

MIR: A TOOL FOR VISUAL PRESENTATION OF WEB ACCESS BEHAVIOR

By

VINODKUMAR P. KIZHAKKE

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2000

ACKNOWLEDGMENTS

I wish to thank my parents for their love and patience and allowing me to make my own mistakes and learn from them.

I wish to thank Dr. Douglas Dankel for allowing me to explore my areas of interest and guiding me through this thesis and for his gracious understanding of the various personal difficulties that punctuated my graduate study.

I wish to express my gratitude to the Gainesville Bahai' community who adopted me during my period of study at the University of Florida. I will be eternally indebted to them.

I wish to thank all my close friends for being there when I needed them.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGMENTS	ii
LIST OF FIGURES	v
ABSTRACT	vi
CHAPTERS	
1. INTRODUCTION	1
Internet Marketing and Consumer Behavior.....	1
Graphical Data Analysis	2
Motivation and Purpose of This Thesis	4
2. BACKGROUND AND RELATED WORK	6
Web Mining	6
Web Content Mining.....	8
Web Usage Mining	9
Pattern Discovery Tools.....	10
Pre-processing Tasks.....	10
Discovery Techniques on Web Transactions	12
Pattern Analysis Tools	14
Visualization Techniques	14
OLAP Techniques.....	15
Data and Knowledge Querying.....	15
Usability Analysis.....	17
WebViz: A Tool for World Wide Web Access Log Analysis.....	17
Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining.....	18
A Spatial Language for the Design of Digital Information Spaces	19
3. MIR–THE CONCEPTS	21
What Is a Metaphor?.....	21
What Is Information?	22
What Is Representation?	23
Metaphors and Data Visualization.....	23

What Is MIR?	25
Case Study: MIR for Visual Presentation of Web Access Behavior	26
MIR Step 1: Choosing the Metaphor-Architecture	27
Why Architecture?	28
MIR Step 2: Representation Mapping	29
Why VRML?	30
4. MIR – THE MODEL IMPLEMENTATION	32
Data Components	32
MIR Step 3: Data Cleansing, Processing and Loading	34
Data Processing and Loading	35
What Is a Session?	35
The MIR system	38
Web client	38
Server	38
VRML Generator	39
The VRML World	39
A Typical Sequence of Operations of the MIR System	42
5. CONCLUSIONS AND FUTURE WORK	47
Influence on Web Site Design	48
Visualizing Data Warehouses	48
Comparative Studies Using Multiple Animations	49
LIST OF REFERENCES	50
BIOGRAPHICAL SKETCH	52

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1: The Broad Classification of the Areas under Web Mining	7
2: Research Areas in Web Usage Mining	10
3: CISE Website represented by Architecture Metaphor	27
4: Physical Layout of Web Pages Visualized	30
5: Sample Record in Web Server Log	32
6: Sample Perl Script for Data Cleansing	35
7: Client Server Architecture of MIR System	38
8: Typical VRML Node	41
9: Date Entry Form	42
10: Session List Form	43
11: Right Camera View of VRML World	44
12: The VRML World from the Avatar's Viewpoint	45

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science

MIR: A TOOL FOR VISUAL PRESENTATION OF WEB ACCESS BEHAVIOR

By

Vinodkumar P. Kizhakke

August, 2000

Chairman: Douglas D. Dankel II

Major Department: Computer and Information Science and Engineering

The Internet has risen as a formidable arena for business. Understanding the behavior of the customer helps the Internet retailer anticipate their wants and also attract new customers. The volumes of data generated by customer traffic in commercial web site are mined to trace patterns for analysis. MIR (Metaphorical Information Representation) proposes a concept for non-statistical representation of data enabling the viewer to experience the data in a three-dimensional information space using metaphors. The Virtual Reality Modeling Language (VRML) is used to implement an interactive representation of the web access data of the Computer and Information Science and Engineering [CISE] Department's web site at the University of Florida using an architecture metaphor. The web site is visualized as a university campus. The tool is primarily designed as a tool for industrial psychologists to make presentations on customer behavior on the Internet in a user-friendly manner.

CHAPTER 1 INTRODUCTION

This chapter provides a preliminary account of the motivation behind and purpose of this thesis. It is organized into two main sections. The first provides a brief overview of Internet marketing and the importance of studying consumer behavior on the net. Data mining is introduced as a method to study Internet consumer behavior, which provides a lead into the second chapter. Second is a general overview of graphical analysis of data. The chapter ends with an outline of the structure of this thesis.

Internet Marketing and Consumer Behavior

Anyone who has purchased any product online has been exposed to the revolution that the Internet has created in retailing. The number of transactions performed on the Internet is constantly increasing at a continuously ascending rate. So, what makes web shopping more attractive to the shopper than physical shopping? Underhill, in his book “Why we buy?” [UNDE1999] gives the following reasons:

- Convenience: The customer can shop from anywhere and at anytime with a computer, a phone jack, and an electrical outlet.
- Information: The web provides extensive research material about most products. New techniques present all the experience about the product short of touching it.
- Speed: Faster access to information is provided with no waiting for catalogs to arrive or a sales person to answer the customer’s call.

- Limitless selection: It is possible to locate and purchase almost anything on the web.

Understanding the consumer's needs, wants, and buying behavior is at the heart of success in the market place. The basis of a business strategy to create new customers and to ensure repeat customers is the understanding of the existing ones. The discipline of consumer behavior in marketing involves the study of the actions performed by decision making units (i.e., customers) in the purchase, usage, and disposal of goods and the factors that influence the decision making process. The evolution of terms in the field of economic psychology like “Internet psychology,” provides a fair idea of the importance that the study of the consumer behavior on the Internet is gaining.

In the case of internet marketing, the study of consumer behavior translates in part to the analysis of the patterns in data generated as the consumer travels through a commercial website and makes purchases. Data mining is often used as a tool for analysis of this data. Data mining is the process of discovering useful and meaningful patterns, profiles, and trends in data using pattern recognition techniques such as neural networks, machine-learning, and genetic algorithms. When applied to the web, this is called web mining, which can be broadly defined as the discovery and analysis of useful information from the World Wide Web.

Graphical Data Analysis

Often the most effective way to describe, explore, and summarize a set of numbers and text is to look at a “picture” of them. The data mining process results in

patterns of data. The patterns are then analyzed either by pure statistical methods involving a direct study of the numbers and text evolving from the mining process or utilizing visuals which may be traditional statistical images like bar graphs and pie diagrams or innovative custom displays pertaining to the industry and the analysis in question. Sometimes a combination of the two is used for maximum understanding of the significance and meaning of the data.

Traditional data graphics visually display measured quantities by means of the combined use of points, lines, a co-ordinate system, numbers, symbols, text, shading, and color. Modern data graphics can be made to do more than just substitute simple figures representative of the numbers. Edward R. Tufte [TUFT1983], whose works are often considered the standards by which data graphics are judged, says that graphical displays should

- Avoid distorting what the data have to say,
- Induce the viewer to think about the substance and significance of the data,
- Make large sets of data coherent in a small space,
- Encourage the human eye to compare and experience the data,
- Reveal the data at several levels of detail, from a broad overview to the fine structure,
- Serve a reasonably clean purpose: description, exploration, presentation, or decoration, and
- Be closely integrated with the statistical and verbal descriptions of a data set.

Designing competent graphical displays of data demands three different skills: substantive, statistical, and artistic. A judicious blend of these skills ensures an effective

data display. Graphical data visualization has long since moved from display boards to the computer screen. These visuals have one of two purposes. First is direct analysis where the aesthetics become unimportant. As long as the person familiar with the data can decipher the meaning of the figure on the screen, the purpose is served. Second is the artistic and meaningful presentation of the data for a general audience unfamiliar with the specifics of the context, which requires ease of understanding. It is towards this purpose that the concepts presented in this thesis and the model implementation is geared.

The use of animation as a communication device at sales meetings and corporate board meetings to communicate complex concepts to employees was the subject of an article in a recent issue of the “AV Video Multimedia Producer” magazine [PLAN2000]. The decision to use web based graphics and animation for visualizing data was foremost during the formulation of this thesis and appears to be reflective of the thought process of the human-computer interface research in the industrial world.

Motivation and Purpose of This Thesis

Today, computerized visual aids for data presentation predominantly use traditional statistical images. Areas like consumer behavior will profit from presentations that convey the meaning and significance of the data in a user-friendlier, interactive manner that does not require a background in statistics. Moreover, these presentations are more visually aesthetic and entertaining. A system that enables such a display on the World Wide Web enables simultaneous access enabling corporate presentations to include personnel not located physically in one location. It also allows the individual user

to scrutinize and explore the presentation according to their specialized needs, enabling a better understanding of the underlying data.

