Web Usage Mining Framework for Mining Evolving User Profiles

PROJECT REPORT

Submitted by

V.D. SRIHARI

Register No: 0820108019

in partial fulfillment for the award of the degree

of

MASTER OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

KUMARAGURU COLLEGE OF TECHNOLOGY

(An Autonomous Institution Affiliated to Anna University, Coimbatore)

COIMBATORE – 641 006

MAY 2010

KUMARAGURU COLLEGE OF TECHNOLOGY

(An Autonomous Institution Affiliated to Anna University, Coimbatore)

COIMBATORE – 641 006

Department of Computer science and Engineering

PROJECT WORK

PHASE II

MAY 2010

This is to certify that the project entitled

Web Usage Mining Framework for Mining Evolving User Profiles

is the bonafide record of project work done by

V.D. SRIHARI

Register No: 0820108019

of M.E. (Computer Science and Engineering) during the year 2009-2010.

V.S. Akshaya

Project Guide

(Mrs. V.S. AKSHAYA)

Head of the Department

(Mrs. P. DEVAKI)

Submitted for the Project Viva-Voce examination held on 18/5/10

18/5/10

Internal Examiner

External Examiner

DECLARATION

I affirm that the project work titled "**Web Usage Mining Framework for Mining Evolving User Profiles**" being submitted in partial fulfillment for the award of M.E Computer Science and Engineering is the original work carried out by me. It has not formed the part of any other project work submitted for award of any degree or diploma, either in this or any other University.

V.D. SRIHARI

0820108019

I certify that the declaration made above by the candidate is true

Signature of the Guide,

V.S. Akshaya

Mrs. V.S. AKSHAYA

Senior lecturer

ACKNOWLEDGEMENT

I express my profound gratitude to our Chairman **Padmabhusan Arutselver Dr. N. Mahalingam B.Sc, F.I.E** for giving this great opportunity to pursue this course.

I would like to begin by thanking to **Dr.S.Ramachandran.**, **Ph.D**, *principal* for providing the necessary facilities to complete my thesis.

I take this opportunity to thank **Dr. S.Thangasamy**, **Ph.D.,** *R & D Dean*, for his precious suggestions.

I take this opportunity to thank **Mrs. P. Devaki M.E., (Ph.D),** *Head of the Department*, Computer Science and Engineering, for his precious suggestions

I register my hearty appreciation to **Mrs. V.S. Akshaya M.E., (Ph.D),** *Senior lecturer,* my thesis advisor. I thank for her support, encouragement and ideas. I thank her for the countless hours she has spent with me, discussing everything from research to academic choices.

I thank all project committee members for their comments and advice during the reviews. Special thanks to **Mrs.V.Vanitha M.E., (Ph.D)** *Assistant professor*, Department of Computer science and Engineering, for arranging the brain storming project review sessions.

I would like to convey my honest thanks to **Mr. T. Chandra Vel Kumar (B.E)**, *Lab Assistant*, all **Teaching** staff members and **Non Teaching** staffs of the department for their support. I would like to thank all my classmates who gave me a proper light moments and study breaks apart from extending some technical support whenever I needed them most.

I dedicate this project work to my **parents** and my **friends**, without their love this work wouldn't possible.

TABLE OF CONTENTS

Estd : 1984

KONGU ENGINEERING COLLEGE

(Autonomous)

PERUNDURAI ERODE 638 052 TAMILNADU INDIA

TÜV SÜD ISO 9001

SILVER JUBILEE YEAR

School of Computer Technology and Applications

Department of Computer Applications

KONGU Assuring the Best

CERTIFICATE

This is to certify that Ms/~~Mr~~/~~Dr~~. ........V.D.SRIHARI... of KUMARAGURU ...COLLEGE OF TECHNOLOGY, COIMBATORE............ has ~~participated~~ / presented the paper titled ..WEB USAGE MINING FRAMEWORK FOR MINING EVOLVING USER PROFILES............ ....in the National Research Conference on **"Challenges & Innovations in Information Technology"** (CIIT - 2010) held on March 25, 2010 at Kongu Engineering College, Perundurai.

Organizing Secretary	Co-Chairman	Chairman	Patron
Ms. D.Chitra	Dr. A.Tamilarasi	Dr. P.Thangaraj	Prof. S.Kuppuswami

VIVEKANANDHA

INSTITUTE OF ENGINEERING AND TECHNOLOGY FOR WOMEN

ELAYAMPALAYAM, TIRUCHENGODE.

Department of Computer Applications

NATIONAL CONFERENCE
ON
" EMERGING TECHNOLOGIES IN ADVANCED COMPUTING AND COMMUNICATION "

13 - March, 2010

CERTIFICATE

This is to Certify that Mr. / Ms. / Dr. V.D. SRIHARI of ME, Kumaraguru college of Technology has participated in the " National Conference on ETACC '10 ", organized by " VIVACIOUS " the professional association of Computer Applications on 13th March 2010 and presented a paper on Web usage Mining framework for Mining Evolving user profiles.

A. Ahamed
HOD & Organizing Secretary

Principal

Chairman & Secretary

ABSTRACT

Customer Relationship Management (CRM) can use data from within and outside an organization to allow an understanding of its customers on an individual basis or on a group basis such as by forming customer profiles. An improved understanding of the customer's habits, needs, and interests can allow the business to profit by, for instance, "cross selling" or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers' preferences and needs forms the basis for effective CRM. As businesses move online, the competition between businesses to keep the loyalty of their old customers and to lure new customers is even more important, since a competitor's Web site may be only one click away. In this paper, we present a complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing an ontology of the Web content. Even though the Web site under study is part of a non-profit organization that does not "sell" any products, it was crucial to understand "who" the users were, "what" they looked at, and "how their interests changed with time," all of which are important questions in Customer Relationship Management (CRM). Hence, we present an approach for discovering and tracking evolving user profiles. We also describe how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior.

ஆய்வுச்சுருக்கம்

வாடிக்கையாளர் உறவு மேலாண்மை ஒரு நிர்வாக அமைப்பில் உள்ள வாடிக்கையாளர்களை தனியாகவோ அல்லது குழுவாகவோ புரிந்து கொள்ள தரவுகளை உள்ளிருந்து மற்றும் வெளியிருந்தும் பயன்படுத்துகிறது. வாடிக்கையாளர்களின் பழக்கங்கள், தேவைகள் மற்றும் ஆர்வத்தை அறிந்து கொண்டால் தொழில் லாபகரமாக இருக்கும். எடுத்துக்காட்டாக வாடிக்கையாளர்களுக்கு தேவையான பொருளை விற்பது. வாடிக்கையாளர்களின் உறவு மேலாண்மைக்கு மிகவும் இன்றியமையானது அவர்களின் விருப்பங்களும், தேவைகளும். இப்பொழுது கணினியுடன் நேரடியாக தொழில்கள் நடைபெற்று வருகின்றன. தொழில் பாட்டி காரணமாக பழைய வாடிக்கையாளர்களை இழக்காமலும் புதிய வாடிக்கையாளர்களை ஈர்ப்பதும் மிகவும் இன்றியமையாதது எதனால் என்றால் போட்டியாளர்களின் இணையதளம் மிக அருகில் உள்ளது. இந்த ஆராய்ச்சியின் மூலமாக முழுமையான வரம்பிற்க்கு மற்றும் இணையதளம் பயன்பாட்டு அமைப்பை அறிகிறோம். இணையத்தள பயன்பாட்டு அமைப்பை அறிவதர்க்காக நாம் ஒரு இணையதளத்தின் வலைப்பகுதியை ஆராயப் போகிறோம். இந்த ஆராய்ச்சிக்காக உபயோகிப்பவரின் பக்கத்தோற்ற வடிவத்தை அறிய வேண்டும். இருப்பினும் நாம் இணையதளம் லாப நோக்கில்லாத நிர்வாக அமைப்பு என்றால் அது எந்த ஒரு பொருளையும் விற்க போவதில்லை அதனால் தனது வாடிக்கையாளர்கள் யார், எதை பார்க்க வருகிறார்கள், அவர்களின் ஆர்வம், காலத்திற்க்கு ஏற்ப எவ்வாறு மாறுகிறது என்ற கேள்வி வாடிக்கையாளர்களின் உறவு மேலாண்மைக்கு முக்கியமானது. இதற்க்காக நாம் இப்பொழுது வாடிக்கையாளர்களின் பக்கத்தோற்ற வடிவத்தை ஆராயப்போகிறோம்.

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

As businesses move online, the competition between businesses to keep the loyalty of their old customers and to lure new customers is even more important, since a competitor's Web site may be only one click away. The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles. These different modes of usage or the so-called mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user clickstreams stored in Web log files. These profiles can later be harnessed toward personalizing the Web site to the user or to support targeted marketing. Although there have been considerable advances in Web usage mining, there have been no detailed studies presenting a fully integrated approach to mine a real Web site with the challenging characteristics of today's Web sites, such as evolving profiles, dynamic content, and the availability of taxonomy or databases in addition to Web logs

1.1 MOTIVATION OF THE PROJECT

- An improved understanding of the customer's habits, needs, and interests can allow the business to profit by, for instance, "cross selling" or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers' preferences and needs forms the basis for effective CRM.

- As businesses move online, the competition between businesses to keep the loyalty of their old customers and to lure new customers is even more important, since a competitor's Web site may be only one click away
- The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles.
- Although there have been considerable advances in Web usage mining, there have been no detailed studies presenting a fully integrated approach to mine a real Web site with the challenging characteristics of today's Web sites, such as evolving profiles, dynamic content, and the availability of taxonomy or databases in addition to Web logs.

1.1 PROBLEM DESCRIPTION

Customer relationship management is more important for an organization even though the organization is a non profit organization. Organization should be able to know the customers who visit their website. The organization should understand the customer habits, needs, interest and preferences. It is also very crucial to understand how interest changes from time to time. Since the business moves online the competitors website is also near and only a click away. Inspite of Large amount of data is available through weblog files which are created during user click streams, the personalization of the websites or web portals to an individual or group of customers is more important. The proposed system addresses the above problem and provides a complete framework with a clustering algorithm to mine a real website and construct dynamic profiles and to track the current profiles against the existing profiles.

1.3 SCOPE

- Any Online Web portal providing multispecialty functionalities will require an User Profiling application like this projects.
- Any Customer Relationship Management (CRM) can use data from within and outside an organization to allow an understanding of its customers on an individual basis or on a group basis such as by forming customer profiles.
- Any Web site or a Web portal that provides access to news, events, resources, company information (such as companies or contractors supplying related products and services) can utilize the functionality of our projects.
- It helps to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time.

CHAPTER 2

2 LITERATURE SURVEY

With the explosive growth of information sources available on the World Wide Web. It has become increasingly necessary for users to utilize automated tools in Find the desired information resources and to track and analyze their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This describes the automatic search of information resources available online i.e. Web content mining and the discovery of user access patterns from Web servers i.e. Web usage mining. In this project, an overview of tools techniques and problems associated with both dimensions. Also presented a taxonomy of Web mining and place various aspects of Web mining in their proper context. There are several important issues unique to the Web paradigm that comes into play if sophisticated types of analyses are to be done on server side data collections. These include integrating various data sources such as server access logs referrer logs user registration or profile information resolving difficulties in the identification of users due to missing unique key attributes in collected data and the importance of identifying user sessions or transactions from usage data site topologies and models of user behavior. The main part of this paper is to discussion of issues and problems that characterize Web usage mining. Furthermore we survey some of the emerging tools and techniques and identify several future research directions.

2.1 Taxonomy of Web Mining

In this section, the taxonomy of Web mining along its two primary dimensions, namely Web content mining and Web usage mining. It also describes and categorizes some of the recent work and the related tools or techniques in each area.

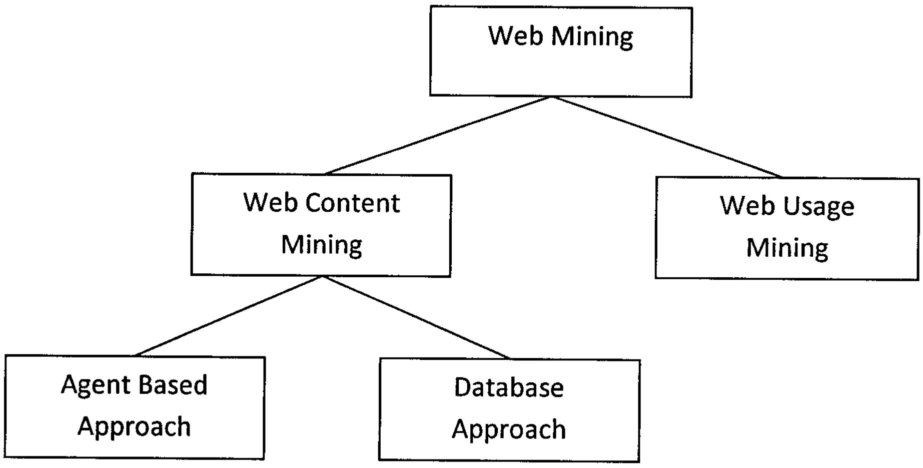


Figure 2.1: Taxonomy of web mining.

2.1.1 Web Content Mining

The heterogeneity and the lack of structure that permeates much of the ever expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and management of Web-based information difficult. Traditional search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB, MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. A recent study provides a comprehensive and statistically thorough comparative evaluation of the most popular search tools.

In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent Web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the Web. Summarize some of these efforts below.

2.1.2 Agent-Based Approach

The agent-based approach to Web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize Web-based information. Generally, the agent-based Web mining systems can be placed into the following three categories:

2.1.2.1 Intelligent Search Agents

Several intelligent Web agents have been developed that search for relevant information using characteristics of a particular domain (and possibly a user profile) to organize and interpret the discovered information. For example, agents such as Harvest, FAQ-Finder, Information Manifold, OCCAM, and Parasite rely either on pre-specified and domain specific information about particular types of documents, or on hard coded models of the information sources to retrieve and interpret documents. Other agents, such as ShopBot and ILA (Internet Learning Agent), attempt to interact with and learn the structure of unfamiliar information sources. ShopBot retrieves product information from a variety of vendor sites using only general information about the product domain. ILA, on the other hand, learns models of various information sources and translates these into its own internal concept hierarchy.

2.1.2.2 Information Filtering/Categorization

A number of Web agents use various information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them. For example, HyPursuit uses semantic information embedded in link structures as well as document content to create cluster hierarchies of hypertext documents, and structure an information space. BO (Bookmark Organizer) combines hierarchical clustering techniques and user interaction to organize a collection of Web documents based on conceptual information.

2.1.2.3 Personalized Web Agents

Another category of Web agents includes those that obtain or learn user preferences and discover Web information sources that correspond to these preferences, and possibly those of other individuals with similar interests (using collaborative filtering). A few recent examples of such agents include the WebWatcher, PAINT, Syskill & Webert, and others. For example, Syskill & Webert is a system that utilizes a user profile and learns to rate Web pages of interest using a Bayesian classifier.

2.2 Database Approach

The database approaches to Web mining have generally focused on techniques for integrating and organizing the heterogeneous and semi-structured data on the Web into more structured and high-level collections of resources, such as in relational databases, and using standard database querying mechanisms and data mining techniques to access and analyze this information.

2.2.1 Multilevel Databases

Several researchers have proposed a multilevel database approach to organizing Web-based information. The main idea behind these proposals is that the lowest level of the database contains primitive semi-structured information stored in various Web repositories, such as hypertext documents. At the higher level(s) meta data or generalizations are extracted from lower levels and organized in structured collections such as relational or object-oriented databases. For example, use a multi-layered database where each layer is obtained via generalization and transformation operations performed on the lower layers. Propose the creation and maintenance of meta-databases at each information providing domain and the use of a global schema for the meta-database. King & Novak propose the incremental integration of a portion of the schema from each information source, rather than relying on a global heterogeneous database schema. ARANEUS system extracts relevant information from hypertext documents and integrates these into higher-level derived Web Hypertexts which are generalizations of the notion of database views.

2.2.2 Web Query Systems

There have been many Web-base query systems and languages developed recently that attempt to utilize standard database query languages such as SQL, structural information about Web documents, and even natural language processing for accommodating the types of queries that are used in World Wide Web searches. Few examples of these Web-base query systems here. W3QL: combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques. WebLog: Logic-based query language for restructuring extracted information from Web information sources. Lorel and UnQL: query heterogeneous and semi-structured information on the Web using a labeled graph data model. TSIMMIS: extracts data from heterogeneous and semi-structured information sources and correlates them to generate an integrated database representation of the extracted information.

2.3 Web Usage Mining

Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs which contains information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts.

Analyzing such data can help these organizations to determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. Analysis of server access logs and user registration data can also provide valuable information on how to better structure a Web site in order to create a more effective presence for the organization. In organizations using intranet technologies, such analysis can shed light on more effective management of workgroup communication and organizational infrastructure.

Finally, for organizations that sell advertising on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users.

Most of the existing Web analysis tools provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools, for example, it is possible to determine the number of accesses to the server and the individual files within the organization's Web space, the times or time intervals of visits, and domain names and the URLs of users of the Web server. However, in general, these tools are designed to deal handle low to moderate traffic servers, and furthermore, they usually provide little or no analysis of data relationships among the accessed files and directories within the Web space.

More sophisticated systems and techniques for discovery and analysis of patterns are now emerging. These tools can be placed into two main categories, as discussed below.

2.3.1 Pattern Discovery Tools:

The emerging tools for user pattern discovery use sophisticated techniques from AI, data mining, psychology, and information theory, to mine for knowledge from collected data. For example, the WEBMINER system introduces a general architecture for Web usage mining. WEBMINER automatically discovers association rules and sequential patterns from server access logs. The algorithms are introduced for finding maximal forward references and large reference sequences. These can, in turn be used to perform various types of user traversal path analysis such as identifying the most traversed paths thorough a Web locality. Pirolli use information foraging theory to combine path traversal patterns, Web page typing, and site topology information to categorize pages for easier access by users.

