ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

DISCOVERING FREQUENT NAVIGATION PATTERNS FOR CONSTRUCTING USER PROFILE: THE CASE OF EBIZ ONLINE SOLUTIONS PLC OFFICIAL WEBSITE

SEPTEMBER
2015

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

DISCOVERING FREQUENT NAVIGATION PATTERNS FOR CONSTRUCTING USER PROFILE: THE CASE OF EBIZ ONLINE SOLUTIONS PLC OFFICIAL WEBSITE

A Thesis Submitted to the College of Natural Science of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

ANTENEH LEGESSE TEKLU

SEPTEMBER
2015

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

DISCOVERING FREQUENT NAVIGATION PATTERNS FOR CONSTRUCTING USER PROFILE: THE CASE OF EBIZ ONLINE SOLUTIONS PLC OFFICIAL WEBSITE

By

ANTENEH LEGESSE TEKLU

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
Million Meshesha (PhD)	Advisor	_____	_____
_____	Examiner	_____	_____

DEDICATION

I would like to dedicate this paper to my father and mother who have been struggling and fighting for my education since I was a little boy. **My parents' congratulations; your dream is now realized!!**

ACKNOWLEDGEMENT

I would like to thank many people for their kindness and assistance during my thesis writing. I sincerely thank my thesis advisor Dr. Million Meshesha for his guidance and encouragement. I owe a lot for his patience, encouragement and heartwarming personality.

I am greatly indebted to Ato Demissew, the owner of Ebiz (e-business) Online Solutions PLC without his help of this research would have been impossible.

I truly appreciate the support of the people at the Ebiz (e-business) Online Solutions PLC. Special thanks are also due to my friends who have one way or another contributed to the success of my thesis.

Most importantly, I thank my father and mother, who have given me great support and encouraged me throughout this difficult but exciting journey.

A heartfelt thanks is also due to my sweetheart Beza. I wouldn't have finished this thesis without her support and love.

Finally, during my stay in the University the support and encouragement of my family, friends and relatives are worth being acknowledged.

Anteneh Legesse

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ARFF	Attribute relation File Format
CGI	Common Gateway Interface
CLF	Common Log Format
CRM	Customer Relationship Management
CSV	Comma-Separated Values
ECLF	Extended Common Log Format
FP-Growth	Frequent Pattern Growth
KDD	Knowledge Discovery in Data
NLP	Natural Language Processing
PLC	Private Limited Company
SEO	Search Engine Optimization
W3C	World Wide Web Consortium
WEKA	Waikato Environment for Knowledge Analysis
WP	Web Personalization
WUM	Web Usage Mining

LIST OF APPENDICES

ABSTRACT

World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. To reduce users browsing time a lot of research has taken place. Ebiz (e-business) being a dynamic and fast growing online service giving organization, it should continuously assess and monitor customers' usage behavior and restructure the website as well as Personalize the pages accordingly.

Web Usage Mining is a type of web mining which applies mining techniques in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching and creating attractive web sites. In this study, an attempt is made to discover useful patterns from the server log files of Ebiz Official website used as input for user profile.

In this research, web usage mining process model suggested by Lalithadevl et al. is used. This model has five distinct phases; Data gathering, Data preprocessing, Navigation pattern discovery, Pattern analysis and visualization, and Pattern applications. Web Server log is used for statistical and pattern discovery. Moreover, TENDER content database is used to create user profile. Next the log files have been preprocessed. That is, data cleaning, data integration, feature selection, data grouping into different web content categories, transaction identification and transformation of the data to Weka understandable format was performed using WUMprep tool, Python programming, MS Access and MS Excel. Finally a total of 42,154 transactions have been prepared for the experiment.

The experiment have been conducted using FP-Growth and K-means algorithms in order to discover interesting patterns of the different web content categories. WebLog Expert is used to yield different useful statistical reports. MS Access 2013 and W3Perl are used to create user profile. The statistical analysis shows that 83.86 % of the user of Ebiz do not browse further than four pages into the site. According to the evaluation result obtained on the user profile, 87.5% of the experts believe that the user profile is useful in all aspect of the questions.

The finding of the study indicates that the Ebiz (e-business) official website should be restructured and the page needs to be personalized using the user profile as a base.

Keywords: Web Usage Mining, Pattern Discovery, Log file, User Profile

CHAPTER ONE

INTRODUCTION

1.1. Background

The World Wide Web (WWW) is the largest distributed information space which has grown to encompass diverse information resources such as service and product catalogs, digital libraries, personal home pages, Usenet news, etc. More noticeably nowadays, the web is considered as the most appropriate environment for business transactions because it is convenient, fast, and inexpensive to use; hence, we see the enormous popularity of electronic commerce and Business-to-Business applications [1].

Although the web is growing exponentially, the individual's capacity to read and digest content is essentially fixed. The web is a collection of semi-structured and structured information sources often visualized as a huge and complex dynamic mesh. Due to information explosion, constantly changing environment, poor understanding of users' needs and preferences, as well as lack of willingness to modify existing web data models, often web users suffer from information overload. Therefore, the full economic potential of the web has not been realized. The ability to access information in the web efficiently and effectively is an enabling technology for realizing its full potential [2].

The result of this explosive growth of the Web has also triggered an increasing demand of Web personalization systems. Personalized information technology services have become a ubiquitous phenomenon, taking advantage of the knowledge acquired from the analysis of the user’s navigational behavior or usage data. Web usage mining (WUM) aims at discovering interesting patterns of usage by analyzing web usage data. This method provides an approach to the collection and pre-processing of those data, and the construction of models representing the behavior and the interests of users. These models can be automatically incorporated into personalization components, without the intervention of any human expert [3].

Some researches only focus on how to meet customers’ information needs from the perspective of functional information, and they exclude most of the other information needs

from their consideration [12, 13, 14]. Research conducted by France et al [4] stated that the attempts of search engines and data mining technology to improve Web information search capabilities to match up various information needs has been limited. This research aims to work on usage mining techniques to create user profile which is an input for personalized recommendations, by customizing the contents of a Web site with respect to the user's need.

Originally, the aim of Web usage mining (WUM) has been to support the human decision making process. Thus, the outcome of the process is typically a set of data models that reveal knowledge about usage patterns of users. WUM typically extracts knowledge by analyzing historical data such as Web server access logs, browser caches, and proxy logs. Using WUM techniques, it is possible to model user behavior, and therefore, to forecast their future movements [5]. In this research the information mined can subsequently be used in order to create user profile.

While each user has individual interests, we can expect a lot of overlap. A trader who is searching for the information about Ethiopian trade system is not alone. Users who want to know and share information get by searching the appropriate documents on the Web site. Hence, a Web site presented according to the entire user community's information access activities is likely to make it easier for new users to find relevant documents quickly which enables users to save time and unnecessary cost.

The knowledge discovered through the usage mining process serves as operational knowledge to create user profile [3]. Realizing the potential of WUM techniques to construct this knowledge, the researcher proposed to create user profile, as an output from analyzing the user's navigational behavior or usage data of Ebiz (e-business) official web site.

1.2. Ebiz (e-business) Online Solutions PLC

Ebiz (e-business) Online Solutions PLC was set up in 2014 to provide integrated information for the Ethiopian business sector [11]. As a shopper has to find a smart route

and good guide to go to Merkato (hence the name tomerkato) and shop there. Hence 2Merkato is introduced to serve as a smart route and good guide to business in Ethiopia. It contains virtually every important information a business person or a potential investor needs while doing business in Ethiopia or trading with businesses located in Ethiopia. The information is provided in easy to navigate formats and contains the regulations, procedures, incentives and statistics of Business Startup in Ethiopia, Investment in Ethiopia, Import to Ethiopia, Export from Ethiopia, Shipping, Ethiopian Tax System and Ethiopian Customs. 2Merkato also serves as a launch pad for people who want to start business in Ethiopia [11].

In addition to these entries, 2Merkato has some dynamic features that serve the local and international business community:

- ***Business Directory***: the directory contains the addresses of more than 7000 businesses, daily updated tender notices and ads: from major newspapers and also posted by users.
- ***Ethiopian Business News***: recognized by Google News as an Ethiopian News source.
- Daily updated Foreign Exchange Information

The main goal of 2Merkato is to be a dependable, accurate, frequently updated source of Ethiopian business information in such a way that every business owner and executive in Ethiopia knows and uses http://www.2merkato.com [11].

The following are the software and platform used for the development and deployment of the website:

- Client side: ASP. NET, HTML, Java Script, JQuery, CSS. AJAX
- Server side: PHP
- DBMS: MySql
- Web Server: Apache
- Operating System: Ubuntu

2Merkato keep several log files on the server. The server operating system has its own log files, web server log, Apache server, has its own log file and almost all the additional application installed including the mail server, database server, etc… have their own log files. In addition to that, the Content Management System (CMS) and the PHP frameworks

that used to manage website content and structure. Moreover, there is a utility developed to transfer data from test web server to production web server.

2Merkato will most likely continue to provide information for the Ethiopian business community in the form of news updates, guides, and tips on best practices of doing business in Ethiopia as well as Ethiopian business current rules and regulations. It also keeps an introduction services similar to online tender publishing.

1.3. Statement of the problem

E-commerce has been growing rapidly keeping the pace with the web. Its rapid growth has made both companies and customers face a new situation. Whereas companies are harder to survive due to more and more competitions, the opportunity for customers to choose among more and more products has increased the burden of information processing before they select which products or services meet their needs [8]. As a result, the need for new marketing strategies such as one-to-one marketing and customer relationship management (CRM) has been stressed both from researches as well as from practical side [9]. One solution to realize these strategies is to create user profile which contain interest of customers that helps customers find the products or services they would like to purchase by producing a list of recommended products or services for each given customer.

Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the web. Web usage mining has become very critical for effectively creating user profile, web site management, creating adaptive web sites, business and support services, personalization, network traffic analysis etc. [10].

The website of 2Merkato has been set up by Ebiz (e-business) Online Solutions PLC to provide integrated information on Ethiopian business. It has an official website namely, WWW.2merkato.com. It is the most direct link a company has to its current and potential customers. In addition to other basic services, the website serves as an e-commerce gateway by providing services such as online advertising, online Business director search, and provides business information for investors, suppliers, and buyers, traders who are engaged or want to engage in business activities in Ethiopia.

The website of 2Merkato serves both the local and international business community. Locally as Ethiopia is new to technology based activity, 2Merkato websites face extra challenge to become at users top list. Moreover, 2Merkato competes globally with other major websites that are providing information about bids, advertisement, and other e-business activities. In order to win this challenging competition, it is required to address various needs of its customers. One of such activity is achieved through analyzing the behaviors of its customers who access the website and streamline the website towards their need.

There are few attempts made to measure and analyze the website access behavior of the customers. Google Analytics is being utilized for statistical analysis of the website usage by the Webmaster. But there is no research conducted on this website. According to the web master and web administrator suggestion there is slow response time, Non-intuitiveness of the page link hierarchy, Language barriers for some website users, not fully optimized search engine, the problem of forecasting effective advertising using the website etc. As Ebiz (e-business) is one of the fastest growing web service giving organization, it needs to carry out continuous research to improve its services quality in order to create adaptive website to satisfy its customer. One of the major such areas is website usage analysis. Moreover, for proper customer segmentation, a study that considers frequent customers in their navigational behavior is essential.

This research therefore aims to create user profile for 2Merkato Web site using web usage mining. The research focuses on the user's activity on a Ebiz Web site; recommend customizing the contents and structuring the presentation of a Web site according to the preferences derived from the user's activity.

To this end, this study attempts to explore and answer the following research questions:

- How to prepare appropriate dataset with relevant attributes for web usage mining?
- Which algorithms are suitable for data preprocessing, pattern discovery and creating user profile?
- What are interesting users browsing behaviors of 2Merkato Web site uncovered by web usage mining?

- How can the captured knowledge be utilized to construct user profile for the Ebiz (e-business) Web site?

1.4. Objective of the study

1.4.1. General objective

The general objective of the research is to construct user profile based on the navigational behavior of users of the 2Merkato official web site of Ebiz (e-business) Online Solutions PLC using web usage mining techniques so as to facilitate personalized recommendation for the website users.

1.4.2. Specific objectives

To achieve the general objective of the research, the following specific objectives shall be addressed.

- To review relevant related literatures to understand the problem area and to identify web usage mining techniques and algorithms.
- To prepare dataset from log web data collected from www.2merkato.com official web site.
- To select appropriate web usage mining tools, techniques and algorithms.
- To describe the website usage statistics based on different parameters.
- To identify users' frequent navigational patterns and discover association rules.
- To construct user profile based on the extracted users navigational behavior and patterns.
- To undertake users acceptance testing about the reliability of the constructed user profile.

1.5. Significance of the Study

The redesign of the whole site (interface, content, structure, usability, etc.) is one of the most important aspects for any institution that wants to survive in the cyberspace. Therefore the output of this study will give an insight to Ebiz (e-business) about the usability of the website design, effectiveness and other business decision making etc. with respect to user navigation patterns. This gives the 2Merkato a chance to improve both the structure,

usability and the content of the website in order to make it more usable and effective. This in turn, helps Ebiz to address a broad customer base and increase its revenue.

As the main goal of Ebiz (e-business) website is to be a dependable, accurate, frequently updated source of Ethiopian business information and also every business owner and executive in Ethiopia knows and uses 2merkato.com. The result of this study will help to know the various demands of customers with regard to the website usage and address those demands. The study will give an opportunity to restructure the website in a more effective way which will help the company to carry out effective promotional and marketing strategy. In addition to satisfying customers' requirement, web server performance can also be improved.

1.6. Scope and Limitation of the Research

This study aims to use the server side of web log files. Actually, there are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. However, in this research, web access log records is used as dataset because literatures and previous researches [15] justify that web access log files is the typical source of navigational behavior. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based on only the server side data [15].

The scope of this research is to construct descriptive models using clustering and association rule discovery techniques so as to statistically describe and classify users and web pages in order to explore Ebiz (e-business) website usage. In this study, the log files are preprocessed, grouped according to navigational behavior using web user navigation patterns and transformed. Then pattern discovery and analysis is made based on which user profile is created according to user's behavior. The study has been conducted on access log files of five months duration that covers from Jul 25, 2014 to Jan 22, 2015.

In this study only frequently accessed pages are used; feature sets having infrequently visited pages are not used because of the collected data is not enough to discover interesting pattern from infrequently visited pages. In order to identify unique users (visitors) the

researcher uses host (IP address), browsers and operating system. However, different users may use the same host (IP address), browsers and operating system which needs further experimentation.

1.7. Research Methodology

Research Methodology is the general principle that guides the research. In order to conduct, a good research, a well-defined approach and principle has to be followed. In this study experimental type of research is used. It is the collection of research design which use manipulation and controlled testing to understand causal processes.

A review of relevant literature has been conducted to assess data mining technology, web usage analysis and user construction profile. Various books, journals, magazines, articles and research papers pertaining to these subject areas have been consulted to understand the potential applicability of web usage mining for construction the user profile in evaluation of web sites, particularly official web site of Ebiz (http://www.2Merkato.com).

1.7.1. Research Design

This study employs a five phase Web Usage Mining process model suggested by Lalithadevl et al. [16]; such as Data gathering, Data preprocessing, Navigation pattern discovery, Pattern analysis and visualization, and Pattern applications.

1.7.2. Data gathering

Web usage data are usually supplied by two sources: trial runs by humans (Explicit) and Web logs (Implicit). The first approach is impractical and rarely used because of the nature of its high time requirement, costly operation and biasedness. Most usage mining systems therefore use log data as their data source [16].

Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from registration forms and rating questionnaires. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages [16].

For this research, usage data of the recent log file from July 25, 2014 to January 22, 2015 is accessed from the web server of Ebiz Online Solutions PLC. The collected data spans five

months. The overall dataset used is very large (over 6.2 GB) containing over 26,000,000 records.

1.7.3. Preprocessing the Web data

Generally, data preprocessing consists of data cleaning, user identification, session identification and path completion, as shown in Figure 1.1 [17].

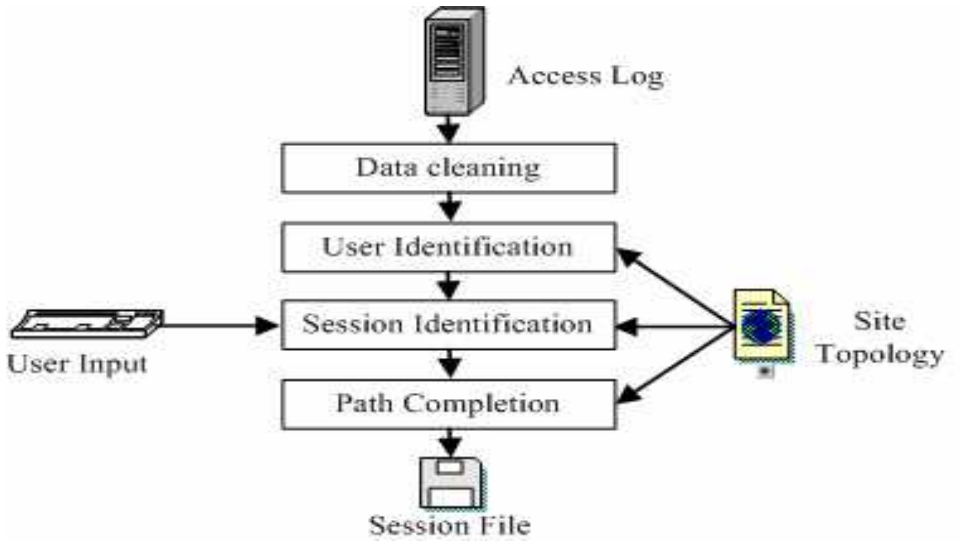

Figure 1.1: Preprocessing Steps for the web data [17]

In this study, however, the first three steps are considered. Path analysis step is not applied since the collected web data for the current study contains no missing page requested.

Data Cleaning

This task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data cleaned: accessorial resources embedded in HTML file, robots' requests and error request [17]. In this research the following log entries are removed in this stage

- ► Entries like gif, JPEG that are downloaded along with user's request.
- ► Unsuccessful status code log entries like <200 and >299.
- ► Automated programs like web robots, spiders and crawlers.

User Identification

A user is identified as the principal using a client to interactively retrieve and provide the resources or access forms. The Web Usage Mining techniques based on the cooperation of users are the easiest ways to deal with this problem [18]. This research use the following heuristics to identify the user:

- Each IP address represents one user;
- For more logs, if the IP address is the same, but the agent log shows a change in User Agent (with Version) or operating system (with screen resolution and processer speed etc.), an IP address represents a different (New) user.

User Session Identification

A user session means a delimited set of user clicks (click stream) across one or more Web servers. The goal of session identification is to divide the page accesses of each user into individual sessions. At present, the methods to identify user session include timeout mechanism and maximal forward reference mainly [18]. The following rules are used to identify user session:

- If there is a new user, there is a new session;
- In one user session, if the refer page is null, there is a new session;
- If the time between page requests exceeds a certain limit (30 minutes), it is assumed that the user is starting a new session.

To undertake the preprocessing task tools, such as WUMprep and Python 2.7 are selected. WUMprep is a java based open source software. It is used to filter raw logs, generate session files, and remove robot accesses. The tool has a configuration editor in which all sorts of activities to be accomplished are set. Once the configuration is saved at the same directory with the raw log file, Perl scripts are available which could be run to perform desired preparation. The Perl scripts include request Filter.pl, sessionize.pl, session Filter and others. The researcher selected the tool regarding the following conditions. Firstly, the tool is very good at detecting and removing non-human requests and automatically downloaded graphics file provided that the user of the tool feed all appropriate data of the log files. Secondly, the tool uses time-based heuristic in developing user sessions of the human requests. Thirdly, it is the only freely available log preparation tool that could be accessed by the researcher [12, 13, 14].

Once WUMPrep is utilized in preparing the raw web log data, the output is user session file in which requests of the same session ID are scattered across the file. Thus, series of text-based preparation should be accomplished to get a data format that is compatible with the selected Weka‘s algorithms. Thus, python programming language was selected for the reason that python is very suitable for text processing tasks. Besides, it is open source software that the researcher could get free of charge [13,14].

1.7.4. Navigational Pattern Discovery

The pattern discovery stage is applying data mining techniques like path analysis, association rule mining, clustering, classification etc., on preprocessed log data. In this study clustering and association technique is considered for pattern discovery. There are two types of clusters to be discovered: usage clusters and page clusters. Clustering usage data is to find visitor groups with common properties, interest or behavior. The aim of clustering web pages is to divide the dataset into groups of pages which have similar content. This study deals with clustering log data which have similar content.

In this study, WEKA software is used for pattern and association rule discovery. WEKA is a Java application that has a collection of visualization tools and algorithms for preprocessing data analysis and predictive modelling, coupled with a graphical user interface for easy access to its functionality. The tool could be used in one of the four modes: Simple Command Line Interface (CLI), Experimenter, Explorer and Knowledge flow [12, 13, 14].

The most common file format for Weka is Attribute-Relation File Format (ARFF) files. Indeed, other file formats such as CSV data files, C45 names files, and Binary serialized instances could be used as input to the tool. ARFF files have three sections: relation, attributes and data. The relation contains the name of the relation with “@Relation” tag. Attributes are defined at the attributes section with the attribute ‘tag and each of the log transactions are written onto the data section of the ARFF file format. Only the header is tagged with “@data”.

The researcher selected Weka data mining tool for pattern discovery on consideration of the following facts.

- Weka is freely available software, licensed under GNU.
- It is easily portable to other platforms by being completely developed in Java.
- Supports various data mining tasks, namely data preprocessing, association, clustering, classification, regression, visualization, and election.
- Allows data files to be read in plain text (CSV, ARFF ...) or directly from relational databases through the JDBC API.
- It is extensible, so new algorithms can be incorporated.
- The researcher is well experienced in using of the tool.

FP-Growth Algorithm is used for pattern discovery. This algorithm is selected because of its efficiency and completeness over other algorithms such as Apriori. Refer to Performance of Apriori vs. FP-Growth Algorithm in Chapter three (Research methods and techniques).

K-means algorism is used because it is very simple and easy to understand and also it follows a very easy way for classifying the items given. Moreover the researcher knows the number of clusters in advance.

In this study **WebLog Expert** is used for statistical analysis. The researcher select this tools because it can be executed any WINDOWS operation system from Windows 2003 itself. It can support the Apache and IIS server logs. It automatically detects the log format and read the GZ and ZIP compressed logs. We can analyze logs from load balanced servers and download logs via FTP and HTTP. By using this software, it is possible to create a report in HTML, PDF and CSG format perhaps can upload reports via FTP and send via email (SMPT or MAPI). The main role of the IP in this software is to country mapping database with additional city, state and organization database. Finally it supports, date macros, multi thread DNS lookup and command line mode [42].

In this study **W3Perl** is used for constructing user profile statistically. The researcher select this tools because it consists of set of Perl Script that can analyze log files for IIS, Apache, FTP, mail etc. supports sessions (length of time visitors spend on your site), RSS stats, referrers, keywords used on search engines, list of error pages invoked, classification of

your visitors by countries, browser stats, screen sizes, real-time statistics, etc. The software is free and licensed under the GNU GPL. http://www.w3perl.com/ describes more details about this tool [41]. The researcher used this tool for the purpose of creating user profile statistically by page view as well as by IP.

1.7.5. Pattern Analysis and Visualization

Navigation patterns, which show the facts of Web usage, need further analysis and interpretation before application. In this final phase the objective is to convert discovered rules, patterns and statistics into knowledge or insight involving the Website being analyzed. Knowledge here is an abstract notion that in essence describes the transformation from information to understanding; it is thus highly dependent on the human performing the analysis and reaching conclusions [16].

The statistical reports and patterns discovered are analyzed using different techniques such as Literature Review, Website structure and content analysis, business knowledge, discussion with technical persons working on the website, and lift interestingness measure. In general the techniques used here are “visualization techniques” and “Usability Analysis”. After analyzing the findings, recommendations are forwarded for future researches as well as to improve the website performance.

1.7.6. Pattern Applications

In this research the results of navigation pattern discovery are applied to construct user profile which can be used for web personalization, Site Modification and integrating recommender system for online business and marketing applications. .

1.8. Organization of the thesis

This Thesis report is organized into six chapters. Chapter one discusses introduction, background of the study, problem statement, objective of the research, significance of the study, scope and limitation of the research and methodology followed. Chapter two is the literature review part in which different conceptual and empirical concepts regarding Web Usage Mining and personalization are discussed. Also related workers are presented to show what is and is not done. Chapter three present research methods and techniques

applied in this study. Chapter four deals with data preparation. In this chapter, data collection, data cleaning, user identification, session identification, feature selection, transaction identification, data transformation and creating user profile are discussed. Moreover, algorithms used for data preprocessing are described. Chapter five deals with experimentation and analysis. Here, different statistical analysis and experimentation of pattern discovery is presented. Chapter six summarizes the research findings and provides concluding remarks and recommendations for further research.

CHAPTER TWO
LITERATURE REVIEW

Data mining has an important place in today's world. It becomes an important research area since the amount of data available in most of the application are increase alarmingly. This huge amount of data must be processed in order to extract useful but hidden information and knowledge. As defined by Han and Micheline [40] Data Mining is the process of discovering interesting knowledge from large amount of data.

Data Mining has various application areas including banking, biology, e-commerce etc… [40]. These are the most well-known and classical application areas. On the other hand, the new data mining applications include processing spatial data, multimedia data, time-related data and World Wide Web.

World Wide Web (WWW) is one of the largest and most widely known data source. Today, WWW contains billions of documents edited by millions of people. The total size of the whole documents can be interpreted in many terabytes [21]. All documents on WWW are distributed over millions of computers that are connected by telephone lines, optical fibers and radio modems. WWW is growing at a very large rate in size of the traffic, in the amount of the documents and in the complexity of web sites. Due to this trend, the demand for extracting valuable information from this huge amount of data source is increasing every day. This leads to new area called Web Mining [21], which is the application of data mining techniques to World Wide Web.

2.1. Overview of Web Mining

There are several reasons for the emergence of web mining. First of all the World Wide Web is a huge, interconnected, semi-structured, widely distributed, highly heterogeneous and hypertext information repository [6]. The Web continues to grow at an incredible rate as information gateway. Many organizations, individuals or societies provide their public information through web. Also, the content of the web pages are much more complex than any other traditional text documents. Today, web pages lack standard structure; they contain more complex style than standardized formats [25].

