WEB USAGE MINING OF ORGANISATIONAL WEB SITES

Craig Peter Oosthuizen

Supervisor: Prof J. Wesson
Co-Supervisor: Dr C. Cilliers

December 2005

Submitted in partial fulfilment of the requirements for the degree of Magister Scientiae in Computer Science & Information Systems at the Nelson Mandela Metropolitan University
Acknowledgements

I would like to thank my supervisors Prof Janet Wesson and Dr Charmain Cilliers for their valuable advice, guidance and encouragement during the course of this dissertation. The many hours that were spent reading through various versions of this document are greatly appreciated. I would like to thank my family for their patience and support during this time.

I would also like to thank the Department of Computer Science and Information Systems for enabling me to conduct this research. I would like to express my gratitude to Dr Thomas Hilmer for his valuable input during the design and evaluation phases. Lastly I would like to thank the Telkom Centre of Excellence programme for providing the necessary finance and technology for this research.
# Table of Contents

Summary.................................................................................................................................................. vi
List of Figures........................................................................................................................................... vii
List of Tables........................................................................................................................................... ix

## Chapter 1: Introduction......................................................................................................................... 1

1.1 Background......................................................................................................................................... 1
1.2 Research Relevance............................................................................................................................ 3
1.3 Problem Statement............................................................................................................................... 3
1.4 Research Questions............................................................................................................................. 4
1.5 Scope and Constraints.......................................................................................................................... 5
1.6 Structure of Dissertation..................................................................................................................... 5

## Chapter 2: Organisational Web Sites................................................................................................. 7

2.1 Introduction........................................................................................................................................ 7
2.2 Organisational Web Sites................................................................................................................... 7
2.2.1 Web Site Structure......................................................................................................................... 8
2.2.2 CS&IS Web Site............................................................................................................................. 9
2.2.3 Log File Structure.......................................................................................................................... 12
2.3 Conclusion....................................................................................................................................... 16

## Chapter 3: Web Usage Mining............................................................................................................ 17

3.1 Introduction....................................................................................................................................... 17
3.2 WUM Phases..................................................................................................................................... 17
3.2.1 Preprocessing................................................................................................................................. 18
3.2.1.1 Data Warehousing.................................................................................................................... 19
3.2.1.2 Data Architecture....................................................................................................................... 20
3.2.1.3 Data Modelling.......................................................................................................................... 21
3.2.2 Pattern Discovery.......................................................................................................................... 22
3.2.2.1 Sequence Analysis....................................................................................................................... 23
3.2.2.2 Association Rules......................................................................................................................... 25
3.2.2.3 Classification............................................................................................................................. 27
3.2.2.4 Cluster Analysis......................................................................................................................... 30
3.2.3 Pattern Analysis............................................................................................................................. 31
Chapter 4: Related Systems.................................................................36
4.1 Introduction..................................................................................36
4.2 123LogAnalyzer...........................................................................36
  4.2.1 Data Storage...........................................................................37
  4.2.2 WUM Algorithms.................................................................37
  4.2.3 Visualisation of Results.........................................................38
4.3 WebLog Expert............................................................................39
  4.3.1 Data Storage..........................................................................40
  4.3.2 WUM Algorithms.................................................................40
  4.3.3 Visualisation of Results.........................................................41
4.4 Absolute Log Analyzer...............................................................42
  4.4.1 Data Storage..........................................................................42
  4.4.2 WUM Algorithms.................................................................43
  4.4.3 Visualisation of Results.........................................................43
4.5 Comparison..................................................................................45
4.6 Conclusion..................................................................................45

Chapter 5: Proposed Model.................................................................46
5.1 Introduction..................................................................................46
5.2 Architecture of the Model............................................................46
5.3 Data Warehouse Design..............................................................48
5.4 Web Mining Algorithm Design...................................................49
  5.4.1 Association Rules.................................................................49
  5.4.2 Sequence Analysis...............................................................50
  5.4.3 Cluster Analysis.................................................................50
5.5 Visualisation Design....................................................................51
  5.5.1 Visualising Association Rules...............................................51
  5.5.2 Visualising Sequence Analysis..............................................52
  5.5.3 Visualising Cluster Analysis.................................................53
5.6 User Interface Design.................................................................54
5.7 Conclusion..................................................................................56
Chapter 6: Implementation of Prototype

6.1 Introduction

6.2 Implementation Tools

6.3 Architecture

6.4 Data Warehouse

6.5 WUM Algorithms

6.5.1 Association Rules

6.5.2 Sequence Analysis

6.6 Visualisation

6.6.1 User Interface

6.6.2 Static Structure of web site

6.6.3 Association Rules

6.6.4 Sequence Analysis

6.7 Data Manipulation

6.8 Conclusion

Chapter 7: Evaluation

7.1 Introduction

7.2 Evaluation Method

7.2.1 Evaluation Procedure

7.2.2 Evaluation Metrics

7.2.2.1 Effectiveness

7.2.2.2 Satisfaction

7.2.2.3 Usefulness

7.2.3 Evaluation Instruments

7.2.3.1 Task List

7.2.3.2 User Satisfaction Questionnaire

7.3 Results and Analysis

7.3.1 Demographics

7.3.2 Effectiveness Results

7.3.3 User Satisfaction Results

7.3.4 Usefulness

7.4 Conclusion

Chapter 8: Conclusions & Recommendations

8.1 Introduction
Summary

Web Usage Mining (WUM) can be used to determine whether the information architecture of a web site is structured correctly. Existing WUM tools however, do not indicate which web usage mining algorithms are used or provide effective graphical visualisations of the results obtained.

WUM techniques can be used to determine typical navigation patterns of the users of organisational web sites. An organisational web site can be described as a site which has a high level of content. The Computer Science & Information Systems (CS&IS) web site at the Nelson Mandela Metropolitan University (NMMU) is an example of such a web site. The process of combining WUM and information visualisation techniques in order to discover useful information about web usage patterns is called visual web mining.

The goal of this research is to discuss the development of a WUM model and a prototype, called WebPatterns, which allows the user to effectively visualise web usage patterns of an organisational web site. This will facilitate determining whether the information architecture of the CS&IS web site is structured correctly.

The WUM algorithms used in WebPatterns are association rule mining and sequence analysis. The purpose of association rule mining is to discover relationships between different web pages within a web site. Sequence analysis is used to determine the longest time ordered paths that satisfy a user specified minimum frequency. A radial tree layout is used in WebPatterns to visualise the static structure of the organisational web site. The structure of the web site is laid out radially, with the home page in the middle and other pages positioned in circles at various levels around it. Colour and other visual cues are used to show the results of the WUM algorithms.

User testing was used to determine the effectiveness and usefulness of WebPatterns for visualising web usage patterns. The results of the user testing clearly show that the participants were highly satisfied with the visual design and information provided by WebPatterns. All the participants also indicated that they would like to use WebPatterns in the future.

Analysis of the web usage patterns presented by WebPatterns was used to determine that the information architecture of the CS&IS web site can be restructured to better facilitate information retrieval. Changes to the CS&IS web site web were suggested, included placing embedded hyperlinks on the home page to the frequently accessed sections of the web site.

Keywords: visual web mining, web usage mining, information visualisation, web usage patterns, radial tree layout, user testing
List of Figures

Figure 1.1: Research structure........................................................................................................ 6
Figure 2.1: Screenshot of the CS&IS web site at NMMU.............................................................. 10
Figure 2.2: Web site structure of CS&IS web site.......................................................................... 11
Figure 2.3: Log file extract............................................................................................................ 13
Figure 3.1: Web usage mining processes....................................................................................... 18
Figure 3.2: Example navigation path represented over the CS&IS web site hierarchy.............. 33
Figure 3.3: Disk Tree technique used to represent usage of a web site..................................... 34
Figure 4.1: 123LogAnalyzer options............................................................................................ 37
Figure 4.2: 123LogAnalyzer generated report.............................................................................. 38
Figure 4.3: WebLog Expert options.............................................................................................. 40
Figure 4.4: WebLog Expert report............................................................................................... 41
Figure 4.5: Absolute Log Analyzer options.................................................................................. 43
Figure 4.6: Absolute Log Analyzer report.................................................................................... 44
Figure 5.1: Architecture for the proposed model........................................................................ 47
Figure 5.2: Snowflake schema of data warehouse....................................................................... 48
Figure 5.3: Visualising association rules...................................................................................... 52
Figure 5.4: Visualising sequences analysis................................................................................... 53
Figure 5.5: Visualising cluster analysis......................................................................................... 54
Figure 5.6: UI Design of WUM prototype (Frequent paths)....................................................... 55
Figure 6.1: WebPatterns architecture.......................................................................................... 59
Figure 6.2: UI of WebPatterns showing association rules visualisation.................................... 65
Figure 6.3: Static structure of CS&IS web site.............................................................................. 66
Figure 6.4: Multicoloured nodes indicating multiple associations............................................. 68
Figure 6.5: Multiple paths to a single web page......................................................................... 69
Figure 6.6: UI of WebPatterns showing sequence analysis visualisation of all paths to all web pages........................................................................................................... 70
Figure 6.7: UI of WebPatterns showing sequence analysis visualisation of all paths to the selected page........................................................................................................... 71
Figure 6.8: Period selection filter................................................................................................ 72
Figure 6.9: URL list filter.............................................................................................................. 72
Figure 7.1: Mean and median percentage of task completion..................................................... 82
Figure 7.2: User satisfaction results............................................................................................. 83
Figure 8.1: Proposed model for the mining and visualisation of web usage data.................... 89
Figure 8.2: Coordinated views of WebPatterns showing associated pages........................... 91
Figure 8.3: Association rules analysis results................................................................. 94
Figure 8.4: Sequence analysis results............................................................................ 95
List of Tables

Table 1.1: Research design including the research questions and associated methods........4
Table 2.1: Log file field descriptions................................................................................13
Table 2.2: Web log file specification using BNF notation..............................................15
Table 3.1: Comparison of sequence analysis algorithms..............................................25
Table 3.2: Comparison of association rule algorithms...................................................27
Table 3.3: Comparison of classification algorithms......................................................29
Table 3.4: Comparison of clustering algorithms.........................................................31
Table 4.1: Analysis report from 123LogAnalyzer.........................................................39
Table 4.2: Analysis report from WebLog Expert..........................................................42
Table 4.3: Analysis report from Absolute Log Analyzer...............................................44
Table 6.1: Extract from Bitmap transaction table.........................................................62
Table 7.1: Summary of tasks........................................................................................78
Table 7.2: Demographic profile of participants............................................................80
Table 7.3: Percentage of task completion.....................................................................81
Table 7.4: Satisfaction results for quality of visualisation and information.................84
Table 8.1: Research achievements and corresponding research questions................88
Chapter 1 – Introduction

1.1. Background

Many users access web sites in order to obtain information. Three factors influence the way users perceive and assess a web site, namely content, web page design and overall site design (Spiliopoulou 2000). The first factor involves the content or information offered by the site. The other two factors are related to the way in which the site makes its content available and understandable to its users. There is a difference between the design of individual web pages and overall site design. A web site is not simply a collection of pages, but a network of related pages, which users will not explore unless they find its structure intuitive.

It is not always clear whether the information contained in web sites is structured efficiently or effectively. It is also not obvious as to how these measures can be determined. When visiting a web site, the only information left behind by users is the trace through the pages they have accessed, also known as a clickstream. All activities that take place on the site are recorded in a file called the web server log (Masseglia, Poncelet and Teisseire 1999). From this large amount of raw data, the designer of the site must determine what the users wanted in the site, what they liked and what disturbed or distracted them.

Data mining is a technique that can be used to analyse large amounts of data in order to discover patterns. Before the information can effectively be mined, it has to be
consolidated. This typically requires the construction of a data warehouse (Impact21 Group 2005). A data warehouse involves combining information from various locations into a central database to allow more precise analysis of patterns.

Web Usage Mining (WUM) is the application of data mining techniques to web clickstream data in order to extract usage patterns (Joshi, Joshi, Yesha and Krishnapuram 1999). The extraction of information about users’ accesses is mainly obtained from analysis of navigation behaviour as observed in web server logs, where all accesses to web pages are recorded (Cooley, Mobasher and Srivastava 1999a). Typical information for each access includes page requested (URL), IP address from which the request originated, time and date of the request, status and referrer field (Géry and Haddad 2003). WUM normally includes three phases: a pre-processing phase wherein the data is extracted from the log files; a pattern discovery phase in which data mining algorithms are used to identify patterns; and a pattern analysis phase in which interesting patterns are filtered (Cooley, Tan and Srivastava 1999b). The output from the final phase (pattern analysis) consists of sequences of accesses with corresponding probabilities.

As more organisations make use of the Internet and the World Wide Web to convey information, the conventional approaches to web site evaluation need to be revised (Cooley et al. 1999a). The Computer Science & Information Systems (CS&IS) Department at the Nelson Mandela Metropolitan University (NMMU) is an example of an organisation which uses a large web site to convey information to current and prospective students. In order for the web site to be used as it was intended, it must be structured correctly.

Before a site structure can be improved, a method is needed to evaluate its current usage. Usability evaluation is the process of collecting data about the usability of a system by a specified group of users for a specific activity within a specified environment (Spiliopoulou 2000). In order to evaluate the usage of a web site, a group of users which represents the intended users would have to be selected. One would then have to identify certain activities that they are expected to perform and record their actions while performing these tasks. Although some companies can afford this kind of evaluation, it is impractical to carry this out each time the quality of the site is
evaluated. It would be ideal to evaluate a site based on the data that is automatically recorded by the web server. By using WUM techniques, it would be possible to extract information from these files to determine the quality of the site structure.

1.2 Research Relevance

The CS&IS Department at the NMMU has an extensive web site which provides users with access to staff information, course information, course lecture notes, exam results and research output. Many of the students rely on the site to access important information about their subjects. It has become a concern as to whether the site is being used, how it is being used and whether the site structure is optimal or if it needs to be reorganised. The only record of accesses to this web site is in large log files stored on the web server.

If the web designer notices a frequently followed route and identifies this route as the expected route, it implies that most users perceive the site structure as the designer intended. However, if a frequently used route is surprising to the designer, it implies that many users navigate differently from how the designer intended. These routes then need to be investigated so the designer can detect pages that are not correctly linked or designed.

1.3 Problem Statement

The primary objective of this research is to propose a model to analyse and view the web usage of an organisational web site. The purpose of this model is to identify firstly, any usage patterns that may exist and secondly, any potential problems with the site structure. This model should enable a web designer to view the web usage patterns of the web site in an interactive and graphical manner. The effectiveness of the proposed model will be demonstrated by developing a prototype, called WebPatterns, to analyse the structure of the CS&IS web site.
1.4 Research Questions

In order to achieve the objectives of this project, the following research questions were proposed:

1. What is an organisational web site?
2. What is web usage mining?
3. What are web usage patterns?
4. Are there any related systems?
5. How will the model be designed?
6. How will a prototype for the model be implemented?
7. How effective and useful is the prototype?
8. What conclusions can be made from this research?

The design for this research is shown in Table 1.1. This table identifies the research methods that will be used to obtain the answers to the proposed research questions. In many cases a literature study forms the foundation of the investigation to answer each of the research questions.

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Research Method</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is an organisational web site?</td>
<td>Data Analysis</td>
<td>Chap 2</td>
</tr>
<tr>
<td>1.1 What is the information architecture of the CS&amp;IS web site?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2 What is the structure of the server log files?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3 What data needs to be mined from the CS&amp;IS web site logs?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. What is web usage mining (WUM)?</td>
<td>Literature Review</td>
<td>Chap 3</td>
</tr>
<tr>
<td>2.1 What is the purpose of WUM?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2 What are the phases of WUM?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. What are web usage patterns?</td>
<td>Literature Review</td>
<td>Chap 3</td>
</tr>
<tr>
<td>3.1 What algorithms can be used to find web usage patterns?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2 What can be determined from these web usage patterns?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3 How can these patterns be visualised?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Are there any related WUM systems?</td>
<td>Extant Systems Analysis</td>
<td>Chap 4</td>
</tr>
<tr>
<td>4.1 What algorithms are used by these related systems?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 How are the patterns stored?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3 How are the patterns visualised?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4 What are the shortcomings of these systems?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. How will the model be designed?</td>
<td>Iterative Design</td>
<td>Chap 5</td>
</tr>
<tr>
<td>5.1 What algorithms will be used?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2 How will the user interface be designed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3 How will the data model be designed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.4 How will the information visualisation be designed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. How will a prototype for the model be implemented?</td>
<td>Implementation</td>
<td>Chap 6</td>
</tr>
<tr>
<td>6.1 What tools are available for implementing the prototype?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2 What implementation techniques will be used to implement the prototype?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. How will the prototype be evaluated?</td>
<td>Evaluation</td>
<td>Chap 7</td>
</tr>
<tr>
<td>7.1 How effective is this prototype?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. What conclusions can be made from this research?</td>
<td>Analysis of Results Conclusions</td>
<td>Chap 7 Chap 8</td>
</tr>
<tr>
<td>8.1 What are the results of the web usage mining?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.2 What recommendations can be made for the CS&amp;IS web site?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.3 How can the proposed model be applied to other web sites?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Research design including the research questions and associated methods
Chapter 1 - Introduction

1.5. Scope and Constraints

The scope of this research will be limited to proposing a WUM model and implementing a prototype based on this model to analyse the site structure of the CS&IS web site. Appropriate algorithms will be investigated and selected to identify web usage patterns from the server log files and the web site will then be evaluated according to these findings. Appropriate visualisation metaphors will be identified and used to display these usage patterns in a graphical format. Based on the findings, appropriate changes will be suggested to improve the information architecture of the CS&IS web site.

1.6. Structure of Dissertation

This dissertation consists of eight chapters. The first chapter provides the context of research, as well as the problem statement (Section 1.3) and relevance of the research (Section 1.2). This is followed by the scope and constraints of the research (Section 1.5). The research design and methods (Section 1.6) specify how the research questions that have been formulated will be answered in the chapters that follow.

