A

Project-I Report

on

ADAPTIVE TECHNIQUES FOR LOG FILE MINING TO ANALYSE USER'S SESSION PATTERN

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Abhishek P. Patil
Shweta S. Wani
Neha B. Patil
Shraddha R. Sonawane
Pooja A. Patil

Under the Guidance of

Mr. Naresh D. Kale

DEPARTMENT OF COMPUTER ENGINEERING
SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)

2016 - 2017

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001 (MS)

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the PROJECT-I entitled *Adaptive Techniques for Log File Mining to Analyse User's Session Pattern*, submitted by

Abhishek P. Patil
Shweta S. Wani
Neha B. Patil
Shraddha R. Sonawane
Pooja A. Patil

in partial fulfillment of the Degree of *Bachelor of Engineering* in *Computer Engineering* has been satisfactorily carried out under my guidance as per the requirement of North Maharashtra University, Jalgaon.

Date: October 8, 2016
Place: Jalgaon

Mr. Naresh D. Kale
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

We thank God almighty for blessing us throughout the work.We would like to thank all those who have contributed to the completion of the project-I and helped us with valuable suggestions for improvement.

We would like to thank our guide, Mr. Naresh D. Kale for all help and support.We are extremely grateful to Prof. Dr. G. K.Patnaik, HOD of Computer Engineering, for providing us with best facilities and atmosphere for the creative work guidance and encouragement.

We are greatly thankful to our principal Prof. Dr. K. S. Wani for providing us facilities. We would also want to thank Mr. Sandip Patil and Mr. Jitendra Patil for there support and guidance during Latex workshop. We thank all staff members of our college and friends for extending their co-operation during our minor project. Above all we would like to thank our parents without whose blessings, we would not have been able to accomplish our goal.

Abhishek P. Patil
Shweta S. Wani
Neha B. Patil
Shraddha R. Sonawane
Pooja A. Patil

Contents

List of Figures

Abstract

The important part of any internet application is system log files. System log files is the purpose of directory in numerous facet of information mining for web users. Web users are not able to locate what they actually want within reasonable amount of time. There is a good style of logs to stock information regarding the search patterns of the users. There can be ample formats of convenience of logs, every of internet application will develop format of its own logs. Generally, IP, date and time of the request, result for the request (with code), dealing size, protocol, request description, browser and software package utilized by the user area unit a number of the vital attributes of each request that get into the record of the log file. The users activity search pattern is examined by the question log files. The users browsing behavior is analyzed using various methodology. Completely different from existing ways, gift system avoids users feedback to the browser and doesnt collect the data which can manufacture privacy problems, e.g. users browsing history, bookmarks, and so on. As a result of the data recorded by server in access logs are utilized to judge the page interest, a referrer-based knowledge pre-processing methodology is administrated to enhance the dependability of the access knowledge and extract the required info for interest estimation.

Chapter 1

Introduction

Web Mining is technique in data mining to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. In Web Mining, data can be collected at the server side, client side, proxy servers, or obtained from an organization's database(which contains business data or consolidated Web data). There are many kinds of data that can be used in Web Mining. According to data analysis objective, web mining can be divided into three different types, which are web usage mining, web content mining and web structure mining, which is shown in the following Figure 1.1.

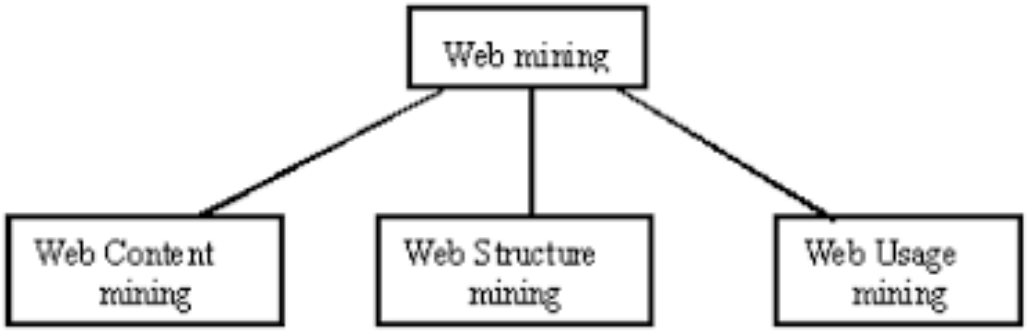

Figure 1.1: Web Mining Classification

Web content mining describes the automatic explore of information resources available online, and involves mining of web data content. Web structure mining is the process of analyzing hyperlink and tree-like structure of a web site using graph theory. Web usage mining is the process of extracting effective information from web server logs. Users show different interests when looking for internet. Some users might be looking at only documentary data, whereas some others might be engaged in multimedia data. Web usage mining (WUM) involves the automatic detection of user access patterns from one or more web servers. Organizations rely on internet for their business work which often generate and collect bulk

size data in their daily practices. Most of this information is generated automatically by web servers and collected in server access logs. The companies can establish better customer manager relationship by giving them exactly what they require. Companies can understand the requirements and serve them accordingly. They can also increase profitability and productivity based on the profiles generated.[1]

Each single access of the web page through the web queries given in the search engine enriches the query log files about the access information of every individual user. The automatic storage of the access information in the query log files of the search engines help to cull out relevant and appropriate information from the web. In every search through the search engine, the user submits a query that leads to the consolidation of the navigational information in the query log file. By this log, the queries can be personalized, based on the users behavioral search pattern.

This chapter includes the introductory part about web mining and web log files used in analysing user's session patter. Section 1.1 includes background, section 1.2 includes motivation, section 1.3 includes problem definition, section 1.4 includes scope and section 1.5 includes objectives.

1.1 Background

One of the key steps in Knowledge Discovery in Databases is to create a suitable target data set for the data mining tasks. In Web Mining, data can be collected at the serverside, clientside, proxy servers, or obtained from an organization's database (which contains business data or consolidated Web data). Each type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, and its method of implementation.[2] There are many kinds of data that can be used in Web Mining. This includes :

- Content

  The real data in the Web pages, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited.

- Structure

  Data which describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure, where the (html) tag becomes the root of the

tree. The principal kind of inter-page structure information is hyper-links connecting one page to another.

- Usage

  Usage: Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of access

- User Profile

  Profile: Data. that provides demographic information about users of the Web site. This includes registration data and customer profile information

Among the various kind of data available, a practicular data can be used to extract information. Certain techniques can be applied on given data set to get the required information. Web Log mining is the outcome of web usage mining which contains information of web access of different users. Several approaches have been proposed for efficient application of the Web Mining Algorithm for web log analysis.

1.2 Motivation

In this world of Information Technology, accessing information is the most frequent task. Every day we have to go through several kind of information that we need and what we do? Just browse the web and the desired information is with us on a single click. Today, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customer 24X7. On the other side visitors are also availing those facilities.In the last fifteen years, the growth in number of web sites and visitors to those web sites has increased exponentially. The number of users by June 30 2016 was 3,611,375,813[4] which is 49.2% of the worlds population. Due to this growth a huge quantity of web data has been generated. To mine the interesting data from this huge pool, data mining techniques can be applied. But the web data is unstructured or semi structured. So we can not apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is used to discover interest patterns which can be applied to many real world problems like improving web sites, better understanding the visitors behavior, product recommendation etc.