This thesis presents a conceptual, virtual reality system that can be used interactively to visually present the behavior of a user visiting a commercial website. Its purpose is to provide a visual analysis tool in the area of web mining. A model implementation of this concept is provided.

The second chapter summarizes the background and academic research in the area of web usage mining with a focus on visualizing web traffic. The third chapter describes in detail the visualization concept used in this thesis. Chapter 4 describes the data components available and the implementation of the MIR system case study. This chapter also details the steps in a typical usage of the MIR system. The last chapter gives conclusive remarks and outlines the areas of future work and development.

CHAPTER 2

BACKGROUND AND RELATED WORK

The background for this thesis draws from various areas of research. The main areas are web usage mining and cognitive science with some insights from market research. The initial portion of this chapter presents an overview of academic research in web mining. Included is the application area and purpose of this thesis. The chapter concludes with descriptions of three specific projects that have inspired this work.

Web Mining

Web mining and its related areas of research developed very recently [COOL1997]. Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This broad definition on the one hand describes the automatic search and retrieval of information and resources available from millions of sites and on-line databases (i.e., web content mining) and on the other hand, the discovery and analysis of user access patterns from one or more web servers or on-line services (i.e., web usage mining).

There are several important issues, unique to the web paradigm, that come into play if sophisticated types of analyses are to be performed on server side data collections. These include the necessity of integrating various data sources such as server access logs, referrer logs, user registration, or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the

importance of identifying user sessions or transactions from usage data, site topologies, and models of user behavior.

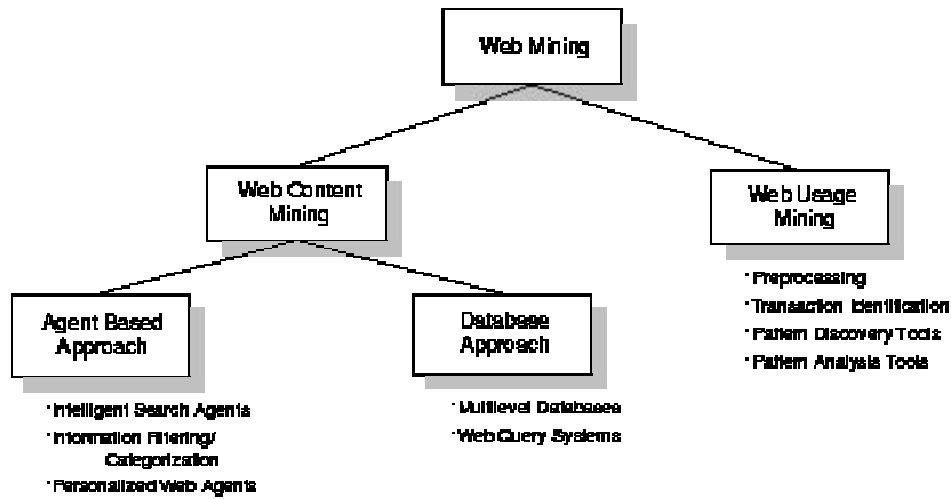


Figure 1: The Broad Classification of the Areas under Web Mining

Web Content Mining

The heterogeneity and lack of structure permeating much of the ever-expanding World Wide Web makes automated discovery, organization, and management of web-based information difficult. Traditional search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents.

These factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as extending database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The two main approaches are:

- Agent based approach: The agent-based approach involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information.
- Database approach: Database approaches have generally focused on techniques for integrating and organizing the heterogeneous and semi-structured data on the web into more structured and high-level collections of resources, like information in relational databases, and using standard database querying mechanisms and data mining techniques to access and analyze this information.

Web Usage Mining

Web usage mining is of interest for this thesis since it is used to model customer behavior. Web usage mining involves the discovery of user access patterns from one or more Web servers. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and is collected in server access logs. Other sources of user information include referrer logs that contain information about the referring pages for each page reference and user registration or survey data gathered via tools such as CGI scripts.

There are a number of advantages to web usage mining including:

- Analyzing such data can help organizations to determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among others.
- Analyzing server access logs and user registration data can provide valuable information on how to better structure a Web site to create a more effective presence for the organization.
- Managing workgroup communication and organizational infrastructure in organizations using intranet technologies can be made more effective with the insight that such analyses provide.
- Analyzing user access patterns helps target ads to specific groups of users for organizations that sell advertising on the World Wide Web.

The following chart represents the techniques for discovery and analysis of the patterns in web usage data. Descriptions follow immediately.

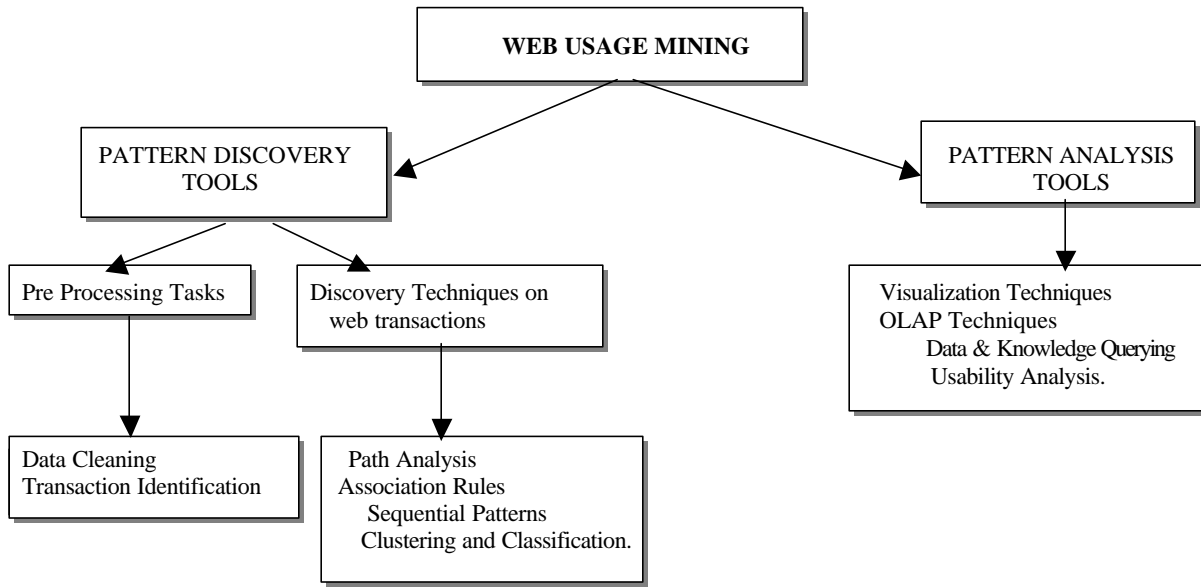


Figure 2: Research Areas in Web Usage Mining

Pattern Discovery Tools

The tools for user pattern discovery use sophisticated techniques from AI, data mining, psychology, and information theory, to mine for knowledge from collected data. For example, the WEBMINER system [COOL1997] introduces a general architecture for web usage mining. WEBMINER automatically discovers association rules and sequential patterns from server access logs. In subsequent sections some of these techniques are discussed in more detail.

Pre-processing Tasks

The processes that are necessary to transform the data collected from various sources into a form to which the actual mining algorithms and methods can be applied include

- **Data cleaning:** Techniques to clean a server log to eliminate irrelevant items are of importance for any type of web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log give an accurate picture of the user accesses of the website. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, jpg, JPG, and map can be removed. Another problem is that of user identification. Using a machine name to uniquely identify users can result in several users being erroneously grouped together as a single user unless details are gathered using HTML forms and CGI scripts. The reason is that unless the document is password protected, the server logs record only the IP-address or machine name of the accessing machine. Forms and scripts request explicit user information that can be used to identify the user. In most web log analyses, the IP-address (or machine name) coupled with the time of access is considered adequate information.
- **Transaction identification:** Before any mining is performed on web usage data, sequences of page references must be grouped into logical units representing web transactions or user sessions. A user session consists of all the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users, as discussed above. Unlike traditional domains for data mining, such as point of sale databases, there is no convenient method for clustering page references into transactions smaller than an entire user

session. Hence, assumptions are made depending on the final purpose of the mining process.

