The Visualisation of Internet Usage

Lee Son, G.S.; Calitz, A.P.
Department of Computer Science and Information Systems, University of Port Elizabeth
P.O. Box 1600, Port Elizabeth 6000, South Africa.
Tel: (041) 504-4254; Fax: (041) 504-2831
E-mail: csbsls@upe.ac.za ; csaapc@upe.ac.za
Topic: Network Management: Policy Based Management

Abstract — The amount of data that is collected and generated by managing Internet usage can be enormous, especially if the data has been collected over an extended period of time. The visualisation of usage statistics generated by reporting tools are important for gaining insight into the information that is contained within the vast amounts of raw data. This data needs to be analysed in order to determine whether the users are adhering to usage policies or whether they are abusing their access by visiting unacceptable websites. The visualisation and accurate interpretation of the data is the key issue to the successful management of Internet usage. Visualisation techniques currently being employed to visualise Internet usage data manage to achieve a certain level of representation but fails to represent the rich details of information that is offered by the data. This paper introduces two new information visualization techniques that can be used by network managers for Internet usage management.

Index terms: Internet usage management, Information visualisation

I. INTRODUCTION

Information overload is considered one of the fundamental human-computer interaction problems today. In an effort to cope with this problem, more and more users are turning to information visualisation (IV). Information visualisation has helped users in a broad range of industries around the world make critical business decisions by allowing them to eliminate the information overload. Instead of wading though endless spreadsheets and text analyses, executives can obtain a quick overview with the graphics offered by visualisation techniques, and still find the level of detail they require [1]. The visual presentations allow users to see patterns they that wouldn't have noticed otherwise.

The use of the Internet and Intranets by all types of enterprises, such as businesses, education institutions and government departments, is growing rapidly [2]. The explosion in use of personal computers with access to the Internet can greatly increase the efficiency, productivity and success of an enterprise. This is done by using the Internet to facilitate useful research or to provide quick answers or to aid effective collaboration between colleagues. However, Internet access has significant negative potential as the continued viability of the Internet and the resulting economic benefits depend on the performance levels of the Internet to meet the demands of existing and emerging applications [3].

Employees can waste considerable working time and network resources by accessing various websites for personal reasons. According to research conducted by International Data Corporation, 30% to 40 % of Internet use in the workplace is not related to business [4]. A solution to the problem is effective and efficient employee Internet usage management.

Managing employees' use of Internet access resources is a sensitive and complex task that is crucial to productivity, profitability and morale in the workplace [5]. Internet usage management consists of two main areas:

- Usage policies; and
- Reporting tools.

Many enterprises have usage policies detailing what is considered acceptable behavior when accessing the Internet, however without reporting tools that mirror the policies they can never truly be enforced. The reporting tool allows management to determine what types of sites individual users or groups of users are visiting, when they were visited, what type of content was sought after by the users, the amount of bandwidth that was consumed in the process and if the visits were in compliance with the usage policy.

These reporting tools are readily available *off the shelf* software, however the graphical reports that are generated by most of these products are static reports that contain vast amounts of information that can be cumbersome to navigate. These reports also offer little efficiency in obtaining details on demand that will aid in making managerial decisions.

Almost all outbound Internet-access reporting products use pre-existing log files as their source of raw data [5]. Log files are tables of highly detailed electronic records that list all hits associated with all outbound activity. The resulting log files are huge in size containing an enormous amount of data to manage. The size of log file that contains usage data for a single day can range from tens of megabytes to hundreds of megabytes depending on the number of users. The use of coordinated visualisations that simultaneously present multiple views of relevant information might be used to address the problems associated with interpreting the log files and Internet usage reports [6].

Exploring and analysing vast volumes of data is becoming increasingly difficult. Information visualisation can help to deal with the flood of information by directly involving the

user in the data exploration and mining process. The basic idea of visual data exploration is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions and directly interact with the data [7]. Visual data mining techniques have proven to be of high value in exploratory data analysis and they also have high potential for exploring large databases.

The objective of this paper project is to investigate the implementation of two new information visualization techniques to determine the usage of the Internet, to provide network managers with a more comprehensive understanding of the data.

II. Internet Usage Data

Internet usage data determines how the Internet on a network is being used and what it is being used for. From this data the issues that can be addressed include identifying which users visit which site at what time, the type of content they were seeking and the amount of bandwidth consumed in the process. The data is obtained from log files kept in proxy servers, firewalls or caching appliances.