1.3 Problem Definition

A log file is a recording of everything that goes in and out of a particular server. It is a concept much like the black box of an airplane that records everything going on with the plane in the event of a problem. The information is frequently recorded chronologically, and is located in the root directory, or occasionally in a secondary folder, depending on how it is set up with the server. The information is updated day by day in log file. The only person who has regular access to the log files of a server is the server administrator, and a log file is generally password protected, so that the server administrator has a record of everyone and everything that wants to look at the log files for a specific server. The log file cant access by others because it is password protected.

Log files can be mined to determine users search pattern. Web log files mining helps in effectively predicting the best related web page to search. While searching a word from search engine it may display some unnecessary links and unrelated data's to user, so to avoid this problem, the conceptual prediction model combines both the web usage and domain knowledge. Using log files better search results are delivered to web users. Web users are able to locate what they actually want within a reasonable amount of time. This serves as an efficient tool towards an intelligent web. The system provides a good model for accessing the information related to particular log from the log files. The discovered knowledge from log files can then be used for many practical web applications such as web recommendations, adaptive web sites, and personalized web search and surfing, by using this knowledge web searching will be more efficient to the users. The system can be used in different type of mining areas of data mining applications.

1.4 Scope

Web usage mining is one of the prominent research. One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining. Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses. Prefetching and caching policies can be made on the basis of frequently accessed pages to improve latency time. Common access behaviors of the users can be used to improve the actual design of web pages and for making other modifications to a Web site. Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

1.5 Objective

The objective of this work is to apply web mining techniques to the different log files fron different web applications. Here the focus is given on mining web access patterns using log files from user session. In general, a web log can be regarded as a sequence of pairs of user identifier and event. Log files are files that list the actions that have been occurred. These log files reside in the web server. In this process, web log files are divided into pieces per mining purpose. Pre-processing can be applied to the original web log files, so that pieces of web log can be obtained. Each piece of web log is a sequence of events from one user or session in timestamp ascending order. The pieces of web log as sequences of events are modelled, and the user session patterns are mined using log files.

1.6 Summary

In this chapter, the background of the system is described. This includes the concept of web usage mining along with the use of log files and its processing. Log files are imporrtant part of the system used for pattern analysis of users. System analysis is described in next chapter.

Chapter 2

System Analysis

Analysis is a collection of data that has been gone through research from different sources e.g. internet, journal, e-book, article and else. The analysis of this system includes study of all things requires to develop it through various sources usch as e-books, internet, journal,etc.The analysis process is very important because it can be used as a guide for a developer to implement the project. The processes included in this step are:

1. Identifying the issue and its problem from the literature survey to get the first concept how the system should be done.
2. Research review about the existing system that has a related to the project which we want to developed.
3. Identifying the existing system's advantages and its weaknesses.
4. Creating the solution about the reviews and conclude it into the project.

This chapter includes the analysis of the system. Section 2.1 includes literature survey, section 2.2 includes proposed system, section 2.3 describes feasibility study, section 2.4 describes risk analysis, section 2.5 describes project scheduling.

2.1 Literature Survey

WWW has become the source for most of the existing information. Search engines have become the gateway of retrieving the requisite information from the web in terms of web snippets with reference to the query logged in by the user. A query is an array of keywords employed to secure information from the web resources even though desirable results are not retrieved at all times. During certain times inappropriate and redundant results are retrieved from the web.

2.2 Proposed System

The commonly used technique to augment the users search experience is the utilization of the knowledge contained within past queries in the log files. Log files are files that list the actions that have been occurred. These log files reside in the web server. A log files, typically, contains information about users, issued queries, clicked results, etc. From this information, knowledge can be extracted to improve the quality in terms of effectiveness and efficiency of their system. The technique used for analysing the serach pattern of user is a part of web usage mining. In the web mining process there are three types of mining they are web content mining, Web structure mining, Web usage mining.

In the web usage mining process, the techniques of data mining are applied so as to discover the trends and the patterns in the browsing nature of the visitors of the website. There is extraction of the navigation patterns as the browsing patterns could be traced and the structure of the website can be designed accordingly.[5] When it is talked about the browsing nature of the user it deals with frequent access of the web site or the duration of using the web site. This information can be extracted from the log file. Only these log files record the session information about the web pages.

Architecture for Processing the Log Files

Here, the log entries are stored into log file and the unnecessary data in the file is removed. User actions and their sessions are identified from the cleaned log file. If the format of the log files is different then, the retrieval of the user sequence and generation of frequent pattern of URLs is a difficult process.

The format of the first log file used in the data set is
<Client IP,Date and Time,URL visited,Status code,Browser >

The format of the second log file is
<Client IP,Http status code,URL visited,Date and Time, Browser >

The format of the third log file is
<Date and time,Client IP,Pages/directories,URL,Method, Username,Http status code,Referrer, Browser Mime type, Filter name,Filter reason,Profile,Interface IP,Interface port, Events,Page views,Bytes transferred,Elapsed time >

The discrepancies exhibited in the format of the log files set a predicament to access the user sequence. So it becomes mandatory to generate the common log file in which the log

entries are organized under the attributes
<Client IP,Date and Time,URL visited,Status code, Browser,OS >

Fig 2.1 shows mining flow of user session pattern

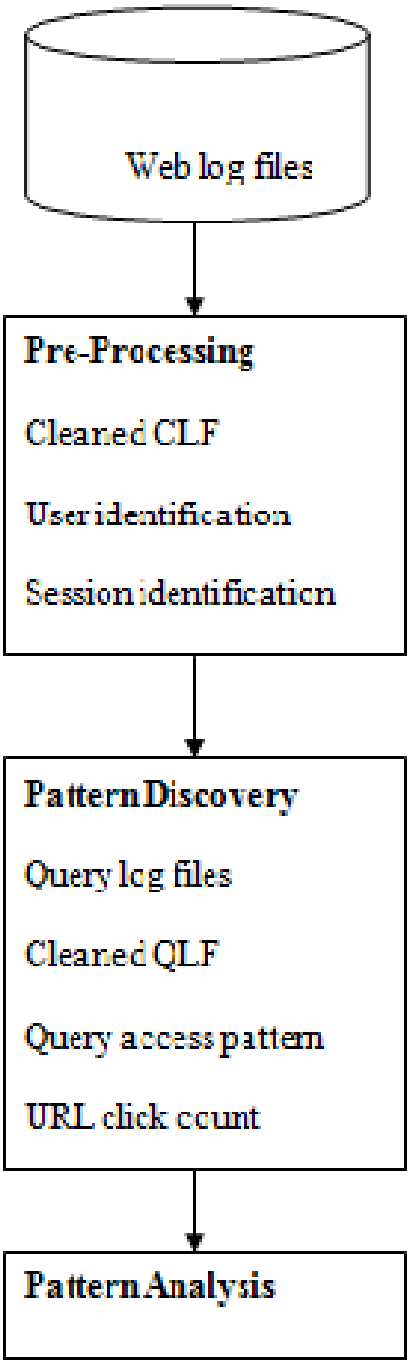

Figure 2.1: Mining Flow of Users Session Pattern

2.3 Feasibility Study

It is essential to evaluate the various aspects before we develop a system. Evaluation should always justify the cost and benefit ratio. If it is found that benefits are less as compared to cost of the project then it better to avoid going further in that project.