Discovery Techniques on Web Transactions

Once user transactions or sessions have been identified, several types of access pattern mining can be performed depending on the needs of the analyst. Some of these discovery techniques are:

- **Path Analysis:** Many different types of graphs can be formed from path analysis. The most obvious is a graph representing the physical layout of a website where web pages are nodes and hypertext links between pages are directed edges. Other graphs can be formed based on the types of web pages visited. For example, the edges might represent the similarity between pages, or the edges could give the number of users traversing from one page to another. Most of the work to date involves determining frequent traversal patterns or large reference sequences from the physical graph layout.
- **Association Rules:** Association rule discovery techniques are generally applied to databases of transactions where each transaction consists of a set of items. In this framework, the problem is to discover all associations and correlations among data items where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items. In the context of web mining, this problem amounts to discovering the correlations among references to various files available on the server by a given client.

- **Sequential Patterns:** The problem of discovering sequential patterns is to find inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. In web server transaction logs, visits by a client are recorded over a period of time. The time stamp associated with a transaction in this case is a time interval determined and attached to the transaction during the data cleaning or transaction identification process. The discovery of sequential patterns in web server access logs allows web-based organizations to predict user visit patterns and helps in targeting advertising aimed at groups of users based on these patterns. By analyzing this information, the web mining system can determine temporal relationships among data items. Another important type of data dependency that can be discovered, using the temporal characteristics of the data, is similar time sequences. For example, we may be interested in finding common characteristics of all clients visiting a particular file within the time period. Or, conversely, we may be interested in a time interval (within a day, or within a week, etc.) in which a particular file is most accessed.
- **Clustering and Classification:** Discovering classification rules allows one to develop a profile of items belonging to a particular group according to their common attributes. This profile can then be used to classify new data items that are added to the database. In web mining, classification techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients, or based on their access patterns. Clustering analysis allows one to group together clients or data items

that have similar characteristics. The clustering of client information or data items on web transaction logs, can facilitate the development and execution of future marketing strategies, both online and off-line. These strategies include automated return mail to clients falling within a certain cluster or dynamically changing a particular site for a client, on a return visit, based on the past classification of that client.

Pattern Analysis Tools

The discovery of web usage patterns, through techniques described above, would not be very useful unless there were mechanisms and tools to help an analyst better understand them. Hence, in addition to developing techniques for mining usage patterns from web logs, there is a need to develop techniques and tools for enabling the analysis of discovered patterns. These techniques draw upon a number of fields including statistics, graphics and visualization, usability analysis, and database querying. This section provides a survey of the existing tools and techniques.

Visualization Techniques

Visualization has been used very successfully in helping people understand various types of phenomena, both real and abstract. Hence it is a natural choice for understanding the behavior of web users. Pitkow and Bharat [PITK1994] developed the WebViz system for visualizing WWW access patterns. This project is one of the inspirations for this thesis and is described in more detail below.

OLAP Techniques

On-Line Analytical Processing (OLAP) is emerging as a very powerful paradigm for strategic analysis of databases in business settings. The research community has recently demonstrated that the functional and performance needs of OLAP require the design of new information structures. This has led to the development of the data cube information model [GRAY1996] and techniques for its efficient implementation. Recent work [DYRE1997] has shown that the analysis needs of web usage data have much in common with those of a data warehouse, hence making OLAP techniques applicable. The access information in server logs is modeled as an append-only history, which grows over time. A single access log is not likely to contain the entire request history for pages on a server, especially since many clients use a proxy server. Information on access requests will be distributed, and there is a need to integrate it. Since server logs grow quite rapidly, it may not be possible to provide on-line analysis. Therefore, there is a need to summarize the log data, perhaps in various ways, to make its on-line analysis feasible. Making portions of the log selectively (in)visible to various analysts may be required for security reasons. These requirements for web usage data analysis show that OLAP techniques may be quite applicable, and this issue needs further investigation. A project that uses a data warehouse and OLAP techniques to study web server logs [BUEC1999] is described later in this thesis.

Data and Knowledge Querying

One of the reasons attributed to the great success of relational database technology has been the existence of a high-level, declarative, query language, which allows an application to express what conditions must be satisfied by the data it needs,

rather than having to specify how to get the required data. Given the large number of patterns that may be mined, there appears to be a definite need for a mechanism to specify the focus of the analysis. Such a focus may be provided in at least two ways. First, constraints may be placed on the database (perhaps in a declarative language) to restrict the portion of the database from which to mine. Second, querying may be performed on the knowledge that has been extracted by the mining process, in which case a language for querying knowledge rather than data is needed. An SQL-like querying mechanism has been proposed for the WEBMINER system [COOL1997].

Usability Analysis

Research in human-computer interactions (HCI) has recently started developing a computational science of usability. The principal goal of this effort is develop a systematic approach to usability by adapting the rigorous experimental method of computational science. The first step is to develop instrumentation methods that collect data about software usability, in a manner akin to instrumentation that has been done for analyzing performance. This data is then used to build computerized models and simulations that explain the data. Finally, various data presentation and visualization techniques are used to help an analyst understand the phenomenon. This approach can also be used to model the browsing behavior of users on the web.

Brief accounts of the three projects/papers that have been the major sources of inspiration for this project now follow.

WebViz: A Tool for World Wide Web Access Log Analysis

This project [PITK1994] is representative of the most common method of representing the World Wide Web – as a graph or network. WebViz visualizes the collection of hypertext documents as a directed cyclic graph. The links in this network-like structure are referred to as paths and represent the hyperlinks between the documents. Each node represents a separate document. This is called the Web-Path Paradigm. WebViz collects frequency and recency information about documents and paths to drive a visual display.

The database back-end is a flat file. In the visual display, the parameters are the thickness and color of nodes and links. Since the main purpose is to record recency (e.g., Was this page touched recently?) and frequency of access (e.g., How many times has this

link been used?), these values are mapped to colors and thickness. For example, a node that has been accessed in the last 60 seconds would be white and the node that has not been accessed in the last 60 seconds would be blue. Visual mapping of the frequency parameter involves varying the thickness of the lines connecting the nodes depending on the number of traversals recorded for that link.

The log data are accessed on a temporal basis. An IP-address is entered into the View Control window and the upper and lower limits of a time interval are also entered. The visualization translates the log data and changes the colors and thickness of the pre-structured graph. The implementation of the visual is done using C++.

The disadvantages of this system is that it relies heavily on the human eye's ability to perceive the difference between the thickness of lines in a network and on a computer screen; therefore, it is probably not a reliable method to use for analysis. The network representation also does not permit much by way of presentation aesthetics.

Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining

The focus of this project is to find a novel way of utilizing data mining techniques on Internet data to discover useful marketing intelligence in electronic commerce scenarios. The most important contribution comes in the form of a formal definition of a generic web log data hypercube and schematic designs for predictive and analytical activities. The project presents a Star or Snowflake warehouse schema that incorporates elements from the web log in combination with data from cookies and demographic data from third party marketing agencies. The applications suggested utilizing this web log cube as a back end are statistical and predictive but not visualization oriented.

This gave rise to the idea of using a data warehouse as a back end for this thesis. But there are no demographic or cookie data available for the case study selected and hence the warehouse is implemented as a single table database into which the cleaned and processed web log data is loaded.

A Spatial Language for the Design of Digital Information Spaces

This thesis performed at the University of Washington [TANN1999] draws on research in spatial metaphor and visualization techniques to propose that a three-dimensional digital environment would provide a more natural interface to that information. This study investigated human perception of form and space as it related to three-dimensional computer models displayed on a computer monitor. It concluded that the basics of form, space, and order could successfully be used to represent information. An information space that represents the data it contains and provides a context for analysis would allow people to filter information both actively and passively and would engender a sense of place among participants. In anticipation of increased presence and communication within electronic space this thesis marks an important milestone in the construction of digital information places that afford both sight and insight.

In the future work section of this thesis is listed the possibility of using virtual reality to design these digital information spaces that form interfaces and representations of data. This sparked the idea of using virtual reality as the tool for visualization in this thesis, the choice being VRML.

Elements of the last three papers and projects listed are combined to form the concept and implementation of MIR. The next chapter outlines the salient features of MIR.

CHAPTER 3

MIR–THE CONCEPTS

This chapter introduces the various components of the acronym MIR (metaphors, information and representation), in their individual contexts of origin, through their use in the realms of computation. After a brief account of the use of metaphors in data visualization, the discussion leads to the description of MIR as a concept combining aspects of these components. A conceptual adaptation of MIR to present the behavior of a user in a commercial website is then presented. A model application of the concept is explored using a case study.

What Is a Metaphor?

Webster’s dictionary gives the meaning of the word metaphor as “a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them.” In its linguistic context of origin, Veale [VEAL1995] defines a metaphor as the act or process of denoting one concept (the tenor) with a sign conventionally tied to another (the vehicle), with the purpose of:

- Emphasizing certain associations of the tenor over other associations (e.g., my dentist is a barbarian),
- Enriching the conceptual structure of the tenor by analogy with another domain (e.g., the CPU is the brain of the computer), and

- Conveying some aspect of the tenor that defies conventional lexicalization (e.g., the leg of the chair, the neck of the bottle).

