# Web Usage Mining of Organisational Web Sites

**Craig P. Oosthuizen**, Janet Wesson & Charmain Cilliers
Department of Computer Science and Information Systems
PO Box 77000, Nelson Mandela Metropolitan University, Port Elizabeth, 6031
Tel: +27 (0)41 504 2323, Fax: +27 (0)41 504 2831
Email: {Craig.Oosthuizen, Janet.Wesson, Charmain.Cilliers}@nmmu.ac.za

Topic: Network Engineering – Modelling and simulation

ABSTRACT – **Web usage mining (WUM) can be used to determine if the information architecture of a web site is structured correctly. Existing WUM tools however, do not indicate which data mining algorithms are being used or provide effective graphical visualisations of the results obtained. The goal of this paper is to discuss the development of a prototype that allows the user to effectively visualise web usage data in order to evaluate the information architecture of an organisational web site. The WUM algorithms that will be used to analyse the web usage data are association rules, sequence analysis and cluster analysis.**

KEYWORDS – **Web usage mining, web site structure, web usage paths, web usage visualisation.**

## I. Introduction

Organisational web sites provide a critical role in making information available to specific user groups. Knowing how well the information architecture is structured is necessary to determine whether the web site needs to be improved. Web usage mining allows large amounts of web usage data, collected by the web server, to be analysed and visualised.

An organisational web site can be described as a site which has a high level of content [12]. The site serves only to provide the information and not any form of promotional content. Examples of such sites are universities, government web sites and corporate intranets. The NMMU Computer Science and Information Systems (CS&IS) departmental web site is an example of such a web site.

This paper discusses the goals of web usage mining and the necessary steps involved in developing an effective web usage mining system. A brief description of web usage mining is given as well as the data that is required for this analysis. The data mining algorithms that will be used for such analysis are discussed and appropriate visualisation techniques for each one proposed. Based on these visualisation techniques, the design of an interactive graphical user interface is presented.

### A. Web Site Structure

Although there are many tools available to aid Web Usage Mining (WUM), this form of web site analysis has shown limited success. This has been attributed to the fact that the need to understand the web site's content and structure is often overlooked [4]. The structure of a web site is required for identifying potentially interesting navigation patterns as navigation patterns can be mapped onto the structure of the web site. This enables the evaluator to interpret each navigation pattern in terms of the specific web site content.

The structure of a web site is created by the hyperlinks between various *pageviews*. These hyperlinks connect the individual web pages and provide a means for the user to navigate through the site. The aim here is to provide an easy to use path for the visitors to follow. This path should allow the user to access the content they are looking for in the fewest possible mouse clicks. The 'three-click' rule states that any page within a site should be no more than three clicks from the homepage [10]. Even although this rule is not official, it should serve as a guideline and assist in developing a successful navigation structure for a web site. The structure of the site and preprocessing the content are interrelated tasks [4]. Web sites are built around basic structural themes which govern the navigational interface of the web site.

### B. Log File Structure

During a user's navigation session, all activity on the web site is recorded in a log file by the web server. A web server can record user accesses in one of two log formats. The first is the common log format which records the host name and the version of the user's web browser. The second is the extended log format, which was introduced partly to support the collection of data.

The log file does not differentiate between various sessions from different users. It is simply a text file which captures each access by a user on a particular day. The fields that have been identified as necessary for the analysis of web usage patterns are shown in Table 1.

| Attribute | Description |
|---|---|
| Date | The date on which the activity occurred. |
| Time | The time the activity occurred. |
| IP Address | The IP address of the client that accessed the server. |
| Username | The name of the user who accessed the server |
| URL | The resource accessed: for example, an HTML page, a CGI program, or a script. |
| User Agent | The browser used on the client. |

**Table 1. Description of web log file attributes.**

## II. Web Usage Mining

The purpose of WUM is to reveal the knowledge hidden in the log files of a web server [7]. As shown in Figure 1, WUM can be broken down into three main phases, namely preprocessing, pattern discovery and pattern analysis. In the first phase, log files are preprocessed in order to retain only the appropriate information. In the second phase, various methods are used to identify interesting patterns. These patterns are then stored so that they can be analysed in the third phase of WUM, to determine which patterns are interesting. This section describes each phase in more detail.

