range, users should be able to make certain deductions concerning visitor patterns. For example, if a page is shown to be an entry page and has a large number of activity (as determined by the link intensity) it could indicate a design issue if that page

was not designed to be an entry page. When the viewpoint is at the closest range to a page, then additional information about the page will be shown by numbers appearing on the page icon (Figure 17d). These numbers correspond to the range of the values for various variables (e.g. failed hits).

Using numbers to represent variable value ranges allows easier comparison with values for different pages. By looking at the number representing that variable on the page icons, users 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 bars where the distortions (both perceived and real) due to relative distances would complicate such comparisons.

Finally, upon page selection, the actual figures from which the numbers appearing on the page icons were derived are depicted in a table in a manner similar to that used in the initial metaphor.

4.4.3 Scalability

It was established earlier that it becomes unfeasible to display all the available information regarding a web site's structure and usage on a single display when that site's size increases beyond a certain size. Previous efforts have addressed this problem using one of two approaches, namely by *aggregation* or *omission*. Aggregation involves combining several items of information into an approximated whole. These approximations have the effect of reducing the number of objects being displayed, thereby simplifying the representation and reducing clutter. Omission, on the other hand, reduces the visual complexity of a display by only depicting those items which are currently deemed to be of interest, and omitting the remaining details. The developed metaphor makes use of both approaches.

With regards to the omission of data, utilising a tree metaphor that implies the presence of links means that the number of links actually rendered is greatly reduced. By using a vertical column for the navigation bar pages, one level of the hierarchy is effectively removed, thereby saving a large amount of screen estate. In addition, by splitting the site into various fans, the number of pages shown in each fan is diminished. Finally, the varying level of detail displayed means that at different viewing ranges, certain items are not depicted. For example, at a medium distance the numbers representing variable ranges are no longer shown on the page icons and moving the viewpoint further back results in the page icons themselves being omitted.

In terms of aggregation, the practice of replacing the lines making up the links for a single branch by polygonal wedges reduces the complexity of the display, while still providing a general impression of the structure of the site.

By varying the level of detail shown according to the current viewing distance another important aspect of scalability is addressed, namely the amount of information shown to the user at any time is controlled so as to prevent information overload.

4.4.4 Context Maintenance

The visualisation consists of a single view and so there is no overview display like the windows directory tree used in the initial metaphor. Instead, the context and focus issue is addressed through the use of a zoom system that enlarges portions of the site while rendering the remainder in a smaller scale. In this manner the user is able to view and concentrate on a section of the site while maintaining context. The zoom system enlarges either entire fans or else single branches making up a fan (Figure 19) .

4.4.5 Data Exploration

Data exploration of the visualisation is achieved through the use of a trackball, which is a widely used mechanism for manipulating view changes of rendered objects. The user rotates the visualisation until the desirable section of a site is identified, then zooms in to that section to discover more detail. Upon wishing to view another part of the site, the user then zooms out and rotates the view, repeating the sequence. Selection of a page icon using the mouse results in the table containing information about that page being displayed.

4.5 Final Resulting Metaphor

The final metaphor makes use of a modified tree representation. A cube representing the home page of the site is attached to the top of a vertical cylinder from which horizontal fans are attached at regular intervals. The lines connecting the fans to the cylinder represent links to pages comprising a global navigation bar that consists of pages that are accessible from the majority of pages on the site. Each fan then represents a sub-tree of the site that has one of the global navigational pages as a root. Navigational bars for sub-sites are represented by their own cylinders. Upon moving the viewpoint closer to the visualisation, the branch fans/polygons are resolved into lines representing links to the individual pages making up a branch. If the line is solid then it is implied that there is a link between the associated page and the closest navigational bar, otherwise if the line is stippled then no such link exists. Relevant non-tree links that are not implied are rendered on demand when a page is selected. Page usage is encoded in the intensity of the line representing the link to that page, with line width corresponding to the usage of the associated page and all the pages comprising the sub-tree of which the original page is the root. When the viewpoint is sufficiently close, pages are represented by various icons depending on the page's type. At the closest level of detail, numbers appear on the page icons representing range values of certain usage variables.

4.6 Initial Experiences and Outcomes of Final Metaphor

A prototype implementation of the final metaphor was carried out using the same development software and running on the same platform as the prototype for the initial metaphor. Once again, simulated data was used and a random site was generated each time the system was run. Feedback was once again obtained from demonstrations to various experts.

The outcome of the initial heuristic evaluation was that the final metaphor represented an improvement over the initial metaphor, especially in terms of structural display, as emphasis is placed on points of importance with regards to web sites such as the home page and navigation bars. These points also serve as reference points aiding in the users' recognition of the various sections of the site that they may be familiar with from visiting the site. The second metaphor also had improved data pattern recognition potential than the initial metaphor due to overall impressions on usage being more readily decipherable from greater ranges and information on areas rather than individual pages being available.

There were some weaknesses detected, however. For instance, there is the impression that all the traffic to a page traveled along the hierarchical tree link to that page, since it is that link that is used to represent the usage for the page. These weaknesses were not deemed to be sufficiently serious to abandon the design and so it was decided to embark on a further, more formal, evaluation.

4.7 Summary

This chapter presented the development of the metaphor that will be used to represent a web site and it's visitor traffic.

Two metaphors were designed, the first of which made use of a windows directory type browser for viewing an overview of the site. Pages of interest could then be identified and viewed in a secondary view, which displayed branches of the site as either a flat hierarchical tree or else as a three dimensional cone tree. Various usage statistics were depicted in this view as bars placed on the page icons.

Upon reviewing the initial metaphor it was decided that it contained flaws that were too serious for its continued adoption. These flaws included the inability to adequately display links that did not form part of the hierarchy as well as the poor scaling of the visualisation to sites of large size.

A second metaphor was then developed that makes use of a modified tree representation. Horizontal fans representing branches of a site are arranged in semi-circles around a vertical column that represents a global navigation bar. The visualisation makes use of varying levels of detail, with more details being displayed the closer the viewpoint is. Making use of the concept of navigation bars allows a large number of links to be implied and therefore not being required to be explicitly shown.

In addition, information concerning areas of a site as opposed to individual pages is available.

Reviewing the second metaphor resulted in the perception that it contained sufficient potential so continued evaluation could take place. The execution of the next part of this evaluation is discussed in the following chapter.

Chapter 5

Intuitiveness Test

It is a desirable trait of a visualisation that a user is readily able to understand it without requiring specialised training or previous experience with it. In other words, the visualisation must make use of a metaphor that is *intuitive*, for if it is not, then it is doubtful whether the visualisation will be adopted into common practice. This is because users may not be willing to expend the time or effort to learn to utilise the visualisation, regardless of its merits. As such, a test was carried out to determine how intuitive the developed metaphor was. This chapter describes that test.

Section 5.1 outlines the aims of the intuitiveness test that was carried out. Section 5.2 then describes the process used to carry out the test. The results are presented and discussed in Section 5.3, which is followed by Section 5.4, which discusses the consequences of intuitiveness test. Finally, Section 5.5 provides a summary of the chapter.

5.1 Test Aims

The main aim of this test was to determine how intuitive users found the metaphor to be, or to be more precise, how intuitive they found the different aspects of the metaphor to be. This was achieved by investigating the success with which various features of the metaphor could be correctly identified without any prior information concerning the visualisation being provided, other than that it was a visualisation of a web site and its usage.

To this end a series of questions were asked to a number of test subjects. These included:

1. What do you think the lines and cubes being displayed represent? (Answer: links and pages)
2. What does the vertical column and the cube on top of it represent? (Answer: the global navigation bar and the home page)

3. What is the significance of the solid and stippled lines? (Answer: solid lines indicate a link to the global navigation bar and stippled lines do not)
4. What is the meaning of the different intensities of the lines? (Answer: line intensity indicates usage of the associated page)
5. What do the different shapes at the end of the lines represent? (Answer: different page types)
6. What do the numbers on the shapes indicate? (Answer: values of different usage variables)
7. How would you obtain more information about a particular page? (Answer: select the page using the mouse)
8. What do the other vertical cylinders indicate? (Answer: the presence of subsite navigation bars)
9. What does the polygon intensity show? (Answer: average usage of the pages making up that branch)
10. What does the line width indicate? (Answer: Usage of a page plus all the pages following it)

The manner in which these question were asked is described in the next section.

5.2 Test Process

The test was carried out on an SGI O2 using the prototype system mentioned in the previous chapter, which utilises simulated usage data and randomly generates a web site after each execution.

5.2.1 User Demographics

A total of six users were used for the intuitiveness test. The subjects, which consisted of four males and two females, possessed varying levels of expertise in web site usage and log file statistics. All the subjects were, however, familiar with web sites in general since they had computer science backgrounds. It was felt that the exercise would prove to be more useful if subjects were found that mimicked the target end users as closely as possible, rather than selecting test subjects at random. However, even though such a visualisation would ultimately be used by site designers who are accustomed to dealing with log file data and usage information that can be inferred from it, two of the participants had no such previous experience.

5.2.2 Test Procedure

The test was carried out as follows. As already mentioned, the test subjects were given no prior information about the visualisation besides that they were viewing a visualisation of web site usage. The procedure itself was informal in order to place the subjects at ease and involved no official questionnaire although the questions contained in Section 5.1 were always asked. Interaction with the system was controlled and driven by the experimenter, who would point out various features of the visualisation and ask the subject to try identify what those features represented (the goal was to test perception, not interaction). After a response was given to each question, the user would be provided with the correct answer if their interpretation was inaccurate. This was to prevent a misconception being carried forward to the remaining questions and potentially having a negative influence on future answers.

5.3 Test Results and Discussion

The results of the intuitiveness test are shown in Table 1, which contains the response of the six subjects to the questions listed in Section 5.1.

		Subjects					
Features		1	2	3	4	5	6
1	Lines and cubes	Yes	Yes	Yes	Yes	Yes	Yes
2	Main cylinder with cube on top	Yes	Yes	No	No	Yes	Yes
3	Solid and stippled lines	No	Yes	No	Yes	No	No
4	Line intensity	Yes	Yes	Yes	Yes	Yes	Yes
5	Page icon shapes	Yes	Yes	No	Yes	Yes	Yes
6	Numbers on page icons	No	No	Yes	Yes	Yes	Yes
7	Obtaining more page details	Yes	Yes	Yes	Yes	Yes	Yes
8	Other vertical cylinders	Yes	Yes	Yes	Yes	Yes	Yes
9	Polygon intensity	Yes	Yes	No	Yes	Yes	Yes
10	Line width	Yes	Yes	Yes	No	Yes	No

Table 1: **Intuitiveness Test Results.** This table contains the results of the intuitiveness test in which subjects were asked to identify what certain features of the site represented. Subjects had to base their answers solely on their intuition, as they had not been provided with any previous information other than that they were viewing a web site usage visualisation. Correct answers are indicated by "Yes" while incorrect answers are indicated by "No".

Overall, the success of users in correctly understanding the majority of the visualisation, despite

being given no information about the metaphor, was positive. The fact that the intensity of the links correspond to the amount of accesses the associated pages received was readily apparent to all the test subjects. In addition, the users were able to immediately detect the presence of sub-sites.

