



P-1906

VISUAL WEB MINING

By

Rajkumar. S

Reg. No. 71204621030

of

Kumaraguru College of Technology

Coimbatore

A PROJECT REPORT

Submitted to the

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

In partial fulfillment of the requirements

for the award of the degree

of

MASTER OF COMPUTER APPLICATIONS

July 2007



P-1906

Kumaraguru College of Technology

Coimbatore – 641006.

Department of Computer Applications

BONAFIDE CERTIFICATE

Certified that this project report titled **Visual Web Mining** is the bonafide work of **Mr. Rajkumar S** who carried out the research under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

N. Jeyakumaran

Project Guide



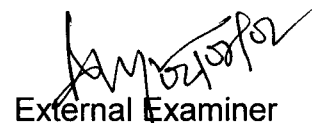
Head of the Department

Submitted for the University Examination held on

02/07/07



Internal Examiner



External Examiner

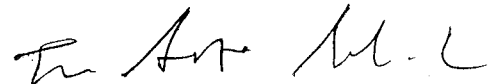
Date: 14/06/07

TO WHOMSOEVER IT MAY CONCERN

This is to certify that Mr **S.Rajkumar** (Reg No: 71204621030) final year student of Master of Computer Applications from Kumaraguru College of Technology, Coimbatore has successfully completed project titled "**VISUAL WEB MINING**" during the period from January 2007 to June 2007.

During this period, his work and conduct was found satisfactory, we wish him success in all his future endeavors.

For **Femtosoft Technologies.**,



Tom Anto Alen.L

Authorized Signatory

ABSTRACT

This project entitled as “**Visual Web Mining**” developed for visualize the web usage mining, integrating the information to the web mining. This Project uses Data Mining and Information Visualization techniques to the web domain. Mining is very important to find out all data sets. Analysis of web site usage data involves two significant challenges. The volume of data arising from the growth of the web and the structural complexity of web sites.

In response to the two challenges, this project proposes Data Mining techniques and applied to large web data sets and use Information Visualization methods on the results. Two or more web pages or web paths are accessed together frequently within a user session, the visualization techniques needs to be able to show all the associated web pages at once.

A disk tree or radial tree can be used to represent hierarchical information. The primary node or home page is located in the centre of the graph. The goal is to correlate the outcomes of mining Web Usage Logs and the extracted Web Structure and proposes several new information visualization diagrams.

The data summarization extracts the fields of session data and displays the results graphically. It displays the visits/day, visits/hour, top browsers, Os, referrer. The hits represents the number of visitors accessed particular website from datewise. It can be displayed as pie chart, barchart, linechart. The frequently visiting websites and search engines can be displayed graphically.

The system is developed in .NET framework with C#.NET as tool for the design of GUI and data validating and processing in the visualization process. These packages manage the front-end screens and the back-end databases.

ACKNOWLEDGEMENT

I would like to express my gratitude and humble thanks to our beloved principal **Dr. Joseph. V. Thanikal**, for having given me the adequate support and opportunity for completing this project work successfully.

I would like to express my deep sense of gratitude to **Dr. M. Gururajan** HOD, Department of Computer Applications for providing moral support towards this project work.

I wish to thank **Mr. A. Muthukumar**, Course Coordinator, Department of Computer Applications for providing his valuable suggestions and encouragement through out my project.

I am deeply indebted to **Mr. N. Jayakanthan**, lecturer, Department of Computer Applications, my internal project guide for offering his guidance, timely encouragement and support to me for the completion of this project.

I would like to express my sincere thanks to **Mr. L. Tom Anto Alen** manager, Femtosoft Technology Solutions Pvt., Limited, for providing all the desired documents and details regarding the various aspects necessary to the proper functioning of the system.

Finally I thank my lovable parents and friends who helped me in many ways during the course of project and have made it great success.

TABLE OF CONTENTS

Topic	Page No.
Abstract	iii
List of Tables	vii
List of Figures	viii
1. Introduction	1
1.1 System Overview	1
1.2 Company Profile	3
2. System Study and Analysis	5
2.1 Problem Statement	5
2.2 Existing System	6
2.2.1 Drawbacks of the Existing System	6
2.3 Proposed System	6
2.3.1 Advantages of the Proposed System	7
2.4 Feasibility Analysis	7
2.4.1 Technical Feasibility	7
2.4.2 Operational Feasibility	8
2.4.3 Economic Feasibility	8
2.5 Users of the System	8
3. Development Environment	10
3.1 Hardware Requirements	10
3.2 Software Requirements	10
3.3 Programming Environment	11
3.3.1 C#.NET	11
3.3.2 MS SQL Server	14

4. System Design and Development	16
4.1 Elements of Design	16
4.1.1 Modular Design	17
4.1.2 Input Design	21
4.1.3 Output Design	21
4.1.4 Database Design	23
4.2 Table Structure	24
4.3 Data Flow Diagram	27
5. Implementation	30
5.1 System Verification	30
5.2 System Validation	31
5.3 Testing	32
5.3.1 Unit Testing	33
5.3.2 Integration Testing	34
5.3.3 System Testing	34
6. Conclusion and Future Enhancement	35
6.1 Conclusion	35
6.2 Future Enhancement	36
Appendices	37
References	46

LIST OF TABLES

Table Description	Page No
Table 4.2.1 LOG TABLE	24
Table 4.2.2 CLEANED LOG TABLE	24
Table 4.2.3 USER LOG TABLE	25
Table 4.2.4 SESSION LOG TABLE	26

LIST OF FIGURES

	Figure Description	Page No
Figure 4.3.1	Level 0 DFD	27
Figure 4.3.2	Level 1 DFD	28
Figure 4.3.3	Level 2 DFD	29

CHAPTER 1

INTRODUCTION

1.1 SYSTEM OVERVIEW

This Project entitled as “Visual Web Mining” developed for visualize the web usage mining using Data Mining concepts and Information Visualization techniques. This Project uses Data Mining and Information Visualization techniques to the web domain in order to benefit from the power of both human visual perception and computing. Data Mining techniques are applied to large web data sets and use Information Visualization methods on the results.

Visual web mining is to correlate the outcomes of mining Web Usage Logs and the extracted Web Structure. This project undergoes several processes for information visualization diagrams and analyzes their utility and elaborate on the architecture of a prototype implementation.

In data cleaning process, the log text file is imported from our system. The log text files are converted to database file through parsing. Parsing is the process of splitting the log file using web patterns and log table is created. The log table consists of non-analyzed resources like jpg, gif, robot files. The non-analyzed resources are removed by filtering and cleaned log table is created from the log table.

The user identification process assigns user id to the unique ip address. It assign user id from cleaned log table. Session identification process gets input from the user identity file for each user id, when the time differences of subsequent page exceeds the threshold time then open the new session otherwise add it to the same session. The useless session is removed using session filtering process.

The data summarization process extracts the fields of session data and displays the results graphically. It displays the visits/day, visits/hour, top browsers, Os, referrer. The hits represents the number of visitors accessed particular website from datewise. It can be displayed as pie chart ,bar chart, line chart.The frequently visiting websites and search engines can be displayed graphically.

Visualization techniques are used to show all the associated web pages at once. A disk tree or radial tree can be used to represent hierarchical information. The primary node or home page is located in the centre of the graph. Visualization process displays URL list, page list, users list, frequently accessed page.

1.2 COMPANY PROFILE

Femtosoft Technology Solutions Private Ltd is one of the fast growing information technology in Chennai. It was started in 1998 and today it is an integrated IT solutions in networking, e-business solutions, client server applications, web- applications development, outsourcing solutions, software development, professional services and embedded systems.

Femtosoft Technology is a leading global consulting and IT services company, offering a wide array of solutions customized for a range of key verticals and horizontals. It works with state-of-art technology to give the most cost-effective solutions for any IT applications. Femtosoft has more than 100 employees and now an aim to become one of India's most respected IT service organizations with a targeted turnover more than Rs.1crore by the year 2006. The company's clients include organizations from the Private sector.

Femtosoft Technology solutions is Web-based & Software Development Service Company providing custom technology solutions to enterprises worldwide, combining proven expertise in technology, and an understanding of emerging business domains, that includes e-business solutions, client server applications, web- applications development, and outsourcing solutions. We develop innovative and creative products, services and concepts providing total information and communication solutions.

Application Management processes are premeditated on ensuring that systems are flexible and evolve with customer business. Their approach is geared to meet their commitments to the customers. Enterprises have significant investments in IT assets to execute their business and derive appropriate returns. These investments are made over a period of time and spread across the technology spectrum; spanning from legacy platforms to client-server systems to more contemporary multi-tier browser based systems.

Femtosoft Technology has been established by a group of professionals and technocrats with decades of experience in software development, project management, and general management with a group of young, energetic, qualified professionals under the supervision and guidance of an experienced management team power the company's activities.