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and

retrieve information as well as to extract useful knowledge. Another reason is that due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, research has encountered a lot of challenges, such as scalability, multimedia, temporal and related issues. As a result, Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the web. A user interacting with the web has a wide diversity of navigational preference, which results in needing different contents and presentations of information [19, 22, 23].

The challenges listed above leads to a research for effective discovery and use of resources in World Wide Web, which leads to new research area called Web Mining. As many believe, it is Oren Etzioni first proposed the term Web mining in his paper [21]. He claimed that the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Many of the researchers cited this explanation in their works. Two different approaches were proposed for defining Web mining. The first approach is a 'process-centric view', which defines Web mining as a sequence of ordered tasks [21]. Second one is a 'data-centric view', which defines web mining with respect to the types of web data that was used in the mining process [20]. As depicted in Figure 2.1, Web data is classified as follows [25]:

- ***Content data****:* Content data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.
- ***Structure data****:* - Structure data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Website together, such as hyperlinks connecting one page to another.
- ***Usage data:*** - Usage data represent a Web site's usage, such as a visitor's IP address, time and date of access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.
- ***User profile data****:* - User profile data provide information about the users of a Web site. A user profile contains demographic information for each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

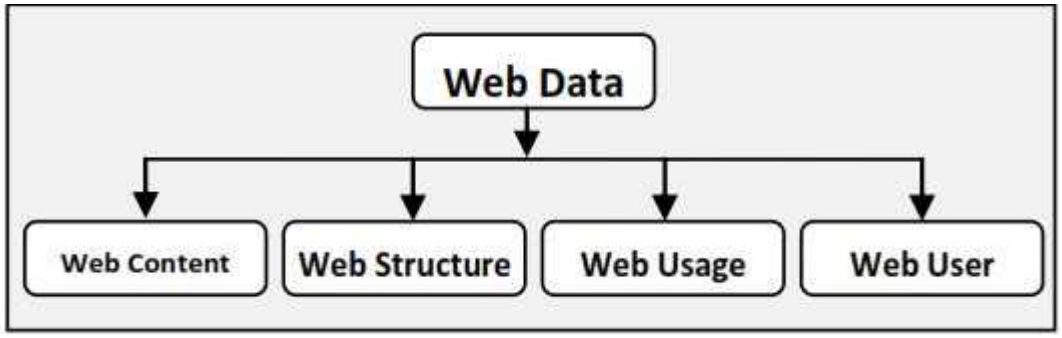

Figure 2.1: Classification of Web Data [28].

Accordingly, Web mining can be classified into three different types [23]: Web content mining, Web structure mining and Web usage mining. The details of web mining taxonomy is presented in figure 2.2.

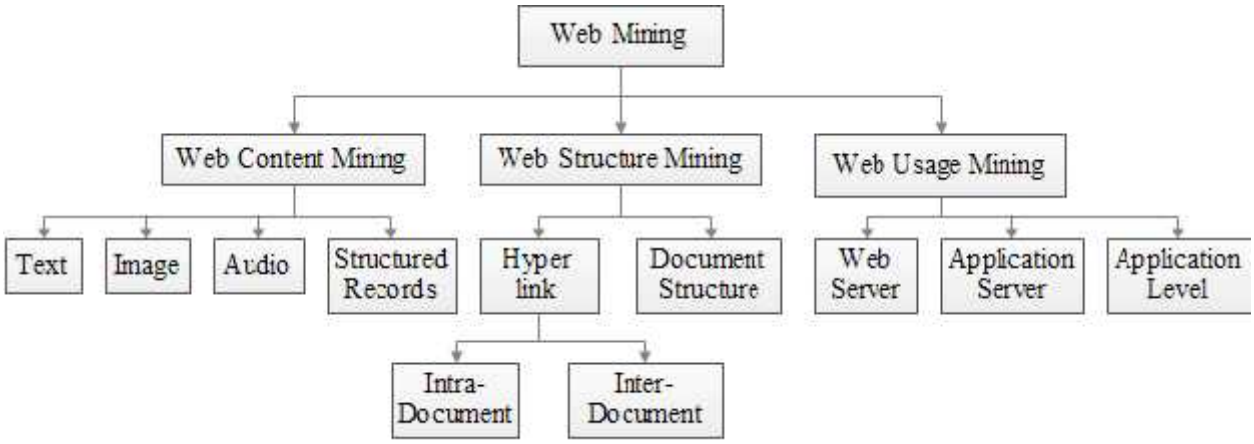

Figure 2.2: Web Mining Taxonomy [23].

2.1.1. Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve the use of techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP) (for further detail refer [22]).

2.1.2. Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages [22]. This can be further divided into two kinds (Hyperlinks and Document Structure) based on the kind of structure information used.

- ***Hyperlinks: -*** A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.
- ***Document Structure***: - In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

2.1.3. Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site [22]. Web usage mining itself can be classified further depending on the kind of usage data considered [22]:

- ***Web Server Data***: - The user logs are collected by Web server. Typical data includes IP address, page reference and access time.
- ***Application Server Data***: - Commercial application servers such as Web logic, Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
- ***Application Level Data***: - New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.

2.2. Location of web log file

Data which is used for web usage mining can be collected from three different levels [24]: Server side, Client side and Proxy side.

Web Server side

Web servers are the common source of data. They store large amounts of information in their log files. These logs generally contain basic information e.g. name and IP of the remote host, date and time of the request etc. The web server stores data regarding request performed by the client. Data can be collected from multiple users on single site. All the click streams are recorded into the web server log file.

Client side

It is the client itself which sends information to a repository regarding the users' behavior. This is done either with an adhoc browsing application or through client side application running standard browsers. Client level data collection can be implemented by using a remote agent (such as Java applets or Java Scripts).

Proxy server side

Information about user behavior is stored at proxy level, thus web data is collected from multiple users on several websites, but only users whose web clients pass through the proxy. Proxy servers collect data of groups of users accessing huge groups of web servers. Proxy level collection is an intermediary between server level and client level. The page load time gets reduce by proxy server, so user experience high performance. In this study, Web Server (HTTP server) log data is used for web usage mining.

2.3. Types of Web Server Logs and their format

Web Server logs are plain text (ASCII) files and are Independent from the server. There are some distinctions between server software, but traditionally there are four types of server logs [25]: Access Log, Error Log, Agent Log and Referrer Log.

The first two types of log files are standard. The referrer and agent logs may or may not be "turned on" at the server or may be added to the transfer log file to create an "extended" log file format [25]. The log file entries of Apache HTTP Server Version 1.3 are discussed below:

2.3.1. Access Log

The server access log records all requests that are processed by the server. The location and content of the access log are controlled by the Custom Log directive. The Custom Log directive is used to log requests to the server. A log format is specified, and the logging can optionally be made conditional on request characteristics using environment variables. The Log Format directive can be used to simplify the selection of the contents of the logs. This section describes how to configure the server to record information in the access log.

Web Access log file is a simple plain text file which record information about each user. Display of log files data in three different format [22].

W3C Extended log file format

W3C log format is default log file format on IIS server. Field are separated by space, time is recorded as GMT (Greenwich Mean Time). It can be customized that is administrators can add or remove fields depending on what information wants to record [22].

NCSA common log file format

NCSA is a fixed ASCII text-based format, so you cannot customize it. The NCSA Common log file format is available for Web sites and for SMTP and NNTP services, but it is not available for FTP sites. Because HTTP.sys handles the NCSA Common log file format, this format records HTTP.sys kernel-mode cache hits [22].

IIS log file format

IIS is a fixed ASCII text-based format, so you cannot customize it. Because HTTP.sys handles the IIS log file format, this format records HTTP.sys kernel-mode cache hits [22]. NCSA and IIS log file format the data logged for each request is fixed.W3C format allows user to choose properties, user want to log for each request.

Here are three log formats considered for access log entries in the case of Apache HTTP Server Version 1.3.They are briefly discussed below [25]:

Common Log Format (CLF)

The configuration of the common log format is given below [25].

LogFormat "%h %l %u %t \"%r\" %>s %b" common

CustomLog logs/access_log common

The log file entries produced in CLF will look something like this:

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

The entries give details about the client who had requested for the web site to the web server

- 127.0.0.1 (%h) - This is the IP address of the client which made the request to the server.
- - (%l) - The hyphen present in the log file entry next to the IP address indicates that the requested information is not available.
- frank (%u) - The user id of the person requesting the document as determined by HTTP authentication
- [10/Oct/2000:13:55:36 -0700] (%t) -The time format resembles like [day/month/year: hour: minute: second zone]
- "GET /apache_pb.gif HTTP/1.0" (\"%r\") - The request sent from the client is given in double quotes. GET is the method used. apache_pb.gif is the information requested by the client. The protocol used by the client is given as HTTP/1.0
- 200 (%>s) - This is the status code sent by the server. The codes beginning with 2 for successful response, 3 for redirection, 4 for error caused by the client, 5 for error in the server
- 2326 (%b) - The last entry indicates the size of the object returned to the client by the server, not including the response headers. If there is no content returned to the client, this value will be "-"

After processing the request of the client in the web server the status code is sent by the web server. There are various status that are exhibited by Apache HTTP Server Version 1.3 as given below [27]:

1xx Info

HTTP_INFO – Request received, continuing process

- 100 Continue – HTTP_CONTINUE
- 101 Switching Protocols -HTTP_SWITCHING_PROTOCOLS
- 102 Processing – HTTP_PROCESSING

2xx Success

HTTP_SUCCESS **–** action successfully received, understood, accepted

- 200 OK – HTTP_OK

- 201 Created – HTTP_CREATED
- 202 Accepted – HTTP_ACCEPTED

3xx Redirect

HTTP_REDIRECT – The client must take additional action to complete the request xx Info

- 301 Moved Permanently – HTTP_MOVED_PERMANENTLY
- 302 Found – HTTP_MOVED_TEMPORARILY
- 304 Not Modified – HTTP_NOT_MODIFIED

4xx Client Error

HTTP_CLIENT_ERROR – The request contains bad syntax or cannot be fulfilled

- 400 Bad Request – HTTP_BAD_REQUEST
- 401 Authorization Required –HTTP_UNAUTHORIZED
- 402 Payment Required – HTTP_PAYMENT_REQUIRED
- 404 Not Found – HTTP_NOT_FOUND
- 405 Method Not Allowed – HTTP_METHOD_NOT_ALLOWED

5xx Server Error

HTTP_SERVER_ERROR – The server failed to fulfill an apparently valid request.

- 500 Internal Server Error – HTTP_INTERNAL_SERVER_ERROR
- 501 Method Not Implemented – HTTP_NOT_IMPLEMENTED
- 503 Service Temporarily Unavailable – HTTP_SERVICE_UNAVAILABLE
- 505 HTTP Version Not Supported – HTTP_VERSION_NOT_SUPPORTED

These status codes that are sent along with the response data is also entered in the log file.

Combined Log Format

The configuration of the combined log format is given below

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"" combined

CustomLog log/acces_log combined

The log file entries produced in combined log format will look something like this:

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"

The additional parameters that are present in the combined log format are discussed below

- *"http://www.example.com/start.html" (\"%{Referer}i\")* - This gives the site that the client reports having been referred from. (This should be the page that links to or includes /apache_pb.gif).

- *"Mozilla/4.08 [en] (Win98; I ;Nav)" (\"%{User-agent}i\")* - This is the information that the client browser reports about itself to the server.

These entries are made in the log file for a combined log format entry [25].

Multiple Access Logs

Multiple access logs can be created simply by specifying multiple Custom Log directives in the configuration file. In this type of log file there are three files created as access log files containing details about the client. It is said to be a combination of common log format and combined log format [25].

The configuration of the multiple access log is given below:

LogFormat "%h %l %u %t \"%r\" %>s %b" common

CustomLog logs/access_log common

CustomLog logs/referer_log "%{Referer}i -> %U"

CustomLog logs/agent_log "%{User-agent}i"

The first line contains the basic CLF information, while the second and third line contains referrer and browser information.

Most of the web servers have the same formats being followed for the log file entry.

2.3.2. Error Log

The information that is contained in most error log entries is the message given below

[Wed Oct 11 14:32:52 2000] [error] [client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test

Whenever an error is occurred while the page is being requested by the client to the web server the entry is made in the error log. The first set of item present in the log entry is the date and time of the message. The second entry lists the severity of the error being reported. The Log Level directive is used to control the types of errors that are sent to the error log by restricting the severity level. The third entry gives the IP address of the client that generated the error. Next is the message itself, which in this case indicates that the server has been configured to deny the client access. The server reports the file-system path of the requested document [25].

2.3.3. Agent Log

Agent log files contain the name of the web browser that was used in each access to the site (for example, *Netscape*, *Opera*) [21].

2.3.4. Referrer Log

Referrer log files list the web page that a user accessed prior to accessing any web page on the site. The information tracked in the referrer log file can be used to determine from *where* (from which site or web page) a user came to this web page [21].

Although these different types of logs are constantly updated by the web servers in the course of the capturing daily web usage information, access logs contain the most interesting patterns waiting to be discovered.

2.4. Approach for Web Usage Mining

When data mining techniques are applied on web usage data in order to extract useful knowledge regarding user behavior, it is known as web usage mining. It is an approach for collecting and preprocessing web usage data, and then constructing models that represent the behavior and interests of users. Such models can automatically be used by personalization system for predicting user's personal interests and thus enhance his surfing experience with the website [28]. As presented by Lalithadevl et al [16] (see figure 2.3) the Web usage mining involves five major functions:

- ***Usage data gathering***: Web logs, which record user activities on Web sites, provide the most comprehensive, detailed Web usage data.
- ***Usage data preparation***: Log data are normally too raw to be used by mining algorithms. This task restores the users' activities that are recorded in the Web server logs in a reliable and consistent way.
- ***Navigation pattern discovery***: This part of a usage mining system looks for interesting usage patterns contained in the log data. Most algorithms use the method of sequential pattern generation, while the remaining methods tend to be rather ad hoc.
- ***Pattern analysis and visualization***: Navigation patterns show the facts of Web usage, but these require further interpretation and analysis before they can be applied to obtain useful results.
- ***Pattern applications***: The navigation patterns discovered can be applied to the following major areas, among others: i) improving the page/site design, ii) making additional product or topic recommendations, iii) Web personalization, and iv) learning the user or customer behavior [16].

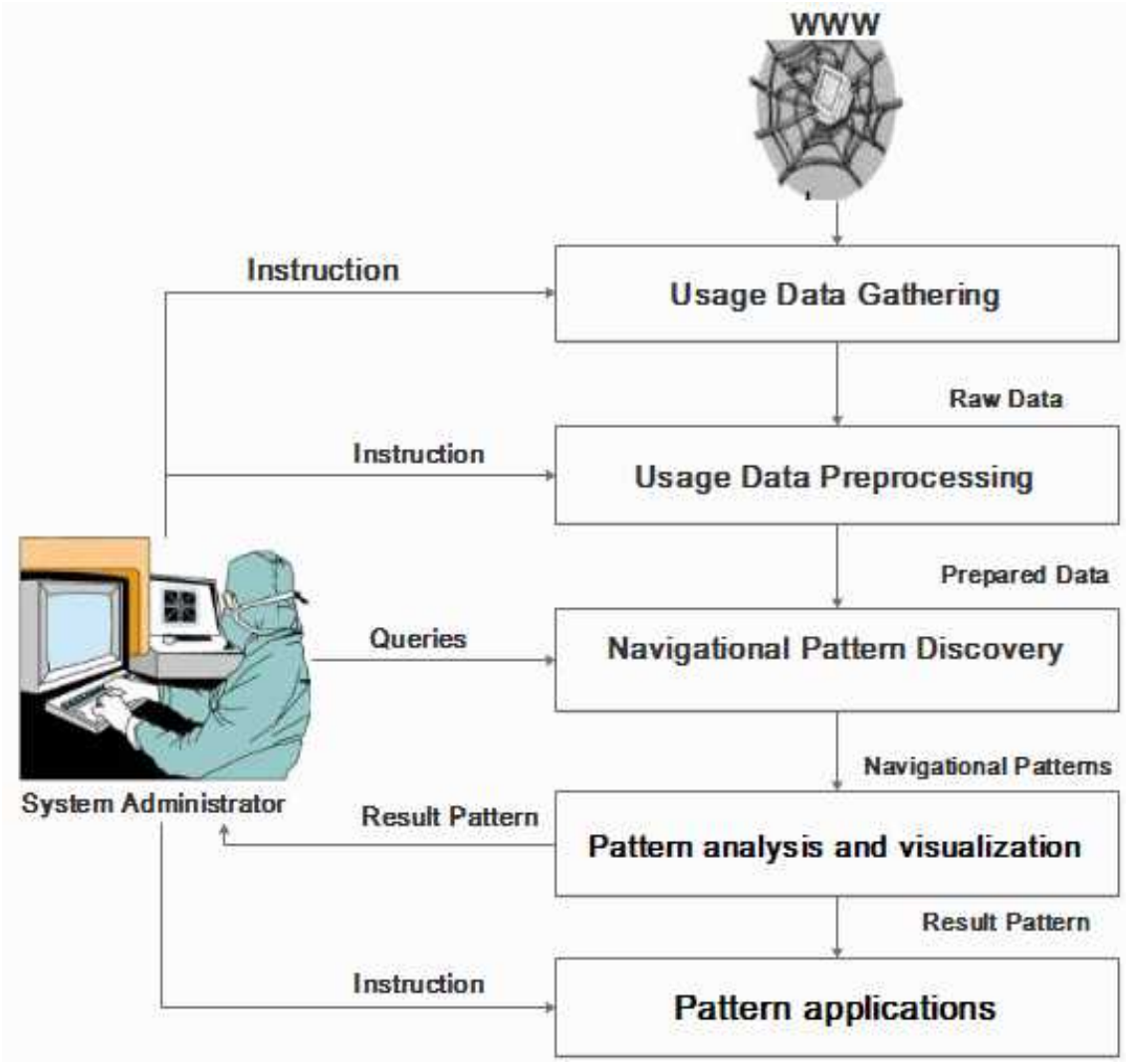

Figure 2.3: Architecture of Web usage mining [16].

A web usage mining can also be divided into the following two types [16]: Personal and Impersonal.

Personal: A user is observed as a physical person, for whom identifying information and personal data/properties are known. Here, a usage mining system optimizes the interaction for this specific individual user, for example, by making product recommendations specifically designed to appeal to this customer.

Impersonal: The user is observed as a unit of unknown identity, although some properties may be accessible from demographic data. In this case, a usage mining system works for a general population, for example, the most popular products are listed for all customers.

2.5. Usage data gathering

In this stage, the usage data is collected from a range of possible sources after which, their contents and structure is recognized. Data is collected from a various sources (see figure 2.4) such as those from web servers, clients connected to server, Web proxy server (intermediary sources), and the third-party databases [28, 16]. These data can be collected either in an implicit or an explicit manner [22].

Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. This helps to study user's behavior at the website [22]. *On the other hand,* ***Explicit data*** comes from registration forms which the user Fill's while signing up with the website. The user rating questionnaires. Additional data such as demographic (i.e. Study of the characteristics of visitors) and application data (example: e-commerce transactions) can also be used. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages [22].

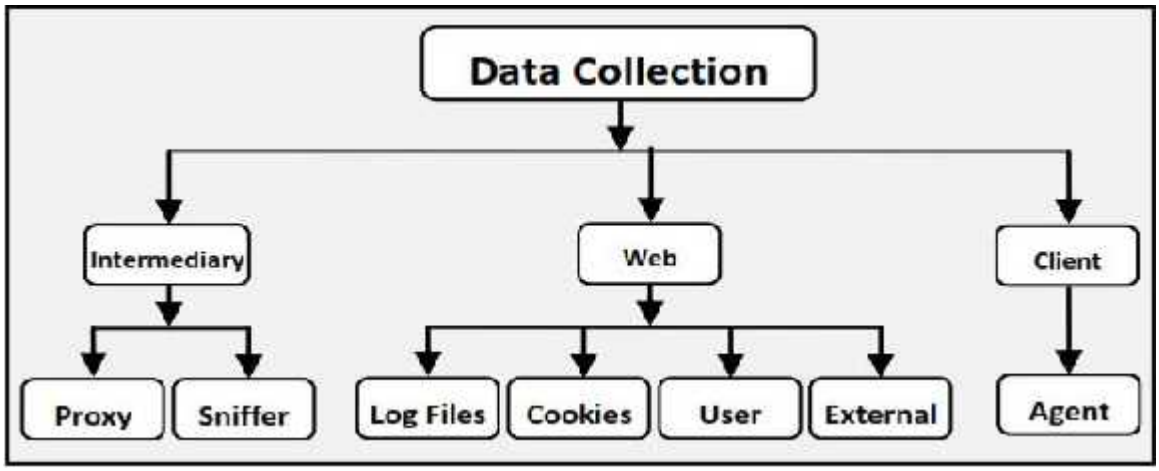

Figure 2.4: Sources for collecting web usage data through web server [28].

2.5.1. Challenges of data collection

Table.2.1 summarizes drawbacks of various data sources, such as web server side, client side and intermediary.

Server-side logs: are considered as a major data source for discovering knowledge about website usage. These logs generally supply the most complete and accurate usage data, but their two draw backs are [16]:

- ► These logs contain sensitive, personal information, therefore the server owners usually keep them closed.
- ► The logs do not record cached pages visited. The cached pages are summoned from local storage of browsers or proxy servers, not from Web servers.

Cookie: is Unique ID generated by web server that is copied as a small file (<= 4KB) on client machine, server also records it for identifying the client later. Cookies are also capable of storing other usage data, but their small size delimits the possible benefits. When client accesses the same website again from same machine, browser reads that Unique ID and sends it to server. Thus server identifies its user with the help of Unique ID that was assigned to user when he visited the website last time. Cookies suffer from two major problems [28];

- ► This approach fails if client disables the cookies in its browser.
- ► Cookies are intended to work with browser on a host, therefore server may misinterpret a user when several users use same computer and visit the same website.

Explicit User: Input data is explicitly provided by users by means of registration forms. But this approach cannot be considered as good because [28]:

- ► It incurs extra burden on user, and it discourages users to visit the website.
- ► Information produced through this approach cannot be trusted because users tend to reveal minimum possible personal information due to privacy issues.

External data: is the data obtained from third party, who maintains user information in its database. But this approach may not suit to privacy and security norms in several countries [28].

Client-side logs: Participants remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose. These are pieces of information generated by a Web server and stored in the users' computers, ready for future access. The drawbacks of this approach are [16]:

- ► The design team must deploy the special software and have the end-users install it.

- ► This technique makes it hard to achieve compatibility with a range of operating systems and Web browsers.

Proxy-side logs: A proxy server takes the HTTP requests from users and passes them to a Web server; the proxy server then returns to users the results passed to them by the Web server. The two disadvantages are [16]:

- ► Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction.
- ► The request interception is limited, rather than covering most requests.

The proxy logger implementation in Web Quilt, a Web logging system, can be used to solve these two problems, but the system performance declines if it is employed because each page request needs to be processed by the proxy simulator [16].

Web server side	
Source	**Drawbacks**
Server log files	• Web-cache • IP-Address misinterpretation
Cookies	• User misinterpretation • May be disabled due to security and privacy issues
Explicit user input	• Discouraging from user's point of view • User may not provide correct and sufficient data
External data	• Faces legal obstacles • Privacy may be compromised
Client side	
Source	**Drawbacks**
Software agent	• May degrade system performance on client side • Very hard for users to accept such agents on their end
Intermediary	
Source	**Drawbacks**
Packet sniffer	• Since data is not logged and hence it may be lost forever • TCP/IP packets may not arrive in sequence • Information cannot be obtained from encrypted packets

Table.2.1: Drawbacks of various data sources

2.6. Data Pre-Processing

The data collected during first stage is usually diverse and voluminous. Therefore it is necessary to preprocess it by filtering unnecessary and irrelevant data, predicting and filling in missing values, removing noise, transforming it into more useful format, and resolving

the inconsistencies [28]. As shown in Figure 2.5, data preprocessing includes data cleaning, user identification, session identification and path completion [32].

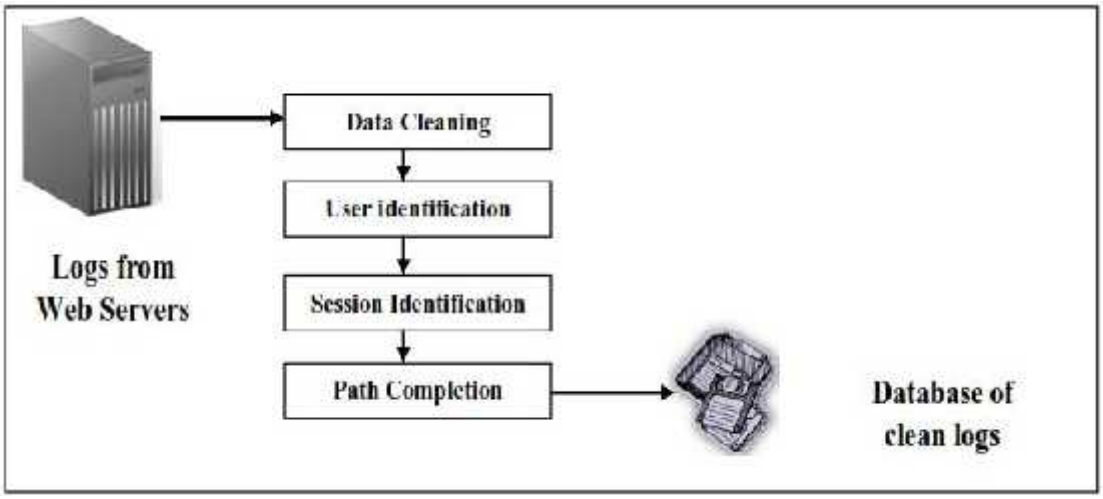

Figure 2.5: Web data preprocessing [32]

2.6.1. Data Cleaning

Data cleaning refers to the process of deleting the data irrelevant to mining log algorithms in web server logs. This is necessary for improving the mining efficiency and optimize memory space usage. As shown in figure 2.6, data cleaning includes elimination of local and global Noise, removal of records of graphics, videos and the format information; removal of records with the failed HTTP status code, robots cleaning [32].