There were a few features which users were unable to correctly identify on their own and needed an explanation. During the testing it was found that although four of the six subjects could intuitively understand that the vertical column represented the global navigation bar, none inferred that links were being implied. However, they did claim that it "made sense" once it was explained that all the pages are understood to be connected to the navigation bar pages. Another point of confusion were the dashed lines used to indicate a page that was not connected to the navigation bar. Suggestions as to the significance of a dashed line as opposed to a solid one included thoughts such as an external link or a temporary link. A decision was made during implementation to draw attention to those pages that were never accessed as they represent potential problems. This was achieved by shading those links red as opposed to the usual yellow. However, some users incorrectly took the red lines to mean that these are the pages with broken links. Finally, approximating a branch with a polygon and using the intensity of the polygon to represent the average values for that branch was not always understood as a few users interpreted the intensity of the polygon to mean the sum of the values for that branch.

5.4 Outcome

The intuitiveness test revealed some weaknesses of the metaphor with regards to ease of understanding. In particular, the fact that solid lines implied that the associated page had links to the global navigation bar was not well identified. In order to determine if this will prove to be a major hindrance in the use of the metaphor further user experiments need to be performed.

5.5 Summary

This chapter described the intuitiveness test carried out to determine whether the metaphor devised in the previous chapter is readily understandable and easy to learn.

The test was undertaken by six participants who had no prior knowledge of the metaphor. The subjects were asked to identify what features of the visualisation represented based on the fact that they were viewing a web site usage visualisation.

The results of the test were promising with the implication of links being the feature that was the least well identified. However, this met expectations as users are unlikely to have encountered such a mechanism before. Although users did claim to understand the concept once it was explained, it is uncertain whether the implication of links will be understood well enough for users to be able

to utilise the visualisation effectively. For that to be answered, a full system implementation would have to be carried out so that full user experiments could be performed. The description of some aspects of this implementation are provided in the next chapter.

Chapter 6

Metaphor Revision and Final System Implementation

The intuitiveness test described in the previous chapter outlined the need to revise certain aspects of the developed metaphor. This has to be accomplished before implementation of the final system can occur.

To date, the devised metaphor has been evaluated using a prototype system that utilises simulated data. The final system, however, will need to make use of real data and will visualise a real web site. This presented complications that were not addressed by the development of the prototype. In particular, a method is required for identifying site features that will feature in the visualisation.

This chapter describes both the metaphor revisions as well as aspects of the system implementation.

Section 6.1 describes the revisions made to the metaphor. Section 6.2 and the sections following it then describe the various components comprising the visualisation system. Section 6.3 describes the web crawler component, while Section 6.4 discusses the web crawler output parser. Next, Section 6.5 talks about the log file parser. This is followed by Section 6.6 which describes certain aspects of the renderer component. Finally, Section 6.7 provides a summary of the chapter.

6.1 Revisions to the Metaphor

After processing the feedback resulting from the intuitiveness test, several changes were made to the metaphor. These consisted of both modifications to the existing representation as well as the addition of features that were deemed to strengthen the metaphor. These are described in turn below.

6.1.1 Metaphor Modifications

After the intuitiveness test, several alterations were made to the metaphor as a result of common misinterpretations. Broken links are now depicted as red as opposed to links to pages that received no traffic. Instead, pages that received zero hits are now shown as grey, the reasoning being that since the navigation bars are grey, grey is the underlying colour of the structure. Hence, if there is no information about a certain section of the site (i.e., no hits for that section) that section defaults to the neutral underlying grey.

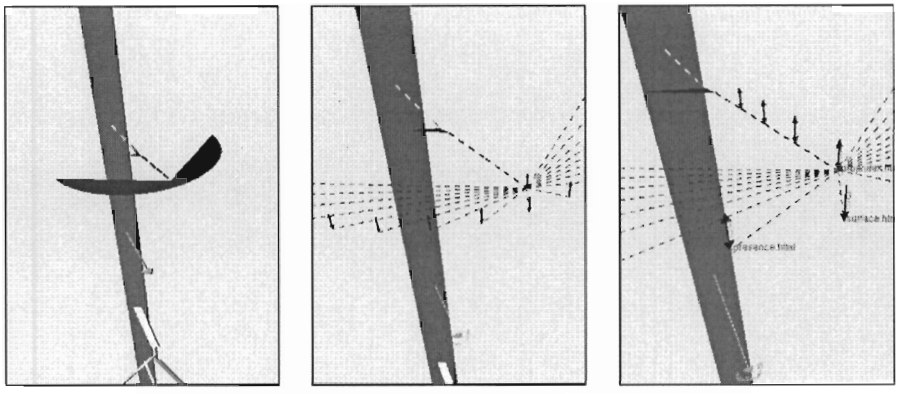
The page icons were found to be unintuitive as arbitrary shapes were chosen to represent the different page types. These have been altered so that the pages icons now consist of:

- Downward pointing arrows, representing entry pages.
- Upward pointing arrows, representing exit pages.
- Double ended arrows, representing entry/exit pages.
- No icons, representing normal pages.

6.1.2 Metaphor Additions

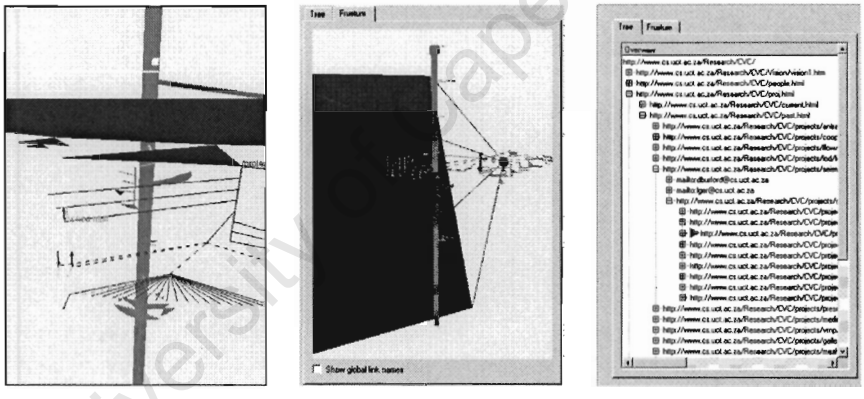
Indicating the usage of a page on the link leading to that page does have a drawback, namely that it is then implied that all the traffic arrived along that link, which may not be the case. To counter this, when an individual page is selected, all the links leading to and that page are shown and the usage indication (shown by link intensity) is distributed among these links. Thus, the user is able to see exactly where the traffic entering this page came from. Links leading from the page are likewise rendered. The colour chosen for links to the page was purple while those leading from the page are shown in blue. This is to coincide with an often used practice of web sites to display the text of links that have not been visited before in blue while those that have already been seen are shown in purple. This new feature is shown in Figure 23.

Displaying fewer explicit links allows us to add another feature. The site structure is rendered as normal (i.e., as a modified tree) but all additional links, such as those that travel up the hierarchy, are also displayed. While the tree is emphasised, these extra links are muted, yet discernible. Thus, the main tree acts as the "focus" of the visualisation, while the remaining (cyclical) links form a "peripheral visualisation" (Figure 22). This approach may not provide any extra usage information, but will give the user a better idea of the connectivity of the site as a whole. If a web site contains mostly links that follow the hierarchy or are due to navigation bars, then all remaining links are displayed. When this is not the case, only those extra links belonging to the currently highlighted page and to all the pages in the containing branch are shown, in order to reduce clutter. Since this



(a) (b) (c)

Figure 20: **Final System Level of Detail.** This figure shows the varying levels of detail of the final system. Note the changes in the pages icons from the prototype system shown in Chapter 4. These screenshots were taken from the Collaborative Visual Computing Laboratory web site (http://www.cs.uct.ac.za/Research/CVC).

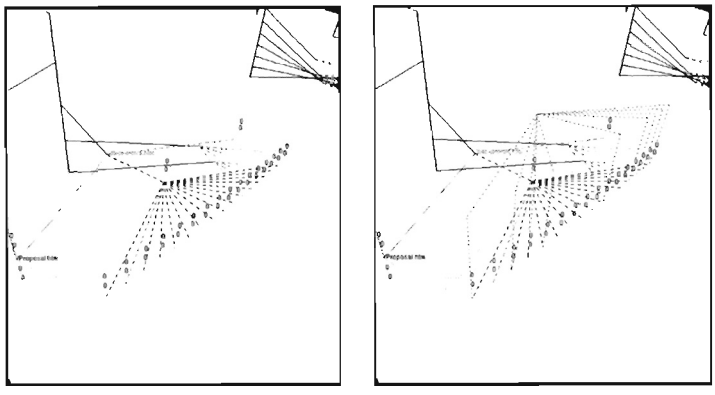


(a) (b) (c)

Figure 21: **Overview.** This figure shows the new overview view which was included to aid in context maintenance. The overview can be toggled between two different views. (a) The normal view. (b) The frustum view showing the entire site plus the current location of the view frustum in (a). (c) The directory tree overview option.

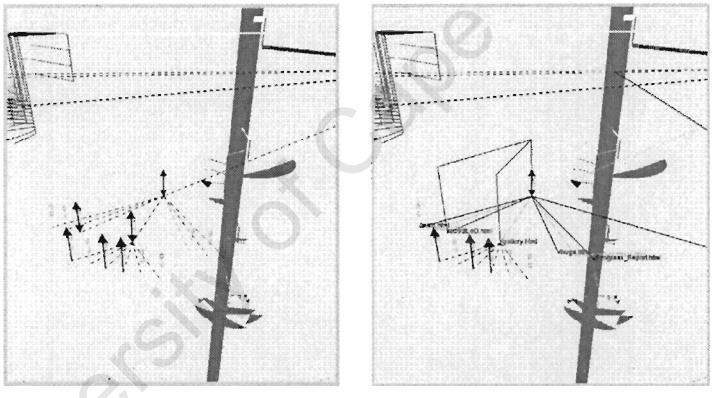
feature could potentially lead to clutter depending on the site structure, users will have the option of turning it off.

The last addition was the inclusion of a second view to provide an overview of the site. It was felt that the context maintenance approach described in Section 4.4.4 of Chapter 4 was inadequate, as even with selected portions of the site enlarged, the user still had to frequently zoom in closer



(a) (b)

Figure 22: **Ghost links.** A new option was added so that non-tree links that could not be implied could be toggled on. (a) The normal view. (b) The same view with non-tree links toggled on.



(a) (b)

Figure 23: **Incoming and Outgoing Traffic.** Upon selection of an individual page, all the links to and from that page are displayed. Incoming links are shown in purple while outgoing links are shown in blue. The intensity of the links varies to indicate the amount of traffic along those links. (a) The normal view. (b) The same view showing incoming and outgoing traffic.

in order to obtain more details. This resulted in much of the surrounding area to be outside the view causing a loss of context. The new view provides a site overview using one of two options, namely a window directory style tree and a view of the entire site with the view frustum of the first view imposed on it (Figure 21). Users are able to toggle between the two options. Although the use of a windows directory tree was abandoned after the development of the initial metaphor,

a main contribution to this was the difficulty in using it to navigate large sites. In the new system, however, the directory tree serves primarily to provide an overview as navigation will take place in the primary view window.

6.2 System Implementation

The final system consists of several components, namely a web crawler, a web crawler output parser, a log file parser and a renderer. Each of these is described in the following sections.

6.3 Web Crawler

The structure of the site being visualised is acquired using a web crawler. The web crawler, which was implemented in Java, runs as a preprocess. It performs a breadth-first traversal of a web site, noting the links present on each page and writing the results to file. External links are not traversed. Once the crawler has completed executing the results are used as input for the web crawler output parser described in the next section.

6.4 Web Crawler Output Parser

The crawler output parser takes the output from the web crawler and parses it, creating a data structure that will represent the site structure. In addition, the parser, which was implemented using C++, identifies certain web site features such as navigation bars based on a set of heuristics.

