International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

PERFORMANCE ANALYSIS OF RMT WEBLOG ANALYZER USING WUM TECHNIQUE

[*1]Ms. Deepika D., [*2] Mrs. Shoba S.A,

[*1] M.Phil Research Scholar, PG & Research Department of Computer Science & Information Technology Arcot Sri Mahalakshmi Women's College, Vellore, Tamil Nadu, India.

[*2] Asst.prof, HOD Of PG & Research Department of Computer Science & Information Technology Arcot Sri Mahalakshmi Women's College, Vilapakkam, Vellore, Tamil Nadu, India.

Abstract –

Web usage mining is crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned. Web usage mining is also helpful for identifying or improving the visitors of a particular Website by accessing the log file of that site. In this paper the focus is on Web usage mining of Log data of an educational institution.

Web usage mining (WUM) also known as Web Log Mining is the application of Data Mining. WUM techniques are applied on large volume of data to extract useful and interesting patterns from Web data, specifically from web logs, in order to improve web based applications. Web usage mining consists of four phases, data source, pre-processing, pattern discovery, and pattern analysis. After the completion of these four phases the user can find the required usage patterns and use this information for the specific needs in a variety of ways such as improvement of the Web application, identifying the visitor's behaviour, customer attraction, Customer retention etc

Key Words: *Customer Relationship Management (CRM), Web usage mining (WUM). Customer attraction, Data mining Techniques*

I. INTRODUCTION

1.1 OVERVIEW OF THE SYSTEM

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services. Web mining is a popular technique for analyzing the visitor's activities in e-service systems.

Web mining research can be classified into 3 types:

a) Web content mining (WCM)
b) Web structure mining (WSM)
c) Web usage mining (WUM).

Web content mining refers to the discovery of useful information from web contents including text, image, audio and video etc., Web structure mining gives the analysis of the out links of a webpage and it has been used for search engine result ranking. Web usage mining refers analyzing search logs and also to learn the user profiles.

1.1.1 Project Scope and Objectives

The objective of the project is to generalize the log file of a web site obtained from a Web Server using Attribute-Oriented Induction technique.

Context:

This can be used for identifying the frequent access pattern for any web site. Accordingly, website can be enhanced.

Web Mining:

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*.

1.1.2 Management and Technical Constraints

- This software will handle only the given log file and will not generate log files for a particular web site.
- It supports only combined and common format.
- It does not provide security while accessing the database.
- It can handle only particular log file and cannot dynamically access log files of different websites.

1.1.3 Computer Networks

Local area network is the interaction between several computer connected to a common server. Remote Monitor is an application that assists the system administrator of a network to manage his network. This is a client – server application to remotely monitor the usage levels of a system. This thesis concentrates on network performance monitoring, a part of network management where the main statistics such as network traffic, interface, system details, memory and process details are depicted as interactive tables and pie charts.

A LAN (Local Area Network) is a private network that is contained within an enterprise. It is a collection of computers connected together by using Network Ethernet Card. The main purpose of LAN is to share users' information and computing resources among employees. A LAN can also be used to facilitate working in groups and for teleconferences.

II. SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

Whenever we are developing any application software to provide network service, if the network is being used to develop the software is unstable then we get unsatisfactory IT services each time the network goes down. So it is necessary to do network management.

The goal of network management is to ensure that the users of network receive the network services with high quality. It is a combination of Fault management, Performance management, Configuration management, Security management and Accounting management.

2.1.1 DISADVANTAGES OF EXISTING SYSTEM

1. Existing system, implemented using the windows platform, so, it causes problem while using the software across the platforms.
2. Existing system shows all the details about the system information in the text and numerical format so, that is less interaction with the user.
3. It does not provide the facility for saving the charts.

2.2 PROPOSED SYSTEM

Web usage mining mines web log records to discover web access pattern of web pages. Analyzing and exploring identifying potential customers for e-commerce enhance the quality and delivery of internet information services to end user and improve web server system performance.