Metaphor comprehension forms a very important area of research in cognitive science, an interdisciplinary study of mind and the nature of intelligence. Applications of cognitive science include among other fields: data processing, human computer interaction, and artificial intelligence. In the area of human computer interface, metaphors facilitate the construction of mental models. A metaphor involves the understanding of how one system works to facilitate the interpretation of a different system. Traditional two-dimensional metaphors often seen on the desktop of a typical computer include the trashcan and the file folder. These have recently been extended to the third dimension to include spatial metaphors such as the virtual office, room, and a city. The city metaphor draws on the inherent knowledge of how cities are structured as a way to understand the organization of information [DIEB1998].

What Is Information?

Information in the computational context is best explained relative to yet another common term in the field, data. They are often used interchangeably, but are quite distinct in their meaning. Data are the collection of discrete elements, collected and stored in a form and medium designed for use by a computer system. Information can be defined as “value added data.” In other words, various application-specific operations are performed on the raw data to make the data usable for various purposes and the result is information. Typical operations include merging, aggregation, derivation, sorting, and

structuring. These actions enrich and enhance the value of the data and make it information. Data processing forms a vital part in any information system.

What Is Representation?

Different researchers, in various contexts, have used diverse expressions to define Representation. From the various definitions studied, the most relevant and easily transferable to the field of computation appears to be the following. Representation is defined as having four components [MARK1999]:

- A represented world: The domain containing the objects to be represented.
- A representing world: The domain containing the actual representations.
- Representing rules: The representing world is related to the represented world through a set of rules that map elements of the represented world to elements in the representing world.
- A process using the representation: The combination of the first three components creates just the potential for representation. The capabilities of a system for representation is complete only when there is a potential and a process.

In other words, an object from the represented world is mapped to the representing world by a representation rule and a process uses this entire system.

Metaphors and Data Visualization

A brief insight into how humans process visual information now becomes relevant. There are a range of cognitive activities that may occur between the time when a person first looks at a symbol and the time when the relevant information is extracted.

The memory of a person can be divided into three parts: long-term memory, short-term memory, and working memory. The perceptual image formed when a person first looks at an image is moved into the working memory. This is the workbench where the metaphor serves as the tool to process the information by matching the perceptual image to one in the long-term memory or the short-term memory. An effective metaphor design initiates a match from the long-term memory [TANN1999].

As discussed earlier, two and three-dimensional visual metaphors have been used in computational areas for many years. The latest addition has been the third dimension and the spatial metaphor. This involves a participatory experience of the user in three-dimensional information space. The natural action of bringing objects of interest closer to see them better can form the basis of an intuitive interface metaphor for information visualization. The following is a list of the observations that can serve to benefit the designs of spatial interfaces to information [BLUM1996]:

- Living and acting in a space is a common experience for all users,
- Spatial structure (landscapes, cities) and artifacts (desks, buildings) offer familiar operations,
- Human memory relies heavily on spatial arrangements and layouts of items,
- Human spatial experience is tightly linked to visual and auditory perception – the primary channels of human computer interaction,
- Spatial structure offers both simplicity at a single level of resolution and hierarchical refinement through operations like zoom, pan, enter, and navigate, and

- Spatial structures convey a wide range of auxiliary information in subtle but intuitive ways, including assistance (e.g., maps, signboards) and visualizing choices for actions (e.g., crossroads, intersections).

Hence, the efficient use of the spatial metaphor to provide an interactive experience of the data serve as a powerful tool to understand its meaning and significance. See Chapter 2 for an account of relevant research work performed in this area.

What Is MIR?

Metaphorical Information Representation [MIR] is a conceptual tool that combines the main aspects of the above-mentioned components in a manner applicable to the area of visual presentation of data. It is best presented as a generalized protocol or a series of ordered steps:

1. Choosing the metaphor: The choice of metaphor entirely depends on the application and the data being studied since it relies on the cognitive psychology of the user to supply the missing elements necessary to effectively understand the information presented. MIR focuses on a three-dimensional spatial metaphor.
2. Representation mapping: The elements in the data (i.e., the representing world) are mapped into the elements in the information space (i.e., the represented world) and the rules of mapping are formed.
3. Data processing: The raw data is re-structured in a form enabling the implementation of the metaphor and its visual presentation in a relatively uncomplicated manner. Ease of understanding of the data without exposure to the

underlying operations is the key goal of the concept. The processed data is stored in a database or data warehouse. The data processing techniques and storage methods will depend on the available technology.

MIR's degree of efficiency depends most on significant quality transfer between the data and the visual representation allowing the user to experience the data in an interactive three-dimensional space and understand its meaning in the context provided by the spatial metaphor.

The demonstration of the design and implementation of a case study using MIR follows. The first two steps of the MIR protocol are described in the rest of this chapter and the data processing step followed by the implementation of the web-based visualization system are described in the next chapter.

Case Study: MIR for Visual Presentation of Web Access Behavior

Visualization of web access data is an area of study that has received relatively less focus among academic circles and the commercial attempts have mostly been a visualization using statistical images like graphs and pie diagrams. A description of the academic research in this area is provided in Chapter 2.

The website chosen for this study and model implementation is that of the Computer and Information Science and Engineering [CISE] Department at the University of Florida. The choice of this site was based on the availability and permissible access to web server logs that are regarded as highly confidential by commercial websites.

MIR Step 1: Choosing the Metaphor-Architecture.

The metaphor chosen to represent the CISE website is Architecture. Specifically, the architecture style chosen closely resembles that of the University of Florida campus with a focus on the CISE department. In other words, the CISE website is modelled as a University campus and a physical landmark, the “Alachua” sculpture adjacent to the entrance to CISE building, is represented to mark the entry point to the website. The walls of the buildings are mapped to a brick texture similar to that of the CISE building exterior.

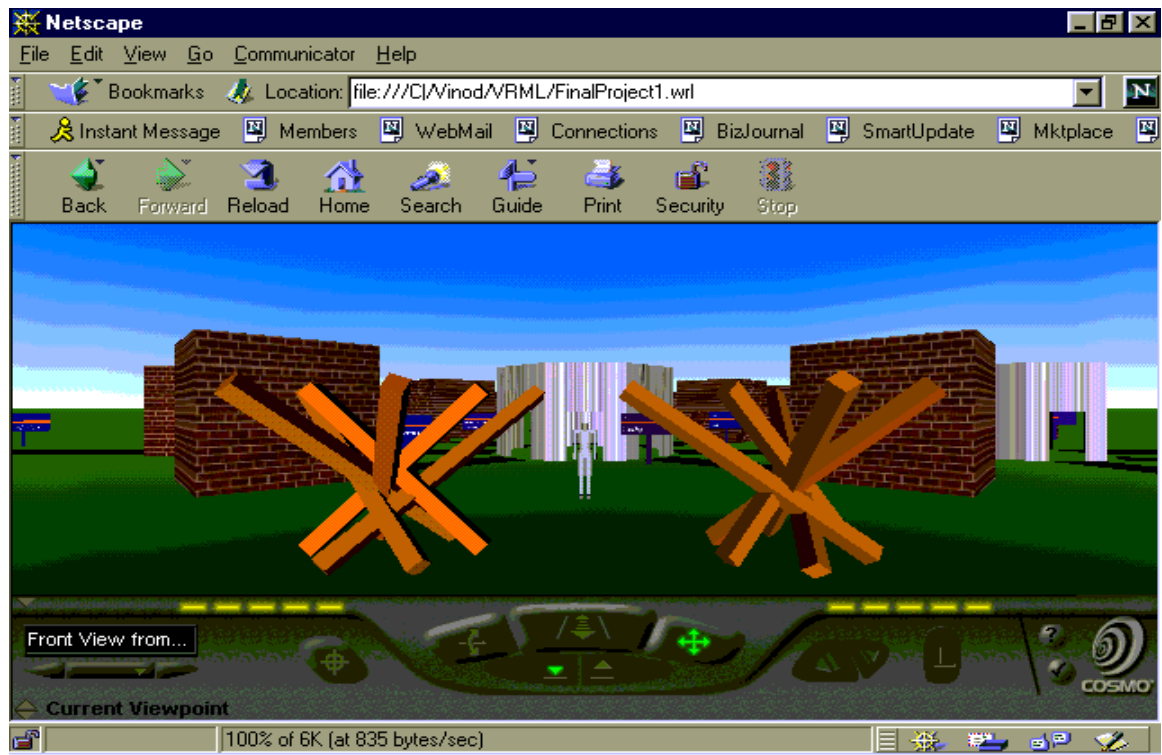


Figure 3: CISE Website represented by Architecture Metaphor

Why Architecture?