6.4.1 Automatic Sub-site Identification

Identifying sub or self-contained web sites can prove to be problematic, as there does not always exist a clear distinction between a sub-site and a "gateway" page (a page containing an index of links). A simplistic approach would be to make the distinction based on the page's URL and on commonly used naming conventions. A URL pointing to an html file called "index.html" or "home.html" would obviously identify the page as the start of a sub-site. If the web site is that of an academic institution, the presence of a tilda followed by a name is likely an indication of a personal web site. URL's that designate directory changes are also potential sign posts of sub-sites, although not particularly reliable ones. Finally, an analysis of the page content itself could lead to clues. The presence of links, navigation bars and certain keywords (such as "Welcome") are all markers of potential sub-sites.

6.4.2 Automatic Navigation Bar Identification

Automatically detecting navigation bars is a complex problem and any solution will be of necessity, heuristically based. The algorithm that was devised only works for navigation bars built using "ordinary" links and fails when encountering navigation bars that are derived from JavaScript or PHP. Again, as the main concern of this research is the manner of depicting pages and navigation bars and not the identification of these bars, the algorithm was regarded as adequate for our needs.

General Case

The algorithm for detecting navigation bars works as follows:

1. Count the number of links to each page in the site.
2. Examine one branch at a time, looking for the most common number of links to pages belonging to that branch.
3. Links which occur as many times as (or more than) the most common number are deemed to be part of a navigation bar. This is based on the fact the pages of a navigation bar will be present on many other pages and that they'll all be present on the same page. Thus, they should all occur frequently and a similar number of times as each other.

While the above successfully identifies large navigation bars, there are problems when the bars become shorter, such as those consisting of 4 pages or less.

In these cases further checks are required. Thus, for each branch:

1. Check that a page(s) exists that occurs more than once in the site. Ignore branch if not.
2. Check that a page(s) exists that has links to its siblings. Ignore branch if not.
3. Make a list for each page of the links it has to its siblings. These lists must be unique, i.e., if two pages have the same list ignore the second list.
4. Count the number of times each list occurs. The list that occurs the most is likely the navigation bar.

A complexity arises when a page has an arbitrary link to a sibling that is not part of a navigation bar. This causes the list counting to be inaccurate. Therefore, in this situation a final step is required: work out the intersection of each list with every other list and the intersection that occurs the most is then the navigation bar.

Frames

From the very nature of frames, we may deduce that a page containing a frame is part of a navigation bar. The difficulty occurs in identifying which frame is the navigation bar and which are merely content holders. It was decided to take the first two frames which contain links and declare one of them as containing the navigation bar links depending on how many links it contains and how much screen estate it is allocated.

Test Cases

Testing whether all the navigation bars (of which there could be thousands) of a web site were successfully detected and no erroneous identifications were made, would be impractical. Instead, the identification algorithm described above was tested on a small test site, which contained various navigation bars. In addition, it was tested using two real web sites. The first, the Collaborative Visual Computing Laboratory (CVC) web site (*http://www.cs.uct.ac.za/Research/CVC*), is a small site consisting of only about 1000 pages. The second site, the Computer Science Department web site (*http://www.cs.uct.ac.za*), is larger and contains over 16000 pages. After running the algorithm on the two sites, the results were searched for known navigation bars to establish whether they had been correctly identified. Furthermore, a series of "spot checks" were performed, whereby random pages were selected, checked manually for navigation bars and then compared with the algorithm output for that page. During final testing, no errors or contradictions were found for either site.

6.5 Log File Parser

A log file parser was implemented in C++ to extract usage information from server log files, such as the total number of *hits* (or visits), each page received. At present, only two other items of data are extracted, these being the number of hits from clients using Internet Explorer browsers and the number of hits from clients using Netscape or Netscape compatible browsers. These two items were arbitrarily chosen as an example of displaying other usage variables besides total hits. Extending the log file parser to retrieve other information such as, say, number of error hits, would be trivial.

In addition to information about the usage of each page, the log file parser also determines information about which category the page falls under. Possible categories include *entry* pages (the first page a user visits when entering a site), *exit* pages (the last page viewed before the user leaves the site). *entry/exit* pages (the first and last page seen) and *normal* pages (pages which are viewed between entry and exit pages). In order to ascertain which category a page belongs to, all the *user sessions* of the site need to be identified. A user session is defined by tracking a single IP address

and building a list of the pages that that IP requested (visited). If the current page being requested is not reachable from the previous page requested, or if that IP has been idle for a certain amount of time, then the previous session is assumed to have ended and a new one is started. This is a standard approach to estimating user sessions used by many log analysis tools (most of which utilise a time-out period of 30 minutes) [39]. Entry pages are then all the pages that begin a session and exit pages are those that end one.

While this method of tracking user sessions is prone to inaccuracies and is only an estimate, it was deemed sufficient for our purposes, as we place more emphasis on displaying data rather than obtaining it.

The results of the log file parser are incorporated into the data structure storing the site structure that was populated by the web crawler output parser. This data structure will be used by the renderer

6.6 Renderer

The renderer makes use of C++, OpenGL and Qt to render the visualisation and provide the user interface. The renderer walks the data structure containing the site structure and usage information and then displays the visualisation. The layout algorithm utilised to position the branches and links is a modified version of that used by [7] for laying out cone trees. This algorithm was adapted to creating a radial layout that is limited to a semicircle. The distance between the fans is based on a metric that is half the width of the broadest fan.

Real data for two sites, the CVC Laboratory site and the Departmental site, were collected. The system interface is shown in Figure 24, while visualisations of the two web sites can be seen in Figure 25.

Once system implementation was complete, the user experiments could be carried out. The design and execution of these will be described in the next chapter.

6.7 Summary

This chapter listed modifications made to the metaphor as well as describing various implementation issues.

Following discoveries made during the intuitiveness test of the previous chapter, existing features of the metaphor were altered. These include using red to indicate broken links and grey to indicate links to pages that received no hits and the adoption of new, more informative page icons. Aside from these modifications, certain features were added. Upon selecting a page traffic to and from that page is indicated by blue and purple lines representing incoming and outgoing links. An optional feature displays non-tree links in a dark grey colour so that they can be perceived without

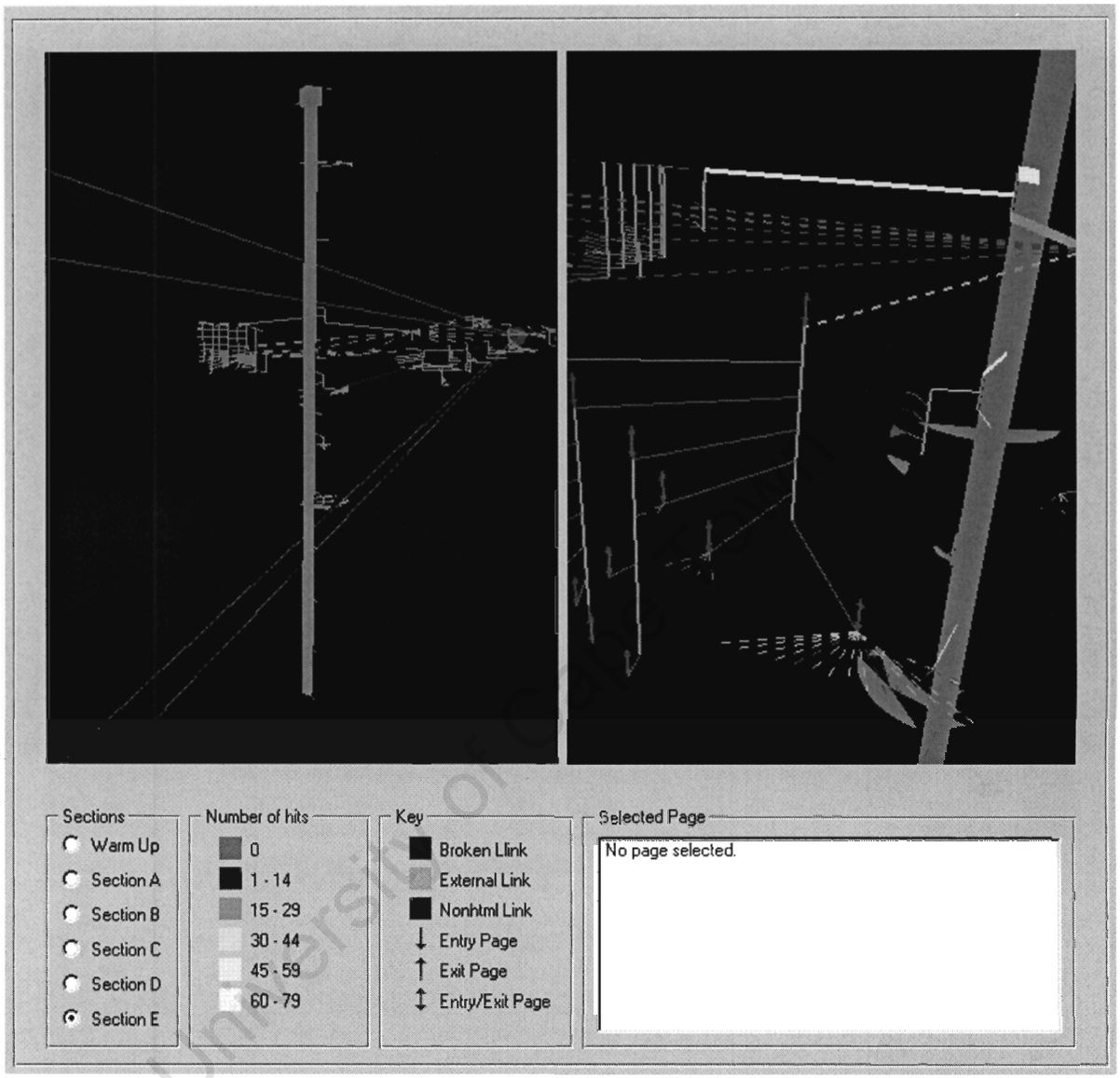
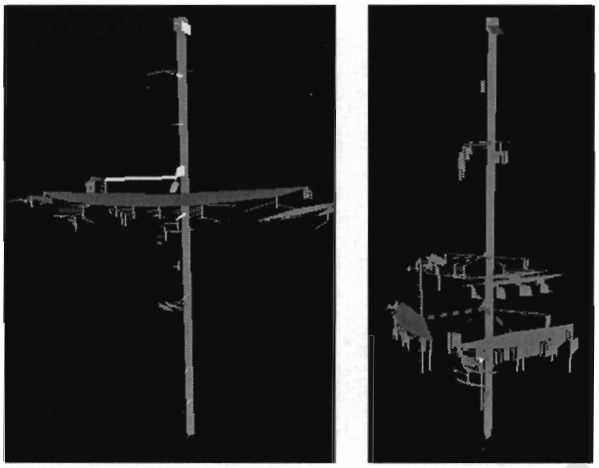


Figure 24: **Web Site Usage Visualisation Interface.** The interface contains two rendering windows: the left window displays the entire site and can be toggled between a directory tree view or a view frustum view (the frustum view is shown in this figure) and is used to provide an overview of the site; the right window – in which a user can zoom, pan and rotate – displays more details. The text box below the right window provides information about the currently selected page.

becoming too conspicuous. Finally, a second overview window was added which display the overall site structure as either a directory tree or as a second view of the metaphor with the viewing frustum



(a) (b)

Figure 25: **Site Visualisations.** (a) Zoomed out view of the Collaborative Visual Computing Laboratory's web site (*http://www.cs.uct.ac.za/Research/CVC*), which contains about 1000 pages. (b) Zoomed out view of the University of Cape Town's Department of Computer Science web site (*http://www.cs.uct.ac.za*), which consists of over 16000 pages.

of the first window superimposed.