2.3.2 Pattern Analysis Tools

Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns. Examples of such tools include the WebViz system for visualizing path traversal patterns. Others have proposed using OLAP techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from server access logs. The WEBMINER system proposes an SQL-like query mechanism for querying the discovered

2.4 Pattern Discovery from Web Transactions

Analysis of how users are accessing a site is critical for determining effective marketing strategies and optimizing the logical structure of the Web site. Because of many unique characteristics of the client-server model in the World Wide Web, including differences between the physical topology of Web repositories and user access paths, and the difficulty in identification of unique users as well as user sessions or transactions, it is necessary to develop a new framework to enable the mining process. Specifically, there are a number of issues in pre-processing data for mining that must be addressed before the mining algorithms can be run. These include developing a model of access log data, developing techniques to clean/filter the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units (i.e. transactions), integration of various data sources such as user registration information, and specializing generic data mining algorithms to take advantage of the specific nature of access log data.

2.4.1 Preprocessing Tasks

2.4.1.1 Data Cleaning

Techniques to clean a server log to eliminate irrelevant items are of importance for any type of Web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses of the Web site. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, jpg, JPG, and map can be removed.

A related but much harder problem is determining if there are important accesses that are not recorded in the access log. Mechanisms such as local caches and proxy servers can severely distort the overall picture of user traversals through a Web site. A page that is listed only once in an access log may have in fact been referenced many times by multiple users. Current methods to try to overcome this problem include the use of cookies, cache busting, and explicit user registration. None of these methods are without serious drawbacks. Cookies can be deleted by the user, cache busting defeats the speed advantage that caching was created to provide and can be disabled, and user registration is voluntary and users often provide false information. Methods for dealing with the caching problem include using site topology or referrer logs, along with temporal information to infer missing references.

Another problem associated with proxy servers is that of user identification. Use of a machine name to uniquely identify users can result in several users being erroneously grouped together as one user. An algorithm checks to see if each incoming request is reachable from the pages already visited. If a page is requested that is not directly linked to the previous pages, multiple users are assumed to exist on the same machine. User session lengths determined automatically based on navigation patterns are used to identify users. Other heuristics involve using a combination of IP address, machine name, browser agent, and temporal information to identify users.

2.4.1.2 Transaction Identification

Before any mining is done on Web usage data, sequences of page references must be grouped into logical units representing Web transactions or user sessions. A user session is all of the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users, as discussed above. A transaction differs from a user session in that the size of a transaction can range from a single page reference to all of the page references in a user session, depending on the criteria used to identify transactions. Unlike traditional domains for data mining, such as point of sale databases, there is no convenient method of clustering page references into transactions smaller than an entire user session.

2.5 Discovery Techniques on Web Transactions

Once user transactions or sessions have been identified there are several kinds of access pattern mining that can be performed depending on the needs of the analyst such as

i. Path analysis

ii. Association rules

iii. Sequential patterns

iv. Clustering and Classification.

2.5.1 Path Analysis

There are many different types of graphs that can be formed for performing path analysis, since a graph represents some relation defined on Web pages (or other objects). The most obvious is a graph representing the physical layout of a Web site, with Web pages as nodes and hypertext links between pages as directed edges. Other graphs could be formed based on the types of Web pages with edges representing similarity between pages, or creating edges that give the number of users that go from one page to another. Most of the work to date involves determining frequent traversal patterns or large reference sequences from the physical layout type of graph. The navigation-content transactions of maximal forward reference transactions or user sessions can be used for path analysis. Path analysis could be used to determine most frequently visited paths in a Web site. Other examples of information that can be discovered through path analysis are:

- 70% of clients who accessed /company/products/file2.html did so by starting at /company and proceeding through /company/whatsnew, /company/products, and /company/products/file1.html;
- 80% of clients who accessed the site started from /company/products; or
- 65% of clients left the site after four or less page references.

The first rule suggests that there is useful information in /company/products/file2.html, but since users tend to take a circuitous route to the page, it is not clearly marked. The second rule simply states that the majority of users are accessing the site through a page other than the main page (assumed to be /company in this example) and it might be a good idea to include directory type information on this page if it is not there already. The last rule indicates an attrition rate for the site. Since many users don't browse further than four pages into the site, it would be prudent to ensure that important information is contained within four pages of the common site entry points.

2.5.2 Association Rules

Association rule discovery techniques are generally applied to databases of transactions where each transaction consists of a set of items. In such a framework the problem is to discover all associations and correlations among data items where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items. In the context of Web mining, this problem amounts to discovering the correlations among references to various files available on the server by a given client. Each transaction is comprised of a set of URLs accessed by a client in one visit to the server. For example, using association rule discovery techniques it can find correlations such as the following:

- 40% of clients who accessed the Web page with URL /company/products/ product1.html, also accessed /company/products/product2.html; or
- 30% of clients who accessed /company/announcements/special-offer.html, placed an online order in /company/products/product1.

Since usually such transaction databases contain extremely large amounts of data, current association rule discovery techniques try to prune the search space according to *support* for items under consideration. Support is a measure based on the number of occurrences of user transactions within transaction logs.

Discovery of such rules for organizations engaged in electronic commerce can help in the development of effective marketing strategies. But, in addition, association rules discovered from WWW access logs can give an indication of how to best organize the organization's Web space. For example, if one discovers that 80% of the clients accessing /company/products and /company/products/file1.html also accessed /company/products/file2.html, but only 30% of those who accessed /company/products also accessed /company/products/file2.html, then it is likely that some information in file1.html leads clients to access file2.html. This correlation might suggest that this information should be moved to a higher level (e.g., /company/products) to increase access to file2.html.

2.5.3 Sequential Patterns

The problem of discovering sequential patterns is to find inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. In Web server transaction logs, a visit by a client is recorded over a period of time. The time stamp associated with a transaction in this case will be a time interval which is determined and attached to the transaction during the data cleaning or transaction identification processes. The discovery of sequential patterns in Web server access logs allows Web-based organizations to predict user visit patterns and helps in targeting advertising aimed at groups of users based on these patterns. By analyzing this information, the Web mining system can determine temporal relationships among data items such as the following:

- 30% of clients who visited /company/products/, had done a search in Yahoo, within the past week on keyword **w**; or
- 60% of clients, who placed an online order in/company/products / product1.html, also placed an online order in /company1/products/product4 within 15 days.

Another important kind of data dependency that can be discovered, using the temporal characteristics of the data, are similar time sequences. For example, Admin may be interested in finding common characteristics of all clients that visited a particular file within the time period [t_1, t_2]. Or, conversely, Admin may be interested in a time interval (within a day, or within a week, etc.) in which a particular file is most accessed.

2.5.4 Clustering and Classification

Discovering classification rules allows one to develop a profile of items belonging to a particular group according to their common attributes. This profile can then be used to classify new data items that are added to the database. In Web mining, classification techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients, or based on their access patterns. For example classification on WWW access logs may lead to the discovery of relationships such as the following:

- clients from state or government agencies who visit the site tend to be interested in the page /company/products/product1.html; or
- 50% of clients who placed an online order in /company/products/product2, were in the 20-25 age group and lived on the West Coast.

In some cases, valuable information about clients can be gathered by the server automatically from the client browsers. This includes information available on the client side in the history files, cookie files, etc. Other methods used to obtain profile and demographic information on clients include user registration, online survey forms, and techniques such as ``anonymous ticketing".

Clustering analysis allows one to group together clients or data items that have similar characteristics. Clustering of client information or data items on Web transaction logs, can facilitate the development and execution of future marketing strategies, both online and off-line, such as automated return mail to clients falling within a certain cluster, or dynamically changing a particular site for a client, on a return visit, based on past classification of that client.

2.6 Analysis of Discovered Patterns

Web site administrators are extremely interested in questions like "How are people using the site?", "Which pages are being accessed most frequently?" etc. These questions require the analysis of the structure of hyperlinks as well as the contents of the pages. The end products of such analysis might include 1) the frequency of visits per document, 2) most recent visit per document, 3) who is visiting which documents, 4) frequency of use of each hyperlink, and 5) most recent use of each hyperlink.

The discovery of Web usage patterns, carried out by techniques described earlier, would not be very useful unless there were mechanisms and tools to help an analyst better understand them. Hence, in addition to developing techniques for mining usage patterns form Web logs, there is a need to develop techniques and tools for enabling the analysis of discovered patterns. These techniques are expected to draw upon from a number of fields including statistics, graphics and visualization, usability analysis, and database querying. In this section it provides a survey of the existing tools and techniques. Usage analysis of Web access behavior being a very new area, there is very little work in it, and correspondingly this survey is not very extensive

2.6.1 Visualization Techniques

Visualization has been used very successfully in helping people understand various kinds of phenomena, both real and abstract. Hence it is a natural choice for understanding the behavior of Web users. Pitkow have developed the WebViz system for visualizing WWW access patterns. A Web path paradigm is proposed in which sets of server log entries are used to extract subsequences of Web traversal patterns called *Web paths.* WebViz allows the analyst to selectively analyze the portion of the Web that is of interest by filtering out the irrelevant portions. The Web is visualized as a directed graph with cycles, where nodes are pages and edges are (inter-page) hyperlinks.

The visualization is composed of two windows, the WebViz control window and the display window. The first provides the analyst with controls to adjust the bindings, select a specific time to view, control the animation, and rearrange the layout. The second window's arrangement allows a document's access frequency to be represented by the width of the node representing it, while the node's color represents it recency of access. Link width and color have corresponding meanings. Temporal manipulation is achieved by either the slider of by playback controls.

2.6.2 OLAP Techniques

On-Line Analytical Processing (OLAP) is emerging as a very powerful paradigm for strategic analysis of databases in business settings. Some of the key characteristics of strategic analysis include 1) very large data volume, 2) explicit support for the temporal dimension, 3) support for various kinds of information aggregation, and 4) long-range analysis, where overall trends are more important than details of individual data items. While OLAP can be performed directly on top of relational databases, industry has developed specialized tools to make it more efficient and effective. Also, the research community has recently demonstrated that the functional and performance needs of OLAP require that new information structures be designed. This has led to the development of the data cube information model and techniques for its efficient implementation.

Recent work has shown that the analysis needs of Web usage data have much in common with those of a data warehouse, and hence OLAP techniques are quite applicable. The access information in server logs is modeled as an append-only history, which grows over time. A single access log is not likely to contain the entire request history for pages on a server, especially since many clients use a proxy server.

Because information on access requests will be distributed, and there is a need to integrate it. Since the size of server logs grows quite rapidly, it may not be possible to provide on-line analysis of all of it. Therefore, there is a need to summarize the log data, perhaps in various ways, to make its on-line analysis feasible. Making portions of the log selectively (in)visible to various analysts may be required for security reasons. These requirements for Web usage data analysis show that OLAP techniques may be quite applicable, and this issue needs further investigation.

2.6.3 Data & Knowledge Querying

One of the reasons attributed to the great success of relational database technology has been the existence of a high-level, declarative, query language, which allows an application to express what conditions must be satisfied by the data it needs, rather than having to specify how to get the required data. Given the large number of patterns that may be mined, there appears to be a definite need for a mechanism to specify the focus of the analysis. Such focus may be provided in at least two ways. First, constraints may be placed on the database (perhaps in a declarative language) to restrict the portion of the database to be mined. Second, querying may be performed on the knowledge that has been extracted by the mining process, in which case a language for querying knowledge rather than data is needed. An SQL-like querying mechanism has been proposed for the WEBMINER system.

2.6.4 Usability Analysis

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

As described in this section, there is an increasing need for, as well as interest in, developing techniques and tools to analyze the usage patterns of information on the Web. Some initial ideas have been proposed, but are still in their nascent stages and much work remains to be done. Believe that the techniques which are most effective will include the following characteristics: (i) will be data driven empirical methods, (ii) will use vast amounts of data for validation, (iii) will use rigorous experimental methods and sound statistical analysis, etc.

CHAPTER 3

3 WEB USAGE MINING SYSTEM

The framework for our Web usage mining is summarized in Figure 3.1, which starts with the integration and preprocessing of Web server logs and server content databases, includes data cleaning and sessionization, and then continues with the data mining/pattern discovery via clustering. This is followed by a post processing of the clustering results to obtain Web user profiles and finally ends with tracking profile evolution.

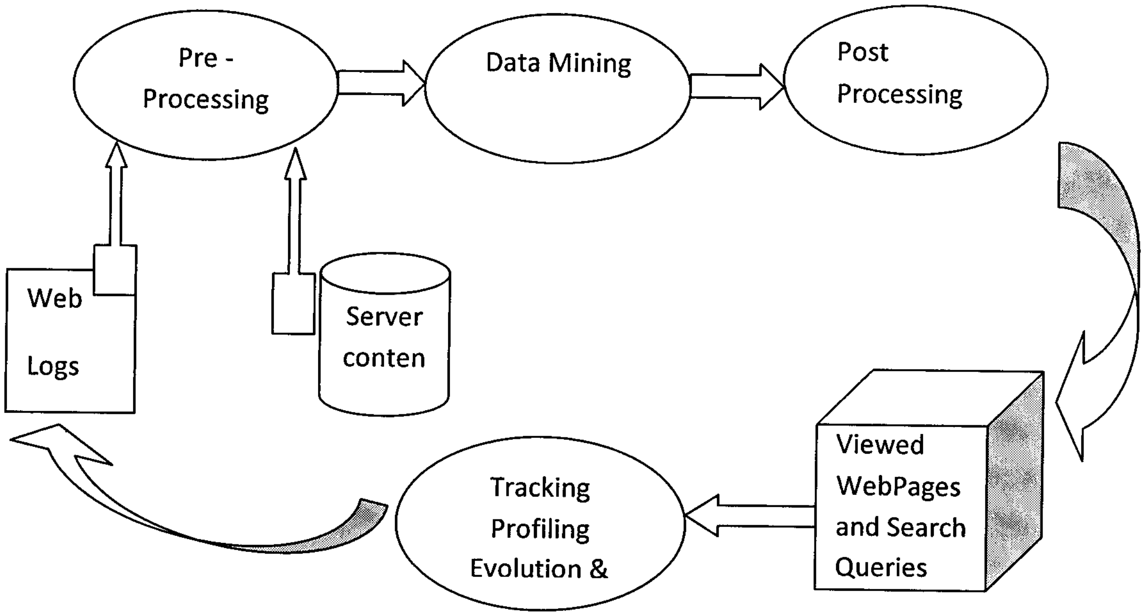

Pre - Processing
Data Mining
Post Processing
Web Logs
Server conten
Viewed WebPages and Search Queries
Tracking Profiling Evolution &

Figure 3.1: Web usage mining process

The automatic identification of user profiles is a knowledge discovery task consisting of periodically mining new contents of the user access log files and is summarized in the following steps:

1. Collection of Web Log and Updation into a Database.
2. Preprocessing the Web Log File to Extract User Sessions.
    a. Adding Semantics with Ontology Concepts.
    b. Mapping of Content Keywords.
3. Clustering of the user sessions by using Hierarchical Unsupervised Niche Clustering (H - UNC)
4. Summarize session clusters/categories into user profiles
5. Tracking current profiles against existing profiles.

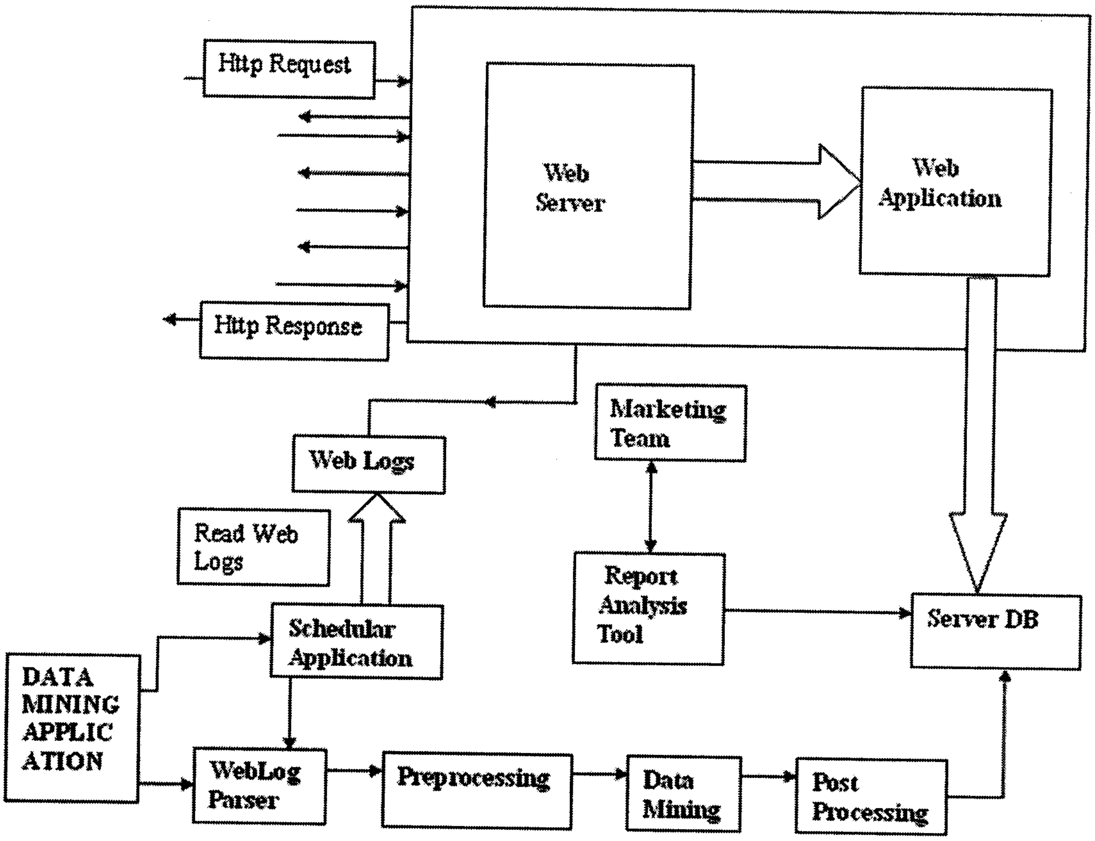


Figure 3.2: Web usage mining system architecture

3.1 Collection of Web Log data and Database Updation

The Clickstreams and URLs that are navigated or visited by the user are updated continuously in the web log file. It was crucial to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time. The web log file is studied and data in the web log file is sorted and analyzed using special tokeniser functionality and delimiter.