2.3.1 Technical Feasibility

For this project the technical feasibility was studied in two aspects. One is hard- ware feasibility and other is software feasibility. The system is very easy to upgrade and maintenance is also very easy. System is developed in Java. The technical issues rising during the feasibility stage of investigation include the fol- lowing points:

1. Does the necessary technology exist to do what is suggestion?
2. Does the equipment require for the proposed system it also available?
3. Can the system be expanded if developed?
4. Are technical guarantees of accuracy, reliability, ease of access and data security?
5. Do the proposed equipments have the technical capacity to hold the data required using the new system?

2.3.2 Economical Feasibility

A system that can be developed technically and that can be used if installed must still be a good investment for the organization. Financial and economic question are raised by analysis during the preliminary investigation for the purpose of estimating the following:

1. The cost to conduct the full system investigation.
2. The cost if hardware and software for the class of application being considered.
3. The benefits in the form of reduced cost that is fewer costly errors.
4. The cost if nothing changes.

2.3.3 Operational Feasibility

The test of feasibility asks if the system will work when it is developed and installed. Are there major barriers to implementation? Here are the questions that will help to test the operational feasibility of the project. Are current business methods acceptable to the user? If they are not, users may welcome the change that will bring bout more operational and useful system. Project follows the operational views of the feasibility.

2.4 Risk Analysis

There are various types of risk are present so different categories of risk are as follows:

- Technical risks:

  Technical risk is simply the risk associated directly with the knowledge base being employed and it's technical aspects including such things as understanding, reproducibility and the like. Exposure to loss arising from activities such as design and engineering, manufacturing, technological processes and test procedures.

- Business risks:

  The term business risk refers to the possibility of inadequate profit or even loss due to uncertainties e.g., changes in tastes, preferences of consumers, strikes, increased competition, change in government policy, obsolesce etc. Every business organization contains various risk elements while doing the business. Business risks implies uncertainty in profits or danger of loss and the events that could pose a risk due to some unforeseen events in future, which causes business to fail.

- Project risks:

  A project risk is an uncertain event that, if it occurs, has a positive or negative effect on the prospects of achieving project objectives. Project risk can be defined as an unforeseen event or activity that can impact the project's progress, result or outcome in a positive or negative way. Effective risk management strategies allow to identify your project's strengths, weaknesses, opportunities and threats. By planning for unexpected events, user can be ready to respond if they arise. To ensure project's success, define how user will handle potential risks so user can identify, mitigate or avoid problems.

2.5 Project Scheduling

The project is schedule by considering the various requirements fullfilment of the customer.Following is the table for project scheduling.

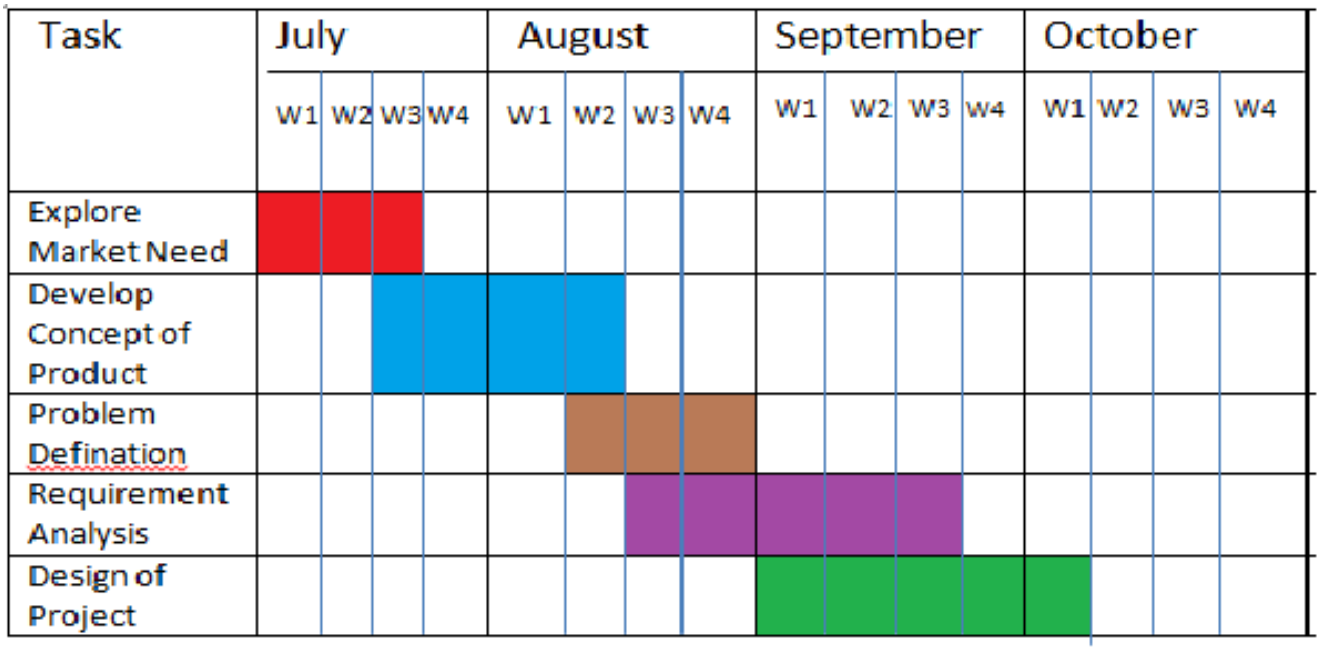

Figure 2.2: Project Scheduling

2.6 Summary

This chapter includes system analysis, the proposed system is described. This includes the steps for processing the log files gathered from the web server. Log files from different web server are merged and cleaned for pattern analysis of user's session. System requirement specification is described in next chapter.

Chapter 3

System Requirement Specification

Understanding the requirements of a problem is one of the most difficult tasks faced by the software engineer. Before begining any technical work, it is good idea to apply set of requirements engineering tasks. These tasks lead to an understanding of what business impact will be on software, what the customer wants, and how end user will interact with the software. Designing and building an elegant computer program that solves the wrong problem serves no one's needs. Thats why it is important to understand what the customer wants before begining to design and build a computer-based system.

This chapter includes the system requirement specification. The section 3.1 describes the hardware requirement, section 3.2 describes software requirement, section 3.3 describes other requirement and constraints and section 3.4 describes performance requirement.

3.1 Hardware Requirement

HArdware requirements are given below:

- Processor : Intel Core i3
- Hard Disk : 40 GB and above
- RAM : 1 GB
- Display : CGA,VGA,SVGA
- Input devices : Mouse,Keyboard

3.2 Software Requirement

Software requirements specifications are typically developed during the last stages of "Requirements Development", which is the initial product development phase in which information is gathered about what requirements are needed and not. This information gathering

stage can include onsite visits, questionnaires, surveys, interviews, and perhaps a return-on-investment (ROI) analysis or needs analysis of the customer or client's current business environment. The actual specificcation, then, is written after the requirements have been gathered and analyzed.