Following these modifications, system implementation was carried out. The system consists of several components: a web crawler that obtains the site structure, a web crawler output parser that identifies site features, a log file parser that extracts usage information from log files and a renderer which displays the visualisation.

Once system implementation was complete, user experiments could be performed. These will be described in the next chapter.

Chapter 7

User Experiments

A visualisation can only be deemed a success if users are able to accurately comprehend and utilise it. In order to evaluate the devised metaphor, a task-orientated user experiment was carried out, wherein test subjects utilised the visualisation to answer a questionnaire.

This chapter describes the design and execution of this experiment.

Section 7.1 explains the aims of the user experiment. Section 7.2 then describes the design of the questionnaire used.The actual process of the experiment is discussed in Section 7.3, after which Section 7.4 presents and discusses the experiment results. This is followed by Section 7.5, which contains the conclusions drawn from the user experiment. Finally, Section 7.6 concludes the chapter with a summary.

7.1 Experiment Aims

The main aim of the user experiment can be broadly stated as to evaluate the effectiveness of the developed visualisation.This can be divided into the more specific aims of determining the extent to which the approach to each of the factors affecting web site usage visualisation (as described in Section 1.3 of Chapter 1) can be deemed successful. These are each discussed in further detail below.

7.1.1 Structure

The first aim is to examine the effectiveness with which the structure of a web site was represented. This intent can be met by investigating whether users could accurately extract and interpret information about the site's structure. Of special interest is the ability of the users to recognise the presence of links that were implied and not explicitly rendered.

7.1.2 Data Representation

This aim is concerned with evaluating the suitability of the chosen means of representing various aspects of web site usage data. In particular, it needs to be determined if users could understand and correctly use information encoded in the line colour, intensity and width, the meaning of the polygonal wedge intensity and the significance of the various page icons.

7.1.3 Scalability

Scalability refers to the ability of a visualisation to gracefully handle sites of increasing scale. In this case, the experiment should determine if users are still able to effectively utilise the visualisation once the site being visualised has increased in size.

7.1.4 Data Exploration (Navigation)

For this factor, the aim is to seek whether users can manipulate and navigate the visualisation in order to obtain the information that they require. This requires that the chosen mechanism of a trackball be examined for effectiveness and ease of use.

7.1.5 Context Maintenance

The final aim is to evaluate the extent to which the overview windows permit users to maintain the context of their present viewing position with respect to the overall site structure. In addition, the two overview options, namely the tree view and the frustum view, can be compared to each other to determine which is the better suited to fulfill the role of context maintenance.

Now that the aims of the user experiment have been stated, the design of the questionnaire, which asks the users to perform tasks that address these aims will be described in the next section.

7.2 Design of the Questionnaire

The questionnaire, the final version of which can be found in Appendix A, required the users to utilise the visualisation to answer several questions. It consisted of twenty six questions, which are split up into several sections with each section containing questions relating to a particular feature of the visualisation. The design intentions behind each section is detailed below.

7.2.1 Training

The first section of the questionnaire was a training section. This section had a two-fold purpose, namely to ensure that the users understood the tutorial (see Section 7.3.3) and to familiarise users

with utilising the visualisation system. As result, the questions in this section were designed to be straight-forward and tested understanding of the basic features of the visualisation such as the structural representation of pages and links and the encoding of usage information in the link intensity.

7.2.2 Structure

Questions on structure were contained in a single section (Section A). The main concern of these questions was to test understanding of the concept of the implication of links, since the explicit structure representation was covered to an extent by the training section.

Section A

The questions in this section consisted of two types. The first of these required users to name all the pages that are directly accessible from a particular page. Most, but not all of these questions involved pages that had access to a navigation bar, whether it was the global bar or else that of a sub-site. If users did not understand the concept of implied links fully, then they would not list the pages comprising the navigation bar in their answer. The second type of question asked users to trace the shortest path between two given pages. This path could include reversal of links via the idea of a "back" button on a browser. Again, in the majority of questions the path involved the use of a navigation bar. Users that did not realise the presence of the navigation bar links would list a much longer path.

7.2.3 Data Representation

Questions concerning the manner in which data was represented was divided into three separate sections, each of which is associated with a particular feature. These are:

- Section B – This section tests the understanding of the polygonal wedge approximation of branches.
- Section C – This section applies to the line width of link representations.
- Section D – This section involves the interpretation of incoming and outgoing traffic to a page as represented in the intensity of incoming and outgoing non-tree links.

Section B

Section B only contained two questions. These required the user to identify the branch/wedge that received the most and least amount of traffic respectively. Data for this question was set up in such a

way that the wedge with the highest usage was deliberately not the largest wedge and visa versa for the lowest usage branch. In order to ensure that the correct answer was not provided by a random guess, users were asked to identify the mechanism which they used to answer the questions.

Section C

This section required users to choose between two sections of the site as to which received more hits. Again the data was manipulated so that the area with the most traffic (i.e., had a root with the widest link/line) actually contained fewer branches and pages. Again, users were required to state how they came by their answer.

Section D

Section D asked a series of questions in which the user had to identify pages in the site from and to which the most traffic for a particular page traveled. In addition, the concept of an entry page was tested by asking why one page received more hits than another even though its parent received less hits.

7.2.4 Scalability, Navigation and Context Maintenance

Questions relating to the concepts of scalability, data exploration and context maintenance were all combined into a single section, namely Section E. This grouping was chosen as in order to evaluate any of these factors requires the user to move from viewpoint to viewpoint.

Section E

All the questions in Section E pertain to a visualisation of the Departmental web site, which can be considered to be reasonably large since it contains over 16 000 pages. Thus, being required to answer questions that involved viewing different areas of this site should provided some feedback as to the effectiveness of the visualisation to handle a site of this size.

Questions occurred in groups of three, with all three corresponding to the same page. The first question would ask the user for some detail of information that would require them to zoom closely to the given page in order to provide the answer. The second question would then require the user to use the overview window of their choice to answer a question involving the relation of the present page to the rest of the site. The third question then merely asks the user to state which of the overview windows they used. The next group of questions would then pertain to a page that was located in a position that would require the user to navigate to another distant area of the site. In this manner, the user was forced to navigate all over a large site and to use the overview windows.

7.2.5 Rating Section

The rating section contained a list of statements to which users were asked to provide a rating based on how strongly they agreed with the statement. These ratings ranged from 1 (strongly disagree) to 5 (strongly agree). The purpose of these statements was to obtain the opinion of the users as to how effectively the visualisation addressed the factors of structure, data representation, scalability, data exploration and context maintenance, although the statements were worded in such a way that the factors were not explicitly stated.

7.2.6 Comments Section

This section simply asked the users to provide any positive, negative or general comments they had which had not been addressed in the previous section.

7.3 Experiment Process

The experiments were performed with a single test subject at a time and were run on an AMD Athlon 650MHz with 256 megabytes of RAM and a GeForce 2 MX graphics accelerator card.

7.3.1 Subject Demographics

The users consisted of twenty six students in the Computer Science Department at the University of Cape Town, of which seven were females and the rest were males. The sample of subjects was deliberately drawn from a background that necessitated a familiarity with computers as well as a general experience of web sites, since the target end users of the visualisation system are expected to have such a profile. In terms of web experience, all but one of the test subjects had created a web site before. Only five of the users had served as a web site administrator before and these five were the only subjects to have made use of a web site usage analysis tool.

7.3.2 Pilot Experiment

A few pilot experiments involving two test subjects were run before the formal experiments took place. This was to ensure that the questionnaire was comprehensible and that the operation of visualisation system ran smoothly. While no serious complications were detected concerning the procedure of the experiment and the system itself, the questionnaire was reworded slightly following feedback from the pilot test subjects. The official experiments then commenced.

7.3.3 Experiment Procedure

Before each experiment started, the user was given a ten minute tutorial on how to interpret and utilise the visualisation system. Following this, the subject answered the training section of the questionnaire which contained the answers for this section at the end. Subjects were required to check their answers and seek clarification if they got any incorrect and failed to understand why. The subjects then proceeded to answer the rest of the questionnaire. For each section of the questionnaire, subjects accessed visualisations of pre-chosen data by selecting the radio button corresponding to that section (see "Sections" radio button panel in Figure 24). An experiment controller was available at all times so that users could ask questions if they were confused or encountered any problems.

No time limit was set for the tasks ad users were asked to take as much time as they required. Once all the users had completed the experiment the results were collected. These are presented and discussed in the following section.

7.4 Results and Discussion

The results of the user experiment will now be summarised and discussed.

7.4.1 Overall Results

The overall results for the entire questionnaire are shown in Figure 26 and Table 2, which contain the individual results and a summary respectively.

	Total	Highest	Lowest	Mean	Median
Score	26	26 (100%)	11 (42.31%)	20.87 (80.27%)	23 (88.46%)

Table 2: **Overall Experiment Scores Summary.** This table contains a summary of the scores obtained for the questionnaire as a whole.

In general, the overall results were promising with an average questionnaire score of 80% and a median of 88.46%. Of particular significance is that all the users who had experience as a web site administrator scored well (scores of 24, 24, 25, 23 and 25 out of 26). This is a very positive result since the target end users of the system will most likely be web site administrators. It is interesting to note, however, that the individual with the highest score also had the least web experience. The subject in question had no experience as a site administrator, had not used site usage analysis tools before and had never even created a web site. In spite of this he managed to obtain a perfect score.

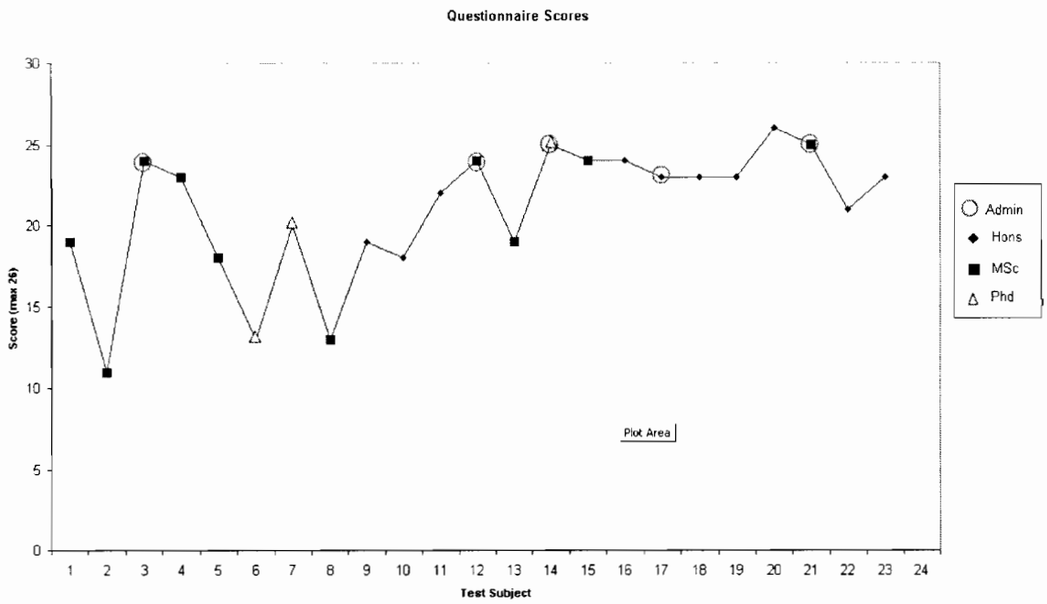


Figure 26: **Overall Experiment Scores.** The individual scores obtained for the entire questionnaire are shown in the graph above.