The analyzed data are taken as updated into the respective columns of the database which maintains the web log file in structured manner. This structured storage facility make it possible for us to implement the datamining algorithm to mine suitable rules and patterns based on which the user profile can be processed.

3.2 Preprocessing the Web Log File to Extract User Sessions

It was crucial to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time. For this reason, we perform clustering of the user sessions extracted from the Web logs to partition the users into several homogeneous groups with similar activities and then extract user profiles from each cluster as a set of relevant URLs.

a. Adding Semantics with Ontology Concepts :

When clustering the user sessions, we exploit the Web site hierarchy to give partial weights in the session similarity between URLs that are distinct and yet located closer together on this hierarchy. The Web site hierarchy is inferred both from the URL address and from a Web site database that organizes most of the dynamic URLs along an “is-a” ontology of items. We also enrich the cluster profiles with various facets, including search queries submitted just before landing on the Web site, and inquiring and inquired companies, in case users from (inquiring) companies inquire about any of the (inquired) companies listed on the Web site, which provide related services. Relying only on Web usage data for user modeling or for personalization can be inefficient, either when there is insufficient usage data for the purpose of mining certain patterns or when new pages are added and thus do not accumulate sufficient usage data at first. The lack of usage data in these cases can be compensated by adding other information such as the content of Web pages or the structure of a Web site. In the keywords that appear in Web pages are used to generate document vectors, which are later clustered in the document space to further augment user profiles. In the Web site’s own hierarchical structure is treated like an implicit taxonomy or concept hierarchy that is exploited in computing the similarity between any two Web pages on the Web site. This allows a better comparison between sessions that contain visits to Web pages that are different and yet semantically related (for example, under the same more general topic). The idea of exploiting concept hierarchies or taxonomies has already been found to enhance association rule mining and to facilitate information searching in textual data.

b. Mapping of Content Keywords :

A dynamic URL is a page address that results from the search of a database-driven Web site or a Web site that runs a script. Unlike static URLs, in which the contents of the Web page do not change, dynamic URLs are typically generated from specific queries to a site’s database. Even though the examples given in the following discussion consistently use the ASP extension, this extension can be replaced by any other dynamic URL extension (such as PHP), without any changes in our generic approach. Although static Web pages tend to have meaningful URLs such as /reports/fall_2003/benefits.html, most dynamic URLs such as /universal.aspx?id=55&codes _id=60 are unfortunately hard to discern or even recognize based only on their URL. We resolved this issue by resorting to available external data that maps database contents to a dynamic resource and its parameter values. The ASP codes in most menus can be mapped during the preprocessing phase to a parent/child structure by using external data, thus mapping URLs to meaningful hierarchical descriptions.

3.3 Clustering of the user sessions by using Hierarchical Unsupervised Niche_Clustering (H - UNC)

We use H-UNC instead of other clustering algorithms is that unlike most other algorithms, H-UNC can handle noise in the data and automatically determines the number of clusters. In addition, evolutionary optimization allows the use of any domain-specific optimization criterion and any similarity measure, in particular a subjective measure that exploits domain knowledge or ontologies. However, unlike purely evolutionary search-based algorithms, H-UNC combines evolution with local Piccard updates to estimate the scale of each profile. The data is continuously pre-processed to produce session lists: A session list si for user i is an item list of URLs visited by the same user. In discovery mode, a session is fed to the learning system as soon as it is available. The ith B-Cell, D-W-B-cell(i), represents the ith candidate profile and encodes a list of relevant URLs. It is matched to incoming sessions using (1-Cosine similarity) as a distance measure. Each profile has its own influence zone defined by sigma square (i). This measures the average dissimilarity between the ith candidate profile and the sessions/antigens that activate the ith B-Cell, D-W- B-cell(i).

3.4 Summarize session clusters/categories into user profiles

After automatically grouping sessions into different clusters, we summarize the session categories in terms of user profile vectors. The kth component/weight of this vector (pik) captures the relevance of URL(k) in the ith profile, as estimated by the conditional probability that URLk is accessed in a session belonging to the ith cluster (this is the frequency with which URLk was accessed in the sessions belonging to the ith cluster). The profiles are then converted to binary vectors (sets) so that only URLs with weights > 0:15 remain. The model is further extended to a robust profile based on robust weights (wij) computed in the UNC algorithm that assign only sessions with high robust weights (that is, wij > wmin) to a cluster's core. The core of a profile consists only of sessions that are very similar to the representative profile.

3.5 Tracking current profiles against existing profiles

Tracking different profile events across different time periods can generate a better understanding of the evolution of user access patterns and seasonality. Note that both profiles and clickstreams are typically evolving, since the profiles are nothing more than summaries of the clickstreams, which are themselves evolving. Each profile pi is discovered along with an automatically determined measure of scale sigma (i) that represents the amount of variance or dispersion of the user sessions in a given cluster around the cluster representative. This measure is used to determine the boundary around each cluster (an area located at a distance _i from the profile pi) and thus allows us to automatically determine whether two profiles are compatible. Two profiles are compatible if their boundaries overlap. The notion of compatibility between profiles is essential for tracking evolving profiles. After mining the Web log of a given period, we perform an automated comparison between all the profiles discovered in the current batch and the profiles discovered in the previous batch by a sequence of SQL queries on the profiles that have been stored in a database.

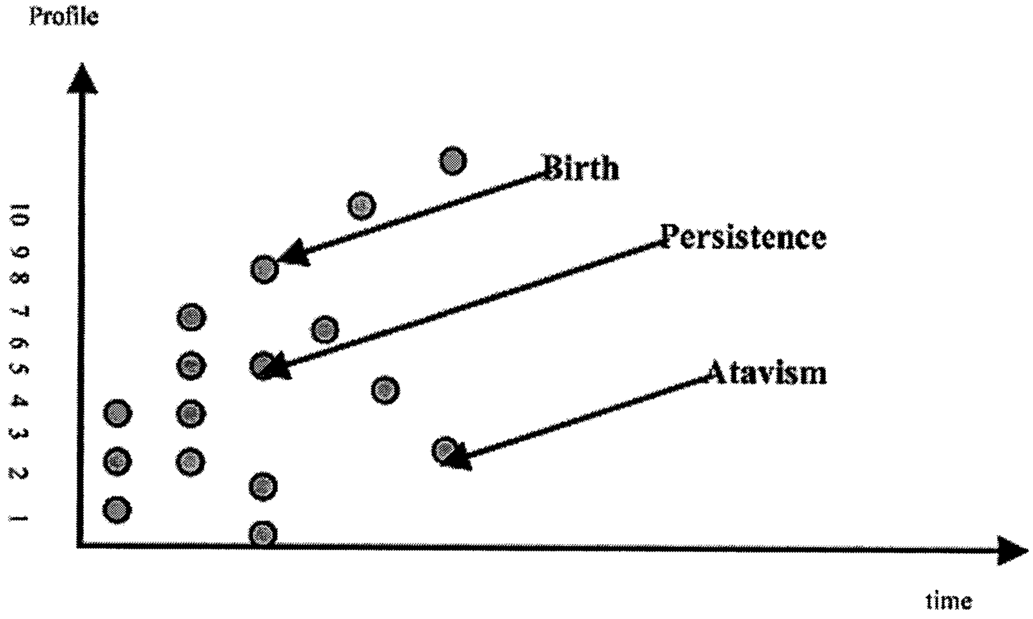


Figure 3.3: Visualization of the profile evolution

CHAPTER 4

4. SYSTEM STUDY

4.1 Existing System Study

Though Web usage mining is a fairly new research topic, many systems and tools are already on the market. Most provide only limited knowledge or information, such as the number of hits, the popular paths/products, etc. Most previous research efforts in Web usage mining have worked with the assumption that the Web usage data is static. However, the dynamic aspects of Web usage have recently become important. This is because Web access patterns on a Web site are dynamic due not only to the dynamics of Web site content and structure but also to changes in the user's interests and, thus, their navigation patterns. Thus, it is desirable to study and discover Web usage patterns at a higher level, where such dynamic tendencies and temporal events can be distinguished.

An improved understanding of the customer's habits, needs, and interests can allow the business to profit by, for instance, "cross selling" or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers' preferences and needs forms the basis for effective CRM. As businesses move online, the competition between businesses to keep the loyalty of their old customers and to lure new customers is even more important, since a competitor's Web site may be only one click away. The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles.

These different modes of usage or the so-called mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user clickstreams stored in Web log files. These profiles can later be harnessed toward personalizing the Web site to the user or to support targeted marketing. Although there have been considerable advances in Web usage mining, there have been no detailed studies presenting a fully integrated approach to mine a real Web site with the challenging characteristics of today's Web sites, such as evolving profiles, dynamic content, and the availability of taxonomy or databases in addition to Web logs.

4.2 Drawbacks of Existing System

- The Existing System consists of Customer Relationship Management functionalities involving Client Server architecture and other off line architectures. So, the Customer Profiling or User Profiling consists of static content only.
- Most previous research efforts in Web usage mining have worked with the assumption that the Web usage data is static.
- There have been no detailed studies presenting a fully integrated approach to mine a real Web site with the challenging characteristics of today's Web sites
- It cannot perform data mining on evolving profiles.
- No strategy is available for dynamic content User Profiling
- It does not support the availability of the specific domain taxonomy or databases in addition to Web logs.

4.3 Proposed System

The proposed System is a complete framework and a summary of our experience in mining Web usage patterns with real-world challenges such as evolving access patterns, dynamic pages, and external data describing an ontology of the Web content and how it relates to the business actors (in the case of the studied Web site, the Companies, contractors, consultants, etc., in corrosion). The Web site in our study is managed by a nonprofit organization that does not sell anything but only provides free information that is ideally complete, accurate, and up to date. Hence, it was crucial to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time. For this reason, we perform clustering of the user sessions extracted from the Web logs to partition the users into several homogeneous groups with similar activities and then extract user profiles from each cluster as a set of relevant URLs. And thus the user profiles and existing user profiles are been tracked.

4.4 Advantages of Proposed System

- Reliable knowledge about the customers' preferences and needs forms the basis for effective CRM.
- It Supports fast pace and large amounts of data available in these online settings
- These profiles can later be harnessed toward personalizing the Web site to the user or to support targeted marketing.
- It present a complete framework and a summary of our experience in mining Web usage patterns with real-world challenges such as evolving access patterns, dynamic pages, and external data describing an ontology of the Web content and how it relates to the business actors
- It was easy to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time.

CHAPTER 5

5. SYSTEM ANALYSIS AND DESIGN

5.1 Feasibility Study

Feasibility study is the high level capsule version of the entire requirement analysis process. The objective of feasibility study is to determine whether the proposed system can be developing with availability resources.

There are three steps to be followed for determining the feasibility study of proposed system.

- Technical Feasibility
- Operational Feasibility
- Economical Feasibility

5.1.1 Technical Feasibility

It is concern with hardware and software feasibility. In this study, one has to test whether the proposed system can be developed using existing technology or not. If new technology is required, what is the likely hood that it can be developed? As per our client requirement the system to be developed should have speed response because of fast and easy communication, programming productivity, reliability, security, scalability, integrity and availability. To meet these requirements as a developer can choose Microsoft Visual studio 2005 as a right choice and use Visual basic .NET. For the Backend for data storage Microsoft SQL Server 2005 Express edition can be used due to its powerful feature and free edition.

5.1.2 Operational Feasibility

Operational feasibility determines whether the proposed system satisfied the user objectives and can be fitted in to the current system operation. The methods of processing and presentation are completely acceptable by the clients because they meet all the user and client requirements. The clients have been involved during the preparation of requirement analysis and design process. The system will certainly satisfy the user objectives and it will also enhance their capability. The proposed system is completely user friendly.

5.1.3 Economical Feasibility

This includes an evaluation of all incremental costs and benefits expected if proposed system is implemented. Cost-benefit analysis which is to be performed during economical feasibility delineates costs for project development and weights them against benefits of the system.