Elimination of Local and Global Noise

Lathwal and Dhawan [32], note that, Web noise is divided into two categories depending on its granularities: Global Noise and Local Noise. Unnecessary objects with high granularities which are larger than individual pages correspond to global noise. This noise includes mirror sites, duplicated Web pages and previous versioned Web pages. Local noise, also called inter-page noise, includes irrelevant items inside a Web page. This noise includes navigational guides, decoration pictures, banner ads etc. It is necessary to remove this type of noise for better results. The local noise also deals with the user background knowledge can be discovered from user's local information collections, such as a stored documents, browsed web pages, and emails [32].

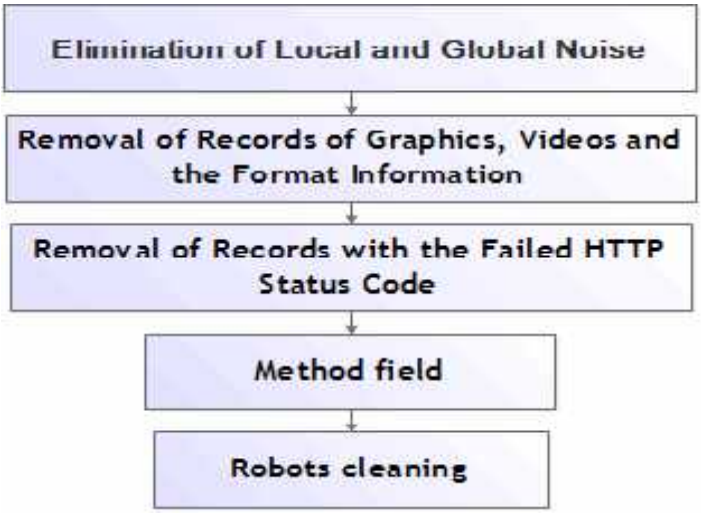

Figure 2.6: Steps in Data Cleaning [32].

The records of graphics, videos and the format information

The records with filename extension of JPEG, GIF, CSS and so on, which can be found in the URI field of the every record, can be removed from the log file. The files with these extensions are not actually the user interested web page; rather it is just the documents embedded in the web page. So it is not necessary to include these files in identifying the user interested web pages [28]. The unnecessary evaluation is eliminated and this process helps in fast identification of user interested patterns.

The records with the failed HTTP status code

The HTTP status code is then considered in the next process for cleaning. In this step the status field of every record in the web access log is examined and the records with status codes over 299 or below 200 are removed. This cleaning process will reduce the evaluation time for determining the users' interested patterns [25].

Method field: Records having value of POST or HEAD in Method field are used in present study for acquiring more accurate referrer information [32].

Robots cleaning: A Web Robot (WR), also called spider or bot, is a software tool that periodically scans a website to extract the content. Web robots automatically follow all the hyperlinks from current web page. Search engines such as Google; use WRs to gather all

the pages from a website in order to update their search indexes. The number of requests from one web robot may be equal to the number of web site's URIs. If the web site does not attract many visitors, then the number of requests coming from all the web robots that have visited the site might exceed that of human generated requests [32].

2.6.2. User Identification

User identification is the process of identifying each of the different users accessing the Web site. Goal of user identification is to mine ever y user's access characteristic, and then make user clustering and provide personal service for the users.

Rules for user identification are [32]:

- Different IP addresses refer to different users.
- The same IP with different operating systems or different browsers should be considered as different users.
- While the IP address, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before, according to the topology of the site.

2.6.3. Session Identification

Session identification defines the number of times the user has accessed a web page. Session identification takes all of the page references for a given user in a log and breaks them up into user sessions. These sessions can be used as data vectors in classification, prediction, clustering and other tasks. Traditional session identification algorithm is based on a uniform and fixed timeout. While the interval between two sequential requests exceeds the timeout, new session is determined [32]. According to some related researches, the value of timeout can be set as 25.5 minutes [32].

2.6.4. Path Completion

It is necessary to determine the existence of important accesses that are not recorded in the access log. Path completion refers to the inclusion of important page accesses that are missing in the access log due to browser and proxy server caching. Similar to user identification, the heuristic assumes that if a page that is requested by the user is not directly linked to the previous page accessed by the same user, the referrer log can be referred to see

from which page the request came. If the page is in the user's recent click history, it is assumed that the user browsed back with the "back" button, using cached sessions of the pages. So each session reflects a full path, including the pages that have been backtracked [32].

2.7. Pattern Discovery and Pattern Analysis

Pattern Discovery is the key component of the web mining. Pattern discovery covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition [24]. Moreover, pattern discovery is about looking for interesting usage patterns out of the prepared log data. This is accomplished using algorithms like sequential pattern generation, association rule mining or clustering. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the web access log. Visualization can also be used in web usage mining, and it presents the data in the way that can be understood by users more easily [15, 40].

2.7.1. Statistical Analysis

Statistical analysts perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. Many web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. By analyzing the statistical information contained in the periodic web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions.

2.7.2. Association Rules

In the web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions.

In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. Association rule mining has been well studied in Data Mining, especially for basket transaction data analysis. Aside from being applicable for e-Commerce, business intelligence and marketing applications, it can help web designers to restructure their web site. The association rules may also serve as heuristic for pre fetching documents in order to reduce user-perceived latency when loading a page from a remote site.

2.7.3. Clustering

Clustering is a technique to group together a set of items having similar characteristics. Clustering can be performed on either the users or the page views. Clustering analysis in web usage mining intends to find the cluster of user, page, or sessions from web log file, where each cluster represents a group of objects with common interesting or characteristic. User clustering is designed to find user groups that have common interests based on their behaviors, and it is critical for user community construction. Page clustering is the process of clustering pages according to the users' access over them. Clustering of pages will discover groups of pages having related content. This information is useful for the Internet search engines and Web assistance providers.

2.7.4. Classification

Classification is the technique to map a data item into one of several predefined classes. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc.

2.7.5. Sequential Pattern

This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis. A sequential pattern discovery analysis in web usage

mining is discovering sequential patterns from web log files has been proposed that provides behavioral marketing intelligence for e-commerce scenarios.

2.7.6. Dependency Modeling

Dependency modeling is another useful pattern discovery task in web mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the web domain. As an example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (i.e., from a casual visitor to a serious potential buyer. There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Modeling of Web usage patterns will not only provide a theoretical framework for analyzing the behavior of users but is potentially useful for predicting future Web resource consumption.

Pattern Discovery is followed by Pattern Analysis. The goal of pattern analysis is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transformed to a format that can be assimilated easily. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations. All these methods assume the output of the previous phase has been structured.

2.8. Pattern Application of Web Usage Mining

The navigation patterns discovered can be applied to the following major areas, among others: i) improving the page/site design, ii) making additional product or topic recommendations, iii) Web personalization, and iv) learning the user or customer behavior [16].

Each of the applications can benefit from patterns that are ranked by their interestingness for the subject under consideration. Web usage mining is used in the following areas [33]:

First, Web usage mining offers users the ability to analyze massive volumes of click stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs.

Second, Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages. Mining that determines frequent access behavior. This can be achieved using web usage mining and identify users needed links to improve the overall performance of future accesses, i.e. to improve the actual design of Web pages, the attractiveness of a Web site, in terms of content and structure and to make other modifications to the site.

Third, Web usage patterns can be used to gather business intelligence to improve Customer attraction, Customer retention, sales, marketing and advertisement, cross sales.

- Mining of web usage patterns can help in the study of how browsers are used and the user's interaction with a browser interface.
- Usage characterization can also look into navigational strategy when browsing a particular site.
- Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web.
- Performance and other service quality attributes are crucial to user satisfaction and high quality performance of a web application is expected.
- Web usage mining of patterns provides a key to understanding Web traffic behavior, which can be used to deal with policies on web caching, network transmission, load balancing, or data distribution.
- Web usage and data mining is also useful for detecting intrusion, fraud, and attempted break-ins to the system.
- Web usage mining can be used in e-Learning, e-Business, e-Commerce, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, Digital Libraries Physical Sciences, Social Sciences, Engineering, Medicine, and Biotechnology.
- Web usage mining can be used in Customer Relationship Management, Manufacturing and Planning, Telecommunications and Financial Planning.

- Web usage mining can be used in Counter Terrorism and Fraud Detection, and detection of unusual accesses to secure data.
- Web usage mining can be used in determination of common behaviors or traits of users who perform certain actions, such as purchasing merchandise.
- Web usage mining can be used in usability studies to determine the interface quality.
- Web usage mining can be used in network traffic Analysis for determining equipment requirements and data distribution in order to efficiently handle site traffic.

2.9. Related Works

During the last few years, local and other researchers have proposed a new unifying area for all methods that apply data mining to Web data, named Web mining. Different modes of usage or 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 [33]. These profiles can later be harnessed towards personalizing the Web site to the user.

Ciesielski and Lalani [34] presented a method to create user profile that is made up of three stages. The first step is to identify the useful data. In second step, K-means algorithm is used for user session clustering. The results show that this method is an effective role to improve the user model. In their work they explain how to understand “who” the users were, “what” they looked at, and “how their interests changed with time, “when” they visit all of which are important questions in Customer Relationship Management (CRM). In their study they present clustering the user profiles. They also describe how the discovered user profiles can be enriched with explicit information.

Research by Cutrell et al., [35] combines keyword and property-value search, allowing users to find information based on whatever they may remember. Among problems and research directions being addressed is whether these designs can be extended to include ’non-personal’ content (information sources of interest that users are not familiar with, such as news and intranet). Authors’ identified three research areas to be explored in

personalization include recommendations, information filtering, and personalized presentation. These three areas are relevant but fairly big, therefore future research direction is suggested to look into personalized presentation based on the user's profile from his searching/navigation activities.

Getahun [14] has conducted a research on a web Usage Pattern Discovery and analysis by region of Ethiopian Airlines official website. Getahun has utilized the server log files and followed the steps: Data collection, Data preprocessing, Pattern discovery and Pattern analysis in his study. He used FP-Growth algorithm for pattern discovery. Moreover, he utilized WUMprep tool, Java programming, Perl and MS Excel were used for the preprocessing task, Google Analytics and MS Excel for statistical analysis and Weka for pattern and association rule discovery. He had described the website statistically, discovered patterns and sequences. His findings show In general, Sheba Miles, Contacts, Network and Baggage related pages are the most frequently accessed pages by the customers in most of the regions. Hence, further attention should be given for these categories in order to make the pages in each category more informative and easily accessible. Moreover, minor restructuring should be made in order to rearrange them according to their access pattern and frequency. He recommended for future researchers need to consider integrated data obtained from Web Server, Proxy Server and Client and also sequential pattern discovery need to consider in order to clearly show the time sequences of access; that is, the order in which the pages are accessed.

Tadele [12] has conducted a research on a web Usage Pattern Discovery of Addis Ababa University website. Tadele has utilized the server log files and followed the steps: raw log preparation, pattern discovery and pattern analysis in his study. He used Apriori algorithm for pattern discovery. Moreover, he utilized Wekaprep tool and python code for data preparation, Mach5 for statistical analysis and Weka for pattern discovery. He had described the website statistically, discovered patterns and sequences. His findings show the daily access trends, top entry and exit pages and pages that cause errors. His findings also show many useful relationships among different pages of the website. According to his study, most users access the university website either directly or via search engines. Finally, he forwarded recommendations for the website to be restructured in a user friendly manner

and make the pages' access as error free as possible. He recommended for future researchers to include error logs and SSL logs in the dataset.

Awet [13] has conducted his study on exploring the navigational behavior of users of Addis Ababa University (AAU) official website. Awet has utilized the server log files and followed the steps: data preprocessing, pattern discovery and pattern analysis in his study. He used Apriori algorithm for pattern discovery. Moreover, he utilized WUMprep tool for data preprocessing, Web Utilization Miner for statistical analysis and pattern discovery. In his research, he has described users' access behavior of the website with different parameters such as navigational pattern, most requested pages, top entry pages, top exit pages, referrer pages, etc. His findings show that the home page of AAU's website is the most frequently accessed and used page. He has also concluded that this same page is the top exit page which indicates that AAU website needs improvement in terms of its content. He also indicated which pages are accessed frequently after home page. According to Awet's study AAU website is reached mostly either by directly typing the URL or via search engine which entails that there is less referral to AAU website from other websites.

The above local researches only focus on how to meet customers' information needs from the perspective of functional information, and they exclude most of the other information needs from their consideration [12, 13, 14]. This initiates the current study which aims to construct user profile based on user' navigational behavior.

CHAPTER THREE

RESEARCH METHODS AND TECHNIQUES

Mining user profiles from vast amounts of historical data stored in server access logs is a possible approach to personalization. Web mining techniques have three operations of interests: clustering (finding natural groupings of users, pages etc.), associations (which URLs tend to be requested together), and sequential analysis (the other in which URLs tend to be accessed).

The study aims to describe an approach to usage-based user profile construction. In particular cluster analysis is used in order to obtain the common interest of user and association analysis to identify URLs tends to be requested together for the purpose of generating user profile.

3.1. Framework of the Study

A general framework for creating user profile based on web usage mining is depicted in Figure 3.1 [15]. This framework distinguishes between the tasks of data preparation and usage mining. This research considers both tasks of data preparation and usage mining in order to obtain user profile. The data preparation tasks result in aggregate structures such as a user transaction file capturing meaningful semantic units of user activity to be used in the mining stage.

The typical steps of creating user profiles correspond to the web usage data mining steps described in chapter two. Generally, User profiles creation process based on web usage mining can be divided into five distinct phases [16]. This framework divide the five phase in to two major tasks. The first two phases (collection of web data and preprocessing of web data) into data preparation and the last three phases (such as pattern discovery, pattern analysis and visualization and pattern applications for formulating user profile) into usage mining respectively.

All these phases, tools and technologies that were used during the web log preparation as well as pattern discovery, and pattern analysis are discussed in subsequent sections.

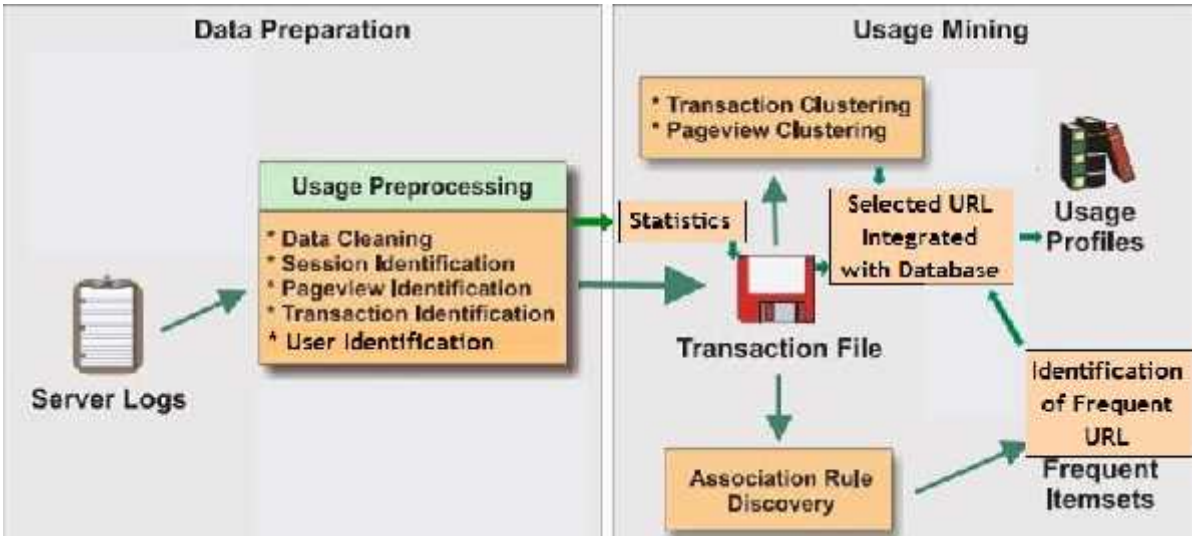

Figure 3.1: A Framework for creating user profile Based on Web Usage Mining [37]

3.2. Collection of Web data

As regards this research, usage data of the recent log files of 150 days which spans over five months' was used, i.e. collected from the web server of Ebiz Online Solutions PLC that official hosts the web site of Ebiz Online Solutions PLC. The overall dataset used is very large (over 6.2 GB) and is sufficient for pattern discovery. Figure 3.2 shows a sample of the information contained in the log file.

213.55.104.254 - - [25/Jul/2014:15:14:04 +0100] "GET /images/mtree/listings/m/3777.jpg HTTP/1.1" 200 9076 "https://www.google.com.et/" "Mozilla/5.0 (Windows NT 6.1) Apple Web Kit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.125 Safari/537.36"

82.145.217.150 - - [25/Jul/2014:15:22:22 +0100] "GET /favicon.ico HTTP/1.1" 200 1459 "http://tender.2merkato.com/tenders/view/64494/53d1accb-d670-415a-9ddb-4dced447fba2" "Opera/9.80 (Android; Opera Mini/7.5.35199/35.3527; U; en) Presto/2.8.119 Version/11.10"

Figure 3.2: Sample web Data of 2merkato.com

3.3. Preprocessing of Web data

The pre-processing steps consist of converting the usage data available into the data abstraction necessary for pattern discovery. As shown in Figure 3.3 data preprocessing includes data cleaning, user identification, session identification, page view identification and transaction identification.

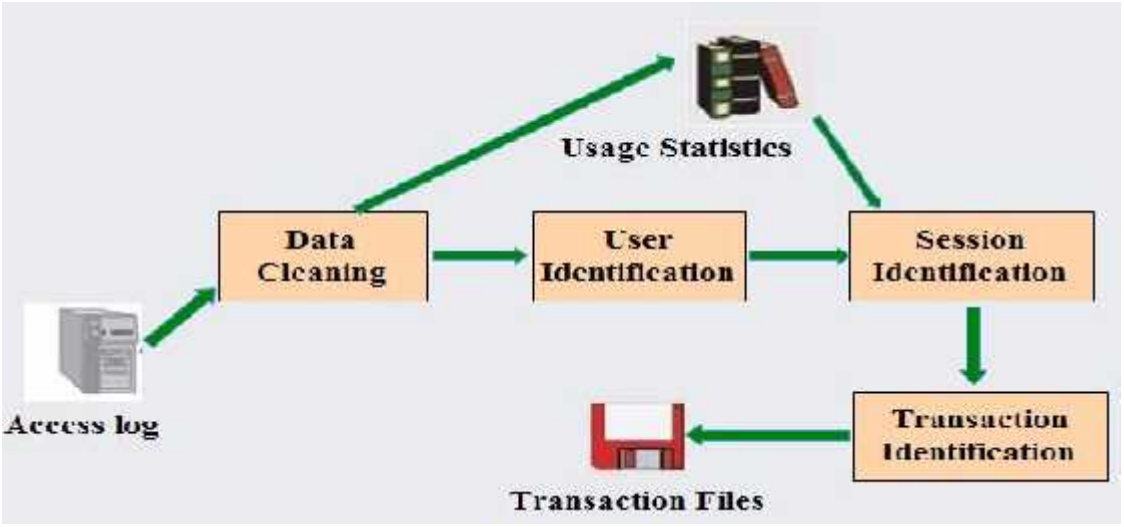

Figure 3.3: Steps in Data Preprocessing for Web Usage Mining [37]

3.3.1. Data cleaning

Data cleaning includes elimination of local and global noise, removal of records of graphics, videos and the format information; removal of records with the failed HTTP status code, robots cleaning [32].

In this research the data cleaning module is intended to clean Web log data by deleting irrelevant and useless records in order to retain only usage data that can be effectively exploited to recognize users' navigational behavior. Since Web log files record all user interactions, they represent a huge and noisy source of data, often comprising a high number of unnecessary records. By removing useless data, we can reduce the size of these files in order to use less storage space and to facilitate upcoming steps. Of course the choice of log data to be removed depends on the ultimate goal of the research. In this case, the goal is to create user usage profiles individual as well as group user usage profiles to offer personalized dynamic links to the site's visitors; hence the system has to keep only log data concerning explicit requests that actually represent users' actions. As a consequence, the data cleaning module has been intended to remove the following requests:

- Access method: this study consider requests with access method "GET". Generally, requests containing a value different from "GET" in the field of the access method do not refer to explicit requests of users but they often concern with CGI accesses,

properties of the Server, visits of robots, etc. Hence, these requests are considered non-significant and, consequently, they are removed from the log file.

- Failed and corrupted requests. These requests are represented by records containing a HTTP error code. A status with value of 200 represents a succeeded request. A status with value different from 200 represents a failed request (e.g. a status of 404 indicates that the requested file was not found at the expected location). Also corrupted lines with missing values in some fields are eliminated in order to clean log files from incomplete information.
- Requests for multimedia objects. Due to the model underlying the HTTP protocol, a separate access request is executed for every file, image, multimedia object embedded in the requested Web page. As a consequence, a single request for a Web page may often produce several log entries that correspond to files automatically downloaded without an explicit request of the same user. Requests for such type of file can be easily identified since they contain a particular URL name suffix, as showed in table 3.1 such as gif, jpeg, jpg, and so on. Whether to keep or remove requests for multimedia objects depends on the kind of Web site to be personalized and on the purpose of the WUM system. In general, these requests do not represent the effective browser activity of the user visiting the site; hence they are deemed redundant and are removed.

Format	Description
\.ico	A file format used for icons in the operating system.
\.gif	A popular format for image files, with built-in data compression.
\.jpg	A file extension indicating a file of JPEG file format; i.e., a digital
\.jpeg	A file format commonly used for image compression; An image file in that format.
\.css	This is a document format which provides a set of style rules which can then be incorporated in an XHTML or HTML document.

Table 3.1: Irrelevant requests, (Extension of URL).

- Requests originated by Web robots. Log files may contain a number of records corresponding to requests originated by Web robots. Web robots (also known as Web crawlers or Web spiders) are programs that automatically download complete Web sites by following every hyperlink on every page within the site in order to update the

index of search engine. Requests created by Web robots are not considered usage data and, consequently, have been removed.

The algorithm depicted in figure 3.4 shows the steps performed on Web Log file in Cleansing stage.

Input: Web Server Log file

Output: Cleaned Log Data Records

Step1: Read Log Record from Web Server Log File

Step2: If (Log Record .url-contains (.gif.jpeg, jpg, css)) AND (found different errors like HTTP 404 or more) then

Remove from web log file.

End of If condition.

Step3: Repeat the above two steps until EOF (Web Server Log File)

Figure 3.4: Algorithm for Data Cleansing.

The algorithm checks in the log data records whether the URL contains .gif, .jpeg, .jpg .css and if found different errors like HTTP 404 or more it remove the log record from web server log file. The action is repeated until it checks all records in the web server log file.

3.3.2. User identification

In this study a client side tracking mechanism is not used, only the IP address, agent and server click stream are available to identify users and server session. In the data preparation stage, the heuristics proposed in Dhawan and Lathwal [32] is used to identify unique user. In this research the rules used for user identification are the following.

- Different IP addresses refer to different users.
- The same IP with different operating systems or different browsers should be considered as different users.
- While the IP address, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before, according to the topology of the site.

Input: Processed Web Log File.

Output: Number of Distinct User.

Step1: User's IP addresses of two consecutive entries are compared.

Step2: If (IP address is same) then

Check user's browser and operating system

If both user's browser and operating system are same then

Consider same user.

Else

Consider new user.

End if

Else

Consider new user.

End if

Step3: Repeat step 1 & step 2 until EOF (Web Log File).

Figure 3.5: Algorithm for User Identification.

The algorithm compares two consecutive cleaned log data records and User's IP addresses for the purpose of user identification. If IP address of the records is the same, then it checks for user's browser and operating system. If both user's browser and operating system are the same, then it considers the two records as same user; otherwise, it consider as new user. However if User's IP addresses is not the same considers the new user.

3.3.3. Session identification

User session can be defined as a set of pages visited by the same user within the duration of one particular visit to a website. It is necessary to divide the log entries of a user into multiple sessions through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user has started a new session. In this research a 30-minute default timeout is considered [15]. Hence the log file, after user identification, may be further divided into sessions for every user. Hence each user's page visits has been split into one or more sessions.

3.4. Pattern Discovery and Analysis of Web data

In this step machine learning or Data Mining techniques are applied to discover interesting usage patterns and statistical correlations between web pages and user groups.

Statistical techniques are the most common methods to extract knowledge about visitors of a Web site. Before considering a particular method for the analysis of sessions file different kinds of descriptive statistical tools (frequency, mean, median etc.) on variables such as page views, viewing time and length of a navigational path can be performed.

In this study WebLog Experts tools is used for web traffic analysis to produce a periodic report containing statistical information such as the most frequently accessed page, average view time of a page or average length of a path through a site. This type of knowledge can be potentially useful for improving the system performance, facilitating the site modification task, and providing support for marketing decision.

This study applies Web Usage Mining for two purpose: The first is general access tracking which is used to analyze the web logs to understand access patterns and trends. The other is Customized usage tracking for constructing user profile as a base to customize web sites to users.

A number of different approaches have been developed dealing with specific aspects of Web Usage Mining for the purpose of automatically discovering user profiles. One of this is cluster analysis. In particular this study considered a usage based creating user profile taking into account similar sessions in terms of viewed pages [15].

3.4.1. Pattern Discovery

The aim of this study is to therefore find out behavioral profiles on the base of grouping similar sessions in terms of viewed pages and association to identify which URLs accessed together.

There are ample of data mining algorithms available which can be used for data mining. Some of the popular data mining tasks applied in this study are *clustering and association rule discovery.*

Clustering

Clustering algorithms find groups of items which are similar. They divide a data set so that records with similar content are in the same group, and groups are as different as possible from each other. Since the categories are unspecified, this is sometimes referred to as unsupervised learning [27].

Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-commerce application or provide personalized web content to the user. On the other hands, clustering of pages will discover groups of pages having related content.

K-means Clustering Algorithm

K-means is one of the common data mining clustering algorithms which can be used for clustering or grouping the items having similar characteristics. In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \leq n$) $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) [43]:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where μi is the mean of points in Si.

In general, K-Means algorithm works as follows [43]:

Step 1. Place K points into the space of the objects being clustered. They represent the initial group centroids.

Step 2. Assign each object to the group that has the closest centroid.

Step 3. Recalculate the positions of the K centroids.

Step 4. Repeat Steps 2 & 3 until the group centroids no longer move.

In this study k-means algorithm is used because it is very simple and easy to understand and also it follows a very easy way for classifying the items given. Moreover the researcher knows the number clusters in advance.

Association rule discovery

In this study two association rule discovery algorithms are experimented; Apriori algorithm and FP growth.