7.4.2 Structure

The individual results for Section A are shown in Figure 27 while Table 3 contains the results summary.

	Total	Highest	Lowest	Mean	Median
Score	9	9 (100%)	1 (11.11%)	6.52 (72.44%)	7 (77.78%)

Table 3: **Structure Section Scores Summary.** This table contains a summary of the results obtained in Section A.

Section A forms arguably the most critical segment of the user experiments since it tests the understanding of the key concept of implying links through the use of navigation bars. This concept is the hardest aspect of the metaphor to grasp as was exposed by the intuitiveness test described in Chapter 5. Bearing this in mind, the scores obtained in this section, with an average score of 73% and a median of 77.78%, were fairly solid.

The questionnaire was set up in such a way that only three questions in this section did not

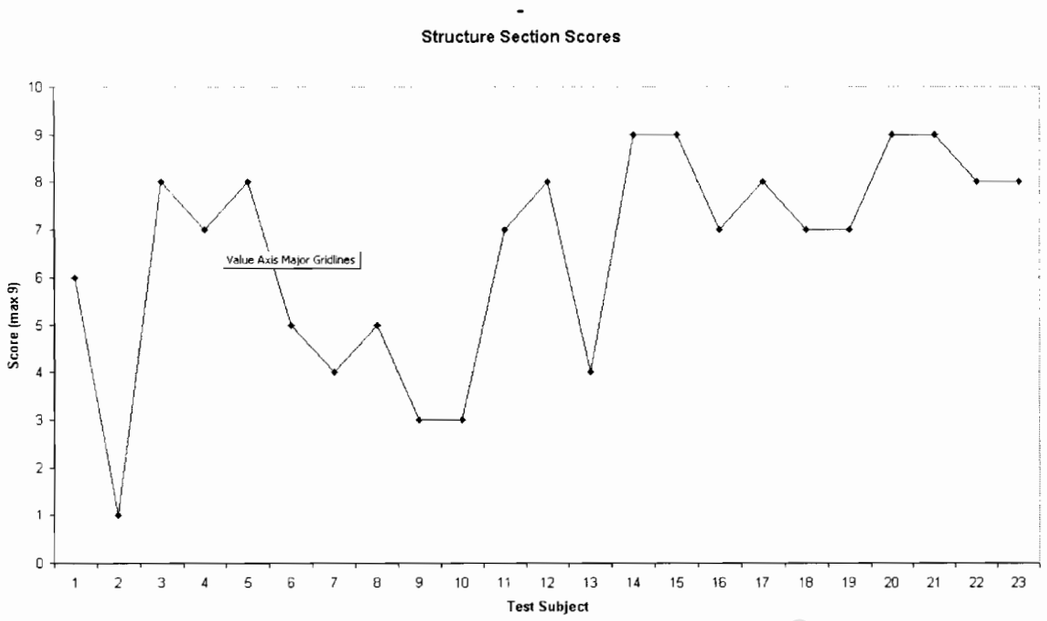


Figure 27: **Structure Section Scores.** This figure shows a graph of the individual scores obtained in the structure section (Section A) of the questionnaire.

involve the implication of links. Thus, if a user did not comprehend the concept of link implication, the highest score they could have obtained was 3 out of 9 (it is highly unlikely that a subject could have provided the correct answer by luck as each answer included multiple pages). Examining the results shows that only three of the twenty three subjects scored 3 or lower. The questions these subjects did get correct were those that did not involve the navigation bars. Thus, the results indicate that the majority of the subjects were able to understand the idea of link implication.

There were two questions which seemed to pose problems for many of the subjects. Both of these involved the use of a navigation bar that was not the global navigation bar. Investigating the pages that users provided in their answers revealed a common misconception that pages linked to this navigation bar were still linked to the global navigation bar. However, the intent of the metaphor design was that it was understood that pages were linked to the closest (in terms of hierarchy) navigation bar.

7.4.3 Data Representation

The combined results of the sections evaluating data representation are contained in Figure 28, which shows the individual results, and Table 4, which summarises them.

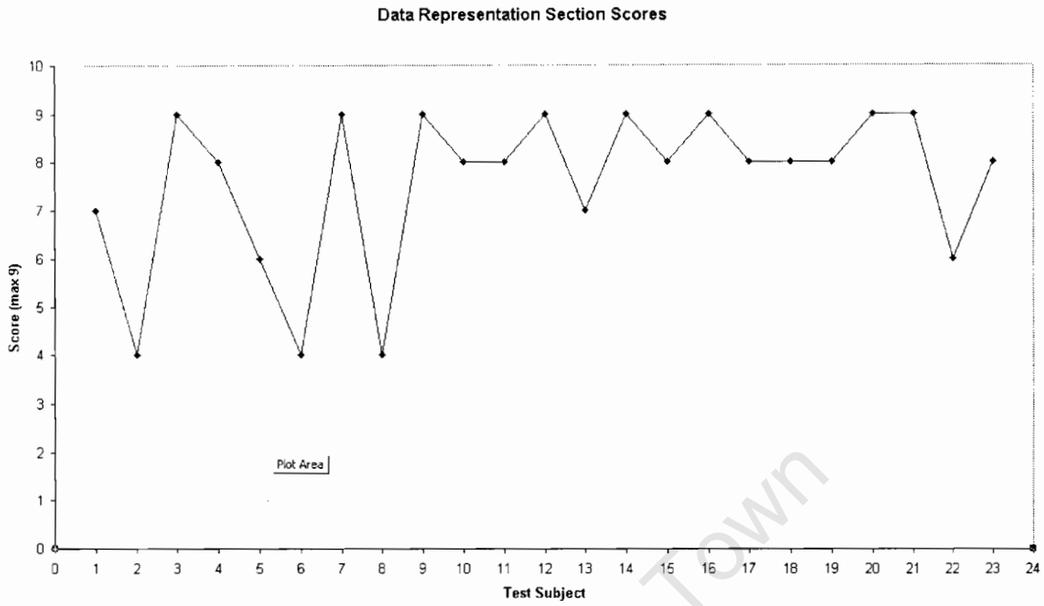


Figure 28: **Data Representation Sections Scores.** The results obtained by the individual test subjects for the data representation sections (Section B, C and D) are presented in the above graph.

	Total	Highest	Lowest	Mean	Median
Score	9	9 (100%)	4 (44.44%)	7.57 (84.11%)	8 (88.89%)

Table 4: **Data Representation Sections Scores Summary.** This table contains a summary of the combined results obtained in Sections B, C and D.

In general, users scored much higher in these sections than in Section A, as is evidenced by the average score of 84% and median of 88.89%. This would seem to validate the choices made during the development of the metaphor with regards to data representation. The scores for each section will be examined in further detail below.

Section B (Polygon Wedge Intensities)

Users performed extremely well in this section. Every one of the twenty six subjects correctly identified both the wedge representing the web site branch with the most usage as well as the wedge indicating the branch that received the least amount of traffic. This would seem to indicate that the approximation of branches and their usage using polygonal wedges is a mechanism that can be readily understood and utilised.

Section C (Line Width)

There was only a single question in this section which seventeen of the twenty three subjects answered correctly. While this represents the majority of the users, the results suggest that the use of line width to extract usage information concerning areas of a site is not as easy to grasp as the polygon wedge approximation.

Section D (Incoming/Outgoing Traffic)

The average score for this section was 77% with a median of 85.71%. The choice of blue and purple as the colours to represent outgoing and incoming links respectively was based on the generally used practice of displaying those text links that have been visited before in purple and those that have not in blue. However, comments by the users (see Section 7.4.6) concerning this section revealed that at high intensities it becomes difficult to distinguish between these two colours. This would seem to account for some of the incorrect answers. Examining the answers more closely, it also becomes apparent that looking at incoming and outgoing links, several users forgot that the link to a page from its parent and the link from it to a child page are also incoming and outgoing links respectively. This is evident by the fact that in questions to which the answer was the link from a parent or to a child, they gave what would be the correct answer if that link was discounted. Thus closer examination of the results would seem to indicate that representing usage information on incoming and outgoing links was successful, with a poor choice of colour harming the effectiveness.

7.4.4 Scalability, Navigation and Context Overview

The factors of scalability, data exploration (navigation) and context overview were evaluated by questions contained in a single section. The results for this section are presented in Figure 29 and Table 5, which contain the individual results and a summary respectively.

Scores for this section were very good, with the average score being 84.75% and the median being 87.5%. However, comments made by users while discussing the experiment after they had

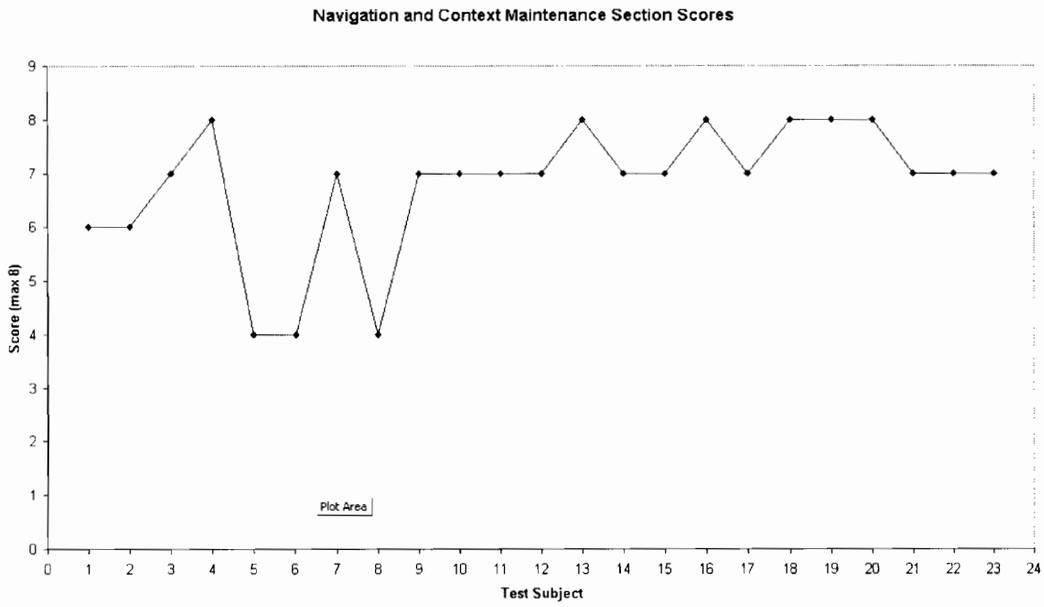


Figure 29: **Navigation, Scalability and Context Maintenance Section Scores.** This graph contains the individual scores obtained by the test subjects in Section E, which examined navigation, scalability and context maintenance.

	Total	Highest	Lowest	Mean	Median
Score	8	8 (100%)	4 (50%)	6.78 (84.75%)	7 (87.5%)

Table 5: **Navigation, Scalability and Context Maintenance Section Scores Summary.** This table contains a summary of the scores obtained in Section E.

completed it revealed that there were a few difficulties experienced, especially with regards to navigation.