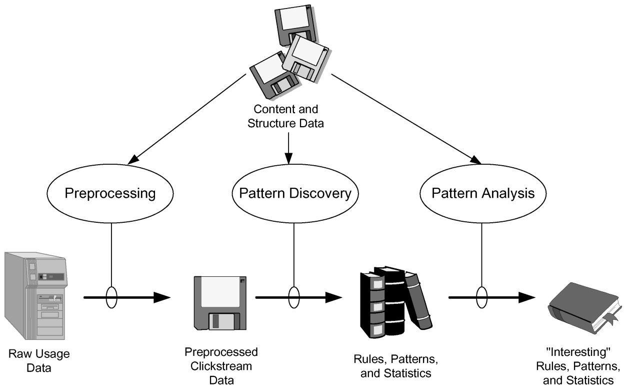

Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

**Figure 1. Web Usage Mining Process [4].**

### A. *Preprocessing*

Preprocessing involves preparing the web log for analysis. This also means identifying different sessions, users, pageviews and clickstreams [7].

Part of the preprocessing phase includes cleaning the server log to eliminate all of the irrelevant items [4]. This can be done by checking the suffix of the URL name and deleting the entries which are of no assistance to the analysis. The log file also needs to be parsed into data fields. Pageview identification determines which page file requests form part of the same pageview and what content was provided. As mentioned before, this step is highly dependent on knowledge of the web site structure and content.

During this phase, the data that is extracted from the log files needs to be stored so that it can be analysed. For this reason a data warehouse is required.

Essentially, the preprocessing phase is performed in order to convert the raw, clickstream web usage data into data that will be used as input to the pattern discovery phase as is shown in Figure 1.

### B. *Pattern Discovery*

The phase of pattern discovery relies on various statistical methods and data mining algorithms to detect interesting patterns [1]. Quantitative statistical methods are the easiest to apply and allow one to determine values such as frequency of visits to a page, average length of a path through a site and the average view time of a page, but no indication of the probabilities of certain patterns. Some of the more useful and appropriate data mining algorithms used for pattern discovery are *sequential patterns, association rules and cluster analysis* (Section III).

### C. *Pattern Analysis*

The last phase in the web usage mining process is pattern analysis. This process involves the user evaluating each of the patterns identified in the pattern discovery phase and deriving conclusions from them. The miner is generally concerned in finding patterns that provide useful information regarding the users' navigation (Section IV).

## III. WEB USAGE MINING ALGORITHMS

In order to extract any usage patterns from the log file, various data mining techniques need to be applied to the data. Criteria can be used to filter the data put to the application of the data mining technique. For example, typical criteria for WUM could be user group, time period or URL.

Data mining techniques are divided into two categories, namely supervised or guided, and unsupervised or unguided techniques [2]. In unsupervised techniques, there is no particular reason or goal for creating the model. However, supervised techniques are generally used for prediction. Since this paper is not focused on predicting user navigation, only unsupervised techniques will be used. Unsupervised techniques include association rules, sequence analysis and cluster analysis.

### A. *Association Rules*

The purpose of association rule mining is to discover relationships between items found in a database of transactions [5].

In terms of WUM, a transaction is described as a collection of time-ordered web page accesses or items, with an item being a single page access. An association would then imply that certain web pages within a web site are frequently accessed within the same user session. Association rules could also be used to discover relationships between various usage paths, which indicate that certain paths are followed frequently within the same user session.

Association rule mining is a two step process. The first step is to find all frequent itemsets (sets of web page accesses) which occur at least as frequently as a pre-determined minimum support count (minimum frequency). The second step is to generate strong association rules from the frequent itemsets which satisfy both the minimum support and minimum confidence [8], where confidence is the percentage of items which satisfy all the conditions in the association rule.