5.2 Global Use-case Diagrams

5.2.1 Use Case Diagram for Log File generation

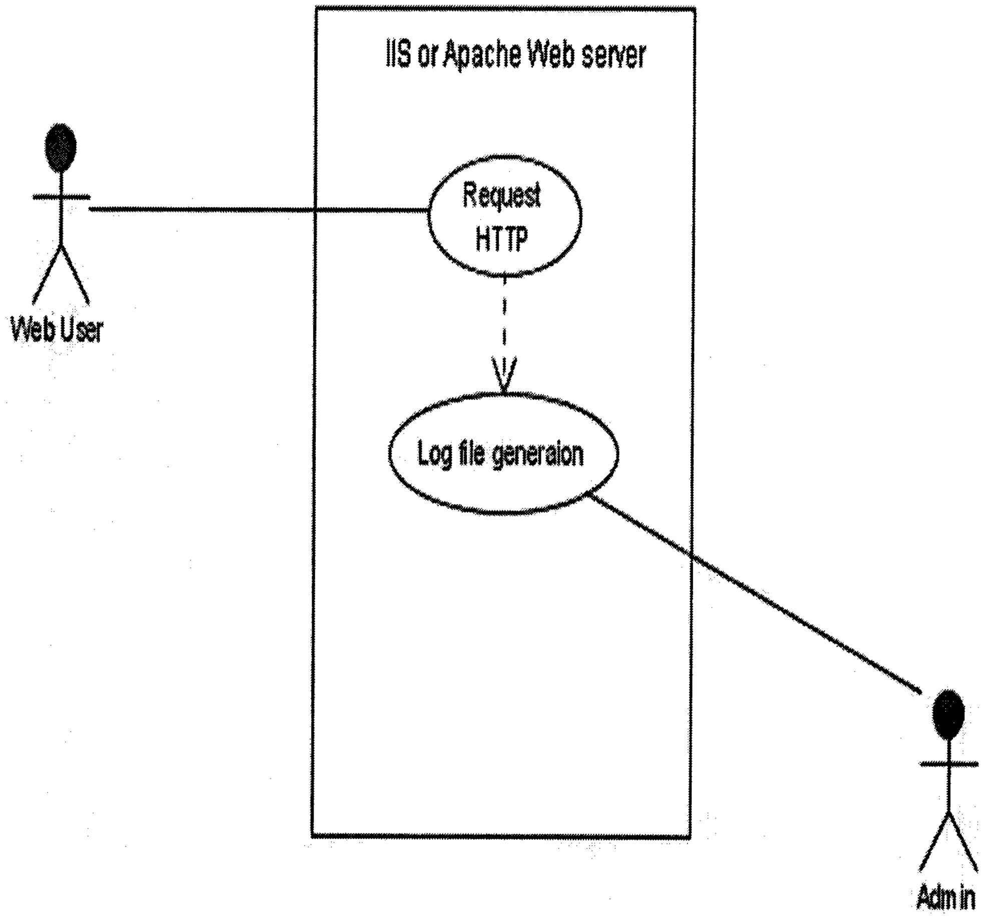


Figure 5.1*:* Use Case Diagram for Log File generation

5.2.2 Use Case Diagram for Processing Log File

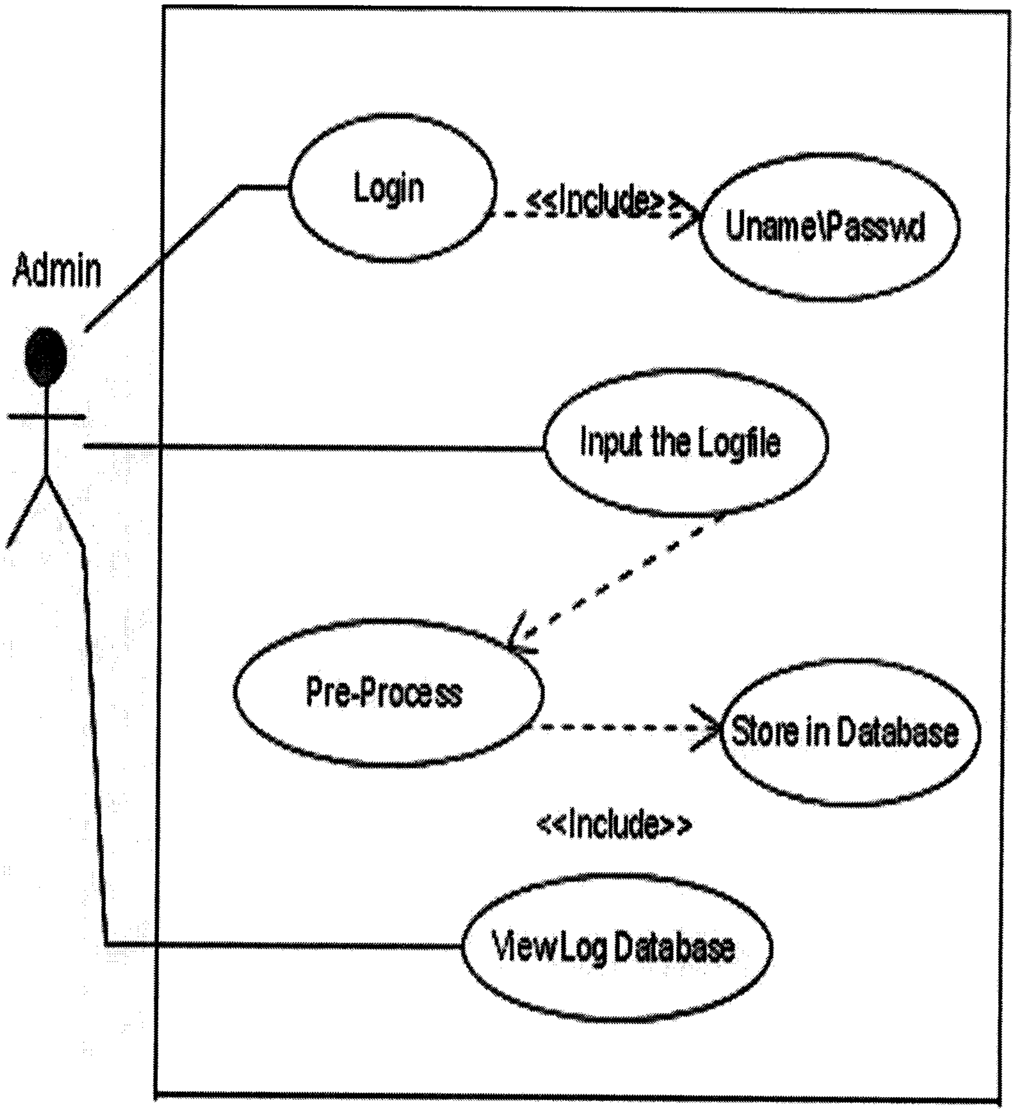


Figure 5.2: Use Case Diagram for Processing Log File

5.2.3 Use Case Diagram for Web Mining

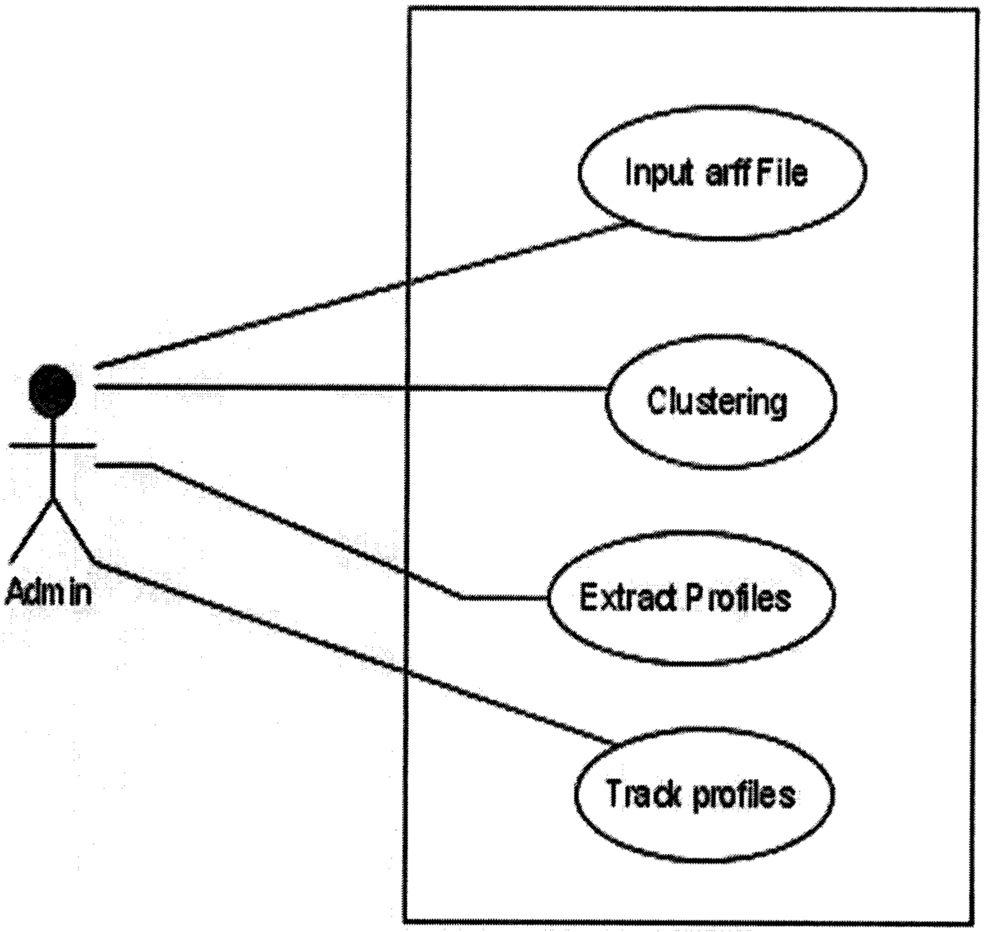


Figure 5.3: Use Case Diagram for Web Mining

5.3 Activity Diagram

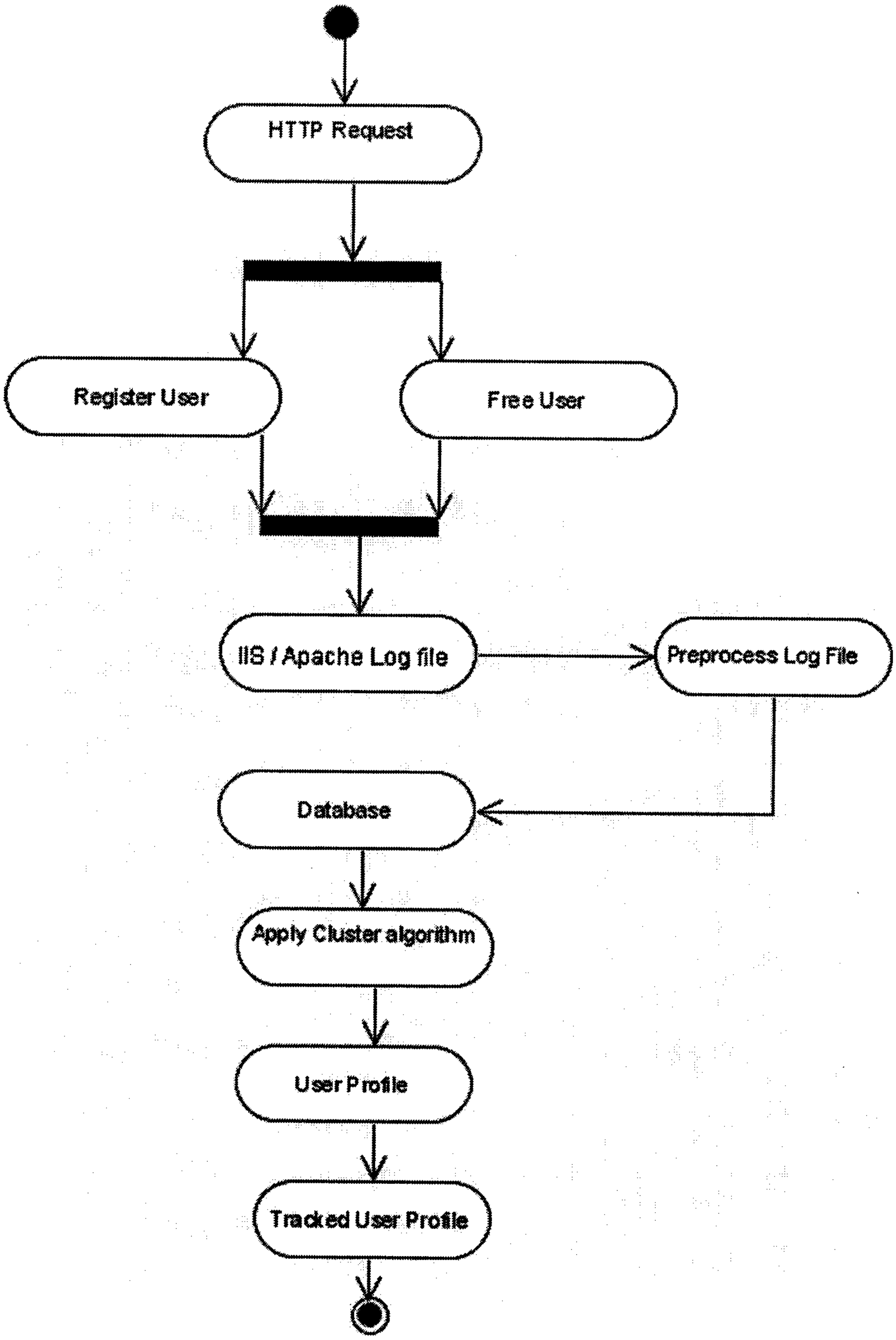


Figure 5.4: Activity Diagram for Web Usage Mining System

5.4 Software Requirements Specification

Requirement Analysis is the first technical step in software engineering process. It is at this point that a general statement of software scope is refined into concrete specification that becomes the foundation for all software engineering activities that follow.

Analysis must focus on information, functional and behavioral domains of the problem. To better understand what is required, models are created and the problem is partitioned. In many cases it is not possible to completely specify a problem at an early stage. Prototyping offers an alternate approach from which requirements can be refined.

A Software Requirements Specification is developed as a consequence of analysis. Review is essential to ensure that the developer and customers have the same perception.

Software Requirements Specification (SRS) is the starting point of the software development activity. The Software Requirements Specification is produced at the culmination of the analysis task. The introduction of the software requirements specification states the goals and objectives of the software, describing it in the context of the computer-based system. The SRS includes an information description, functional description, behavioral description, validation criteria.

The purpose of this document is to present the software requirements in a precise and easily understood manner. This document provides the functional, performance, design and verification requirements of the software to be developed. \

This is the only document that describes the requirements of the system. This is meant for use by the developers and will also be the basis for validating the final delivered system.

What is a requirement?

A requirement is a statement about what the proposed system will do that all stakeholders (users, customers, developers and their management) agree must be made true in order for the customer's problem to be adequately solved.

Types of requirements

Requirements can be divided into two major types. Requirements document normally include both. The following are the two major categories of requirements.

- Functional requirements.
- Non-functional requirements.

5.4.1 Functional requirements

The functional requirements describe what the system should do, i.e., the services provided for the users and for the systems. The functional requirements can be further categorized as follows:

- What ***inputs*** the system should accept, and under what conditions. This includes data and commands both from users and from other systems.
- What ***outputs*** the system should produce, and under what conditions, Outputs can be to the screen or printed. They can also be transmitted to other systems, such as special I/O devices, clients or servers.

- What data the system should ***store*** that other systems might use. This is really a special kind of output that will eventually become an input to other systems. Data which is stored for the exclusive use of this system can be left until design.
- What ***computations*** the system should perform. The computations should be described at the level of all the readers can understand.
- The ***timing and synchronization*** of all the above. Not all systems involve timing and synchronization. This category of functional requirements is of most important in hard real-time systems.

An individual requirement often covers more than one of the above categories.

Input:

A Web Log File generated from User Clicks in the Web Server.

Output:

1. Homogeneous Clusters of User Profiles.
2. Evolving User Profiles

Storage:

The Web Log file after pre-processing is saved in the Database for further processing.

5.4.2 Non-Functional Requirements

Non-functional requirements are constraints that must be adhered to during development. They limit what resources can be used and set bounds on aspects of the software's quality.

One of the most important things about non-functional requirement is to make them verifiable. The verification is normally done by measuring various aspects of the system and seeing if the measurements confirm to the requirements. Non-functional requirements are divided into several groups: The group of categories reflects the five qualities attributes

1. Usability
2. Efficiency
3. Reliability
4. Maintainability
5. Reusability

These requirements constrain the design to meet specified levels of quality.

The second group of non-functional requirements categories constrains the environment and technology of the system like,

1. Platform
2. Technology to be used

5.4.3 Hardware Requirements

- ➢ Processor: At least P-IV or later
- ➢ Primary Memory (RAM): 512 MB or higher
- ➢ Secondary Memory (Hard Disk):80 GB SATA
- ➢ Monitor: Plug and Play Monitor
- ➢ Keyboard: 101/102
- ➢ Mouse:ps/2 compatible mouse
- ➢ NIC : 10/100 Network Interface Card

5.4.4 Software Requirements

- Operating system: windows xp / 2000 / 2003
- Front End: Microsoft Visual Studio 2005
- Database: Microsoft Sql Server 2005
- Web Servers: IIS 5.1, Apache