Apriori Algorithm

Apriori algorithm is a popular data mining algorithm or technique which can be used to finding or discovering association rules or frequent item sets. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database [40].

To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space. Apriori property states that, all nonempty subsets of a frequent itemset must also be frequent [40].

As presented in Algorithm 3.1, in Apriori algorithm involves two steps [40]: the join step and the prune step. During the join step Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself. This set of candidates is denoted Ck. Ck is a superset of Lk, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in Ck. A scan of the database to determine the count of each candidate in Ck would result in the determination of Lk (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to Lk). Ck, however, can be huge, and so this could involve heavy computation. To reduce the size of Ck, the Apriori property is used as follows. Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1)-subset of a candidate k-itemset is not in Lk-1, then

the candidate cannot be frequent either and so can be removed from Ck. This subset testing can be done quickly by maintaining a hash tree of all frequent itemsets.

Association rules can be generated from frequent itemsets as follows [40]:

- For each frequent itemset l, generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule:

  "*s => (l-s)*" if $\frac{\textit{Support count (l)}}{\textit{Support count (s)}} \geq \textit{min conf}$,

  Where min conf is the minimum confidence threshold.

Algorithm 3.1: The Apriori algorithm

Input:

- D, database of transactions;
- min_sup, the minimum support count threshold

Output: L, frequent itemsets in D

Method:

```
(1)  L1=find_frequent_1-itemsets(D);
(2)  for(k=2; Lk-1!=null;k++){
(3)      Ck=apriori_gen(Lk-1);
(4)      for each transaction t Є D{ // scan D for counts
(5)          Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)      for each candidate c Є Ct
(7)          c.count++;
(8)  }
(9)    Lk={c Є Ck | c.count≥min_sup}
(10)       }
(11)       Return L=UkLk

Procedure apriori_gen(Lk-1:frequent(k-1)-itemsets)
      (1)  for each itemse l t 1 Є Lk-1
      (2)  for each itemset l2 Є Lk-1
      (3)   if(l1[1]=l2[1])^(l1[2]=l2[2])^…^(l1[k-2]=l2[k-2])^(l1[k-1]<l2[k-1])
      then{
      (4)          c=l1xl2; //join step: generate candidates
      (5)       if has_infrequent_subset(c,Lk-1) then
      (6)          delete c; //prune step: remove unfruitful candidate
      (7)       else add c to Ck;
      (8)          }
      (9)       Return Ck;

   procedure has_infrequent_subset(c: candidate k-itemset; Lk-1: frequent (k-1)-itemsets);
   //use priori knowledge
      (1)  for each (k-1)-subset s of c
      (2)      if s !Є Lk-1 then
      (3)       Return TRUE;
      (4)  Return FALSE;
```

Apriori algorithm is costly both space-wise and time-wise [40].

The frequent item sets or patterns found from this algorithm can be used to analyse the web user trends or interest or behavior. The popular application of this algorithm is “market basket analysis” This is a classic algorithm which can be used in data mining domain for purpose of learning association rules [14, 43].

FP Growth Algorithm

FP-Growth algorithm one of the popular and efficient data mining algorithms which can used for finding the association rules or frequent item sets or interesting patterns. To overcome the inefficiency of Apriori, an algorithm named frequent pattern growth (FP-Growth) is introduced. FP-Growth adopts a divide-and-conquer strategy. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree, which retains the itemset association information. It then divides the compressed database into a set of conditional databases (a special kind of projected database), each associated with one frequent item or “pattern fragment,” and mines each such database separately. With FP-Growth algorithm, no candidate generation is required [40].

As presented in Algorithm 3.2, in FP-Growth algorithm, the first scan of the database is the same as Apriori, which derives the set of frequent items (1-itemsets) and their support counts (frequencies). The set of frequent items is sorted in the order of descending support count. This resulting set or list is denoted L. An FP-tree is then constructed as follows. First, create the root of the tree, labeled with “null.” Scan database D a second time. The items in each transaction are processed in L order (i.e., sorted according to descending support count), and a branch is created for each transaction [40]. In general, when considering the branch to be added for a transaction, the count of each node along a common prefix is incremented by 1, and nodes for the items following the prefix are created and linked accordingly [40].

To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links. In this way, the problem of mining frequent patterns in databases is transformed to that of mining the FP-tree [40].

A study on the performance of the FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm

[40]. The main advantage of FP-Growth algorithm is that it uses compact data structure and avoids repeated database scan.

Algorithm 3.2: The FP-growth algorithm for discovering frequent itemsets without candidate generation works as follows [40];

Input: *D*, a transaction database; and *min sup*, the minimum support count threshold.
Output: The complete set of frequent patterns.
// Construct FP-tree
Scan the transaction database *D* once. Collect *F*, the set of frequent items, and their support counts.
Sort *F* in support count descending order as *L*, the *list* of frequent items.
Create the root of an FP-tree, and label it as "null."
For each transaction *Trans* in *D* do the following.
Select and sort the frequent items in *Trans* according to the order of *L*.
Let the sorted frequent itemlist in *Trans* be [*p*|*P*], where *p* is the first element and *P* is the remaining list.
Call insert tree([*p*|*P*], *T*),which is performed as follows.
If *T* has a child *N* such that *N.item-name=p.item-name*, then
increment*N*'s count by 1;
else
create a new node *N*
let its count be 1
its parent link be linked to *T*
itsnode-link to the nodes with the same *item-name* via the node-link structure.
If *P* is nonempty
callinsert_tree(*P*, *N*) recursively.
// The FP-tree is mined by calling FP growth(*FP tree, null*), which is implemented as follows.
procedure FP growth(*Tree*, a)
if *Tree* contains a single path *P* then
for each combination (denoted as) of the nodes in the path *P*
generate pattern $\beta \cup \alpha$ with *support count = minimum support count of nodes in* β ;
else
for each *ai* in the header of *Tree*
generate pattern $\beta = ai \cup \alpha$ with *support count = ai*:*support count*;
construct β's conditional pattern base and then β's conditional FP tree *Tree*β ;
if *Tree* $\beta \neq \Phi$ then
call FP-growth(*Tree* , β)

In this study FP-Growth algorithm is used in web usage to find frequent item sets from given web server log files. “Web usage mining” includes mining web of data that is web server log data in order to find the *frequent web access patterns or frequently accessed web pages.* This help to analyse the web user behaviour so as to help improve the performance of web applications and also help improve business. Here the frequently accessed web page patterns are discovered which helps to analyse web usage.

By this result we can pre-fetch and keep frequently accessed web page sequence in cache so as to create user profile and reduce the page access time in future.

Comparison of Algorithms

The data mining algorithms like FP-Growth and Apriori algorithms are used here in this study. The performance comparison of both the algorithms in web mining or in generating association rules or finding interesting patterns is done here. In general FP-Growth algorithm is known as an efficient data mining algorithm. Here we have considered two parameters to compare the performance of FP-Growth and Apriori algorithms. One of the two parameters is “time taken” for finding interesting patterns or association rules and the other one is the “number of patterns” found. The Apriori algorithm will take less time than FP-Growth but it will result in very few numbers of interesting patterns. But FP-Growth finds more number of interesting patterns and hence it is the most efficient data mining algorithm in respect of number of interesting patterns than Apriori [14, 43].

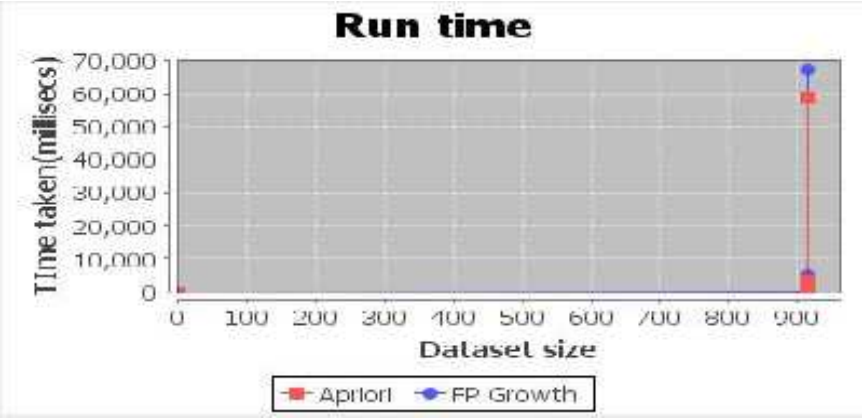

Figure 3.6: Graph showing time taken by both algorithms

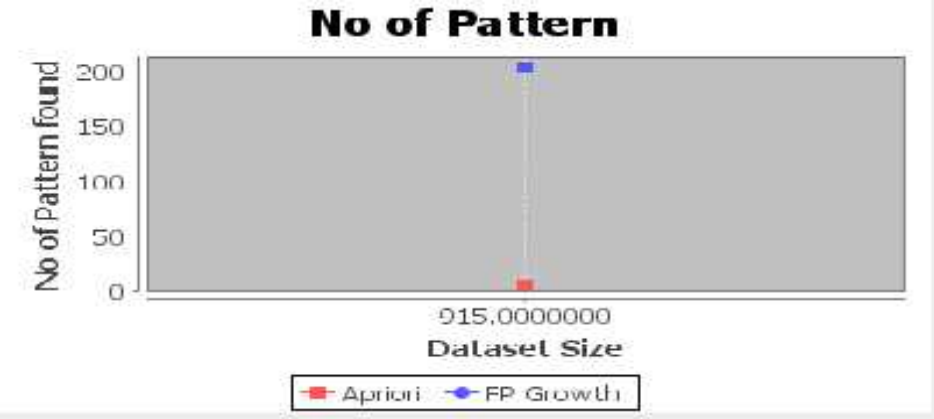

Figure 3.7: Graph showing the number of patterns found by both algorithms

Here we can see the graphs drawn for showing the time taken by both the algorithms for finding the patterns and also the graph for showing the number of patterns found by both the algorithms. The FP-Growth algorithm is more efficient than Apriori algorithm to discover the interesting patterns as it helps to discover huge number of interesting patterns as compared to Apriori algorithm.

3.4.2. Pattern Analysis

The pattern analysis is performed to identify interesting patterns or rules from frequent item sets which are found in previous pattern discovery step. One of the common forms of pattern analysis consists of “knowledge query mechanism” like “SQL”. In general, the techniques used here is “visualization techniques”.

Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

The statistical reports and patterns discovered are analyzed using different techniques such as Literature Review, Website structure and content analysis, business knowledge, discussion with technical persons working on the website, lift interestingness measure, etc.

CHAPTER FOUR

DATA PREPARATION

Preprocessing Web usage data is still imperfect, mainly due to the difficulty to identify users accurately in the absence of registration forms and cookies, and due to log requests that are missing because of caching. A web log file, as an input to the preprocessing phase of WUM, large in size, contains number of raw and irrelevant entries and is basically designed for debugging purpose [38]. Consequently, web log file cannot be directly used in WUM process. Preprocessing of log file is complex and laborious job and it takes 80% of the total time of web usage mining process as a whole [39]. Paying due attention to preprocessing step, improves the quality of data [40], so as to improve the efficiency and effectiveness of pattern discovery.

For this research, the log file containing information about all web requests to the Ebiz's official website from Jul 25, 2014 to Jan 22, 2015 is collected. This was the only log data available on the server as of the access date. Each line in the web usage log file contains information about one web resource request, the time of request, the URL requested, as well as other information (IP, web browser info, etc.). The raw web log file contained over 26,000,000 (twenty six million) web requests. The log file is divided into manageable size which is nineteen different log files each contain on average 1.2 million records.

4.1. Data Preprocessing

In order to prepare the web log data for the mining process, the web log file needed to be cleaned of irrelevant requests, each relevant request needed to be assigned to a visit session, and the resulting file had to be transformed to a format that could be fed into the mining algorithm.

For cleaning the web log file of irrelevant requests and creating sessions in it, WUMprep, a freely available tool for web log data preparation is used, which consists of a set of Perl scripts. This preparation process consist of cleaning data, identifying sessions, identifying users, selecting feature, identifying transactions and transforming the data. These steps are customized from the basic log file preprocessing steps as presented in figure 4.1.

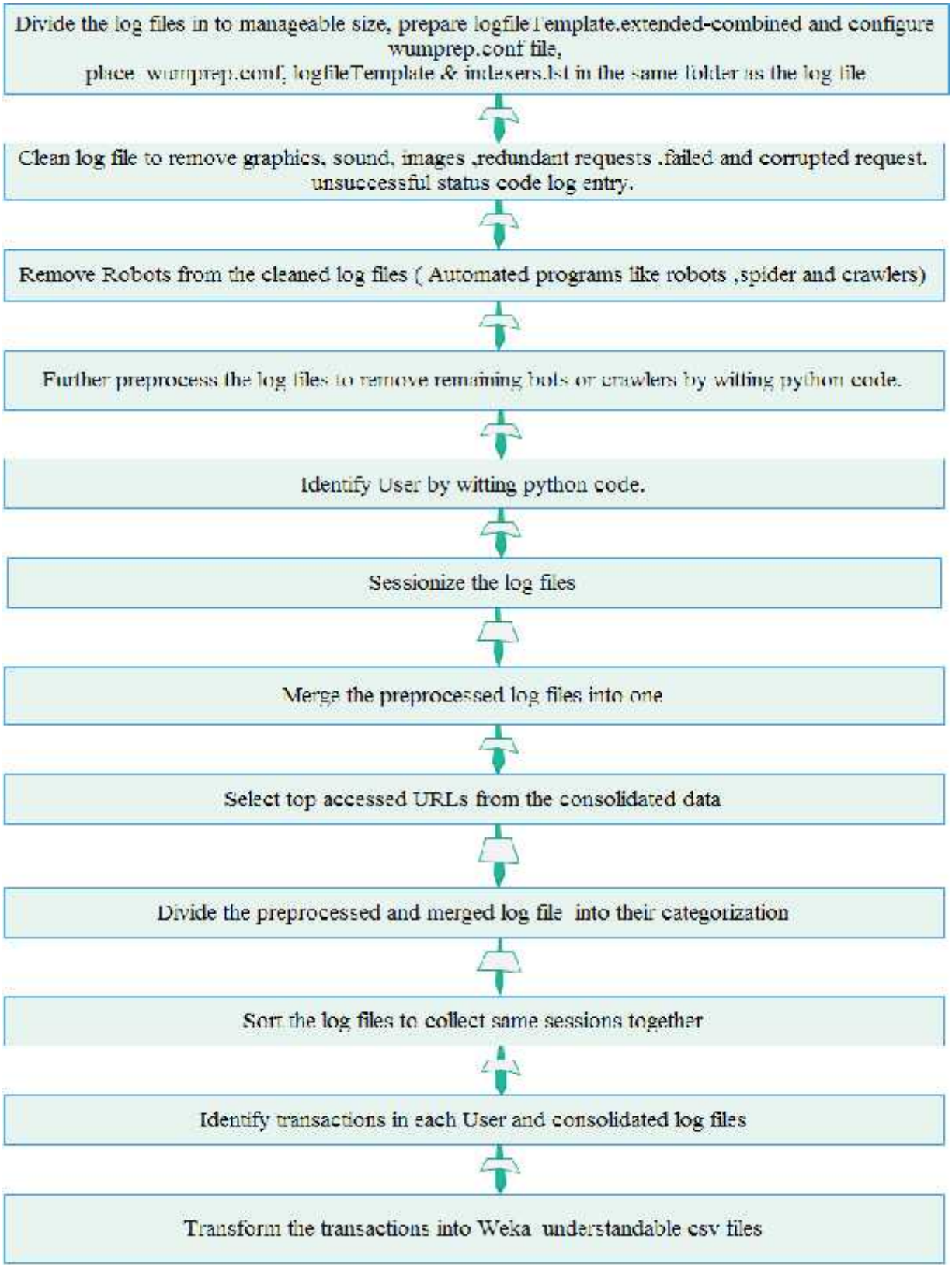

Figure 4.1: Log File Preprocessing Steps

4.2. Removing Irrelevant Requests

The first step in preparing a log file for data mining is removing irrelevant requests such as tensions .ico, .gif, .GIF, .jpeg, .JPEG, .jpg, .JPG, .js, .css, .pdf, .txt, .png, and .flv that are embedded in a web page. For the purpose of finding rules or patterns in web usage data, this study only interested in the pages or documents the visitors visit when traversing a website. In order to remove irrelevant resources from the web log file, the researcher used a WUMprep Perl logFilter.pl scripts, as shown in Figure 4.2 which removed on the average 62 % irrelevant requests from the log file, out of the original 100 %, leaving 38 % web page or document requests.

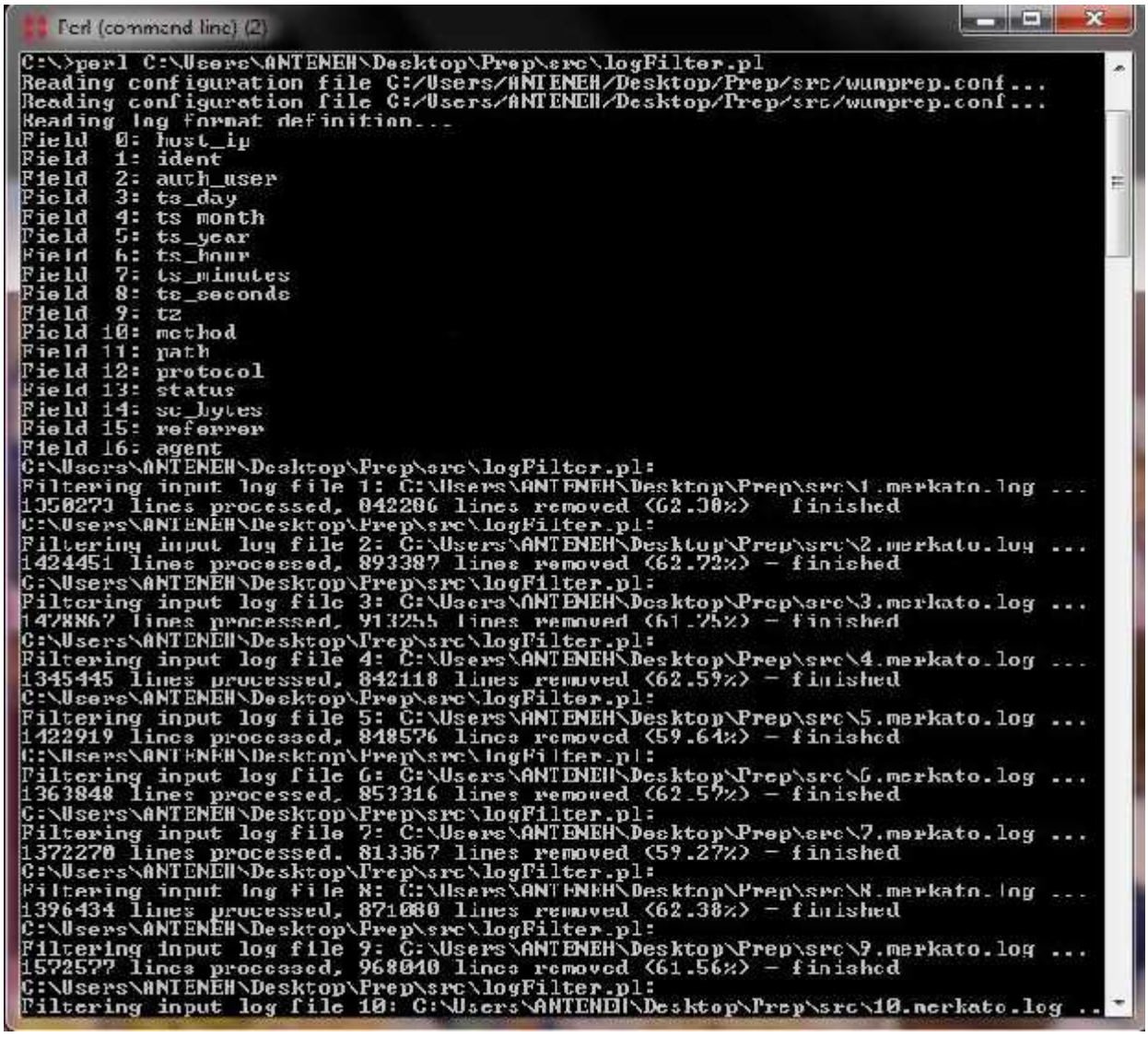

Figure 4.2: Screen shot of the preprocessing for removing irrelevant request

4.3. Removing automatic requests

Various software robots, indexers, and spiders make automated requests while crawling the World Wide Web and creating their own databases. These requests are logged in the web log file but are not representative of the behavior of actual visitors to the website and would make noise in the analysis process.

To remove automatic requests the researcher was used a WUMprep's Perl removeRobots.pl script, which recognizes robot requests based on the configuration (See figure 4.3). However, since the script was somewhat outdated and would not be able to recognize all robot requests, the researcher did additional algorism by python programming language for cleaning of the rest robot requests (For detail see appendix III).

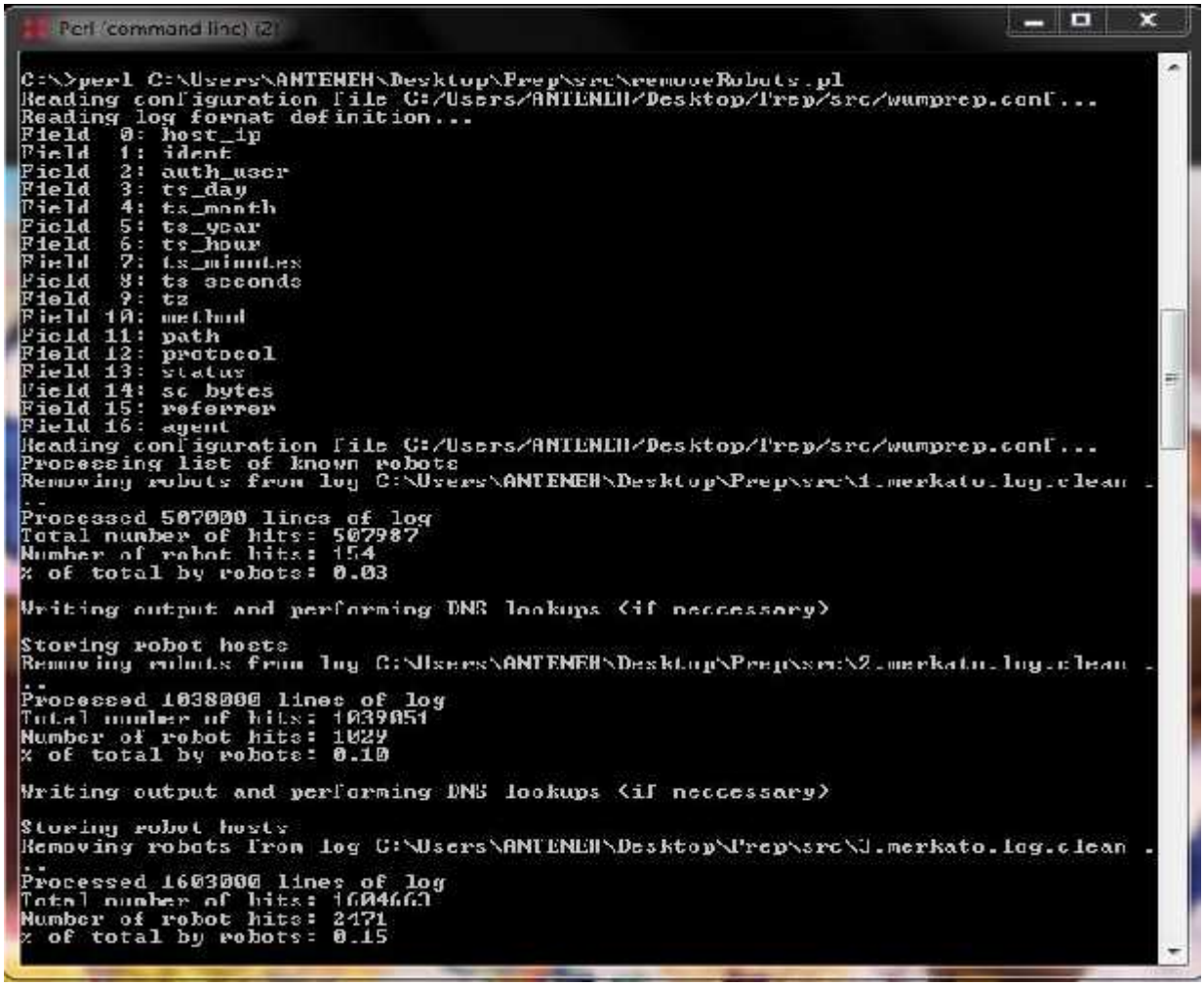

Figure 4.3: Screen shot of the preprocessing in removing automatic request

4.4. Session Identification

The time duration spent on web pages are called Session. To identify the new session the researcher was used a WUMprep's Perl sessionize.pl script which divides the log file into separate user sessions (See figure 4.4). When if the IP address, browser version and operating system are same the referrer information should be taken. A new user session is identified if the URL in the Refer URI – field is a larger interval usually more than 30 minutes between the accessing times on a given record.

Figure 4.4: Sample Screen for Sessionized Log file

4.5. User Identification

In this stage the individual user is identified using their IP address, browser and operating system. The researcher build the algorithm by using python for the User Identification. When the algorism run it starts reading the entry in the sever log file. If the IP address is new then consider it as new user. If the IP address already exists but either the browser or operating system differs then it is also considered as different users. In this study 60,330 users have been identified. Source code written for user identification is given below (for detail see appendix II).

```
if singlhit["agent"] == z["agent"] and newuser > 0:
        if z["agent"] in newclientagent:
            for ip_add, coun in sorted(holduserno.items()):
                 if p == ip_add :
                    out_file.write("%5d\t%s" % (coun ,line))
                    break
                 else:
                     continue
         else:
                 newclientagent.append(singlhit["agent"])
                 user += 1
                 out_file.write("%5d\t%s" % (user ,line))
                 holduserno[p] = user
                 newuser += 1
elif singlhit["agent"] == z["agent"]:
        if z["agent"] in newclientagent:
            for ip_add, coun in sorted(holduserno.items()):
                if p == ip_add :
                    out_file.write("%5d\t%s" % (coun ,line))
                else:
                     continue
        else:
            newclientagent.append(singlhit["agent"])
            user += 1
            out_file.write("%5d\t%s" % (user ,line))
            holduserno[p] = user
            newuser += 1
else:
        if z["agent"] in newclientagent:
            for ip_add, coun in sorted(holduserno.items()):
                if p == ip_add :
                    out_file.write("%5d\t%s" % (coun ,line))
                else:
                    continue
         else:
            user += 1
            newclientagent.append(singlhit["agent"])
            out_file.write("%5d\t%s" % (user ,line))
            holduserno[p] = user
            newuser += 1
```

Figure 4.5 below shows part of the result produced by the above implementation in User Identification.