The output that is produced by an association rule would be of the form `XY (C, S)`; where `X` and `Y` are the items that are associated, `C` is the level of confidence and `S` is the level of support. In the case of web usage mining, `X` and `Y` would be the URL for the specific web pages.

For example, consider the output `/DegreesAndCourses.asp,/UndergradSubjects.asp (85,60)`. This can be interpreted as 85% (confidence) of navigation sessions that contain "degrees and course list" also contain "undergrad subject list". Sixty percent (support) of all the navigation sessions contain both these items.

### B. *Sequence Analysis*

Sequence analysis is the process of determining the longest time ordered paths that satisfy a user specified minimum frequency. In the case of web usage mining, sequences are the web usage paths. Since the log files represent a user's interaction on a web site, the goal is to discover patterns in the form of sequences [3].

A user supports a sequence *s* if *s* is contained in a session for this user. The problem with mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support. Sequences satisfying the minimum support are called *large sequences*.

The output that is produced by sequence analysis would be of the form `(AB…C) X%`, where `A`, `B` and `C` are the URL's of the web pages that form the sequence, and `X` is the frequency with which this sequence occurs.

For example, consider the output `(/HomePage.asp, /SubjectList.asp, SubjectDetails.asp), 35%`. This can be interpreted as 35% of the users followed the path "Home Page", "Subject List", then "Subject Details" during their navigation sessions.

### C. *Cluster Analysis*

Clustering is a discovery process that groups data into sets such that the similarity of the items within a group is maximised and similarity between items of different groups is minimised [9].

Clustering makes it possible to discover dense and sparse regions and, therefore, overall distribution patterns among data. Clustering of users tends to establish groups of users exhibiting similar browsing patterns [11]. It is useful to use cluster analysis when there are many objects with no natural groupings. Once clusters have been detected, other methods can then be used to analyse them and interpret what they mean.

The output that is produced by cluster analysis would be of the form `(ABC) X`, where `A`, `B` and `C` are the URL's of the web pages within a particular cluster and X is the frequency with which it occurs.

For example, consider the output `(/HomePage.asp, /StaffList.asp, /StudentDetails.asp) 90%`. This can be interpreted as "Home Page", "Staff List" and "Student Details" web pages form a cluster based on the specified criteria and this cluster occurs in 90% of the user sessions.

## IV. Visualising Web Usage Mining Results

Information visualisation is the process of creating visual interfaces to help users understand and navigate through complex information spaces [6]. The challenge in information visualisation is how to create a visual metaphor which presents the information in a meaningful way.

In order to select an appropriate visualisation technique, the data type needs to be taken into consideration. The data that is visualised in WUM is web usage data, which consists of sets of URLs and web usage paths. For each of these, the appropriate frequencies also need to be displayed.

Another factor that has to be considered is the importance of showing the structure of the web site being analysed, and not only the usage data. A problem with visualising web site structures is that they are not necessarily hierarchical in nature. Because of the many hyperlinks that may exist which link web pages with each other, the structure may represent a connected graph rather than a tree. One way in which this problem is addressed is by looking at the usage data when drawing the structure diagram. For each web page within the web site, only one link leading into it is displayed. This is the link that is followed with the highest frequency. By doing this, all the web pages are displayed but the connected graph is reduced to a hierarchical structure.

### A. *Visualising Association*

Since the purpose of association rules is to show that two or more web pages or web paths are accessed together frequently within a user session, the visualisation technique needs to be able to show all the associated web pages at once.

A disk tree or radial tree can be used to represent hierarchical information as shown in Figure 2. The primary node or home page is located in the center of the graph. Each successive descendant falls on concentric rings spanning out from the center. On each of these rings, the nodes are allocated space according to how many leaf nodes fall under it [13]. That is, nodes with a large number of descendants are allocated more space than those with fewer descendants. The advantage of the disk tree is that the area in which the tree is displayed is used more efficiently than a hierarchical tree. This allows a larger tree with more levels to be displayed in a smaller space.

In order to show the association between various page accesses or usage paths, visual cues can be used. For instance, pages or paths which are associated can be coloured similarly.