5.5 Sequence Diagram

5.5.1 Sequence Diagram for Web User:

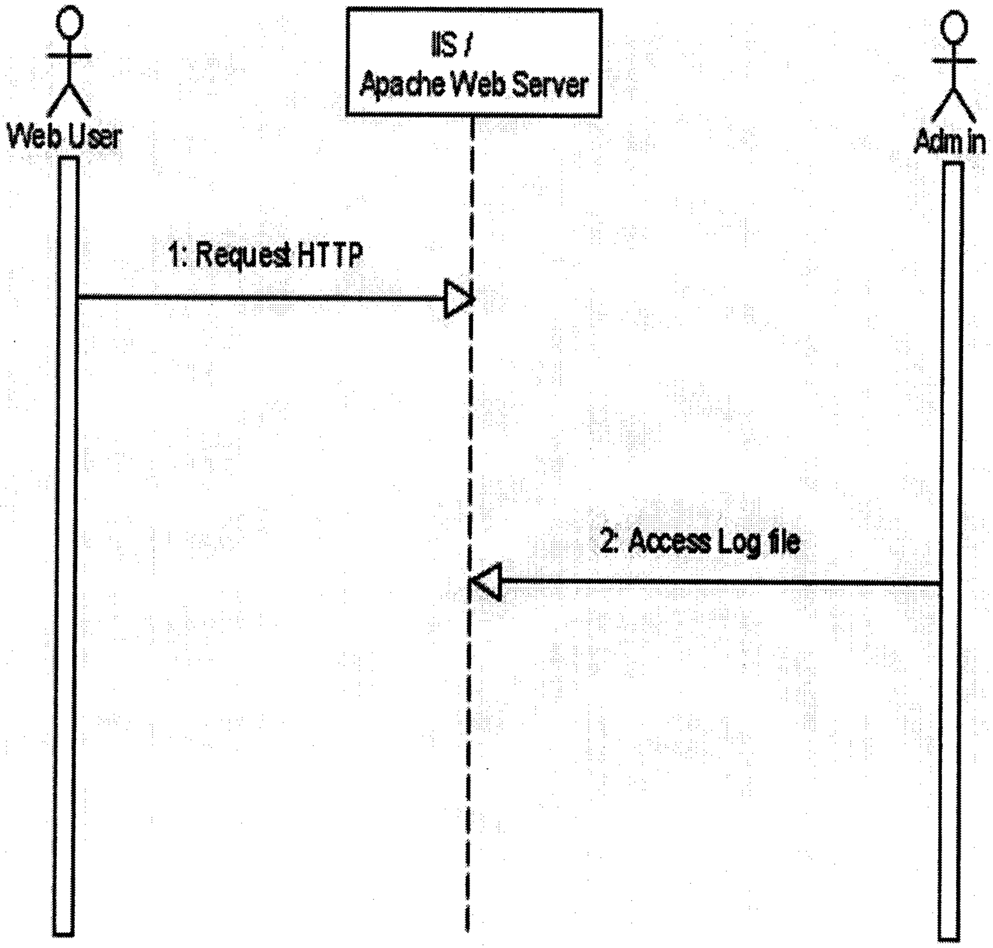


Figure 5.5: Sequence Diagram for Web User

5.5.2 Sequence Diagram for processing Web Log:

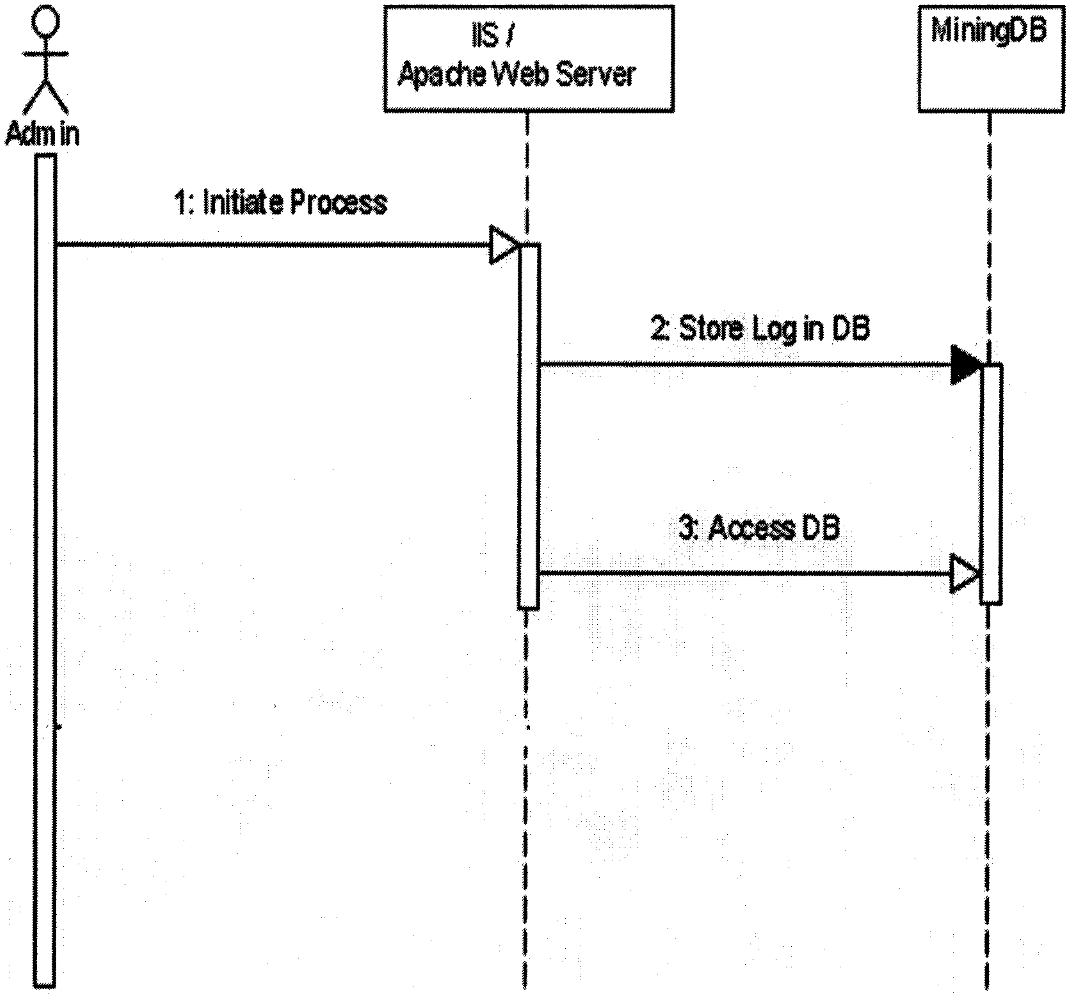


Figure 5.6: Sequence Diagram for Admin processing Log File

5.6 Class Diagram

5.6.1 Class Diagram for Web Usage Mining System:

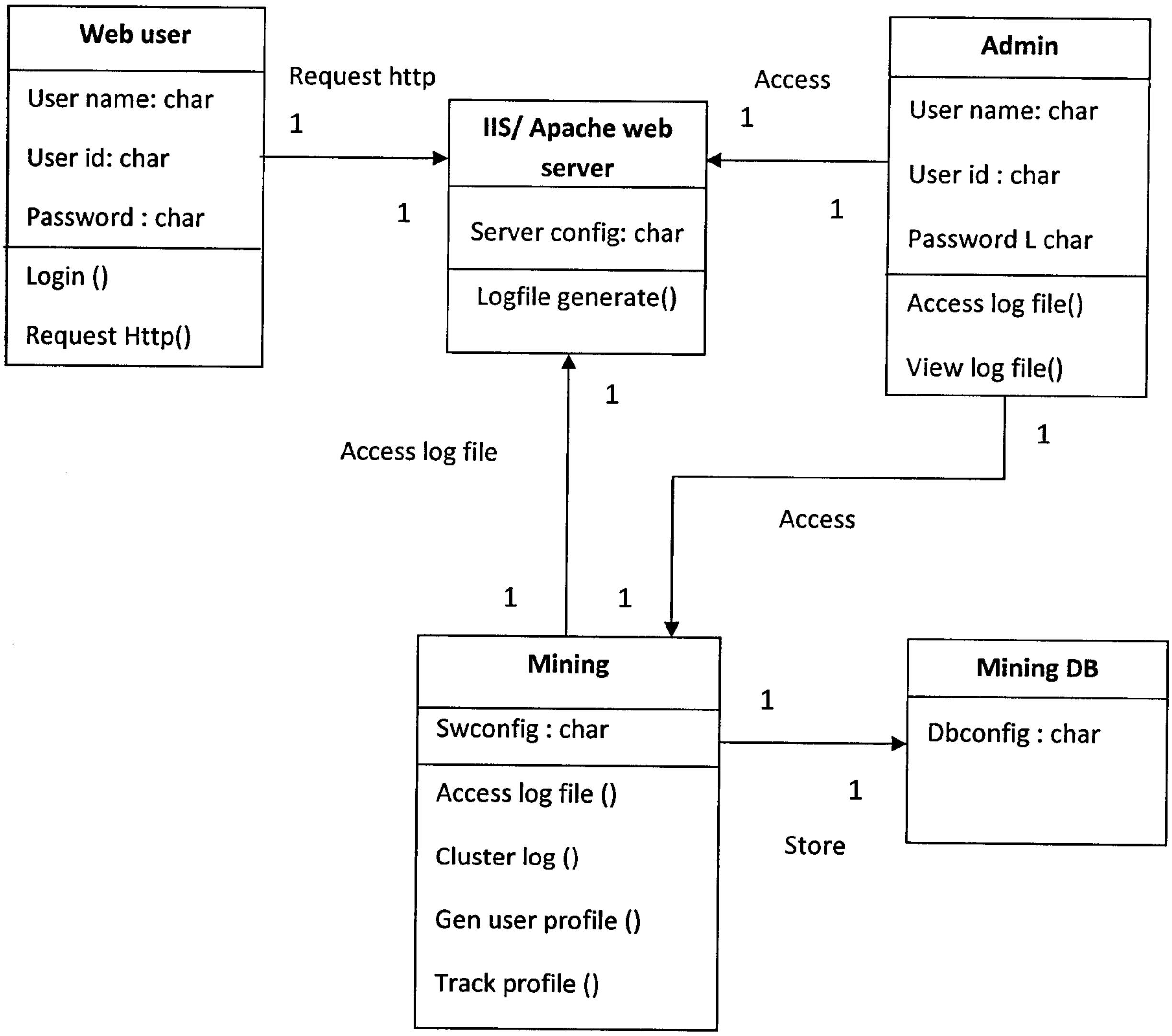


Figure 5.7: Class Diagram for Web Usage Mining System

CHAPTER 6

6. SYSTEM DESIGN METHODOLOGY

6.1 Hierarchical Unsupervised Niche Clustering Algorithm (HUNC)

The Hierarchical Unsupervised Niche Clustering Algorithm (HUNC) HUNC is a divisive hierarchical version of UNC. HUNC has proved its effectiveness when compared to other clustering methodologies. In a recent experiment, HUNC profiles were compared to the profiles resulting from traditional pattern discovery, where the entire usage data from all time periods is used to discover usage patterns in one shot. The latter can be considered as the best output possible since all usage data is mined at once.

However, HUNC has proved that it too can discover profiles that are as good (or better) than using the traditional one-shot method. Most importantly, HUNC has the critical advantage of enabling scalability in handling very large usage data that makes it impossible to mine all patterns in one shot.

Steps for Implementation of Hierarchical Unsupervised Niche Clustering Algorithm

1) Let Resolution Level (L) $= 1$ Initialise the Cluster representive p_i and corresponding Scale σ_i.

2) Repeat until L = $L_{\max}$ or Cluster Cardinality (C.C) $N_i < N_{split}$ **or** Scaleσ_i. $<\sigma_{split}$

3) Increase Resolution level L = L + 1.
4) For each parent cluster represent p_i found at level (L-1).
5) If cluster cardinality $N_i > N_{split}$ or cluster scale $\sigma_i > \sigma_{split}$
6) Reapply UNC on only data records x_j assigned to the cluster representative p_i

Steps for Implementation of Unsupervised Niche Clustering Algorithm

1) Randomly select an initial population of $N_{p.}$

2) Set initial scale σ_i = small fraction (1/10).

3) Update the distance d_{ij} and Robust weight $W_{ij\,.}$

   Where $W_{ij\,=}{}^{-e\,-\,dij\,/\,(2\,\sigma_i\,)}$

4) For i=1 to $N_p/2$ do.

5) Select randomly a candidate parent p_i and p_k from population.

6) Obtain children c1 and c2 by crossover and mutation.

7) Update the scale value, apply deterministic crowding as replacement policy to fill new population

8) Replace the parent if the child fitness is greater.

About Genetic Algortihm:

In Genetic Algorithm, we start with an initial population and then we use some genetic operators on it for appropriate mixing of exploitation and exploration. A simple genetic algorithm consists of an initial population followed by selection, crossover, and mutation. Selection operation selects the best results among the chromosome through some fitness function. The idea of the crossover operation is to swap some information between a pair of chromosomes to obtain a new chromosome. In mutation, a chromosome is altered a little bit randomly to get a new chromosome. The simple structure of the GA is:

```
GA( )
{ Initialize population;
Evaluate population;
While termination criterion not reached
{
select solutions for next population;
perform crossover & mutation;
evaluate population;
}
}
```

We apply genetic algorithm on the input urls taken from the web log data .

Crossover and Mutation

Breeding between two matrices produces two new matrices as the offspring of the parental matrices. From each matrix, select two rows, one from the first and the other from the second. Take a cut point randomly, which is a point between 1 and the highest column number. Divide each row (on which we are applying the crossover) in two parts. The first part is one before the cut point and the second part is the elements in the rows after the cut point. When the two rows meet, these two parts act differently and generate two new rows of the offspring. Elements before the cut point are swapped with each other. Suppose the chosen cut point is 3. If, in the first row before the cut point, the elements are 1 2 3 and, in the second row, the elements before the cut point are 5 6 7, then in the two offspring rows, generated from them, the first three elements of the first offspring will be 5 6 7 and the first three elements of the second row will be 1 2 3.For the elements after the cut point, mating is done in a different fashion. We search the first parental row after the cut point in the first matrix for which the elements are common with the elements after the cut point of the first row in second matrix. Then, the order in which the common elements occur in the row of the second matrix is listed. Reposition the common elements in the order as they occur in the second matrix. For the elements, which are not common with the row of the second matrix, keep them in the same position as they are in the first parental row before the crossover. Thus, the elements after the first three elements of the offspring one are generated. Now, the entire row of the first offspring is formed. In the same manner, the elements of the second offspring are generated from the elements after the cut point of the second parental row and the elements after the cut point of the first parental row. Thus, two new offspring rows forming two existing parental rows are generated and we have two more options to choose the best among them. By applying the procedure on each row of the matrix, two new offspring matrices are generated, making four genes to choose the best

The Fitness Function :

The fitness function is used to rank the quality of a chromosome. A fitness value is assigned to a chromosome by a fitness function and a chromosome is evaluated with this value for survival. The fitness function, used in this problem, is as below

Fitness Function Value $= f_i = \sum_j w_{ij} / \sigma_i$

6.2 System Flow Diagram

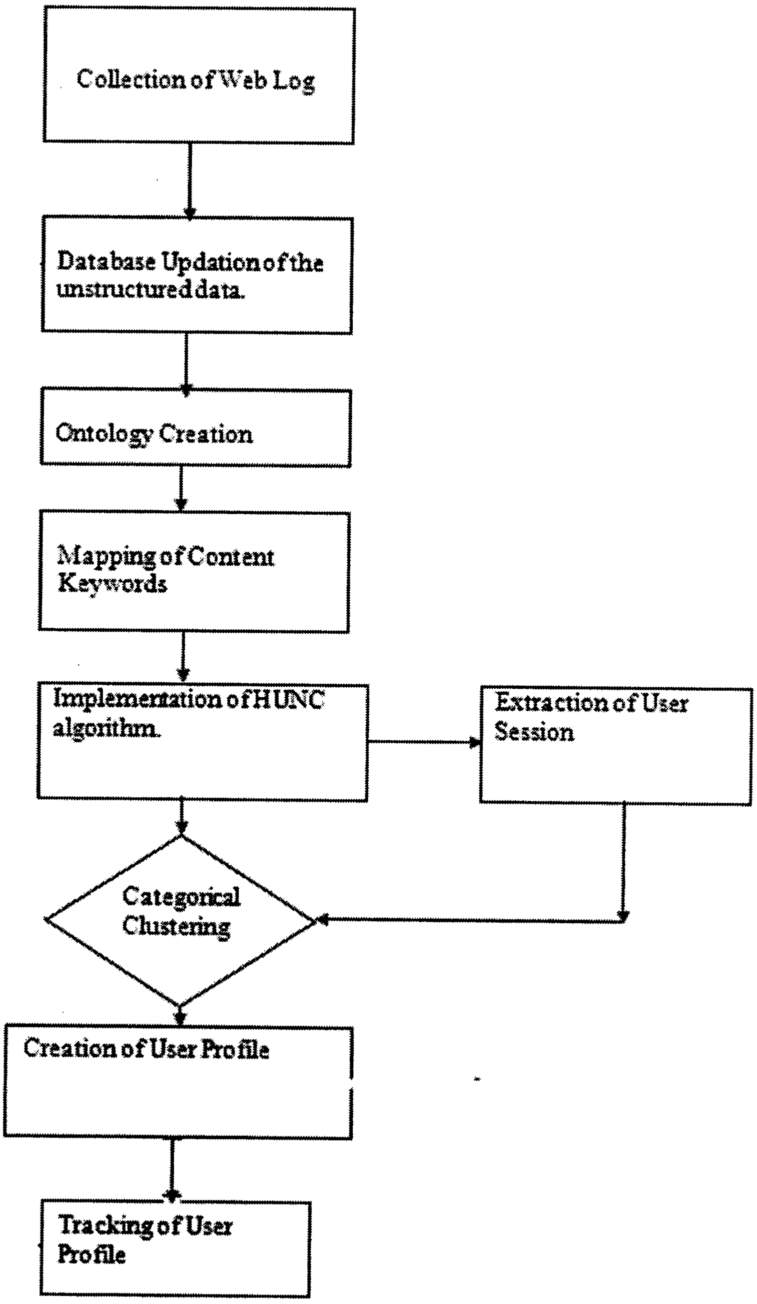

Collection of Web Log
Database Updation of the unstructured data.
Ontology Creation
Mapping of Content Keywords
Implementation of HUNC algorithm.
Extraction of User Session
Categorical Clustering
Creation of User Profile
Tracking of User Profile

Figure 6.1 System Flow Diagram

CHAPTER 7

7. CONCLUSION & FUTURE ENHANCEMENTS

7.1 Conclusion

The framework for mining, tracking, and validating evolving multifaceted user profiles on Web sites that have all the challenging aspects of real-life Web usage mining, including evolving user profiles and access patterns, dynamic Web pages, and external data describing an ontology of the Web content. A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries, and inquiring and inquired companies. The choice of the period length for analysis depends on the application or can be set, depending on the cross-period validation results. Thus the web mining and tracking will be very helpful to know about the customer's interests and the need. It also helps to personalize the web page and know how the interest of user changes from time to time.

7.2 FUTURE ENHANCEMENTS

Even though we did not focus on scalability, the latter can be addressed by following an approach similar to, where Web clickstreams are considered as an evolving data stream, or by mapping some new sessions to persistent profiles and updating these profiles, hence eliminating most sessions from further analysis and focusing the mining on truly new sessions.

CHAPTER 8

8 APPENDIX

8.1 Sample Code

Tracking Profile Code

```
Public Class frmTrackProfiles
  Dim objX As New clsConnection
  Private Sub btnView_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles btnView.Click
    Try
      TTPProces.Show("Processing...", btnView, 0, 0)
      findRec()
      updateBirth()
      updateDeath()
      UpdatePersistence()
      UpdateAtavism()
      Dim qryData As String = "select * from ProfEvol"
      frmGridDisplay.QryRecy = qryData
      frmGridDisplay.getRecords(qryData)
      frmGridDisplay.ShowDialog()
      TTPProces.Hide(btnView)
    Catch ex As Exception
      MessageBox.Show(ex.Message.ToString, "btnView.Click")
    End Try
  End Sub

  Sub findRec()
    Try
      objX.exeQuery("Truncate Table IP_Month")
```

```
Dim qryIp As String = "Select distinct(ClientIp) from tbl_ImportedData2"
Dim dsetIp As New DataSet
dsetIp = objX.getDataset(qryIp)
Dim qryMin As String = "Select min(date_time) from tbl_ImportedData2"
Dim dsetMin As New DataSet
dsetMin = objX.getDataset(qryMin)

Dim qryMax As String = "Select max(date_time) from tbl_ImportedData2"
Dim dsetMax As New DataSet
dsetMax = objX.getDataset(qryMax)
Dim minDate As Date = dsetMin.Tables(0).Rows(0).Item(0).ToString.Trim
minDate = minDate.AddMonths(-1)
Dim maxDate As Date = dsetMax.Tables(0).Rows(0).Item(0).ToString.Trim
Dim diff As Integer = 0
Dim cnt As Integer = 0
diff = DateDiff(DateInterval.Month, minDate, maxDate)
For cnt = 1 To diff
  Dim cutDate As String = ""
  Dim cutDateEnd As String = ""
  Dim cutMonth As String = ""
  Dim endDate As String = ""

  cutMonth = Mid(Format(minDate.AddMonths(cnt), "yyyy-MM-01"), 6, 2)
  If cutMonth = "01" Then
    endDate = "31"
  ElseIf cutMonth = "02" Then
    endDate = "28"
  ElseIf cutMonth = "03" Then
    endDate = "31"
  ElseIf cutMonth = "04" Then
    endDate = "30"
  ElseIf cutMonth = "05" Then
    endDate = "31"
```

```
ElseIf cutMonth = "06" Then
    endDate = "30"
ElseIf cutMonth = "07" Then
    endDate = "31"
ElseIf cutMonth = "08" Then
    endDate = "31"
ElseIf cutMonth = "09" Then
    endDate = "30"
ElseIf cutMonth = "10" Then
    endDate = "31"
ElseIf cutMonth = "11" Then
    endDate = "30"
ElseIf cutMonth = "12" Then
    endDate = "31"
End If

Dim cutTemp As String = ""
cutDate = Mid(Format(minDate.AddMonths(cnt), "yyyy-MM-01"), 1, 10)
cutTemp = Mid(cutDate, 1, 7)
cutDateEnd = Mid(Format(minDate.AddMonths(cnt), "yyyy-MM-" &
        endDate), 1, 10)

Dim qryPres As String = ""
qryPres = "Select distinct(ClientIp) from tbl_ImportedData2 where
date_Time >= '" & cutDate & "' and Date_Time < '" & cutDateEnd & "'"
Dim dsetPres As New DataSet
dsetPres = objX.getDataset(qryPres)


Dim cntIns As Integer = 0
For cntIns = 0 To dsetPres.Tables(0).Rows.Count - 1
    Dim qryIns As String = ""
    qryIns = "Insert into IP_Month Values ('" &
      dsetPres.Tables(0).Rows(cntIns).Item(0).ToString.Trim & "' , '" &
cutTemp & "')"
    objX.exeQuery(qryIns)
  Next
Next
Dim qryMonths As String = ""
Dim dsetMonths As New DataSet
Dim rowcnt As Integer = 0
Dim qryCreate As String = ""
objX.exeQuery("drop table ProfEvol")

qryMonths = "Select distinct(PresMonth) from IP_Month order by
PresMonth"
dsetMonths = objX.getDataset(qryMonths)
qryCreate = "Create Table ProfEvol ( IpAdd Varchar(20) Not null, "

For rowcnt = 0 To dsetMonths.Tables(0).Rows.Count - 1
  qryCreate = qryCreate & "Prof_"
  qryCreate = qryCreate &
Mid(dsetMonths.Tables(0).Rows(rowcnt).Item(0).ToString.Trim, 1, 4) & "_"
  qryCreate = qryCreate &
Mid(dsetMonths.Tables(0).Rows(rowcnt).Item(0).ToString.Trim, 6, 2) & "
Varchar(20)  , "
Next
qryCreate = Mid(qryCreate, 1, Len(qryCreate) - 2)
qryCreate = qryCreate & ")"
objX.exeQuery(qryCreate)


objX.exeQuery("Truncate table ProfEvol")
Dim cntIp As Integer = 0
For cntIp = 0 To dsetIp.Tables(0).Rows.Count - 1
  Dim qryIns As String = ""
```

```
            qryIns = "Insert into ProfEvol (IpAdd) values ('" &
dsetIp.Tables(0).Rows(cntIp).Item(0).ToString.Trim & "')"
            objX.exeQuery(qryIns)
          Next
        Catch ex As Exception
          MessageBox.Show(ex.Message.ToString, "findRec()")
        End Try
      End Sub

      Sub updateBirth()
        Try
          Dim qryMonYr As String = ""
          Dim dsetMonYr As New DataSet
          qryMonYr = "Select distinct(PresMonth) from IP_Month order by PresMonth"
          dsetMonYr = objX.getDataset(qryMonYr)

          Dim ipAddr As String = ""
          Dim dsetIpAddr As New DataSet
          ipAddr = "select ClientIp  from IP_month"
          dsetIpAddr = objX.getDataset(ipAddr)

          Dim qryBirth As String = ""
          Dim dsetBirth As New DataSet
          Dim rowCnt As Integer = 0
          qryBirth = "Select ClientIp , Min(PresMonth) from IP_Month group by
          ClientIp"
          dsetBirth = objX.getDataset(qryBirth)

          For rowCnt = 0 To dsetBirth.Tables(0).Rows.Count - 1
            Dim varMonColumn As String = ""
            Dim qryUpdate As String = ""
            varMonColumn = "Prof_" &
          Mid(dsetBirth.Tables(0).Rows(rowCnt).Item(1).ToString.Trim, 1, 4) & "_" &
      Mid(dsetBirth.Tables(0).Rows(rowCnt).Item(1).ToString.Trim, 6, 2)
        qryUpdate = "Update ProfEvol set " & varMonColumn & " = 'Birth' where
      IpAdd = '" & dsetBirth.Tables(0).Rows(rowCnt).Item(0).ToString.Trim & "'"
        objX.exeQuery(q; yUpdate)
      Next
    Catch ex As Exception
      MessageBox.Show(ex.Message.ToString, "updateBirth()")
    End Try
  End Sub

  Sub updateDeath()
    Try
      Dim qryMonYr As String = ""
      Dim dsetMonYr As New DataSet
      qryMonYr = "Select distinct(PresMonth) from IP_Month order by PresMonth"
      dsetMonYr = objX.getDataset(qryMonYr)

      Dim ipAddr As String = ""
      Dim dsetIpAddr As New DataSet
      ipAddr = "select ClientIp  from IP_month"
      dsetIpAddr = objX.getDataset(ipAddr)

      Dim qryBirth As String = ""
      Dim dsetBirth As New DataSet
      Dim rowCnt As Integer = 0
      qryBirth = "Select ClientIp , Max(PresMonth) from IP_Month group by
      ClientIp"
      dsetBirth = objX.getDataset(qryBirth)

      For rowCnt = 0 To dsetBirth.Tables(0).Rows.Count - 1
        Dim varMonColumn As String = ""
        Dim qryUpdate As String = ""
        Dim prev As String = ""
```