CHAPTER 2

SYSTEM STUDY AND ANALYSIS

2.1 PROBLEM STATEMENT

The Site provides links to Web sites and access to content and services from third parties, including users, advertisers, affiliates and sponsors of the Site. Visual Mining is not responsible for the availability of content provided on, third party Web sites. The project should refer to the policies posted by other Web sites regarding privacy and other topics.

Visual Mining is not responsible for third party content accessible through the Site, including opinions, advice, statements and advertisements, and understands that you bear all risks associated with the use of such content. Visual Mining is not responsible for the quality of third party products or services and fulfilling any of the terms of your agreement with the seller, including delivery of products or services and warranty obligations related to purchased products or services.

2.3.1 Advantages of proposed system

- Web usage can be graphically displayed as a output report.
- WebPages developed are validated at both the server side as well as the client side.
- Separate tables are used to store separate information Data manipulation is easily done.
- Users can view the status of the request made by him.
- The users can provide feedback about the links of the website.
- The generation of report can be customized, which helps to improve the visual web mining performance.

2.4 Feasibility Analysis

A feasibility study is conducted to select the best system that meets performance requirements. This entity of identification description, an evaluation of candidate systems, and the selection of the best system for the job.

- Economic Feasibility.
- Technical Feasibility.
- Behavioral Feasibility.

2.4.1 Technical Feasibility

Technical analysis centers on the existing computer system (hardware, software etc.) and to what extent it can support the proposed addition. This involves financial considerations to accommodate technical enhancement. If the budget is a serious constraint, then the project is judged not feasible.

2.4.2 Operational Feasibility

An estimate should be made of how strong a reaction the user is likely to have toward the development of a computerized system. It is common knowledge the computer installations have something to do with turnover, transfers and changes in employee job status. Therefore it is understandable that the introduction of a candidate system requires special effort to educate, sell, and train the staff on new ways of conducting business.

2.4.3 Economic Feasibility

Economic analysis is the most frequently used method for evaluating the effectiveness of the candidate system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that benefits outweigh costs, and then the decision is made to design and implement the system. Otherwise, further justification or alterations in the proposed system will have to be made if it is to have a chance of being approved.

2.5 Users of the System

The users of the proposed system have been categorized as below and each of the user categories will have a set of rights which manage their use of the proposed system.

- Data cleaning
- User Identification
- Data summarization
- Visualizing

In data cleaning process, the log text file is imported from our system. The log text files are converted to database file through parsing. Parsing is the process of splitting the log file using web patterns and log table is created. The log table consists of non-analyzed resources like jpg, gif, robot files. The non-analyzed

resources are removed by filtering and cleaned log table is created from the log table.

The user identification process assigns user id to the unique ip address. It assign user id from cleaned log table. Session identification process gets input from the user identity file for each user id, when the time differences of subsequent page exceeds the threshold time then open the new session otherwise add it to the same session. The useless session is removed using session filtering process.

The data summarization process extracts the fields of session data and displays the results graphically. It displays the visits/day, visits/hour, top browsers, Os, referrer. The hits represents the number of visitors accessed particular website from datewise. It can be displayed as pie chart ,barchart, linechart. The frequently visiting websites and search engines can be displayed graphically.

Visualization techniques are used to show all the associated web pages at once. A disk tree or radial tree can be used to represent hierarchical information. The primary node or home page is located in the centre of the graph. Visualization process displays URL list, page list, users list, frequently accessed page.

CHAPTER 3

DEVELOPMENT ENVIRONMENT

3.1 HARDWARE REQUIREMENTS

The hardware support required for deploying the application

Server Configuration

Processor	:	Pentium 4 or above
RAM	:	Minimum 512 MB
Hard Disk	:	20GB or more

Client Configuration

Processor	:	Pentium 3/4
RAM	:	Minimum 256 MB

3.2 SOFTWARE REQUIREMENTS

The software support required for deployment is

Operating System	:	Windows XP
IDE	:	Visual Studio.Net 2003
Language	:	C#.NET
Database	:	Microsoft SQL Server 2000
Platform	:	.NET Framework
Application	:	C# Windows Applications

3.3 Programming Environment

3.3.1 C#.NET

C#.NET supports many new or improved object-oriented language features such as inheritance, overloading, the overrides keyword, interfaces, shared members, constructors, new data type, structured exception handling and delegates.

C#.NET is a pillar of the .NET Framework. It is a high-level programming language for the .NET Framework, and provides the easiest point of entry to .NET. The Microsoft. C Sharp namespace contains classes that support compilation and code generation using the C#.NET language.

C++ Developers can create multi threaded, scalable Applications using explicit free threading. Any CLS-Compliant Languages can use the classes, objects and components are creating in C#.NET.

C#.NET gets a new set of rich UI components called Win forms .The System. Win forms. forms name space contains classes for creating Windows-based applications that take full advantage of the rich user interface features available in the Microsoft Windows operating system C#.NET Win forms have built-in support for resizing controls as the user resize the form at runtime.

C#.NET provides an environment that's common to all languages, known as Integrated Development Environment (IDE). The IDE provides tools for designing, executing, debugging the Applications. The Form Designer is one of the most improved areas of C#.NET. IDE is used to create applications that have good appearance and user friendly. Windows form is the new platform for Windows application development, based on the .NET Framework. This framework provides a clear, object-oriented, extensible set of classes that enables to develop rich Windows applications. Additionally, Windows Form can act as local user interface in a multi-tier distributed solution.



P-1906

In C#.NET, the project groups become more integral to the development process, and they are called solutions. The Solution Explorer is displaying the Windows applications. In the Solution Explorer the solution has been given the same name, Windows Applications1, as the contained project. The top line in the Solution Explorer holds the name of the current solution, and below it is the line containing the name of the contained project, which in this case is also Windows Application.

The Solution Explorer displays the files in the current projects. For example, Form1.cs stores the form in the application, and Assembly Info.cs holds data about the current assembly. Windows Forms is the new platform for Microsoft Windows application development, based on the .NET Framework. This framework provides a clear, object oriented, extensible set of classes that enable to develop rich Windows application. Additionally, Windows Forms can act as the local user interface in a multi-tier distributed solution. Forms have properties that determine aspects of their appearance such as size, color and aspects of their behavior of resizable.

Each form is designed to perform a specific task like display data, give instructions, set up process etc., when we design a form, it become an interface that the user sees on the screen. Graphical control which include the label, Buttons, textboxes, scrollbars etc.

An event is an action which can respond to, or "handle" in code. Events can be generated by a user action, such as clicking the mouse or pressing a key, program code or by the system. For example, most objects will handle a click event, if a user clicks a form a code in the form's click event handler is executed.

The label is used to display read only text as for the user is concerned and the caption is altered in design time. In project it is used to display field names. Textboxes are commonly used for accepting user input or entering data and displaying the data. A Button can be clicked by using the mouse, enter key, or spacebar if the button has focus. It is used to get simple responses from the user or to invoke special functions on forms.

It contains a textbox and a list box. This allows the user to select an item from the dropdown list box, or to type in a selected in the textbox. Group box control is used to group the selected controls, so that each group is identified separately, in another group box the navigation buttons are displayed. Link label control is used for navigation purpose. It is used for giving link from one module to another.

The properties window exposes the various characteristics of a project namely form, classes and modules. The entire object that makes up the application all packed is as project. The property window displays the property of the selected control.

Code in C#.NET is written in class. The class should be in public.

Variables declared in general section are called global variables. These variables can be accessed throughout the program.

The Windows Forms checkbox control indicates whether a particular condition is on or off. It is commonly used to present a Yes/No or True/False selection to the user. The checkbox controls in groups are used to display multiple choices from which the user can select one or more.

To make the application more user-friendly toolbar is added to the form. A toolbar is represented by the `System.Windows.Forms.ToolBar` class. The application can have more than one toolbar in a form. A toolbar contains one or more buttons, represented by the `ToolBar Button` class. The image or an icon embedded as the display for each toolbar button. For this purpose, an `Image List` control is used and it acts as the container of the images.

3.3.2 Microsoft SQL Server

Microsoft SQL Server is a relational database management system (RDBMS) produced by Microsoft. Its primary query language is Transact-SQL, an implementation of the ANSI/ISO standard Structured Query Language (SQL) used by both Microsoft and Sybase. SQL Server is commonly used by businesses for small- to medium-sized databases, but the past five years have seen greater adoption of the product for larger enterprise databases.

Microsoft SQL Server uses a variant of SQL called T-SQL, or Transact-SQL, an implementation of SQL-92 (the ISO standard for SQL, certified in 1992) with many extensions. T-SQL mainly adds additional syntax for use in stored procedures, and affects the syntax of transaction support. (Note that SQL standards require Atomic, Consistent, Isolated, Durable or "ACID" transactions.) Microsoft SQL Server and Sybase/ASE both communicate over networks using an application-level protocol called Tabular Data Stream (TDS). The TDS protocol has also been implemented by the FreeTDS project in order to allow more kinds of client applications to communicate with Microsoft SQL Server and Sybase databases.