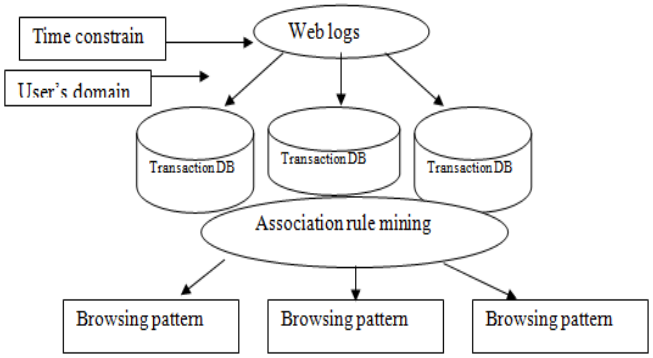

Figure 2.1: Design of Web log system

The log file contents are retrieved from text file and tokens are separated by using String Tokenizer. The contents are then stored into a database.

1 Unwanted Tuples are then removed and stored in another table.
2 Aggregate functions are used for extracting the required tuples. SQL Queries are passed to database

LOG FILE

Log files are files that contain a record of website activity. Every time a person visits the website, a log file is updated with the visitor's information by the web server. These log files can be downloaded and used to generate useful statistics.

An access of a web page or a file will generate a "Hit" on the web server. For example, if a web page contains 10 pictures, a visit on that page will generate 11 "hits" on the web server, one hit for the web page, 10 hits for the pictures. If a visitor viewed 5 web pages on the web site, each page contain 10 pictures, the web server will record:

WEBLOG FILES

Web Server log files are simple text files that are automatically generated every time someone accesses the Website. Every "hit" of the Web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of text for each hit to the website. This contains information about who was visiting the site, where they came from, and exactly what they were doing on the particular Web site.

2.2.1 ADVANTAGES OF PROPOSED SYSTEM

- GUI Representation of Web Performance with the use of Web Charts
- GUI Representation of Network performance with the use of tables and charts.
- Advantage of save the graphs (Line Chart & Pie Chart).
- Monitoring the system performance using the tables.

III. SOFTWARE REQUIREMENT SPECIFICATIONS

3.1 JAVA

Java was conceived by James Gosling, Patrick Naughton, Chris Warth, Ed Frank and Mike Sheridan and SUN Micro Systems Incorporation in 1991. It took 18 months to develop the first working version. This language was initially called "OAK", but was renamed "JAVA" in 1995. Before the initial implementation of OAK in 1992 and the public announcement of Java in 1995, many more contributed to the design and evolution of the language.

Java is a powerful but lean object oriented programming language. It has generated a lot of excitement because it make it possible to program for internet by creating applets, programs that can be embedded in web page. The context of an applet is limited only by one's imagination. For example, an applet can be animated with sound, an interactive game or a ticker tape with constantly updated stock prices. Applets can be just little decoration to liven up web page, or they can be serious applications like word processors or spreadsheet.

Java builds on the strength of C++. It has taken best features of C++. It has added garbage collection, multi threading and security capabilities. The result is that Java is actually a platform and easy to use.

Java is actually a platform consisting of three components:

1. Java programming language
2. Java library of classes and interfaces
3. Java virtual machine

One of the biggest advantages java offers is that it is portable. An application written in java will run on all the major platforms. Any computer with a java-based browser can run the applications or applets written in java programming languages. A programmer no longer has to write one program to run on a UNIX machine, and so on. Developers write code once; Java code is compiled into byte codes rather than a machine language. These byte codes go to the Java virtual machine, which executes them directly into the language that is understood by the system.

3.2Packages, Interfaces and Application Programming Interface

Java allows to groups classes in a collection called packages. Packages are convenient way of organizing the classes and libraries. Packages can be nested. A number of classes having same kind of behavior can be grouped under a package.