The qualities of architecture that are considered the most important are commodity, firmness, and delight[VITR1960]. Commodity refers to appropriateness of form to function, firmness addresses structural integrity, and delight refers to the ability to surprise. These three words also speak of the essence of a good interface between humans and information [MITC1995]. MIR is an interface tool between humans and information and since the qualities are similar to those of architecture, to effect a quality transfer between the two is relatively easy to understand and convey. Keeping with the requirement of MIR for a spatial metaphor, architecture fits the bill quite literally providing significant scope for the implementation of the information space. Moreover, navigation through a website can be very easily compared to movement between various buildings within a campus. Added familiarity in this case comes from the visual representation being close to that of a University campus with elements from the University of Florida. The choice of the architecture metaphor hence provides a very rich context for the design of an aesthetically pleasing, virtual information space for the user to intuitively navigate and experience the data and thereby allowing the study of the behaviour of a particular user during a particular session. The path taken by a user (i.e., the underlying data) is presented visually by the movement of an animated AVATAR that traverses the various buildings and pathways. The entire campus is visualized as a collection of single story buildings without any roofs so that the movement of the AVATAR can be clearly seen by varying the camera angles and view points depending on the viewer's preference.

MIR Step 2: Representation Mapping

Once the metaphor is selected, it is time to design the representation of the information.

In this case, the elements are as follows:

- The represented world: This domain consists of the actual web pages, the hyperlinks and the titles of the pages.
- The representing world: This domain consists of buildings of different shapes with different entrances and exits, walkways and paths, and signboards.
- Representation Rules: The mapping rules are as follows:
 - i. The web pages that have only one hyperlink are mapped to rectangular buildings that have one entrance and one exit.
 - ii. The web pages that have more than one hyperlink are mapped to buildings that have a circular cross section (i.e., cylindrical with as many entrances and exits as there are hyperlinks leading into and out of them).
 - iii. The hyperlinks themselves are mapped to paths that connect the various buildings. The paths have no walls on their sides to increase the visibility of the AVATAR.
 - iv. The titles of the web pages are mapped to signboards similar to those on the University of Florida campus.
- The Process that uses the representation: The process that is using this representation is that of visualization and animation.

Not all the pages in the CISE website are represented and visualized . Individual user pages are not considered and the pages presented are those having had many recent

visits for the purpose of clear visualization and movement. The following diagram shows the physical layout of the web pages visualized.

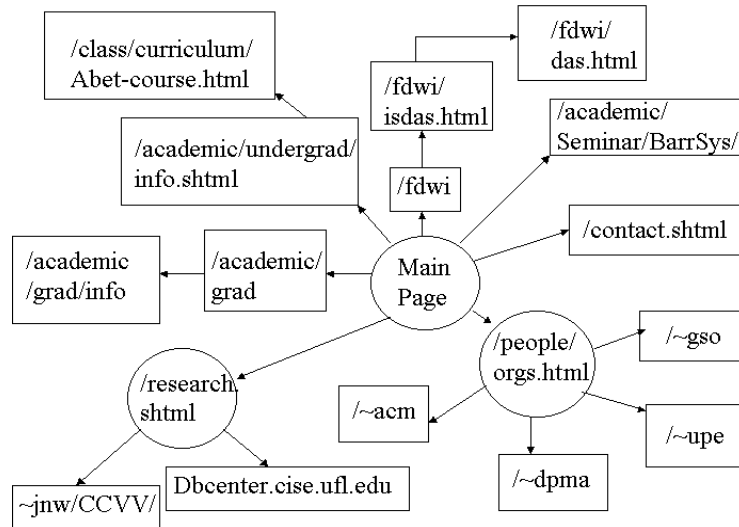


Figure 4: Physical Layout of Web Pages Visualized

The actual data processing and implementation forming the Step 3 of this MIR model is described in the next chapter

Why VRML?

The visual presentation in this MIR model is done as web-based animation. The tool used is the Virtual Reality Modeling Language. It is essentially a 3D interchange format and it is a 3D analog to HTML. This means that VRML serves as a simple, multiplatform language for publishing 3D worlds on the web. It provides the technology that integrates three dimensions, text, and multimedia combined with scripting languages and Internet capabilities enabling interactive application design and development.

It was also the tool of choice for this project because the VRML files is essentially a plain text file making it easier to generate on the fly in response to user queries. A VRML generator builds the files and it is implemented using Perl.

CHAPTER 4

MIR – THE MODEL IMPLEMENTATION

This chapter describes the example implementation of the metaphor proposed and discussed in the previous chapter. The data components available, the cleaning and loading of the data into the backend database, the adding of value to the existing data to simplify the generation of the corresponding VRML world, and, finally, the module that performs the generation of the VRML world are described in detail.

Data Components

The sample scenario chosen, as explained in the previous chapter, is the visualization of the web access traffic of the Computer and Information Sciences and Engineering [CISE] Department website at the University of Florida. The server log of the CISE Department uses the “Extended Log Format” as created by the National Center for Supercomputing Applications (NCSA). A typical record in the web server log, otherwise known as the transfer log, appears as follows. The individual components are [MENA1999]:

```
210.218.175.96 - - [05/Jun/2000:05:01:21 -0400] "GET
/academic/grad/info HTTP/1.1" 200 34117
"http://www.cise.ufl.edu/academic/grad" "Mozilla/4.0 (compatible; MSIE
5.0; Mac_PowerPC)"
```

Figure 5: Sample Record in Web Server Log

- Host Field [210.218.175.96]: This is the first field of the extended log format. This is usually the remote server making the request from the website. It is either in the format of a domain name (rain.cise.ufl.edu) or an IP address (128.227.205.225).
- Identification Field [-]: This field is almost always a hyphen due to the fact that it is never used.
- Authuser Field [-]: This field was designed to contain the authenticated user name that a visitor may need to gain access to a protected directory accessible only by a password.
- TimeStamp [05/Jun/2000:05:01:21 -0400]: This field contains the date, time, and the Greenwich Mean Time (GMT). The format for the date is DD/Mon/YYYY and the format for the time is HH:MM:SS. The GMT uses a plus or minus sign to indicate the number of hours the local time for the server differs from Greenwich meantime.
- HTTP Request ["GET /academic/grad/info HTTP/1.1"]: This field contains three parts. The first part contains a command. The values that can be in this part are "GET," "POST," or "HEAD." The GET method tells the server which document the visiting browser is requesting. The POST tells the server to expect some data and what program or script can be used to handle it. The HEAD command works exactly like the GET method except that the server only returns the <HEAD> section of any document. The second part of this field actually contains the name of the file being requested (/academic/grad/info/). The third part contains the name and version of the protocol.

- Status Code Field [200]: This field in describes the outcome of the transaction requested by a status code.
- Transfer Volume [34117]: This field provides the total number of bytes transferred by the server to the client during the transaction.
- Referrer Log [["http://www.cise.ufl.edu/academic/grad"](http://www.cise.ufl.edu/academic/grad)]: This field contains the URL from which the requests for the web page originated. If the website was found through a search engine like Yahoo!, this field records the keywords that were used to locate the site.
- Agent Log ["Mozilla/4.0 (compatible; MSIE 5.0; Mac_PowerPC)"]: This field records the name and version number of the browser making the request. It also contains the name of search engine tracks.

The different fields are delimited by a single white space and are recorded in ASCII text format.

MIR Step 3: Data Cleansing, Processing and Loading

Continuing the MIR steps from the previous chapter, the data cleansing process is performed next. The server logs record every visit and transaction that takes place on that server. The data are voluminous. To extract useful data, the first step is to eliminate all unnecessary data. In the case of this implementation, two sets of Perl script perform these operations. The first set deletes records in the log that are records of accesses of images (GIFs, JPEGs, etc.). Matching the appropriate text pattern in the HTTP request field and deleting the record achieves this. The pattern is changed and the script is run

many times to eliminate all the common patterns. This reduces the size of the log file. A sample script appears as follows:

```
$logfile = "server.log";  
open(LOGFILE, $logfile) || die "Error! : Cannot open $logfile: $!\n";  
while(<LOGFILE>){  
  @fields=split;  
  print unless $fields[4]=~/gif/;  
}
```

Figure 6: Sample Perl Script for Data Cleansing

The second step is to use a different script to select the records needed for this project from the now smaller log file, placing these records into another ASCII text file. The patterns are the locations and names of the web pages to be visualized. The script is similar to the one above except that the lines selected are inserted into another temporary file. The temporary file now contains all the records pertaining to the visits to the web pages selected for this project.