Microsoft SQL Server also supports Open Database Connectivity (ODBC). The latest release SQL Server 2005 also supports the ability to deliver client connectivity via the Web Services SOAP protocol. This allows non-Windows Clients to communicate cross platform with SQL Server. Microsoft has also released a certified JDBC driver to let Java Applications like BEA and IBM WebSphere communicate with Microsoft SQL Server 2000 and 2005.

Features of SQL Server

- Simplify the integration of back-end systems and data transfer.
- Derive additional value from data using sophisticated data mining tools.
- Obtain results quickly using Microsoft English Query, which allows users to pose questions in English instead of using Structured Query Language (SQL) or Multidimensional Expressions (MDX).
- Create business-to-business (B2B) and business-to-consumer (B2C) Web sites, analyze Web site trends, and implement personalization automatically using Microsoft Commerce Server 2000 and SQL Server 2000.
- Improve productivity with T-SQL enhancements.
- Take advantage of complete, end-to-end analysis capabilities-including data mining-with the integrated SQL Server 2000 extensible Analysis Services.
- Deliver robust, scalable database applications rapidly using the improved SQL Server 2000 development tools.
- Take full advantage of your hardware resources by running multiple, isolated applications on a single computer using SQL Server 2000 multi-instance support.

CHAPTER 4

SYSTEM DESIGN AND DEVELOPMENT

4.1 ELEMENTS OF DESIGN

System Design is the most creative and challenging phase in the development of a software system. Design implies to a description of the final system and the process by which it is developed. The first step is to determine what input data is needed for the system and then to design a database that will meet the requirements of the proposed system. The next step is to determine what outputs are needed from the system and the format of the output to be produced.

During the design of the proposed system some areas where attention is required are

- What are the inputs required and the outputs produced?
- How should the data be organized?
- What will be the processes involved in the system?
- How should the screen look?

The steps carried out in the design phase are as follows

- Modular Design
- Input Design
- Output Design
- Database Design

4.1.1 Modular Design

It is always difficult for any System Development team to grasp a system without breaking it into several smaller systems. These smaller systems will be a part of the original system yet they will be independent in the sense that they will incorporate within them the major functionalities of the proposed system.

A software system is always divided into several subsystems which make it easier to develop and perform tests on the whole system. The subsystems are known as the modules and the process of dividing an entire system into subsystems is known as Decomposition.

The modules identified for the proposed Visual Web Mining are as below

- Admin Login
- Data Cleaning
- User Identification
- Session Identification
- Session Filtering
- Data summarization
- Visualization

4.1.1.1 Admin Login

Admin login provided by the system is used to allocate various users who can browse through the web site. Administrator track and allocate the request sent by clients to the respective service personal. The job process is taken care by the personal and the status will be intimated to the administrator. The administrator stores the detail about the visits/day, visits/hour, top browser, OS, referrer.

4.1.1.2 Data Cleaning

The log file is the source of this module. It gets raw log data as input and produces the cleaned log data as the output. The Log data which is in the form of text file is imported from our system. The Log text files are converted in to database file by parsing.

Parsing

In data cleaning process, the log text file is imported from our system. The log text files are converted to database file through parsing. Parsing is the process of splitting the log file using web patterns and log table is created. The log table consists of non-analyzed resources like jpg, gif, robot files. The non-analyzed resources are removed by filtering and cleaned log table is created from the log table. The unwanted file in our log file stored like robots. These robot files are eliminated through data cleaning.

4.1.1.3 User Identification

Each and every system has a unique ip address .The system can have more than one Os or web browser. Different browser or Os consist of different ip in the same system. The user identification process assigns user id from the cleaned log data.

4.1.1.4 Session Identification

The unique ip is assigned in the user identification process. The session id is assigned from the unique ip. The accessing time for each ip address will differ, so unique ip can have different session id.

Session identification process gets input from the user identity file for each user id, when the time differences of subsequent page exceeds the threshold time then open the new session otherwise add it to the same session.

4.1.1.5 Session Filtering

The session filtering process gets the input from the session data and it removes the useless session, session of length one. The session filtering process gets the input from the session data. The useless session is removed using session ip. This process is mainly used for visualization purposes.

4.1.1.6 Data summarization

The data summarization process extracts the fields of the session data and display the results graphically. It displays the visits/day, visits/hour, top browser, OS, referrer etc. The hits represent the number of visitors accessed our website. The browser represents the count eg: mozilla firefox, internet explorer, and opera are viewed in chart.

- Hits
- Browsers
- Operating System

Hits

The hits represent the number of visitors accessed particular website. It can be displayed as pie chart, bar chart, line chart. It consists of report type such as daily, month wise, date wise.

Browsers

The browser represents the count of the web browser used and also identifies the top web browser used. For e.g.: mozilla fire fox, internet explorer, and opera are viewed in chart.

Operating System

The Operating System represents the type of the platform used frequently. For e.g.: windows vista/XP/2000/Me/98/95, UNIX, Macintosh, Linux are viewed chart.

4.1.1.7 Visualizing

Two or more web pages or web paths are accessed together frequently within a user session, the visualization techniques needs to be able to show all the associated web pages at once. A disk tree or radial tree can be used to represent hierarchical information. The primary node or home page is located in the centre of the graph. It displays URL list, users list, page list, frequently accessed page list.

4.1.2 Input design

The input design is the process of converting the user-oriented inputs into computer-based format. The goal of designing input data is to make sure that the automation is easy, logical and free from errors. Input design is one of the most expensive phases of the operation of computerized system and is often the major problem of a system.

The input design requirements such as user friendliness, consistent format and interactive dialogue which provide users with timely help and correct messages are given high priority.

In the project, the code generation page is made with several easy to use options. For example, to insert a node, we need not write the entire code. Only giving the name will insert all the necessary code itself.

In addition, syntax checking option is provided to check for the valid code entry. Checking often will reduce the wrong code and the input is easy now. In this software, importance is given to automation code generation system, which is an important factor in developing efficient and user-friendly software.

The input form in this project is

- User Login page

It is used by the various users of the system and the system restricts access to the data based on the type of user logged in.

4.1.3 Output Design

Output design generally refers to the results and information that are generated by the system for many end-users. The output is the main reason for developing the system and the basis on which they evaluate the usefulness of the application.

In the project, the design view i.e., visualization is the output page available. Moreover, the data can be display the results graphically. This process displays the visits/day, visits/hour, top browser, Os, referrer etc can be displayed as pie chart, bar chart, line chart.

The output reports produced by this project are as follows

- Hits Report

The hits represent the number of visitors accessed particular website. It can be displayed as pie chart, bar chart, line chart. It consists of report type such as daily, month wise, date wise.

- Browser Report

The browser represents the count of the web browser used and also identifies the top web browser used. For e.g.: mozilla fire fox, internet explorer, and opera are viewed in chart.

- Operating System Report

The Operating System represents the type of the platform used frequently. For e.g.: windows vista/XP/2000/Me/98/95, UNIX, Macintosh, Linux are viewed in chart.

- Visualization Report

Visualization techniques are used to show all the associated web pages at once. The primary node or home page is located in the centre of the graph. Visualization process displays URL list, page list, users list, frequently accessed page.

4.1.4 Database Design

Database design deals with the table structure and organization. The purpose of the database is to enable easy access of information for the user. The general theme behind databases is to handle the information as an integrated one.

While designing a database, we have to make decisions regarding how best to take some system in the real world and model it in a database. This process consists of deciding which tables to create and what columns they will contain as well as the relationships between tables. A database is an integrated collection of user related data stored with minimum redundancy, serves many users/application quickly and efficiently.

A database is a collection of inter-related data stored with minimum redundancy to serve many users quickly and efficiently. The general objective of database design is to make the data access easy, inexpensive and flexible to the user. An elegantly designed database can play a strong foundation for the whole system.

A database system is basically a computerized record keeping system, i.e., it is a computerized system whose overall purpose is to maintain information and make that information available on demand. DBMS is collections or inter related data and set of programs that allow several users to access and manipulate the data. Its main purpose is to provide users with an abstract view of the data, i.e. the system hides certain details of how the data is stored and maintained.