Packages are imported into the required java programs using the implements keyword. Interfaces provide a mechanism that allows unrelated classes to implement the same set of methods. An interface is a collection of method prototypes and constant values that is free from dependency on a specific class.

Application programming interface (API) forms the heart of any java program. These API'S are defined in corresponding java packages and are imported to the program.

3.3 JSP

The Java Server Page (JSP) technology is an extensible technology that enables the generation of dynamic content for a Web client. A JSP page contains:

HTML and XML tags to format the Web document. Document designers can edit and work with the JSP page without affecting the generation of the content.

3.4 SERVLETS

Java servlets are small, platform independent server-side programs that programmatically extend the functionality of the web server. The Java Servlet API provides a simple framework for building applications on web servers. This API is described in the Java Servlet API Specification (currently version 2.1) by the Java Software Division of Sun Microsystems Inc.

3.5 TOMCAT

Apache Tomcat is the servlet container that is used in the official Reference Implementation for the Java Servlet and Java Server Pages technologies. The Java Servlet and Java Server Pages specifications are developed by Sun under the Java Community Process.

Apache Tomcat is developed in an open and participatory environment and released under the Apache Software License. Apache Tomcat is intended to be a collaboration of the best-of-breed developers from around the world.

Apache Tomcat powers numerous large-scale, mission-critical web applications across a diverse range of industries and organizations. Almost every Tomcat release is initially released as an Alpha release. After a week or so of testing a vote is held to gather views on the stability of the release. If a major issue is identified during testing, then the vote may not take place and the release will remain as an Alpha release.

3.6 JDBC

JDBC is the set of interfaces for connecting to SQL table. With JDBC, we can query and update the data stored in the SQL tables within our Java programs. By this way, any Java object can be saved into SQL tables. This Java API is essential for EJB, the core API in J2EE.

JDBC creates a programming-level interface for communicating with databases in a uniform manner similar in concept to Microsoft's Open Database Connectivity (ODBC) component, which has become the standard for personal computers and LANs.

3.7 Network programming

One of Java's greatest strength is painless networking. The Java network library designers have made it quite similar to reading and writing files, except that the "file" exists on a remote machine and the remote

machine can decide exactly what it wants to do about the information you're requesting or sending. The programming model you use is that of a file; in fact, you actually wrap the network connection (a "socket") with stream objects, so you end up using the same method calls as you do with all other streams.

3.8 Socket

A *socket* is a communication end point- an object through which a windows sockets application sends or receives packets across the network. A socket has a type and is associated with a running process, and it may have a name. Currently, sockets generally exchange data only with other sockets in the same "Communication Domain", which uses the internet protocol suite. Two types of socket are

1. Stream socket
2. It provides data flow without record boundaries-a stream of bytes. Streams are guaranteed to be delivered and to be correctly sequenced and unduplicated.

The Oracle Server

The Oracle Server is an object-relational database management system that provides an open, comprehensive, and integrated approach to information management. An Oracle Server consists of an Oracle database and an Oracle Server instance. The following sections describe the relationship between the database and the instance.

Structured Query Language (SQL)

SQL (pronounced SEQUEL) is the programming language that defines and manipulates the database. SQL databases are relational databases; this means simply that data is stored in a set of simple relations. A database can have one or more tables. And each table has columns and rows. We can define and manipulate data in a table with SQL commands. We can use data definition language (DDL) commands to set up the data. DDL commands include commands to creating and altering databases and tables.

Database Structure

An Oracle database has both a physical and a logical structure. Because the physical and logical server structures are separate, the physical storage of data can be managed without affecting the access to logical storage structures.

Physical Database Structure

An Oracle database's physical structure is determined by the operating system files that constitute the database. Each Oracle database is made of three types of files: one or more data files, two or more redo log files, and one or more control files. The files of an Oracle database provide the actual physical storage for database information.

IV. SYSTEM DESIGN

Systems design is the process or art of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. One could see it as the application of system theory to product development. There is some overlap and synergy with the disciplines of system analysis, system architecture and system engineering.