Scalability

As none of the questions in this section were directly based on the size of the web site, evaluation of scalability relies heavily on users' comments (see Section 7.4.6). These comments indicated that cluttering was encountered while viewing the site at the highest level of detail due to the number of page icons being displayed. However, since cluttering only occurred at the highest level of detail, overcoming the problem should not prove to be too difficult. The fact that users were able to score

highly (as indicated by the high average score and median) while answering questions that required users to provide details about certain pages for a relatively large web site consisting of over 16000 pages, suggests that the metaphor scaled fairly effectively.

Navigation

Based on comments made by users, the trackball mechanism used provided many problems. Users found it difficult to navigate around the visualisation which affected their performance in answering questions in this section. Although the navigation controls were explained to the users during the tutorial, many had never encountered the concept of a trackball before. Thus, upon being required to move the viewpoint from a page at the top of the site structure to a page at the bottom, instead of rotating until the bottom page was in view and then zooming in, users attempted to "fly through" the structure by zooming until they reached the bottom page. However, due to its nature, a trackball does not permit zooming beyond its midpoint. As such when users reached the middle of the trackball they would "get stuck" when they tried to zoom further. The experiment controller had to frequently aid users who had become stuck and explain the mechanics behind utilising the trackball again.

Context Maintenance

The high scores obtained in this section again indicate that users were able to maintain context fairly well, as a number of the questions required users to place the page that they were currently viewing in context with the rest of the site. The ratings given to the statement concerning context maintenance in the ratings section (Section 7.4.5) indicated that only one person disagreed with that they had a good idea of where the page that they were currently viewing was located in terms of the rest of the site. Five subjects were neutral, but the remainder, and the vast majority, all agreed or strongly agreed.

An interesting exercise is to compare the usage of the two overview windows that were available. A summary of the results of the tree versus frustum overview are shown in Table 6.

	Easier to understand	Easier to use	No. of times used to answer a question	No. of correct answers given when used
Tree view	12	18	59	49 (83.05%)
Frustum view	7	3	31	27 (87.1%)

Table 6: **Tree versus Frustum Overviews.** This table contains ...

The tree was the more popular of the two views and was used to answer more questions. It was

also generally felt to be more usable and understandable. The frustum view did perform slightly better in terms of the number of correct answers provided while using it. It could be argued that these figures support the claim that users conceptually think of web sites as trees and thus find it easier to relate to a tree view. However, the results could also be a consequence of more exposure to directory tree type views. When asked about their preference, a number of users stated that they liked both views and would ideally prefer to use the two views in conjunction with each other.

7.4.5 Rating Section

The subjects responses to the statements in the ratings section are summarised in Table 7.

Question	Highest	Lowest	Mean
A.8.1 (Scalability)	5	2	3.42
A.8.2 (Context Maintenance)	5	2	4
A.8.3 (Exploration)	5	1	3.33
A.8.4 (Data Representation)	5	2	3.46
A.8.5 (Structure)	5	3	4.38
A.8.6 (Usage)	5	2	4

Table 7: **Ratings Section Summary.** This table contains answers given in the Ratings section of the questionnaire (Appendix A.8) in which users were asked to rate positive statements concerning various aspects of the visualisation according to how strongly the agreed (indicated by a rating of 5) or disagreed (rating of 1) with the statement.

The responses in this section were generally positive. The question regarding the ease with which information could be found (A.8.3) received the worst rating while the question about concerning how well the visualisation gave the user an idea of the site structure was rated the highest.

7.4.6 Comments

While the comments obtained in this section varied, several did recur.

Negative Aspects

Common comments made in this section include:

- The visualisation became cluttered when the user zoomed in to the highest level of detail, due to the page icons and the selected page indicator (a big pink arrow showing the currently selected page) obscured details.

- Navigation is difficult and requires time to learn.
- Selecting a page is difficult as the user is required to be quite precise in clicking on the page icon.
- It is difficult to distinguish between the colours of the incoming and outgoing links.

Most of these issues are due to implementation and do not point to serious flaws in the metaphor design. As such, they should be relatively easy to address.

Positive Aspects

Among the comments made in this section was the opinion that the site structure was easy to see and that the overview windows, particularly the tree view, were very useful.

7.5 Conclusions

On the whole, the results of the user experiment were positive enough to indicate that the metaphor contains potential. Particularly promising was the fact that those users with the most amount of web experience, i.e., those that had utilised previous site usage analysis tools and had served as web site administrators all scored highly.

While several issues were encountered, many of these were either implementation related or else could be solved by modifying the implementation. For example, the problem of cluttering being experienced at the highest level of detail could be addressed by adjusting the distance at which the highest level of detail occurs. In addition, perhaps only the page icon of the selected page should be shown with the others being rendered upon a mouse-over action.

One item that did become apparent was that the trackball mechanism used for navigation would either have to be replaced or else users would require extensive training in effectively utilising it.

7.6 Summary

This chapter presented the user experiments that were run to evaluate the metaphor.

The experiments required users to answer an questionnaire using the visualisation. The questionnaire was divided into sections, each of which tested understanding of a different aspect of the visualisation.

The experiment involved twenty six users from a computer science background with varying degrees of web experience. The test subjects were given a short tutorial on the visualisation before starting the experiment.

The results of the user experiment were generally promising although several minor issues would have to be addressed.

Chapter 8

Conclusion

Analysis of web site usage has become both a major business and a growing field of research. This is a natural outcome of the increasing prevalence of the internet and its ability to reach more people. Since the success of a web site is measured by the amount of visitor traffic it receives, there is a continual drive to evaluate and improve web site designs. Such an evaluation requires the collection and interpretation of web site usage information, which is usually carried out by the web site administrators and site designers. Obtaining some form of this usage information is a fairly easy process. Portraying the information in such a way that useful inferences can be made, however, is a more complicated problem. This dissertation presented a novel approach to creating such a portrayal.

8.1 Problem Description

To date, the most common method of obtaining web site usage information is through the use of log file analysis. Log files are text files created by web servers to store requests and their outcomes between servers and client browsers. These text files can then be parsed and usage information extracted from them such as the details about the request, as well as information about the client who sent the request.

As a result of their wide-spread use there are many tools which perform log file analysis. However, they present their results using tables and two dimensional graphs such as bar and pie charts. These results therefore contain no reference to the actual layout or link structure of the site being analysed. This lack could lead to the full potential of the data extracted from the log files not being utilised, as the layout of a web site is a vital consideration when investigating that site's usage [17]. Consequently, there existed a requirement for a method of presenting web site usage information in conjunction with information about that site's structure,

Previous efforts have been made to address this problem. These past works have utilised a variety of representations for web sites including:

- Cyclic graphs – networks of node and link representations which correspond to pages and the links between them.
- Classical hierarchical trees - hierarchical trees of parent and leaf nodes with the home page as the root.
- Cone trees – hierarchical trees in which children nodes are laid out in circles around their parents.
- Radial views – hierarchical trees in which the levels of the tree are positioned in concentric circles.
- Hyperbolic trees – hierarchical trees laid out in hyperbolic rather than cartesian space.

Upon investigating existing attempts, it was discovered that they all suffered from a common weakness, namely that they approached web site representations by treating them as standard trees or graphs. Utilising such an approach, however, precludes the use of those features which are unique to web sites. It stands to reason that a representation that did incorporate such features would possess advantages over current representations. Thus, there existed a need for a method of representing the structure of a web site that does not treat it as a standard tree or graph but as a special entity.

The problem being addressed was thus two-fold: firstly, devise a representation of web site usage information that incorporates a visualisation of that site's structure, and secondly, represent the site structure in a manner that takes advantage of the unique properties of web sites. A tool that was implemented to utilise such a representation could then be used by web site administrators to better evaluate the usage of their sites.

Creating a specialised representation of a web site's usage and structure to solve the problem stated above involved the following steps: investigate web sites and determine which features can be exploited for visualisation purposes, develop an evaluation framework by which to measure past works and the representation being designed, develop the representation taking into account lessons learned from investigation of previous efforts, and finally evaluating the resulting representation.

The first of these steps is discussed in the next section.

8.2 Web Site Features

While web sites may vary greatly in purpose, appearance and structural connectivity, they do share a number of features. These include a base URL, links connecting the various pages that make up the

site and optional text, images, animations, audio and video files. In addition many sites possess an organisational homepage, a global navigation bar and the inclusion of smaller, self-contained web sites. The global navigation bar and the navigation bar included in many sub-sites was of particular interest to this research, as both result in a large amount of redundancy in the web site with regards to links. This redundancy would prove to be a useful feature during the visualisation design.

The next step in the development process was to devise an evaluation framework by which the resulting representation could be appraised.

8.3 Evaluation Framework

Once certain web site features were extracted and classified as being exploitable, the design of the visual representation, or metaphor, could begin. However, in order to evaluate the resulting metaphor, some manner of evaluation framework was first required. This framework takes the form of several factors that were identified after reviewing visualisation literature and investigating previous web site visualisations. The factors can be listed as follows:

- *Structure Representation* – This involves the choice of arrangement chosen to depict the structure of a web site. The chosen configuration should accurately portray the site's structure without confusing the user.
- *Data Representation* – This concerns the manner in which information (such as page accesses) are encoded in the visualisation. Data should be represented in such a way that their values are easy to ascertain and that interesting patterns are readily apparent.
- *Data Exploration* – This relates to the mechanism of navigating around the visualisation in order to examine various parts of the site, as well as to the technique/s used for allowing users to obtain more details about items of interest. Navigation in the visualisation should be consistent and easy to perform.
- *Scalability* – This refers to how well the visualisation scales with size. As it is possible to encounter web sites consisting of several thousand pages, the visualisation should be able to gracefully handle sites of reasonable size.
- *Context Maintenance* – With the size of large web sites it is unlikely that one would be able to survey an entire site in great detail in the same view at the same instant. Instead, users examine subsections of a site, which are usually depicted in some type of magnified display. This issue therefore deals with a user being able to keep track of the relative position of the subsection, or area, of the site they are currently viewing with regards to the layout of the rest of the site.

With this framework in place, development of the metaphor could then commence.

8.4 Metaphor Development

The initial attempt at designing a web site representation made use of a windows directory type browser. The primary focus of this attempt was to address the first aim described in Section 8.1, namely to incorporate usage statistics currently available in most log analysis products into a visualisation of the site structure. The directory tree browser view is used for obtaining an overview of the site. Pages of interest could then be identified and viewed in a secondary view, which displayed branches of the site as either a flat hierarchical tree or else as a three dimensional cone tree. Various usage statistics were depicted in this view as bars placed on the page icons.

Once the design of the metaphor was complete a prototype system was implemented and an initial evaluation took place. This evaluation took the form of a heuristic evaluation that involved a web site usage expert from industry. The results of the evaluation revealed some serious weaknesses in the metaphor, especially with regards to the evaluation framework factors of structure representation and scalability. The metaphor was thus abandoned and development of a second metaphor then started.

In developing the second metaphor, more focus was given to the second aim established in Section 8.1, as a conscious effort was made to incorporate and take advantage of web site features. The resulting metaphor uses a modified tree representation. Horizontal fans representing branches of a site are arranged in semi-circles around a vertical column that represents a global navigation bar. The visualisation makes use of varying levels of detail, with more details being displayed the closer the viewpoint is. Utilising the concept of navigation bars allows a large number of links to be implied and therefore not explicitly shown. In addition, information concerning areas of a site as opposed to individual pages is available.

Upon completion, the second metaphor then underwent the same heuristic evaluation as the first. This time however, the results were much more promising so it was decided to proceed with an implementation of a full system and further evaluations.

8.5 Metaphor Evaluation

The visualisation was evaluated by means of two sets of experiments, the first of which tested the intuitiveness of the metaphor and the second which tested the users ability to use the metaphor.