4.2 TABLE STRUCTURE

Table 4.2.1 LOG TABLE

Field Name	DataType	Width	Description
L_IP	Varchar	20	Primary key,Log ip
L_Identity	Varchar	20	Log identity
L_User	Varchar	20	Log user
L_Date	DateTime	8	Log date
L_DateTime	DateTime	8	Log date time
L_TimeZone	DateTime	8	Log time zone
L_Method	Varchar	10	Log method
L_Path	Varchar	120	Log path
L_Protocol	Varchar	10	Log protocol
L_Status	Int	8	Log status
L_Bytes	Int	8	Log bytes
L_Referer	Varchar	120	Log referrer
L_Agent	Varchar	120	Log agent

Table 4.2.2 CLEANED LOG TABLE

Field Name	Type	Width	Description
L_IP	Varchar	20	Primary key,Log ip
L_Identity	Varchar	20	Log identity
L_User	Varchar	20	Log user
L_Date	DateTime	8	Log date
L_DateTime	DateTime	8	Log date time
L_TimeZone	DateTime	8	Log time zone
L_Method	Varchar	10	Log method
L_Path	Varchar	120	Log path
L_Protocol	Varchar	10	Log protocol
L_Status	Int	8	Log status
L_Bytes	Int	8	Log bytes
L_Referer	Varchar	120	Log referrer
L_Agent	Varchar	120	Log agent

Table 4.2.3 USER LOG TABLE

Field Name	Type	Width	Description
L_Userid	Varchar	20	Primary key ,Log user identification
L_IP	Varchar	20	Foreign key, Log ip
L_Identity	Varchar	20	Log identity
L_User	Varchar	20	Log user
L_Date	DateTime	8	Log date
L_DateTime	DateTime	8	Log date time
L_TimeZone	DateTime	8	Log time zone
L_Method	Varchar	10	Log method
L_Path	Varchar	120	Log path
L_Protocol	Varchar	10	Log protocol
L_Status	Int	8	Log status
L_Bytes	Int	8	Log bytes
L_Referer	Varchar	120	Log referrer
L_Agent	Varchar	120	Log agent

Table 4.2.4 SESSION LOG TABLE

Field Name	Type	Width	Description
L_SessionId	Varchar	20	Primary key, Log session identification
L_UserId	Varchar	20	Foreign key, Log user identification
L_IP	Varchar	20	Foreign key, Log ip
L_Identity	Varchar	20	Log identity
L_User	Varchar	20	Log user
L_Date	DateTime	8	Log date
L_DateTime	DateTime	8	Log date time
L_TimeZone	DateTime	8	Log time zone
L_Method	Varchar	10	Log method
L_Path	Varchar	120	Log path
L_Protocol	Varchar	10	Log protocol
L_Status	Int	8	Log status
L_Bytes	Int	8	Log bytes
L_Referer	Varchar	120	Log referrer
L_Agent	Varchar	120	Log agent

4.3 DATA FLOW DIAGRAMS

Data flow diagrams are graphical representation depicting information regarding the flow of control and the transformation of data from input to output. The DFD may be used to represent the system or software at any level of abstraction. In fact, DFD can be partitioned into levels.

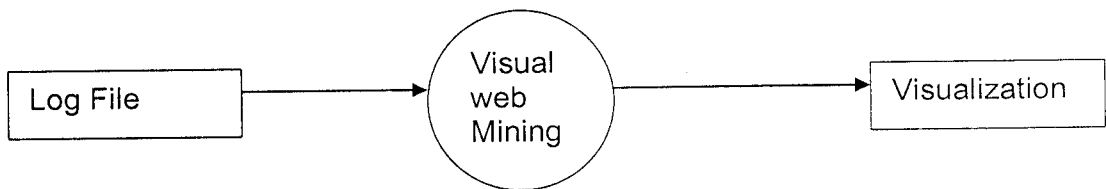
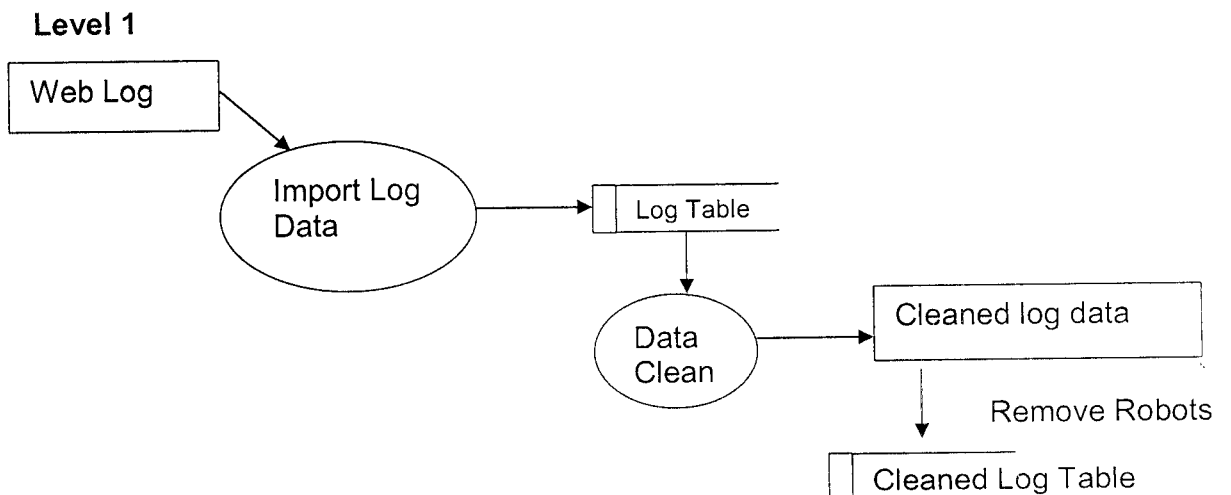