The literature study is divided into three chapters (Chapter 2 – 4). Figure 1.1 shows how these three chapters contribute to the design and specification of the proposed model. Chapter 2 provides a general discussion of organisational web sites. A description of the CS&IS web site at the NMMU is then provided, as well as the structure of the web site and the log files.

Chapter 3 describes WUM, including the various phases that are involved and a discussion of the algorithms that can be used for the pattern discovery phase. An investigation into the various visualisation techniques used for WUM is also provided.

Chapter 4 provides a comparative analysis of three related WUM systems. This chapter documents their strengths and shortcomings in terms of data analysis and information visualisation. The comparison of the different systems aids the specification of the requirements for the WUM model and the prototype.

Chapter 5 documents the design and specification of the proposed WUM model including the design of the data warehouse, the user interface and the algorithms. Appropriate visualisation techniques are also discussed.
Chapter 6 documents the implementation of a prototype, called WebPatterns, for the model specified in Chapter 5. Chapter 7 provides an evaluation of WebPatterns by using it to analyse the web usage patterns of the CS&IS web site at the NMMU.

Chapter 8 concludes the dissertation by examining whether the objectives of the research have been achieved. The research conducted is reviewed and suggestions for future research are discussed. Suggested improvements to the information architecture of the CS&IS web site are also documented.
Chapter 2 – Organisational Web Sites

2.1. Introduction

The first question of Table 1.1 aims to define an organisational web site. The characteristics of organisational web sites are discussed in Section 2.2. Since the CS&IS web site is used as an example of an organisational web site, knowledge of its information architecture is required (Section 2.2.2). In order to effectively mine the log files of the CS&IS web site, the structure of the log file needs to be investigated (Section 2.2.1 and 2.2.3).

2.2. Organisational Web Sites

An organisational web site is classified as a site which has a high level of content (University of Saskatchewan 1999). Visitors to the site generally have extrinsic motivation for visiting. The site serves only to provide information and not any form of promotional content. Examples of such sites are universities, government web sites and corporate intranets. The NMMU CS&IS departmental web site is an example of such a web site and will be used as a case study in the evaluation of the proposed model (Chapter 6).
2.2.1. Web Site Structure

WUM has shown limited success despite the availability of many WUM tools (Cooley 2003). This has been attributed to the fact that the need to understand the web site’s content and structure is often overlooked. The structure of a web site is useful for identifying potentially interesting patterns, as observed navigation patterns can be mapped onto the structure of the web site (Chen, Sun, Zaiane and Goebel 2004). This enables the evaluator to interpret each observed navigation pattern in terms of the specific web site structure.

In order to provide a clear understanding of the structure of a web site and the log files, certain terms need to be defined, namely user, pageview, clickstream, user session and server session. A user is defined as an individual that accesses files from a web server using a web browser. A pageview consists of several items that are visually rendered in a web browser to construct a single web page. This view of a web site can consist of frames, text, graphics and scripts amongst others. A clickstream is a series of pageview requests made by a user during a navigation session. A user session is defined as the time-delimited set of pageviews across the entire Web (Mortazavi-Asl 1999). A server session is defined as the subset of the user session for a specific web server (Géry and Haddad 2003). A server session is the grouping together of consecutive pages requested by the same user. A user session is created when a new IP address is observed in the log file. Subsequent requests from the IP address are added to the session as long as the time between two consecutive requests does not exceed a pre-defined threshold. A recommendation is that thirty minutes of inactivity indicates the end of a session (Mortazavi-Asl 1999). If the session duration exceeds the threshold, the current session is closed and a new session is created. Since this research is focused on user activity on a single web site, the terms user session and server session may be used interchangeably as both relate to the activity of a user on a single web server, that is the CS&IS web site.

The structure of a web site like that of the CS&IS web site is created by 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 user to follow. This path should allow the users 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 home page (Mightymedia 2004). Even although this rule is not a recognised standard, it should serve as a guideline and assist in developing a successful navigation structure for a web site.

Determining the structure of the site and preprocessing the content are interrelated tasks (Cooley 2003). Web sites are built around basic structural themes which govern the navigational interface of the web site.

Three important structures can be used to build a web site (Mightymedia 2004):

- Sequences,
- Hierarchies, and
- Webs.

The simplest way to organise information is to place it in a sequence. In a sequence, the only links are those that support a linear navigation path. Sequences are ideal for training sites in which the user is expected to follow a specific series of web pages. A hierarchy is the best method for structuring complex bodies of information. Hierarchies are only practical with well-organised information. A web organisational structure poses very few limitations on the navigation pattern. This structure provides links to and from all web pages but often causes confusion. Due to the substantial connectivity, the expected navigation pattern is very hard for the user to predict.

The majority of web sites make use of a combination of all three structures in their design.

### 2.2.2. CS&IS Web Site

Figure 2.1 shows a screenshot of the home page for the CS&IS web site. On the left hand side are the main links that are available from the home page. These include ‘Degrees and Courses’, ‘Staff’, ‘Students’, ‘Research’ and ‘CoE’. The main navigation menu on the left hand side and the title bar at the top remain on the screen during the navigation of the site, and only the main frame changes with the relevant information.
Figure 2.1: Screenshot of the CS&IS web site at NMMU (28 November 2005)

Figure 2.2 shows the web site structure of the CS&IS web site at NMMU. This web site incorporates in its design all three structures discussed in Section 2.2.1. The links between the home page, the staff page and the staff member page are an example of a sequential structure. The home page is sub-divided into the five sections and is an example of a hierarchy. The embedded hyperlinks from the research supervision page to the projects and staff members’ pages are an example of a web-like structure. The web pages of the CS&IS web site are dynamic. This means that the information displayed in each pageview is obtained from an SQL Server Database by means of Active Server Pages (ASP) (Bennett, Wesson and Barnard 2000). Each pageview within the site also consists of a number of frames, usually one for the title, and one for the menu items and another for the main content. It is not uncommon for a pageview to consist of two or three frames.

In Figure 2.2, each pageview is represented by a block. Colours are used to group related blocks together. The bolded heading in each block represents the title of the web page. Below the title of each web page is an URL. Each URL contains the location of the relevant ASP page and some are also followed by a condition such as ‘DegreeID=?’. In this case the ‘?’ represents the particular course the user wishes to view.
Chapter 2 – Organisational Web Sites

Figure 2.2: Web site structure of CS&IS web site (28 November 2005)
Chapter 2 – Organisational Web Sites

The home page of the CS&IS web site is shown on the left-hand side in light yellow. To the right of this block there are five more blocks representing the five links available from the home page, namely ‘Degrees and Courses’, ‘Staff’, ‘Students’, ‘Research’ and ‘CoE’. Each of these is represented using a different colour, namely orange, light blue, dark green, purple and light green respectively. Moving towards the right-hand side are the subsequent web pages the user can follow in order to reach a target page. All web pages are coloured according to the block to which they belong. In the case where there is a block containing only a title and no URL and which is a different colour from the block to which it is linked, this indicates that there is an embedded hyperlink to a different block. For example, the light blue ‘Staff Member’ block connected to the orange ‘Lecturer’ block indicates the presence of a hyperlink from the ‘Lecturer’ page to the ‘Staff Member’ page.

2.2.3. Log File Structure

During a user session, all navigation 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 format used for the CS&IS web site is the extended log format.

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. Table 2.1 gives a list of the various fields available for an extended log file and a description of each field. The fields that have been identified as relevant to the analysis of web usage patterns are: date, time, c-ip, cs-username, cs-uri-stem and cs(user agent). These fields are highlighted in the table. The reason for these fields being used is that they provide required information regarding the identity of the user, the information being requested from the site, the time and date of the request as well as the browser being used by the client.
<table>
<thead>
<tr>
<th>Log file field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>The date on which the activity occurred.</td>
</tr>
<tr>
<td>Time</td>
<td>The time the activity occurred.</td>
</tr>
<tr>
<td>c-ip (Client IP Address)</td>
<td>The IP address of the client that accessed the server.</td>
</tr>
<tr>
<td>cs-username (User Name)</td>
<td>The name of the user who accessed the server.</td>
</tr>
<tr>
<td>s-ip (Server IP)</td>
<td>The IP address of the server on which the log entry was generated.</td>
</tr>
<tr>
<td>s-port (Server Port)</td>
<td>The port number to which the client is connected.</td>
</tr>
<tr>
<td>cs-method (Method)</td>
<td>The action the client was trying to perform.</td>
</tr>
<tr>
<td>cs-uri-stem (URI Stem)</td>
<td>The resource accessed, such as an HTML page.</td>
</tr>
<tr>
<td>cs-uri-query (URI Query)</td>
<td>The query, if any, the client was trying to perform; that is, one or more search strings for which the client was seeking a match.</td>
</tr>
<tr>
<td>sc-status (Protocol Status)</td>
<td>The HTTP status code returned to the client, indicating whether the file was successfully retrieved and if not what error message was returned.</td>
</tr>
<tr>
<td>cs(User-Agent)</td>
<td>The browser used on the client.</td>
</tr>
<tr>
<td>s-sitename</td>
<td>The Internet service that was running on the client computer.</td>
</tr>
<tr>
<td>s-computername</td>
<td>The name of the server on which the log entry was generated.</td>
</tr>
<tr>
<td>sc-win32-status</td>
<td>The status of the action, in terms used by Windows.</td>
</tr>
<tr>
<td>sc-bytes</td>
<td>The number of bytes sent by the server.</td>
</tr>
<tr>
<td>cs-bytes</td>
<td>The number of bytes received by the server.</td>
</tr>
<tr>
<td>time-taken</td>
<td>The length of time the action took.</td>
</tr>
<tr>
<td>cs-version</td>
<td>The protocol (HTTP, FTP) version used by the client. For HTTP this will be either HTTP 1.0 or HTTP 1.1.</td>
</tr>
<tr>
<td>cs(cookie)</td>
<td>The content of the cookie sent or received, if any.</td>
</tr>
<tr>
<td>cs(referer)</td>
<td>The site that directed the user to the current site.</td>
</tr>
</tbody>
</table>

Table 2.1: Log file field descriptions

An extract of a log file from the CS&IS website is shown in Figure 2.3. Examples of fields that were identified as being necessary for WUM are labelled. The top of the log file contains a header which consists of information on the software and version as well as the date, time and the fields that are captured. Below the header are consecutive line entries where each line entry relates to a single request on the web server.
Upon further inspection of the log files, it was noted that the username for the user initiating the request is not always recorded in the log file. This may have implications during the analysis of the log files.

**Log File Specification**

This section describes the log file specification using Backus-Naur (BNF) notation. An explanation of the symbols used in BNF notation is given in Marcotty and Ledgard (1986). It was necessary to formally describe the log file to enable a more seamless extraction of the data during the preprocessing phase (Section 3.3.1).

An extended log file contains a sequence of lines containing ASCII characters. Each line may contain a *directive* or an *entry* (Hallam-Baker and Behlendorf 1996). Lines beginning with the # character contain directives. The following directives are defined:

- **Software**: `<string>`
  - Identifies the software which generated the log.

- **Version**: `<integer>.<integer>`
  - The version of the extended log format used.

- **Date**: `<date> <time>`
  - The date and time on which the log file was recorded.

- **Fields**: `{<specifier>}`
  - Specifies the fields recorded in the log.

Log file entries consist of a sequence of fields relating to a single HTTP transaction. Each frame or image within a web page which is requested from the web server is recorded in the log file. Since each web page could have a number of frames and multiple images, the number of entries in each log file can be quite large. Fields are separated by a space. If a field is unused in a particular entry, a dash “-” indicates the omitted field. A directive records information about the logging process. The grammar for the extended log file format used by the CS&IS web server is given in Table 2.2.
Chapter 2 – Organisational Web Sites

Table 2.2: Web log file specification using BNF notation

Table 2.2 does not define `<char_except_whitespace>` and `<char>`. All of the standard ASCII characters are defined as members of `<char>`, and all characters but "white space" are defined as `<char_except_whitespace>`.

Some problems that are typically encountered with the data captured in a log file are the following (Srivastava, Cooley, Deshpande and Tan 2000):

- Single IP/Multiple Server Sessions: Internet service providers usually have a set of IP addresses through which users access the web. A single proxy may have several users accessing the same web site.

- Multiple IP address/Single Server Session: Some Internet service providers randomly assign each request from a user to one of several IP addresses. This means that a single user could have multiple IP addresses.

- Multiple IP address/Single User: A user may access a web site from different machines. This means a single user will have a different IP address which makes tracking repeated visits by the same user difficult.

- Multiple Agent/Single User: A user that uses more than one browser even on the same machine will appear as multiple users.

The problems mentioned above are extremely difficult to identify and it is also difficult to take them into consideration when analysing a log file. This is due to the fact that additional knowledge is required such as server and proxy details, cache information as well as how many users are using each machine and which browsers each user is using.
For the purposes of this research, it will be assumed that each unique IP address with the same agent represents a different user and multiple agents with the same IP address represent multiple users.

2.3. Conclusion

The purpose of this chapter was to define organisational web sites. The CS&IS web site is an example of an organisational web site that provides users with information. The structure of this site was determined to be a combination of sequences, hierarchies and webs (Section 2.2.1). An extended log file format is used to record all user accesses. In order to determine the structure of the log file, a formal specification of the log file format was provided. The fields from the log file identified as relevant for WUM are date, time, ip address, username, URL and user agent (Section 2.2.3). Due to the problems identified in differentiating between users of a web site, a combination of ip address, username and client agent will be used to uniquely identify a user of the CS&IS web site.

The next chapter discusses WUM as well as the various data mining algorithms which can be used for analysing log file data and the appropriate visualisation techniques which can be used to graphically represent the web usage patterns identified.
Chapter 3 – Web Usage Mining

3.1. Introduction

The aim of the second research question in Table 1.1 is to define web usage mining (WUM) and discuss each of the phases. WUM is the application of data mining algorithms to web clickstream data in order to extract web usage patterns. WUM of large web sites may require a data warehouse to store the log file data; WUM algorithms to determine the web usage patterns; and appropriate visualisation techniques to present the WUM results.

Several WUM algorithms are discussed and the most appropriate ones are selected based on the characteristics of the data available from the log files (Section 2.2.3). The results of WUM need to be visualised in order to assist with their interpretation and analysis. This chapter provides a discussion of various visualisation techniques and selects the most appropriate ones, based on the web usage patterns produced by the selected WUM algorithms.

3.2. WUM Phases

The purpose of WUM is to reveal the knowledge hidden in the log files of a web server (Eirinaki and Vaziargiannis 2003). There are two main goals in the application of this discovered knowledge (Mortazavi-Asl 1999):

- General access pattern tracking for understanding access patterns and trends.
- Customised usage tracking for adapting and personalising browsing experience for the user.
Chapter 3 – Web Usage Mining

This research is focused on the proposal of a WUM model for organisational web sites for the former of the above two goals (Section 1.3).

As shown in Figure 3.1, 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 relevant information (Section 3.2.1). In the second phase, various methods are used to identify interesting patterns from the retained information (Section 3.2.2). These patterns are then presented for analysis in the third phase of WUM (Section 3.2.3).

![Figure 3.1: Web usage mining processes (Cooley 2003)](image)

### 3.2.1. Preprocessing

Preprocessing involves preparing the server log for analysis. This also means identifying different user sessions, users, pageviews and clickstreams (Eirinaki and Vaziargiannis 2003).

Part of the preprocessing phase includes cleaning the server log to eliminate all of the irrelevant items (Cooley 2003). This can be done by checking the suffix of the URL name and deleting the entries which are of no assistance to the analysis, such as JPG and GIF, amongst others. The log file also needs to be parsed into the data fields identified in Table 2.1. Pageview identification determines which page file requests
form part of the same pageview and what content was provided. As mentioned in Section 2.2.1, this step is highly dependent on knowledge of the website structure and content.

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 shown in Figure 3.1.

During the preprocessing phase, the data that is extracted from the log files needs to be stored so that the WUM algorithms can be applied to it during the pattern discovery phase. For this reason a data warehouse is required. Section 3.2.1.1 defines the structure of the data warehouse while Section 3.2.1.2 describes the data architecture used and Section 3.2.1.3 discusses the modelling of the data to be stored.

3.2.1.1. Data Warehousing

A data warehouse is a multidimensional database designed for query and analysis (Oracle 2002). It usually contains historical data derived from transaction data but can include data from other sources. Data warehousing provides architectures and tools that allow data to be systematically organised and understood so that strategic decisions can be made (Han and Kamber 2001). In order to integrated historical information and support complex multidimensional queries, a data warehouse is required.

Data warehousing is described as the process of constructing and using a data warehouse which is subject-oriented, integrated, non-volatile and time-variant. High performance of integrated heterogeneous database systems is supported by data warehousing since the data is copied, preprocessed, integrated, annotated, summarised and restructured into one semantic data store.

Underneath the data warehouse is a maze of connections and transformations connecting the data warehouse with other systems (Thearling 2000). Because data in an organisation is often created and stored in functionally specific systems, the data may need to be replicated and moved between these systems and the data warehouse.

The following sections discuss the data warehousing architectures involved in storing the data that is required by the preprocessing phase of WUM.
3.2.1.2. Data Architectures

When creating a data warehouse, a relevant data architecture needs to be selected. This is dependent on the type of data the data warehouse will store. Three types of data architectures are (Thornthwaite 1998):

- Single layer data architecture,
- Two layer data architecture, and
- Three layer data architecture.

In the single layer data architecture, the data is stored only once. This minimises the data storage requirements since transactions are performed on the actual database in real-time; that is operational transactions and query processing occurs concurrently on the same database. This means, however, that at certain times parts of the database will be locked due to read-write transactions and will thus be unavailable, which could lead to time delays on other transactions.

The two layer approach provides a solution by utilising two copies of the same data. These are the real time data and the derived data. Operational transactions take place on the real time data while the derived data is used for performing queries. This approach addresses the issue of contention of data usage but introduces a new problem in that it often results in high levels of data duplication. Despite this, it is the most commonly used approach in data warehousing.