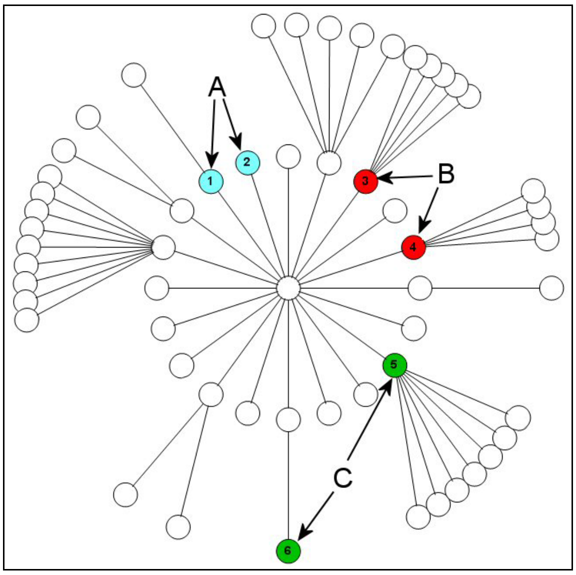


**Figure 2. Visualising Associations using a Radial Tree**

In Figure 2, three associations are shown and are labelled A, B and C. Each association is coloured differently. These indicate that the web pages within each association are accessed frequently together.

### *B. Visualising Sequences*

When visualising user navigation sequences it is also essential to show the structure of the web site. Radial trees were therefore also selected as an appropriate visualisation technique. Navigation sequences, however, not only indicate which pages were accessed but they also include the order in which they were visited. This means that the start and end points need to be represented using directed edges.

Navigation sequences can be visualised using radial trees by colouring all the pages visited in a specific sequence using the same colour and then using directed arrows to show the order in which they were visited. Figure 3 indicates how a radial tree can be used to visualise user navigation sequences.

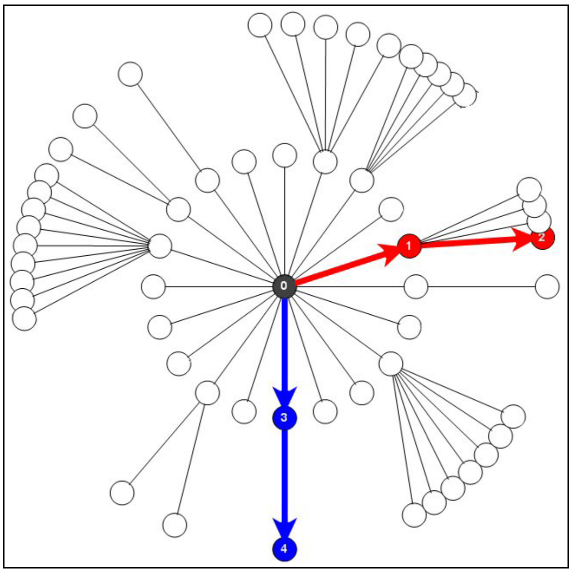


**Figure 3. Visualising Sequences using Radial Trees**

Two separate sequences are shown in Figure 3, the first showing a path from 0 to 1 to 2, and the second showing a path from 0 to 3 to 4. These two sequences represent the most popular paths accessed by the users.

### *C. Visualising Clusters*

Visualising clusters can be achieved using the same technique as visualising associated web pages. The web pages that form part of a specific cluster are then coloured similarly to show their grouping.

Figure 4 shows how clusters can be visualised using radial trees. The six highlighted nodes, 0 through 5, in Figure 4 form part of a cluster. The interpretation of a specific cluster is determined by the criteria selected when performing the clustering.