4.1 OVERALL SYSTEM DIAGRAM

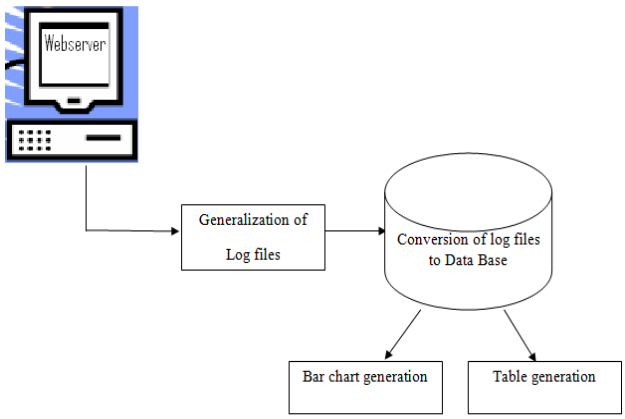

Figure 4.1: Overall System Diagram (Part I)

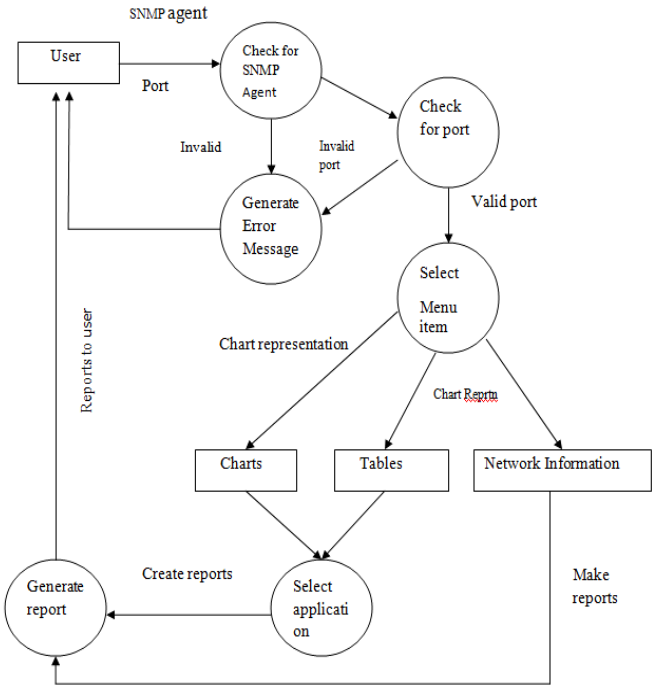

Figure 4.2: Overall System Diagram (Part II)