Log files are automated records consisting of a highly detailed list of all hits associated with all outbound activity. A hit is as any Web browser related interaction whose purpose is to display a Web page in the browser. A hit also includes all individual elements of information that appear as a result of the interaction. Examples include each individual graphic that is displayed, advertising banners and audio and video files.

There are two log file format standards to consider when collecting Internet usage data: the World Wide Web Consortium (W3C) extended log file and Microsoft's Internet Security and Acceleration (ISA) Server [8] log format. ISA Server log files consist of tables of information with each field referring to an item related to the data that makes up a recorded hit. The three primary fields include the user name, the Uniform Resource Locator (URL) and the time-stamp. These fields indicate which user visited which site at what time. The remaining fields of the log files depend on the type of log file and what information is chosen to be logged.

Log files are useful in determining the usage of the Internet, however a direct correlation between the information included in the log files and users' usage patterns are not easily obtained. For example, the type of Web sites that are being visited by users cannot be directly interpreted, as the URL only gives an indication of where the Web site is and not what it is about. Added to this limitation is the fact that not all URL's indicate actual Web site visits. A majority of URL's listed in the log file are hits pointing to individual pieces of information, since most Web sites are graphic-rich [9]. To illustrate this point and help differentiate between hits and visits consider two users visiting two different Web sites.

User 1 visits http://www.whitehouse.gov/text/index.html. This is a text-only page with no images, banners or advertisements. The log file registers only one hit. In this case one hit equals one visit. User 2 visits http://www.cnn.com/. This is complex page with 22 images, banners and advertisements. This time the log file registers 23 hits, one click and 22 data items. In this case 23 hits equal one visit. Looking at the log file it would also seem as though User 2 was much more active than User 1, even though the amount of user-initiated activity is the same in both cases, namely one click of the mouse.

To accurately analyse the data so as to determine whether or not users are adhering to usage policies directly from the log file is a difficult and laborious task as the log files are difficult to interpret. To solve this problem a visual representation of the data must be presented using some form of information visualisation technique.

III. Information Visualisation

Information visualisation (IV) is defined as the use of computer-supported interactive, visual representations of data to amplify cognition [10], where cognition is the acquisition or use of knowledge. The purpose of IV is to gain insight. The main goals of this insight are discovery, decision-making and explanation. IV reduces the search for data by grouping or visually relating information, effectively, thereby compacting information into a small space [11]. By incorporating numerous techniques that promote and establish a visual environment, users can gain a better and more comprehensive understanding of the data that is being visualised.

In addition to the visualisation technique, it is necessary to use interaction and distortion techniques so that the data can be explored effectively [7]. Interaction techniques allow for direct interactions with the visualisations and allow for the visualisations to be dynamically changed according to exploration objectives. The aim of distortion techniques is to help in the data exploration process by providing means for focusing on details while preserving an overview of the data. The basic idea behind distortion techniques is to display portions of the data with a high level of detail, while others are shown with a lower level of detail. An example of a distortion technique is the Fisheye View [12].

These techniques can allow for the hierarchical searching of data by using overviews to locate areas that require more detailed searching. They can promote zooming capabilities in order to improve the user's perception of a data element. At the same time they allow for the drilling down of large data sets in order to display details-on-demand on certain data elements. Information visualisation also facilitates the identification of patterns in data.

IV. Current IV Techniques utilised for managing Internet Usage

The types of visualisation techniques currently employed by Internet usage management tools, such as ISA Server [8] reports and Cyfin Reporter [13], are graph-based techniques. The ISA Sever reports make use of to two-dimensional line graphs, (see Figure 1), and bar charts, (see Figure 2), to help illustrate usage statistics. Figure 1 depicts a line graph of traffic measured in MB over a course of 24 hours. Figure 2 shows a bar chart listing the most active users over a certain period of time. Figure 3 is an example of another Internet usage report generated by Cyfin. The report in Figure 3 is also limited to bar graphs.