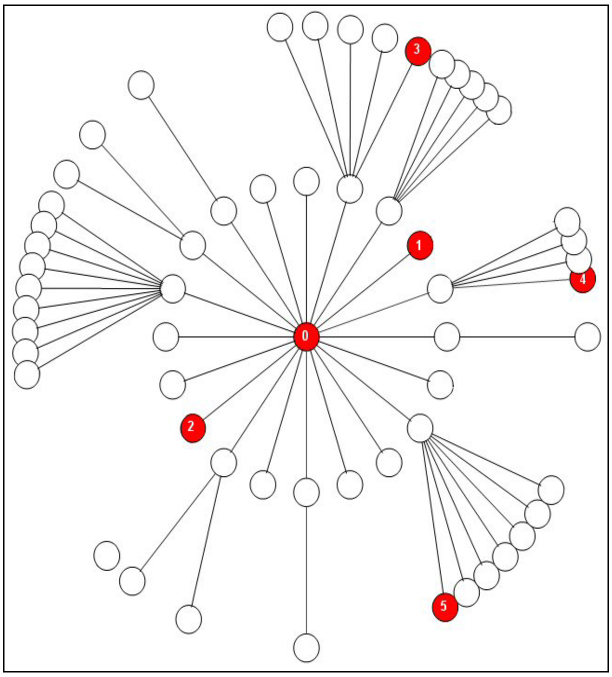
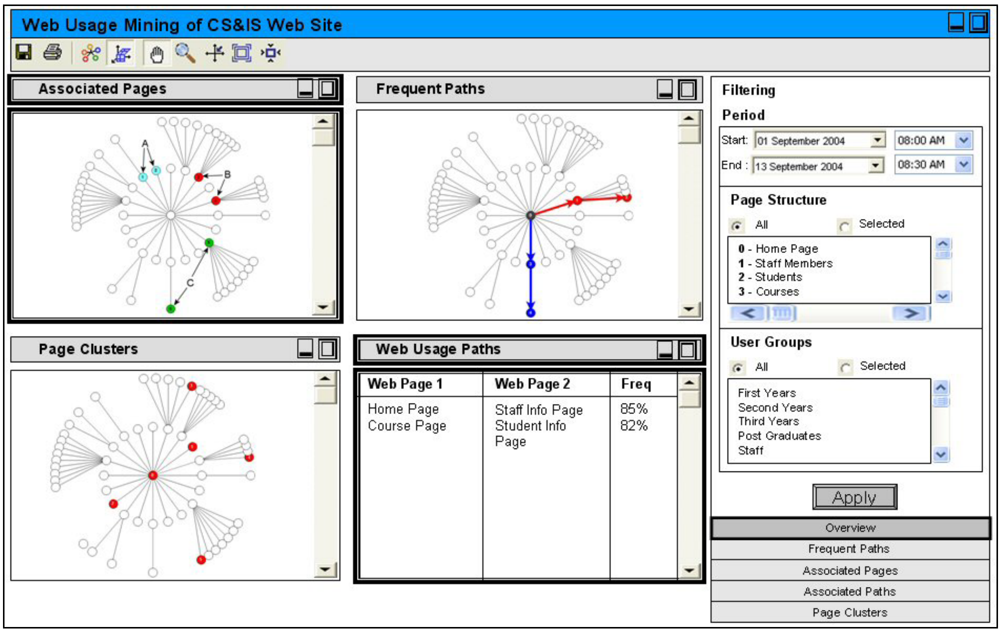


**Figure 4. Visualising Clusters using Radial Trees**



Period

Start: 01 September 2004 08:00 AM

End : 13 September 2004 08:30 AM

Page Structure

● All ○ Selected

**0** - Home Page
**1** - Staff Members
**2** - Students
**3** - Courses

User Groups

● All ○ Selected

First Years
Second Years
Third Years
Post Graduates
Staff

| Web Page 1 | Web Page 2 | Freq |
|---|---|---|
| Home Page | Staff Info Page | 85% |
| Course Page | Student Info Page | 82% |

**Figure 5. UI design of WUM prototype (Multiple Views)**

## V. User Interface (UI) Design

An iterative design approach was adopted in the UI design of the prototype. A conceptual model extraction focused on producing low-level prototype of the user interface to be evaluated by the users. This iterative process enabled the low-level prototype to be improved using the feedback from the users.