Data Processing and Loading

The cleaned data is now processed and additional fields added to facilitate ease of animation. Visualization is done on a per session basis, so the primary step is to define what a session is.

What Is a Session?

A session is the transaction identification process described in Chapter 2. The cleaned data is aggregated into distinct units for analysis. The analytic process in this

case, is the visualization. For purposes of this implementation, all the visits from a particular IP-address to the CISE website on one day are identified as being one session. Visits to pages not visualized are eliminated and hence a session is reduced to a continuous traversal through the pages chosen. Later sections describe how these assumptions are translated into movement to ensure clarity in the visual presentation.

The cleaned data is loaded into a Microsoft Access database since it allows easy import of delimited text into tables. The manipulation and session splitting is done using update queries and Visual Basic scripts. The records are sorted by IP address, then grouped by IP address and sorted by ascending order of time. This groups the data selected from the server logs of a single day into sessions. These sessions now are to be named for identification.

A field is added in every record in the temporary table and an update query is used to add the session name. The conventions elected for the session name is: The date + a 3 digit sequential number. Hence, a session occurring on 06/05/2000 will have a session id as 06052000 + a 3 digit sequential number (e.g., 06052000019). Using this convention, a maximum of 999 sessions per day can be grouped and identified.

Yet another field is added which is a single character string. This is referred to as a page code and is used to identify the sequence of pages visited in a session. The following table gives the page codes allocated to the pages visualized in this implementation. Since the number of pages is less than 26, the codes are alphabetical. By increasing the length of the page code string, any number of pages can be included.

Table 1: Page Codes for the Web Pages of the CISE Website

S.No	URL of Page	Page Code
1	http://www.cise.ufl.edu/	a
2	http://www.cise.ufl.edu/contact.shtml	b
3	http://www.cise.ufl.edu/academic/seminar/BarrSys/	c
4	http://www.cise.ufl.edu/fdwi/	d
5	http://www.cise.ufl.edu/fdwi/das.html	e
6	http://www.cise.ufl.edu/fdwi/isdas.html	f
7	http://www.cise.ufl.edu/academic/undergrad/info.shtml	g
8	http://www.cise.ufl.edu/class/curriculum/abet-course.html	h
9	http://www.cise.ufl.edu/academic/grad/	i
10	http://www.cise.ufl.edu/academic/grad/info/	j
11	http://www.cise.ufl.edu/research.shtml	k
12	http://www.cise.ufl.edu/~jnw/CCVV/	l
13	http://www.dbcenter.cise.ufl.edu/	m
14	http://www.cise.ufl.edu/people/orgs.shtml	n
15	http://www.cise.ufl.edu/~acm/	o
16	http://www.cise.ufl.edu/~dpma/	p
17	http://www.cise.ufl.edu/~gso/	q
18	http://www.cise.ufl.edu/~upe/	r

After the session-id and page code are added to each of the records in the temporary table, a few of the fields are chosen to be loaded into the Sybase backend of the MIR system. The fields selected are: Session-Id, Page Code, IP-Address, Date, and URL of the visited Page. This selection is based on the requirement for this implementation. If the visualization incorporates the referral log values also, then it can be included. The design is very versatile. The selected fields are then exported as a text file with the “|” as the column delimiter and a “&” as the row delimiter. This text file is then imported into a Sybase table using the bulk copy routine. Now, the data has been cleaned and loaded and is ready for use in the MIR system.

The MIR system

The MIR system is designed using a client-server architecture and is implemented using Perl and CGI. The client-server interaction is shown Figure X.

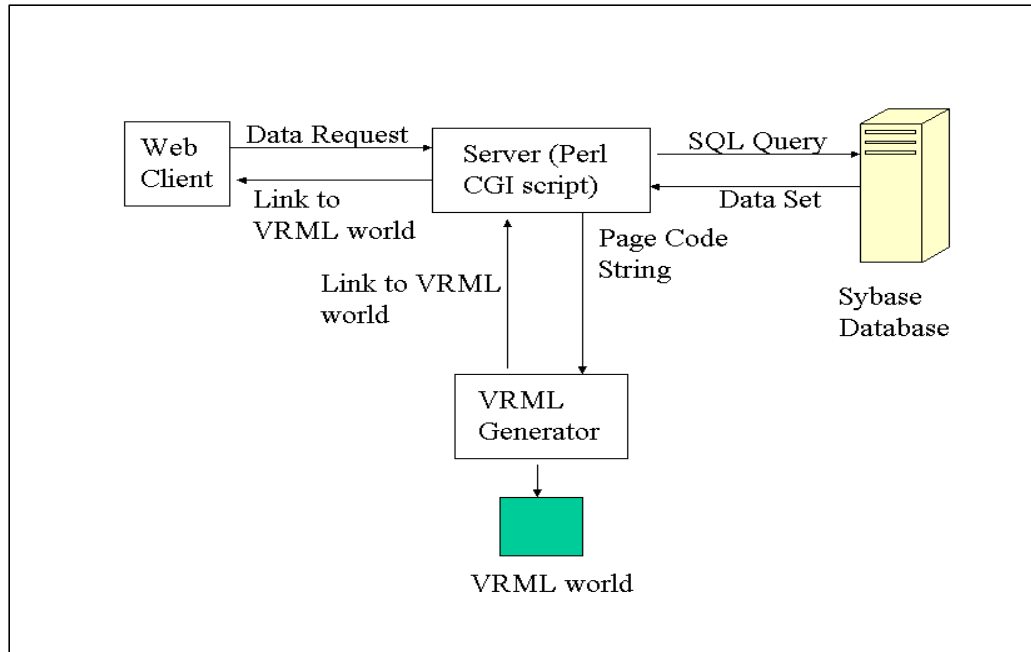


Figure 7: Client Server Architecture of MIR System

Web client

The client is a web page. The user is prompted to enter a date. This date is communicated to the server. A VRML browser plugin (Cosmo Player offered by SGI) is installed to enable the visualization of the VRML world generated. The client receives a link to the VRML world. The world is displayed by clicking on the link.

Server

The server is a CGI script written in Perl. This receives the date from the web client page. A connection to the Sybase back-end using the Perl DBI module is made and

an SQL statement with the date as parameter is executed. The results, a listing of the sessions for that day, are displayed on a webpage. The user then selects one session and re-enters that value. This is now sent to the server where another SQL statement is executed. The data set for the particular session is retrieved and a page code string is constructed by concatenating the page codes of all the pages in that session in order. This string is then sent to the VRML generator.

VRML Generator

The VRML generator is also a CGI script written with Perl subroutines. This receives the page code string from the server. The string is unpacked into its components. Perl hash data structures are used to store position and orientation values for the various pages visualized. The values are retrieved and the corresponding lines of VRML code are generated as lines of text. The code generated determines the path taken by the AVATAR for the traversal of the world. These lines are then inserted between existing code files that compose the standard portions of the world. The new file is renamed and its permissions are changed so that it can be viewed on the WWW. The hyperlink to the world is returned to the server, which is then sent to the browser-the web client.

The VRML World

The Virtual Reality Modeling Language (VRML, pronounced as *vermal*) is a file format for describing interactive three-dimensional objects and worlds. It was first started as an experimental project at Silicon Graphics, Inc., in 1989 [CARE1997].

Here a world is a model of a 3D space, which can contain 3D objects, light and sound sources, and backgrounds; in other 3D systems this is often called a scene. Objects can be built from solid shapes, from text, or from primitive points, lines, and faces. Objects can be grouped into more complex objects, used multiple times, translated, and rotated. Objects can trigger events, which can be routed to other events or to scripts written in JavaScript or Java. Within VRML worlds, one can trigger sounds, move objects along paths, and link to HTML or other VRML targets. The experience for someone browsing a VRML world can be active or passive, depending on design.

A VRML file contains a file header, comments, and nodes. Nodes, which describe objects, may have names and contain fields and values. The VRML world is a directed acyclic graph of nodes. Nodes can contain other nodes (some type of nodes may have “children”) and may be contained in more than one node (they may have more than one “parent”), but a node must not contain itself. The world can be split into small scene sub-graphs which makes it relatively easy to create large and complicated worlds. There are some world builders available that make it easier for a non-technical person to create basic worlds, but hand coding is considered the best way to get maximum customization. This project is entirely hand coded except for the creation of the AVATAR which was done using a freely available VRML world builder named VRML Pad from Parallel Graphics, Inc.