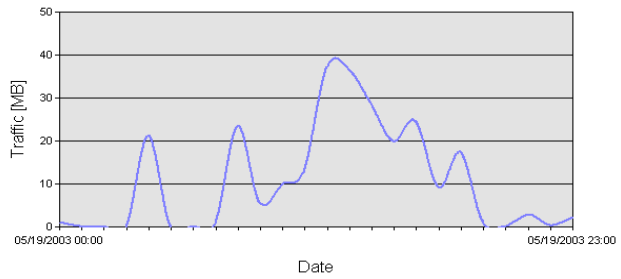

Figure 1: ISA Server Line Graph depicting Traffic for 1 day

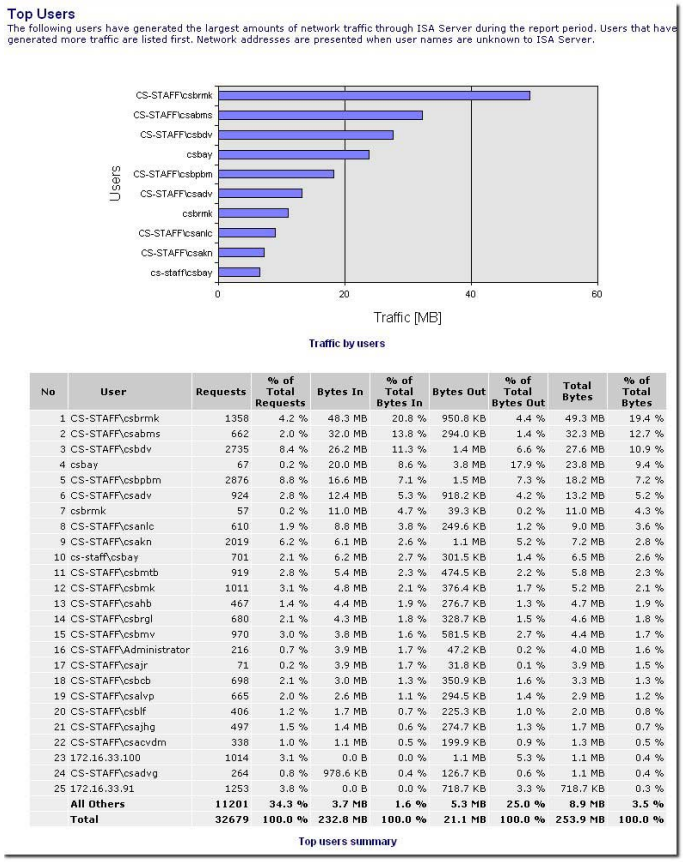

Figure 2: ISA Server Bar Chart of most active users

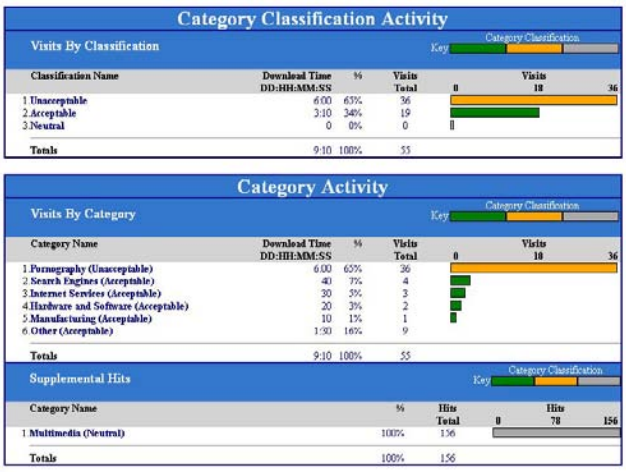

Figure 3: Cyfin report's use of bar charts

There are no interaction or distortion techniques associated with the either the bar charts or the line graphs, thus restricting the effectiveness of the techniques employed. Both graph-based techniques are also restricted to two dimensions, but Internet usage data is comprised of multivariate datasets which include multiple dimensions such as time, hit name, hit type and hit size. Thus there is significant potential to utilise other types of visualisation techniques to best suit the visualisation to Internet usage data.

IV. Analysis and Design of IV Techniques To Better Visualise Internet Usage Data

In order to make a decision on what visualisation techniques to employ the dimensions of the usage data dataset must first be defined. Each dataset is a subset of all of the collected usage data for a specified period of time and includes the following dimensions:

- Timestamp;
- Website URL;
- Website category; and
- Kilobytes per hour.

The time when each website was visited is recorded in the timestamp, with the URL indicating where the user went and the category detailing what content was being sought. The dimension Bytes per Hour collects the total number of bytes downloaded in an hour and answers the question of how much bandwidth the user made use of.