4.2 USE CASE DIAGRAM

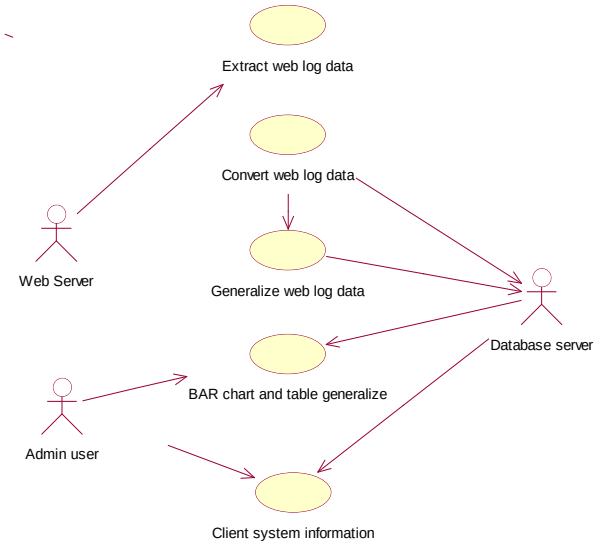

Figure 4.3: Use Case Diagram

4.3 ACTIVITY DIAGRAM

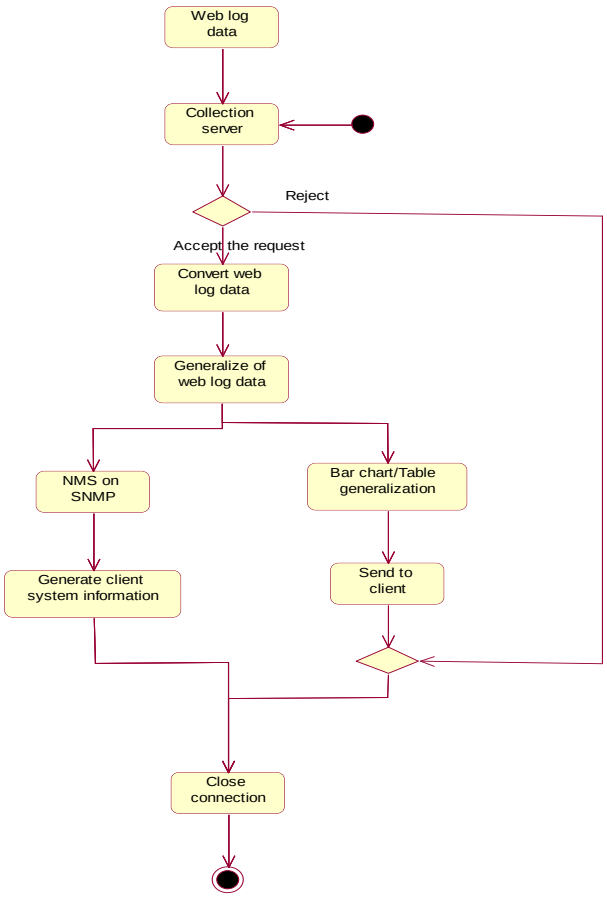

Figure 4.3: Activity Diagram

V. IMPLEMENTATION OF WUM

5.1 DETAILED PROCESS OF WUM

Step 1: Data preprocessing

Data preprocessing has a fundamental role in Web Usage Mining applications. It has different tasks:

(a) **Data Cleaning** -This step consists of removing all the data tracked in web logs that are useless for mining purposes.

(b) **Session Identification and Reconstruction-** This step consists of (i) identifying the different users' sessions from the usually very poor information available in log files and (ii) reconstructing the users' navigation path within the identified sessions.

(c) **Content and Structure Retrieving -** *Web content refers to the discovery of useful information from web contents including text, image, audio and video etc., structure retrieving gives the analysis of the out links of a webpage and it has been used for search engine result ranking.*

(d) **Data Formatting** - Once the previous phases have been successfully completed, data are properly formatted before applying mining techniques. So stored data extracted from web logs into a relational database.

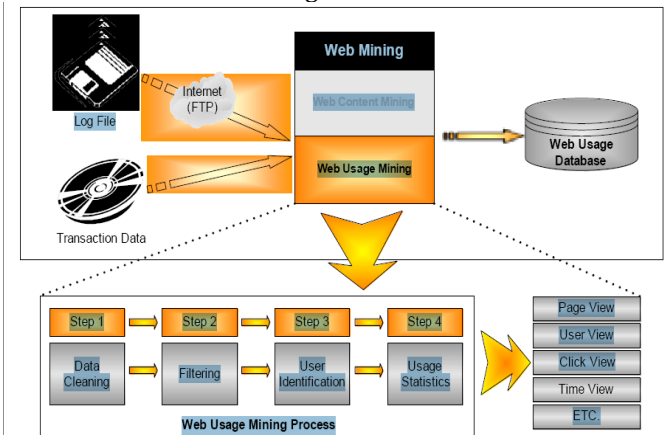

Figure 5.1: Phases of WUM

Step 2: Mining Algorithms

Process of mining algorithm or pattern discovery:

(a) **Statistical Analysis:** Statistical techniques are the most common method to extract knowledge about visitors to a Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. Many Web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site.