Figure 4.5: The result of User Identification

4.6. Feature Selection

For the purpose of pattern discovery (association and clustering rules) only the URL attributes are relevant for WUM. Hence, the other attributes are removed from the dataset. But as the number of unique URLs/paths in the log file was too large (over 175,000), the researcher then decided to categorize those Page/URL in to six different categories. Accordingly, Page/URLs whose categories are article, company profile, news, tender, user and index are shown below in table 4.1. At this point the Page/URL that were not used for the behavior analysis are excluded; pages are related to administration pages such as admin.html, admin_orders_edit.html, admin_users_edit.html, admin_users_view.html etc.

As can be seen from Table 4.1 below, from the sample 96,407 records of log file, ARTICLE category takes 186 lines of log recodes which is 0.09 percent of the total record. The */articles/starting-a-business* page take 82 lines of record from its categories which is 44.09 percent of the ARTICLE category.

Catagories	Page	Occurence	Percentage	Total %
ARTICLE	/articles/starting-a-business	82	44.09	0.09
	/articles/customs	30	16.13	0.03
	/articles/general-info	26	13.98	0.03
	/articles/industry	24	12.90	0.02
	/articles/labour-law	24	12.90	0.02
COMPANY PROFILE	/directory/index.html	8,570	43.89	8.89
	/directory/index.php	4,351	22.28	4.51
	/directory/advsearch	3,252	16.65	3.37
	/directory/add-listing	2,139	10.95	2.22
	/directory/15308-autoetnet-ethiopian-car-dealer	1,215	6.22	1.26
NEWS	/news/page:3	1,533	50.54	1.59
	/news/page:5	717	23.64	0.74
	/news/page:7	435	14.34	0.45
	/news/page:9	210	6.92	0.22
	/news/page:11	138	4.55	0.14
TENDER	/tenders/view/43337	179	0.24	0.19
	/tenders/view/64295	167	0.23	0.17
	/tenders/users/subscription	154	0.21	0.16
	/tenders/users/login	125	0.17	0.13
	/tenders/users/language	94	0.13	0.10
	/tenders/users/fax_subscription	13	0.02	0.01
USER	/users/login	24,246	68.93	25.15
	/users/register	5,690	16.18	5.90
	/users/upgrade	1,721	4.89	1.79
	/users/password_reset	1,565	4.45	1.62
	/users/account	944	2.68	0.98
	/users/payment	734	2.09	0.76
	/users/renew	271	0.77	0.28
	/users/upgrade2	5	0.01	0.01
INDEX	/index.php	37,753	0.76	39.16
TOTAL		**96,407**		

Table 4.1: Page Categories and their occurrence

4.7. Transaction Identification

Once the features are selected, the next step is to identify the transaction from the consolidated log file, as mentioned in Section 4.6. For the experiments, transactions were identified for each

categories data. To identify the transactions, the first step was to sort each data in order to arrange the same session IDs together. As a result, the log files were sorted with MS Excel using the columns (Date, Session ID and Time) as sort keys respectively (See table 4.3). The resulting file was saved as a separate text file. Note that in this case the first column holds Transaction ID and the next six columns hold the Page/URLs (ARTICLE, COMPANY PROFILE, NEWS, TENDER, USER, and INDEX).

Web Access Log File	Number of Transactions
Access 2014	2,080,753
Access 2015	143,025

Table 4.2: Number of Transactions Used from Web Access Log Files

Session ID	IP	Date	Page/URLs	SOR	Referr
842322:1	31.13.99.115	25/Jul/2014:15:34:45	INDEX	493	?
842322:1	31.13.99.115	25/Jul/2014:15:34:45	INDEX	493	?
842322:10	188.165.194.217	25/Jul/2014:15:51:37	INDEX	388	http://ww
842322:10	188.165.194.217	25/Jul/2014:15:51:37	TENDER	388	?
842322:100	213.55.107.71	25/Jul/2014:17:16:45	INDEX	446	?
842322:100	213.55.107.71	25/Jul/2014:17:16:45	INDEX	446	?
842322:1000	213.55.73.110	26/Jul/2014:08:56:54	COMPANY PROFILE	8456	http://ww
842322:1000	213.55.73.110	26/Jul/2014:08:56:54	TENDER	8456	?
842322:10000	82.145.216.22	28/Jul/2014:10:25:15	INDEX	8628	?
842322:10000	82.145.216.22	28/Jul/2014:10:25:15	ARTICLE	8628	?
842322:10000	82.145.216.22	28/Jul/2014:10:38:57	ARTICLE	8515	?

Table 4.3: Sample Transaction for clustering task

As depicted in table 4.4, corresponding to each Transaction ID, if a specific Page/URLs is accessed, it is indicated with “A” Otherwise, “?” is used to signify that the Page/URLs does not exist in the transaction. Note that “?” is used to represent non-accessed Page/URLs in the transactions so that Weka ignores it. The URLs in the sorted log file was compared with the six top accessed Page/URLs to check whether the two Page/URLs match or not.

Session ID	IP	ARTICLE	COMPANY PROFILE	NEWS	TENDER	USER	INDEX
842322:1	31.13.99.115	?	?	?	A	?	A
842322:10	188.165.194.217	?	?	?	A	A	A
842322:100	213.55.107.71	?	A	A	A	?	A
842322:1000	213.55.73.110	A	A	A	A	A	A
842322:10000	82.145.216.22	A	?	A	?	?	A
842322:10001	213.55.109.121	A	A	A	A	A	A
842322:10002	172.246.114.2	?	?	?	A	A	A
842322:10003	82.145.217.226	?	A	?	A	?	A
842322:10004	92.128.113.113	?	?	?	A	A	A
842322:10005	82.145.220.145	?	A	?	?	?	A

Table 4.4: Sample Transaction for Association

4.8. Data transformation

The last step in Data Preprocessing was to transform the transactions to a Weka understandable format. As a result, the files in section 4.7 above were edited with MS Excel in order to remove the Transaction ID column, add the column labels, ARTICLE, COMPANY PROFILE, NEWS, TENDER, USER, and INDEX. Then the resulting file, which is a URL Matrix, was saved in a separate folder with ".csv" extension. See table 4.5 below for sample dataset consisting of Page/URLs Matrix.

ARTICLE	COMPANY_PROFILE	NEWS	TENDER	USER	INDEX
?	?	?	?	?	A
?	?	?	A	?	A
?	?	?	?	?	A
?	A	?	A	?	?
A	?	?	?	?	A
A	A	A	A	?	A
?	?	?	A	A	A
?	A	?	?	?	A
?	?	?	A	A	A

Table 4.5: Sample Dataset for Associations csv Format

This dataset is used for WUM pattern discovery using association rule mining algorithms such as FP-Growth with Weka software.

Session ID	IP	Date	Page/URLs	SOR
842322:10	188.165.194.217	25/Jul/2014:15:51:37	TENDER	388
842322:1000	213.55.73.110	26/Jul/2014:08:56:54	COMPANY PROFILE	8456
842322:1000	213.55.73.110	26/Jul/2014:08:56:54	TENDER	8456
842322:10000	82.145.216.22	28/Jul/2014:10:25:16	ARTICLE	8628
842322:10000	82.145.216.22	28/Jul/2014:10:38:57	ARTICLE	8515
842322:10001	213.55.109.121	28/Jul/2014:10:25:32	TENDER	43840
842322:10001	213.55.109.121	28/Jul/2014:10:26:26	TENDER	12173
842322:10001	213.55.109.121	28/Jul/2014:10:30:16	TENDER	7163
842322:10001	213.55.109.121	28/Jul/2014:10:31:36	TENDER	7257
842322:10001	213.55.109.121	28/Jul/2014:10:39:38	COMPANY PROFILE	9192

Table 4.7: Sample Dataset for clustering csv Format

This dataset is used for WUM pattern discovery using clustering mining algorithms such as K-means with Weka software.

CHAPTER FIVE

EXPERIMENTATION AND ANALYSIS

In this study, an experiment is conducted using statistical analysis, clustering and association rule discover so as to mine patterns in web usage data of Ebiz (e-business) Online Solutions PLC official website.

The experiment is conducted using Laptop computer with Intel Core i7 processor that has 2.6 GHz processor speed and 8 GB RAM (memory) size. The Operating System is Windows 8.0 Pro (64 - bit). Weka 3.7.9 knowledge discovery software is used for pattern discovery, WebLog Expert and MS Excel 2013 for Statistical Analysis as well as MS Access 2013 and W3Perl for constructing user profile.

5.1. Experimental setup

In this research, a complete statistical analysis has been implemented through WebLog Expert. This statistical information is used to produce a periodic report from the site such as information about user's popular pages, average visit time of a page, average time of users' browsing through a site, average length of a navigational path through a site, common entry and exit pages and high-traffic days of site.

As a matter of fact statistical technique for pattern discovery perform a sketchy analysis on preprocessed data but obtained knowledge can be useful. For instance detecting entry points which are not home page or finding the most common invalid URL lead to enhance system performance and security and also facilitate the site topology modification task.

W3Perl is used for constructing user profile statistically. W3Perl are mainly used to construct two kinds user profile statistically i.e. single user profile (Host) and aggregate user profile (pages). Aggregate user profile (pages) is used to identify users with similar preference and interest of page (URL). Such knowledge is especially used for automated return mail to users falling within a certain aggregate user profile. Single user profile (Host) is used to identify the interest of a particular user. This knowledge is used for dynamically changing a particular site for a user, on a return visit, based on past classification of that user (provide personalized Web content to the users).

In this research K-means clustering algorism is experimented. Clustering techniques used to discover one of interesting cluster i.e. page cluster. Clusters of Web pages contain pages that seem to be conceptually related according to the users' perception. In this research the knowledge that is obtained from clustering is useful to identify which page the most frequent one is for to construct user profile. This profile is used for market segmentation in ecommerce, designing adaptive Websites, personalized web page, and designing recommender systems.

Once sessions have been identified association rules can be used to relate pages that are most often referenced together in a single server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests.

Pattern analysis and constructing user profile process two fundamental goals. The first goal is to extract the interesting rules, patterns or statistics from the output of the pattern discovery process by filtering the irrelative rules or statistics. Another aim of this analysis is to obtain some information can offer valuable insights about users' navigational behavior. For example we can understand the number of users that started from a page and proceeded through some certain pages and finally visited their goal page. Also, we can obtain some information about page popularity or some pages that contain the most information for a visitor. The exact analysis is done by Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data, Discussion with the expertise, content information can be used to filter out patterns containing pages of a certain usage type, content type.

The tender user profile is constructed using the result obtained from the experiment, and integrate with the tender database. Finally, the usefulness of users profile has evaluated.

5.2. Statistical Analysis

In this research, for statistical analyses WebLog Expert is used. This tool uses different terms such as accesses, request, etc. Accesses correspond to hits on HTML pages. This is an accurate way to know how popular your website is. Keeping track of request (hits) is a way of

measuring traffic to a website that can be misleading. This is a very important distinction between Requests and Accesses. When a page is requested from the server, it may contain, for example, four images. When this occurs, the access log will records five "hits" - one for the HTML Web page and four for the four images. However, when one attempt to track statistics, the interest on HTML web page requested, not in the number of images that are requested.

Below are some important statistical reports extracted from WebLog Expert and summarized for the period of July 25, 2014 – January 22, 2015. During the given period, as shown in table 5.1 the website was visited 2,340,929 times with an average daily request of 12,862, accesses 12,218, hosts 1,431 and traffic (in MB). Table 5.1 below shows the General Statistics and the description of the common URLs (Appendix I) that are in the statistical reports respectively.

Global access	Total	External
Requests	2,340,929	2,340,929
Accesses	2,223,778	2,223,778
Documents	582	582
Traffic (Gb)	26.18	26.18
Number of different sites	65,628	65,628
Number of different pages used	162,922	

Average	Days	Weeks
Requests	12,862	90,035
Accesses	12,218	85,529
Documents	3	22
Hosts	1,431	1,431
Traffic (Mb)	147	1,031

Table 5.1: General Statistics

5.2.1. Most Frequently Accessed Pages

Frequently accessed web pages are web pages that are most frequently visited by users. They indicate what the hottest pages in this web site are. Frequently web pages are created based on the usage statistics. Out of the five months log file, figure 5.1 presents the top ten most frequently accessed pages. Special attention should be given to these pages in terms of performance optimization and business promotion.

Most Popular Pages

	Page	Hits	Incomplete Requests	Visitors	Bandwidth (KB)
1	http://www.2merkato.com/	283,124	12	115,217	6,254,216
2	http://www.2merkato.com/tenders/	90,772	0	55,796	991,170
3	http://www.2merkato.com/login/	75,120	0	33,461	1,033,471
4	http://www.2merkato.com/feed/rss/	35,053	0	30,300	131,590
5	http://www.2merkato.com/tenders/page:2/	41,295	0	29,151	440,497
6	http://www.2merkato.com/news/	48,755	0	24,766	489,737
7	http://www.2merkato.com/search/	53,150	0	22,956	171,753
8	http://www.2merkato.com/tenders/page:3/	28,925	0	21,823	305,962
9	http://www.2merkato.com/users/login/	29,307	0	18,379	59,931
10	http://www.2merkato.com/tenders/page:4/	22,632	0	17,567	240,443
	Subtotal	**708,133**	**12**	**N/A**	**10,118,775**
	Total	**2,778,634**	**12**	**N/A**	**30,682,346**

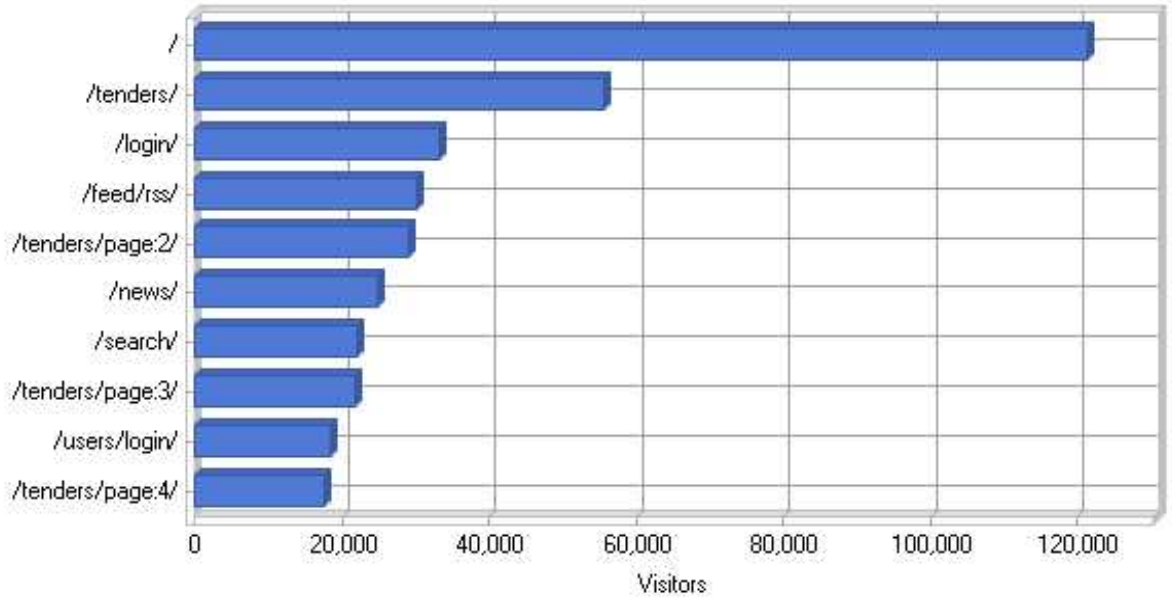

Figure 5.1: Most Frequent access pages

As shown in Figure 5.1 above, the home page, Tenders pages and login pages are the most three frequently accessed pages. This is because the Home page and Tenders pages are the default entry pages and the login page is used to track the real and the subscribed customer by web administrator.

5.2.2. Page Views per Visitor

Looking at page views per visitor can tell what content on your site is the most popular. Figure 5.3 below depicts summary of page views per visitors.

No	Page Views per Visitor	Visitors	Percentage Visitors	Persentage Anlysis of Page View
1	1 page view	332,919	55.20	83.86
2	2 page views	94,828	15.72	
3	3 page views	51,832	8.59	
4	4 page views	26,184	4.34	
5	5 page views	16,365	2.71	16.14
6	6 page views	15,315	2.54	
7	7 page views	7,623	1.26	
8	8 page views	6,110	1.01	
9	0 page views	5,993	0.99	
10	9 page views	5,192	0.86	
Subtotal		**562,361**	**93.25**	
Total		**603,079**	**100.00**	

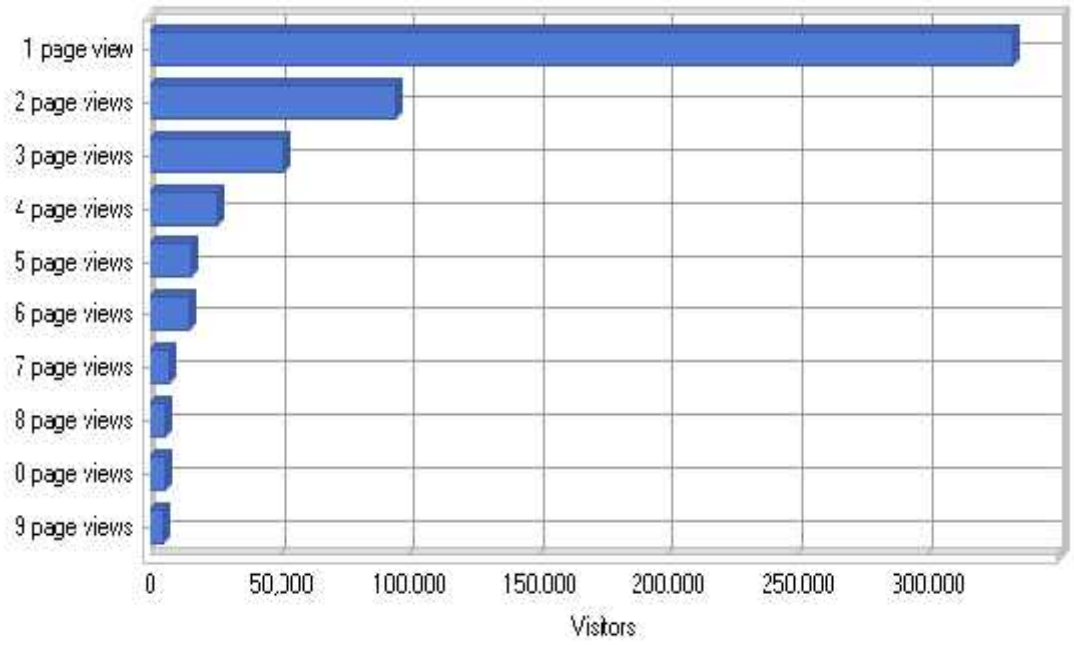

Figure 5.2: Page Views per Visitor

As shown in Figure 5.2 above, 83.86 percent of the user do not browse further than four pages into the site. Therefore the web master should be tactful to ensure that most important information is contained within four pages of the common site entry points.

5.2.3. Top Entry Pages

Entry pages are the first pages visited by a user on the site. Figure 5.3 below shows the number of visitors and top entry pages.

Top Entry Pages

	Page	Visitors
1	http://www.2merkato.com/	94,280
2	http://www.2merkato.com/tenders/	26,639
3	http://www.2merkato.com/feed/rss/	22,961
4	http://www.2merkato.com/news/	13,171
5	http://www.2merkato.com/component/com_ninjarsssyndicator/feed_id,1/format,raw/	11,022
6	http://www.2merkato.com/feed/atom/	10,531
7	http://www.2merkato.com/news/feed/rss/	8,310
8	http://www.2merkato.com/users/login/	5,422
9	http://www.2merkato.com/administrator/	5,368
10	http://www.2merkato.com/news/alerts/3167-ethiopia-government-announced-civil-servants-salary-adjustment/	5,109
	Subtotal	**203,591**
	Total	**597,086**

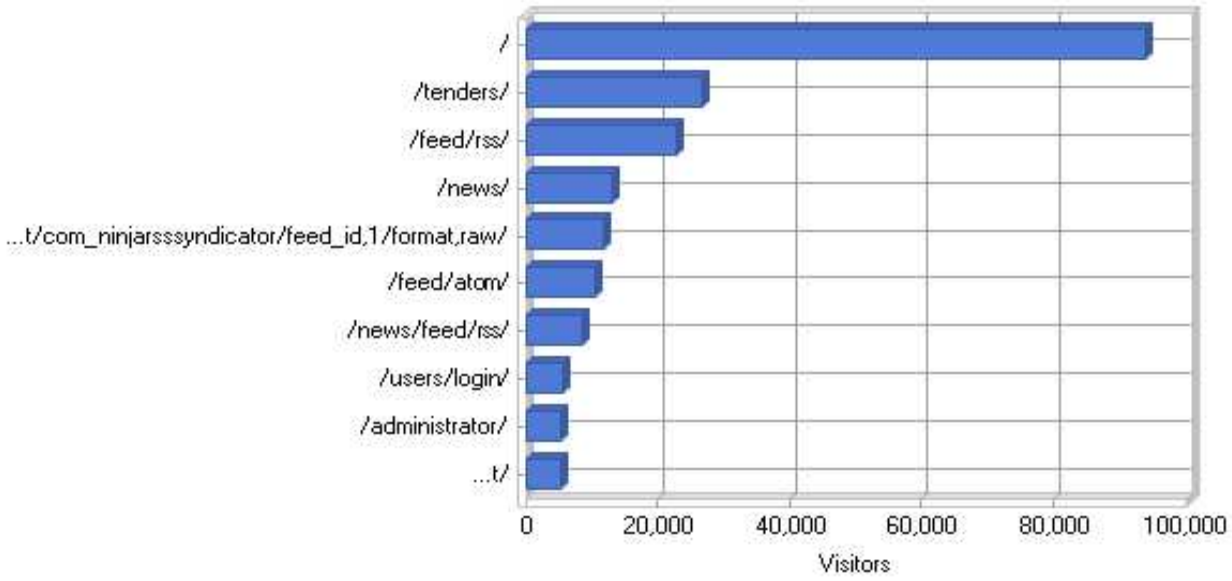

Figure 5.3: Top Entry Pages

As shown in figure 5.3 above, ignoring the admin pages such as feed/rss/…./feed_id 1/format.raw, /news/feed/rss/. The home page, tenders and news pages are the three top most landing pages. This shows that these pages required by many users and they are relatively better optimized. These pages are the most frequently used as entry pages to Ebiz (e-business)

website. Therefore, it is required to make these pages resourceful and more interesting in order to minimize the bounce rate.

5.2.4. Top Exit Pages

Top exit pages specify the last page visited by a user on the site. These are the exit points where visitors left the website. In some cases, a page can be both landing and exit page. Figure 5.4 below shows the top ten exit pages of visitors Ebiz (e-business) website.

Top Exit Pages

	Page	Visitors
1	http://www.2merkato.com/	89,898
2	http://www.2merkato.com/feed/rss/	24,477
3	http://www.2merkato.com/tenders/	18,817
4	http://www.2merkato.com/component/com_ninjarsssyndicator/feed_id,1/format,raw/	11,831
5	http://www.2merkato.com/search/	11,840
6	http://www.2merkato.com/news/	11,176
7	http://www.2merkato.com/feed/atom/	8,851
8	http://www.2merkato.com/log-in/	8,738
9	http://www.2merkato.com/news/feed/rss/	8,211
10	http://www.2merkato.com/articles/tax/types/61-personal-income-tax-in-ethiopia/	5,731
	Subtotal	**179,370**
	Total	**597,086**

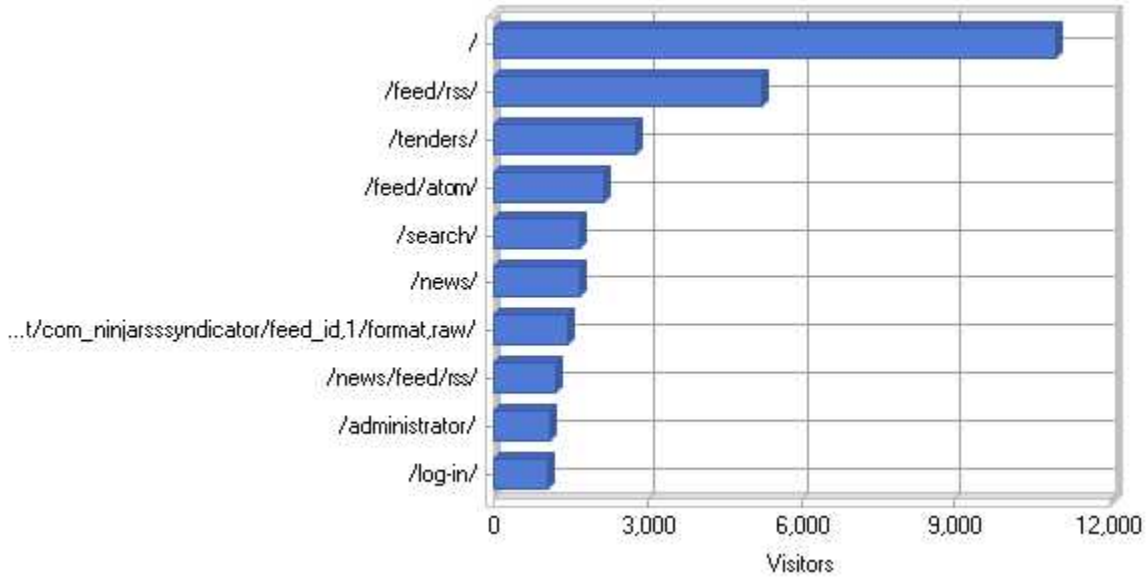

Figure 5.4: Top Exit Pages

Ignoring the admin pages such as feed/rss/…./feed_id 1/format.raw, /news/feed/rss/. Figure 5.4 above clearly presents that the home page, tenders and news pages are the three top most exit pages. The tender pages and the news pages are in the top exit list pages because users complete their transaction here.