Software requirements for running the application:

- Front end : JDK 1.7
- Web Server : Tomcat Web Server 6.0
- Opearting System : Windows 7, 8 and Linux

3.3 Other Requirement and Constraints

The following are attributes of software:

1. The software should be efficient with minimum overhead on the system
2. Each time the required output must be generated with usefull information.
3. The output should be accurate and efficient

3.4 Performance Requirement

The machines must be atleast of third generation of computers with sufficient amount of physical memory. The software should be compatible with most of the operating systems. The software should consume less time and memory.

3.5 Summary

This chapter describe the software requirement specification along with the various requirements needed for system that includes hardware requirement, software requirement, preformance requirement and other requirements & constraints. Next chapter describes the system design.

Chapter 4

System Design

When any application is developed then the system design plays very important role for successful development. In system design the actual layout of the complete system is designed and by that design the development proceeds so that the time required for development is reduced and task becomes easy.

This chapter includes design of the system. Section 4.1 contain the system architecture, section 4.2 includes the data flow diagram and section 4.3 describes the various UML diagrams.

4.1 System Architecture

System architecture describes the various modules present in project which are interrelated to each other.

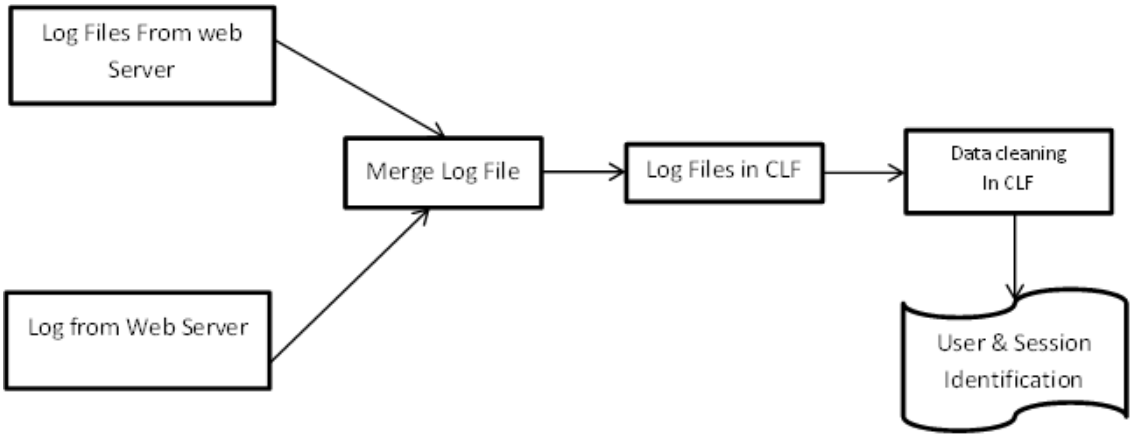

Figure 4.1: Process of Log Files CLF Mining

4.2 Data Flow Diagrams

The Data Flow Diagram (DFD) is the graphical representation of the processes and the flow of data among them. A data flow diagram illustrates the processes, data stores, external entities and the connecting data flows in a system. It is a common practice to draw a context-level data flow diagram first which shows the interaction between the system and outside entities. This context-level DFD is then "exploded" into a detailed DFD.

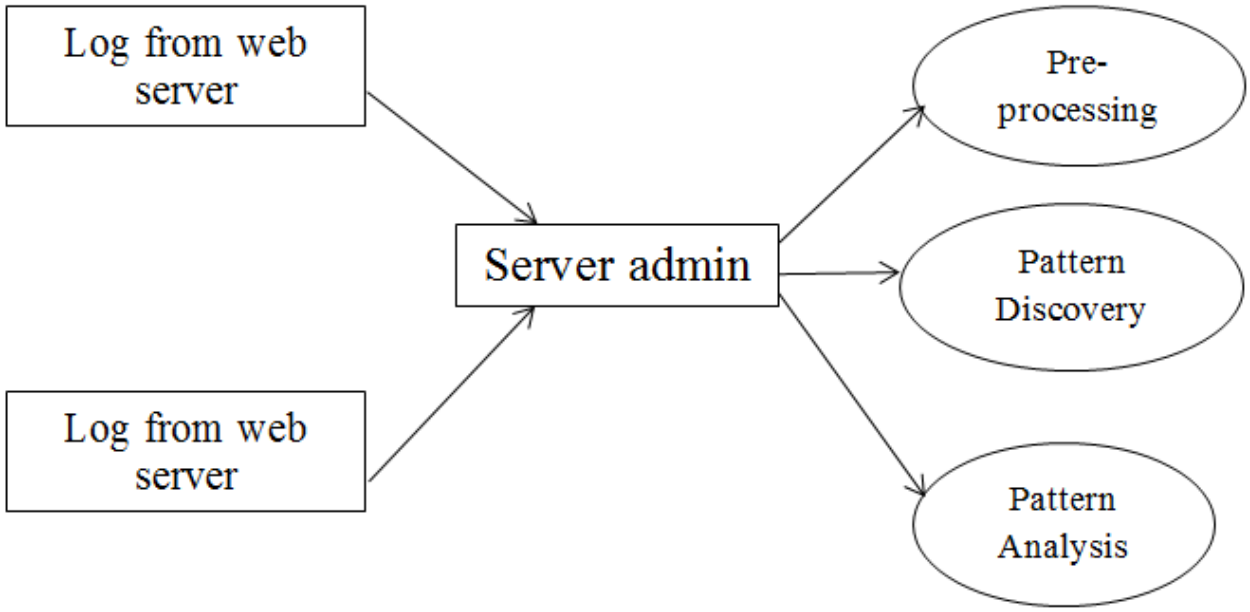

Figure 4.2: DFD (Level 0)