(b)**Clustering***:* Clustering is a technique which groups together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to discovered. (i.e.) usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing

patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users.

(c)**Classification***:* Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category.

(d)**Association Rules:** Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold.

(e)**Sequential Patterns***:* The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups.

(f)**Dependency Modeling:** Dependency modeling is another useful pattern discovery task in Web Mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain.

Step 3: Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process as described in Figure 3. The motivation behind pattern analysis is to filter out uninteresting rules or Patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL.

5.2 MODULES DESCRIPTION

5.2.1. Extracting web log files.

Extracting the log files from different web servers with various formats.

5.2.2. Converting web log files.

Converting information from text files (it is a file which is created by the log analyzer) and storing those webs based available in the file to database.

5.2.3. Generalization web log data

Posting of all data to the appropriate tuples.

5.2.4. Bar chart generation

Based on the information available in the database, it's going to build the required chart. (Example. Daily, Weekly, Web Pages viewed etc.)

5.2.5. Table generation:

Based on the information available in the database, its going to build the required information on the table on the database. Example. Daily, Weekly, Web Pages viewed etc.,

5.2.6 User Login

The user must provide his username and password at the time of login. The username is verified against the password that is stored in the database. Once the login is incorrect, an error message is provided. Otherwise the user can successfully logon to the monitoring system.

5.2.7 User Registration

Once the registration and login is over, the client should be selected where the user has to monitor. Once the user has made a successful registration or login, the server will take him to the next screen where all the clients connected to the server will be displayed.

VI. EVALUATION RESULT:

System implementation is a stage in the project where the theoretical design is turned into working system. It is the process of converting a new system into operation. The implementation phase of software development is concerned with translating design specification into source code. The most crucial stages is giving the users confidence that the new system will work effectively and efficiently.

Step 1: Extract the Web log File from the Web Server
Step 2: Convert the Web Log Files
Step 3: Generalize the Web Log Files using Algorithms
Step 4: Analyze the Web Site Performance using Bar Chart/ Table Generation
Step 5: Using admin login, click the menu
Step 6: The user request for data retrieval from Network is received by the collection server