The first test involved six subjects, who with no previous experience or knowledge of the metaphor, were asked to identify and interpret various features of the metaphor. The results show

that the test subjects were able to correctly interpret most of the metaphor. The concept of link implication did confuse most of the users though.

The second test involved a task-oriented user experiment. The experiment required users to answer a questionnaire using the visualisation. This questionnaire was divided into sections, each of which tested understanding of a different aspect of the visualisation.

The experiment involved twenty six users from a computer science background with varying degrees of web experience. The test subjects were given a short tutorial on the visualisation before starting the experiment. Once the experiment was complete the results were collected and analysed. These are presented in the next section.

8.6 Evaluation Results

The results of the evaluations were generally very promising although a few areas of concern were identified. The results will be discussed in terms of the factors making up the evaluation framework outlined in Section 8.3.

8.6.1 Structure Representation

The results of the two experiments indicated that the metaphor is able to represent the structure of a web site in a manner that is meaningful to users. The intuitiveness test revealed that even subjects who had never seen the metaphor before and had received no training whatsoever could correctly identify the homepage, global navigation bar and page and link representations.

The user experiment showed that the concept of link implication was clear to most users, as only three of the twenty three subjects could not correctly answer any questions involving link implication. The remaining subjects were able to utilise the concept to correctly answer the majority of questions requiring the realisation that links were present that were not explicitly shown, as was indicated by the solid scores for the structure representation section of the questionnaire. With increased exposure to the link implication idea and further training it is believed that these scores will improve further.

8.6.2 Data Representation

With regards to data representation, the intuitiveness test indicated that the concept of varying line intensity to represent usage was readily understood. The varying levels of detail displayed by the visualisation did not seem to cause any difficulties. There were some misconceptions regarding items such as the colour of the wedges and the significance of the varying line width. However, the user experiment showed that once these features were explained, they were effectively used as the

scores in the data representation sections of the questionnaire were quite high. One problem that was identified was the use of colours for portraying usage of incoming and outgoing links, as the chosen colours (blue and purple) were not easily distinguishable at certain intensities. This problem should be easily remedied however.

8.6.3 Scalability

Scalability could be said to be generally well handled due to the combined use of branch approximations, link implications and varying levels of detail. Users scored very well in the section of the questionnaire requiring them to provided details about individual pages of a relatively large web site consisting of over 16000 pages. There was some cluttering experienced at the highest level of detail, however. Reducing the amount of this clutter through the use of various techniques such as altering the distance at which details are visible and using mouse-over operations, should not prove to be too difficult.

8.6.4 Context Maintenance

Context maintenance was well addressed according to the test subjects' responses and comments. No user claimed to have experienced a loss of context. In addition, upon being asked about the extent to which they agreed to the statement that they were always aware of where the page they were currently viewing was located in terms of the overall site structure, only one subject disagreed (rating of 2 out of 5), five subjects were neutral (3 out of 5) and the rest of the twenty six subjects agreed or strongly agreed. In terms of preferences, the majority of users found the directory tree overview easier to both use and understand with a minority preferring the frustum view.

8.6.5 Data Exploration

Data exploration was unfortunately not examined during the intuitiveness test as the focus of that test was to investigate users' perception of the various features of the visualisation. The user experiments did expose the fact that a trackball is not intuitive to use and a large number of subjects experienced problems navigating around the visualisation. Performances did improve after multiple explanations were provided on how to effectively use the mechanism. What became apparent was that either a better mechanism needs to be found or else more focused training on how to use the trackball is required.

8.7 Recommendations

Overall, the results of the evaluation turned out to be quite positive. Although there were a few minor issues encountered, none of these should prove to be serious enough to severely hinder the use of the metaphor. A particularly pleasing aspect of the evaluation was that those users who most closely resembled the targeted end users, namely those with web site administration experience, all scored highly. The aims of this dissertation to create a visualisation of web site usage that incorporated the site structure but that treated web sites as unique entities and not as general trees or graphs, can thus be said to have been met,

It is recommended therefore that further research be carried out regarding the metaphor, in order to improve upon it. Some suggested avenues for this research are listed in Section 8.9.

8.8 Observations

A few observations were made during the course of the user experiments.

The tendency of test subjects to choose the tree overview window to answer questions which would be easier to answer with the frustum overview (i.e., would require less manipulation) seems to suggest that users are likely to utilise a mechanism that they are more familiar with than an unfamiliar one that may be more suited to the task at hand.

The experiences of the users performing the experiments would seem to indicate that trackballs are not inherently intuitive or easy to use. At the very, least they do not appear to be suited for manipulating non-solid appearing objects (as opposed to seeming solid enclosed mesh models). However, further research would need to be performed in order to confirm this.

8.9 Future Work

Future work can be separated into several categories, namely:

- Work that needs to be performed on the current system in order to address certain weaknesses,
- Investigation of whether the link implication concept can be adapted to be used in other metaphors,
- Extensions to the metaphor to add to its functionality, and
- Other applications which the metaphor could be used for.

These are discussed in turn below.

8.9.1 Current Concerns

The user experiments identified certain issues that would have to be addressed.

Cluttering needs to be reduced at the highest level of detail, which could be achieved by making use of mouse-over operations to reveal page icons of those pages that are not currently selected. In addition, users should be able to toggle the currently selected page indicator on and off.

A viable alternative to the trackball mechanism utilised for navigation has to be found. Since a large number of users encountered problems while attempting to "fly" through the visualisation, perhaps a better choice would be to implement a free moving camera system that would permit such navigation behaviour. An additional button could then be available for users to reset the camera position at any time.

8.9.2 Incorporating Link Implication in Other Metaphors

Should training or prolonged exposure increase the ability of users to apply the concept of implicating links, then by reducing the number of links required to be displayed, it may make other techniques used and then abandoned in the past more viable. It could also improve existing metaphors provided a method was found of incorporating it.

8.9.3 Extensions to the Metaphor

Future possible extensions include incorporating provision for viewing alterations over time. Such a provision could provide feedback with regards to the effect of changes to structure in terms of usage. One viable method of illustrating time alterations 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.

In addition, a method of identifying navigation bars resulting from scripting languages needs to be developed. A possible approach would be to utilise image processing techniques, as navigation bars usually form uniform, contiguous patterns, which could be identified, especially if they appear on multiple pages.

8.9.4 Future Applications

By making certain modifications, the metaphor can be adapted to be used in areas other than site usage visualisation. Two possible examples are described below.

Visualising Web Site Queries

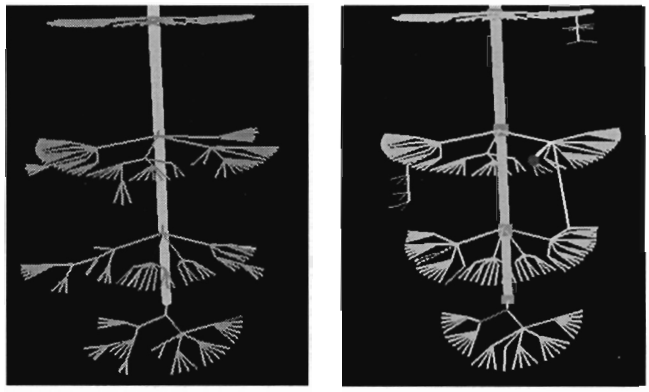
Many web sites contain their own search engine, which the user utilises to find pages of interest to them. The results returned by these engines typically contain the relevant links and a small section of text in which the search keyword was found. There is also normally a relevancy score attached to each result. The problem with such results is that the user is provided with very little information to correctly identify those pages with the desired contents.

The metaphor devised for visualising web site usage could be adapted to visualise these query results instead. The structure of the site will be visualised as usual. In this instance, however, the intensity of the links will correspond to the relevancy the associated pages scored in terms of the search results. Those pages not listed in the results would be a neutral grey colour, as shown in Figure 30a. Showing the user where the result pages are located in the site by displaying them on a visualisation of the site structure, allows the user to infer more information about which of the pages returned by their search contain the desired content. As an example, suppose a user was browsing a site looking for technical support for a product they had purchased. The user then enters the products name into the search engine and waits for the results. The engine then returns two pages which both mention that product. By looking at the visualisation the user can immediately see that one page is not what they are looking for as it is located in the "purchase order" section of the site. The user then elects to follow the second link, which according to the visualisation, is located in a more promising branch of the site. It should be noted that the visualisation is not intended to replace traditional text-based result displays but rather to supplement them as an additional aid.

Visualising Web Site Navigation

The current tools utilised by web browsers to aid the user in navigating web sites, such as the forward and back buttons and the history list, are extremely limited. Aside from sometimes behaving unpredictably, they provide the user with a poor idea of his/her current location. In addition, by not giving some indication of the site's structure they restrict the user from making well informed choices as to which future links to take. By modifying our metaphor slightly, we can create a visualisation that, when used in conjunction with a normal browser, will greatly improve the user's ability to effectively navigate a site.

Taking the current metaphor, the page currently being viewed can be indicated by a red sphere. This sphere is then analogous to the "You are here" icon utilised by many maps. Pages which are accessible from the current one are then displayed as green links, green being the colour normally associated with proceeding. A recent history of the pages already visited is shown by yellow links, the intensity of which indicates recency, i.e., pages just visited are bright yellow whereas pages viewed earlier are darker yellow. In cases where usage information is available, that data can be



(a) (b)

Figure 30: **Future Applications.** (a) *Visualising Web Site Queries*. The left-hand figure provides an idea of what a visualisation of the search results of a web site query would look like using our metaphor. The links to those pages containing the search item are shown in various shades of blue, the intensity of which correspond to the relative relevancy of those pages. Unrelated pages are indicated by neutral grey links. (b) *Visualising Web Site Navigation*. This image illustrates how our metaphor could be adapted to visualise web site query navigation. The red sphere indicates the user's current location, while the green links show pages accessible from the current one. A history of the user's visits are shown by the yellow links, the intensity of which correspond to the recency that the user viewed those pages.

encoded in the thickness of the lines representing links. The user would thus be able to see which pages were popular with other visitors and use this information to supplement their own navigation. Figure 30b provides an idea of how an implementation of a web site navigation tool would look.

Appendix A

Web Site Usage Visualisation Questionnaire

A.1 Web Experience

The following questions seek to determine your level of experience in creating and maintaining web sites.

1. Have you ever made a web site? If yes, was it a personal, commercial or informative web site?
2. Have you ever made use of a web site usage analysis tool (sometimes known as a log file analyzer)?
3. Have you ever served as a web site administrator?

A.2 Training

The questions in the remaining sections ask you to examine and interpret various features of the visualisation. For each section you will have to click on the matching button on the bottom left of the screen. Start by selecting the warm up option. This section contains a few simple questions to test whether you understood the tutorial.

1. Is there a direct link from page A to page E (i.e., you can reach E from A via a single click)?
2. Is there a direct link from page B to page E?
3. Why is the link to page C a different colour to the other links (i.e., red)?

4. Approximately how many hits did page E get?
5. Which page received more hits; D or G?
6. What does the grey vertical column (line) at page H represent?

A.3 Section A

Select the "Section A" radio button located in the "Sections" box in the left bottom corner.

1. Name all the pages that can be directly reached (i.e., via a single click/link) from the following pages:
   (a) 28
   (b) 10
   (c) 12
   (d) 11
   (e) 19
2. A path between pages is defined by the links which need to be taken to get from one page to another. Assume that you can reverse a link (e.g. from page 33 to get to page 28) by pressing the "back" button on a browser (i.e., treat links as two-way links). An example of a path is from page 36 to page 10, which consists of 2 links, i.e., from 36 "back" to 32 and then from 32 back to 10 (36-32-10). Another example would be from page 1 to page 26, consisting of 3 links, i.e., 1-5-17-26.