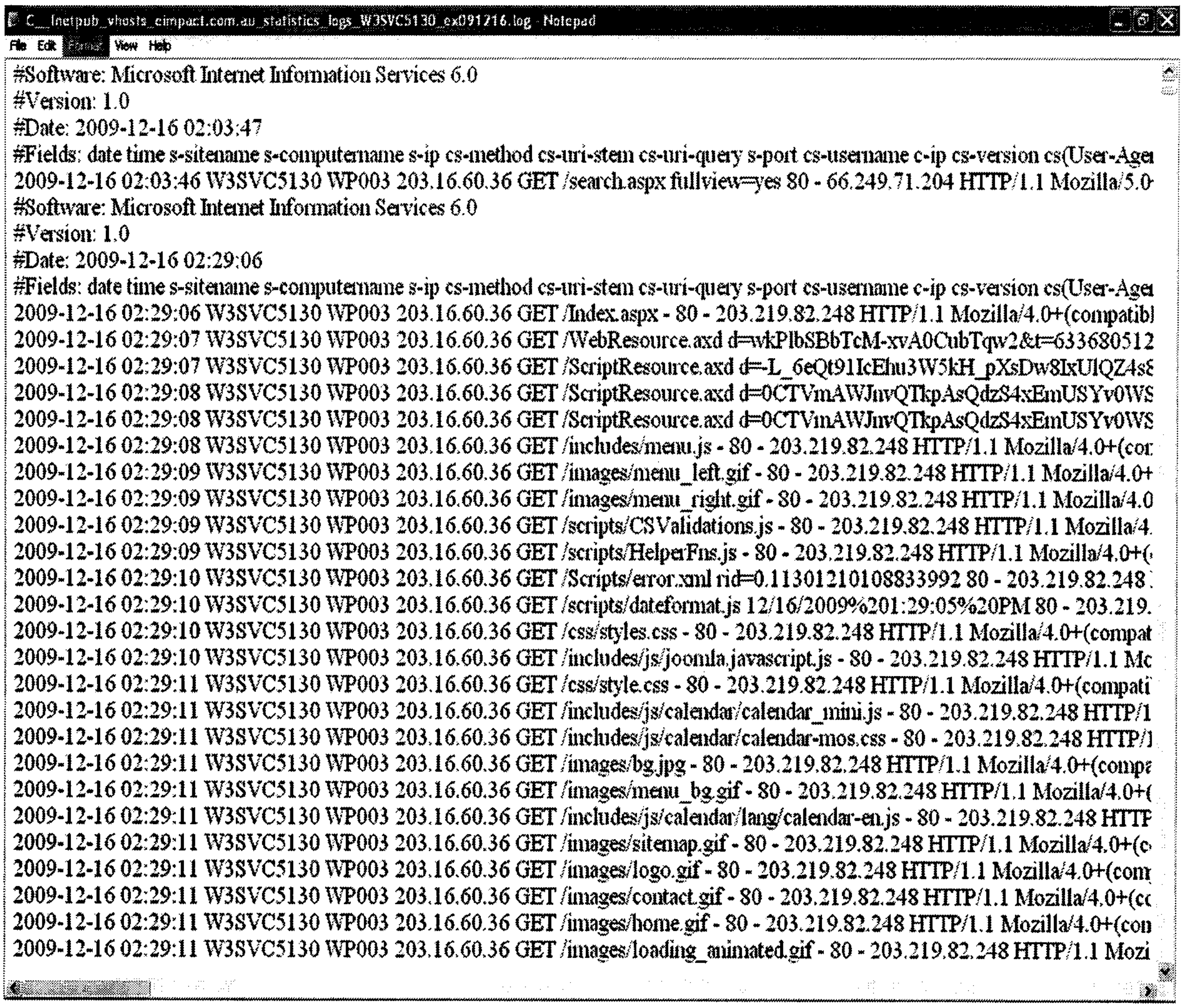
```
            Dim curSt As String = ""
            varMonColumn = "Prof_" &
        Mid(dsetBirth.Tables(0).Rows(rowCnt).Item(1).ToString.Trim, 1, 4) & "_" &
        Mid(dsetBirth.Tables(0).Rows(rowCnt).Item(1).ToString.Trim, 6, 2)
            prev = checkBirth(varMonColumn,
        dsetBirth.Tables(0).Rows(rowCnt).Item(0).ToString.Trim)
            If prev.Trim = "" Then
                curSt = "Death"
            Else
                curSt = "Birth/Death"
            End If
            qryUpdate = "Update ProfEvol set " & varMonColumn & " = '" & curSt &
        "' where IpAdd = '" &
        dsetBirth.Tables(0).Rows(rowCnt).Item(0).ToString.Trim & "'"
            'MsgBox(qryUpdate)
            objX.exeQuery(qryUpdate)
        Next
    Catch ex As Exception
        MessageBox.Show(ex.Message.ToString, "updateDeath()")
    End Try
End Sub

Function checkBirth(ByVal Column As String, ByVal Ip As String) As String
    Dim status As String = ""
    Try
        Dim qryStatus As String = ""
        Dim dsetStatus As New DataSet
        qryStatus = "select " & Column.Trim & " from ProfEvol where IpAdd = '" &
        Ip.Trim & "'"
        dsetStatus = objX.getDataset(qryStatus)
        status = dsetStatus.Tables(0).Rows(0).Item(0).ToString.Trim
    Catch ex As Exception
        MessageBox.Show(ex.Message.ToString, "checkBirth()")
        End Try
        Return status
    End Function
     Private Sub fr nTrackProfiles_KeyDown(ByVal sender As Object, ByVal e As
            System.Windows.Forms.KeyEventArgs) Handles Me.KeyDown
        Try
            If e.KeyCode = Keys.Escape Then
                Me.Close()
            End If

            If e.KeyCode = Keys.Enter Then
                SendKeys.Send("{TAB}")
            End If
        Catch ex As Exception
            MessageBox.Show(ex.Message.ToString, "Me.KeyDown")
        End Try

    End Sub

    Private Sub frmTrackProfiles_Load(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles MyBase.Load
        Try

        Catch ex As Exception
            MessageBox.Show(ex.Message.ToString, "frmTrackProfiles_Load")
        End Try
    End Sub

    Sub UpdatePersistence()
        Try

            Dim qryMonth As String = ""
            Dim dsetMonth As New DataSet
```

```
qryMonth = "select distinct(PresMonth) from IP_Month order by PresMonth"
dsetMonth = objX.getDataset(qryMonth)

Dim qryCol As String = ""
Dim dsetCol As New DataSet
Dim cnt As Integer = 0

qryCol = "Select Column_Name from information_schema.Columns where
table_name = 'ProfEvol'  and Column_Name not in ('IPAdd')"
dsetCol = objX.getDataset(qryCol)
Dim qryIp As String = "Select IpAdd from ProfEvol where IpAdd not in
(select IpAdd from ProfEvol where " &
dsetCol.Tables(0).Rows(0).Item(0).ToString.Trim & " = 'Birth/Death')"
Dim dsetIp As New DataSet
Dim rowCnt As Integer = 0
dsetIp = objX.getDataset(qryIp)

For rowCnt = 1 To dsetMonth.Tables(0).Rows.Count - 2
   Dim qryChk As String = ""
   Dim inRow As Integer = 0

   For inRow = 0 To dsetIp.Tables(0).Rows.Count - 1
      qryChk = "Select Count(ClientIp) from IP_Month where PresMonth = '"
& dsetMonth.Tables(0).Rows(rowCnt).Item(0).ToString.Trim & "' and
ClientIp = '" & dsetIp.Tables(0).Rows(inRow).Item(0).ToString.Trim & "'"
      Dim dsetChk As New DataSet
      dsetChk = objX.getDataset(qryChk)

      If dsetChk.Tables(0).Rows(0).Item(0).ToString.Trim = 1 Then
         Dim varMonth As String = ""
         Dim qryUpd As String = ""
         varMonth = "Prof_" &
Mid(dsetMonth.Tables(0).Rows(rowCnt).Item(0).ToString.Trim, 1, 4) & "_"
```

```
        & Mid(dsetMonth.Tables(0).Rows(rowCnt).Item(0).ToString.Trim, 6, 2)
              qryUpd = "Update ProfEvol set " & varMonth & " = 'Persistence'
        where IpAdd = '" & dsetIp.Tables(0).Rows(inRow).Item(0).ToString.Trim &
        "'"
              objX.exeQuery(qryUpd)
            End If
          Next
        Next
      Catch ex As Exception
        MessageBox.Show(ex.Message.ToString, " UpdatePersistence()")
      End Try
    End Sub

    Sub UpdateAtavism()
      Try
        Dim qryCol As String = ""
        Dim dsetCol As New DataSet
        Dim cnt As Integer = 0
        qryCol = "Select Column_Name from information_schema.Columns where
        table_name = 'ProfEvol'  and Column_Name not in ('IPAdd')"
        dsetCol = objX.getDataset(qryCol)

        Dim qryIp As String = "Select IpAdd from ProfEvol where IpAdd not in
        (select IpAdd from ProfEvol where " &
  dsetCol.Tables(0).Rows(0).Item(0).ToString.Trim & " = 'Birth/Death')"
        Dim dsetIp As New DataSet
        dsetIp = objX.getDataset(qryIp)
        Dim rowCnt As Integer = 0
        Dim fstCol As String = ""
        Dim secCol As String = ""
        Dim qrySel As String = ""
        Dim dsetSel As DataSet
      Dim rCol As Integer = 0

      For rowCnt = 0 To dsetIp.Tables(0).Rows.Count - 1
        For rCol = 1 To dsetCol.Tables(0).Rows.Count - 2
          fstCol = ""
          secCol = ""
          qrySel = ""
          fstCol = dsetCol.Tables(0).Rows(rCol).Item(0).ToString.Trim
          secCol = dsetCol.Tables(0).Rows(rCol + 1).Item(0).ToString.Trim
          qrySel = "Select " & fstCol & " , " & secCol & " from ProfEvol where
      IpAdd = '" & dsetIp.Tables(0).Rows(rowCnt).Item(0).ToString.Trim & "'"
          dsetSel = New DataSet
          dsetSel = objX.getDataset(qrySel)
          If dsetSel.Tables(0).Rows(0).Item(0).ToString.Trim = "" And
      dsetSel.Tables(0).Rows(0).Item(1).ToString.Trim = "Persistence" Then
            Dim qryUpd As String = ""
            qryUpd = "Update ProfEvol set " & secCol & " = 'Atavism' where
      IpAdd = '" & dsetIp.Tables(0).Rows(rowCnt).Item(0).ToString.Trim & "'"
            objX.exeQuery(qryUpd)
          End If
        Next
      Next

    Catch ex As Exception
      MessageBox.Show(ex.Message.ToString, " UpdateAtavism()")
    End Try
  End Sub
End Class
```

8.2 Screenshots

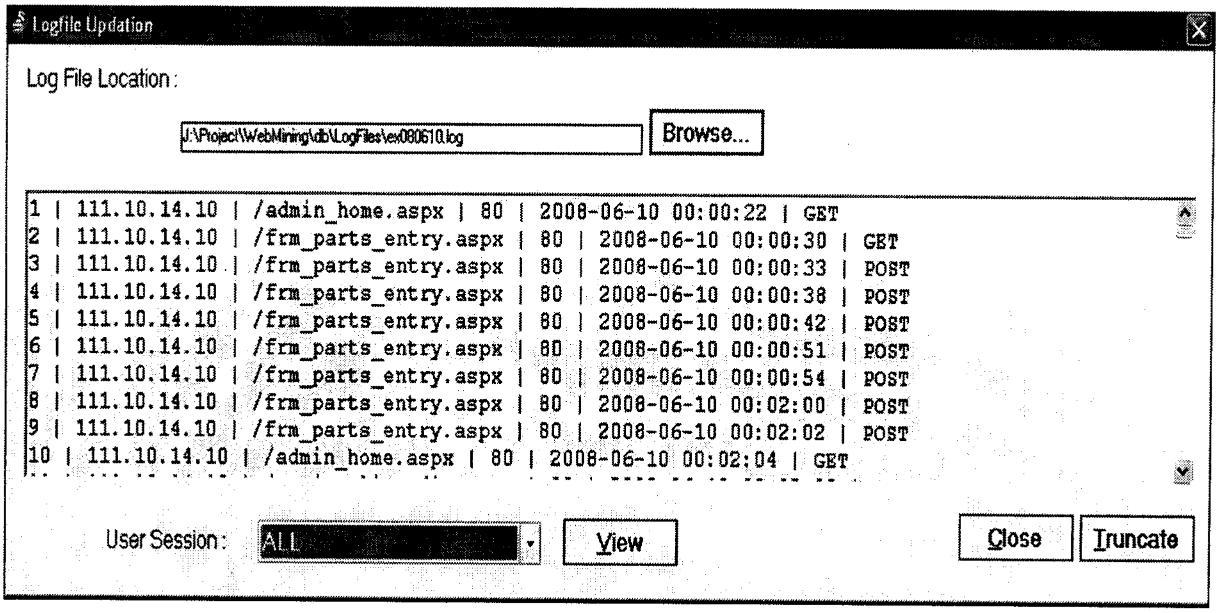
Sample Web Log File


C__Inetpub_vhosts_eimpact.com.au_statistics_logs_W3SVC5130_ex091216.log - Notepad
File
Edit
Format
View
Help