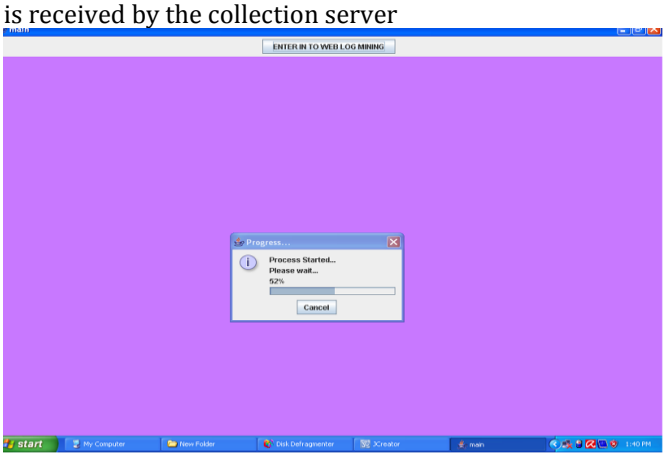

Figure 6.1Home Page

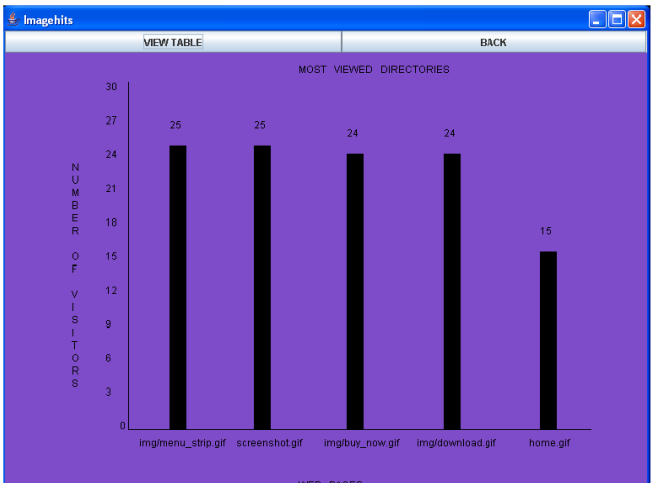

Figure 6.2 View table

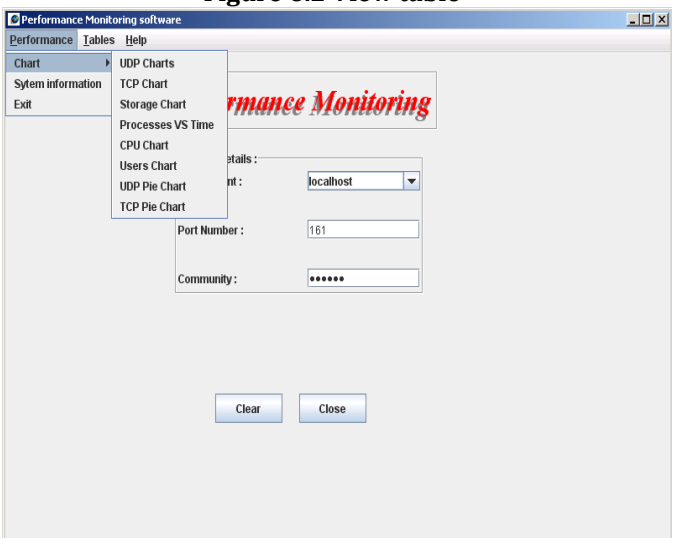

Figure 6.3 Performance Monitor

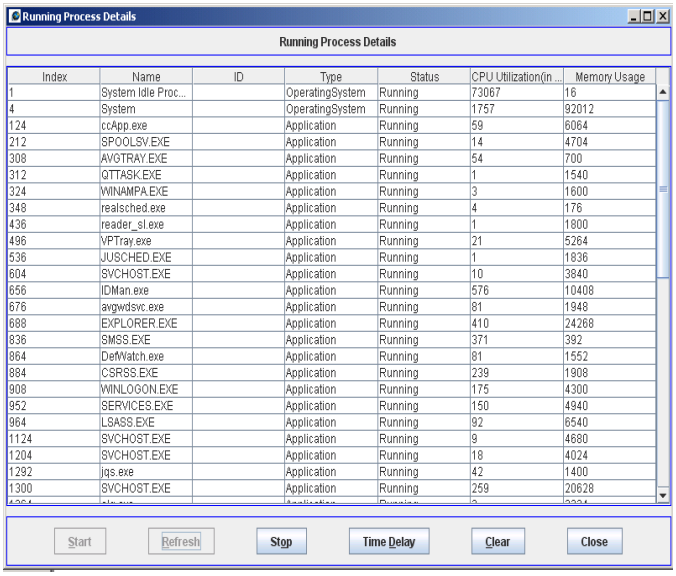

Figure 6.4 Running process detail

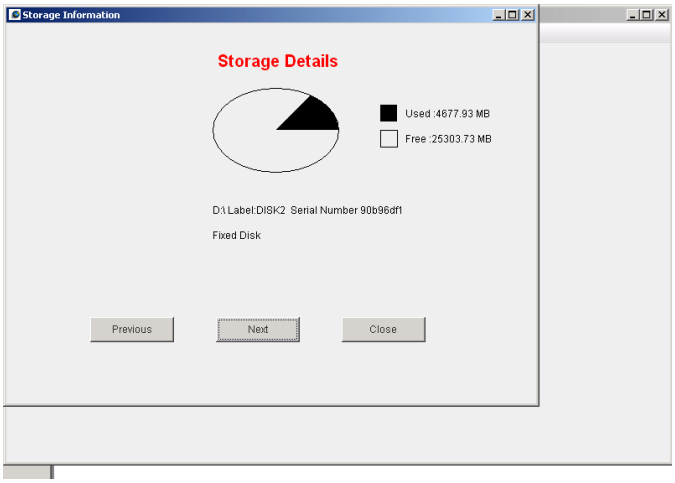

Figure6.5 Storage Details

Maintenance covers a wide range of activities, including correcting coding, design errors, updating documentation and test data, and upgrading user support. Many activities classified as maintenance or actually enhancement. Maintenance means restoring something to its original conditions. Unlike hardware, however software does not wear out, it is corrected. In contract, enhancement means adding, modifying or redeveloping the code to support changes in the specification. It is necessary to keep up with changing user needs and the operational environment.

CONCLUSION

Web Usage Mining is an active field for research and Web Usage Mining applications are being used in some famous Websites. This project presents an implementation of the Web Usage Mining. Web Server log files are mined in order to analyze the Web Usage pattern. The methodology employs *Data Preprocessing, Mining Algorithms and Pattern Analysis.* Data Processing phase for the Web Usage Mining is a challenging task. By applying mining algorithms to the Web log file, the relationship between the accessed pages can be mined. The Web usage patterns and user behavior are analyzed by using the mining algorithms. The results from this project can be used by Web administrator and Web masters in order to improve Web services and performance through the improvement of Web sites, including their contents, structure, presentation and delivery. The current remote monitoring tools are based on the SNMP protocol. Most of the commercial network components have embedded SNMP agents. Because of the universality of the Internet with TCP/IP protocol, the transport of management information for SNMP management, which is TCP/IP based is resolved automatically. In addition, most of the popular host operating systems come with the TCP/IP suite and thus are amenable to SNMP management.