The website categories can be defined as three categories; Acceptable, Unacceptable and Other; or they can be more detailed, defining each website in categories such as News and Media, Entertainment, Sport and Search Engines. In order to categorise a website the URL is checked against a list containing thousands of already defined URL's. Cyfin Reporter [13] was used to obtain these categories. Another benefit of parsing the log files through Cyfin was that it reduces the size of the dataset by listing website visits and not each individual website hit.

Datasets can be defined for individual users or for groups of users. Each group consists of a number of users. For example, by selecting the group 'Staff' all of the user id's that are staff members is selected. A dataset is also determined by the time frames consisting of a starting date and time and an ending date and time.

With the dataset defined various IV techniques to visualise the data were employed. The choice of what techniques to employ were based on issues such as the size of the dataset, how much data can be displayed in the technique and the ease at which the technique could be interpreted. Specific techniques that were investigated included geometric-projection techniques as the aim of geometric projection techniques is to find interesting projections of multidimensional data sets. It is essentially the visualisation of geometric transformations and projections of the data [14].

Analysing the dataset reveals that the data follows a linear structure based on time. One of the better techniques to use for this type of structure would be a line graph, but line graphs are limited in size and interaction. Displaying a time line depicting each minute of a day would be too large to fit

on the screen and would fail to represent the rich details of information that is offered by the data.

The IV technique developed to visualise usage data is based on a line graph fitted onto a clock-face. 12 o'clock on the clock-face represents the starting time of the dataset time period and with 11:59 representing the ending time. This idea is taken from the spiral graphs developed by Weber, Alexa and Müller [15]. The spiral graph technique is a geometric projection IV technique that supports the analysis of time-series data with each revolution of the graph representing a certain time period.

Each website that is visited is then represented as a marker on the graph. The position of the marker corresponds to the time the website was visited and the colour of the marker either set to indicate the website category of the website acceptability. Clicking on marker will reveal the URL for that specific website providing details-on-demand. Figure 4 illustrates an example of this technique with maker colour set to individual website categories.

The dataset for the example in Figure 4 contains usage data for a single user over a one day time period. Every website that the user visited in that period is recorded in the dataset. In this example the dataset consists of 184 website visits. From the graph it can immediately be seen when the user was most active as well as obtaining a general idea of what type of content was being sought. It can be seen that the user was most active between 11:30 and 12:00 and between 13:30 and 15:15. The majority of the markers are blue in colour which indicates that the user was mostly visiting websites from the category associated with that colour. In this example it would indicate that user was mostly visiting News and Media websites, which can be deemed acceptable or unacceptable depending on the usage policy. This one day graph provides an overview of all 184 website visits.

By changing the time scale of the graph from minutes in a day to minutes in an hour a more detailed look at the data can be obtained. Figure 5 shows an example of an hour graph based on the hour 14:00 – 14:59 obtained from Figure 4.

Moving from the one day graph down to the hour graph and the back to the day graph is known as drilling down and rolling up. The drill-down operation navigates from less detailed data to more detailed data while roll-up operates by climbing up to a less detailed view [16].

Graphs can drill-down and roll-up based on the hierarchy of the time in the dataset. The hierarchy follows as *hour* < *day* < *week* < *month* < *year*. This implies that usage data can be viewed by hour, by day, by week and by year depending on the timescale of the dataset.

Further investigation of the usage data can be achieved by manipulating the graph to show only certain website visits that fall into certain website categories. For example, to view all unacceptable websites that were visited only the website markers that set unacceptable are displayed on the graph. In this way the data can be filtered an analysed in greater detail.

Comparing multiple users or groups can also be achieved using this technique. Overlaying the usage data of one user on top of that of another user would not be possible as that would obscure the view of the data at the bottom, however by first decreasing the size of one graph and then overlaying it onto another would provide a means for comparison to take place. Figure 6 illustrates this point. The outer ring contains usage data for user 1 while the inner ring contains usage data for user 2 during the same time period. It can be seen from this graph how similar of different the two users' Internet surfing patterns are.