5.2.5. Top website access by country

The knowledge of how the website is being accessed by different countries could be useful in order to improve the website access by different geographic locations. Figure 5.5 depicts the top ten countries that most frequently access Ebiz (e-business) website.

Most Active Countries

	Country	Hits	Visitors	% of Total Visitors	Bandwidth (KB)
1	Ethiopia	142,169	26,608	26.94%	1,566,164
2	France	72,359	22,343	22.63%	732,259
3	United States	54,883	20,202	20.48%	550,077
4	Ukraine	19,533	4,500	4.57%	429,417
5	China	16,738	4,229	4.28%	373,859
6	France	7,325	2,428	2.46%	158,982
7	Germany	5,130	2,275	2.30%	85,975
8	India	33,831	1,896	1.92%	660,461
9	United Kingdom	37,214	1,666	1.69%	88,101
10	Russian Federation	4,615	1,279	1.30%	102,522
	Subtotal	**393,805**	**87,465**	**88.57%**	**4,748,628**
	Total	**454,631**	**98,760**	**100.00%**	**6,451,993**

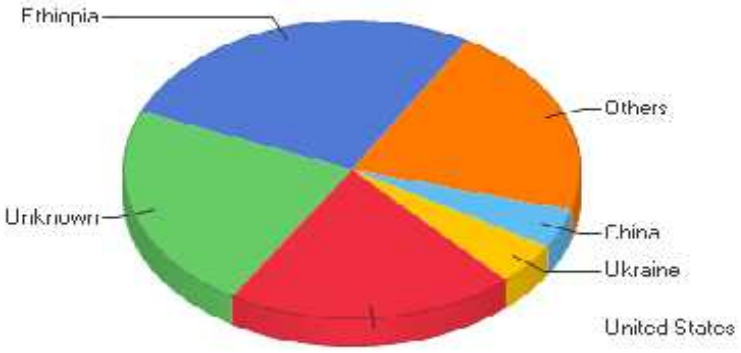

Figure 5.5: Most Active Countries

As can be seen from figure 5.5, the top most three countries in which the website is frequently accessed are: Ethiopia, France and United States. It is required to analyze why some countries such as India, which are supposed to be among the top five lists are missing in order to improve the website access in that country.

5.2.6. Top Website Referrals

Referrals are another useful parameter in measuring website accessibility. Referrals are sources that redirect the users to the given website. These could be search engines, social media or other websites. If a given website is characterized by many referrals, it implies that the website is popular. Figure 5.6 below summarizes the top ten referrals to Ebiz (e-business) website.

Top Referring Sites

	Site	Visitors
1	No Referrer	427,555
2	http://tender.2merkato.com	4,968
3	http://search.tb.ask.com	4,141
4	http://r.search.yahoo.com	2,506
5	http://www.google.com	2,256
6	http://mereja.com	1,897
7	http://news.google.com	1,825
8	http://lm.facebook.com	1,595
9	http://www.google.com.et	1,257
10	http://www.aigaforum.com	1,185
	Subtotal	**449,185**
	Total	**465,355**

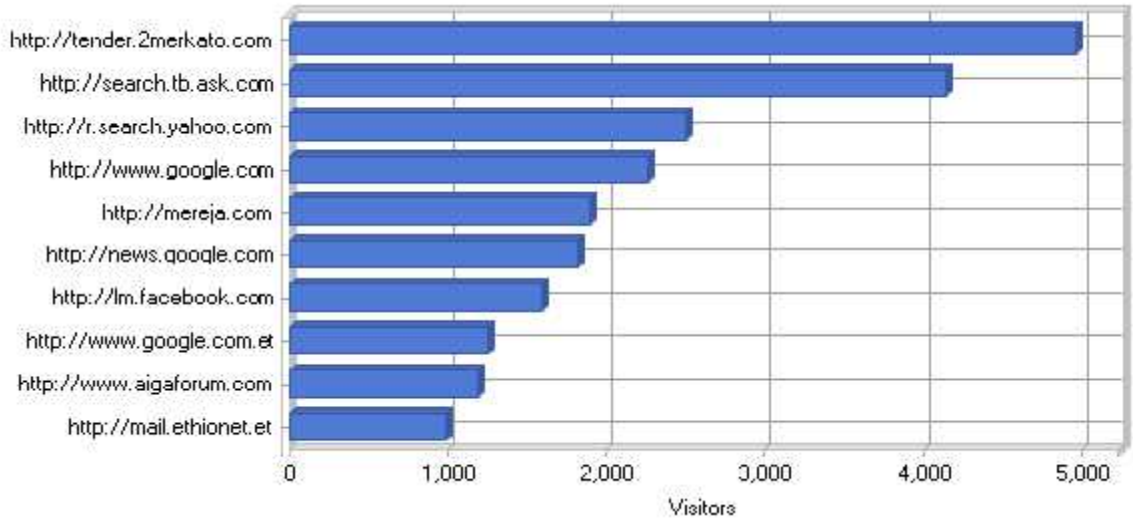

Figure 5.6: Top Referring Sites

As it is shown in figure 5.6, tender.2merkato.com is the predominant referral to Ebiz (e-business) website because it is the initial domain name of the website when it was first launched. In addition the web site is well known for its up-to-date and complete information on the tender.

5.2.7. Top Browsers

The type of web browsers that are used by the users to access the website is another useful parameter that is worth knowing. Figure 5.7 below shows the top ten web browsers that are used by the users to access Ebiz (e-business) website.

Most Used Browsers

	Browser	Hits	Visitors	% of Total Visitors
1	Google Chrome	713,107	147,764	23.14%
2	Firefox	721,145	129,692	20.31%
3	Opera	376,378	122,481	19.18%
4	Internet Explorer	454,580	48,817	7.65%
5	Android Browser	139,045	29,651	4.64%
6	Mobile Safari	79,801	22,150	3.47%
7	FeedBurner/1.0 (http://www.FeedBurner.com)	17,004	16,036	2.51%
8	Recorded Future	16,379	11,494	1.80%
9	Safari	35,343	11,288	1.77%
10	Others	300,019	5,814	0.91%
	Subtotal	**2,852,801**	**545,187**	**85.38%**
	Total	**3,301,628**	**638,505**	**100.00%**

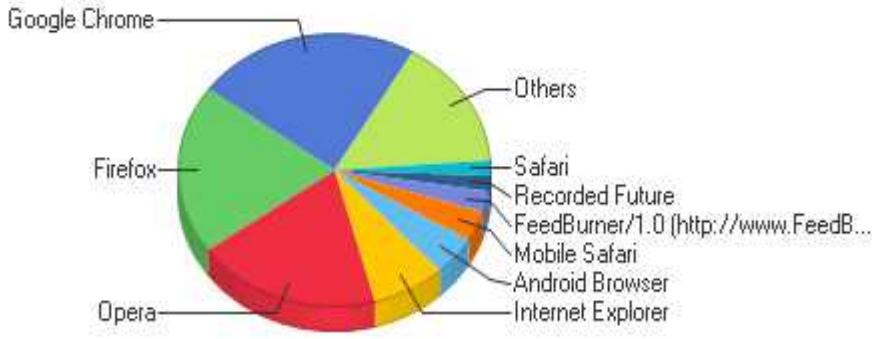

Figure 5.7: Top Ten Browsers

As shown in the above figure 5.7, it has been found that most of the Web browsers are used by most of the users to visit the Website. The result clearly shows that Google Chrome, Firefox and Opera are the top most browsers used by users to access the website. This helps the

website designers to consider browsers compatibility while developing the application in order to increase the effectiveness of the Website.

5.2.8. Access Trend Analysis

In this study users access trend analysis is undertaken by month, by day and by hour of a day as discussed below.

5.2.8.1. Trend Analysis by Month

As shown in figure 5.8 below, the highest access to the website was recorded in October followed by September and August and the lowest in July. As discussed with experts, this is logical as October is one of the peak seasons because most government companies buy different goods by this month. The lowest access in July is probably because most government companies close their fiscal year budget and start the new budget year by this month.

Activity by Month

Month	Hits	Page Views	Visitors	Bandwidth (KB)
Jul 2014	120,163	111,177	19,920	1,474,557
Aug 2014	561,703	514,729	106,981	6,519,681
Sep 2014	637,809	599,008	93,375	7,145,875
Oct 2014	646,796	604,903	100,011	8,733,062
Nov 2014	392,538	367,987	72,502	4,347,907
Dec 2014	530,564	495,115	111,540	6,655,690
Jan 2015	458,174	427,984	98,750	5,482,435
Total	**3,347,747**	**3,120,903**	**603,079**	**40,359,209**

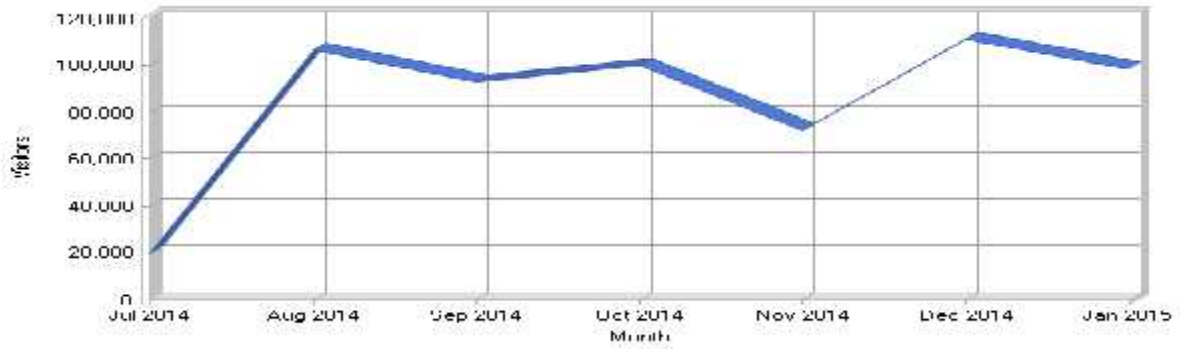

Figure 5.8: Activity by Month

5.2.8.2. Trend Analysis by Day

The daily web data for the entire weeks tells about the number of Hits, Pages, Visits, and Bandwidth (KB) that have been visited. Details of user's daily activities are presented in Figure 5.9.

Daily Activity

Date	Hits	Page Views	Visitors	Average Visit Length	Bandwidth (KB)
Tue 1/13/2015	21,537	20,073	5,033	00:33	253,779
Wed 1/14/2015	33,390	30,916	7,846	00:13	387,956
Thu 1/15/2015	33,678	31,464	7,082	05:18	370,260
Fri 1/16/2015	30,386	28,332	6,914	05:37	337,556
Sat 1/17/2015	28,324	24,974	5,154	05:57	320,813
Sun 1/18/2015	22,080	20,742	5,166	05:21	231,898
Mon 1/19/2015	26,550	24,888	5,900	05:33	347,413
Tue 1/20/2015	37,838	35,470	7,215	00:32	458,705
Wed 1/21/2015	38,462	35,964	7,828	05:49	416,905
Thu 1/22/2015	10,806	9,096	2,210	05:03	121,136
Subtotal	**281,011**	**262,619**	**60,108**	**05:53**	**3,226,505**
Total	**3,347,747**	**3,120,903**	**603,079**	**06:15**	**40,359,209**

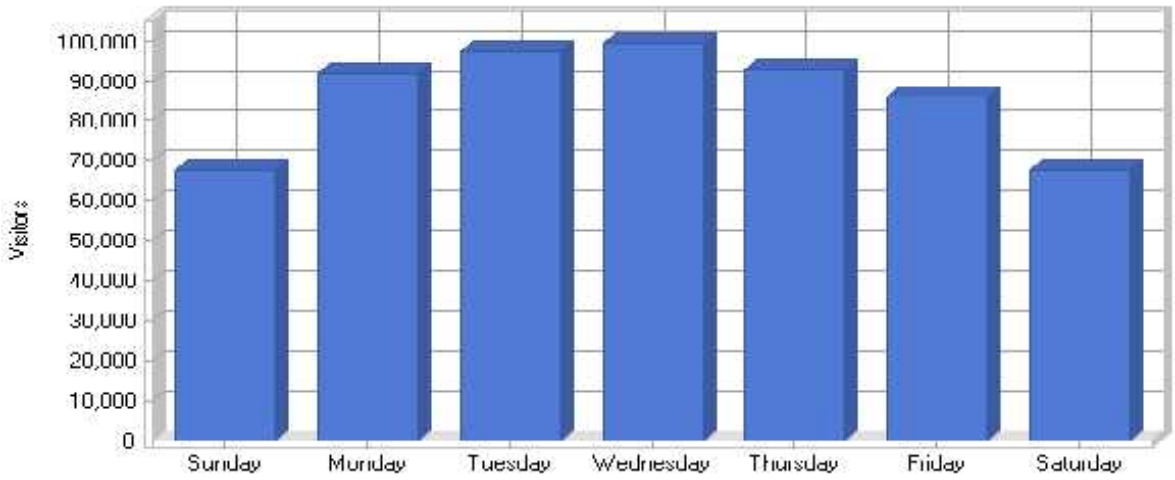

Figure 5.9: Daily Activity

As shown in Figure 5.9, the maximum Hits per Day were 38,462, the maximum visits per Day were 7,828 and the maximum Bandwidths (KB) per Day were 458,708. Figure 5.9 shows above present's daily access trend of Ebiz (e-business) website, it receives far more visits from Monday to Saturday. In general, access to the website is lower during weekends and higher during weekdays. On Sundays, it is at its lowest level. This reinforces further the earlier finding that the Website is mainly used by working people, employees or company because users (company) trend to access the website while they are on duty than when they are off duty.

5.2.8.3. Trend Analysis by Hour of Day

The amount of requests per hour can help to detect times of very high load. If that is the case one could decide to increase the server capacity so that the service does not break down during these times. On the other hand, one can see times during which the dataset is not requested very often so the capacity of the server could be decreased to save resources and money. An observation of the time statistics over a time period can help to reason about the popularity of a dataset. Analysis of hourly access trend of the website is shown in Figure 5.10 below.

Activity by Hour of Day

Hour	Hits	Page Views	Visitors	Bandwidth (KB)
00:00 - 00:59	58,456	54,603	12,088	756,371
01:00 - 01:59	55,550	51,614	12,536	832,443
02:00 - 02:59	63,383	57,993	14,287	910,598
03:00 - 03:59	76,041	71,127	17,909	993,524
04:00 - 04:59	136,952	130,686	21,731	1,392,324
05:00 - 05:59	141,973	131,923	29,100	1,596,796
06:00 - 06:59	170,840	156,438	34,731	1,986,899
07:00 - 07:59	171,430	155,879	35,641	2,102,386
08:00 - 08:59	178,939	162,846	36,120	2,280,691
09:00 - 09:59	167,728	154,552	31,738	2,105,535
Subtotal	**1,221,292**	**1,127,661**	**245,881**	**14,957,570**
Total	**3,005,478**	**2,778,634**	**598,139**	**38,021,667**

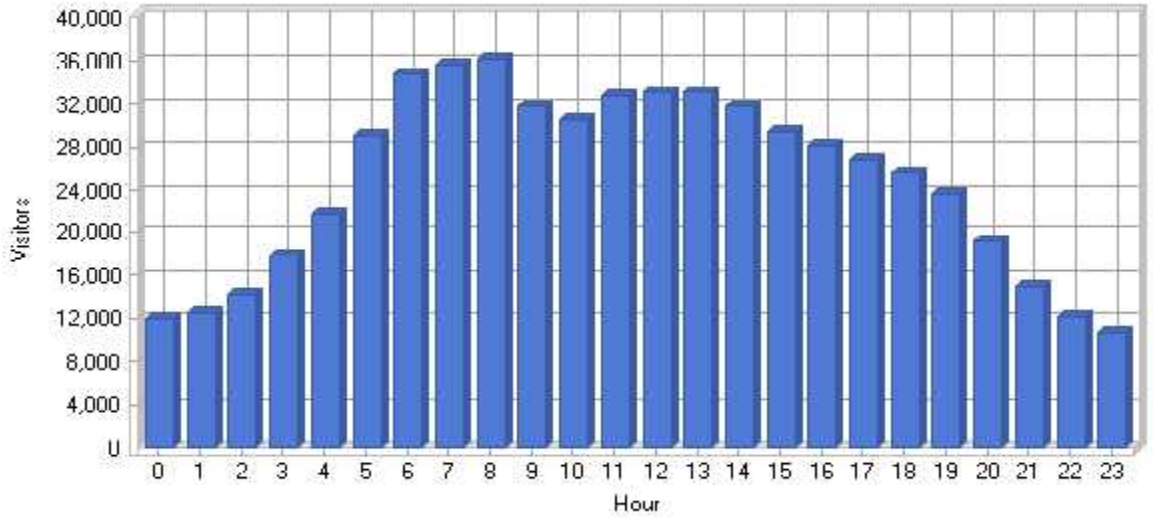

Figure 5.10: Activity by Hour of Day

From figure 5.10 we found that, access to the website is at its lowest during the nights and at its peak during daytime (Web Server time which is GMT). It rises during the morning and

reaches its maximum in the afternoon. Then it starts falling down in the evening. This is because majority of the website users are active during daytime and at rest during the night.

5.2.8.4. User profile by Unique IP Address

User Profile is information about the user including demographic attributes (age, income, etc.), and preferences that are gathered either explicitly (through registration forms) or implicitly (through Web server logs). In this research user profile gather implicitly by processing the server log file using tools called **W3Perl** .This tool consists of set of Perl Script that can be customize the way you want to be the output look like. The researcher customize the Perl script to create user profile in such way that each user(host) requests, percentage of the total request, accesses, percentage of the total accesses, different accesses for each user, daily visits , traffic (download) ,and percentage of the total traffic of the pages. Under the output, as shown in Table 5.2 for each user (IP) are the first top ten frequent pages with its occurrence.

num	Hosts	Requests	Percentage	Accesses	Percentage	Different accesses	Daily visits	Traffic	Percentage
>	212.71.251.162	131378	6.0 %	131378	6.3 %	597	105	53.7 Mb	0.2 %
>	14.140.53.178	105997	4.8 %	104898	5.0 %	6179	87	838.9 Mb	3.3 %
>	46.105.17.34	70106	3.2 %	70106	3.4 %	1	2	302.3 Mb	1.2 %
>	66.249.93.79	28330	1.3 %	28268	1.4 %	17044	97	19.7 Mb	0.1 %
>	193.201.224.94	25804	1.2 %	25804	1.2 %	3	5	112.4 Mb	0.4 %
>	41.82.26.228	19877	0.9 %	19745	0.9 %	14349	1	525.1 Mb	2.1 %
>	80.78.250.25	19202	0.9 %	19202	0.9 %	1	1	82.8 Mb	0.3 %
>	64.31.25.56	18842	0.9 %	18842	0.9 %	1	1	81.7 Mb	0.3 %
>	41.82.61.5	18685	0.9 %	18650	0.9 %	12498	1	488.6 Mb	1.9 %
>	195.238.108.103	18598	0.8 %	18598	0.9 %	1	2	80.2 Mb	0.3 %

Table 5.2: Top Ten User Profile among 60,330

From this output we can understand user behavior and personalizing the page for each user is crucial to fulfill their interest without asking explicitly. Table 5.3 further presents expanded user profile, including pages and their occurrence under each category by host (IP address).

N0	Hosts		Requests	Percentage	Accesses	Percentage	Different accesses	Daily visits	Traffic	Percentage
1	**212.71.251.162**		131,378	6.0 %	131,378	6.3 %	597	105	53.7 Mb	0.2 %
	Pages	**Occurence**								
	/tenders/categories/upcat	50,327								
	/tenders/cron/task:send/se	30,194								
	/tenders/cron/task:queue/s	30,142								
	/email_queues/cron/35bca	15,090								
	/admin/users/login	5,033								
	/tenders/fax/53d31f11-0f38	1								
	/tenders/fax/53d441f2-555	1								
	/tenders/fax/53d442e1-c28	1								
	/tenders/fax/53d44b4a-1e1	1								
	/tenders/fax/53d8375a-d2b	1								
2	**14.140.53.178**		105,997	4.8 %	104,898	5.0 %	6,179	87	838.9 Mb	3.3 %
	Pages	**Occurence**								
	/index.html	747								
	/tenders/index.html	394								
	/tenders	391								
	/tenders/page:2	365								
	/tenders/page:3	361								
	/tenders/page:4	359								
	/tenders/page:5	357								
	/tenders/page:6	356								
	/tenders/page:7	350								
	/tenders/page:8	350								
3	**46.105.17.34**		70,106	3.2 %	70,106	3.4 %	1	2	302.3 Mb	1.2 %
	Pages	**Occurence**								
	/administrator/index.php	70,106								
4	**66.249.93.79**		28,330	1.3 %	28,268	1.4 %	17,044	97	19.7 Mb	0.1 %
	Pages	**Occurence**								
	/users/login	49								
	/tenders/pixel/54050a2f-83	43								
	/tenders/index.html	38								
	/tenders/pixel/53d83d7a-6	36								
	/tenders/pixel/53eab36f-42	36								
	/tenders	35								
	/tenders/pixel/54026731-b	33								

Table 5.3: Top Ten Expanded User Profile among 60,330

5.2.8.5. User profile by Page View

The researcher in this case customize the Perl script to create aggregate user profile in such way that as shown in Table 5.4 each page, occurrence, percentage of the total occurrence, traffic (download), percentage of the total traffic and number of hosts (users) accesses the displayed page.

	Pages	Occurence	Percentage	Traffic (Gb)	Hosts
>	/index.html	180749	8.7 %	3.93	17325
>	/tenders	49573	2.4 %	0.53	8015
>	/search	32265	1.6 %	0.10	6605
>	/news	29086	1.4 %	0.29	5633
>	/tenders/page:2	23020	1.1 %	0.25	5024
>	/tenders/index.html	39164	1.9 %	0.42	5008
>	/login	49910	2.4 %	0.67	4247
>	/tenders/page:3	15922	0.8 %	0.17	4216
>	/users/login	22228	1.1 %	0.04	3928
>	/tenders/page:4	12745	0.6 %	0.14	3798

Table 5.4: Aggregate User Profile

In line with pages categories depicted in table 5.4. Table 5.5 shows for each page category the first top ten frequent hosts (users) with their occurrence.

	Pages	Occurence	Percentage	Traffic (Gb)	Hosts
>	/index.html	180749	8.7 %	3.93	17325
	Hosts / Occurence: 192.35.222.163 6333; 37.57.231.125 2011; 109.239.235.195 1558; 107.183.64.47 1534; 65.36.241.77 1444; 91.207.7.182 1427; 109.239.235.213 1380; 23.22.131.24 1206; 109.239.235.222 1128; 38.108.108.178 1101				
>	/tenders	49573	2.4 %	0.53	8015
	Hosts / Occurence: 213.55.107.68 393; 14.140.53.178 391; 213.55.107.111 297; 213.55.107.91 254; 208.69.40.107 253; 66.249.89.24 210; 213.55.104.203 203; 197.156.119.2 201; 213.55.107.111 192; 213.55.107.77 186				
>	/search	32265	1.6 %	0.10	6605
	Hosts / Occurence: 78.180.96.92 2836				

Table 5.5: Expanded Aggregate User Profile

5.3.Pattern Discovery and Analysis

Weka was used to discover patterns for the aggregate dataset for which Transactions were identified and transformed for Weka. Both Association and Clustering have been experimented, algorithms used in this study are *K-means clustering algorithm* for clustering, whereas *Apriori and FP-Growth for association rule discovery.*

K-means clustering algorithm for clustering is selected because the number of cluster known in advance. For association discovery both Apriori and FP-Growth algorithms were tested by heuristically selecting different minimum support. After conducting several experiments, the researcher found that the patterns discovered by both algorithms give similar result. However, FP-Growth algorithm is faster than Apriori algorithm by several magnitudes. Hence, in this research, FP-Growth Algorithm was used for pattern discovery.

After conducting several experiments, the researcher found that the patterns discovered by both association and clustering algorithms give similar result.

Note that in each experiment below the threshold values for minimum support and confidence are selected by heuristics after conducting several experiments with different values. That is, the values that delivered optimum number of rules are selected and used.

5.3.1. Pattern Discovery and Analysis Using Clustering Algorism

In order to determine the navigation patterns the researcher apply the K-Means data mining method. K-Means (or centroid method) is a method of aggregation/clustering which computes a list of groups (clusters) of objects with similar characteristics from a data set. Specifically, the K-Means method is based on obtaining a number of K groups, set at the beginning of the process. In this study, the groups to be computed are sequences of users browsing the web site of the Ebiz, so as to determine the resources that were most frequently accessed.

The researcher present in Table 5.6 the results of K-Means for the real log data files of the Ebiz using the WEKA framework. Since there are five clusters of pages (such as tender, article, company profile, user and news) excluding the administrator categories, the size of clusters, K of k-means is defined with five number of clusters.

No Cluster	Cluster	Class attribute: Page/URLs	Clustered Instances	Incorrectly clustered instance	Incorrectly clustered instance (Percent)	Remark
5	0	TENDER	28,372 (57%)	22,678	45.9004%	Excluding INDEX Page I have 5 Categories
	1	ARTICLE	3,771 (8%)			
	2	COMPANY PROFILE	7,189 (15%)			
	3	USER	3,628 (7%)			
	4	NEWS	6,447 (13%)			

Table 5.6: Result of K-Means Cluster Algorism

The clustering result depicted in Table 5.6 shows that 45.9% of the instances are correctly clustered. Out of the total instances, 57% of the instances clustered in Cluster 0, which has brought together a total of 28,372 lines. The second and third largest Clusters are Cluster 2 and cluster 4, with 15% and 13%, respectively.