The three layer architecture splits the data into the real time data, the reconciled data and the derived data. Firstly, the data is reconciled from the real time data. Secondly, information that is required by the user is derived from the reconciled data. The reconciled data layer combines the data from various operational sources into a single, logical image of the data model. The derived data layer is summarised data which is useful for predefined queries.

The log files provided by the NMMU CS&IS web site are not an operational database. During the course of a specific day, entries are recorded in a log file as they occur. At the end of each day, the log file is saved. Since no duplicate data is required to
accommodate operational processing concurrent with WUM, a single layer data architecture will thus be implemented.

3.2.1.3. Data Modelling

Before a data warehouse can be designed and populated with data, the data warehouse needs to be modelled. The data warehouse design process consists of the following steps (Han and Kamber 2001):

1. Choose the process to model. In the case of this research, the web usage data collected by the web server of the NMMU CS&IS web site will be used.
2. Choose the grain of the process. The granularity or frequency of the data is usually determined by the time dimension. The lower the granularity, the more records there will be in the fact table. In this case, the granularity is time measured to the level of seconds (access time).
3. Choose the dimensions that will apply to the fact table record. These dimensions include date, time, URL accesses and client id. These dimensions provide filtering access to the aggregated data.
4. Choose the measures that will populate each fact table record. Measures are typically aggregated numeric values such as number of accesses and number of user sessions.

A data warehouse can be designed in one of three ways, namely using a star schema, snowflake schema or fact constellation schema (Han and Kamber 2001).

Star Schema

The most common data schema used is the star schema. In this case, there is a large central fact table containing the bulk of the data with no redundancy, and a set of smaller tables, one for each dimension (Han and Kamber 2001). The schema diagram represents a star shape with the dimension tables displayed in a radial pattern around the central fact table. Using the star schema enables relatively quick query execution but it requires more storage space due to redundancy of information.
**Fact Constellation Schema**

This type of schema makes use of multiple fact tables which allows the separation of detail from the aggregated values. A fact constellation can have its dimensions normalised, but it is not always required to do so. A disadvantage of this design is that fact tables share dimension tables making it a complicated design. Another disadvantage is that simple queries also need to be subdivided before they can be answered. An advantage of this schema is that it does not have sparse fact tables. Another advantage of a fact constellation schema is that it can model multiple, interrelated subjects (Han and Kamber 2001).

**Snowflake Schema**

The snowflake schema is a variation of the star schema, where some dimension tables are normalised in order to reduce redundancies (Levene and Loizou 2003). This further splits the data into additional tables. The main advantage of the snowflake schema is the reduced disk storage requirements due to reduced redundancies while the disadvantages include the need for more dimension tables and the joins on foreign keys between these dimensions. This requires more complex queries which can result in slower response time.

A snowflake schema was selected to model the data to be stored in the data warehouse. This decision was based on the fact that the web usage data in the log files consists of multiple dimensions; namely, date, time, IP address, username, URL and agent (Figure 2.3). A snowflake schema also has the advantages of reduced redundancies and less disk storage requirements.

**3.2.2. Pattern Discovery**

The pattern discovery phase relies on various statistical methods and data mining algorithms to detect interesting patterns (Vanzin and Becker 2003). 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. However, no indication of the probabilities or frequencies of certain patterns can be determined by these statistics.
There are several kinds of data mining techniques that can be used. These are divided into two categories, namely supervised (guided) and unsupervised (unguided) techniques (Berson, Smith and Thearling 2004). An unsupervised technique is one where there is no particular reason or goal for creating the pattern. These are more suited to general access pattern tracking (Section 3.2) and are relevant to this research. Supervised techniques are generally used for prediction and are applicable for customised usage tracking.

Some of the more useful and appropriate data mining algorithms used for pattern discovery are sequence analysis, association rules, classification rules and cluster analysis. A classification rule is an example of a supervised mining algorithm while unsupervised algorithms include sequence analysis, association rules and cluster analysis.

One of the challenges in the pattern discovery phase is identifying the target page (Srikant and Yang 2001). For some web sites it is easy to differentiate between content pages and navigation pages. In these cases the target pages for a visitor can be considered as the set of content pages requested by the visitor. However, for some web sites the distinction is not as clear. In these cases a time threshold can be defined and pages where the visitor spent more time than the specified threshold are considered to be target pages. For the purposes of this research, the target page is assumed to be the last web page in a user session.

### 3.2.2.1. Sequence Analysis

The objective of sequential pattern analysis is to discover navigation patterns in the form of sequences (Büchner, Baumgarten, Anand, Mulvenna and Hughes 1999). The discovery of patterns using sequence analysis involves showing that the presence of a set of web pages is followed by another web page within a specified time period. The length of a sequence is the number of web pages in the sequence (Agrawal and Srikant 1995). Sequence analysis enables the analysis of user navigation patterns. A user supports a sequence \( s \) where \( s \) is a set of web pages, 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. In terms of WUM, the most common use of sequential pattern analysis is to identify frequent
navigation patterns within a web site. Sequences satisfying the minimum support are called large sequences (Joshi 1997).

Examples of the information that can be obtained from the output of the sequence analysis algorithm could be as follows (with reference to Figure 2.2):

- 60% of users who accessed /webpages/individualcoursepage, did so by starting at /webpages/staffpages, and proceeding through /webpages/individualstaffpage.
- 55% of users who accessed the site did so by starting at /webpages/academicpage.

The first observed pattern implies that there is useful information in /webpages/individualcoursepage, but since the users take a roundabout route to get there, it may not be clearly marked. The second observed pattern indicates that a large number of users are entering the site through a page other than the home page.

Two algorithms which can be used for mining sequential patterns from the web logs are Apriori and PrefixSpan (Han and Pei 2000).

**Apriori**
The objective of the Apriori algorithm is to find all web pages that occur frequently within a user session. In each pass of the algorithm, the set of frequent sequences from the previous pass is used to generate the candidate or potential sequences and then measure their support by making a pass over the database. At the end of the pass the support of the candidates is used to determine the frequent sequences. In the first pass, the output of the l-itemset phase is used to initialise the set of frequent 1-itemsets, where an n-itemset is a set of n web pages.

**PrefixSpan**
The objective of the PrefixSpan algorithm is to determine all sequences by first scanning a database $S$ for all length-1 sequences, where a length-$n$ sequence is a sequence consisting of $n$ web pages. This derives the set of length-1 sequential patterns. The search space can then be partitioned into the number of unique length-1 sequences. The subsets of sequential patterns can be mined by constructing corresponding
databases and mining each recursively. The sequential patterns with prefix $a$ can be mined from the prefix $a$-projected database. This is the collection that contains only those subsequences prefixed with the first occurrence of $a$. By scanning the $a$-projected database once, all the length-2 sequential patterns prefixed with $a$ can be found. Recursively, all sequential patterns with prefix $a$ can be partitioned into their unique groups. This process is then repeated for all sequences prefixed with $b$ and then for $c$ and so on. By doing this, sequential patterns are identified by exploring local frequent patterns.

The advantages and disadvantages of these sequence analysis algorithms are tabulated in Table 3.1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrefixSpan (Han and Pei 2000)</td>
<td>Projects database based on only frequent prefix subsequences.</td>
<td>Bears similar non-trivial, inherent costs as the Apriori method.</td>
</tr>
</tbody>
</table>

Table 3.1: Comparison of sequence analysis algorithms

As a result of the preprocessing phase (Section 3.2.1), the log file data was stored in time-ordered user sessions. Upon investigation of the sequence analysis algorithms in Table 3.1, it was decided that analysis of the time-ordered user sessions in the data warehouse could be performed without the above-mentioned algorithms. The transformation of the log file data into these time-ordered user sessions represents a large portion of the sequence analysis process. The remaining analysis was required to scan through these user sessions in order to determine the maximal sequences as well as the frequency with which these sequences occur. The implementation of this sequence analysis algorithm is discussed in more detail in Section 6.5.2.

3.2.2.2. Association Rules

*Association rule* mining identifies which web pages are most frequently accessed together in a single user session (Han and Kamber 2001). Relationships between pages which are not necessarily connected together via hyperlinks can also be determined.
Determining association rules 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, also known as coverage. The second step is to generate strong association rules from the frequent itemsets which satisfy both the minimum support and minimum confidence (Cooley 2003), where confidence is the percentage of items which satisfy all the conditions in the association rule.

The main problem with generating association rules is that there are so many possible rules (Borgelt 2004). Therefore, efficient algorithms are needed that check only a subset of all rules. The Apriori algorithm (Agrawal and Srikant 1994), uses support and confidence measures to assess association rules. These measures are discussed in more detail below.

An example of the output of association rule mining could be as follows (with reference to Figure 2.2):

- 40% of the users who accessed 
  `/webpages/staffpages/staffmember1.html`
  also visited
  `/webpages/staffpages/staffmember2.html`

Association rule mining discovers relationships between different web pages found in a database of user sessions (Cooley et al. 1999a). Association rules are often in the form XY where X and Y are web pages. The intuitive meaning of this rule would be that sessions containing X tend to contain Y with a certain degree of support and confidence.

A formal definition of association rule mining is as follows:

Let $I = \{i_1, i_2, ..., i_m\}$ where $m \leq$ total number of web pages in the web site, be a set of web pages. Let $D$ be a set of user sessions where each user session $T$ is a set of web pages such that $T$ is a subset of $I$. Each user session is associated with an identifier TID. An association rule is an implication of the form $A \Rightarrow B$, where $A$ and $B$ are subsets of I, and $A \cap B$ is the empty set. This rule holds for all user sessions in $D$ with support $s$, where $s$ is the percentage of user sessions in $D$ that contain $A$ and $B$. This rule has confidence $c$ in the set of user sessions $D$, where $c$ is the percentage of user sessions in $D$ containing $A$ that also contain $B$. 

Two algorithms which can be used to generate association rules are Apriori and Market Basket Analysis.

**Apriori**
The Apriori algorithm, which was discussed in the section on sequential analysis (Section 3.2.2.1), can also be used for discovery of association rules.

**Market Basket Analysis**
In the online shopping environment a technique known as *market basket analysis* is used to find regularities in the shopping behaviour of customers of online stores (Borgelt 2004). This technique can be applied to web logs in order to analyse the browsing habits of the users of a web site, by finding associations between the web pages viewed by the users (Han and Kamber 2001).

This technique is based on a counting algorithm, with probabilities being computed by taking ratios among various counts (Bloor Interactive 1999). The algorithm builds a co-occurrence table and then generates rules about the likelihood of web pages occurring together. These rules are usually expressed in the form “if...then...”. Market Basket Analysis is not interested in why certain web pages are related, but simply that they are related. The results of Market Basket Analysis would be that pages A and B frequently occur together in a single user session.

The advantages and disadvantages of these association rule algorithms are tabulated in Table 3.2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apriori</td>
<td>Same as for sequential pattern discovery (Table 3.1).</td>
<td></td>
</tr>
<tr>
<td>Market Basket Analysis</td>
<td>Produces simple and easy to understand results.</td>
<td>Only works for discrete data values otherwise binning is necessary.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Works best when web pages occur in roughly the same number of user sessions (Bloor Interactive 1999).</td>
</tr>
</tbody>
</table>

**Table 3.2: Comparison of association rule algorithms**

As can be seen from Table 3.2, Market Basket Analysis is not as effective when web pages do not occur in approximately the same number of user sessions. Apriori will thus be used as the association rule algorithm for analysing web usage patterns.
3.2.2.3. Classification

Classification is the process that maps a web page into one of several predetermined classes (Eirinaki and Vazirgiannis 2003). In the web environment, a class usually represents different user profiles and classification is performed using specific characteristics that denote each user category.

For a database $D$ and a sequence $A$, the support or frequency of $A$ in $D$ denoted by $\text{support}_D(A)$ is the number of users in $D$ whose sequences contain $A$ as a subsequence. The minimum support denoted by $\text{min\_support}$, is a user specified threshold that is used to define frequent sequences. A sequence occurs frequently in $D$ if its support in $D$ is at least $\text{min\_support}$. The rule $A \Rightarrow B (c\%)$ (meaning $A$ implies $B$), and involving sequence $A$ and sequence $B$, is said to have confidence $c$ if $c\%$ of the user sessions that contain $A$ also contain $B$.

In the context of the web, classification is useful to develop a profile of users belonging to a particular class or category. In order to do this, it is necessary to identify features that best describe the given class or category. Classification of the web server logs may lead to the discovery of the following relationship:

- 50% of users who accessed \\
  /webpages/undergradstudents.asp \\
  were first year students.

Classification can be done using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers or k-nearest neighbour classifiers.

**Decision Tree Classifiers**

Decision tree induction involves developing a pattern that describes a distinct attribute, called a class, of an entity in terms of other attributes of a web page, called the observed attributes (Bonchi, Giannotti, Manco, Renso, Nanni, Pedreschi and Ruggieri 2001). A pattern is constructed for a set of web pages, called the training set, whose class value is known and can be used to predict the unknown class value of web pages in another set. An example of this rule would be that users follow pages $A$, $B$ and $C$ to get to target page $D$ in observed data. Therefore one can predict that if a user follows pages $A$, $B$ and $C$, they will go to page $D$. 

28
Naïve Bayesian Classifiers

Naïve Bayesian techniques attempt to find the probability that a condition will be satisfied, given only that it is known how many times it was satisfied in the past, and how many times it was not satisfied (Bloor Interactive 1999). It is the probability that condition $A$ will be satisfied, given that condition $B$ already is and is written as $P(A|B)$. The result of this algorithm could be that 30% of the users that visit page $A$ are in the 18-25 age group and live in Port Elizabeth.

k-Nearest Neighbour

The purpose of k-nearest neighbour classification is to find the $k$ nearest neighbours of a query $q$ in the training set and predict the class label of $q$ as the most frequent one occurring in the $k$ neighbours (Domeniconi and Gunopulos 2002). This classification produces continuous and overlapping neighbourhoods and uses a different neighbourhood for each query. If there are $J$ classes and $l$ training observations, and the training observations consist of $n$ feature measurements $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and the known class labels $j = 1, 2, \ldots, J$. The goal is to predict the class label of a given query $q$.

The output produced is the same as that for Naïve Bayesian Classifiers. The results of the k-Nearest Neighbour algorithm could be that 20% of users that visit page $A$ are first year students who are registered for a computer course.

The advantages and disadvantages of these classification algorithms are tabulated in Table 3.3.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Trees</td>
<td>Can handle raw data (Brooks 1997). Requires little preprocessing.</td>
<td>Mistake at higher level causes all lower levels to be incorrect.</td>
</tr>
<tr>
<td></td>
<td>Produces easy to understand rules. Provide clear indication of which fields are most important for prediction and classification. Can handle large number of fields (Bloor Interactive 1999).</td>
<td>Does not easily handle non-numeric data. Computationally expensive.</td>
</tr>
<tr>
<td>Naïve Bayesian Classifiers</td>
<td>Works well with numeric data. Easy to use (Bloor Interactive 1999). Requires only one pass through the data. Works with partial data.</td>
<td>Performs very poorly in cases of high correlation. Limited to discrete variables.</td>
</tr>
<tr>
<td>k-Nearest Neighbour</td>
<td>Can utilise entire data source rather than require sampling for training (Brooks 1997). Good for discovering clusters.</td>
<td>Requires large amounts of memory. May be overly sensitive to closely matching records.</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of classification algorithms
Since these algorithms do not easily handle non-numeric data such as the log file data (Chapter 2), enumeration of the log files are required before analysis can be performed. It was decided that classification is beyond the scope of this research, due to the fact that additional information regarding the web pages as well as the users of the web site is required. These algorithms will therefore not be considered for the proposed model.

3.2.2.4. Cluster Analysis

Cluster analysis is a technique which groups together web pages with similar characteristics, such that the similarity of the web pages within a group is maximised and similarity between web pages of different groups is minimised (Han, Karypis, Kumar and Mobasher 1997). Cluster analysis makes it possible to discover dense and sparse regions and, therefore, overall distribution patterns among web pages. Clustering is useful when there are many web pages with no natural groupings.

Clustering can also be used as a preprocessing step for other algorithms, such as classification, which would then operate on the detected clusters. Clustering is an example of unsupervised learning which means that it does not rely on predefined training examples. Typical requirements for clustering include: scalability, discovery of clusters with arbitrary shapes, high dimensionality and minimal requirements for domain knowledge (Han and Kamber 2001).

An example of cluster analysis could show that the web usage data can be grouped into the following two clusters, namely:

- Undergraduate students, who frequently visit the web page /webpages/undergradstudents.asp, and
- Postgraduate students, who frequently visit the web page /webpages/postgradstudents.asp

Some of the common algorithms used for cluster analysis are automatic cluster detection and k-Means.

Automatic Cluster Detection
Automatic clustering works by identifying distinguishing characteristics of web pages, and then partitioning the web pages representing these characteristics, along logical boundaries (Bloor Interactive 1999). Distance measures, which may be user-defined, can be used to determine the closeness between web page accesses in any two different user sessions. An example of output from automatic cluster analysis could be that three main clusters exist, namely users who frequently visit page $A$, those who frequently visit page $B$ and those who frequently visit page $C$.

**k-Means**

k-Means is a commonly used clustering technique which searches for a nearly optimal partition with a fixed number of clusters (Michaud 1997). First, an initial partition with the set number of clusters is created. Then the number of clusters remains the same and the partition is improved iteratively. Each web page is handled sequentially and reassigned to the cluster such that the partitioning criterion is most improved. The procedure ends when no improved partitioning is achieved. Since the k-Means procedure works on a fixed number of clusters, the procedure needs to be repeated with different numbers of clusters in order to obtain a final solution. An example of output of the k-Means algorithm could be that page $A$ forms part of a cluster which contains mostly undergraduate students. This implies that page $A$ was visited mostly by undergraduate students.