Figure 4.3.1: Level 0 DFD



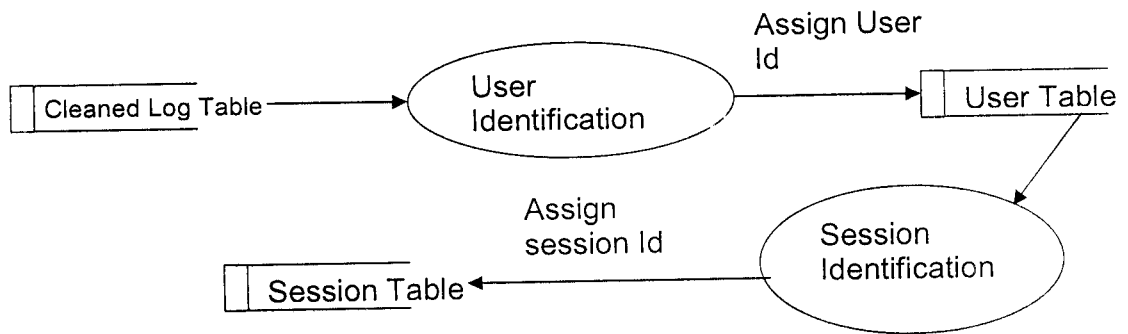


Figure 4.3.2: Level 1 DFD

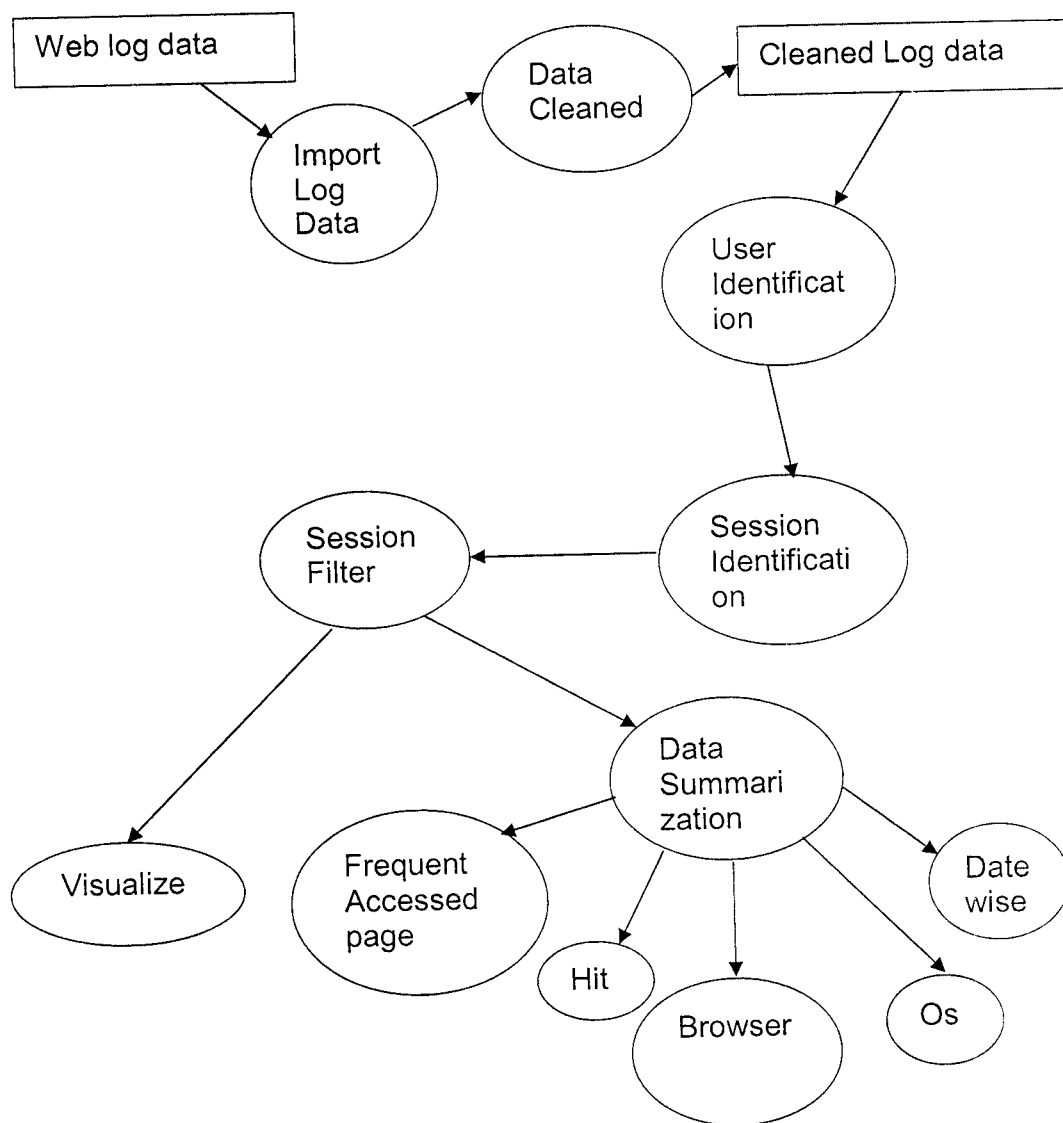


Figure 4.3.3: Level 2 DFD

CHAPTER 5

IMPLEMENTATION

System Implementation is the part of the software engineering life cycle, where, the design artifacts are converted to a working application. Coding is done in this stage using an apt framework and programming language, which would solve the specific problem the best way. Once the design is coded into a working application, it has to be verified, validated and tested in detail. The tested product if successful is deployed in the user environment.

5.1 SYSTEM VERIFICATION

System Verification is an important element of software quality assurance and represents the ultimate review of specification, design and loading. No system design is ever perfect communication problems, programmers negligence are time constraints create error that must be eliminated before the system is ready for user acceptance testing. Following system testing is acceptance testing or running the system with live data by the actual user.

The verification of the system is a means of assessing or measuring the system to determine quality. System verification makes a logical assumption that if all parts of the system are correct, the system will successfully be implemented. Inadequate testing or non-testing leads to errors that may not appear until months later. This creates the time lag between the cause and appearance of the problem and the effects of the system errors on files and records within the system. Hence, the aim of testing is to create bug free reliable and secured system. The implementation of newly designed package is an important phase in adopting a successful new system the implementation of the package involves testing, user training.

Acceptance and changeover the objective of testing is to discover errors. To fulfill these objectives a series of test steps-viz.-unit test, integration, validations and system test were planned and executed. A program may function perfectly in isolation but fall when interfaced with other modules. The approach is to test entity wish successively large ones, up to the system level. The test data were collected from the concern and the live data were used to test data.

5.2 SYSTEM VALIDATION

Validation answers the question "Am I building the right product?" This checks whether the developer is moving towards the right product, whether the development is moving towards the actual intended product that was agreed upon in the beginning. Validation also determines if the system complies with the requirements and performs functions for which it is intended and meets the organization's goals and user needs. It is traditional and is performed at the end of the project. In data access, it checks whether we are accessing the right data, in terms of data required to satisfy the requirement.

Validation is performed after a work product is produced against established criteria ensuring that the product integrates correctly into the environment. It determines the correctness of the final software product by a development project with respect to the user needs and requirements.

Functional validation is done in the Visual web Mining to check whether each of the functions are done correctly as expected in every page. Each control in a Screen is designed to do some function. These functions are checked against the requirements stated for them. For example, clicking 'Parse' button should take the corresponding action of Parsing the data into the database. Clicking the Assign user id icon should assign the user id internally and that are being currently displayed. This level of validation can continue to all the controls in the system. This checking is usually done after the system is developed so that all activities that are affected can be checked.

Field level validation is done in Visual Web Mining to check whether each of the fields either accepts the data as expected and do the client side validation of data entered. For example, a field level validation on a text box would check against the type of data entered and follow rules such as length of entry etc. The data type validation checks are conducted after the form is submitted. If the validation check fails then the processing stops and the control returns back to the original form that was submitted.

The validation is done in a step by step process. First the screen is loaded with the controls. When the user moves between controls on the screen, the validation events for the control that lost the focus are fired and appropriate error messages (if any) are displayed. If the user generates a form save request, the entire form is evaluated for any validation controls that are not valid. If even one control is not valid, the form will not be submitted.

5.3 TESTING

Testing is a critical element of software quality and assurance and represents the ultimate review of specification design and coding. It is a vital activity that has to be enforced in the development of any system. This could be done in parallel during all the phases of system development. The feedback received from these tests can be used for further enhancement of the system under consideration. The testing phase conducts test using the Software Requirement Specification as a reference and with the goal to see whether the system satisfies the specified requirements.

The proper choice of test data is an important as the test itself. If test data as input are not valid or representative of the data to be provided by the user, then the reliability of the output is suspect. Test data may be artificial. Properly created artificial data should provide all combinations of values and formats and make it possible to test all logic and transaction path subroutines.

The testing done will differ in nature and will have different objectives at each level. The focus of all testing will be to find errors, but different type of errors are looked for at each level.

The levels of testing in this project will be:

- Unit Testing
- Integration Testing
- System Testing

5.3.1 Unit Testing

Unit testing is the lowest level of testing and its function is to test the functionality of basic unit of software in isolation. This is a white box testing where the most detailed investigation of the internal functions of every individual unit is carried out.

The unit test plan describes the features and functionality that is to be tested for each unit. The purpose of unit testing is to find errors, which could be data or logic related errors and also prove that the individual units are robust and fit for purpose they are developed.

The typical tests that will be carried out during unit test include:

- Data validation to check valid and invalid data entered into a text box.
- Field length check to check the maximum length of the field.
- Errors handling to check appropriate front-end validations are being carried out.
- Database validations to check the data entered in the front end is stored into appropriate table in the database.
- Test to ensure that all paths are traversed and branching takes properly.
- Verify operation outside range values.
- Verify operation at normal value range.

- Ensure that all loops are terminated successfully.
- Identify and remove abnormal termination of all loops.

5.3.2 Integration Testing

An integration test plan outlines the process and procedure to be followed for integration testing. Integration testing involves the process of testing two or more tested units that have been fully integrated. The integration testing should look for errors in the following.

- The interfaces between the tested units.
- The function that can be performed by the integrated unit.

Programs are invariably related to one another and interact in total system. Each program is tested to see whether it confirms to related program in the system. Each portion of the system is tested against the entire module with both test and live data before the entire system is ready to be implemented.

A bottom-up strategy will be followed for integration testing. This would involve integrating the bottom units with the calling units and test the calling functions. Bottom-up test assures that the lower level modules are tested before testing the higher-level modules, which invoke them. The global variables were traced such that they hold data related to the current module.

5.3.3 System Testing

System testing is actually a series of different tests, whose primary purpose is to fully exercise the computer-based system. This helps in verifying that all the system elements have been properly integrated and perform the allocated functions. It verifies the entire product after having integrated all software and hardware components, and validates it according to the original project requirement.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

The Project mainly make easy way to visualize the web usage. Our project uses the web mining of organizational web sites. Mining is very important to find out the all data sets. We propose a new effective and memory efficient pruning technique, which, unlike other previous proposals, does not require the whole set of closed patterns mined so far to be kept in the main memory. More importantly, association rules extracted from closed item sets have been proven to be more meaningful for analysts, because all redundancies are discarded.

The goal of this project is to discuss the development of a visual web mining prototype called Web Patterns which allows the user to effectively visualize web usage patterns. The system should provide a confirmed security. Existing system does not have visualization part. Thus the proposed system should overcome these demerits by visualizing the web usage.