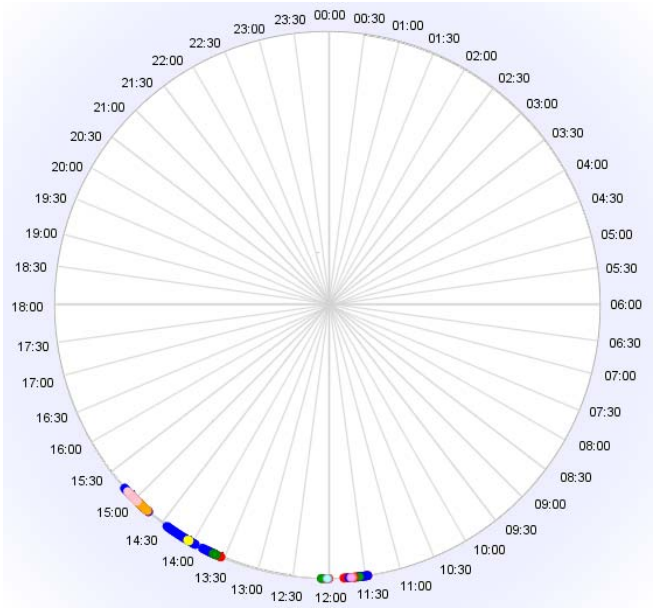

Figure 4: One Day time period graph of a single user

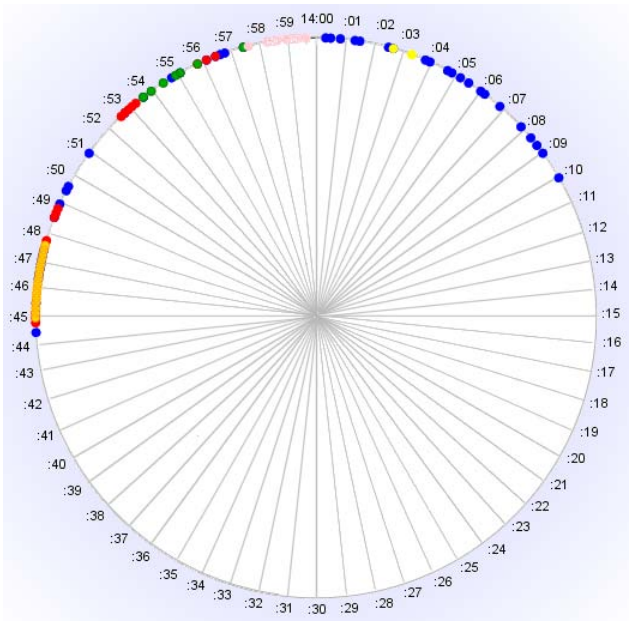

Figure 5: One Hour time period graph of a single user

A single user or group can also be compared to themselves using this technique. By showing multiple days or weeks on the same graph usage patterns can emerge and be identified.

Other capabilities of the technique include the addition of interaction techniques such as zooming, rotating and panning the graph. Zooming is not only a means to display data objects larger but also a means for more details of the data to be represented at higher zoom levels. Zooming combines filtering with limited increases in detail [17]. Rotation can occur in three dimensions, rotating along the x, y and z axes of the graph. This rotation allows the graph to be viewed in

different ways, from which new insights into the data can be gained.

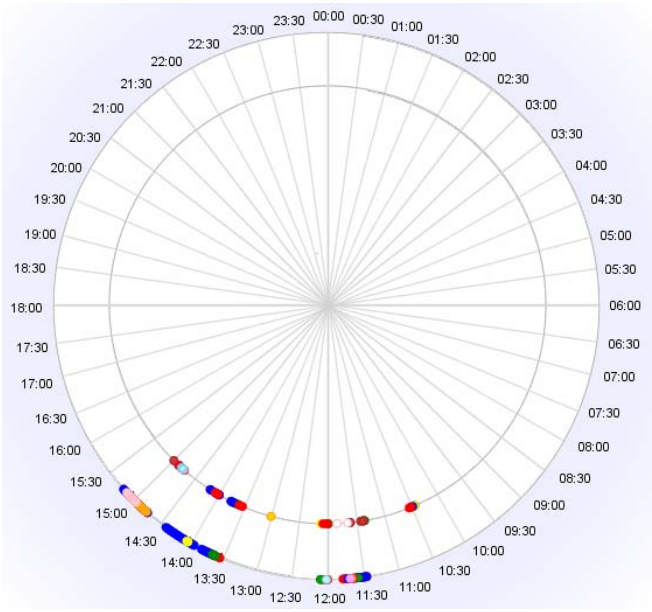

Figure 6: Comparison of two users for the same time period

The ability to see the entire dataset in a single view is the main advantage of this technique over those that are currently in place. This view allows for patterns in the data to be easily identified. The technique also provides the user with the ability to interact with the data, allowing for effective data exploration.