The advantages and disadvantages of these clustering algorithms are tabulated in Table 3.4.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic Cluster Detection</td>
<td>Works well with categorical, numerical, and textual data (Brooks 1997).</td>
<td>Can be difficult to choose the correct distance measures and weights.</td>
</tr>
<tr>
<td></td>
<td>Easy to apply (Bao 2003).</td>
<td>Sensitive to initial parameters.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can be hard to interpret the resulting clusters.</td>
</tr>
<tr>
<td>k-Means</td>
<td>Relatively scaleable.</td>
<td>Mean needs to be defined (Smith and Ng 2003).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The number of clusters needs to be specified.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noisy data and outliers cause problems.</td>
</tr>
</tbody>
</table>

Table 3.4: Comparison of clustering algorithms

Since automatic cluster detection works well with textual data and is therefore suited to web usage data, it will be used for cluster analysis in the proposed model.

**3.2.3. Pattern Analysis**
Chapter 3 – Web Usage Mining

The final phase in the WUM process is pattern analysis. This is an essential phase of WUM since the WUM algorithms in the pattern discovery phase generate a large number of patterns (Vanzin and Becker 2003). Pattern analysis involves the user evaluating each of the patterns identified in the second phase and deriving conclusions from these patterns. The user is generally concerned in finding patterns that provide useful information regarding the users’ navigation. Visualisation of these results allows them to be more easily interpreted by the user. The process of combining WUM and information visualisation (IV) techniques in order to discover useful information about web usage patterns, is called visual web mining (Chen et al. 2004; Youssefi, Duke and Zaki 2004).

IV is the process of creating visual interfaces to help users understand and navigate through complex information spaces (Eick 2001). The challenge in IV is how to create a visual metaphor which presents the information in a way that makes sense. In terms of the web, there are three significant issues that IV designers face. These are:

- **Scale**: There are a large number of web sites which vary in complexity. These may also have numerous pageviews and accommodate many visitors per day. In order to perform effective analysis, IV systems need to be able to process extremely large data sets.

- **Dimensional Complexity**: It is now technically and financially feasible for any organisation operating a web site to gather rich, fine-grain, online web usage data sets. Integrating online and offline activity increases the dimensionality of the problem. By correlating site activity with enterprise data, many analysis problems can be solved.

- **Range of analyses and navigation tasks**: Before attempting to visualise information, it is important to understand the goals of the user of such visualisation tools. The user of a WUM tool would be the site designer or developer. The developer’s goals generally involve determining which pages on the site are viewed most frequently, where and how the site visitor is most likely to find interesting content, and which pages are the most common entry or exit points. In terms of the goals of this research, the visualisation techniques should focus on showing the path and flow through the site by the visitors in order to assist designers in building a more effective site.
These issues were taken into consideration when designing the visualisation of the results of the WUM algorithms.

### 3.2.3.1. Visualising Web Usage Patterns

Most web log analyser applications are capable of outputting the general statistics regarding the usage of the site in a graphical format. Producing statistical data can easily be represented using either bar, line or pie graphs. It is, however, more of a challenge to display the navigation paths, association rules and clustering results in an intuitive way, since once these data analysis techniques have been applied, the results can often be extensive and difficult to interpret (Chi 2002). The impact of these results can be maximised by providing the structure of the web site being analysed, together with the results. By visualising the structure, the user can determine if the web site is structured correctly (Eick 2001).

One way to show web usage patterns is to make use of a tree metaphor, by superimposing a trace of the web pages visited on top of the site map (Eick 2001). By graphically comparing the structure of the web site as a tree of paths to the sequence of navigation patterns, one is able to clearly visualise the path taken by the user to achieve a specific task, as well as determine if the path taken was in fact intended by the designer (Spiliopoulou 2000). Figure 3.2 shows an example of this. In Figure 3.2 the CS&IS website is represented as a tree hierarchy and the navigation path for a specific user is represented as a dotted line.
Figure 3.2 represents the user’s navigation during a single user session. The start page and target page have been indicated. One can see that the user began the navigation session at the CS&IS home page. The target page was L, namely the Project Page. It would seem obvious that the user would follow the path A,D,L in order to get to the target page; however the path that was followed was A,B,G,B,H,B,A,D,L. The decision taken by the user to follow the longer path could be a result of poor information architecture. The benefit of this visualisation method is that the navigation path is clearly visible at a glance. A variation of this type of graph displays edges with numbers which represent the number of users who navigated from one page to another. This is useful in determining the volume of traffic to various areas of the web site.

Hierarchical trees are frequently laid out radially, with the home page in the middle and the rest of the pages at varying depths in circles around the centre node (Eick 2001). A disk tree algorithm can be used to convert a hierarchical tree to a radial tree using a breadth first search (Chen et al. 2004). The idea behind this technique is that for each web page, only one link to it is represented, namely the primary link and that embedded links are not shown. A disk tree layout (Figure 3.3) also makes better use of the screen real estate than the vertical layout.
Figure 3.3 shows a disk (radial) tree visualisation used to show one week’s worth of web site usage. In this case, the volume of traffic throughout the web site is indicated using green lines of varying thickness. The thicker the line between two web pages, the more traffic occurred between these pages.

The disk tree visualisation has the advantage of being able to visualise the web site structure compactly (Chi, Pitkow, Mackinlay, Pirolli, Gossweiler and Card 1998). The circular layout is also aesthetically pleasing to the eye since there is no crossing of links.

3.3. Conclusion

A single layer data architecture was selected as the most appropriate architecture for web usage data, since this data is kept in log files (Section 3.2.1.2) and does not require duplication. A snowflake schema was identified to model the data to be stored in the data warehouse. This decision was based on the characteristics of web usage data and the advantages of the snowflake schema.

Several data mining techniques were investigated as part of the pattern discovery phase (Section 3.2.2). Three algorithms were selected based on their suitability for WUM, namely Apriori for Association Rules, Automatic Cluster Detection for Cluster Analysis and the Sequence Analysis algorithm; which uses the time-ordered sequences provided by the preprocessing of the log files.

A radial tree was selected as the most appropriate visualisation technique to visualise the results of the selected WUM algorithms (Section 3.2.3.1). Colour and other visual cues are used to represent the results of the various WUM algorithms. A radial tree enables the structure of the web site to be visualised concurrently with the web usage patterns.

In order to obtain a better understanding of WUM, several related systems are investigated in Chapter 4.
Chapter 4 - Related Systems

4.1. Introduction

The fourth research question in Table 1.1 is concerned with an investigation into related WUM systems. In order to answer this question, three related systems were evaluated. These systems are 123LogAnalyser (ZY Computing Inc 2003), WebLog Expert (Alentum Software Inc 2004) and Absolute Log Analyser (BitStrike Software 2004). The criteria that were used when evaluating these systems correspond to the three phases of WUM (Figure 3.1). This chapter aims to determine for each system, the type of data storage that takes place in the preprocessing, what WUM algorithms are used in the pattern discovery and how these patterns are visualised in the pattern analysis. The limitations of each system are discussed.

4.2. 123LogAnalyzer

123LogAnalyzer (ZY Computing Inc 2003) is a Windows-based program which can read the major log file formats from both UNIX and Windows platforms. It is simple and its intuitive interface requires no technical knowledge.
4.2.1. Data Storage

One useful feature of 123Log Analyzer is the program's ability to analyse log file archives (such as ZIP or GZ) without the need to extract the files to the client machine first. Retrieving and analysing compressed logs from a remote location can also save some download time and hard drive space on the client machine. 123LogAnalyzer does not, however, allow multiple log files to be in the same archive. In addition to allowing files to be manually added for analysis, 123LogAnalyzer also allows the files to be downloaded directly from a remote location via FTP or HTTP. The log file types that are accepted as input are .log and .txt.

123Log Analyzer performs the analysis directly on the log files without duplicating the data. For this reason, no separate data warehouse is required.

4.2.2. WUM Algorithms

Once log files have been added for analysis, various filters can be applied in order to perform an in-depth and precise analysis of the data. Figure 4.1 shows the filtering options available.

![Figure 4.1: 123LogAnalyzer options](image)
These filters can enable the user to fine-tune the results. For example, the user can determine how many users visited a particular page and also what other pages they visited, what their browsing sequences are, which files they downloaded, which website they were referred from, and what keywords they used to find the website being analysed. Combinations of these filters can also be used.

It is not evident, however, from the investigation or from the documented reports, exactly what WUM algorithms are used for this analysis and only descriptive statistics are provided.

4.2.3. Visualisation of Results

Once the analysis has been performed, an HTML report is generated and is displayed in the user’s Internet browser. This HTML report is shown in Figure 4.2. The information is categorised in the left hand column and the results are displayed on the right hand side.

![Figure 4.2: 123LogAnalyzer generated report](image-url)
Where applicable, simple 2D bar and line graphs of the reported information are also provided. The information provided in this report is shown in Table 4.1.

<table>
<thead>
<tr>
<th>Information Provided</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Statistics</td>
<td>This shows a daily visit report graph and a statistics table for the time period entered for the range of dates in the analyse window. It also provides a summary of the activity statistics.</td>
</tr>
<tr>
<td>Activity Statistics</td>
<td>This displays the number of visitors, unique IP’s, amount of bandwidth used, and the number of hits, broken down by Time Increment, Day of the Week, and Hour of the Day for the time period entered. Analyses by Pageview per Day and by Visitor Stay Length can also be viewed.</td>
</tr>
<tr>
<td>Visitor/IP Information</td>
<td>This displays the IP addresses, Domain Name, Country, Time of Last Access, and IP Ownership information for the site's visitors, broken down by Access Time, Hits, Bandwidth, Stay Length, and Authenticated Visitors.</td>
</tr>
<tr>
<td>Resource Accessed</td>
<td>This provides web pages viewed, files downloaded, directories that were accessed, and images that were accessed during the time period, broken down by Page Views, Browsing Sequences, Downloaded Files, Accessed Directories, Accessed Images, Top Entry Pages and Files, and Top Exit Pages and Files.</td>
</tr>
<tr>
<td>Referrer Statistics</td>
<td>This report shows which Domains and URLs the visitors come from according to Referring Domains and Referring URLs.</td>
</tr>
<tr>
<td>Search Engine Statistics</td>
<td>This report displays the search engines that referred visitors to the site, the phrases and keywords visitors searched for broken down by Top Search Engines, Keywords, and Each Search Engine.</td>
</tr>
<tr>
<td>Geographic Region</td>
<td>This report displays a Most Active Countries graph and table showing which Countries the visitors come from during the time period.</td>
</tr>
<tr>
<td>Browsers and Platforms</td>
<td>This report shows which browsers and platforms were used by visitors who visited during the time period.</td>
</tr>
<tr>
<td>Technical Information</td>
<td>An explanation of errors encountered is provided.</td>
</tr>
<tr>
<td>Marketing Performance Report</td>
<td>The Marketing Report helps analyse the online marketing campaign. 123LogAnalyzer tracks costs, sales, and profits by visitors, by bandwidth, and by page views and downloads. Information regarding marketing costs and tracked pages needs to be entered.</td>
</tr>
</tbody>
</table>

Table 4.1: Analysis report from 123LogAnalyser

4.3. **WebLog Expert**

WebLog Expert (Alentum Software Inc 2004) is a web log analyser which can reveal important statistics regarding a web site’s usage.
4.3.1. Data Storage

Web Log Expert supports the W3C Extended log format that is the default log format of Microsoft IIS 4/5/6. It also supports the Combined and Common log formats of Apache web server. It supports compressed log files (.gz and .zip) and can automatically detect the log file format. If necessary, log files can also be downloaded via FTP or HTTP. Analysis is performed directly on the log files and no separate data warehouse is required.

4.3.2. WUM Algorithms

Figure 4.3 shows the options window of WebLog Expert. All analysis options, reporting settings and log file download settings are available from this screen. It is possible to schedule an analysis to take place automatically.

It is, however, not apparent which WUM algorithms are used for this analysis and only descriptive statistics are provided.

![Figure 4.3: WebLog Expert options](image)
Once the log files have been selected there is an include/exclude filter, which allows the user to select what information should be included or excluded from the analysis.

### 4.3.3. Visualisation of Results

WebLog Expert produces an easy-to-read HTML report, similar to that produced by 123LogAnalyzer (Section 4.2.3). This HTML report contains various categories which can be navigated from the menu on the left hand side. The textual results together with simple bar and line graphs are then displayed on the right hand side as is shown in Figure 4.4. If an analysis has been scheduled, the generated report can be automatically emailed to the web designer, if required.

![Figure 4.4: WebLog Expert report](image)
The information provided in the HTML report is summarised in Table 4.2.

<table>
<thead>
<tr>
<th>Information Provided</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General statistics</strong></td>
<td>total and average hits, total and average page views, total and average number of visitors, total and average bandwidth</td>
</tr>
<tr>
<td><strong>Activity statistics</strong></td>
<td>daily, by hours of the day, by days of the week and by months</td>
</tr>
<tr>
<td><strong>Access statistics</strong></td>
<td>statistics for pages, files, images, directories, entry pages, exit pages, paths through the site and file types</td>
</tr>
<tr>
<td><strong>Information about visitors</strong></td>
<td>hosts, top-level domains, countries, states, cities, organisations, authenticated users</td>
</tr>
<tr>
<td><strong>Referrers</strong></td>
<td>referring sites, URLs, search engines (including information about search phrases and keywords)</td>
</tr>
<tr>
<td><strong>Browsers, operating systems and spiders statistics</strong></td>
<td>most frequently used browsers and operating systems as well as frequently detected spiders</td>
</tr>
<tr>
<td><strong>Information about errors</strong></td>
<td>error types, detailed 404 error information</td>
</tr>
<tr>
<td><strong>Tracked files statistics</strong></td>
<td>activity and referrers</td>
</tr>
</tbody>
</table>

Table 4.2: Analysis report from WebLog Expert

4.4. Absolute Log Analyzer

Absolute Log Analyzer (BitStrike Software 2004) is a client-based log file analysis tool, designed for web traffic analysis. Unlike the previous two systems, the main screen of Absolute Log Analyzer is the same screen in which the report is displayed. Firstly, log files need to be added to the analysis and the results are then displayed. Apart from the graphical user interface, Absolute Log Analyzer also has a command line interface.

4.4.1. Data Storage

Absolute Log Analyzer allows log files to be downloaded via FTP. The analyser can recognise the majority of log file formats (Microsoft IIS and Apache) automatically. It also has the facility to manually specify your own format for non standard log files. It will analyse compressed log files (.gz and .zip) and can recompress them to minimise drive space usage. The analyser imports data into the highly optimised proprietary database. This allows the user to incrementally update the statistics as new log files become available and makes it simple to zoom in on a particular quarter, month, week, or day and even view all of these statistics in the same table, so that any trends can be seen.
4.4.2. WUM Algorithms

Figure 4.5 shows the settings for the analysis using Absolute Log Analyzer. These settings are used to tailor various aspects of the analysis and are categorised by the tabs at the top of the window.

It is not apparent, however, which WUM algorithms are used for the analysis, nor is there any way of selecting alternate algorithms. Only descriptive statistics are provided.

![Workspace settings](image)

**Figure 4.5: Absolute Log Analyzer options**

4.4.3. Visualisation of Results

As discussed previously, the report for Absolute Log Analyzer is displayed in the main screen as shown in Figure 4.6. There is no option to export the full report or subsections of it into HTML format. The menu on the left displays the available statistics which have been categorised similarly to the previous systems. The textual results are shown in the right hand window and where applicable, a graphical representation is shown in the window at the bottom of the screen. The report information is summarised in Table 4.3.
### Information Provided

<table>
<thead>
<tr>
<th>Information Provided</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traffic</strong></td>
<td>Reports in this group describes general site information and includes summary report, list of visits during particular day and three types of information about requests to the server, used bandwidth, visits, page hits, downloads, images, bots and spiders, web server errors.</td>
</tr>
<tr>
<td><strong>Pages</strong></td>
<td>This group contains information about page views on the site. Particular reports are most visited pages, entry pages, exit pages, single access pages and paths through site.</td>
</tr>
<tr>
<td><strong>Downloads</strong></td>
<td>This group is dedicated to resources with type ‘Download’. Here reports about downloaded files, used bandwidth, referrers and referrer relevancy analysis can be found.</td>
</tr>
<tr>
<td><strong>Images</strong></td>
<td>Here information about banners or other important image files is found.</td>
</tr>
<tr>
<td><strong>Referrers</strong></td>
<td>Contains complete information about site’s referrers such as how many visitors comes from a particular page, server or search engine, which search words and phrases have been used to find the site and so on.</td>
</tr>
<tr>
<td><strong>Audience</strong></td>
<td>Provides more information about visitors, including where they reside, which operating system uses and more.</td>
</tr>
<tr>
<td><strong>Bots and Spiders</strong></td>
<td>Information about bots, spiders, crawlers and other non-human activity on the site.</td>
</tr>
<tr>
<td><strong>Technical Information</strong></td>
<td>Broken link reports and page not found errors can be found here.</td>
</tr>
</tbody>
</table>

Table 4.3: Analysis report from Absolute Log Analyzer
4.5. **Comparison**

The three systems investigated all provide descriptive statistics regarding the activity of the server in terms of files requested, referring web sites and peak traffic times. These systems also provide useful bar and line graphs for representing the statistics generated.

These systems all provide information regarding frequent navigation paths through the site; however they do not sufficiently illustrate browsing characteristics such as similarities between the multiple sessions for a single user or files accessed frequently between various sessions. These systems also do not provide a graphical representation of the users’ navigation paths. The textual representation of these paths provided by these systems is difficult to interpret since these are simply combinations of the URLs visited.

4.6. **Conclusion**

This chapter evaluated three related systems by investigating the WUM algorithms, the data mining models, as well as the visualisation techniques used by each system. The three systems evaluated were 123Log Analyzer, WebLog Expert and Absolute Log Analyzer.

None of the three systems are able to satisfy the goals that were established for this research, in terms of applying WUM algorithms to web usage data and visualising the patterns produced (Section 4.5).