The UI of the system was designed to enable the user to view the web usage patterns of users over a specified period. The system interface illustrated in Figure 6 shows the visualisation technique that displays the structure of the web site being analysed as well as the results of one of the three web usage mining algorithms. Figure 6 also shows the three main views of the interface. The placement of these remains constant during the user interaction. The three views are the graphical view on the left which shows the visualisation of the results, the textual view at the bottom which provides the textual results of the algorithms and the filtering view on the right which allows the user to select what data is analysed by the WUM algorithms. The graphical view displayed is dependent on which WUM algorithm is selected. For example, the currently selected view in Figure 6 shows the most frequent paths followed by the users.

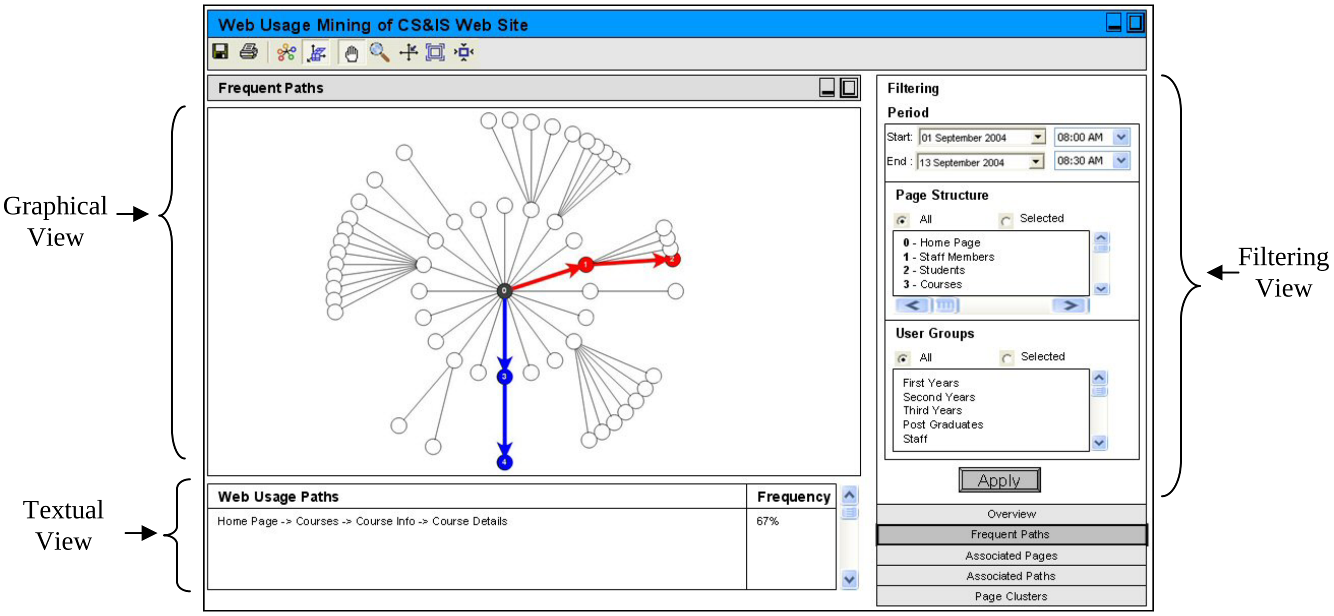
There are five buttons below the filtering menu which allow the user to switch between any of the web usage mining algorithms. There is also the option of viewing all three algorithms results on the same screen for comparison as is illustrated in Figure 5.

The user can also filter the data according to specific criteria using the filtering menu. The 'Period' option allows the user to specify the date and time range to be included in the analysis. The 'Page Selection' option makes it possible for the user to specify how much of the web site is displayed in the graphical view. This is useful if the user only wants to perform analysis on a section of the web sites.

At the bottom of the filtering menu is an 'Apply' button which applies all the desired filtering to the data and refreshes the graphical and textual views. In this view, the graphical and textual views are coordinated using multiple view techniques. Any of the three graphical views can be selected by clicking on the desired view. The selected view is then displayed as is shown in Figure 6.