Typical VRML code for a node is shown in Figure 8.

```

Transform {

  children [
    DEF Building Shape{
      appearance Appearance{
        material Material {}
        texture ImageTexture {url "brick.jpg" }
      }

      Geometry IndexedFaceSet {
        solid FALSE
        coord Coordinate { point [0 0 0, 1 0 0, 1 0 -1, 0 0 -1, 0 1 0, 1 1 0,
                                   1 1 -1, 0 1 -1, 0.33 0 0, 0.33 0.5 0, 0.66 0.5 0,
                                   0.66 0 0, 1 0 -0.33, 1 .5 -0.33, 1 .5 -0.66,
                                   1 0 -0.66, 0.66 0 -1, 0.66 0.5 -1, 0.33 0.5 -1,
                                   0.33 0 -1,]}
        coordIndex [ 0 4 5 1 11 10 9 8 -1,
                     1 5 6 2 15 14 13 12 -1,
                     2 6 7 3 19 18 17 16 -1,
                     3 7 4 0 -1,]
      }
    ]
  }
}

```

Figure 8: Typical VRML Node

The code in Figure 8 is an example of a Transform node that permits this object to be translated or rotated as needed. The appearance of the object is specified by an appearance node. The texture is taken from an image (brick.jpg). The IndexedFaceSet node is used to build the object from the co-ordinates of the tile in which it fits. The coordIndex specifies the sequence in which the co-ordinates are to be connected and filled to build the shape, which in the above case is one of the buildings in the campus “world” built for this project. VRML models movement by routing events from one object to another.

A Typical Sequence of Operations of the MIR System

This section describes the typical sequence that a user performs while using the model implementation of the MIR system.

Step 1: The user enters a date from the Date Entry Form. The sessions on this date are ones that the user wants to study.

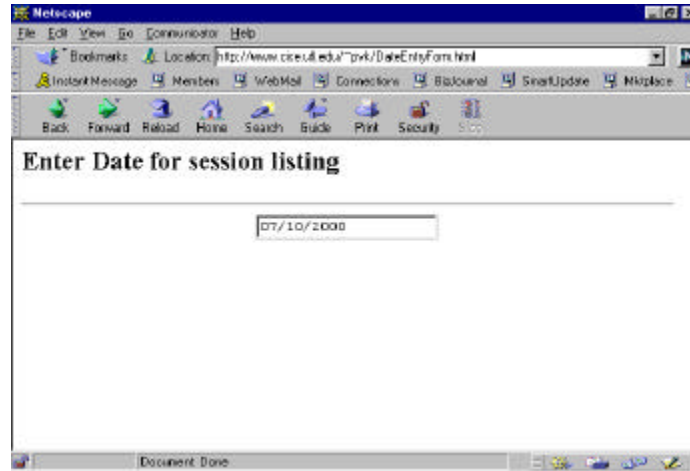
A screenshot of a Netscape browser window. The title bar says 'Netscape'. The address bar shows 'http://www.cireull.edu/~prvk/DateEntryForm.html'. The browser has a menu bar (File, Edit, View, Go, Communicator, Help) and a toolbar with icons for Back, Forward, Reload, Home, Search, Guide, Print, Security, and Stop. Below the toolbar, the page title is 'Enter Date for session listing'. There is a large text input field with the date '07/10/2000' entered. The status bar at the bottom says 'Document Done'.

Figure 9: Date Entry Form

Step 2: The list of sessions for the particular date are displayed. One of the session names is re-entered by user into the form.

The Sessions for 07/10/2000 are

07102000008	07/10/2000	195.37.70.1
07102000001	07/10/2000	166.90.67.253
07102000002	07/10/2000	202.96.227.53
07102000003	07/10/2000	216.34.109.190
07102000005	07/10/2000	165.247.45.134
07102000006	07/10/2000	134.50.237.227
07102000009	07/10/2000	128.227.162.44
07102000004	07/10/2000	209.117.183.130
07102000007	07/10/2000	149.170.190.138

Enter Session Name:

Document Done

Figure 10: Session List Form

Step 3: The user clicks the “Generate VRML” button on the form.

Step 4: A form displays the details of the session selected with a “Click here to see VRML world” button.

Step 5: The VRML world is displayed.

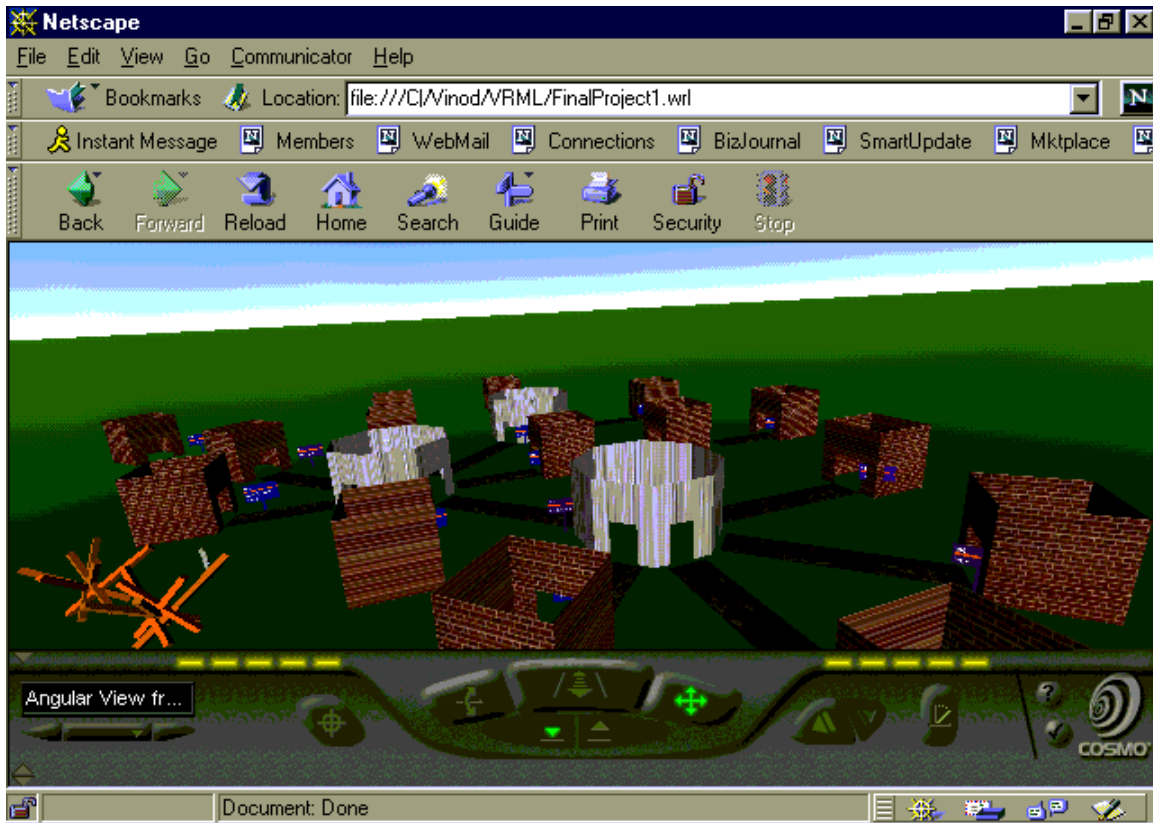


Figure 11: Right Camera View of VRML World

At any point the user can move the viewpoint within the scene to provide an alternative visualization of the world. Figure X provides another view from the AVATAR's viewpoint.