The next chapter discusses the design of a model for the WUM of organisational web sites.
Chapter 5 – Proposed Model

5.1. Introduction

The objective of this chapter is to answer the fifth research question in Table 1.1 by proposing a model for WUM. The specification of the proposed model includes the data model design, the algorithm design, the information visualisation techniques and the design of the user interface (UI). An iterative design approach was used to design the UI using the visualisation techniques identified in Chapter 3. This enabled low-level prototypes to be developed and evaluated by the users in order to produce high-level designs based on the feedback obtained. The data warehouse design (Section 5.3) explains the relationships between the data elements of the model and illustrates how they are modelled using a data schema. Section 5.4 describes the WUM algorithms used to analyse the web usage data in order to extract web usage patterns. This section also discusses the techniques used to visualise the results of these algorithms. The UI design describes the design of the user interface to visualise the web usage patterns identified using the WUM algorithms (Section 5.5).

5.2. Architecture of the Model

Han and Kamber (2001) propose an architecture for WUM systems. This is discussed in terms of a 3-tier architecture on which data mining systems are based. The architecture of the proposed WUM model is derived from this architecture and is shown in Figure
5.1. The proposed model consists of three layers which correspond to the three phases of WUM shown in Figure 3.1.

These three layers are:

1. The Data Layer: the data warehouse containing the preprocessed data obtained by preprocessing the log files kept by the NMMU CS&IS web server.
2. The Application Layer: comprising a web mining server which applies the WUM algorithms to the data in the data warehouse. This facilitates the pattern discovery phase.
3. The Presentation Layer: consists of a graphical user interface with which the user interacts with the WUM system and views the results. This layer enables the user to perform effective pattern analysis.

The presentation layer makes use of web services to interact with the application layer (Figure 5.1). The application layer in turn retrieves the necessary data from the data layer via a JDBC connection. The implementation of the web services will be discussed further in Chapter 6. In using this model, the user would make the desired selections, such as which algorithm to use as well as the filtering criteria, via the graphical user interface. These selections are passed to the web mining server by means of a web
service. Here the choices made by the user are used to apply the selected algorithm and the appropriate filtering criteria to the data in the data warehouse. The data in the data warehouse is passed to the web mining server via the JDBC connection, the algorithm is applied and the results are sent to the user interface via the web service where they are graphically displayed.

The components of each of the three layers of this model are now discussed in more detail.

5.3. Data Warehouse Design

The output of the preprocessing phase of WUM is clean data, stored in a data warehouse ready for analysis (Section 3.2.1). The fields that were identified from the log files as being necessary for WUM are: date, time, client ip, client username, URL and user agent (Section 2.2.3). The snowflake schema shown in Figure 5.2 was selected based on the web usage data being stored and used to model the data recorded by the web server to facilitate easy storage and retrieval of the data.

The Fact Table contains the measures (Section 3.2.1.3) as well as the keys for each of the dimensions. These measures include number of accesses (NoAccesses) and number of sessions (NoSessions). These measures are used to determine the support of various discovered rules, during the pattern discovery phase. The dimensions Client, Time, Access and Pattern allow the Fact Table to be filtered on these dimensions. The
attributes provide a description for each dimension. The path attribute in the Pattern Dimension Table is determined by the sequence of URLs accessed in each user session.

5.4. Web Mining Algorithm Design

As discussed in Section 3.2.2, there are many data mining algorithms available. Not all algorithms can be applied to all types of data. An initial choice of WUM algorithm for each of the three categories namely, association rules, sequential analysis and cluster analysis, was selected. These choices were:

2. Sequential Analysis: The sequences accessed in the user sessions.

These algorithms will be discussed in more detail, highlighting the format of the output of each algorithm. This information is needed to assist with designing the appropriate type of visualisation technique for visualising the results.

5.4.1. Association Rules

The purpose of association rule mining is to discover relationships between web pages found in a database of user sessions (Section 3.2.2.2). An association between certain web pages would imply that these web pages are frequently accessed within the same user session.

The design of the Apriori algorithm is based on an implementation by Yibin, Hamilton and Liu (2000). Changes to this implementation to allow it to suit web usage data are discussed in Chapter 6.

The output that is produced by the Apriori algorithm on web usage data would be of the form:

\[ (P_1, P_2, \ldots, P_n) \ (c, f) \]

where

\[ 1 < n < maxNodes; \ (P_1, P_2, \ldots, P_n) \] are the URLs of the web pages that are associated with one another in a single user session; \( c \) is the level of confidence and \( f \) is the level of support (frequency).
An example of output from the Apriori algorithm could be: (/DegreesAndCourses.asp,/UndergradSubjects.asp) (85,60). This is interpreted as 85% (confidence) of user sessions that contain the web page “/DegreesAndCourses.asp” also contain the web page “/UndergradSubjects.asp”. Sixty percent (support) of all the user sessions contain both these web pages.

5.4.2. Sequence Analysis

Sequence analysis is the process of determining the longest time ordered paths that satisfy a user specified minimum frequency. Since the log files represent a user’s interaction on a web site, the goal is to discover patterns in the form of sequences of page accesses (Section 3.2.2.1).

The output that is produced by sequence analysis of web usage data would be of the form:

\[(P_1, P_2, \ldots P_n) \ f\]

where

\[1 < n < maxNodes; (P_1, P_2, \ldots P_n)\] are the URLs of the web pages that form the sequence, and \(f\) is the frequency (support) 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 “/HomePage.asp”, “/SubjectList.asp”, followed by “/SubjectDetails.asp” during their user sessions.

5.4.3. Cluster Analysis

Clustering is a discovery process that groups web pages into sets such that the similarity of the web pages within a group is maximised and similarity between web pages of different groups is minimised (Section 3.2.2.4).

The output that is produced by cluster analysis of web usage data would be of the form:

\[(P_1, P_2, \ldots P_n) \ f\]
where

\[ 1 < n < maxNodes; \ (P_1, P_2, \ldots, P_n) \] are the URLs of the web pages within a particular cluster and \( f \) is the frequency with which this cluster occurs.

For example, consider the output \((/\text{HomePage.asp}, /\text{StaffList.asp}, /\text{StudentDetails.asp})\) 40\%. This can be interpreted as the “/HomePage.asp”, “/StaffList.asp” and “/StudentDetails.asp” web pages form a cluster based on the specified criteria and this cluster occurs in 40% of the user sessions.

5.5. Visualisation Design

Visual web mining allows WUM results to be interpreted more effectively (Section 3.2.3). In order to select an appropriate IV technique, the data type and format of the results needs to be analysed. The data that is visualised in visual web mining is web usage information, which consists of sets of URLs and relevant frequencies (Section 5.4). 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 embedded hyperlinks that may exist which link web pages with each other, the structure may represent a connected graph (web structure) rather than a tree (Section 3.2.3.1).

The radial tree visualisation metaphor discussed in Section 3.2.3.1, works well for relatively small sites with tens to hundreds of pages (Eick 2001). Most organisational web sites can be classified as relatively small, so this technique was chosen to represent the static web site structure when visualising the results of the WUM algorithms.

5.5.1. Visualising Association Rules

Since the purpose of association rules mining is to show that two or more web pages are accessed together frequently within a user session, the visualisation technique needs to be able to show all the associated web pages concurrently.
The primary node or home page is located in the centre of the graph (Figure 5.3). Each successive descendant falls on concentric rings spanning out from the centre. As discussed in Section 3.2.3.1, the advantage of radial trees 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 web page accesses, visual cues can be used. For instance, pages which are associated can be coloured similarly. In Figure 5.3, 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. For example, when page 1 occurs in a user session, page 2 frequently occurs in the same session and vice versa.

5.5.2. Visualising Sequence Analysis

When visualising user navigation sequences it is also essential to show the structure of the web site. 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 5.4 indicates how a radial tree can be used to visualise user navigation sequences.

Two separate sequences are shown in Figure 5.4; the first showing a path from 0 to 1 to 2 (coloured in red), and the second showing a path from 0 to 3 to 4 (coloured in blue). These two sequences represent the most popular paths that occur in the user sessions.

5.5.3. Visualising Cluster Analysis

Visualising clusters can be achieved using the same technique as visualising associated web pages (Section 5.5.1). The web pages that form part of a specific cluster are then coloured similarly to show their grouping.

Figure 5.5 shows how clusters can be visualised using radial trees. The six nodes coloured in red, namely nodes 0 through 5, form part of a cluster. The interpretation of a specific cluster is determined by the criteria selected when performing the cluster analysis.
5.6. **User Interface Design**

An iterative design approach was used to design the user interface (UI) of the proposed model (Section 5.2). Conceptual model extraction was used to obtain feedback from users and improve the UI design.

The UI was designed to enable the user to view the web usage patterns over a specified period and consists of three coordinated views (Figure 5.6). These views are coordinated so that any changes to one of them are reflected in the others. The three views displayed are the graphical view on the left (bordered in red) which shows the visualisations of the WUM results; the textual view at the bottom (bordered in green) which provides the WUM results in textual format (web usage paths); and the filtering view on the right (bordered in blue) which allows the user to select what data should be analysed by the WUM algorithms (filtering menu).
Chapter 5 – Proposed Model

Figure 5.6: UI design of WUM prototype (Frequent paths)
The graphical view illustrated in Figure 5.6 depicts a radial tree used to display the structure of the web site being analysed, as well as the results of the sequence analysis algorithm. The graphical view displayed is dependent on which WUM algorithm is selected.

There are three buttons below the filtering menu, which allow the user to switch between any of the WUM algorithms, namely frequent paths, associated pages and page clusters. 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 used. The ‘Page Structure’ option allows 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 site.

At the bottom of the filtering menu is an ‘Apply’ button which applies the desired filtering to the data, invokes the selected WUM algorithm and refreshes the coordinated views.

5.7. Conclusion

The goal of this chapter was to propose a model for WUM of organisational web sites. The proposed model consists of three layers, namely the data layer, the application layer and the presentation layer. The data layer comprises a data warehouse which stores web usage data (Section 5.3). The characteristics of the web usage data (Section 2.2.3) were used to select the three WUM algorithms namely, sequence analysis, association rules and cluster analysis (Section 5.4).

The output of the algorithms was analysed to select an appropriate visualisation technique. A radial tree was selected as the visualisation technique for the results of all three WUM algorithms (Section 5.5). In order to facilitate visual web mining, the three WUM algorithms, together with the selected visualisation technique, were integrated in a graphical UI (Section 5.6). This graphical UI allows selection of the WUM algorithms as well as filtering of the results.

The next chapter discusses the implementation of a prototype, called WebPatterns, to demonstrate the effectiveness of the proposed model.
Chapter 6 – Implementation of Prototype

6.1. Introduction

The objective of this chapter is to demonstrate the effectiveness of the model proposed in Chapter 5. This will be done by implementing a prototype, called WebPatterns. The selection of a suitable implementation tool is discussed (Section 6.2) as well as the implementation of the three-tier architecture of the model proposed in Chapter 5, namely the data warehouse (Section 6.4), the WUM algorithms (Section 6.5) and the visualisation techniques (Section 6.6). Any modifications that were made during the implementation of the prototype are also documented.

6.2. Implementation Tools

The model proposed in Chapter 5 consists of three layers (Section 5.2), namely the data layer (data warehouse), application layer (WUM algorithms) and presentation layer (UI). In order to implement WebPatterns, appropriate implementation tools needed to be selected for each of these three layers.

The log files created by the CS&IS web site are large and as a result, a data centre for this data was required. Oracle was selected to host the data warehouse since it can store large volumes of data and provides a centralised location for the data. A data centre which was already operational on a Sun server (Van Tonder 2005) was used to store the
data warehouse. The computational speed and fast data access provided by the Sun server were also favourable features.

The Sun server also provided the means to store the web mining server of the application layer. For this reason, the WUM algorithms were implemented using Java (Sun Microsystems 2005) in order to be compatible with the web server.

Two tools were considered for the implementation of the UI of the presentation layer, namely Microsoft’s C#.NET (Microsoft Corporation 2005) and Sun Microsystems’ Java. The presentation layer was required to connect to the application layer using web services. This was necessary so that the desired filtering criteria could be supplied to the WUM algorithms, and that the results of these algorithms could in turn be sent back to the UI in order to be visualised. The visualisation technique that was selected in Chapter 5 to represent the web usage information was a radial tree (Section 5.5). Therefore, the only data representation requirements that the presentation layer needed to satisfy is to render radial trees by applying the necessary layout algorithm.

Several data manipulation requirements were identified for WebPatterns. These include (Carr 1999):

- **Filtering:** This is the process of reducing the data set based on the attributes. The data set to be analysed can be reduced by setting certain parameters and specifying desired ranges.

- **Details-on-demand:** While exploring a data set, the user can view the details of a particular data item by clicking or hovering the mouse over it and have the details displayed in a popup or auxiliary window.

Based on the requirements of the presentation layer, several charting components were considered and their features compared in terms of data visualisation, data manipulation and data warehouse connectivity. These components included Nevron Chart (Nevron LLC 2004), NetCharts Pro (Visual Mining 2004), FlowChart.Net Pro (MindFusion Ltd 2005) and Guess (Hewlett Packard 2004). A comparison of the tools that were considered and the features they support can be found in Appendix A.
Only two charting components were found that provide a way of rendering radial trees, namely FlowChart.Net and Guess. Both of these tools support the data representation and manipulation requirements. A feasibility study of these two tools was conducted using demo versions of each. Based on the level of support provided by each tool, FlowChart.Net Pro was selected as the most suitable option for visualising the WUM results (Appendix A). Since FlowChart.Net is a Microsoft.Net component, the presentation layer of WebPatterns was implemented using Microsoft C#.Net (Appendix A).

6.3. Architecture

This section provides a discussion of the implementation of the architecture of the model described in Chapter 5. The architecture of WebPatterns shown in Figure 6.1 is based on the architecture of the model proposed in Figure 5.1.

The data layer consisting of the data warehouse and the application layer consisting of the WUM algorithms were implemented on the same host, namely the Sun V40z Server. The operating system used is Solaris 10 which makes use of containers to host different applications and services. For WebPatterns, the containers used were the data warehouse and the WUM algorithms.

![Figure 6.1: WebPatterns architecture](image-url)
The data warehouse which stores the preprocessed log file information was created with Oracle 10g using the snowflake schema shown in Figure 5.2. This data warehouse will be discussed in more detail in Section 6.4. The web mining server is run by Sun Java System Application Server which is hosted on the Sun Fire V40z Server. Included in this web mining server is the knowledge base which contains the WUM algorithms. The data layer and application layer interface was implemented using a JDBC connection.

The presentation layer comprising the UI of WebPatterns is stored on the client workstation. The UI interacts with the application layer by means of web services. During interaction with WebPatterns, the user is required to select which of the visualisations are to be viewed, as well as the various filtering options required by the selected visualisation. These choices are then passed to the Web Mining Server via a web service. The Web Mining Server then uses the parameters provided to query the data warehouse via the JDBC connection and obtain the results. These results, in the format discussed in Section 5.4, are then sent back to the UI by means of a web service, where they are visualised using the selected technique.

6.4. Data Warehouse

In order for the WUM algorithms to be applied to the web usage data obtained from the log files, a certain amount of preprocessing first needs to take place. Once the data in the form of daily log files is obtained, the required fields (Figure 2.1) need to be extracted and the various user sessions identified. Before this data can be preprocessed, the snowflake schema shown in Figure 5.2 needs to be implemented in order to store the data.

Various problems exist with log files as discussed in Section 2.2.4. Since the IP address and user agent fields are not always reliable in distinguishing between various users and because the username of the user is not always recorded in the log file, all three of these fields were taken into consideration when creating the user sessions. This means that a change in any one of these fields is considered as being an access by a different user.
The log files record each user’s URL accesses throughout each day, but are only saved at the end of each day. For this reason, the data stored in the data warehouse is not real-time data. The WebPatterns prototype does, however, allow log files to be preprocessed and added to the database as soon as they become available. This enables the data warehouse to remain up to date. In order for the administrator to add data to the database, the user needs to select the location where the log files are stored and select which log files to use. WebPatterns then determines the most recent data in the data warehouse and processes the log files from this point on. This is possible, since each log file name records the date on which it was saved. This process is not, however, an automated one and the administrator is responsible for making sure that as log files become available, they are processed.

The next section discusses the implementation of the WUM algorithms in the application layer, and how these algorithms are used to manipulate the data in the data warehouse.

6.5. WUM Algorithms

Section 5.4 detailed the design of each of the selected WUM algorithms, as well as the format of the output provided by each algorithm. During the design of the association analysis (Section 5.4.1) and cluster analysis (Section 5.4.3) algorithms, it was noted that there are several similarities with respect to the results provided, as well as the graphical representation of these results. The results of the cluster analysis do not provide significantly more information to the user than the association analysis can provide, using the same filtering. For this reason, a decision was made not to implement the cluster analysis algorithm in the prototype. The WUM algorithms that were implemented in WebPatterns are association rules and sequence analysis.

6.5.1. Association Rules

The algorithm used for the association rule analysis is the Apriori algorithm (Section 3.2.2.2). An existing implementation (Yibin et al. 2000) was implemented and
modified. This algorithm is very efficient and produced results in the format required by the presentation layer.

The initial implementation was designed to take integer values as input, not strings such as URLs. For this reason the URLs in the data warehouse needed to be coded as integer values. This implementation also required various text files in order to run, namely a configuration file containing fields for the number of records, number of web pages and minimum support, and a transaction file which contains the log file information.

Modifications were made to allow the number of records and number of web pages to be determined from the data in the data warehouse and the minimum support to be obtained from the filtering menu in the UI, thus eliminating the need for a configuration file. In order to eliminate the need for a transaction file, a new table with the same format as the transaction file was added to the data warehouse. Each entry in this table records the date/time of the session as well as details of which web pages were accessed during the session. The session details are recorded in bitmap form, that is, either ‘0’ or ‘1’ is recorded for each of the web pages in the web site. A ‘0’ indicates that a web page was not accessed during the user session and a ‘1’ indicates that it was accessed. Figure 6.1 shows a few entries from this bitmap table.