In order to visualise the last dimension of the dataset, namely Kilobytes per Hour, a separate visualisation technique is used. This is mainly due to the fact that the lowest level of granularity of this dimension is an hour, while the other dimensions in the data set is reduced to seconds.

Like the technique used to visualise the actual websites, the technique used to visualise the kilobytes per hour follows a circular nature. This technique is variation of a standard radar plot.

A radar plot is a simple means to illustrate a multivariate dataset, in which the multiple measurements of the variables are plotted and linked on equally spaced radii extending from the centre of a circle to form a radar [18]. Each radius stands for an hour of a day with the value on the radius depicting the number of kilobytes downloaded in that hour. Figure 7 shows an example of this technique. The plot indicates that the user was active between 11:00 and 15:00 and how much bandwidth was consumed during that period. The filled area of the graph depicts the total amount of bandwidth the user downloaded for the day.

Comparing users and groups can also be achieved with this technique. By using multiple colours and transparencies two or more datasets can be overlaid without losing any detail. Figure 8 illustrates how this is done.

Using these two techniques for visualising Internet usage data provides network managers with the ability to quickly discover what sites the users visited, what content users sought, when the users visited and how much bandwidth was used.

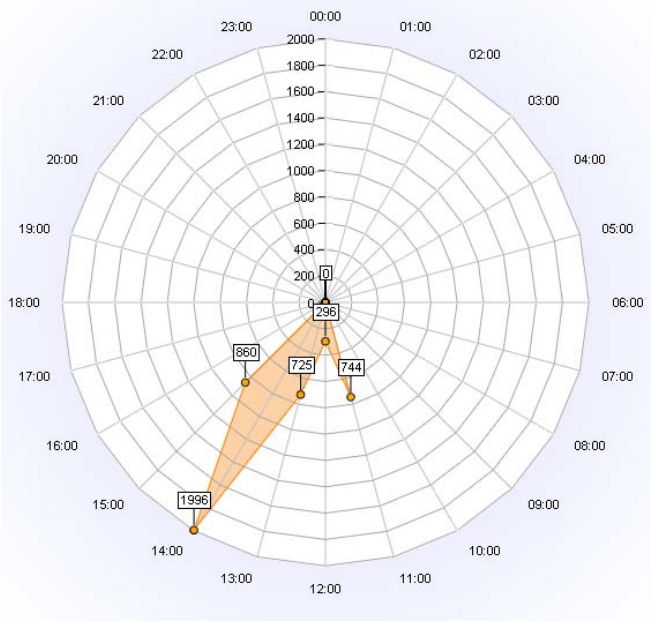

Figure 7: Radar plot showing Kilobytes per Hour for a One Day time period

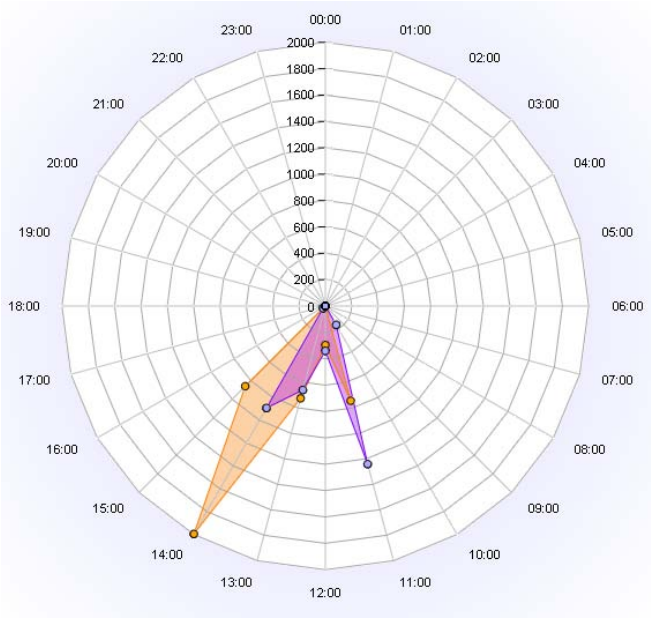

Figure 8: Comparison of two users bandwidth usage for the same time period

V. IMPLEMENTATION

Implementation of the techniques were developed using C#.NET [19], SQL Server [20] and Nevron Chart for .NET [21]. The .NET platform was chosen as the development tool as it provides support for rapid application development allowing quick prototypes to be built. The log file data that is parsed through Cyfin is maintained in a SQL Server database. The use of SQL Server allows for the scalability of the data and also provides security and robustness. Nevron Chart for .NET is a visualisation component for .NET. It has the ability to display dynamic charts in a professional manner while at the same time allowing for the creation of new types of graphs.