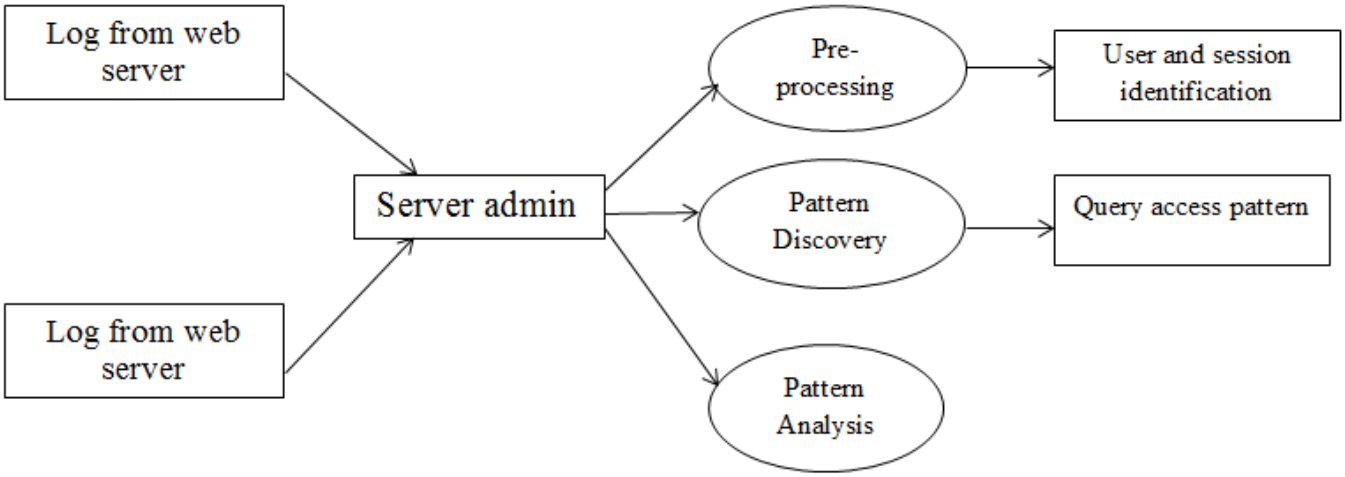

Figure 4.3: DFD (Level 1)

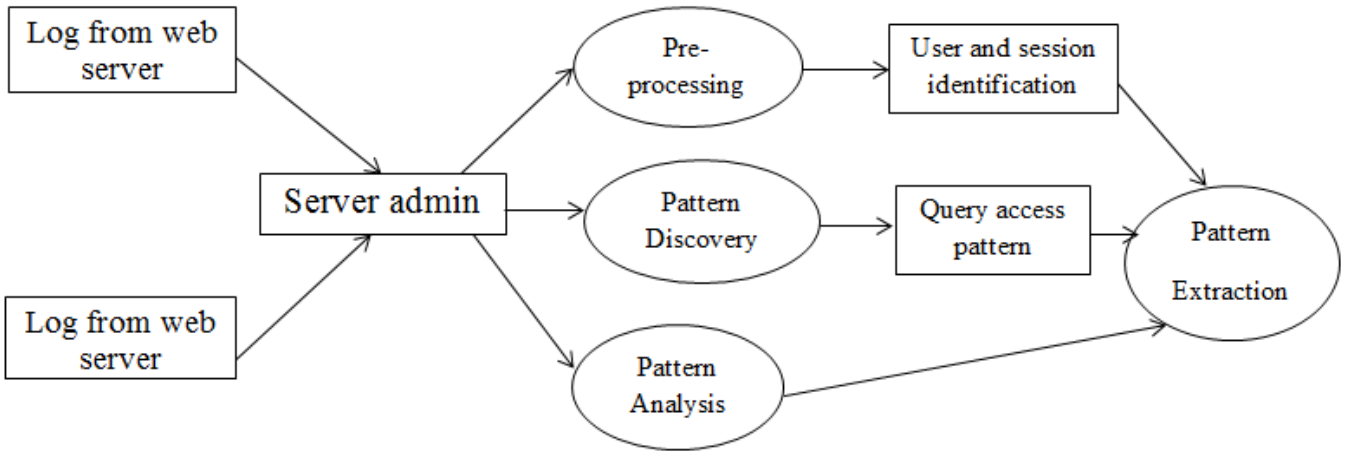

Figure 4.4: DFD (Level 2)

4.3 UML Diagrams

There are various types of UML diagrams that are present that are useful to completely understand the system from both the developer and the user side. These diagrams easily describe the actual function of the developed system to other developers also and the user can easily understand how the system works.

4.3.1 Use Case Diagram

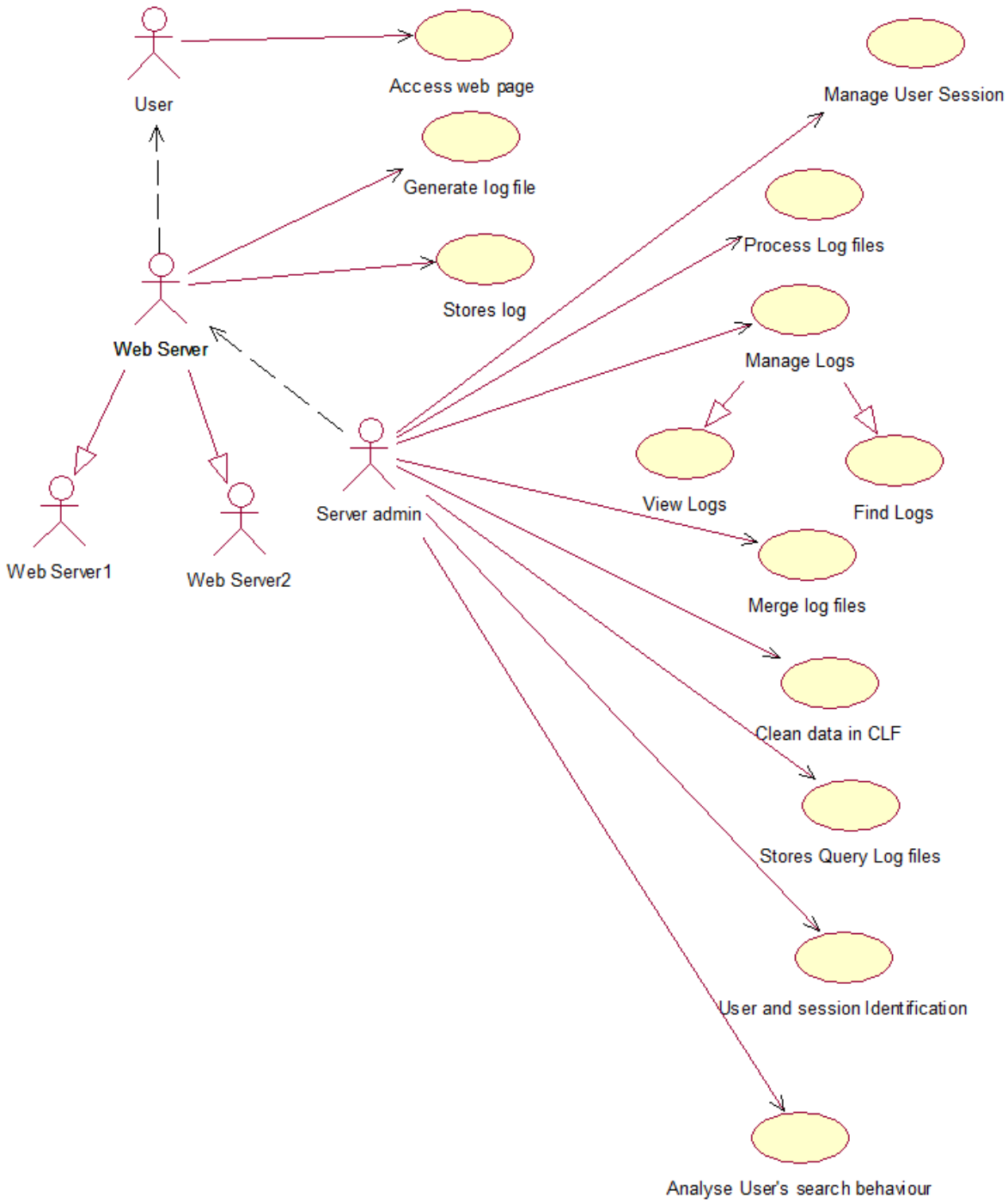

Figure 4.5: Use Case Diagram

Above is the use case diagram which shows the scenario generating the log files for analysing the user pattern. This diagram contains various actors such as user, sever admin and web

server. The various functionality of actors are accessing web pages, generating log files,storing logs, manage use sessions, process log files, manage log, merge log files, clean data in CLF, store query log files, analyse user's search behaviour.