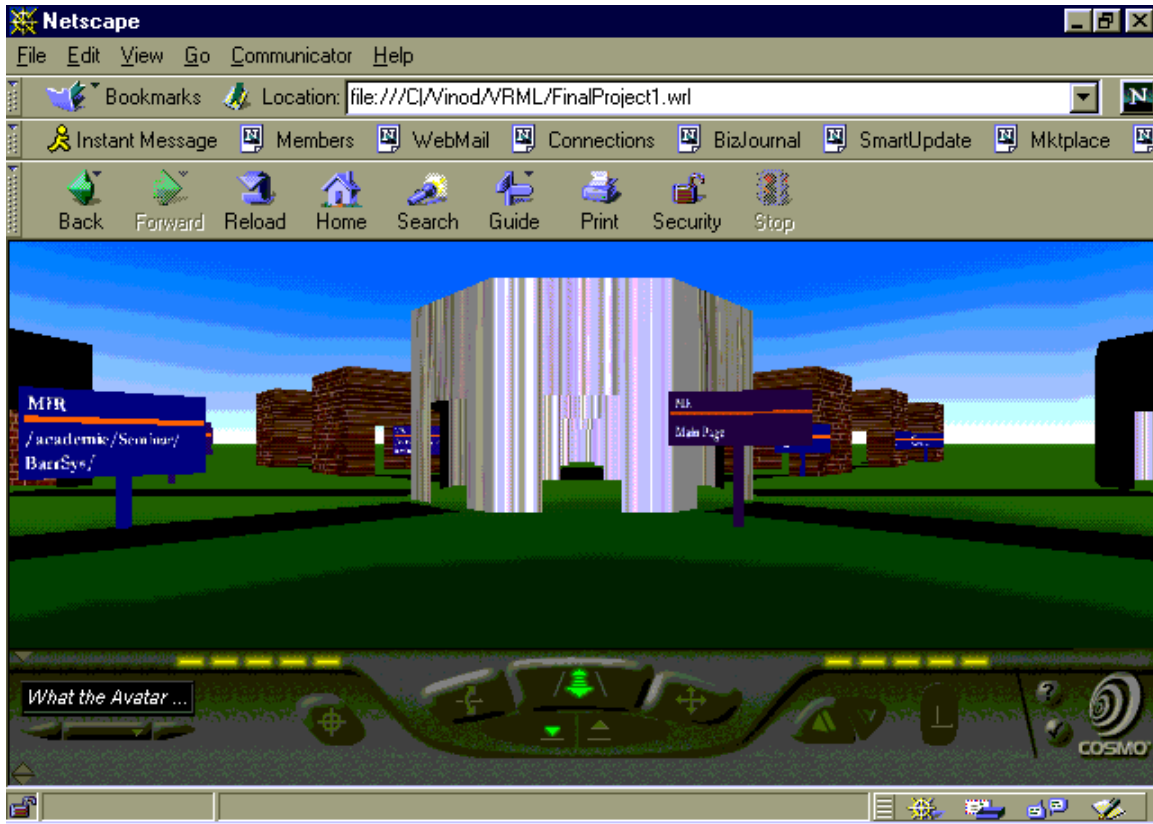


Figure 12: The VRML World from the Avatar's Viewpoint

Step 5: Click on the AVATAR to initiate its motion. Links to the individual pages are presented in the “buildings” if the user wants to access them by a mouse click.

The MIR protocol is thus applied to create a visual representation of the CISE website using an architecture metaphor and the user navigates through this information space. The following chapter will present some concluding remarks and areas of future expansion and research.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

This thesis proposes the use of metaphors to represent information for computerized visualization. The use of spatial metaphors is stressed. It utilizes the idea that the basics of form, space, and order along with an understanding of human perception and visual processing can be effectively used to implement a spatial metaphor to form an interface to data. The application of virtual reality to provide an information space for exploration of the data is demonstrated in the case study. This study uses a university campus metaphor to visually represent the behavior of a visitor to the website of the Computer and Information Science and Engineering [CISE] department at the University of Florida. The implementation is coded using Perl, CGI, Sybase, and VRML. The idea, in its present form, is intended more as aide to the industrial psychologist to be used in his presentations on Internet consumer behavior than for a statistician. It is also more a tool for commercial websites than educational ones.

There ideas expressed in this thesis can be extended and modified to enable this concept to move from a presentation tool to an analysis tool—conforming more to its future application as a visual data mining (or visual Web mining) aide. Some of the following are specific to Web mining whereas others are applicable to the bigger area of web mining.

Influence on Web Site Design

MIR can be used as a visual aide to design websites which will not only be effective for the commercial purpose for which it is intended, but which will provide a rich context of animation to study the use of that website. If the architecture metaphor is used, the traditional tree structure used in ordering the hierarchy of the web pages translates better than a structure where all the pages are linked to one main page and all navigation is directed towards that one page. Developing the MIR tool in conjunction with the web designer will be the most efficient way to ensure ease of understanding of the final visual.

Visualizing Data Warehouses

The visualization of data in a warehouse using statistical images is available commercially. But the images used are statistical translating data into graphs, curves, etc. MIR can be used for the same purpose. Moreover, the metaphor can be extended to the design of the warehouse schema. Consider Web mining:, if a warehouse backend is being used to store the web access data, a snowflake schema may be proposed with each dimension in the snowflake schema representing the traffic and demographic data pertaining to one page. In other words, if the architecture metaphor I used, just as a room in a house or a building in a campus represent one web page, a dimension in the underlying schema is representative of one web page and contains the data for that web page. The top level is the visual of the room, the second level is the one dimension of the backend warehouse schema that defines the data about accesses to that particular room (i.e., web page) and the third and bottom level is the actual table that holds the data.

Comparative Studies Using Multiple Animations

In the example implementation, only one AVATAR was used to depict the movement. But more than one can be used effectively. For example, if the users at a particular IP-address can be identified, sessions of the two individual users can be run simultaneously so that the similarities and differences in their behavior can be studied. This coupled with demographic data about the users, can be used effectively by a commercial site to alter their marketing strategy. Metaphorical mappings to other properties of the animated figure like speed and color can also be used effectively. To make the statistician also happy, traditional information representations can be combined with the metaphorical so that the user has a choice to see the statistical version if he wants. For example, a toggle between the three-dimensional architecture representation and a two-dimensional network representation with bar graphs in the third dimension could be an effective analysis tool.

MIR provides an insight into the possibility of exploration of information spaces visually to enable a general audience to experience data. The possibilities of extending the concepts presented here will make room for more research and development.

LIST OF REFERENCES

- [BLUM1996] Blumenthal, B. and Kuhn, W. "Spatialization: Spatial Metaphors for User Interaction." In Proceedings of Computer Human Interaction(pp. 346-347). Vancouver, BC: ACM Press. 1996
- [BUEC1998] Buechner, A. G. and Mulvenna, M. "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining," SIGMOD Record, 27(4):54-61, December 1998.
- [CARE1997] Carey, R and Bell, G. The Annotated VRML 2.0 Reference Manual. Reading, MA: Addison-Wesley Developers Press, 1997.
- [COOL1997] Cooley, R., Mobasher, B., and Srivastava, J. "Web Mining: Information and Pattern Discovery on the World Wide Web; WEBMINER – a survey," Tech Report, Department of Computer Science, University of Minnesota, Minneapolis, 1997
- [DIEB1998] Dieberger, A. and Frank, U. "A City Metaphor to Support Navigation in Complex Information Spaces." Journal of Visual Languages and Computing. 9:597-622, 1998
- [DYRE1997] Dyreson, C. "Using an Incomplete Data Cube as A Summary Data Sieve," Bulletin of the IEEE Technical Committee on Data Engineering:19-26, March 1997.
- [GRAY1996] Gray, J., Bosworth. A., Layman A., and Pirahesh, H. " Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-tab, and Subtotals." In IEEE 12th International Conference on Data Engineering, 152-159, 1996.
- [MARK1999] Markman A. Knowledge Representation. London: Lawrence Erlbaum Associates, Publishers, 1999.
- [MENA1999] Mena, J. Data Mining Your Website, Boston:Digital Press, 1999.
- [MITC1995] Mitchell, W. The City of Bits, Cambridge: MIT Press. 1995.
- [PLAN2000] Planec, P. " The Invisible Team: I Know You're Out There, But Does Your Company?" AV Video and Multimedia Producer Magazine, pp. 27-28, May 2000.

- [PITK1994] Pitkow, J and Bharat, K. "Webviz: a Tool for World Wide Web Access Log Analysis," In Proc. of the 1st International Conf on the World Wide Web, 1994.
- [TANN1999] Tanney, S. "A Spatial Metaphor for the Design of Digital Information Spaces," Masters Thesis, University of Washington, Seattle, 1999.
- [TUFT1983] Tufte, E. The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press, 1983.
- [UNDE1999] Underhill, P. Why We Buy: The Science of Shopping, New York: Simon & Schuster, 1999.
- [VEAL1995] Veale, T. "Metaphor, Memory and Meaning: Symbolic and Connectionist Issues in Metaphor Comprehension," Dissertation, Trinity College, Dublin, 1995.
- [VITR1960] Vitruvius, P. (unknown, BC). De architectura. (M. Morgan, Trans.) Vitruvius: The Ten Books on Architecture. New York: Dover Publications. 1960.

BIOGRAPHICAL SKETCH

Vinodkumar P. Kizhakke was born in the village of Chendamangalam in the state of Kerala, in southern India on the 28th of September, 1972. He completed his bachelor's degree in electronics and communication engineering in 1993 from Bharathiar University, Coimbatore, India. He proceeded to work as a research and development engineer with Phillips Medical Systems. He then started graduate study in computer engineering at the University of Florida. His areas of interest are databases, graphics and visualization.