<table>
<thead>
<tr>
<th>date_time</th>
<th>session_details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-01-14-07:22:11</td>
<td>1 0 0 1 1 0 0 0 1 0 1 1 1 0 0 0</td>
</tr>
<tr>
<td>2004-01-14-08:00:03</td>
<td>1 1 0 0 0 0 1 0 0 1 0 0 1 0 1 0 1</td>
</tr>
<tr>
<td>2004-01-14-08:53:17</td>
<td>0 0 1 1 1 0 0 0 0 1 0 1 0 1 0 1 0 1</td>
</tr>
</tbody>
</table>

Table 6.1: Extract from Bitmap transaction table

This bitmap table is derived from the existing data and can be updated at the same time as new data is added to the data warehouse by appending new entries to the end of the table. This is possible since successive time-ordered log files do not contain overlapping information.

When the Apriori algorithm is run, the results are sent back to the presentation layer as an array containing the associated web pages and the corresponding frequencies. The format of these results makes it easy to visualise the results. The visualisation of the results is discussed in Section 6.6.3.
6.5.2. Sequence Analysis

Sequence analysis requires that the target page is determined for each user session. When implementing the sequence analysis algorithm (Section 5.4.2), there was no predefined method to determine which page in a user session is the target page. For this reason, the target page was considered to be the last page in a session.

Each time a web page is refreshed, either by the user or automatically by the browser, it is recorded as a request for the same web page. This duplication of pages causes paths to be extremely long and contain redundant information. For this reason, adjacent duplicate requests for a single page are eliminated from a path.

Each path is stored in a hash table which contains the path as well as a count of the number of times the path occurs. When analysing the data, a new path is first checked against the existing paths in the hash table. If this path already exists, then its count is incremented. If it does not yet exist in the hash table, then it is added with a count of 1. The use of the hash table allows for paths to be compared very efficiently, thus speeding up the execution of the algorithm. The output of the sequence analysis algorithm is an array containing the list of paths, as well as an array containing the corresponding frequencies. In this format the results can then be easily used by the presentation layer to visualise the paths as discussed in Section 6.6.4.

6.6. Visualisation

Chapter 5 documented the design decisions made regarding the UI and visualisation of the results of the WUM algorithms. Based on iterative feedback from the users and due to certain limitations of the charting tool (FlowChart.Net Pro), some changes were made to the UI and the visualisation techniques. This section discusses the implementation of the visual components of the prototype, focusing on the required modifications.

The initial design involved the use of a web based metaphor. A web-based tool has the advantage of being accessible from any workstation with an Internet connection.
During iterative design this concept was not supported by the users. The users suggested that a stand-alone tool be created instead, since this is the standard procedure for WUM tools. The users also indicated that a web-based tool would not be required.

**6.6.1. User Interface**

The UI in WebPatterns has been kept simple with the aim of allowing tasks to be performed as efficiently as possible. The UI is divided into three main screen elements as shown in Figure 6.2:

- The left-hand section which contains the graphical view.
- The right-hand top section which provides filtering options.
- The right-hand bottom section which displays the textual results.

The iterative prototyping methodology was used in order to obtain user feedback during the implementation of the prototype. This feedback resulted in changes being made to the implemented UI (Figure 6.2) from what was initially designed (Figure 5.6). Some of the key differences between the initial UI design and the final implementation are discussed below.

The buttons on the bottom right of the UI design shown in Figure 5.6 were replaced with a tab control, since the latter provides a more intuitive form of navigation between the two visualisations. The filtering options on the right-hand side were changed slightly (Figure 6.2). Based on feedback from potential users of WebPatterns, it was also decided not to allow the user to exclude areas of the web site from analysis, since this implied re-rendering of the radial tree each time which changed the position of the nodes and did not provide a consistent representation of the web site.

As can be seen in Figure 5.6, the location of the textual results was intended to be below the graphical view. In order to provide a larger area for the visualisation, the textual results were moved to the right, below the filtering options, as shown in Figure 6.2.
Figure 6.2: UI of WebPatterns showing association rules visualisation
6.6.2. Static structure of web site

In order to view the web usage patterns provided by each algorithm in relation to the information architecture of the web site, a static structure diagram of the web site had to be created. According to the design specification in Chapter 5, this structure diagram is represented in the form of a radial tree. FlowChart.Net Pro (MindFusion Ltd 2005) allows hierarchical trees to be drawn manually or programmatically using boxes for nodes and arrows for links. This hierarchical tree is then converted to a radial tree representation. This is the process that was used to create the structure diagram for the CS&IS web site. The initial static structure of the web site shown in Figure 6.3 was derived from Figure 2.2 and is discussed in Section 5.5. As discussed in Section 5.5.1, the home page is positioned in the centre of the radial tree and each successive descendant falls on concentric rings spanning out from the centre.

![Figure 6.3: Static structure of CS&IS web site](image-url)
FlowChart.Net Pro allows the radial tree diagram to be exported to a (.xml) file and stored using the `XmlWriter.Write` method. This implies that when the structure diagram is required for a visualisation, it simply needs to be loaded from the file using the `XmlReader.Read` method, instead of being redrawn each time.

### 6.6.3. Association Rules

The results obtained from the Apriori algorithm consists of an array containing the largest set of web pages that are accessed together at least as often as the minimum threshold (Section 6.5.1). In some cases, this means that a single set of web pages is returned. However, the results obtained at the various stages of the algorithm could also provide interesting information to the user. These stages are the combinations of two pages, combinations of three pages and so on, as shown in Figure 6.2. For this reason, it was decided to visualise the results of each of these stages as well. The user can highlight a desired view by clicking on it and the corresponding textual results are then displayed in the textual view.

In order to avoid the visualisation becoming too cluttered, the number of associations that can be displayed concurrently has been limited to five. The user can then scroll through the rest of the associations by using the `Next 5` and `Previous 5` buttons. These buttons are located below the textual results in Figure 6.2. Each of the associations is colour coded, using a rainbow colour scale (Ed Huai-hsin Chi 1999), in order to efficiently map the textual results to those being visualised. This colour scale uses red to indicate the highest frequency and blue to indicate the lowest frequency currently being visualised.

Depending on the results, some web pages may occur in more than one association. However, a limitation of FlowChart.Net Pro is that any node on the graph can only be coloured with a single colour. This does not allow multiple associations to be represented for a single web page. In order to overcome this problem, multiple circles were drawn over each other, with decreasing size of the circle representing decreasing frequency as illustrated in Figure 6.4. In this figure, the node numbered “1” can be seen to have three colours, namely, red to indicate its association with node number “3”, yellow to indicate its association with node number “5” and lastly cyan to indicate its association with node number “9”. The outermost colour (red) represents the
association with the highest frequency, while the innermost colour (cyan) represents the association with the lowest frequency.

![Multicoloured nodes indicating multiple associations](image)

**Figure 6.4: Multicoloured nodes indicating multiple associations**

### 6.6.4. Sequence Analysis

Visualisation of the sequence analysis results required indicating not only which web pages occur on the path, but also the order in which the web pages were accessed. In order to show which web pages occur in a specific path the relevant nodes are coloured according to the rainbow colour scale used in visualising the association rules (Section 6.6.3). Due to the limitation of being able to colour a specific node using only one colour, only the target web page for each path is coloured. In the case where one web page occurs as the target page for more than one path, the same technique of using multiple circles, as discussed in Section 6.6.2, is used to show these results. This is shown in Figure 6.5 where two paths have web page number “3” as their target page. In this case the most frequent path is indicated by the outermost colour, while the innermost colour represents the less frequent of the two paths.
Figure 6.5: Multiple paths to a single web page

In order to indicate the sequence in which the web pages were accessed, coloured arrows are used. Another limitation of FlowChart.Net Pro is that once an arrow is drawn over a specific link, any subsequent arrows drawn over the same link simply overlap each other, thereby concealing the previously drawn arrows. In order to show more than one path across a specific link, Bezier curves (Sederberg and Farouki 1992) were used as shown in Figure 6.5. Using these curves prevents any arrows from overlapping one another.

When the user interacts with this visualisation, he/she has the option of viewing all the frequent paths to any web page, as shown in Figure 6.6. Alternatively, a specific target page can be selected from the URL list in order to view the frequent paths to that page, as is illustrated in Figure 6.7. This feature was not envisioned during the design phase and was implemented based on suggestions from potential users, which were obtained during user testing. As with the association rules visualisation, only five paths are displayed at a time, and these are also colour coded with the results in the textual view. Should the user wish to view more results, the Next 5 or Previous 5 buttons can be used (Figure 6.6).
Figure 6.6: UI of WebPatterns showing sequence analysis visualisation of all paths to all web pages
Figure 6.7: UI of WebPatterns showing sequence analysis visualisation of all paths to the selected page
6.7. Data Manipulation

The filtering view remains consistent throughout the interaction with WebPatterns. The three filtering options provided are:

1. Period selection,
2. URL selection, and
3. Threshold selection.

The period selection shown in Figure 6.8 allows the user to specify the date range for the data to be analysed. Once the date range is selected, the desired URL needs to be selected from the URL List shown in Figure 6.9. The user can also reduce the number of results returned, by specifying the minimum threshold (also illustrated in Figure 6.9). The threshold value is used by the WUM algorithms in order to eliminate all results that have a frequency less than the specified value.
Depending on which WUM algorithm is being visualised, the URL list has two possible purposes. During the association analysis, a web page is selected from the URL list in order to determine the pages with which it is associated. This is done by highlighting the URL in the list. During sequence analysis, the user has the option of viewing the most frequent paths to All web pages or viewing the most frequent paths to a selected target page. These two options are available by selecting the Show All button, or by selecting the desired target page from the URL list and selecting the Show Selected button.

6.8. Conclusion

This chapter demonstrated that a visual web mining prototype, namely WebPatterns, could be effectively implemented based on the model proposed in Chapter 5.

The required data warehouse in the data layer was created using Oracle on a Sun server. The web mining server of the application layer containing the WUM algorithms is also stored on this server. The WUM algorithms were implemented using Java. FlowChart.NET Pro was identified as the most appropriate implementation tool to visualise the results of the WUM algorithms using a radial tree layout. Since FlowChart.Net Pro is a Microsoft.Net component, C#.NET was used to implement the UI in the presentation layer.

Iterative user feedback during the implementation stage, as well as limitations of the visualisation tool selected, resulted in several changes being made to the UI (Section 6.6). Key changes that were implemented include, changing the interaction technique used to select the desired visualisation; moving the textual view to the right-hand side to maximise the graphical view area; and changing the filtering view to include an URL list.

The following chapter discusses the evaluation and testing of WebPatterns and the results of this evaluation. The results of this evaluation will be used to confirm the effectiveness and usefulness of WebPatterns.
Chapter 7 – Evaluation

7.1. Introduction

This chapter aims to answer the seventh research question in Table 1.1 regarding evaluating the usefulness of WebPatterns. This in turn will allow the effectiveness of the model proposed in Chapter 5 to be confirmed. The usefulness can be determined from the perceived ease of use and usability of WebPatterns. These measures are derived from the effectiveness and satisfaction of WebPatterns. This chapter documents the evaluation process, including the procedure followed (Section 7.2.1), metrics used (Section 7.2.2) and instruments required (Section 7.2.3). Once the results have been analysed and interpreted, appropriate conclusions are drawn.

7.2. Evaluation Method

In order to conduct the evaluation of WebPatterns, and thereby confirm the effectiveness of the model discussed in Chapter 5, certain steps need to be followed. This section discusses the instruments used, the procedures followed and the metrics gathered from the evaluation procedure. The evaluation instruments include the task list and the user satisfaction questionnaire used. The evaluation procedure discusses the evaluation environment and hardware equipment required as well as the procedures followed. The evaluation metrics include the results obtained from the evaluation which will be used to make conclusions regarding the usefulness of WebPatterns.
7.2.1. Evaluation Procedure

Before the evaluation can be performed, the correct participants need to be selected. It is important that these participants have the correct experience, as well as knowledge of the problem domain. WebPatterns is a WUM tool which applies WUM algorithms to web usage data to extract web usage patterns and then visualise these patterns. For this reason, the participants are required to be knowledgeable in web site design and management. It is also of added value if the participants have used existing WUM tools previously. For this evaluation process, a recommended minimum of eight participants is required (Rubin 1994).

Once suitable participants were selected, an evaluation task list (See Appendix B) was designed (Section 7.2.3.1). This was completed by each participant, with the purpose of evaluating the effectiveness of WebPatterns. The second part of the evaluation required a user satisfaction questionnaire to be designed (Section 7.2.3.2). This was completed by the participants once they had completed the evaluation task list.

In order to allow the participants to concentrate on the specific tasks, the evaluation was conducted in a usability lab. The equipment that was used in the evaluation comprised a single computer with WebPatterns installed on it. The participants interacted with WebPatterns using a keyboard and mouse. The monitor resolution was set to 1280x1024 in order to provide the maximum effect of the interface.

The author assumed the role of test administrator for the purposes of this evaluation. During the evaluation, the test administrator sat next to the participants in order to make observations, record any comments and provide assistance if necessary. Comments and assistance provided were also written down. In order to encourage the participants to make note of any criticism, they were allowed to write these down, should they have felt uncomfortable discussing them with the test administrator.

The procedure that was followed for each evaluation is given below:

- Upon arrival for the evaluation, the participant was greeted by the test administrator.
- The participant was seated at the computer and the test administrator proceeded to give an overview of the research and WebPatterns.
• The evaluation procedure was explained to the participant, as well as a brief explanation of what exactly was required.
• The participant then commenced the evaluation with the test administrator making observations.
• Once the participant had completed the task list, the participant proceeded to complete the satisfaction questionnaire.
• An informal discussion followed where the participant was allowed to make any comments and the test administrator mentioned any observations made.
• The participant was then thanked for participating in the evaluation.

The evaluation results, as well as any recorded comments or assistance provided were then captured in a spreadsheet.

7.2.2. Evaluation Metrics

Three metrics were identified for the evaluation of WebPatterns. The first two were obtained directly from the evaluation, namely effectiveness and satisfaction. The third, namely the usefulness of WebPatterns, was derived by combining the first two metrics (Davis 1989) and was used to confirm the effectiveness of the model.

7.2.2.1. Effectiveness

Effectiveness is a measure based on the accuracy and completeness with which a user completes a specific task using WebPatterns. Various measures of effectiveness include percentage task completion, frequency of errors, frequency of assists given to the participant by the test administrator and frequency of accesses to help or documentation (Rubin 1994).

Completion rate is the percentage of participants who completely and correctly answered a specific task. The frequency of errors refers to the number of tasks a participant was unable to correctly complete. Assists refer to tasks where the test administrator had to intervene in order to allow the test to proceed. Since assists are
recorded during the evaluation, they need to be taken into consideration when analysing task completion rates.

The combination of these measures was used to achieve an overall effectiveness value for WebPatterns. From this, any problems with WebPatterns that would prevent a user from being able to complete a specific task, were identified.

### 7.2.2.2. Satisfaction

Satisfaction refers to a user’s subjective response when using WebPatterns (Rubin 1994). A participant’s satisfaction with WebPatterns, could be suggestive of its potential future use. Questionnaires commonly used to measure user satisfaction include QUIS (Chin, Diehl and Norman 1988) and SUMI (Kirakowski 1996) and make use of scales such as Likert (Usability by Design 2004). These questionnaires obtain measures of satisfaction, usefulness and ease of use. The results were used to determine the overall satisfaction of participants when using WebPatterns.

### 7.2.2.3. Usefulness

The usefulness of WebPatterns would provide insight into whether participants would use WebPatterns to support their desired tasks. According to Davis (1989), usefulness is influenced by the perceived ease of use of WebPatterns. The ease of use of WebPatterns was obtained from the effectiveness and satisfaction measured during the evaluation. If participants could effectively and satisfactorily complete their tasks using WebPatterns, then their perceived ease of use would be high. Comments and feedback by the participants with respect to the use of WebPatterns, either during or after the evaluation, could also add to the measure of usefulness. Positive feedback could be an indication that WebPatterns would be used as a WUM tool in future.

### 7.2.3. Evaluation Instruments

In order to perform the user testing, two evaluation instruments were required namely, the task list and the user satisfaction questionnaire. The task list was used to measure the effectiveness of WebPatterns, while the user satisfaction questionnaire was used to
determine the satisfaction of the participants in using WebPatterns to perform specific tasks.

7.2.3.1. Task List

The task list (Appendix B) was created in such a way that the participants were not only required to determine the results of the WUM algorithms, but where applicable, to also give their interpretation of the results. A summary of the tasks is provided in Table 7.1. The participants were also required, in some cases, to indicate whether they used the graphical view or the textual view to obtain the results.

<table>
<thead>
<tr>
<th>Task ID</th>
<th>Description</th>
<th>Visualisation Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>Determine the most frequently visited paths to ALL of the web pages.</td>
<td>![Image]</td>
</tr>
<tr>
<td>Task 2</td>
<td>Determine the most frequently visited paths to a specific web page.</td>
<td>![Image]</td>
</tr>
<tr>
<td>Task 3</td>
<td>Determine which web pages are associated with a given web page.</td>
<td>![Image]</td>
</tr>
<tr>
<td>Task 4</td>
<td>Interpret a given association by using the structure diagram.</td>
<td>![Image]</td>
</tr>
</tbody>
</table>

Table 7.1: Summary of tasks
Chapter 7 – Evaluation

The answers of each task were evaluated in order to determine the overall effectiveness with which the tasks were performed using WebPatterns. The participants were also required to indicate whether they used the graphical view or the textual view to obtain the results. The usefulness of each view in assisting participants in performing the tasks was then determined.

7.2.3.2. User Satisfaction Questionnaire

The user satisfaction questionnaire (Appendix C) was compiled to allow the participants to indicate the satisfaction with which they were able to complete the tasks in the task list as well as their satisfaction with the results provided by WebPatterns. Due to the fact that web usage data is unique to WUM, and since the effectiveness of each visualisation also needed to be determined, standard satisfaction questionnaires could not be used.