4.3.2 Class Diagram

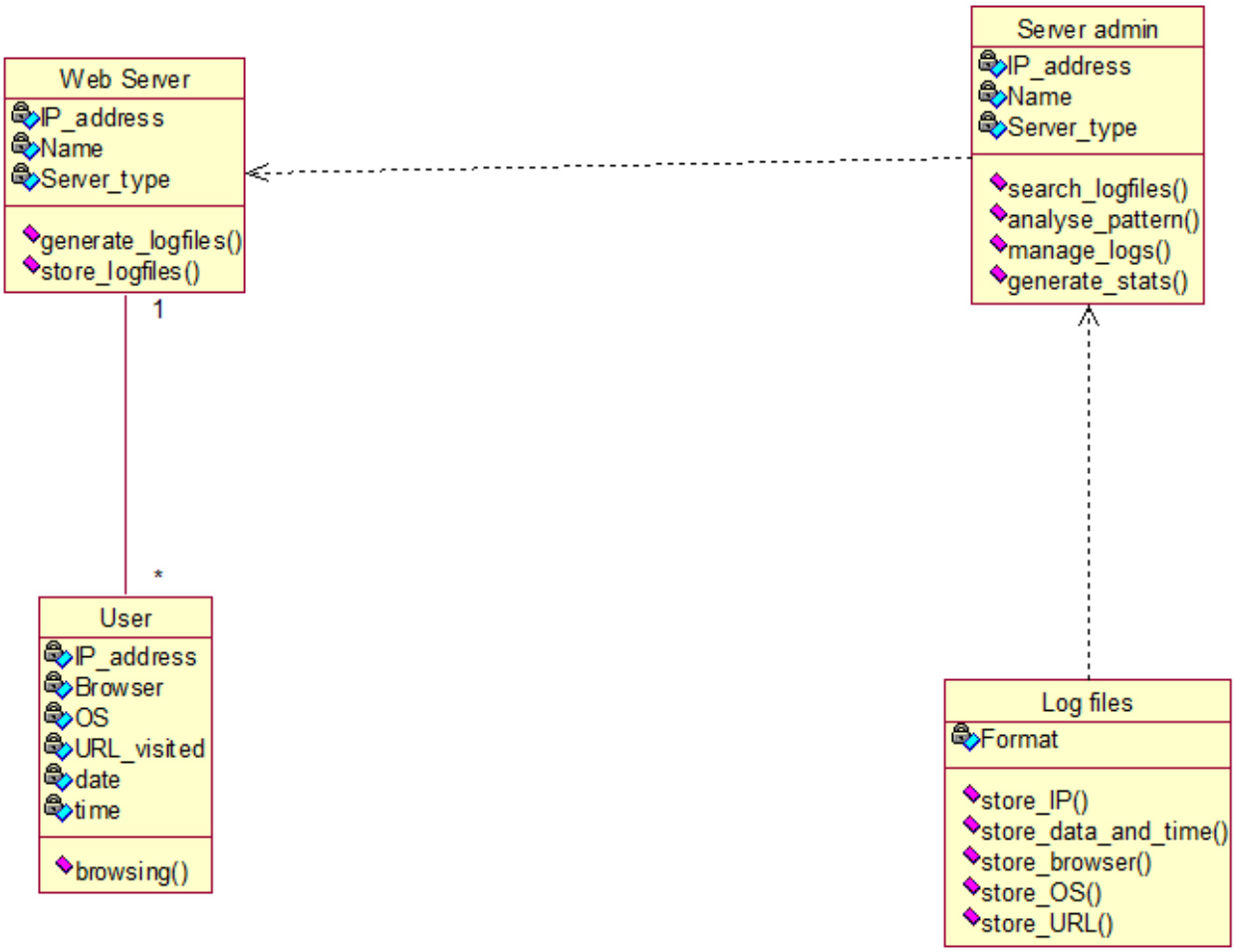

Figure 4.6: Class Diagram

Above is the class diagram showing various classes included in system. The classes Web server, User, Server admin and Log files.