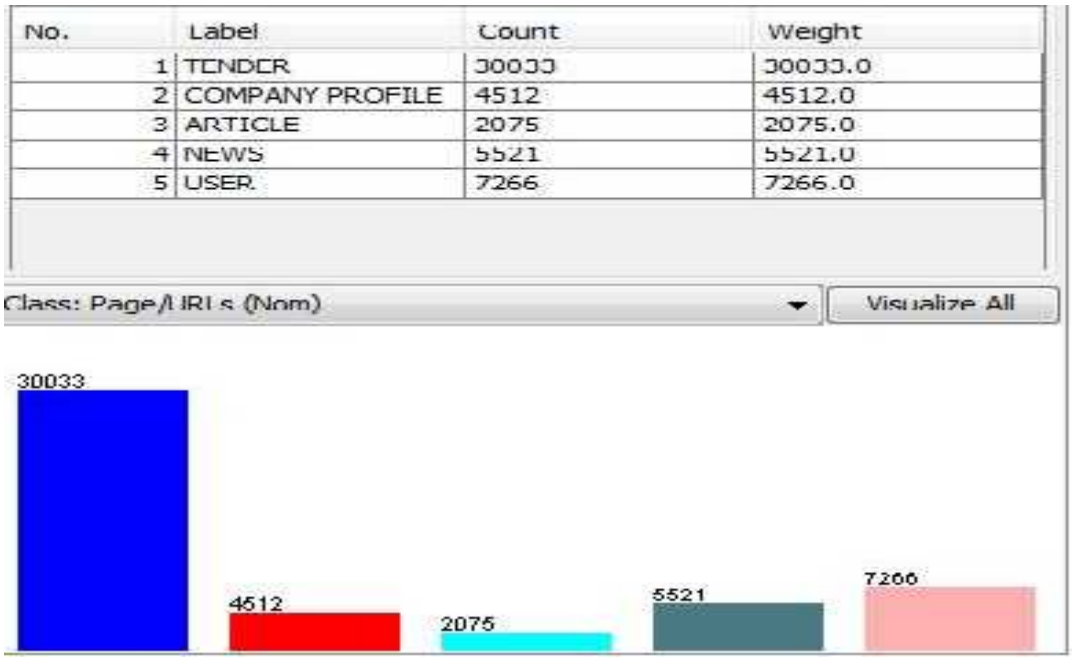

Figure 5.11: Results Resource of Accessing the Categories

As depicted in figure 5.11 the resource that has taken the most values of weight in Cluster 0(TENDER), Cluster 3(USER), Cluster 4(NEWS) and Cluster 2(COMPANY PRFILE) respectively.

Analyzing the weight of the Resource of Accessing the Categories, the resources arguably the second most commonly accessed resources are represented in Cluster 3, but the issue with that Cluster 3(USER) is only has 3,628. Nor does it appear that the second most common navigational sequence is represented by Cluster 2(COMPANY PRFILE), since the result is distorted by the large number of values present in the Weka K-means Clustering.

In section 5.2.8.5 the most common resources are those which have been accessed more often given the set of IP addresses. Beyond the statistics of in section 5.2, the data mining techniques employed here can give valuable insights into the navigational patterns and accesses to the resources.

The analysis of Table 5.6, allows us to conclude that, regardless of the total number of accesses, 57% of the user of Ebiz website tend to access resources whose URL is /tender/view/…/.html (TENDER). This is supported by the fact that this resource is the one that have taken the most accessed line of all the Cluster.

5.3.2. **Pattern Discovery and Analysis Using Association Algorism**

To discover association rule, an experiment is conducted using FP-Growth association rule discovery algorism. After conducting several experiment using different minimum support and minimum confidence, a minimum Support of 0.01 and Minimum Confidence of 0.75 provide but results home page attribute included.

Rule 1. USER=A 6984 ==> TENDER=A 6286 <conf :(0.9)> lift :(1.39) lev :(0.04) [1773] conv :(3.54)

This rule states that about 90% of visitors who visited the USER pages (/users/login, /users/register, /users/upgrade, /users/password_reset, /users/account, /users/payment and /users/renew) page also accessed the TENDER page (/tenders/view/, /tenders/users/subscription, /tenders/users/login, /tenders/users/fax_subscription and /tenders/users/language). This is because majority of users subscribed the Ebiz (e-business) website is tender page. As a result the login page takes users to the tender page directly.

Rule 2.USER=A INDEX=A 6123 ==> TENDER=A 5486 <conf :(0.9)> lift :(1.39) lev :(0.04) [1529] conv :(3.4)

This rule states that about 90% of visitors who visited the USER pages (/users/login, /users/register, /users/upgrade, /users/password_reset, /users/account, /users/payment and /users/renew) page and INDEX (index.html) also accessed the TENDER page (/tenders/view/, /tenders/users/subscription, /tenders/users/login, /tenders/users/fax_subscription and /tenders/users/language).

This is because almost all user subscribed the Ebiz (e-business) website is tender page the login page takes users to the tender page directly and the user after accessing the root directory directly go to the tender menu. In some cases, users might go to the home page after accessing the TENDER information in order to get further information or perform further transaction.

Rule 6. NEWS=A USER=A 677 ==> TENDER=A 527 <conf :(0.78)> lift :(1.2) lev :(0) [89] conv :(1.59)

This rule states that about 78% of visitors who visited the USER pages (/users/login, /users/register, /users/upgrade, /users/password_reset, /users/account, /users/payment and /users/renew) page and NEWS (/news/page 1,2…) also accessed the TENDER page (/tenders/view/, /tenders/users/subscription, /tenders/users/login, /tenders/users/fax_subscription and /tenders/users/language).

Rule 7. COMPANY_PROFILE=A NEWS=A 816 ==> TENDER=A 619 <conf :(0.76)> lift :(1.17) lev :(0) [91] conv :(1.46)

This rule states that about 78% of visitors who visited the COMPANY_PROFILE pages (/directory/1, 2, 3,) and NEWS (/news/page 1, 2…) also accessed the TENDER page (/tenders/view/, /tenders/users/subscription, /tenders/users/login, /tenders/users/fax_subscription and /tenders/users/language).

Rule 8. ARTICLE=A NEWS=A 593 ==> INDEX=A 448 <conf :(0.76)> lift :(0.98) lev :(0) [-7] conv :(0.94)

This rule states that about 78% of visitors who visited the ARTICLE pages (/articles/starting-a-business, /articles/customs, articles/general-info, /articles/labour-law, /articles/labour-law) and NEWS (/news/page 1, 2…) also accessed the INDEX page (index.html).

5.3.3. Tender Aggregate and Single User Profile

In this study Web usage mining techniques are applied to identify users' navigational behavior. In order to improve website design, usability, and performance and personalize a Web site, the system should be able to distinguish between different users or groups of users. This process is called user profiling and its objective is the creation of an information base that contains the preferences, characteristics, and activities of the users.

This study apply two further options in order to identify either regarding each user as a member of a group and addressing the interest of each user individually. When addressing the users as a group, the method used (based on rules and patterns) is the construction of aggregate user profiles otherwise single user profile.

In addition to the relevant URLs that are extracted from the statistically in section 5.2.8.5 assigned to each profile, we can extract explicit information about the need of the users in each profile. Hence, for each profile, we accumulate all the frequently accessed URL. This allows us to describe each profile in terms of a set of significant URLs and IP address.

Moreover, the relevant URLs that are extracted using statistical analysis are assigned to each profile; in addition we can extract information about which URLs tend to visit by the user. To confirm this fact we need additional methods, these are clustering and association rules discovery. From section 5.3.1 and 5.3.2 the researcher now successfully identified TENDER URL accessed frequently. Besides from the pattern analysis, we can understand the TENDER resource is the most accessed one and hence need special attention. Therefore in this study TENDER is selected for developing user profile.

To develop the user profile we need filter tender URL and integrate with the content of tender web page in each single day accessed by the user when browsing the web site. The researcher identified the tender URL with the corresponding IP (user) and frequency by the customized Perl script. Only considering the tender URL which contain "view" with number example "/*tenders/view/66303*". The number with tender is the unique identification for tender for that particular content. The researcher develop the database by Access 2013 in order to identify

which user accessed which tender URL and which tender URL is the common interest of the users.

The researcher develop three table such us TENDER (TENDER_ID, CONTENT), HOST (HOST_ID, HOST (IP)) and ACCESSED (ID, TENDER_ID, URL, HOST, FREQUENCY, CONTENET, and HOST_ID) as shown in Figure 5.12.

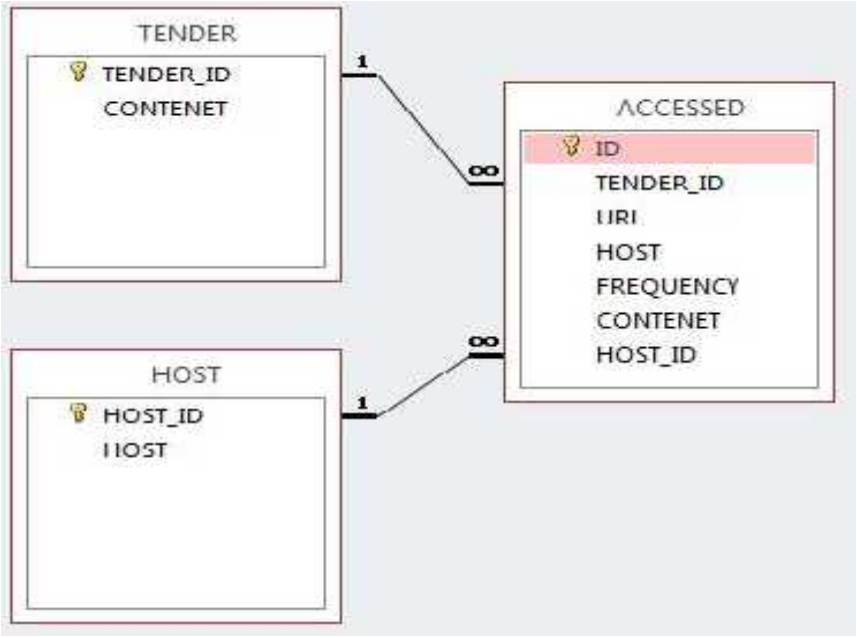

Figure 5.12: Relation Diagram for tables

5.3.4. Single User Profile

The researcher develop user profile by creating relationship between HOST table with ACCESSED table. As shown in Table 5.7 and Table 5.8 the result is single user profile (HOST_ID and HOST). Here we can identify each user with his interest.

HOST_ID	HOST
HOST_1011	197.156.77.131
HOST_1012	197.156.77.63
HOST_1015	82.145.211.99
HOST_1016	82.145.216.123
HOST_1017	82.145.218.154
HOST_1018	82.145.210.15
HOST_1022	82.145.211.117
HOST_1023	82.145.217.247
HOST_1036	82.145.209.115
HOST_1040	82.145.216.98
HOST_1041	82.145.218.181

Table 5.7: Single User Profile

Table 5.8: Expanded Single User Profile

5.3.5. Aggregate User Profile

The researcher develop aggregate user profile by creating relationship between TENDER table and ACCESSED table. As shown in Table 5.9 and Table 5.10 the result is aggregate user profile. Here we can identify aggregate user with their interest.

TENDER	CONTENET
63831	Commercial Bank of Ethiopia (CBE) invites all interested bidders to supply Learning Management S
63833	Commercial Bank of Ethiopia (CBE) invites bidders to supply the Mobile Money solution (Mobile W
63969	Sugar Corporation invites bidder for the Supply, Installation and Commissioning of GRP pipes and F
64119	African Green Revolution Forum (AGRF) invites bidder for the provision of the following goods and
64295	REQUEST FOR PROPOSAL: Consultancy services to conduct a Country Assessment of Essential Comm
64391	Call for expressions of interest (EoI) for Consultancy Services for the Construction of a Cold Room S
64392	Bid invitation for construction works
64425	Notification of Award of Works Contract: Design and Construction of F6 Junction - F4 Junction Road
64444	ISLAMIC RELIEF (IRE) invites bidders for the procurement of construction works
64456	GIZ invites bidders for the purchase of External Hard Disk, for offsite backup purpose
64491	የቡሌ ሆራ- ዩኒቨርሲቲ ለ2007 የበጀት አመት ከዚህ በታች በተጠቀሱት ሎቶ ያሉ ሪቃዎችን በዉረታ አወዳድሮ ለመግዛት ይፈልጋል

Table 5.9: Aggregate User Profile

TENDER	CONTENET
63831	Commercial Bank of Ethiopia (CBE) invites all interested bidders to supply Learning Mana

ID	HOST	FREQUENCY	HOST_ID	Click to Add
15839	82.145.209.155	2	HOST_740	
15840	197.252.1.221	2	HOST_5634	
15841	41.186.56.12	2	HOST_5633	
15842	197.156.86.73	2	HOST_5632	

63833	Commercial Bank of Ethiopia (CBE) invites bidders to supply the Mobile Money solution

ID	HOST	FREQUENCY	HOST_ID	Click to Add
5734	205.177.226.54	4	HOST_5797	
15836	98.250.162.252	2	HOST_5799	
15837	107.197.26.80	2	HOST_5798	
15838	84.108.9.184	2	HOST_5649	

63969	Sugar Corporation invites bidder for the Supply, Installation and Commissioning of GRP p

ID	HOST	FREQUENCY	HOST_ID	Click to Add
3560	213.55.109.81	6	HOST_255	
15829	82.145.209.155	2	HOST_740	
15830	197.156.111.14	2	HOST_671	
15831	213.55.106.159	2	HOST_5413	

Table 5.10: Expanded Aggregate User Profile

5.4. Evaluating the profile

The constructed users profile is finally evaluated by experts including web master, database administers, costumer and marketing manager. The evaluations tries to identify the extent of users' acceptance of the user profile constructed in this study. Summary of evaluation of experts' acceptance testing result is presented in table 5.11 below.

Evaluation of user profile		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1	User profile is helping visitors to quickly find relevant information on the web.	4	1	2	1	
2	User profile and companies user database are same.	5	1	2		
3	User profile is useful for user satisfaction.	7	1			
4	User profile Improving website design and usability.	8				
5	User profile can improving customer relation and loyalty.	8				
6	User profile can increase cross-reference by recommending pages related to the ones being considered.	7	1			
7	User profile can making results of information retrieval/search more aware of the context and user interest.	6	1	1		
8	User profile is an adaptive and easy to use.	3	3	1	1	
Sum		**48**	**8**	**6**	**2**	
Percentage		**75%**	**12.5%**	**9.4%**	**3.25%**	

Table 5.11: Evaluation of User Profile

According to the evaluation result obtained (see table 5.11 above) on the user profile, 87.5% of the expertise believe that the user profile is useful in all aspect of the questions. The rest 9.4% and 3.25% expertise are neutral and disagree respectively. This indicates that the created user profile is helpful in improving website design and usability, visitors to quickly find relevant information on the web, increase cross-reference by recommending pages related to the ones being considered etc..

Summary of the Findings

The research summarizes the analysis of the general usage of the Ebiz (e-business) website log dataset as follows:

The statistical analysis in section 5.2.1., indicated that the Home Page, Tender Page and Users Page are the most frequently accessed pages.

Experimentation with Ebiz (e-business) website log data shows the following strong and interesting Association Rules.

USER *==>* ***TENDER***

- *(USER pages (/users/login, /users/register, /users/upgrade, /users/password_reset, /users/account, /users/payment and /users/renew)) ==> (TENDER pages (/tenders/view/, /tenders/users/subscription, /tenders/users/login, /tenders/users/fax_subscription and /tenders/users/language)).*

This shows that users who access the USER pages also access the TENDER pages with a confidence level of over 90%.

(USER, INDEX) ==> TENDER

- *(USER pages (/users/login, /users/register, /users/upgrade, /users/password_reset, /users/account, /users/payment and /users/renew) page , INDEX (index.html)) ==> (TENDER page (/tenders/view/, /tenders/users/subscription, /tenders/users/login, /tenders/users/fax_subscription and /tenders/users/language)).*

This shows that users who access the USER and INDEX pages also access the TENDER pages with a confidence level of over 90%.

The clustering experimentation analysis of Ebiz (e-business) website log data allows us to conclude that regardless of the total number of accesses, 57% of the users of Ebiz tend to access resources whose URL is /tender/view/…/.html (TENDER). This finding is supported by the fact that tender page is the one that has taken the most accessed line of all Clusters.

The research summarized statistical, cluster and association analysis of Ebiz (e-business) website log data shows the following findings:

The statistical analysis showed that home (INDEX), tender and user page are the most frequently accessed pages and the association analysis discovered that 90 % of the users who accessed the USER pages are found to access the TENDER pages. Besides, 90% of the users who access the USER and INDEX pages have also accessed the TENDER pages. Hence, the above analysis clearly showed that TENDER page is the most frequently accessed page. Clustering analysis has also confirmed the above conclusion. Therefore, TENDER is successfully identified to be the most frequently accessed resource. As a result, it is selected for the construction of user profile.

This study created implicitly TENDER user profile by integrating tender database with the tender resource via by identification and analysis of frequent navigational pattern using web usage Mining technique from the log file of Ebiz (e-business) official website (2merkato.com).

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

The log files contain useful information for constructing user profile based on user modelling using web usage mining. The study focuses on the three main interdependent tasks for performing WUM; which are Preprocessing, Pattern Discovery and Pattern Analysis so as to construct user profile. The general website access statistics, pattern discovery, pattern analysis and construction of user profile was made. The statistical analysis and pattern discovery, have done based on 150 days of Ebiz (e-business) website usage.

In this paper, the complete statistical analysis of web server log files has been done by using WebLog Expert tool. This study has adopted an efficient user profile constructor tool. The adopted W3Perl is an open source Perl script help to find the user access behavior and also shows the percentage a user access behavior from all user statically. The log files were thoroughly preprocessed in order to make them suitable for the mining task. Preprocessing of these log files was a challenging and time taking task mainly because of the log file size. After the processing and data transformation was completed, several experiments were conducted to discover the most frequently accessed resource in order to select for the construction of user profile. Accordingly, the research found the useful information about the user of Ebiz (e-business) website.

The statistical analysis showed home (INDEX), tender and user page are the most frequently accessed pages and the association analysis discovered that 90 % of the users who accessed the USER pages are found to access the TENDER pages. Besides, 90% of the users who access the USER and INDEX pages have also accessed the TENDER pages. Hence, the above analysis clearly showed that TENDER page is the most frequently accessed page. Clustering analysis has also confirmed the above conclusion. Therefore, TENDER is successfully identified to be the most frequently accessed resource. As a result, TENDER is selected for the construction of user profile.

This study constructed implicitly TENDER User profile (single and aggregate) by integrating tender database with the tender resource via by identification and analysis of frequent navigational pattern using web usage Mining technique from the log file of Ebiz (e-business) official website (2merkato.com). Single user profile contains single user identification ID, HOST (IP) and different contents of the tender with its frequency whereas the aggregate profile also contains multiple user identification ID and HOST (IP) with its frequency overseas single tender content, which is the interest of multiple user.

Several results indicated that overseas visitors mostly access pages related to TENDER page. Therefore, the website administrator should ensure that the information in these pages is accurate. Also, once a visitor is identified his interest (TENDER), dynamically providing more links to pages related to tender or personalize the page would help users quickly access without asking desired information.

This research has attempted to answer the stated research questions and has achieved the goals. As a result, almost all of the problems stated in the statement of the problem, can be addressed to some degree by properly implementing the recommendations for website improvement. Conducting further researches recommended for future research can improve the website to its optimal level.

The strength of this research compared to similar other researches is that, this research introduces an identification and analysis of frequent navigational pattern integrating with company's tender database to construct user profile which can have a great contribution towards web personalization, Site Modification and integrating recommender system for online business and marketing applications. Moreover, the study introduces a set of association rules that achieves highest coverage for the dataset independent of the website's topology. It is based on usage patterns discovery and analysis describing users' behaviors and predicting what users will like based on their similarity to other users with relying on analyzing the content. According to the evaluation result obtained on the user profile, 87.5% of the expertise believe that the user profile is useful in all aspect of the questions.

However, in this study the user profile is representing pages that are frequently accessed by users. There are also pages that are infrequently visited which may have interesting patterns that can make the users profile complete.

6.2. Recommendations

While this study has revealed a number of interesting access patterns, further research needs to be conducted to expand the applicability of the result of this study.

Larger data sets with different resources

Using large sized web access log files might allow the data mining algorithms to discover more access patterns. However, in this study experiments consider transactions covering only five month visited server weblog file. Future researches need to consider integrated data obtained from Web Server, Proxy Server and Client so as to identify users' access and browsing patterns. Also, in this research, server logs with successful status are used. Future researches can consider those log files with error status to analyze both server and client side error patterns.

Visitor Identification

In this study, host (IP address), browsers and operating system are used to identify unique user (Visitors). However, different users may use the same host (IP address), browsers and operating system. Hence, we recommend as future research direction to use user's name while login to the system, cookies or more other sophisticated techniques to identify visitors so as to improve the end result.

Domain Knowledge

Incorporation of knowledge about the web site structure and path completion can provide more accurate visit information. The path completion task helps in filling in the missing page accesses. Incorporation of path completion with knowledge of website structure will provide more accurate information about user visits.

Removal of the web robots

The results obtained are to some extent inaccurate due to entries of web robots that did not access the WUMprep and the developed python algorism. Using better approaches like identifying web robots through the web browser information for their removal will improve results.

Infrequently Visited Pages

In this study only frequently accessed page are analyzed but feature sets having infrequently visited pages may discover interesting patterns of web access. So future research needs to consider them so as to make the users profile more complete.

Website Improvement

- Having of this pertinent information will help Ebiz (e-business) to develop more effective promotions better internet accessibility, inter-company communication and structure, and productive marketing skills. The webmaster may decide to make the entire Web site adaptive and customize it according to the profile of users. There for the Ebiz (e-business) official website should be restructure and personalized the page according to the whole rule set and user profile.
- After having information about access page views per visitor and entry page the user don't browse further than four pages into the site, therefore the web designer should be tactful to ensure that most important information is contained within four pages of the common site entry points.

Reference

[1] J. Pitkow, "Summary of WWW characterizations," *Web Journal,* vol. 2, no. 12, pp. 3-13, 1999.

[2] S. Cyrus, and B. Farnoush, "Effcient and Anonymous Web-Usage Mining for Web Personalization," *Amit Basu,* vol. 5, no. 02, pp. 1-48, 2002.

[3] R. Girargi and L. Balby Marinho, "A domain model of web recommender systems based on usage mining and collaborative filtering," Springer-Verlag, London, 2006.

[4] T. France, D. Yen , J.C. Wang, and C. M. Chang, "Integrating search engines with data mining for customer-oriented information search," *Information Management and Computer,* vol. 10(5), no. 03, p. 242–254, 2002.

[5] R. Baraglia and F. Silvestri, "Dynamic personalization of web sites without user intervention," *Communications of the,* vol. 50, no. 2, pp. 63-67, 2007.

[6] A. Rachit, "A Review Paper on Web Usage Mining and Pattern Discovery," *Journal of information, knowledge and research in computer engineering,* vol. 02, no. 02, pp. 279-284, 2013.

[7] S. K. Pani , L. Panigrahy, V.H.Sankar, B. K. Ratha, A.K.Mandal, and S.K.Padhi, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs," *International Journal of Instrumentation, Control & Automation (IJICA),* vol. 1, no. 1, pp. 15-23, 2011.

[8] J.Kim , B. Lee , M. Shaw , H. Chang, M. Nelson, "Application of decision-tree induction techniques to personalized advertisements on internet storefronts," *International Journal of Electronic Commerce,* vol. 5, no. 3, pp. 45-62, 2001.

[9] S. W. Changchien, and T. Lu, "Mining association rules procedure to support on-line recommendation by customers and products fragmentation," *Expert Systems with Applications,* vol. 20, no. 2, pp. 325-335, 2001.

[10] G. Sudhamathy and C. Jothi Venkateswaran, "An Efficient Hierarchical Frequent Pattern Analysis Approach for Web Usage Mining," *International Journal of Computer Applications,* vol. 43, no. 15, pp. 1-7, 2012.

[11] 2Mekato, "About 2Merkato," Ebis, 03 Sept 2014. [Online]. Available: http://www.2merkato.com/indx/about 2mekato. [Accessed 03 Sept 2014].

[12] A. Tadele, "Web usage pattern discovery: the case of Addis Ababa University official website," MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[13] F. Awet, "Web Usage: Exploring Navigational Behavior of Users, the Case of the Official website of Addis Ababa University," MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[14] N. Getahun, "Web usage pattern discovery and analysis by region: the case of Ethiopian airlines official website," MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2014.

[15] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns," *Knowledge and Information Systems,* vol. 1, no. 1, pp. 5-32, 1999.

[16] B.Lalithadevi , A.Merry Ida and W.Ancy Breen, "A New Approach for Improving World Wide Web Techniques in Data Mining," *International Journal of Advanced*, vol. 3, no. 1, pp. 243-251, Research in Computer Science and Software Engineering.

[17] A. J. L. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques," *Journal of Theoretical and Applied Information Technology (JATIT),* vol. 1, no. 1, pp. 67-73, 2010.

[18] R. Ankit Kharwar, A. Chandni Naik, and K. Niyanta Desai, "A Complete Pre Processing Method for Web Usage Mining," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 10, pp. 638-641, 2013.

[19] L. Shaily, B. Mehul, and M. Darshak, "Pre Pre-Processing: Procedure on Web Log File

for Web Usage Mining," *International Journal of Emerging Technology and Advanced Engineering,* vol. 2, no. 12, pp. 419-423, 2012.

[20] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," in *In Proceedings of the 9th IEEE International Conference on Tools with AI (ICTAI, 97),*, Minnesota, 1997.

[21] O. Etzioni, "The World Wide Web: Quagmire or gold mine," *Communications of the ACM,* vol. 39, no. 11, pp. 65-68, 1996.

[22] Beatric, Ryan Fernandes, Leo. J. Peo, Nikhila Kamat, and Sergius Miranda, "New Approaches to Web Personalization Using Web Mining Techniques," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 2, pp. 2195-2201, 2014.

[23] A. Sheetal Raiyani, S. Jain, and A.G. Raiyani, "Advanced Preprocessing using Distinct User Identification in web log usage data," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 1, no. 6, p. 2278 – 1021., 2012.

[24] J. Srivastva, R. Cooly, M. Deepande, and Pang-MingTan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *ACM SIGKDD Explorations,* vol. 1, no. 2, pp. 12-23, 2000.

[25] N. Lakshmi, R. Sekhara Rao, and S. Satyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 2, no. 4, pp. 274-279, 2013.

[26] N. Singh, A. Jain, and R. Shringar Raw,, "Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 3, no. 4, pp. 137-147, 2013.

[27] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai,, "Analysis Of Web Logs And Web User In Web Mining," *International Journal of Network Security & Its Applications*

(IJNSA), vol. 3, no. 1, pp. 99-110, 2011.

[28] R. Shukla, S. Silakari and P.K. Chande, "Web Personalization Systems and Web Usage Mining: A Review," *International Journal of Computer Applications*, vol. 72, no. 21, pp. 6-13, 2013.

[29] M. Eirinaki, and M. Vazirgirgiannis, "Web Mining for Web Personalization," *ACM Transactions on Internet Technology,* vol. 3, no. 1, p. 1–27, 2003.