Visual WUM makes use of visualisations of the results of WUM algorithms to make these results easier for the user to interpret. The problem with using relatively new visualisation techniques in visual WUM tools, is that there is no standard questionnaire for evaluating the effectiveness of the visualisation and the WUM tool itself (Grinstein, Hoffman, Laskowski and Pickett 2001). A framework for evaluating such visual mining tools was proposed by (Marghescu, Rajanen and Back 2004). This framework recognises quality of use as a key concept to evaluate WUM tools. Quality of use includes:

- Quality of interaction;
- Quality of visualisation; and
- Quality of information.

The user satisfaction questionnaire was then compiled based on these measures of quality of interaction, visualisation and information. This questionnaire was completed by the participants once they had completed the task list. The results of this questionnaire were then used to determine the quality of use of WebPatterns.
7.3. Results and Analysis

Once the user evaluation was performed, the results were analysed and interpreted. This section presents the results and discusses how they were interpreted.

7.3.1. Demographics

The demographic profile of the participants is shown in Table 7.2. The eight participants were selected based on the user profile provided in Section 7.2.1. Due to the lack of female web masters, only males were included in the evaluation. This is, however, in line with current industry demographics, which indicate that men make up the majority of the workforce in the IT industry (ITAA 2003). While 50% of the participants had more than six years experience as web masters, the other half of the participants had some experience in web design and WUM.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Years Experience</th>
<th>Web Design Software Used</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 7.2: Demographic profile of participants

7.3.2. Effectiveness Results

The effectiveness of WebPatterns was determined by means of the task completion rate of the task list. The task list required the participants to use both the visualisation techniques namely, association rules and sequence analysis. Using each visualisation the participants were required to perform a series of tasks and record their answers. These tasks represented were common to WUM. In cases where results were obtained from the UI, the participants were requested to indicate whether they used the graphical view or the textual view to obtain the answer.

From the results of the task list, it was evident that task 1 (Table 7.1) accounted for 31% of the task list, task 2 accounted for 46% and task 3 accounted for 23%. These
proportions were used to determine the weighted task completion rate for each participant as follows:

Total completion rate for a participant (\(\%\)) = \(T_1 \times 0.31 + T_2 \times 0.46 + T_3 \times 0.23\)

where \(T_i\) = percentage of task completion for a participant for task \(i\) where \(i = 1,2,3\).

The overall task completion rate was very high (Table 7.3), with all participants completing all tasks correctly, except for participant 4 who obtained a 76.92\% overall task completion rate. Participant 1 required some assistance, but only as far as providing an explanation for certain terminology.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Percentage (% of Task Completion)</th>
<th>Total % of Task Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1 (31%)</td>
<td>T2 (46%)</td>
</tr>
<tr>
<td>1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>75%</td>
<td>100%</td>
</tr>
<tr>
<td>5</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>6</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>7</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>8</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Median % of Task Completion</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Mean % of Task Completion</td>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 7.3: Percentage of task completion

From further investigation of the task results, it was evident that 62.5\% of the participants obtained their answers for Task 1 by using the graphical view, while the rest used the textual view only to obtain the results.

For Task 2, 56.25\% of the participants used the graphical view or a combination of the both views to obtain the results, while 43.75\% used the textual view only. The reason for some participants choosing the textual view, was attributed to the fact that although the frequencies of paths was obtainable from tooltips on the paths, the users were not familiar with this technique and thus preferred the textual view.

Task 3 provided similar results with 50\% of the participants using the graphical view or a combination of the two views, and 50\% using the textual view to obtain the results.
Although associated page frequencies were available from tooltips on the associated pages, some of the participants elected to use the textual view instead.

The final question of the task list was intended to encourage the participants to make interpretations from the static web site structure diagram. When asked whether they thought there were sufficient links provided between certain associated pages, 75% of the participants indicated that more links between these associated pages should be provided than were currently available. Their motivation for this was based on the fact that these pages contain related content and links between them should therefore be provided. These suggestions are discussed further in Section 8.4.

Figure 7.1 provides a summary of the mean and median completion rates. The overall mean completion rate was 97% and the overall median completion rate was 100%. This provides a strong indication that the participants could correctly complete the majority of the required tasks using WebPatterns. Individual evaluation results are provided in Appendix E.
7.3.3. User Satisfaction Results

User satisfaction was obtained by means of a satisfaction questionnaire which was completed once the task list was completed. The questionnaire (Appendix C) consisted of five sections, namely:

1. Overall reaction to WebPatterns
2. Overall quality of interaction
3. Overall quality of visualisation
4. Quality of individual visualisations
5. Quality of information for each visualisation

Each of these sections consisted of questions which the participants were required to rate using a five-point Likert scale (Usability by Design 2004). This scale uses a rating of 1 through 5 to indicate the participants’ satisfaction for specific criteria. A rating of 1 indicates strongly disagree while a rating of 5 indicates strongly agree. The median and standard deviation for each question for all participants was then calculated in order to determine the overall rating of satisfaction.

Figure 7.2 provides a summary of the mean results of the satisfaction questionnaire. The complete set of results is provided in Appendix E. From the results shown in Figure 7.2, it is evident that the participants were highly satisfied with WebPatterns (mean rating $\geq 4.4$ out of 5).

![Figure 7.2: User satisfaction results](image)

84
Table 7.4 shows the satisfaction results for each visualisation. From these results it is evident that there was almost no difference in satisfaction between the two visualisations. The overall median for quality of visualisation and quality of information was >= 4.5, indicating that the participants were highly satisfied with WebPatterns.

<table>
<thead>
<tr>
<th>Part 3b: Quality of Visualisations</th>
<th>Associated Pages</th>
<th>Frequent Paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>Std Dev</td>
<td>Median</td>
</tr>
<tr>
<td>3.6. The layout of the screen was effective.</td>
<td>5.00</td>
<td>0.46</td>
</tr>
<tr>
<td>3.7. The size of the visualisation was adequate.</td>
<td>4.50</td>
<td>0.74</td>
</tr>
<tr>
<td>3.8. Locating information elements on the screen was easy.</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>3.9. Description of the data is sufficient.</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>3.10. The use of colours in the visualisation was helpful.</td>
<td>5.00</td>
<td>0.52</td>
</tr>
<tr>
<td>3.11. Terminology related to the tasks was useful.</td>
<td>4.00</td>
<td>0.52</td>
</tr>
<tr>
<td><strong>Median for each visualisation</strong></td>
<td><strong>4.50</strong></td>
<td><strong>5.00</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 4: Quality of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
</tr>
<tr>
<td>4.1. The information is presented in a clear and understandable way.</td>
</tr>
<tr>
<td>4.2. I am satisfied with the usefulness of the information obtained.</td>
</tr>
<tr>
<td>4.3. The information obtained is easy to interpret.</td>
</tr>
<tr>
<td>4.4. I think that the information I obtained is interesting.</td>
</tr>
<tr>
<td>4.5. The information I obtained is new for me.</td>
</tr>
<tr>
<td>4.6. I obtained the information that I needed for the task.</td>
</tr>
<tr>
<td>4.7. I am satisfied with the information provided by the tool.</td>
</tr>
<tr>
<td>4.8. The quality of information is good.</td>
</tr>
<tr>
<td><strong>Median for each visualisation</strong></td>
</tr>
</tbody>
</table>

Table 7.4: Satisfaction results for quality of visualisation and information

7.3.4. Usefulness

As discussed in Section 7.2.2.3, usefulness is influenced by the perceived ease of use of WebPatterns and can be determined using a combination of effectiveness and
satisfaction. Based on the results provided in Sections 7.3.2 and 7.3.3, it can be deduced that WebPatterns is a useful WUM tool.

In order to obtain more insight into the usefulness of WebPatterns, participants were asked if they would use WebPatterns to assist in similar tasks and to motivate their answers. All the participants indicated that they would like to use WebPatterns in the future. The reasons they provided included:

- “Would be useful for both intranet and public NMMU web sites. Provides additional information that other log analysing tools do not provide. I would use it in conjunction with these.”
- “It has an enjoyable interface.”
- “It would be interesting to see results obtained from web sites that I webmaster.”
- “It provides useful and interesting results.”
- “The information is very useful for web developers.”
- “It provides interesting and potentially useful information.”

It is evident from these comments that the participants found WebPatterns easy to use and thought the information and visualisations provided were useful. They were eager to view results for web sites they had developed or ones that they manage.

Although the participants did not experience difficulties in using WebPatterns, a comment was made in relation to its learnability and the radial tree metaphor used by one participant. This comment was:

- “There is a cognitive disconnect between the physical website and the logical visualisation. So one needs to become familiar with the more abstract way of looking at web usage data.”

This comment was not surprising, since the radial tree representation used in WebPatterns is intended to be used by web masters who are familiar with the web site they are analysing. They are also expected to learn all aspects of this tool before using it to perform their daily tasks, thus they will be “familiar with the more abstract way of looking at web usage data”. This will allow them to fully understand the results provided by WebPatterns and take full advantage of its features.
7.4. Conclusion

WebPatterns was evaluated to confirm the effectiveness of the model proposed in Chapter 5. The evaluation procedure (Section 7.2.1), the metrics used in the analysis (Section 7.2.2) and the evaluation instruments required (Section 7.2.3) were documented. The metrics used were effectiveness and user satisfaction.

The evaluation instruments, comprising the task list and the satisfaction questionnaire, were designed in order to obtain measures for the identified metrics. Based on the results of the evaluation of WebPatterns (Section 7.3), it was concluded that the participants perceived WebPatterns to be a very useful and effective WUM tool for analysing web usage patterns. This perceived usefulness enabled the effectiveness of the proposed model to be confirmed.

The following chapter documents the conclusions of this research and provides recommendations for future work.
Chapter 8 – Conclusions and Recommendations

8.1. Introduction

The goal of this chapter is to conclude this dissertation by reviewing the research conducted (Section 8.2) and highlighting the practical and theoretical implications of this research by providing a discussion for future research (Section 8.3). The objective of this research was to develop a model to analyse and view the web usage of an organisational web site. A further goal was to demonstrate that a prototype could be developed from the proposed model. The prototype was implemented to analyse the web usage of the CS&IS Department at the NMMU and visualise the results.

8.2. Research Achievements

In order to achieve the objectives of this research, certain research questions were formulated (Section 1.4). The research achievements are divided into two sections, namely the theoretical achievements which include the proposed model, and the practical achievements which include the implementation of a prototype. The theoretical achievements address research questions 1 to 5 and the practical achievements address research questions 6 to 8. The research questions associated with each achievement are listed in Table 8.1.

The dissertation was structured so as to answer these research questions using the research methods proposed in Table 1.1.
Chapter 8 – Conclusions and Recommendations

<table>
<thead>
<tr>
<th>Theoretical Achievements: Proposed Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is an organisational web site?</td>
</tr>
<tr>
<td>2. What is web usage mining?</td>
</tr>
<tr>
<td>3. What are web usage patterns?</td>
</tr>
<tr>
<td>4. Are there any related systems?</td>
</tr>
<tr>
<td>5. How will the model be designed?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practical Achievements: Implementation of Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. How will a prototype for the model be implemented?</td>
</tr>
<tr>
<td>7. How effective is the model?</td>
</tr>
<tr>
<td>8. What conclusions can be made from this research?</td>
</tr>
</tbody>
</table>

Table 8.1: Research achievements and corresponding research questions

8.2.1. Theoretical Achievements

The theoretical achievement of this research was the development of a model for the mining and visualisation of web usage data (Figure 8.1). In order to propose the model, the structure of the web site as well as the log files being analysed had to be determined (Chapter 2). An investigation into WUM together with the phases involved was conducted (Chapter 3). This information was then used to propose the model (Chapter 5).

The CS&IS web site was identified as an example of an organisational web site as it provides its users with a large quantity of information. The structure of the CS&IS web site was analysed (Section 2.2.2) and the log files recorded by the web server were examined (Section 2.2.3). The extended log format records information regarding each URL accessed on the web site. The fields identified as being necessary for WUM were: date, time, client ip address, username URL and client agent.

An investigation into WUM was conducted. This included identifying the various phased involved in analysing web usage data. These phases are preprocessing, pattern discovery and pattern analysis (Section 3.2). A single layer data architecture was selected, while a snowflake schema was chosen to model the data in the data warehouse. From the various pattern discovery techniques that were investigated, three techniques were identified as being appropriate to web usage data, namely sequential patterns, association rules and cluster analysis (Section 3.3.2).
There are several WUM algorithms that can be used for each of the three pattern discovery techniques identified (Section 3.4). Based on the web usage data being analysed, the most appropriate WUM algorithms were selected to be: Apriori for association rules, automatic cluster detection for cluster analysis and the time ordered sequences available in the data warehouse for sequence analysis.

The final phase of WUM is pattern analysis. In order to assist the user in interpreting the patterns discovered, pattern analysis makes use of visualisation techniques which show the web usage patterns as well as the structure of the web site (Section 3.2.3.1). A radial tree representation for the static structure of the web site was selected to be the most appropriate visualisation technique. Colour and other visual cues are used to show the results of the WUM algorithms.

In order to determine how the three phases of WUM are incorporated, several related systems were investigated (Chapter 4) in terms of the data storage, the WUM algorithms and the visualisation of the results. Several limitations were identified for all three systems investigated. These limitations included that it was not clear which WUM
algorithms were being used to analyse the data. There was also no data model and the analysis appeared to be performed directly from the log files. In terms of visualisation of the results, only simple bar and line graphs were used and no web site structure was provided to assist with interpretation of the results. It was concluded that the three systems investigated would not be able to satisfy the goals of this research and that a new WUM model needed to be proposed.

The literature review in Chapter 2 and 3, together with the investigation into several related systems in Chapter 4 were used to propose the model in Chapter 5. The architecture of the proposed model (Section 5.2) consists of three layers which were based on the three phases of WUM. These three layers are the data layer, application layer and the presentation layer. The data layer contains the data warehouse; the application layer contains the WUM algorithms; and the presentation layer contains the visualisation techniques and the user interface. The WUM algorithm design focused on determining the format of the output so as to assist with the visualisation of the results (Section 5.4). An iterative design approach was used for the visualisation design (Section 5.5) and the UI design (Section 5.6). This facilitated the development of low-level prototypes which incorporated the visualisation techniques proposed. Based on feedback from the users, high-level designs were produced.

8.2.2. Practical Achievements

The practical achievement of this research was the implementation of the prototype (Chapter 6), called WebPatterns, which allowed the effectiveness of the model to be demonstrated. The data warehouse of the data layer and the web mining server of the application layer were implemented using Oracle on a Sun server (Section 6.2). A list of requirements was then compiled for the presentation layer. Based on the data manipulation and data representation requirements identified, FlowChart.Net Pro was selected as the most appropriate implementation tool for the proposed visualisation techniques. C#.NET was therefore selected as the implementation tool for the implementation of the presentation layer.

The UI is divided into three main views (Figure 8.2), namely the graphical view (bordered in red), the filtering view (bordered in blue) and the textual view (bordered in green).
The implementation of the WUM algorithms (Section 6.5) was discussed along with the implementation of the appropriate visualisation techniques (6.6). The three coordinated views (Figure 8.2) of the UI were also discussed. Key changes that were made to the implementation as a result of iterative user testing were also documented. These included changing the interaction technique used to select the desired visualisation, moving the textual view to the right-hand side to maximise the graphical view area, and changing the filtering view to include an URL list.

To confirm the effectiveness and usefulness of WebPatterns, the prototype was evaluated by means of user testing (Chapter 7). A task list was used to determine the effectiveness of WebPatterns (Section 7.2.3.1) and an evaluation questionnaire was used to measure the user satisfaction (Section 7.2.3.2). The combined results of the effectiveness and satisfaction were then used to determine the perceived usefulness of WebPatterns.

The mean task completion rate for the task list was 97% (Table 7.3), and the evaluation questionnaire indicated a satisfaction rating of above 4.4 out of 5 (Figure 7.2). This
indicates that the participants were highly satisfied with WebPatterns. These evaluation results, when combined with the opinions of the participants, indicate that WebPatterns could be an effective and useful WUM tool for analysing web site usage.

8.3. Implications and Limitations

The theoretical implication of this research is that a WUM model was proposed for WUM of organisational web sites. This model can be applied to any web site provided the necessary information is recorded in the log files.

The practical implication of WebPatterns is that it demonstrated the effectiveness of the proposed model. The evaluation of WebPatterns indicated that it could be used as a WUM tool to analyse the web usage of organisational web sites. Another practical implication is that different association rule or sequence analysis algorithms can be implemented. The only constraint of the visualisation techniques used is that the output of these algorithms should remain in the same format.

A limitation of WebPatterns is that no facility is provided to automatically render the structure of a particular web site. The initial web site structure needs to be created using FlowChart.Net Pro, before analysis can proceed. Another restriction of WebPatterns is that because the web site structure was converted to a hierarchical structure in order to represent it in the form of a radial tree, some embedded links are not indicated on the structure diagram. For this reason, the user analysing the web site needs to have sufficient knowledge regarding the web site structure of the site being analysed.

8.4. Recommendations

The goals of this research included using the proposed model in order to provide recommendations regarding the information architecture of the CS&IS web site. In order to do this, WebPatterns was used to analyse the web usage data of the CS&IS web site for 2004 and determine if any changes needed to be made to facilitate information retrieval by the user. This analysis is divided into two sections based on the WUM techniques used, namely association rules and sequence analysis.
8.4.1. Association Rules

The purpose of applying the association rules algorithm to the CS&IS web usage data is to determine whether there are sufficient links between web pages that are frequently accessed together. It was assumed that the web pages which are most frequently associated with the home page would be those that are directly linked to the home page, those being the five categories available from the menu (Figure 2.1). It was therefore surprising to see that one of the most associated pages to the home page (node 0) is in fact three levels into the web site, namely the individual course page (node 14 in Figure 8.3). This was evident from the result (0, 14) 4.44%. This does, however, fall within the three-click guideline for access to pages within a site (Section 2.2.1). Upon further investigation it was noted that this web page was accessible by firstly selecting degrees and courses (node 1), followed by selecting undergraduate (node 6) and finally selecting the individual course (node 14). A possible recommendation would be to place a link on the home page directly to the undergraduate web page. This would allow the user to find the course they were looking for in fewer mouse clicks.