VII. Conclusions and Future Work

Information visualisation is characterised by the need for designers to invent a way to transform data into graphical representation [14]. This representation needs to express the important properties of the data and express how different items are related to one another. The visualisation techniques currently being employed to visualise Internet usage data manage to achieve a certain level of this representation but at the same time fail to represent the rich details of information that is offered by the data. The use of different types of IV techniques will be able to solve this problem. Evaluation and usability tests of the two techniques discussed in this paper will be carried out to determine whether they offer better insight into Internet usage data than those techniques that are currently in use. These new techniques will hopefully be shown to be easily interpretable, thus promoting better usage management.

VIII. References

[1] Borzo, J. (2004): Get the Picture. Special Report to the Wall Street Journal, Dow Jones WebPrint Service, 12 January, 2004.

[2] Wavecrest Computing (2002): Wavecrest Computing: Internet Monitoring: Options for Managing Internet Use, URL : www.wavcrest.net

[3] Cross-Industry Working Team (2000): Internet Service Performance: Data Analysis and Visualization. Reston, Virginia, URL: http://www.xiwt.org.

[4] IDC (2000): International Data Corporation : September 2000, URL: www.idc.com

[5] Wavecrest Computing (2002): Wavecrest Computing: Policy-Based Approaches to Internet Monitoring in the Workplace. URL : www.wavcrest.net

[6] Hochheiser, H. and Shneiderman, B. (2002): Coordinating Overviews and Detail Views of WWW Log Data. Institute for Systems Research and Institute for Advanced Computer Studies, University of Maryland.

[7] Keim, D.A (2002): Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, Vol. 8, No. 1, January – March 2002.

[8] Microsoft (2002): Internet Security and Acceleration Server.

[9] Wavecrest Computing (2002): Wavecrest Computing: Internet Monitoring: Hits vs. Visits: Internet Monitoring the Right Way, URL : www.wavcrest.net

[10] Card, S.K., Mackinlay, J.D. and Shneiderman, B. (eds) (1999): Readings in Information Visualization: Using Vision to Think. San Francisco, California, Morgan Kaufmann Publishers, Inc.

[11] Wesson, J.L. and Van Greunen,D. (2002): Visualisation of usability data: Measuring task efficiency. Proc. Annual Research Conference of the SA Institute of Computer Scientists and Information Technologists (SAICSIT 2002), Port Elizabeth. ACM International Conference Proceedings Series, KOTZE, P., VENTER, L. and BARROW, J. (eds). SAICSIT.

[12] Furnas, G.W. (1986): Generalized Fisheye Views. Human Factors in Computing Systems CHI ’86 Conference Proceedings, pp. 16-23.

[13] Wavecrest Computing (2004): Cyfin Reporter.

[14] Zhang, H. (2000): Mining and Visualization of Association Rules over Relational DBMSs. University of Florida

[15] Weber, M., Alexa, M. and Muller, W. (2001): Visualizing Time Series on Spirals. Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01).

[16] Han, J. and Kamber, M. (2001): Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, California.

[17] Carr,D.A. (1999): Guidelines for Designing Information Visualization Applications. Proceedings of the 1999 Ericson Conference on Usability Engineering.

[18] Fienberg, S.E. (1979): Graphical Methods in Statistics. The American Statistician, vol. 33, no. 4, pp. 165-178.

[19] Microsoft .NET Framework (2004): Microsoft Visual C#.NET

[20] Microsoft SQL Server (2004): Microsoft SQL Server 2000

[21] Nevron LLC (2004): Enterprising Charting Solutions.

Biography - Graeme S. Lee Son obtained his B.Sc Honours from the University of Port Elizabeth in 2002, and is currently working towards his Masters degree in Computer Science. His dissertation is entitled “The Visualisation of Internet Usage”.