4.3.3 Sequence Diagram

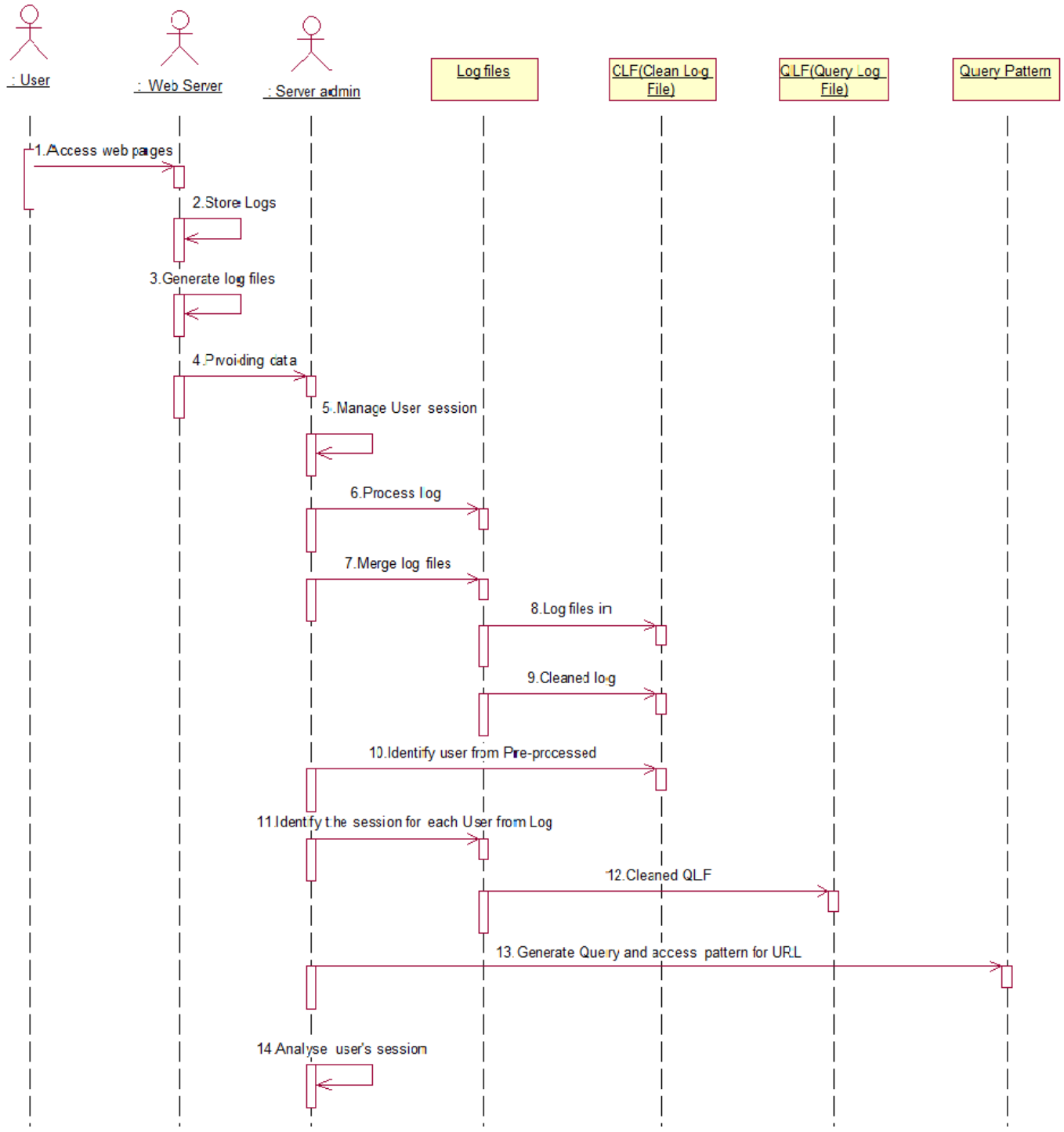

Figure 4.7: Sequence Diagram

Above is the sequence diagram showing the event scenario which depicts interaction between various objects arranged in time sequence. It clearly shows the flow of the system which includes the accessing the web page by users upto analyse user's session.

4.3.4 Collaboration Diagram

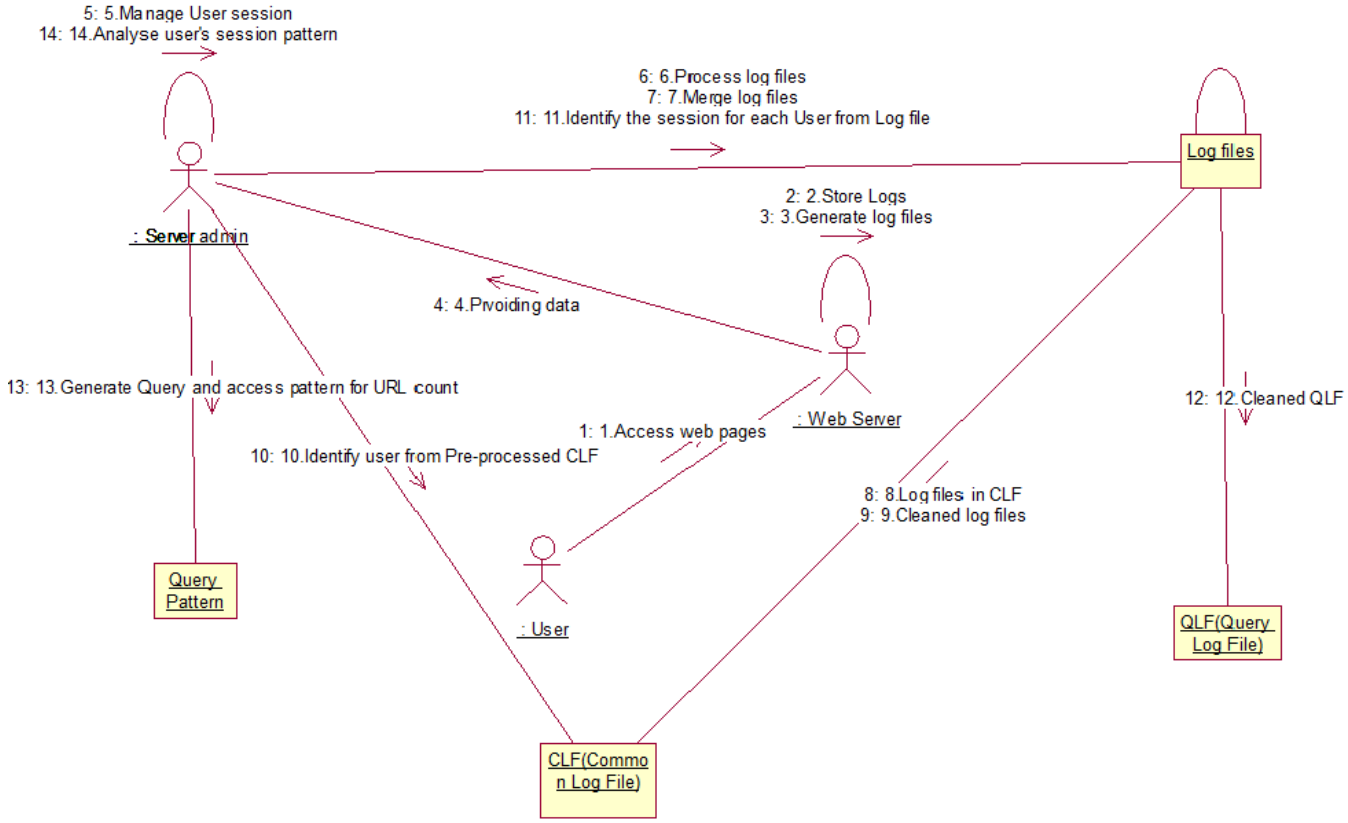

Figure 4.8: Collaboration Diagram

Above is the collaboration diagram which shows the relationships and interaction about various object. The objects includes log files, (QLF)Query log files, CLF(Commomn log file) and query pattern.