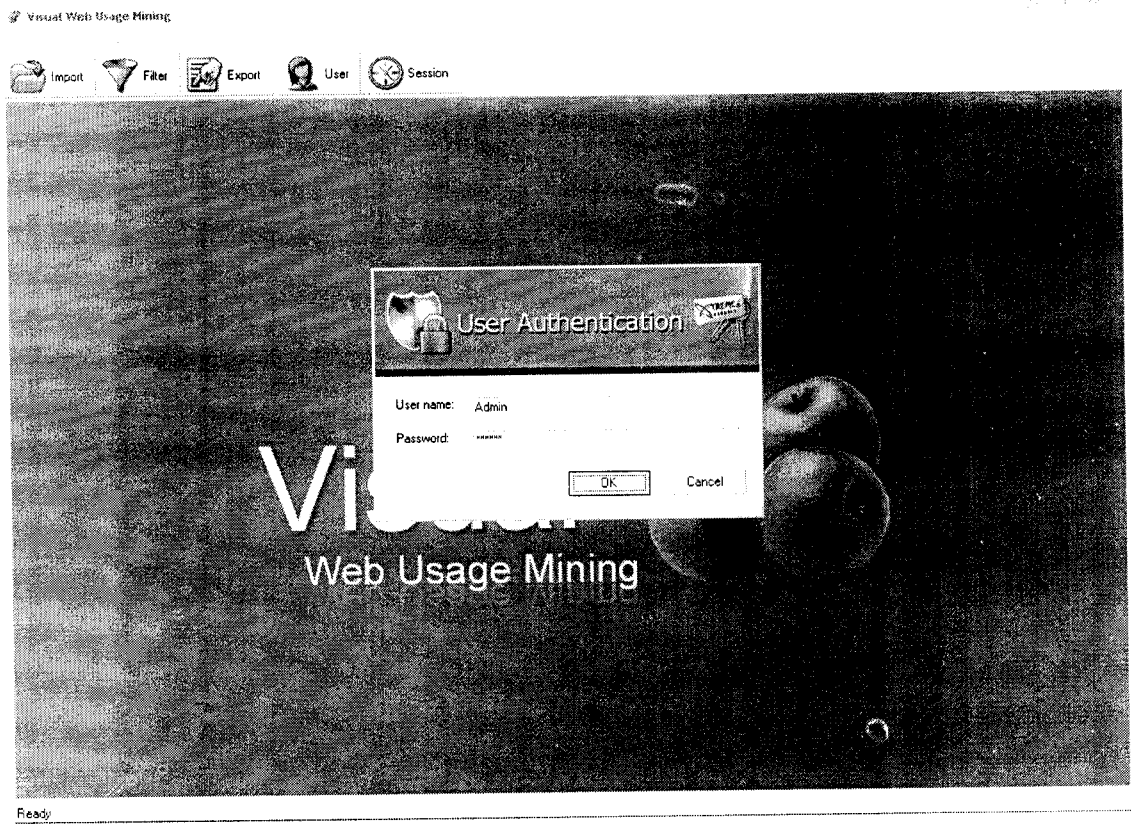
6.2 FUTURE ENHANCEMENT

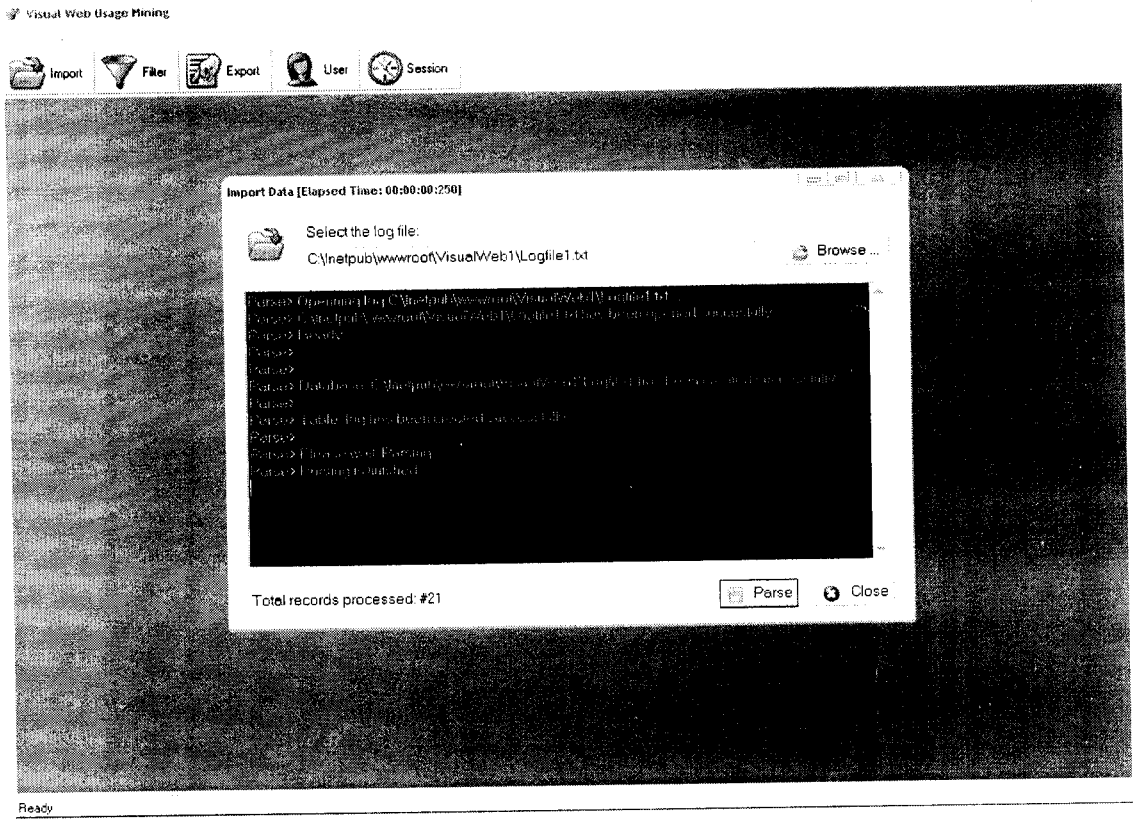
Demonstrating the utility of web mining can be done by making exploratory changes to web sites, e.g., adding links and then extracting, visualizing and interpreting changes in access patterns. This may also require running our implementation on logs obtained over longer period of time.

This convoluted analysis is necessary to discover useful patterns and understand the navigational behaviors of web site visitors, whether to improve web site structures, provide intelligent on-line tools or offer support to human decision makers.