## VI. Conclusions & Future Work

Web usage mining is required to allow web developers to analyse the information architecture of organisational web sites. Existing tools fail to illustrate the algorithms being used and effectively visualise their results. The data required for WUM was identified as date, time, IP address, URL, username and user agent. WUM algorithms were selected based on the type of the specific data collected by the web server. The algorithms selected were association rules, sequence analysis and cluster analysis. Appropriate visualisations techniques for each algorithm were proposed. A radial tree was selected as the most appropriate visualisation technique for all three algorithms. An iterative design approach was used to develop a WUM prototype. This process resulted in the design of an interactive user interface which can be used to visually represent the web usage of an organisational web site. The effectiveness of this interface still remains, however, to be evaluated.



Period

Start: 01 September 2004 08:00 AM

End : 13 September 2004 08:30 AM

Page Structure

● All ○ Selected

0 - Home Page
1 - Staff Members
2 - Students
3 - Courses

User Groups

● All ○ Selected

First Years
Second Years
Third Years
Post Graduates
Staff

| Web Usage Paths | Frequency |
|---|---|
| Home Page -> Courses -> Course Info -> Course Details | 67% |

**Figure 6. UI design of WUM prototype (Frequent Paths)**

## ACKNOWLEDGMENT

We would like to thank the Telkom Centre of Excellence programme and the Department of Computer Science and Information Systems at the Nelson Mandela Metropolitan University South Campus for making this research possible.

## REFERENCES

[1] BECKER, K. and VANZIN, M. (2003): Discovering interesting Usage Patterns in Web-based Learning Environments. *Utility, Usability and Complexity of e-Information Systems*:57-73.8-9 December 2003.

[2] BERSON, A., SMITH, S. and THEARLING, K. (2004): An Overview of Data Mining Techniques. In *Building Data Mining Application for CRM*.

[3] BÜCHNER, A.G., BAUMGARTEN, M., ANAND, S.S., MULVENNA, M.D. and HUGHES, J.G. (1999): Navigation Pattern Discovery from Internet Data. *Proc. WEBKDD '99*, San Diego, CA.August 1999.

[4] COOLEY, R. (2003): The use of web structure and content to identify subjectively interesting web usage patterns. *ACM Transactions on Internet Technology(TOIT)* **3**(2):93-116.May 2003.

[5] COOLEY, R., MOBASHER, B. and SRIVASTAVA, J. (1999): Data Preparation for Mining World Wide Web Browsing Patterns. University of Minnesota. Minneapolis, USA.

## BIOGRAPHY

Craig Oosthuizen received his B.Sc Hons degree in 2003 from the University of Port Elizabeth. He is presently doing his M.Sc in Computer Science at the Nelson Mandela Metropolitan University. His current field of research involves using data mining algorithms to visualise web usage navigation.

[6] EICK, S.G. (2001): Visualizing online activity. *Communications of the ACM* **44**(8):45-50.2001.

[7] EIRINAKI, M. and VAZIARGIANNIS, M. (2003): Web mining for web personalization. *Acm Transactions on Internet Technology(TOIT)* **3**(1):1-27.February 2003.

[8] HAN, J. and KAMBER, M. (2001): *Data Mining: Concepts and Techniques*. San Diego, CA, Morgan Kaufmann Publishers.

[9] HAN, S., KARYPIS, G., KUMAR, V. and MOBASHER, B. (1997): Clustering Based on Association Rule Hypergraphs. University of Minnesota. Minneapolis, MN.

[10] MIGHTYMEDIA (2004): Website Design. http://www.website-design-101.co.uk

[11] SRIVASTAVA, J., COOLEY, R., DESHPANDE, M. and TAN, P.-N. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations Newsletter* **Vol I**(2):12-23.January.

[12] UNIVERSITY OF SASKATCHEWAN (1999): Web Design for Instruction. **2003**.

[13] USABILITY FIRST (2002): Usability Glossary. http://www.usabilityfirst.com/glossary/main.cgi?function=display_term&term_id=818