4.3.5 Component Diagram

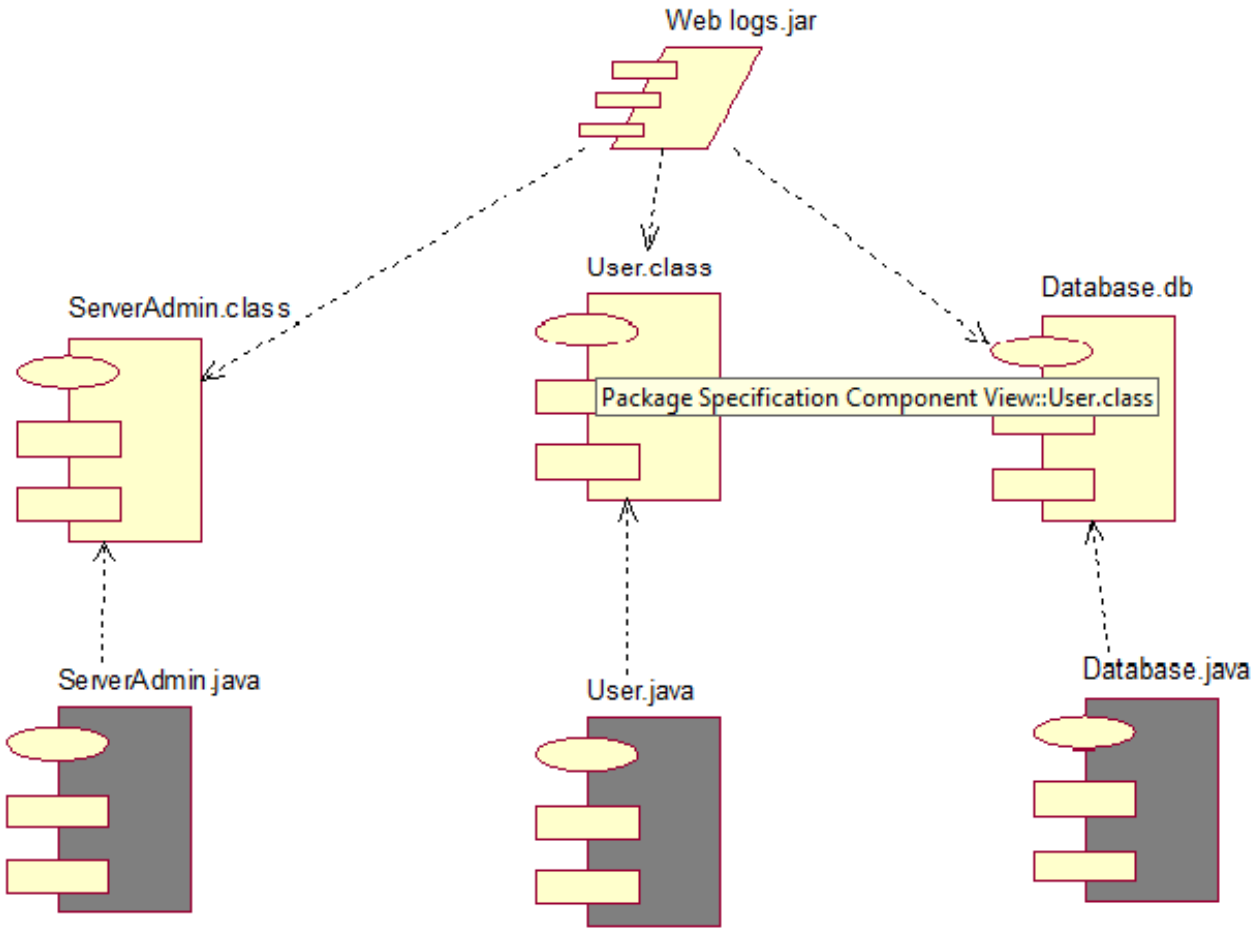

Figure 4.9: Component Diagram

Above is the component diagram which shows various components such as Web logs, Server-Admin, User and Database.

4.3.6 Deployment Diagram

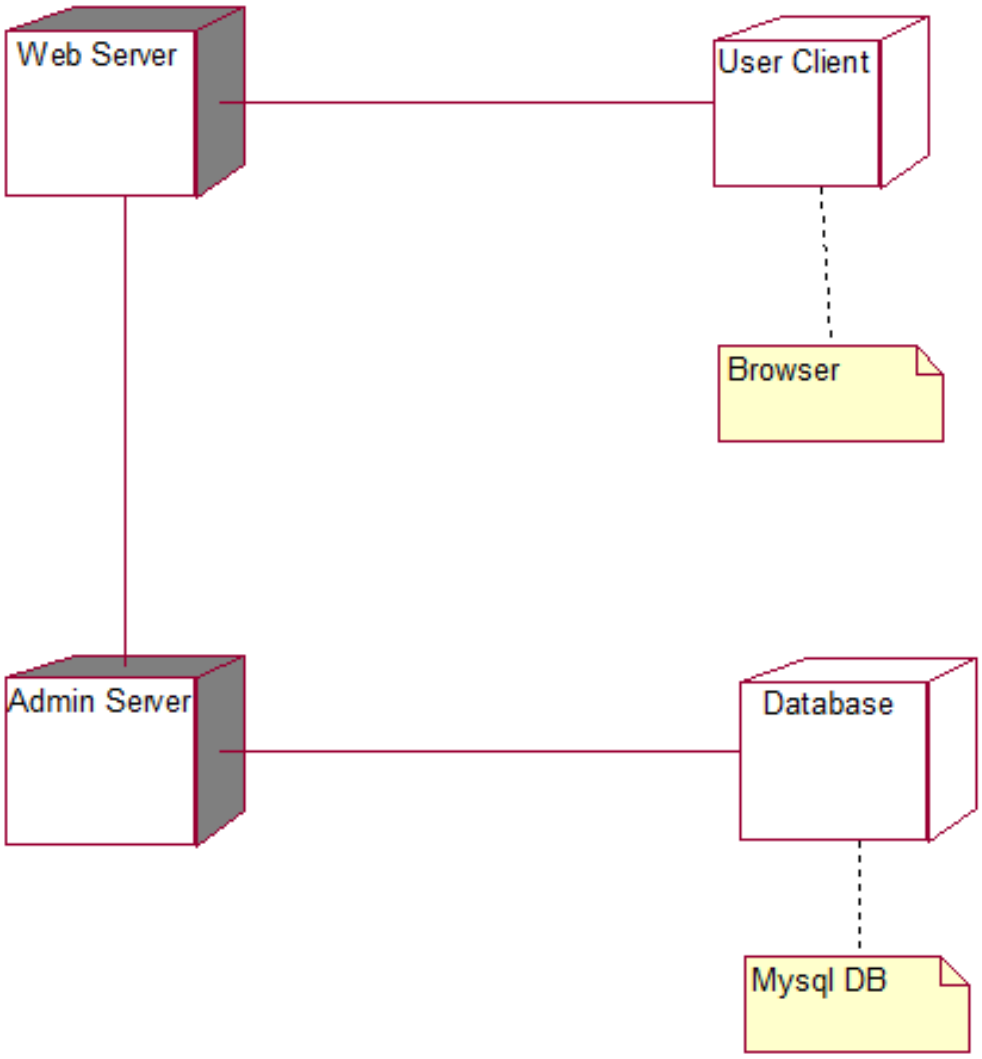

Figure 4.10: Deployment Diagram

Above is the deployment diagram which models the physical deployment of artifacts on nodes. This diagram includes various nodes Web server, Admin server, User Client and Database.

4.4 Summary

This chapter describes the designing of the system with the help of various UML diagrams and DFDs.

Chapter 5

Conclusion

As a result of lots of research and development, quality of web related services has improved significantly. But because of unprecedented growth of the Web and abundance of information, Web Users are not able to locate what they actually want within a reasonable amount time. This work has proposed one efficient tool towards an intelligent web. The proposed tool provides a good model for accessing the information related to particular log from the log files. The tool can be used in different type of mining areas of data mining applications. But due to the fast development in the technology and explosion in the number of users, Web Mining area still gives lots of research opportunities. Mining log files will provide the desired information. This information can be out to use by companies, organizations, businesses and individuals.

Bibliography

[1] Chandana S. Khatavkar(2015), A Hybrid Approach For Clustering Weblog(2015). International Journal Of Advanced Research In Computer Science And Software Engineering, Volume 5,Issue3.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In Proc. ACM KDD, 1994.

[3] R. Pamnani, P. Chawan Web Usage Mining: A Research Area in Web Mining

[4] http://www.internetworldstats.com

[5] Mehak(2013), Web Usage Mining: An Analysis, Journal Of Emerging Technologies In Web Intelligence, Vol. 5, No. 3