   Bearing in mind that multiple paths are possible between the same two pages, write down the shortest path between the following pages:
   (a) page 5 to page 11
   (b) page 20 to page 7
   (c) page 31 to page 5
   (d) page 18 to page 21

A.4 Section B

Select the "Section B" radio button.

1. Identify the wedge which was visited the most (i.e., received the most no. of hits).
2. Identify the wedge which received the least amount of hits.
3. On what did you base your above answers?

A.5 Section C

Select the "Section C" radio button.

1. Which branch (section) received more hits, A or B?
2. On what did you base the above answer?

A.6 Section D

Select the "Section D" radio button. Select page 14 by right-clicking on the link (line) to it.

1. From which page in the site did **most** of the traffic for page 14 come?
2. From which page in the site did **least** of the traffic for page 14 come?
3. To which page in the site did **most** of the traffic from page 14 go to?

Select page 30 by right-clicking on the link (line) to it.

1. To which page in the site did **most** of the traffic from page 30 go to?
2. From which page in the site did **most** of the traffic for page 14 come?

Deselect page 30 by right-clicking anywhere other than on a link (page).

1. Which page received more hits. 14 or 30? Does this surprise you in light of your answer to your answer to the previous question? If so, what is the explanation for this?

A.7 Section E

Select the "CS Site" radio button.

For the following questions it may be helpful to locate pages by typing their full URL's in the "Find Page" box (found in the bottom right hand corner) and pressing enter (Note: If the URL points to a directory remember to add a trailing slash). Information about the currently selected page appears in the "Selected Page" box, which is above the "Find Page" box.

You may be asked questions which state that you must use the left hand window only. Note that the left hand window consists of two parts:a directory tree type visualisation and a zoomed out representation of the right window, showing its frustum. The view can be switched between these two by clicking on the "Tree" or "Frustum" tabs. Either view may be used for answering "left hand window only" questions.

A "global link" is a link which is part of the *global navigation bar* or menu. The global navigation bar is the main vertical column to which the home page is attached as opposed to a local navigation bar which is every other vertical column. Thus the global link for the page **http://www.cs.uct.ac.za/Courses/CS300W/DP/prolog_manual/Contents.html** is the link to the page **http://www.cs.uct.ac.za/Staff/** (type the above link in the "Find Page" box and use either of the left views to confirm this).

The following questions concern a visualisation of the Department of Computer Science web site which consists of about 16000 pages.

1. What type of page (entry/exit/entry-exit/normal) is **http://www.cs.uct.ac.za. clyness/Personal/**?
2. Using only the left hand window identify which global link the above page is a (great, etc.) grandchild of?
3. Which of the two left hand views did you use to answer the above question **Tree** or **Frustum**?
4. What type of page (entry/exit/etc) is **http://www.cs.uct.ac.za/%7Egaz/teach/hci/default.htm**?
5. Using the left hand window only: Within the global navigation bar, does the global link that is the grandparent/great-grandparent, etc. of the above page occur before or after the global link **http://www.cs.uct.ac.za/Students/** (Links in the global navigation bar appear in order, i.e., the first link appears at the top and so on.)?
6. Which of the two left hand views did you use to answer the above question **Tree** or **Frustum**?
7. How many hits did the page **http://www.cs.uct.ac.za/Research/DNA/about.html** receive?

8. Using the left hand window only: Trace a path from the home page to the above page (write down the URL's)
9. Which of the two left hand views did you use to answer the above question **Tree** or **Frustum**?
10. What page type (entry/exit/etc.) is **http://www.cs.uct.ac.za/courses/CS400W/**?
11. Using the left hand window only: Does the global link which is the (great, etc) grandparent of the above page lead to more or fewer pages (i.e., has more children, grandchildren, great-grandchildren, etc.) than the global link **http://www.cs.uct.ac.za/Information_for_Prospective_St**?
12. Which of the two left hand views did you use to answer the above question **Tree** or **Frustum**?
13. Which of the two left hand views did you find easier to *understand*?
14. Which of the two left hand views did you find easier to *use*?

A.8 Ratings

The next few questions require you to give the response which you feel is the most accurate. The possible answers range from 1 to 5, where 1 means you strongly disagree with the statement and 5 means that you strongly agree with it.

1. The system was still usable even with a fairly large site (the CS site)
2. I had a good idea of where the page/s I was currently viewing were located in terms of the rest of the site.
3. It was easy to find the information I needed.
4. The information provided was easy to understand.
5. The visualisation provided me with a good idea of the structure of the web site being visualised.
6. The visualisation provided me with a good idea of how much all the pages were visited/viewed.

A.9 Comments

1. List the most negative aspects of the visualisation.
2. List the most positive aspects of the visualisation.

3. List any further comments you may have.

Bibliography

[1] Active Concepts Products. Funnel web professional. http://www.activeconcepts.com/prod.html.

[2] Apple. Apple hotsauce. No longer supported.

[3] G. D. Battista, R. Lillo, and F. Vernacotola. Ptolomaeus, the web cartographer. http://www.dia.uniroma3.it/ ptolemy/.

[4] A. Buchner and M. Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *ACM SIGMOD Record*, 27(4):54–61, 1998.

[5] S. K. Card, J. D. Mackinlay, and B. Schneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kuafman Publishers, Inc. San Francisco, California, 1999.

[6] S. K. Card, T. P. Moran, and A. Newell. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum, 1983.

[7] J. Carriere and R. Kazman. Interacting with huge hierarchies: Beyond cone trees. Proceedings of the IEEE Symposium on Information Visualization, pages 74–81, Atlanta, Georgia, USA, October 1995.

[8] J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through connectivity. Proceedings of the 6th International World Wide Web Conference, 1997.

[9] E. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, and S. Card. Visualizing the evolution of web ecologies. Proceedings of the Conference on Human Factors in Computing Systems, CHI '98, pages 400–407, Los Angeles, CA, 1998.

[10] E. H. 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. Proceedings of ACM CHI 2000 Conference on Human Factors in Computing Systems, pages 161–168, The Hague, The Netherlands, 2000.

[11] CLEARWeb, Inc. Clearweb. http://www.clearweb.com/.

[12] J. Conklin. Hypertext: An introduction and survey. *IEEE Computer*, 20(9):17–41, September 1987.

[13] Dynamic Diagrams Inc. http://www.dynamicdiagrams.com/Home.htm.

[14] J. Eighmey. Profiling user responses to commercial web sites. *Journal of Advertising Research*, 37(3):59–66, 1997.

[15] J. Goldberg. On interpreting access statistics. http://www.cranfield.ac.uk/docs/stats. Cranfield Computer Centre, Cranfield University.

[16] N. P. Group. Nature neuroscience. http://www.nature.com/neuro/info/site_guide.

[17] 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.

[18] Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, /2000.

[19] I. Herman, M. Delest, and G. Melancon. Tree visualisation and navigation clues for information visualisation. *Computer Graphics Forum*, 17(2):153–165, 1998.

[20] 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.

[21] IBM Haifa Research Lab. Mapuccino. http://www.alphaworks.ibm.com/tech/mapuccino.

[22] InContext Systems. Webanalyzer v2.0. http://www.incontext.com/WAinfo.html.

[23] Inxight Software. Tree studio. http://www.inxight.com/.

[24] IXACTA. Ixsite. http://www.ixacta.com/products/ixsite/.

[25] P. Kahn. Mapping web sites. http://www.dynamicdiagrams.com/seminars/mapping/map6.htm, May 1999. Dynamic Diagrams Seminar.

[26] G. Landow. *Hypertext: The Convergence of Contemporary Critical Theory and Technology*. The John Hopkins University Press, 1992.

[27] J. Mackinlay, S. Card, and G. Robertson. Perspective wall: Detail and context smoothly integrated. Proceedings of the ACM SIGCHI '91 Conference on Human Factors in Computing Systems, pages 173–179, New Orleans, LA, April 1991. ACM, ACM.

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

[29] Merz.com. Netscope. No longer available.

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

[31] J. Mogul and P. Leach. Simple hitmetering and usage-limiting for http, 1997. RFC 2227.

[32] S. Mukherjea and J. Foley. Visualizing the world-wide web with the navigational view builder. *Computer Networks and ISDN Systems 27, Elsevier Science*, 1995.

[33] M. Mulvenna, S. Anand, and A. Buchner. Personalisation on the net using web mining. *Communications of the ACM*, 43(8):123–125, August 2000.

[34] T. Munzner. Drawing large graphs with h3viewer and site manager. Proceedings of the Symposium on Graph Drawing GD '98, pages 384–393, Springer-Verlag, 1998.

[35] T. Munzner. Exploring large graphs in 3d hyperbolic space. *Computer Graphics and its Applications*, 8(4):18–23, July/August 1998.

[36] T. Munzner and P. Burchard. Visualizing the structure of the world wide web in 3d hyperbolic space. Proceedings of VRML '95, pages 33–38. ACM, ACM, 1995.

[37] J. Nielsen. *Designing Web Usability*. New Riders.

[38] P. Pirolli, J. Pitkow, and Rao. Silk from a sow's ear: Extracting usable strutures from the web. Conference Proceedings on Human Factors in Computing Systems (ACM CHI '96), 1996.

[39] J. Pitkow. In search of reliable usage data on the www. Proceedings of 6th International World Wide Web Conference, Santa Clara, CA, April 1997.

[40] 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.

[41] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey. *Human–Computer Interaction*. Addison–Wesley, 1994.

[42] 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, pages 189–194, 1991.

[43] Sane Solutions. Nettracker enterprise. http://www.sane.com/products/NetTracker.

[44] M. Sarker and M. H. Brown. Graphical fisheye views. *Communications of the ACM*, 37(12):73–84, December 1994.

[45] B. Schneiderman. Tree visualization with tree–maps: A 2–d space–filling approach. *ACM Transactions on Graphics*, pages 92–99, January 1992.

[46] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. Technical Report UMCP-CSD CS-TR-3665, College Park, Maryland 20742, U.S.A., 1996.

[47] M. Spiliopoulou. Web usage mining for web site evaluation. *Communications of the ACM*, 43(8):127–134, August 2000.

[48] M. Spiliopoulou, C. Pohle, and L. Faulstich. Improving the effectiveness of a web site with web usage mining. Proceedings of the Workshop on Web Usage Analysis and User Profiling, WEBKDD '99, pages 51–56, San Diego, California, 1999.

[49] E. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut, 1983.

[50] E. Tufte. *Envisioning Information.* Graphics Press, Cheshire, Connecticut, 1990.

[51] E. Tufte. *Visual Explanations.* Graphics Press, Cheshire, Connecticut, 1997.

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

[53] Visual Web. Visual web tool. http://world.isg.de/World/2_Internet/Visual_Web/index.html.

[54] Vividence. http://www.vividence.com.

[55] WebCriteria. Webcriteria siteprofile. http://www.webcriteria.com.

[56] WebTrends. Webtrends log analyzer. www.webtrends.com.

[57] G. Wills. Nicheworks – interactive visualization of very large graphs. Graph Drawing '97 Conference Proceedings. Springer-Verlag Lecture Notes in Computer Science, 1997.

[58] A. Wood, R. Beale, N. Drew, and B. Hendley. HyperSpace: A World-Wide Web visualiser and its implications for collaborative browsing and software agents, 1995.