FUTURE WORK

As a future enhancement of this project, web pages can be pre-fetched depending on the usage patterns. Pre-fetching can improve the web performance at a great level. Further, the method for analyzing sparse data can be used in the study of Web log access, use of different similarity Association Rules and conclude about the most suitable alternatives for knowledge extraction from Web log data. Finally the project can be extended to access and process the external web servers with appropriate access rights. SNMP accommodates the management of devices that do not implement the SNMP software by means of proxies. A proxy is an SNMP agent that maintains information on behalf of one or more non-SNMP devices.

REFERENCES:

[1] Bianka M. M. T. Gonçalves1, Isabel Cristina Italiano2, and João Eduardo Ferreira1,Data updating between the operational and analytical databases through dw-log algorithm, Proceedings of the 9th International Database Engineering & Application Symposium (IDEAS'05), 1098-8068/05, IEEE, 2005.

[2] Bong-Joon Choi, IL Kim, Kyoo-Seok Park," A Study on Web-Usage Mining Control System of using Page Scroll".

[3]Cooley: Jaideep, Srivastava t, Robert, Mukund Deshpande, Pang-Ning Tan,"Web Usage Mining: Discovery and Applications of Usage Patterns from Web to 23.Data", SIGKDD Explorations. Jan 2000. Volume 1.

[4]R.Cooley, B.Mobasher and J.Srivastava, Data Preparation for Mining WWW Browsing Patterns, Journal of Knowledge and Information systems, 1

[5] Craig P. Oosthuizen, Janet Wesson & Charmain Cilliers,"Processing Web Logs in order to Mine Web Usage Patterns".

[6]Federico Michele Facca and Pier Luca Lanzi," Recent developments in Web Usage Mining Research".

[7] Y.Fu,M.Creado and M.Shih,Adaptive Web site by Web Usage Mining, International Conference on Internet Computing, as Vegas, Pg.no 28-34, June, 2001.

[8]Hisayoshi Kato, Hironori Hiraishi and Fumio Mizoguchi,"Log summarizing for web access data using data mining techniques", IEEE transaction 2001

[9]Jan Kerkhofs Prof. Dr. Koen Vanhoof Danny Pannemans, A Case Study of Web Usage Mining on Proxy Servers, Published Limburg University Centre, July 30, 2001.