8.4.2. Sequence Analysis

The CS&IS web site is relatively highly connected, allowing users to navigate easily through various areas of the web site. For this reason it is difficult to predict usage patterns. It was, however, assumed that in the majority of cases the user would navigate the site using the menu structure provided on the left-hand side of all the web pages. This menu structure categorises the web site into five areas, namely Course information, Staff information, Student information, Research information and Centre of Excellence (Figure 2.1).

By analysing the sequence analysis results (Figure 8.4), it was evident from the most frequent paths accessed, that the individual course page (node 14) occurred in many of these frequently accessed paths. This supports the suggestion in 8.4.1 that a link to the individual course page be placed on the home page for easier access by the users. From the frequent paths identified, it was also evident that users seldom navigate back in order to try alternative paths, for example (0, 1, 14, 15) 0.72%. This is an indication that the current information architecture sufficiently facilitates the location of information by the user. This could also be attributed to the fact that embedded links are also provided between related categories. This allows the users to easily find what they are looking for, without having to navigate back to the home page and select an alternative category.
Figure 8.3: Association rules analysis results
Figure 8.4: Sequence analysis results
8.5. Future Research

This dissertation has identified several possibilities for future research in WUM of organisational web sites. Research that could extend the mining of web usage data and the visualisation of the results are discussed below.

Theoretical future research would be to apply other WUM algorithms to the web usage data. Various WUM visualisation techniques could also be investigated to display the results of these algorithms. By implementing more than one visualisation technique for each of the algorithms, a comparison could be made to determine the most effective visualisation.

Practical future research would be to use WebPatterns to analyse the CS&IS web site once the suggested changes have been made to the information architecture. This would enable the user to determine whether the suggested changes were indeed successful in facilitating easier navigation of the web site.

Further practical research could include using WebPatterns to analyse the new NMMU web site (www.nmmu.ac.za). Since this web site is the combination of several existing organisational web sites it would be interesting to determine how easily users are able to find the desired information.

Additional practical research could be conducted to allow WebPatterns to automatically extract the web site structure, especially for larger web sites. This could be done by using web crawlers to perform a breadth-first search of the web site based on the links provided. This would eliminate the need and possible human error involved in manually creating the web site structure diagram.

8.6. Summary

This dissertation provided an analysis of the various aspects involved in WUM. It also discusses how WUM can be used to improve the information architecture of a web site based on the analysis and visualisation of its usage. This dissertation specified the
design of a WUM model including a data warehouse to store the web usage data, the WUM algorithms to perform the analysis and appropriate visualisation techniques to visualise the results of the algorithms. The development of WebPatterns demonstrated the effectiveness of the proposed model. An evaluation of WebPatterns concluded that the participants could effectively complete the required tasks and that they were highly satisfied with the information and visualisations provided. These results indicate a high level of usefulness, thereby confirming the effectiveness of the proposed model. This dissertation makes a significant contribution towards the field of WUM with the successful proposal of a WUM model and the development of WebPatterns to analyse the web usage data of organisational web sites.
References


BROOKS, P. (1997): Data Mining Today, DBMS.


References


MICROSOFT CORPORATION (2005): C#.NET.


SUN MICROSYSTEMS (2005): Java.


### Appendix A – Comparison of Charting Components

<table>
<thead>
<tr>
<th>Feature</th>
<th>Tool</th>
<th>Nevron Chart</th>
<th>NetCharts Pro</th>
<th>FlowChart.Net Pro</th>
<th>Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language</strong></td>
<td></td>
<td>C#</td>
<td>Java</td>
<td>C#</td>
<td>Java</td>
</tr>
<tr>
<td><strong>Data Representation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radial/Disk Tree</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Data Manipulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Details-on-demand</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Connect to Oracle</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Save and Export Image</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rapid Prototyping</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 6.1: Comparison of charting components (✓ = Supported; ✗ = Not Supported)
Appendix B – Evaluation Task List

Effectiveness of WebPatterns

Using WebPatterns, please complete the following questions:

A. Frequent Paths
1. a. For the period 1 May 2004 – 14 July 2004, with Threshold = 1%, which were the 3 most frequently followed paths to ALL pages? Observe and record these paths.

<table>
<thead>
<tr>
<th>Frequent Paths (Page Numbers)</th>
<th>%</th>
<th>Path Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Which URL occurs as the target URL in the majority of the paths? ____________________________

c. How did you determine this?

<table>
<thead>
<tr>
<th>Answer</th>
<th>(√)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent paths list</td>
<td></td>
</tr>
<tr>
<td>Observed in graph</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

If other, please explain: ____________________________________________________________

2. a. For the period 1 January 2004 – 30 April 2004, and for the target URL 5 - /coe_web/coehomepage.asp, with Threshold = 0%, which were the 3 most frequently followed paths? Observe and record these paths.

<table>
<thead>
<tr>
<th>Frequent Paths (Page Numbers)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

b. Which was the longest path used to reach the ‘5 - /coe_web/coehomepage.asp’ page? Indicate only the frequent path page numbers. ___________________________________

c. How did you determine this?

<table>
<thead>
<tr>
<th>Answer</th>
<th>(√)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent paths list</td>
<td></td>
</tr>
<tr>
<td>Observed in graph</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

If other, please explain: ____________________________________________________________

d. What is the shortest path available to find ‘5 - /coe_web/coehomepage.asp’ page from ‘0 - /webpages/homepage.asp’? Indicate only the path’s page numbers. ____________________________
e. How did you determine this?

<table>
<thead>
<tr>
<th>Answer</th>
<th>(√)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent paths list</td>
<td></td>
</tr>
<tr>
<td>Observed in graph</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

If other, please explain: ________________________________________________________________

f. Was the most frequently used path to the ‘5 - /coe_web/coehomepage.asp’ page identified in (a) above, the shortest path available, in terms of number of nodes followed? ______

B. Associated Pages

3. a. For the period 1 January 2004 – 31 December 2004, with Threshold = 2%, which pages were most frequently associated with the ‘3 - /webpages/studentpage.asp’ page? Observe and record these associated page numbers where applicable.

(N = Number of Pages)

<table>
<thead>
<tr>
<th>N</th>
<th>Associated Pages (Page Numbers)</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. How did you determine these pages were associated?

<table>
<thead>
<tr>
<th>Answer</th>
<th>(√)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated Pages list</td>
<td></td>
</tr>
<tr>
<td>Observed in graph</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

If other, please explain: ________________________________________________________________

4. Without using the system, consider that the following pages are associated when answering a) and b) below: (2, 3, 4)

a. Would you expect these pages to be associated together? (Y/N) ________________

b. If Yes, why?

<table>
<thead>
<tr>
<th>Answer</th>
<th>(√)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar Name</td>
<td></td>
</tr>
<tr>
<td>They occur on the same path</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

If other, please explain: ________________________________________________________________
c. Do you feel there are sufficient links between these pages? (Y/N) _______________

If not, which additional links would you like to see?
Please record 3 of these links if applicable:

<table>
<thead>
<tr>
<th>Linked Pages (Page Numbers)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C – Evaluation Questionnaire

WebPatterns Evaluation:  
Post-Test Questionnaire

**Questionnaire No:**  
**Date:**  

**Instructions:**  
Please enter your personal details below. Please complete this questionnaire by circling the numbers which most appropriately reflect your impressions about using this computer system.  
Not applicable = NA

<table>
<thead>
<tr>
<th>Full Name:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Tel No:</td>
<td></td>
</tr>
<tr>
<td>Home Language:</td>
<td></td>
</tr>
<tr>
<td>Gender:</td>
<td>□ Male</td>
</tr>
<tr>
<td>Age:</td>
<td>□ 20-30</td>
</tr>
<tr>
<td>Occupation:</td>
<td></td>
</tr>
<tr>
<td>Years Web Experience:</td>
<td></td>
</tr>
<tr>
<td>Web Design/Web Master Experience:</td>
<td></td>
</tr>
<tr>
<td>Yes (List which tools):</td>
<td></td>
</tr>
<tr>
<td>No:</td>
<td></td>
</tr>
</tbody>
</table>
### 1. Overall:

<table>
<thead>
<tr>
<th></th>
<th>strongly disagree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overall, the system was easy to use.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1.2. Overall, I could effectively complete all the tasks.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1.3. Overall, I could efficiently complete all the tasks.</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

### 2. Quality of interaction: (Overall)

<table>
<thead>
<tr>
<th></th>
<th>strongly disagree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1. Overall, I was satisfied with the quality of the interaction.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.2. WebPatterns is easy to interact with.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.3. I learned to use the system with ease.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.4. The system responds quickly to my inputs.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.5. The number of errors was minimal.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.6. The error messages provided were clear and useful.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.7. WebPatterns provides sufficient interaction methods.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.8. The filtering menu was easy to use.</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

### 3a. Quality of Visualisation: (Overall)

<table>
<thead>
<tr>
<th></th>
<th>strongly disagree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. Overall, the layout of the screens was effective.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.2. Overall, the size of the visualisations was adequate.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.3. Overall, the use of colours was effective.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.4. Overall, interpreting the visualisations was easy.</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.5. Overall, the organisation of the information of the visualisation is good.</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
## 3b. Quality of visualization:

<table>
<thead>
<tr>
<th>Description</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6. The layout of the screen was effective.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>3.7. The size of the visualisation was adequate.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>3.8. Locating information elements on the screen was easy.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>3.9. Description of the data is sufficient.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>3.10. The use of colours in the visualisation was helpful.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>3.11. Terminology related to the tasks was useful.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
</tbody>
</table>

## 4. Quality of information:

<table>
<thead>
<tr>
<th>Description</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. The information is presented in a clear and understandable way.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.2. I am satisfied with the usefulness of the information obtained.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.3. The information obtained is easy to interpret.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.4. I think that the information I obtained is interesting.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.5. The information I obtained is new for me.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.6. I obtained the information that I needed for the task.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.7. I am satisfied with the information provided by the tool.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
<tr>
<td>4.8. The quality of information is good.</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
<td>1 2 3 4 5 NA</td>
</tr>
</tbody>
</table>
5. Comments

5.1. Would you use WebPatterns again? Y/N______________________________________________

Please explain your answer above.

___________________________________________________________________________________
___________________________________________________________________________________
___________________________________________________________________________________
___________________________________________________________________________________

5.2. Any other comments

___________________________________________________________________________________
___________________________________________________________________________________
___________________________________________________________________________________
___________________________________________________________________________________
___________________________________________________________________________________
___________________________________________________________________________________

Thank you very much!
## Appendix D – Task List Results

<table>
<thead>
<tr>
<th>Question No</th>
<th>Participant 1</th>
<th>Participant 2</th>
<th>Participant 3</th>
<th>Participant 4</th>
<th>Participant 5</th>
<th>Participant 6</th>
<th>Participant 7</th>
<th>Participant 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 a 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 a 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 a 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 b</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Section</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>75%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2 a 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 a 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 a 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 b</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 d</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 f</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Section</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>3 a 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3 a 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3 a 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Section</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>33%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>77%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
## Appendix E – Evaluation Questionnaire Results

<table>
<thead>
<tr>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part 1: Overall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 Overall, the system was easy to use</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>1.2 Overall, I could effectively complete all the tasks</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.52</td>
</tr>
<tr>
<td>1.3 Overall, I could efficiently complete all the tasks</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>5.00</td>
<td>4.00</td>
<td>4.00</td>
<td>5.00</td>
<td>4.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>0.58</td>
<td>0.00</td>
<td>0.58</td>
<td>0.00</td>
<td>0.58</td>
<td>0.00</td>
<td>0.58</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Part 2: Quality of Interaction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Overall, I was satisfied with the quality of the interaction</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.64</td>
</tr>
<tr>
<td>2.2 WebPatterns is easy to interact with</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.71</td>
</tr>
<tr>
<td>2.3 I learned to use the system with ease</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.71</td>
</tr>
<tr>
<td>2.4 The system responds quickly to my inputs</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5.00</td>
<td>0.52</td>
</tr>
<tr>
<td>2.5 The number of errors was minimal</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.35</td>
</tr>
<tr>
<td>2.6 The error messages provided were clear and useful</td>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>5.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.7 WebPatterns provides sufficient interaction methods</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4.50</td>
<td>0.89</td>
</tr>
<tr>
<td>2.8 The filtering menu was easy to use</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.52</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>5.00</td>
<td>4.00</td>
<td>4.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.00</td>
<td>4.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>0.76</td>
<td>0.82</td>
<td>0.58</td>
<td>0.35</td>
<td>0.38</td>
<td>0.00</td>
<td>0.49</td>
<td>0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Part 3a: Overall Quality of Visualisation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 Overall, the layout of the screens was effective</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5.00</td>
<td>0.41</td>
</tr>
<tr>
<td>3.2 Overall, the size of the visualisations was adequate</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.75</td>
</tr>
<tr>
<td>3.3 Overall, the use of colours was effective</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4.00</td>
<td>0.41</td>
</tr>
<tr>
<td>3.4 Overall, interpreting the visualisations was easy</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.82</td>
</tr>
<tr>
<td>3.5 Overall, the organisation of the information of the visualisation is good</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>5.00</td>
<td>4.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>0.55</td>
<td>0.84</td>
<td>0.89</td>
<td>0.45</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Part 3b: Quality of Association Rules Visualisations

<table>
<thead>
<tr>
<th>Part 3b</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6 The layout of the screen was effective</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5.00</td>
<td>0.46</td>
</tr>
<tr>
<td>3.7 The size of the visualisation was adequate</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.50</td>
<td>0.46</td>
</tr>
<tr>
<td>3.8 Locating information elements of the screen was easy</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>3.9 Description of the data is sufficient</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>3.10 The use of colours in the visualisation was helpful</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>3.11 Terminology related to the task was useful</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.52</td>
</tr>
</tbody>
</table>

**Median** 5.00 4.00 4.50 5.00 4.00 4.50 4.50 4.50

**Standard Deviation** 0.52 0.63 0.55 0.52 0.52 0.55 0.52 0.55

## Part 4: Quality of Association Rules Information

<table>
<thead>
<tr>
<th>Part 4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 The information is presented in a clear and understandable way</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.74</td>
</tr>
<tr>
<td>4.2 I am satisfied with the usefulness of the information obtained</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>4.3 The information obtained is easy to interpret</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.00</td>
<td>0.71</td>
</tr>
<tr>
<td>4.4 I think that the information I obtained is interesting</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.52</td>
</tr>
<tr>
<td>4.5 The information I obtained is new for me</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4.50</td>
<td>0.89</td>
</tr>
<tr>
<td>4.6 I obtained the information that I needed for the task</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.46</td>
</tr>
<tr>
<td>4.7 I am satisfied with the information provided by the tool</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>4.8 The quality of information is good</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5.00</td>
<td>0.52</td>
</tr>
</tbody>
</table>

**Median** 5.00 4.00 4.00 5.00 5.00 4.00 4.50 4.50

**Standard Deviation** 0.35 0.71 0.35 0.35 0.74 0.35 0.71 0.53
### Appendix E – Evaluation Questionnaire Results

#### Part 3b: Quality of Sequence Analysis Visualisations

<table>
<thead>
<tr>
<th>Description</th>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6 The layout of the screen was effective</td>
<td></td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.46</td>
</tr>
<tr>
<td>3.7 The size of the visualisation was adequate</td>
<td></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>0.46</td>
</tr>
<tr>
<td>3.8 Locating information elements of the screen was easy</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>3.9 Description of the data is sufficient</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.53</td>
</tr>
<tr>
<td>3.10 The use of colours in the visualisation was helpful</td>
<td></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.35</td>
</tr>
<tr>
<td>3.11 Terminology related to the task was useful</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.52</td>
</tr>
</tbody>
</table>

#### Part 4: Quality of Sequence Analysis Information

<table>
<thead>
<tr>
<th>Description</th>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 The information is presented in a clear and understandable way</td>
<td></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.35</td>
</tr>
<tr>
<td>4.2 I am satisfied with the usefulness of the information obtained</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.00</td>
<td>0.52</td>
</tr>
<tr>
<td>4.3 The information obtained is easy to interpret</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.50</td>
<td>0.74</td>
</tr>
<tr>
<td>4.4 I think that the information I obtained is interesting</td>
<td></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>4.5 The information I obtained is new for me</td>
<td></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4.50</td>
<td>0.89</td>
</tr>
<tr>
<td>4.6 I obtained the information that I needed for the task</td>
<td></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5.00</td>
<td>0.46</td>
</tr>
<tr>
<td>4.7 I am satisfied with the information provided by the tool</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.50</td>
<td>0.53</td>
</tr>
<tr>
<td>4.8 The quality of information is good</td>
<td></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5.00</td>
<td>0.52</td>
</tr>
</tbody>
</table>

**Participant** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **Median** | **Standard Deviation**
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
3.6 The layout of the screen was effective | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | 5.00 | 0.46
3.7 The size of the visualisation was adequate | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 5.00 | 0.46
3.8 Locating information elements of the screen was easy | 5 | 4 | 4 | 5 | 4 | 5 | 4 | 5 | 4.50 | 0.53
3.9 Description of the data is sufficient | 5 | 4 | 4 | 5 | 5 | 4 | 5 | 5 | 5.00 | 0.53
3.10 The use of colours in the visualisation was helpful | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5.00 | 0.35
3.11 Terminology related to the task was useful | 5 | 4 | 4 | 4 | 5 | 5 | 4 | 4 | 4.00 | 0.52
**Median** | **5.00** | **4.00** | **4.50** | **5.00** | **5.00** | **5.00** | **5.00** | **4.50** | **4.00** | **5.00** | **4.50**
**Standard Deviation** | **0.00** | **0.52** | **0.55** | **0.52** | **0.52** | **0.41** | **0.52** | **0.55** | **0.00** | **0.52** | **0.55**