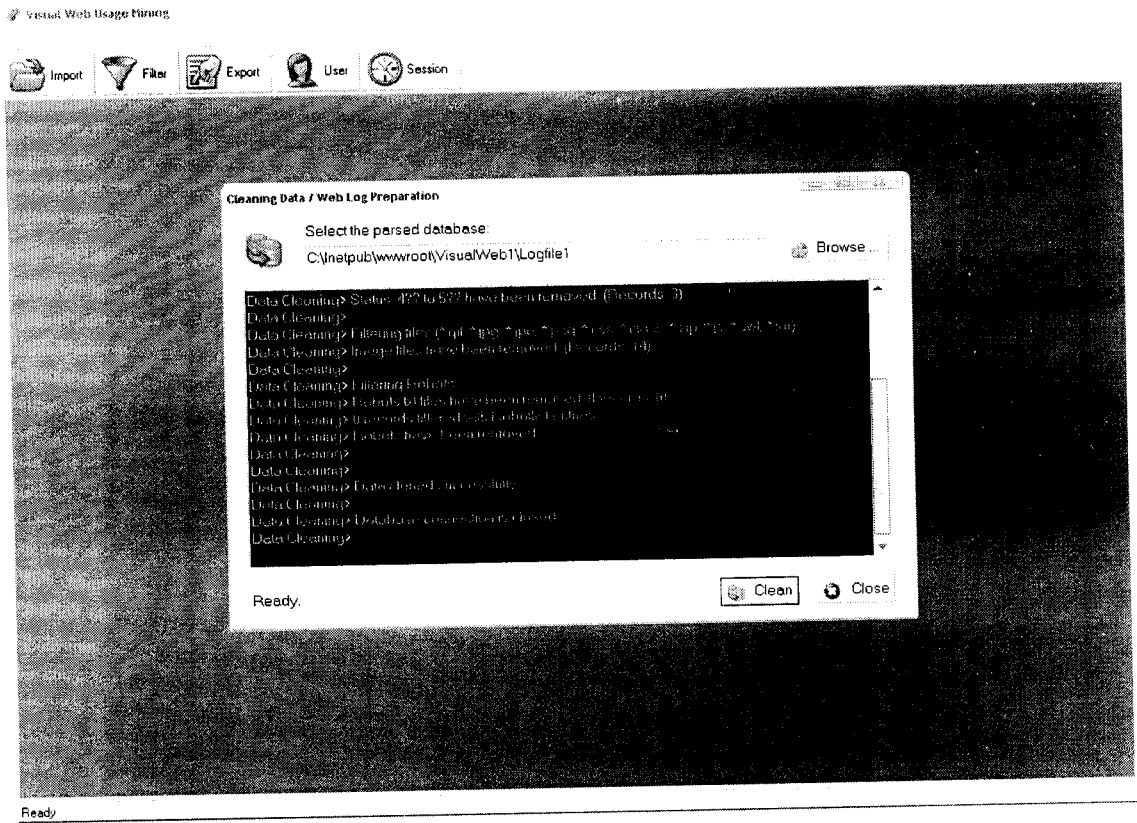
APPENDICES

LOGIN SCREEN

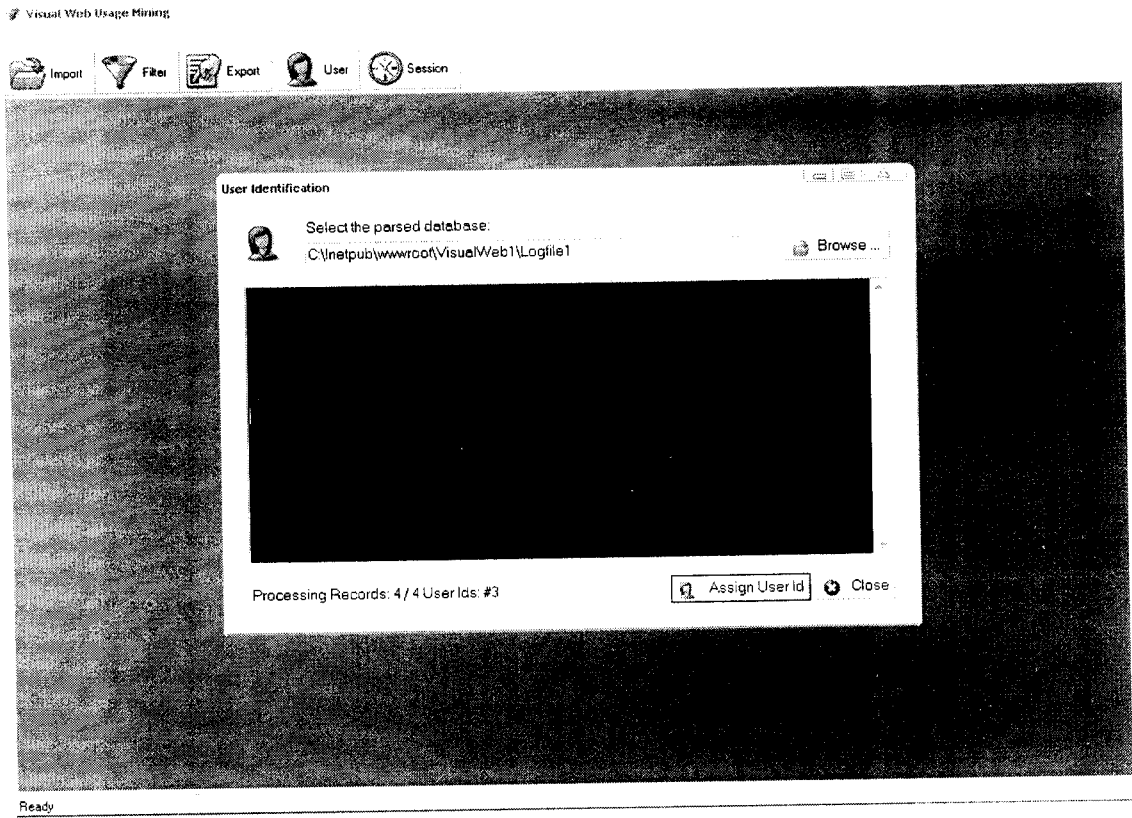




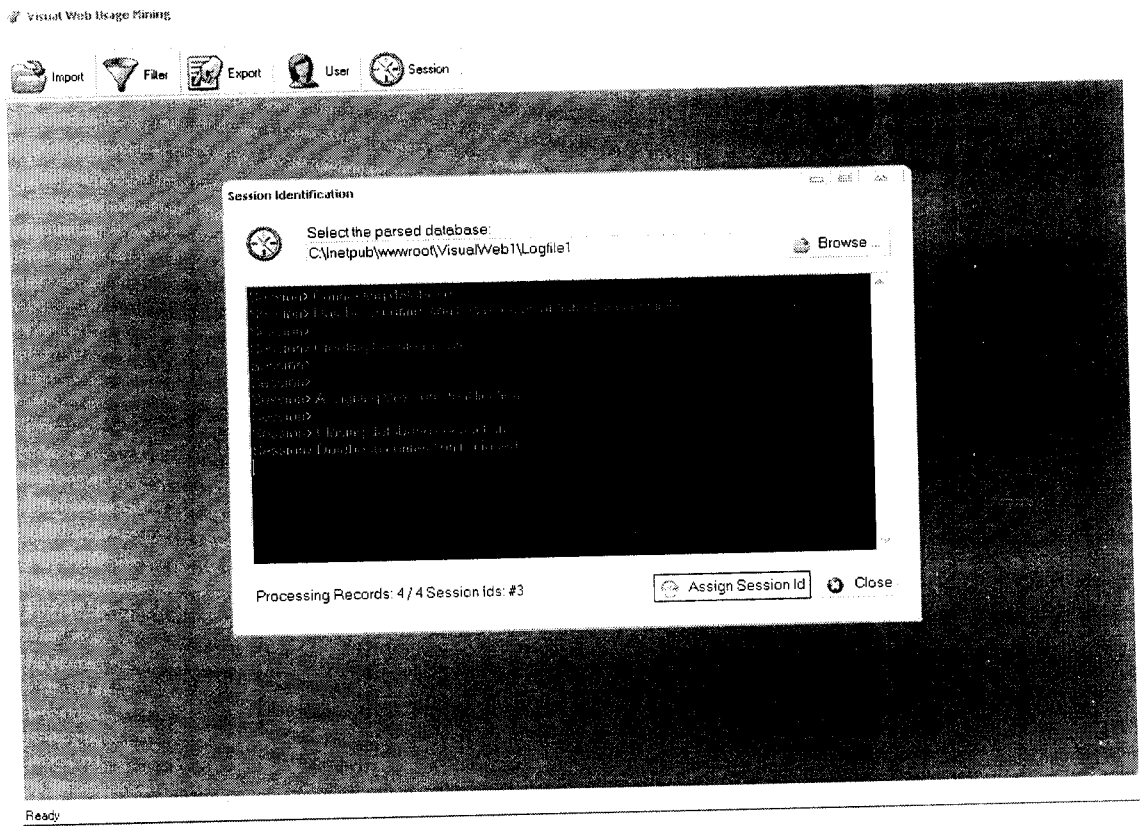
Parsed State



Data Cleaning



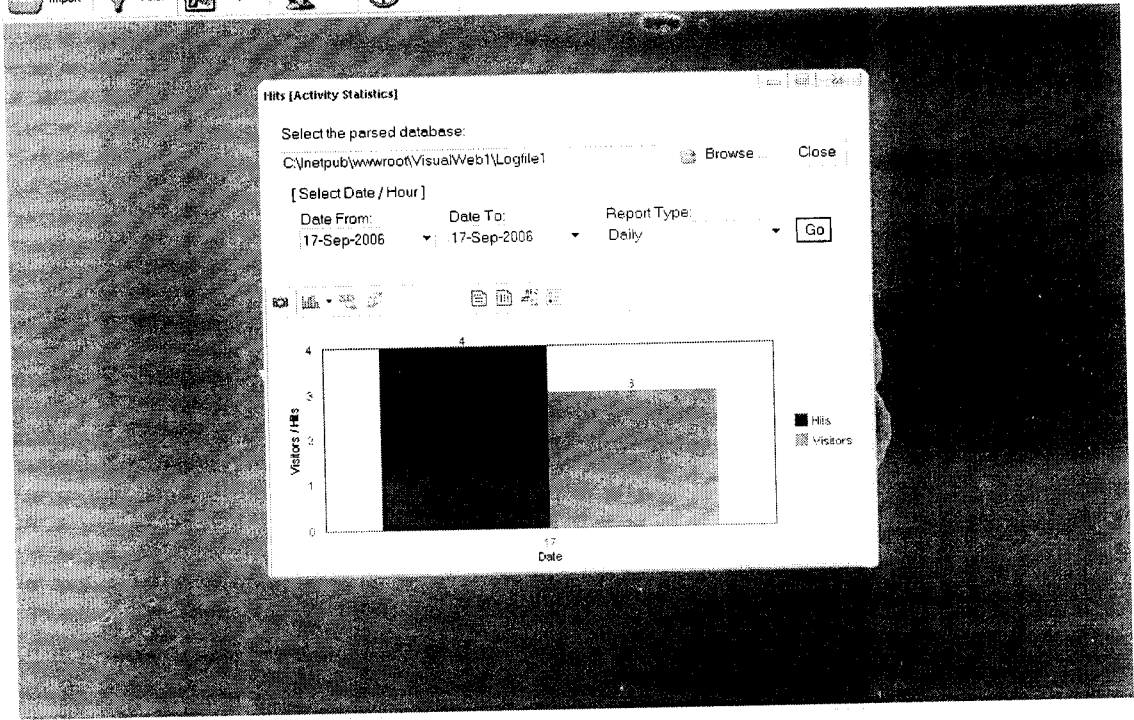
User Identification



Session Identification

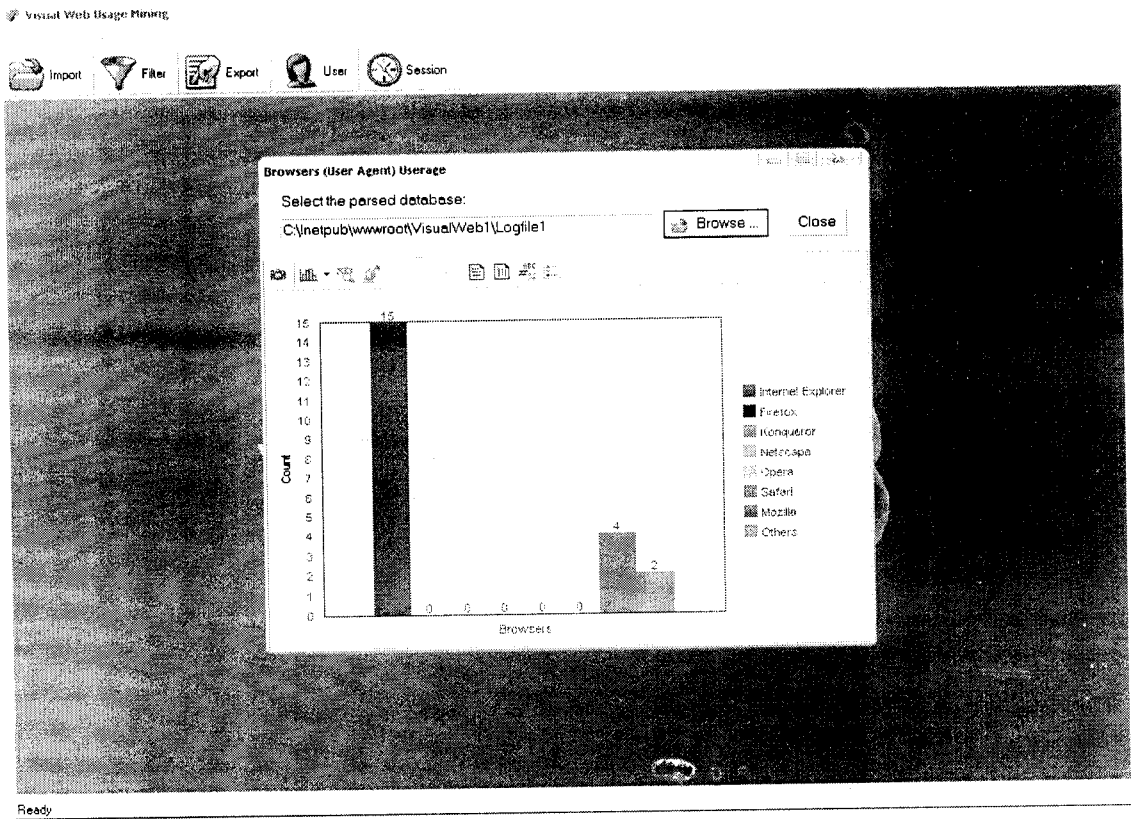
Visual Web Usage Mining

Import Filter Export User Session

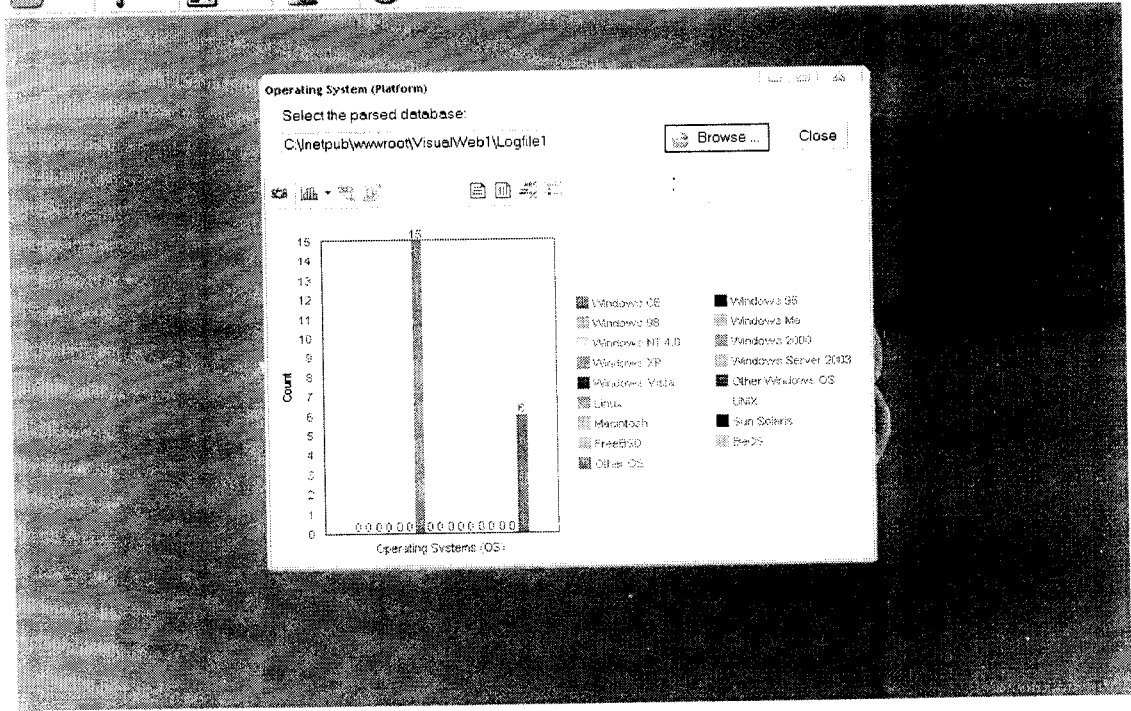


Ready

Hits Report

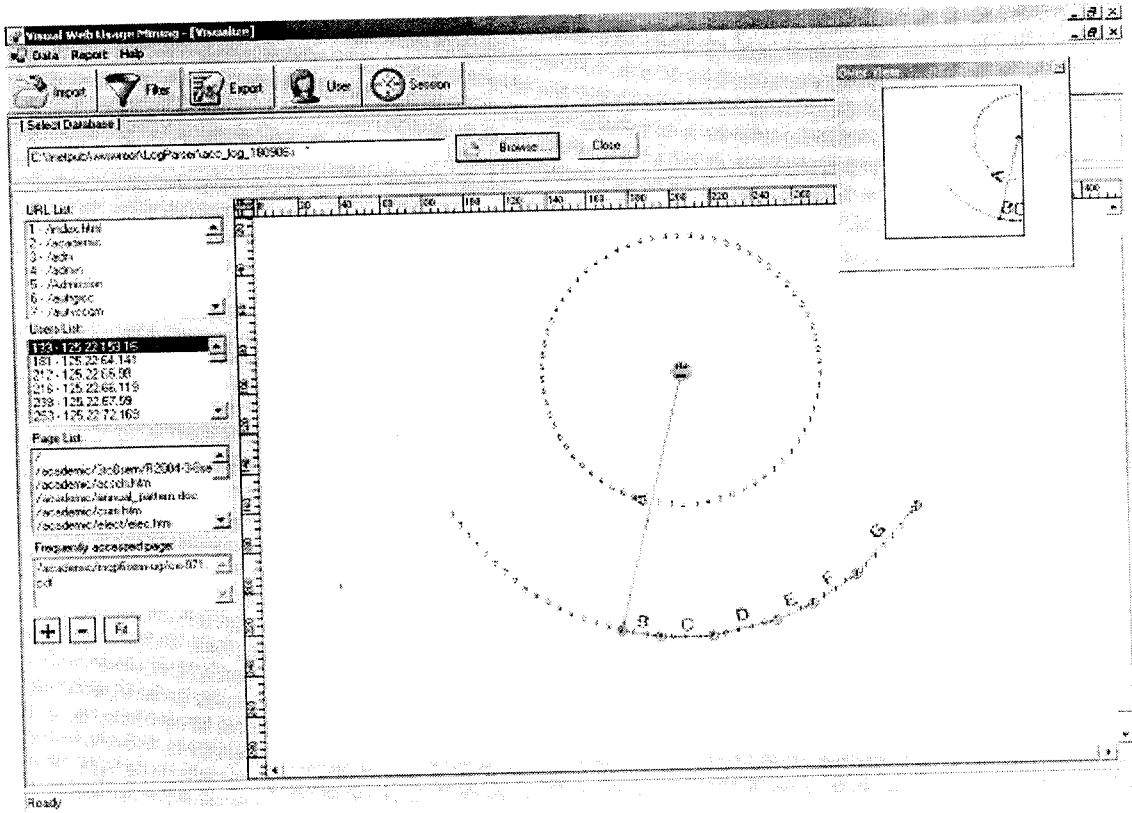


Operating system Report



Ready

Browsers Report



Visualization

REFERENCES

1. Jeff Ferguson "C# Programming Bible", WILEY, Second Edition 2003.
2. Peter Wright "Beginning Visual C# 2005 Express Edition – From Novice to Professional" Apress 2005.
3. Shapiro, Jeffrey "Microsoft SQL server 2005: The Complete Reference", Tata McGraw-Hill Publishing Company Limited.
4. Roger S. Pressman – Software Engineering – A Practitioners approach, Fourth Edition, Tata McGraw-Hill Publications, 1998.

Web References

1. www.dotnet247.com
2. www.w3schools.com
3. www.devarticles.com
4. www.developer.com