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2009-12-16 02:03:47
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Age
2009-12-16 02:03:46 W3SVC5130 WP003 203.16.60.36 GET /search.aspx fullview=yes 80 - 66.249.71.204 HTTP/1.1 Mozilla/5.0
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2009-12-16 02:29:06
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Age
2009-12-16 02:29:06 W3SVC5130 WP003 203.16.60.36 GET /Index.aspx - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(compatibl
2009-12-16 02:29:07 W3SVC5130 WP003 203.16.60.36 GET /WebResource.axd d=wkPlbSBbTcM-xvA0CubTqw2&t=633680512
2009-12-16 02:29:07 W3SVC5130 WP003 203.16.60.36 GET /ScriptResource.axd d=-L_6eQt91IcEhu3W5kH_pXsDw8IxUlQZ4s
2009-12-16 02:29:08 W3SVC5130 WP003 203.16.60.36 GET /ScriptResource.axd d=0CTVmAWJnvQTkpAsQdzS4xEmUSYv0WS
2009-12-16 02:29:08 W3SVC5130 WP003 203.16.60.36 GET /ScriptResource.axd d=0CTVmAWJnvQTkpAsQdzS4xEmUSYv0WS
2009-12-16 02:29:08 W3SVC5130 WP003 203.16.60.36 GET /includes/menu.js - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(co
2009-12-16 02:29:09 W3SVC5130 WP003 203.16.60.36 GET /images/menu_left.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+
2009-12-16 02:29:09 W3SVC5130 WP003 203.16.60.36 GET /images/menu_right.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0
2009-12-16 02:29:09 W3SVC5130 WP003 203.16.60.36 GET /scripts/CSValidations.js - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.
2009-12-16 02:29:09 W3SVC5130 WP003 203.16.60.36 GET /scripts/HelperFns.js - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(
2009-12-16 02:29:10 W3SVC5130 WP003 203.16.60.36 GET /Scripts/error.xml rid=0.11301210108833992 80 - 203.219.82.248
2009-12-16 02:29:10 W3SVC5130 WP003 203.16.60.36 GET /scripts/dateformat.js 12/16/2009%201:29:05%20PM 80 - 203.219.
2009-12-16 02:29:10 W3SVC5130 WP003 203.16.60.36 GET /css/styles.css - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(compat
2009-12-16 02:29:10 W3SVC5130 WP003 203.16.60.36 GET /includes/js/joomla.javascript.js - 80 - 203.219.82.248 HTTP/1.1 Mo
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /css/style.css - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(compati
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /includes/js/calendar/calendar_mini.js - 80 - 203.219.82.248 HTTP/1
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /includes/js/calendar/calendar-mos.css - 80 - 203.219.82.248 HTTP/1
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/bg.jpg - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(comp
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/menu_bg.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /includes/js/calendar/lang/calendar-en.js - 80 - 203.219.82.248 HTTP
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/sitemap.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(c
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/logo.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(com
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/contact.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(c
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/home.gif - 80 - 203.219.82.248 HTTP/1.1 Mozilla/4.0+(con
2009-12-16 02:29:11 W3SVC5130 WP003 203.16.60.36 GET /images/loading_animated.gif - 80 - 203.219.82.248 HTTP/1.1 Mozi
```

The web Log File is the input file. This log file Contain the information about the date of log file created, site name, computer name, server ip address, client ip address, url address, server to client bytes send , client to server byte send, etc..... It contains all the clicks made by the user in the website. Above is the real web Log file of the site www.eimpact.com.au.

Collection of Web Log data and Database Updation

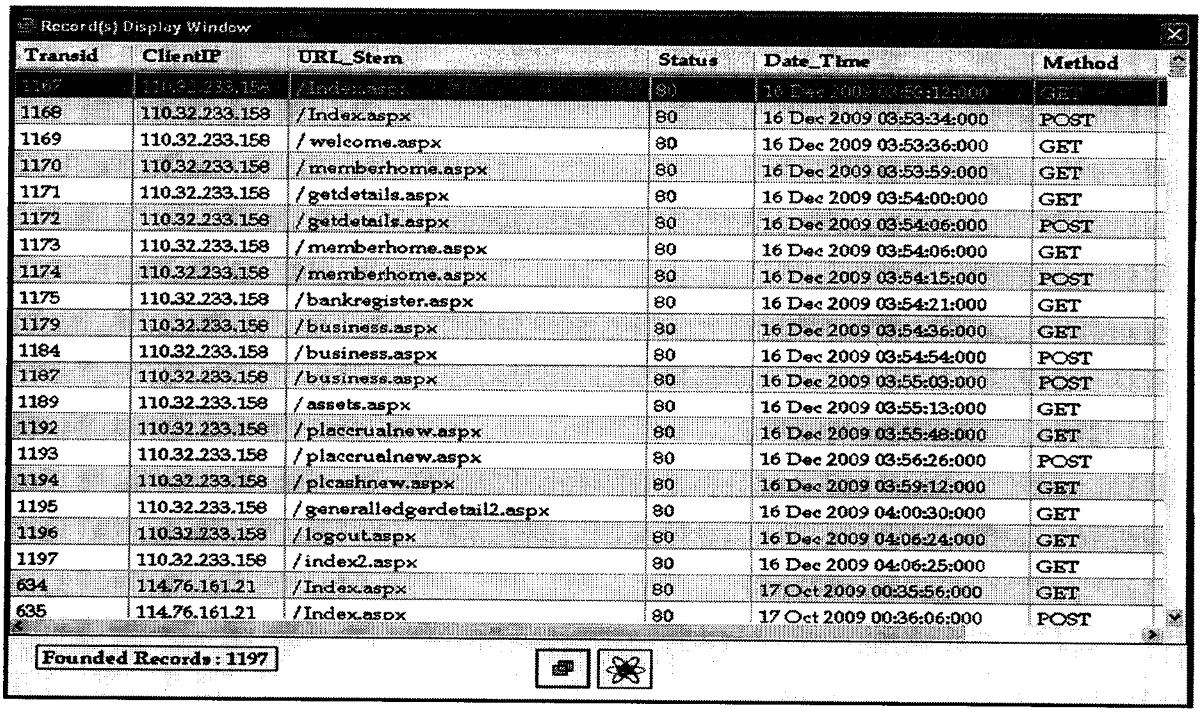


J:\Project\WebMining\db\LogFiles\ex080610.log

User Session : ALL

```
1 | 111.10.14.10 | /admin_home.aspx | 80 | 2008-06-10 00:00:22 | GET
2 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:00:30 | GET
3 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:00:33 | POST
4 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:00:38 | POST
5 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:00:42 | POST
6 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:00:51 | POST
7 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:00:54 | POST
8 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:02:00 | POST
9 | 111.10.14.10 | /frm_parts_entry.aspx | 80 | 2008-06-10 00:02:02 | POST
10 | 111.10.14.10 | /admin_home.aspx | 80 | 2008-06-10 00:02:04 | GET
```

The above log files are been updated into the database. Each line in the log file are been scanned and the required data which are essential to web mining are been moved to the database.

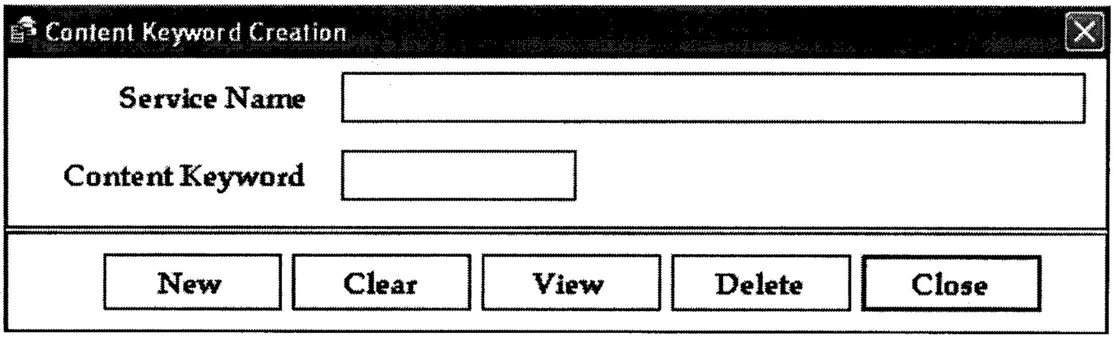


Transid	ClientIP	URL_Stem	Status	Date_Time	Method
1167	110.32.233.158	/Index.aspx	80	16 Dec 2009 03:53:12:000	GET
1168	110.32.233.158	/Index.aspx	80	16 Dec 2009 03:53:34:000	POST
1169	110.32.233.158	/welcome.aspx	80	16 Dec 2009 03:53:36:000	GET
1170	110.32.233.158	/memberhome.aspx	80	16 Dec 2009 03:53:59:000	GET
1171	110.32.233.158	/getdetails.aspx	80	16 Dec 2009 03:54:00:000	GET
1172	110.32.233.158	/getdetails.aspx	80	16 Dec 2009 03:54:06:000	POST
1173	110.32.233.158	/memberhome.aspx	80	16 Dec 2009 03:54:06:000	GET
1174	110.32.233.158	/memberhome.aspx	80	16 Dec 2009 03:54:15:000	POST
1175	110.32.233.158	/bankregister.aspx	80	16 Dec 2009 03:54:21:000	GET
1179	110.32.233.158	/business.aspx	80	16 Dec 2009 03:54:36:000	GET
1184	110.32.233.158	/business.aspx	80	16 Dec 2009 03:54:54:000	POST
1187	110.32.233.158	/business.aspx	80	16 Dec 2009 03:55:03:000	POST
1189	110.32.233.158	/assets.aspx	80	16 Dec 2009 03:55:13:000	GET
1192	110.32.233.158	/placcrualnew.aspx	80	16 Dec 2009 03:55:48:000	GET
1193	110.32.233.158	/placcrualnew.aspx	80	16 Dec 2009 03:56:26:000	POST
1194	110.32.233.158	/plcashnew.aspx	80	16 Dec 2009 03:59:12:000	GET
1195	110.32.233.158	/generalledgerdetail2.aspx	80	16 Dec 2009 04:00:30:000	GET
1196	110.32.233.158	/logout.aspx	80	16 Dec 2009 04:06:24:000	GET
1197	110.32.233.158	/index2.aspx	80	16 Dec 2009 04:06:25:000	GET
634	114.76.161.21	/Index.aspx	80	17 Oct 2009 00:35:56:000	GET
635	114.76.161.21	/Index.aspx	80	17 Oct 2009 00:36:06:000	POST

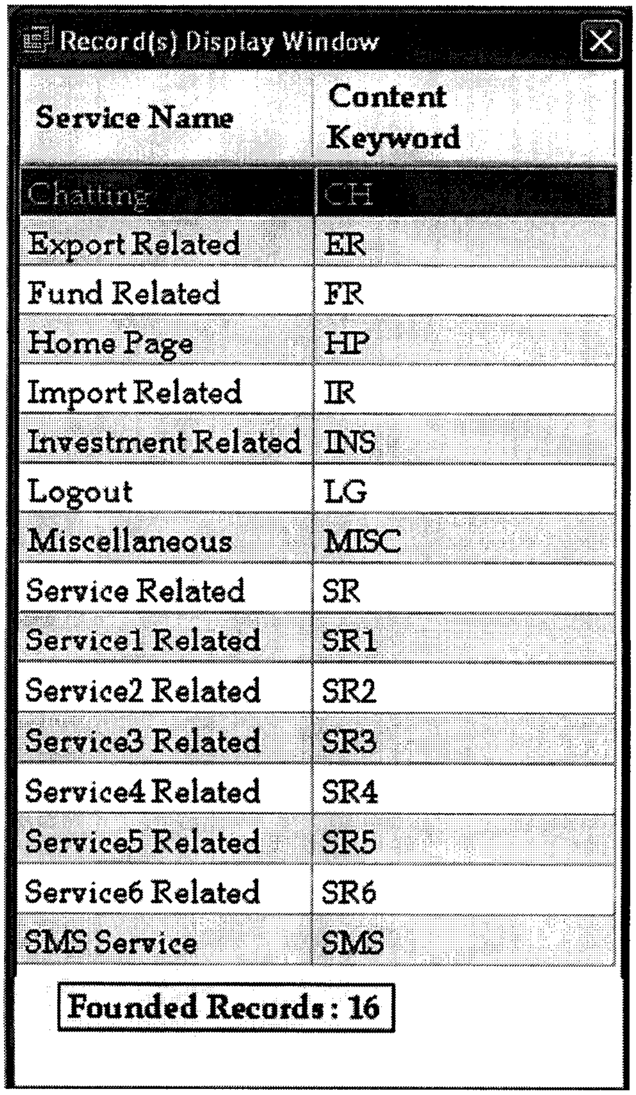
Founded Records : 1197

The updated data of the log file which are saved in the database are shown above.

Adding Semantics with Ontology Concepts



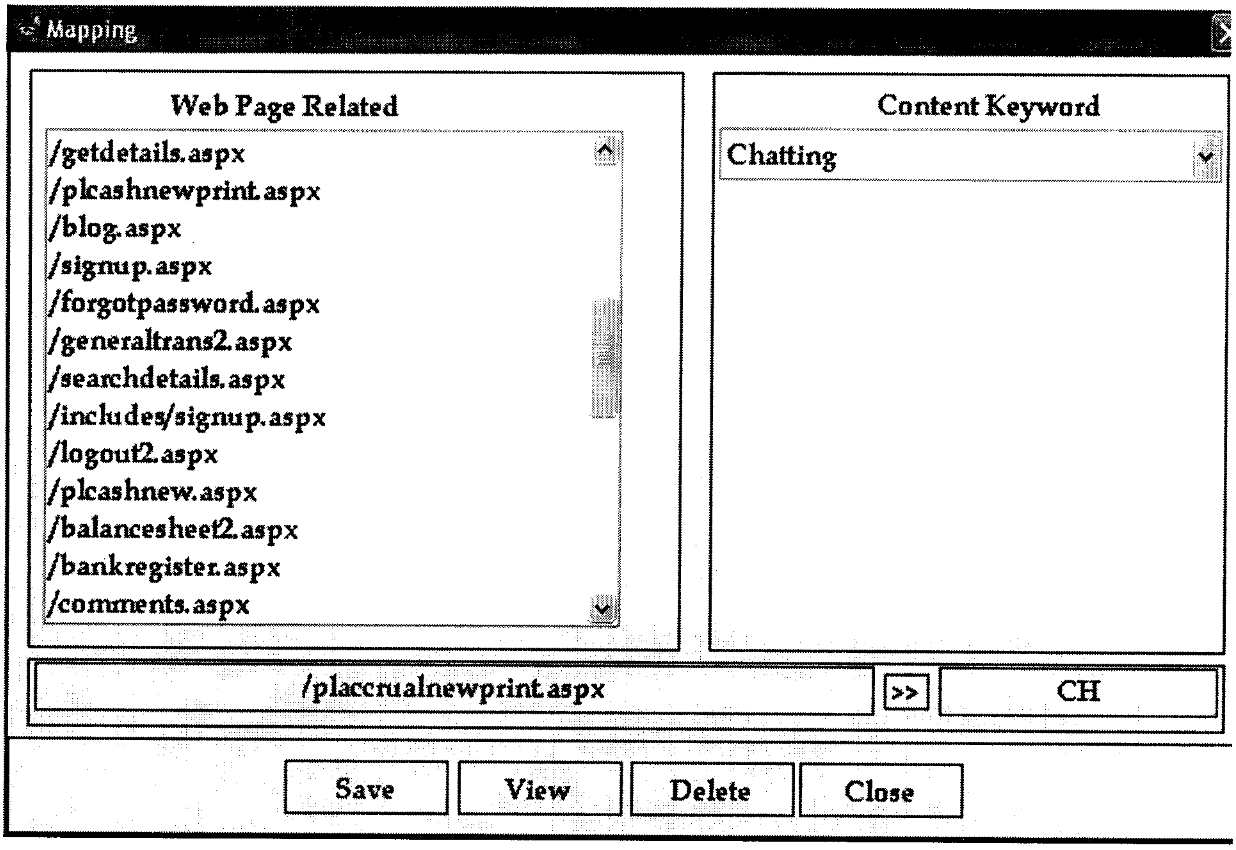
The content key creation is creating the service name and the content keyword for that service name. These content keys are been created to map to the urls.



Service Name	Content Keyword
Chatting	CH
Export Related	ER
Fund Related	FR
Home Page	HP
Import Related	IR
Investment Related	INS
Logout	LG
Miscellaneous	MISC
Service Related	SR
Service1 Related	SR1
Service2 Related	SR2
Service3 Related	SR3
Service4 Related	SR4
Service5 Related	SR5
Service6 Related	SR6
SMS Service	SMS

Founded Records : 16

Mapping of Content Keywords



/getdetails.aspx
/plcashnewprint.aspx
/blog.aspx
/signup.aspx
/forgotpassword.aspx
/generaltrans2.aspx
/searchdetails.aspx
/includes/signup.aspx
/logout2.aspx
/plcashnew.aspx
/balancesheet2.aspx
/bankregister.aspx
/comments.aspx

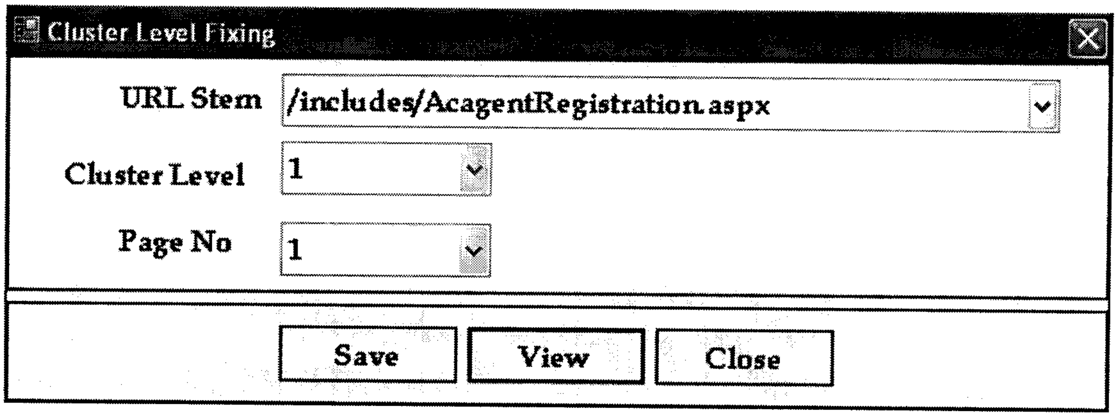
Chatting

/placcrualnewprint.aspx

CH

The content key which are been created are been mapped with the URL's which are been in our website or the web portals. This is an ontology concept, so that this is easy to the user to identify the URL that user have clicked.

Cluster Level Fixing



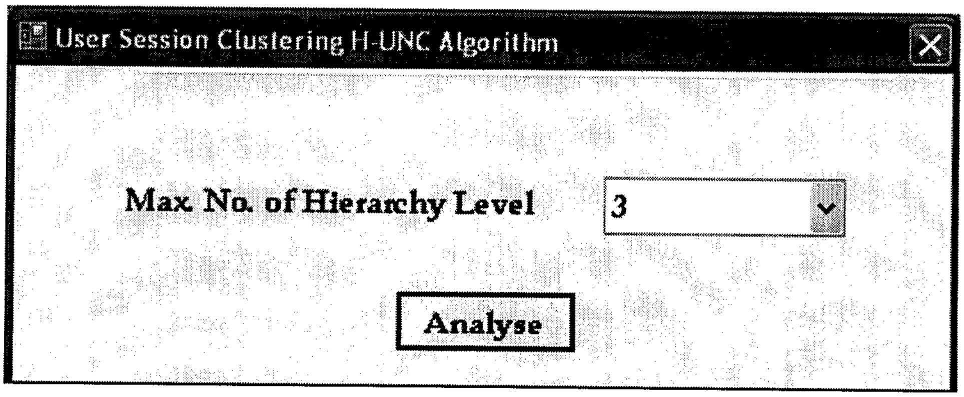
URL Stem /includes/AcagentRegistration.aspx

Cluster Level 1

Page No 1

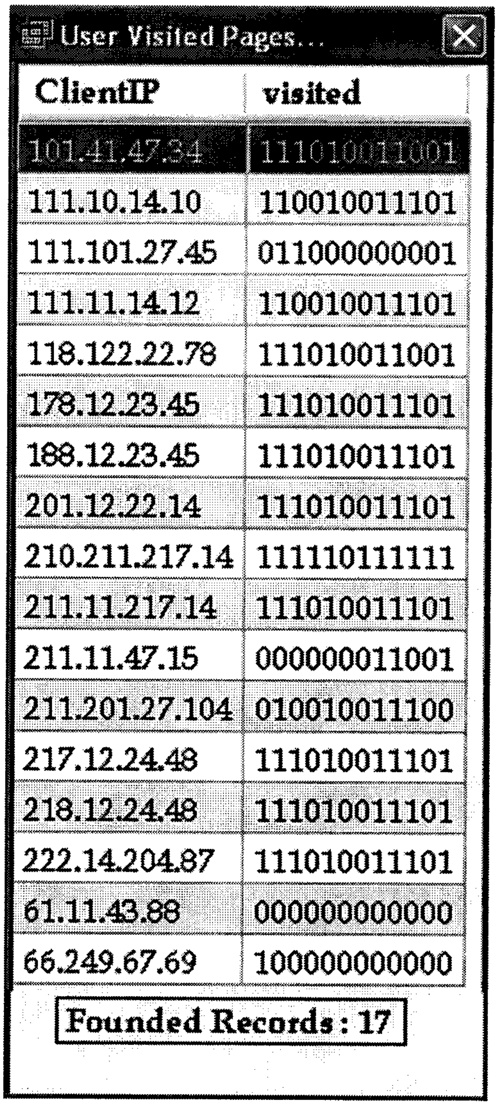
The cluster level fixing is been used to fix the cluster level and the unique page number of the URL'S which are been in the website or the web portal

Clustering of the user sessions by using Hierarchical Unsupervised Niche Clustering (H - UNC)



Max. No. of Hierarchy Level 3

The input for the H-unc is the Binary section vectors, It start with the minimum value 1 to the maximum number of hierarchy level specified. The output of the H-unc will be the user profiles.

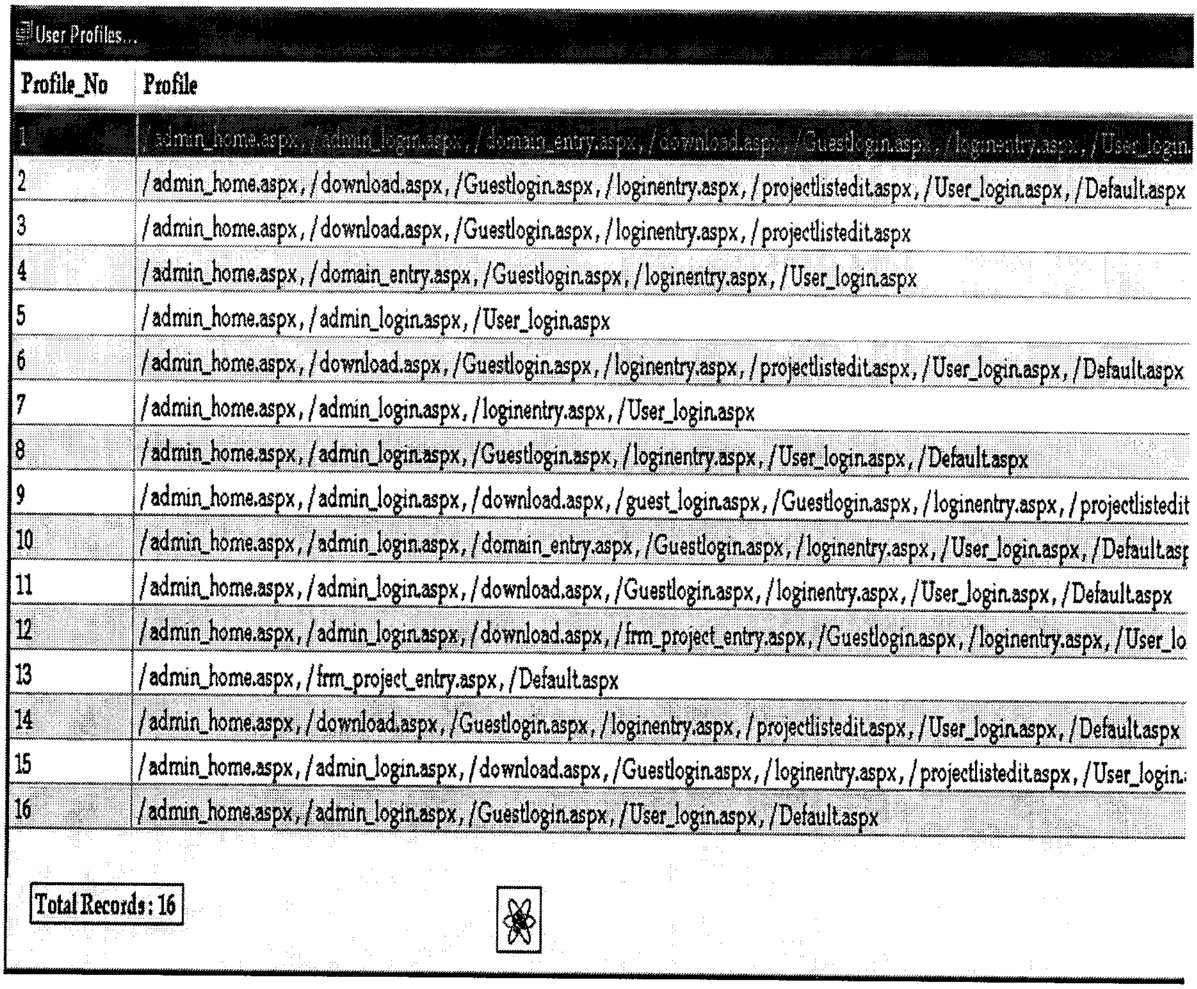


User Visited Pages...

ClientIP	visited
101.41.47.34	111010011001
111.10.14.10	110010011101
111.101.27.45	011000000001
111.11.14.12	110010011101
118.122.22.78	111010011001
178.12.23.45	111010011101
188.12.23.45	111010011101
201.12.22.14	111010011101
210.211.217.14	111110111111
211.11.217.14	111010011101
211.11.47.15	000000011001
211.201.27.104	010010011100
217.12.24.48	111010011101
218.12.24.48	111010011101
222.14.204.87	111010011101
61.11.43.88	000000000000
66.249.67.69	100000000000

Founded Records : 17

The above figure shows the client IP and the binary chromosome, the binary chromosome the 1 are visited page and 0 are non visited page by the particular client IP address.



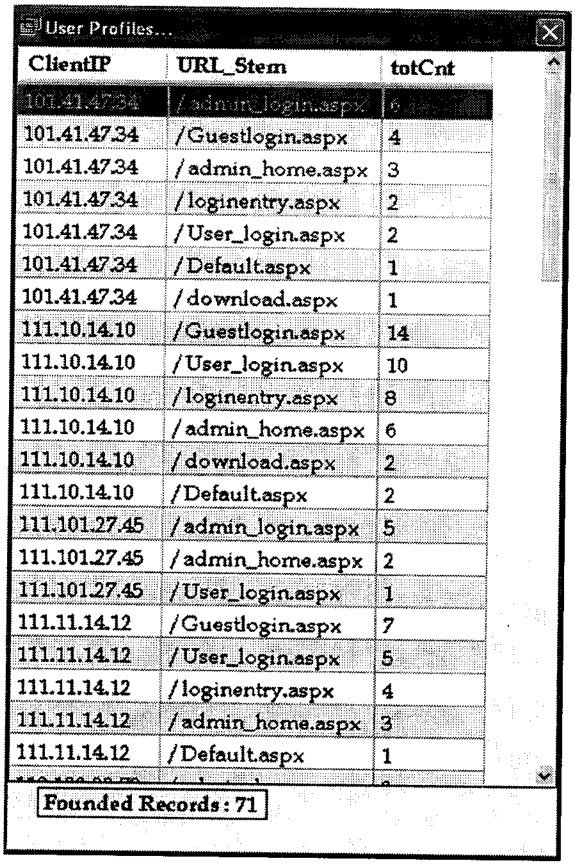
User Profiles...

Profile_No	Profile
1	/admin_home.aspx, /admin_login.aspx, /domain_entry.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /User_login.
2	/admin_home.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /projectlistedit.aspx, /User_login.aspx, /Default.aspx
3	/admin_home.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /projectlistedit.aspx
4	/admin_home.aspx, /domain_entry.aspx, /Guestlogin.aspx, /loginentry.aspx, /User_login.aspx
5	/admin_home.aspx, /admin_login.aspx, /User_login.aspx
6	/admin_home.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /projectlistedit.aspx, /User_login.aspx, /Default.aspx
7	/admin_home.aspx, /admin_login.aspx, /loginentry.aspx, /User_login.aspx
8	/admin_home.aspx, /admin_login.aspx, /Guestlogin.aspx, /loginentry.aspx, /User_login.aspx, /Default.aspx
9	/admin_home.aspx, /admin_login.aspx, /download.aspx, /guest_login.aspx, /Guestlogin.aspx, /loginentry.aspx, /projectlistedit
10	/admin_home.aspx, /admin_login.aspx, /domain_entry.aspx, /Guestlogin.aspx, /loginentry.aspx, /User_login.aspx, /Default.asp
11	/admin_home.aspx, /admin_login.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /User_login.aspx, /Default.aspx
12	/admin_home.aspx, /admin_login.aspx, /download.aspx, /frm_project_entry.aspx, /Guestlogin.aspx, /loginentry.aspx, /User_lo
13	/admin_home.aspx, /frm_project_entry.aspx, /Default.aspx
14	/admin_home.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /projectlistedit.aspx, /User_login.aspx, /Default.aspx
15	/admin_home.aspx, /admin_login.aspx, /download.aspx, /Guestlogin.aspx, /loginentry.aspx, /projectlistedit.aspx, /User_login.
16	/admin_home.aspx, /admin_login.aspx, /Guestlogin.aspx, /User_login.aspx, /Default.aspx

Total Records : 16

The above figure is the output of the H-Unc which gives the various user profiles which are been generated.

Summarize session clusters/categories into user profiles

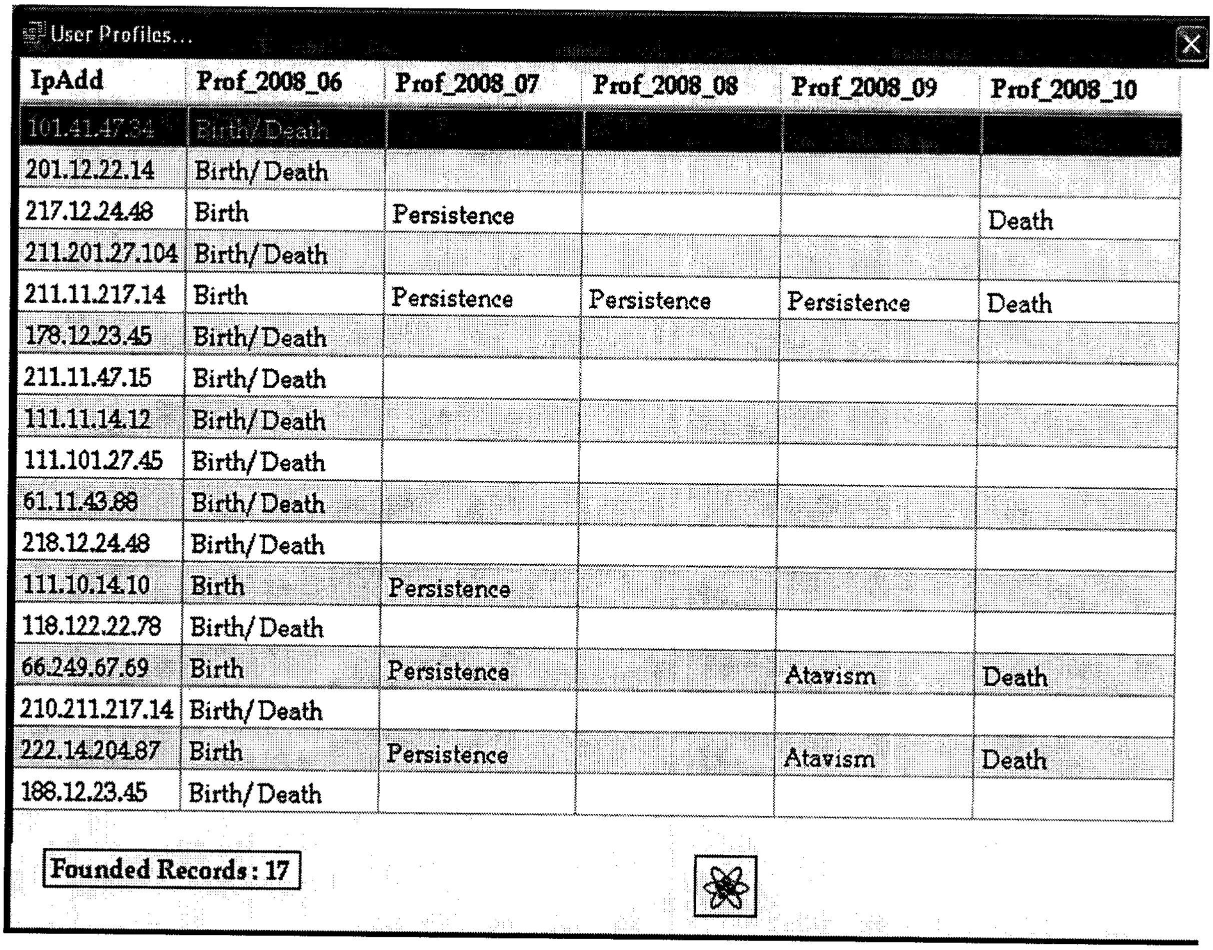


ClientIP	URL_Stem	totCnt
101.41.47.34	/admin_login.aspx	6
101.41.47.34	/Guestlogin.aspx	4
101.41.47.34	/admin_home.aspx	3
101.41.47.34	/loginentry.aspx	2
101.41.47.34	/User_login.aspx	2
101.41.47.34	/Default.aspx	1
101.41.47.34	/download.aspx	1
111.10.14.10	/Guestlogin.aspx	14
111.10.14.10	/User_login.aspx	10
111.10.14.10	/loginentry.aspx	8
111.10.14.10	/admin_home.aspx	6
111.10.14.10	/download.aspx	2
111.10.14.10	/Default.aspx	2
111.101.27.45	/admin_login.aspx	5
111.101.27.45	/admin_home.aspx	2
111.101.27.45	/User_login.aspx	1
111.11.14.12	/Guestlogin.aspx	7
111.11.14.12	/User_login.aspx	5
111.11.14.12	/loginentry.aspx	4
111.11.14.12	/admin_home.aspx	3
111.11.14.12	/Default.aspx	1

Founded Records : 71

The summarization of the user profile which gives the detail summarization of the each Page visited by the each client IP address and the total time of visit to the particular page.

Tracking current profiles against existing profiles



IpAdd	Prof_2008_06	Prof_2008_07	Prof_2008_08	Prof_2008_09	Prof_2008_10
101.41.47.34	Birth/Death				
201.12.22.14	Birth/Death				
217.12.24.48	Birth	Persistence			Death
211.201.27.104	Birth/Death				
211.11.217.14	Birth	Persistence	Persistence	Persistence	Death
178.12.23.45	Birth/Death				
211.11.47.15	Birth/Death				
111.11.14.12	Birth/Death				
111.101.27.45	Birth/Death				
61.11.43.88	Birth/Death				
218.12.24.48	Birth/Death				
111.10.14.10	Birth	Persistence			
118.122.22.78	Birth/Death				
66.249.67.69	Birth	Persistence		Atavism	Death
210.211.217.14	Birth/Death				
222.14.204.87	Birth	Persistence		Atavism	Death
188.12.23.45	Birth/Death				

Founded Records : 17

The user profiles are been tracked and the user are been classified into four different types they are

- Birth – Newly arrived IP address

- Persistence – The IP address with are visited in the regular interval of time.

- Atavism – The IP address with are visited In the Irregular interval of time.

- Death – The IP address which never visit again.

CHAPTER 9

9. REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web,” Proc. Ninth IEEE Int’l Conf. Tools with AI (ICTAI ’97), pp. 558-567, 1997.

[2] O. Nasraoui, R. Krishnapuram, and A. Joshi, “Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator,” Proc. Eighth Int’l World Wide Web Conf. (WWW ’99), pp. 40-41, 1999.

[3] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, “Extracting

Web User Profiles Using Relational Competitive Fuzzy Clustering,” Int’l J. Artificial Intelligence Tools, vol. 9, no. 4, pp. 509-526, 2000.

[4] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” SIGKDD Explorations, vol. 1, no. 2, pp. 1-12, Jan. 2000.

[5] M. Spiliopoulou and L.C. Faulstich, “WUM: A Web Utilization Miner,” Proc. First Int’l Workshop Web and Databases (WebDB ’98), 1998.

[6] T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, “From User Access Patterns to Dynamic Hypertext Linking,” Proc. Fifth Int’l World Wide Web Conf. (WWW ’96), 1996.3.

[7] Gary cornell & Jonathan Morrison, “ Programming VB.NET “ Pares publishers

[8] Matt.j.Crounch, “ VB.NET programming“ Pearson Education, 2003 Edition.

[9] Steven volzner , "VB.NET programming Languages"

[10] Complete Database Programming using Microsoft SQL SERVER 2000 By Patrick Mitchell - Pearson Education, 2003 Edition.

[11] Elias.M.Award, " System Analysis and Design " Galgotia Publication Pvt.Ltd.1991

[12] Roger.S.Pressman, Software Engineering McGraw-Hill International Editions, 1991

[13] Gary cornell & Jonathan Morrison, " Programming VB.NET" Pares publishers

[14] Matt.j.Crounch, " VB.NET programming" Pearson Education, 2003 Edition.