[30] K.Nirosha V.Karthick, and Mrs.E.Jaya, "An Evaluation of Personalization Systems using Web Mining Techniques," *International Journal of Computer Science and Information Technology Research,* vol. 2, no. 2, pp. 265-270, 2014.

[31] P. Mehtaa, B. Parekh, K. Modi, and P. Solanki, "Web Personalization Using Web Mining: Concept and Research Issue," *International Journal of Information and Education Technology,* vol. 2, no. 5, pp. 510-512, 2012.

[32] S. Dhawan and M. Lathwal, "Study of Preprocessing Methods in Web Server Logs," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 5, pp. 430-433, 2013.

[33] K. Amsaveni, S. Vydehi, and M.Phil., "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites," *International Journal of Computer Trends and Technology,* vol. 3, no. 4, pp. 573-575, 2012.

[34] V. Ciesielski and A. Lalani, "Data Mining of Web Access Logs From an Academic Web Site," in *Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03),* VC, alalani, 2003.

[35] E. Cutrell, D. C. Robbins, S. T. Dumais, and R. Sarin, "Fast, Flexible Filtering with Phlat — Personal Search and Organization Made Easy," in *Personal Information Management*, Montréal, Québec, Canada, 2006.

[36] S. Sen ,B. Padmanabhan ,A. Tuzhilin , N. White and R. Stein, "On the Analysis of Web

Site Usage Data: How Much Can We Learn About the Consumer From Web Logfiles?," *European Journal of Marketing: Special Issue on Marketing in Cyberspace,* vol. 32, no. 7/8, pp. 125-136, 1998.

[37] C.P. Sumathi, R. Padmaja ,and T. Santhanam, "An Overview Of Preprocessing Of Web Log Files For Web Usage Mining," *Journal of Theoretical and Applied Information Technology,* vol. 34, no. 1, pp. 88-95, 2011.

[38] N. Khasawneh and C. C. Chan, "Active User-Based and Ontology- Based Web Log Data Preprocessing for Web Usage Mining," in *International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)*, Loanden, 2006.

[39] P. Zidrina, "Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining," in *24th Int. Conf. lnformafion Technology Interfaces*, Cavtat, Croatia, 2002.

[40] Jiawei Han and Micheline, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publisher, 2006.

[41] D. Patel, P. Kalpesh , and A. Patel, "Sessionization –A Vital Stage in Data Preprocessing of Web Usage Mining-A Survey," *International Journal of Engineering Research and Applications (IJERA),* vol. 2, no. 1, pp. 327-330, 2012.

[42] V. Jayakumar, and K. Alagarsamy, "Analysing Server Log File Using Web Log Expert In Web Data Mining," *International Journal of Science, Environment and Technology,* vol. 2, no. 5, pp. 1008-1016, 2013.

[43] C. L. Mugali, A.A. Maniyar, and P. Dandannavar, "Pre-Processing and Analysis of Web Server Logs," *International Journal of Innovative Research in Advanced Engineering (IJIRAE),* vol. 2, no. 8, pp. 46-55, 2015.

APPENDIX

Appendix I

URL Description

URL	Description
/administrator/index.php	Administrator page
/index.html	Root directory
/2merkato.com/administrator/index.php	Administrator page daily update the tender
/tenders	Tenders default page
/login	Users login page
/tenders/categories/upcat	Tenders categories page
/tenders/index.html	Tender tracking page
/index.php	The home page
/search	Search different page
/tenders/cron/task:send/	The tenders send to the subscribed user page
/tenders/cron/task:queue/secret	The tenders authentication page
/news	Business news information page
/tenders/page:2	Tenders page
/users/login	Users login page
/register	Users register page
/tenders/page:3	Tenders page

Appendix II

Python code for user identification

```
import re
import sys
from datetime import datetime, date, timedelta
from collections import Counter


#######################################################################
# Define the day of interest in the Apache Extended log format.   #
#######################################################################

try:
  daysAgo = int(sys.argv[1])
except:
  daysAgo = 1
  theDay = date.today() - timedelta(daysAgo)
  apacheDay = theDay.strftime('[%d/%b/%Y:')

#############################################
# Regex for the Apache common log format.  #
#############################################

parts = [
    r'(?P<host>\S+)',                   # host %h
    r'(?P<userid>\S+)',                 # indent %l (unused)
    r'(?P<user>\S+)',                   # user %u
    r'\[(?P<time>.+)\]',                # time %t
    r'"(?P<request>.*)"',               # request "%r"
    r'(?P<status>[0-9]+)',              # status %>s
    r'(?P<size>\S+)',                   # size %b (careful, can be '-')
    r'"(?P<referrer>.*)"',              # referrer "%{Referer}i"
    r'"(?P<agent>.*)"',                 # user agent "%{User-agent}i"
]
pattern = re.compile(r'\s+'.join(parts)+r'\s*\Z')

#################################
# Regex for a feed request.      #
#################################

feed = re.compile(r'/all-this/(\d\d\d\d/\d\d/[^/]+/)?feed/(atom/)?')

# Change Apache log items into Python types.
def pythonized(d):

# Clean up the request.
  d["request"] = d["request"].split()[1]

# Some dashes become None.
  for k in ("userid", "user", "referrer", "agent"):
    if d[k] == "-":
```

```
      d[k] = None

# The size dash becomes 0.
  if d["size"] == "-":
     d["size"] = 0
  else:
    d["size"] = int(d["size"])

# Convert the timestamp into a datetime object. Accept the server's time
zone.
  time, zone = d["time"].split()
  d["time"] = datetime.strptime(time, "%d/%b/%Y:%H:%M:%S")

  return d

# Is this hit a page?
def ispage(hit):
  # Failures and redirects.
  hit["status"] = int(hit["status"])
  if hit["status"] < 200 or hit["status"] >= 300:
    return False

  # Feed requests.
  if feed.search(hit["request"]):
    return False

  # Requests that aren't GET.
  if hit["request"][0:3] != "GET":
    return False
##############################################
# Images, sounds, etc.                       #
##############################################

  #if hit["request"].split()[1][-1] != '/':
    #return False

# Must be a page.
  return True


####################################
#IP Adderess Extractor             #
####################################



def extract_ip(line):
    return line.split()[0]


########################################
#IP Adderess Counter                   #
########################################
```

```
def increase_count(ip_dict, ip_addr):

    if ip_addr in ip_dict:
       ip_dict[ip_addr] += 1

    else:
       ip_dict[ip_addr] = 1


    return  ip_dict

#######################################
#Read the in put file                 #
#######################################

def read_ips(infilename):
    res_dict = {}
    log_file = file(infilename)
    for line in log_file:
        m = pattern.match(line)
        if m is None :
            continue
        hit = m.groupdict()
        if ispage(hit):
           ip_addr = extract_ip(line)
           increase_count(res_dict, ip_addr)
           tpages.append(pythonized(hit))
        else:
           continue

    return res_dict , tpages

#######################################
#Write the out put on the text file   #
#######################################

def write_ips(outfilename, res_dict):

        p = []
        user = 0
        ip_cont = []
        ip_use = {}
        log_file = file(infilename)
        out_file = file(outfilename, "a")


#############################################################
# Uniq user  Identification  by Uniq IP                     #
#############################################################
```

```
        for line in log_file:
            k =line.split()
            if line.isspace():
                continue
            p= k[0]
            m = pattern.match(line)
            if m is None :
                continue
            hit = m.groupdict()

            for ip_addrs, count in
sorted(res_dict.iteritems(),key=valuesort,reverse=False)[:]:

                 if  p in ip_addrs and count == 1:
                    if ispage(hit):
                       user += 1
                       ip_cont.append(p)
                       out_file.write("%5d\t%s" % (user ,line))
                    else:
                        continue

                 else:
                     continue

        return  user

def user_identification(outfilename, res_dict, tpages , user):

          singlhit = []
          newclientagent = []
          holduserno = {}
          newuser=0
          p=[]
          log_file = file(infilename)
          out_file = file(outfilename, "a")
          for line in log_file:
            k =line.split()
            if line.isspace():
                 continue
            p= k[0]
            m = pattern.match(line)
            if m is None :
                 continue
            hit = m.groupdict()

            for ip_addr, count in
sorted(res_dict.iteritems(),key=valuesort,reverse=False)[:]:

                 if  p in ip_addr and count == 1 :
                   break

                 if  p in ip_addr and count > 1 :
                      for z in tpages:
```

```
                        if p == z["host"]:
                            if ispage(hit):
                                singlhit = pythonized(hit)
                            else:
                                 continue
                            if singlhit["agent"] == z["agent"] and
newuser > 0:
                                 if z["agent"] in newclientagent:
                                    for ip_add, coun in
sorted(holduserno.items()):
                                        if p == ip_add :
                                          out_file.write("%5d\t%s" %
(coun ,line))
                                          break
                                        else:
                                          continue

                                 else:


newclientagent.append(singlhit["agent"])
                                     user += 1
                                     out_file.write("%5d\t%s" % (user
,line))
                                     holduserno[p] = user
                                     newuser += 1

                            elif singlhit["agent"] == z["agent"]:

                                  if z["agent"] in newclientagent:

                                      for ip_add, coun in
sorted(holduserno.items()):
                                          if p == ip_add :
                                              out_file.write("%5d\t%s" %
(coun ,line))

                                          else:
                                              continue
                                  else:


newclientagent.append(singlhit["agent"])
                                     user += 1
                                     out_file.write("%5d\t%s" % (user
,line))
                                     holduserno[p] = user
                                     newuser += 1

                            else:

                                  if z["agent"] in newclientagent:
```

```
                                        for ip_add, coun in
sorted(holduserno.items()):
                                            if p == ip_add :
                                                out_file.write("%5d\t%s" %
(coun ,line))

                                            else:
                                                continue
                                    else:

                                       user += 1

newclientagent.append(singlhit["agent"])
                                       out_file.write("%5d\t%s" % (user
,line))
                                       holduserno[p] = user
                                       newuser += 1

                           else:
                               continue

                    else:
                        continue

            out_file.close()

############################################################
#To order the out put in decreasing/incresasing order   #
############################################################

def valuesort(tuple):
  return tuple[1]


############################################################
#To enter the input file and the the out put file       #
############################################################

res_dict = {}
ip_dict = {}
tpages = []
pages = []
infilename = raw_input ('የቴክስት ፋይሉን ስም ያስገቡ: ')
outfilename = raw_input ('የቴክስት ፋይሉን ስም ያስገቡ: ')
res_dict ,tpages = read_ips(infilename)
user = write_ips(outfilename, res_dict)
user_identification(outfilename, res_dict, tpages , user)
```

Appendix III

Python code for robot remover from the record

```
import re
import sys

#This regular expression is the heart of the code.
#Python uses Perl regex, so it should be readily portable


COMBINED_LOGLINE_PAT = re.compile(
  r'(?P<origin>\d+\.\d+\.\d+\.\d+) '
+ r'(?P<identd>-|\w*) (?P<auth>-|\w*) '
+ r'\[(?P<date>[^\[\]:]+):(?P<time>\d+:\d+:\d+) (?P<tz>[\-
\+]?\d\d\d\d)\] '
+ r'"(?P<method>\w+) (?P<path>[\S]+) (?P<protocol>[^"]+)"
(?P<status>\d+) (?P<bytes>-|\d+)'
+ r'( (?P<referrer>"[^"]*")( (?P<client>"[^"]*")(
(?P<cookie>"[^"]*"))?)?)?\s*\Z'
)

#Patterns in the client field for sniffing out bots
BOT_TRACES = [
    (re.compile(r".*http://help\.yahoo\.com/help/us/ysearch/slurp.*"),
"Yahoo robot"),
    (re.compile(r".*\+http://www\.google\.com/bot\.html.*"), "Google
robot"),

(re.compile(r".*\+http://about\.ask\.com/en/docs/about/webmasters.shtml.
*"), "Ask Jeeves/Teoma robot"),
    (re.compile(r".*\+http://search\.msn\.com\/msnbot\.htm.*"),"MSN
robot"),

(re.compile(r".*http://www\.entireweb\.com/about/search_tech/speedy_spid
er/.*"), "Speedy Spider"),
    (re.compile(r".*\+http://www\.baidu\.com/search/spider_jp\.html.*"),
"Baidu spider"),

(re.compile(r".*\+http://www\.gigablast\.com/spider\.html.*"),"Gigabot
robot"),
    (re.compile(r".*http://www\.uptimerobot\.com/.*"), "Uptime robot"),
    (re.compile(r".*\+http://www\.freewebmonitoring\.com/bot\.html.*"),
"Freewebmonitoring robot"),
    (re.compile(r".*\+http://www\.bing\.com/bingbot\.htm.*"), "Bing
robot"),
    (re.compile(r".*\+http://www\.exabot\.com/go/robot.*"), "Exa
robot"),
    (re.compile(r".*BOT\ for\ JCE.*"), "JCE robot"),
    (re.compile(r".*\+http://yandex\.com/bots*"), "Yandex robot"),
    (re.compile(r".*\+http://ahrefs\.com/robot/.*"), "Ahrefs robot"),
    (re.compile(r".*\+http://www\.flamingosearch\.com/bot.*"), "Flamingo
robot"),
```

```
    (re.compile(r".*http://help\.naver\.com/robots/.*"), "Naver Corp
robot"),
    (re.compile(r".*\+http://go\.mail\.ru/help/robots.*"), "RU Bot
robot"),
    (re.compile(r".*http://www\.majestic12\.co\.uk/bot\.php\?\+.*"),
"MJ12 robot"),
    (re.compile(r".*NerdyBot.*"), "Nerdy robot"),
    (re.compile(r".*ContextAd Bot 1.0.*"), "ContextAd robot"),
    (re.compile(r".*FeedBot.*"), "Feed robot"),
    (re.compile(r".*Facebot/ 1.0.*"), "Face robot"),
    (re.compile(r".*UnisterBot;\ crawler@unister\.de.*"),
"crawler@unister robot"),
    (re.compile(r".*Twitterbot/1.0.*"), "Twitter robot"),
    (re.compile(r".*support\.voilabot@orange-ftgroup\.com.*"), "Voila
robot"),
    (re.compile(r".*http://www\.reklama-i-rabota\.ru/.*"), "Reklama
robot"),
    (re.compile(r".*http://www\.opensiteexplorer\.org/dotbot.*"),
"Opensiteexplorer robot"),
    (re.compile(r".*\+http://OpenLinkProfiler\.org/bot.*"),
"OpenLinkProfiler robot"),
    (re.compile(r".*\+http://www\.google\.com/mobile/adsbot\.html.*"),
"Google2 robot"),
    (re.compile(r".*\+http://fulltext\.sblog\.cz/.*"), "Seznam robot"),
    (re.compile(r".*http://www\.moz\.com/dp/rogerbot.*"), "Roger
robot"),
    (re.compile(r".*http://support\.paper\.li/entries/20023257-what-is-
paper-li.*"), "Support robot"),
    (re.compile(r".*\+http://www\.seoengine\.com/seoengbot\.htm.*"),
"Seoeng robot"),
    (re.compile(r".*BufferBot.*"), "Buffer robot"),
    (re.compile(r".*http://squirro\.com/squirrobot/.*"), "Squirro
robot"),
    (re.compile(r".*\+http://www\.g2reader\.com/.*"), "G2reader robot"),
    (re.compile(r".*\+http://import\.io.*"), "Import robot"),
    (re.compile(r".*http://linkfluence\.net/.*"), "Linkfluence robot"),
    (re.compile(r".*http://www\.wesee\.com/bot/.*"), "Wesee robot"),
    (re.compile(r".*http://www\.feedspot\.com.*"), "Feeds robot"),
    (re.compile(r".*finbot.*"), "Fin robot"),
    (re.compile(r".*\+http://www\.grouphigh\.com/bot\.html.*"),
"Grouphigh robot"),
    (re.compile(r".*\+http://www\.seoengine\.com/seoengbot\.htm.*"),
"Seoengine robot"),
    (re.compile(r".*\+http://tweetmeme.\com/.*"), "Tweetmeme robot"),
    (re.compile(r".*http://everyonesocial\.com/.*"), "Everyonesocial
robot"),
    (re.compile(r".*http://linkfluence\.net/.*"), "Linkfluence robot"),
    (re.compile(r".*\+http://tweetmeme\.com/.*"), "Tweetmeme robot"),
    (re.compile(r".*mailto:crawling@ubermetrics-technologies\.com.*"),
"Mailto robot"),
    (re.compile(r".*http://intro\.squirro\.com/squirrobot/.*"), "Squirro
robot"),
```

```
    (re.compile(r".*\+http://www\.career-x\.de/bot\.html.*"), "Career
robot"),
    (re.compile(r".*\+http://yandex\.com/bots.*"), "Yandex robot"),

(re.compile(r".*\+http://www\.sogou\.com/docs/help/webmasters\.htm#07.*"
), "Sogou robot"),
    (re.compile(r".*http://commoncrawl\.org/faq/.*"), "Commoncrawl
robot"),
    (re.compile(r".*\+http://import\.io.*"), "Import robot"),
    (re.compile(r".*http://www\.Nutch\.de/.*"), "Nutch robot"),
    (re.compile(r".*\+http://OpenLinkProfiler\.org/bot.*"), "OpenLink
robot"),
    (re.compile(r".*\+https://www\.mojeek\.com/bot\.html.*"), "Mojeek
robot"),
    (re.compile(r".*http://www\.trendiction\.de/bot.*"), "Trendiction
robot"),
    (re.compile(r".*http://www\.crystalsemantics\.com/service-
navigation/imprint/useragent/.*"), "Crystalsemantics robot"),

(re.compile(r".*\+http://www\.archive\.org/details/archive\.org_bot.*"),
"Archive robot"),
    (re.compile(r".*\+http://www\.linkdex\.com/bots/.*"), "Linkdex
robot"),
    (re.compile(r".*\+http://www\.seokicks\.de/robot\.html.*"),
"Seokicks robot"),
    (re.compile(r".*\+http://webmeup-crawler\.com/.*"), "Webmeup
robot"),
    (re.compile(r".*http://www\.googlebot\.com/bot.html.*"), "Google3
robot"),
    (re.compile(r".*http://ow\.ly/zXWuP.*"), "Ow robot"),
    (re.compile(r".*http://showyou\.com/crawler.*"), "Showyou robot"),
    (re.compile(r".*http://www\.reklama-i-rabota\.ru/.*"), "Reklama
robot"),
    (re.compile(r".*\+http://cliqz\.com/company/cliqzbot.*"), "Cliqz
robot"),
    (re.compile(r".*Searcharoo\.NET.*"), "Searcharoo robot"),
    (re.compile(r".*\+http://www\.profound\.net/urlappendbot\.html.*"),
"Profound robot"),
    (re.compile(r".*\+http://www\.datagnion\.com/bot\.html.*"),
"Datagnion robot"),
    (re.compile(r".*http://disq\.us/8jhg3h.*"), "Disq robot"),
    (re.compile(r".*\+http://www\.ScreenerBot\.com.*"), "Screener
robot"),
    (re.compile(r".*Facebot\/1.0.*"), "Face robot"),
    (re.compile(r".*\+http://www\.xovibot\.net/.*"), "Xovi robot"),

(re.compile(r".*RediffNewsBot;betatest;contact:searchops@rediff\.co\.in.
*"), "RediffNews robot"),
    (re.compile(r".*http://fb\.me/18xkWmYAi.*"), "Fb robot"),
    (re.compile(r".*http://nutch\.apache\.org/bot\.html.*"), "Apache
robot"),
    (re.compile(r".*\+http://www\.200please\.com/bot.*"), "Please
robot"),
```

```
    (re.compile(r".*Phantom\.js\bot.*"), "Phantom robot"),
    (re.compile(r".*http://www\.rssing\.com.*"), "Rssing robot"),
    (re.compile(r".*Phantom\.js\ bot.*"), "Phantom robot"),
    (re.compile(r".*http://ow\.ly/zXWmn\Livelapbot\/0.1.*"), "Livelap
robot"),
    (re.compile(r".*Feedspotbot:\http://www\.feedspot\.com.*"), "Feeds
robot"),
    (re.compile(r".*http://www\.reklama-i-rabota.\ru/.*"), "Reklama
robot"),
    (re.compile(r".*DomainTools.*"), "Domain robot"),
    (re.compile(r".*mailto:\crawling@ubermetrics-technologies\.com.*"),
"Mailto robot"),

(re.compile(r".*\+http://www\.easou\.com/search/spider\.html\.com.*"),
"Easou robot"),
    (re.compile(r".*\+http://tweetedtimes\.com.*"), "Tweetedtimes
robot"),
    (re.compile(r".*Googlebot-Image/1.0.*"), "Googl3 robot"),
    (re.compile(r".*Gecko/20061204firefox/2.0.0.1.*"), "Gecko robot"),
    (re.compile(r".*Exabot-Thumbnails.*"), "Exa robot"),
    (re.compile(r".*\uMBot-FC/1.0;\mailto:\crawling@ubermetrics-
technologies\.com.*"), "UM robot"),
    (re.compile(r".*\+http://www\.baidu\.com/search/spider\.html.*"),
"Baidu robot"),
    (re.compile(r".*\+http://www\.easou\.com/search/spider\.html.*"),
"Easou robot"),
    (re.compile(r".*\+http://www\.proximic\.com/info/spider\.php.*"),
"Proximic robot"),
    (re.compile(r".*360Spider.*"), "360 robot"),
    (re.compile(r".*YisouSpider.*"), "Yisou robot"),
    (re.compile(r".*spider.*"), "Spider robot"),
    (re.compile(r".*MPDP-ALR-Search-Bot.*"), "MPDP robot"),
    (re.compile(r".*xintellibot.*"), "Xintelli robot"),
    (re.compile(r".*Gecko/20061204firefox/2.0.0.1.*"), "Gecko robot"),
    (re.compile(r".*\+http://tweetedtimes\.com.*"), "Tweetedtimes
robot"),
    (re.compile(r".*http://disq\.us/8jjocf.*"), "Disq robot"),
    (re.compile(r".*acapbot/0.1;treat.*"), "Acap robot"),
    (re.compile(r".*KomodiaBot/1.0.*"), "Komodia robot"),
    (re.compile(r".*Insitesbot/1.0.*"), "Insite robot"),
    (re.compile(r".*Livelapbot/0.1.*"), "Livelap robot"),
    (re.compile(r".*mailto:\crawling@ubermetrics-technologies\.com.*"),
"Mailto robot"),
    (re.compile(r".*\+http://webmeup-crawler\.com/.*"), "Webmeup
robot"),
    (re.compile(r".*www\.radian6\.com/crawler.*"), "Radian6 robot"),
    (re.compile(r".*fr_crawler.*"), "Radian6 robot"),
    (re.compile(r".*\+http://www\.grapeshot\.co\.uk/crawler\.php.*"),
"Grapeshot robot"),
    (re.compile(r".*\+http://www\.brandwatch\.net.*"), "Brandwatch
robot"),
    (re.compile(r".*\+http://www\.alexa\.com/site/help/webmasters.*"),
"Alexa robot"),
```

```
    (re.compile(r".*\+http://filterdb\.iss\.net/crawler/.*"), "Filterdb
robot"),
    (re.compile(r".*simplenewz\.com\crawler\-\admin@simplenewz\.com.*"),
"Fimplenewz robot"),
    (re.compile(r".*http://crawler\.sistrix.\net/.*"), "Sistrix robot"),
    (re.compile(r".*FRCrawler.*"), "FRC robot"),
    (re.compile(r".*admin@simplenewz\.com.*"), "Admin robot"),
    (re.compile(r".*\+http://www\.netseer\.com/crawler\.html;.*"),
"Netseer robot"),
    (re.compile(r".*http://crawler\.sistrix\.net/.*"), "Sistrix robot"),
    (re.compile(r".*GermCrawler.*"), "Germ robot"),
    (re.compile(r".*ops@percolate\.com.*"), "Ops robot"),
    (re.compile(r".*GT-S5301\Build/IMM76D.*"), "GT robot"),
    (re.compile(r".*http://jetsli\.de/crawler.*"), "Jetsli robot"),
    (re.compile(r".*MSIECrawler.*"), "MSIEC robot"),
    (re.compile(r".*http://jetsli\.de/crawler.*"), "Jetsli robot"),
    (re.compile(r".*www\.integromedb\.org/Crawler.*"), "Integromedb
robot"),
    (re.compile(r".*WebTarantula\.com\Crawler.*"), "WebTarantula
robot"),
    (re.compile(r".*URLfilterDB-crawler/1.1.*"), "URLfilterDB robot"),
    (re.compile(r".*Oxyme\.Search - Web crawler.*"), "Oxyme robot"),

(re.compile(r".*Healthbot\/Health\_and\_Longevity\_Project\_(HealthHaven
\.com).*"), "*Health robot"),
    (re.compile(r".*\+http://www\.diffbot\.com.*"), "Diff robot"),
    (re.compile(r".*\+http://archive\.org/details/archive\.org_bot.*"),
"Archive robot"),
    (re.compile(r".*\+http://blekko\.com/about/blekkobot.*"), "Blekko
robot"),
    (re.compile(r".*\+http://www\.meanpath\.com/meanpathbot\.html.*"),
"Meanpath robot"),
    (re.compile(r".*\+http://overx50\.com.*"), "Overx50 robot"),
    (re.compile(r".*niki-bot.*"), "Niki robot"),
    (re.compile(r".\+http://www\.accelobot\.com.*"), "Accelo robot"),
    (re.compile(r".Bichoo Spider.*"), "Bichoo robot"),
    (re.compile(r".JOC Web Spider.*"), "JOC  robot"),
    (re.compile(r".Xaldon WebSpider.*"), "Xaldon  robot"),
    (re.compile(r".New-Sogou-Spider/1.0.*"), "New  robot"),
    (re.compile(r".http://blog\.wasalive\.com/wasalive-bots/.*"), "Bog
robot"),
    (re.compile(r".http://www\.setooz\.com/oozbot\.html.*"), "Setooz
robot"),
    (re.compile(r".Screaming Frog SEO Spider/2.40.*"), "SEO  robot"),
    (re.compile(r".Blogshares Spiders.*"), "SEO  robot"),
    (re.compile(r".Spider.*"), "Spider robot"),
    (re.compile(r".Screaming Frog SEO Spider/2.55.*"), "SEO robot"),
    (re.compile(r".http://www\.ranks\.nl/.*"), "Ranks robot"),
    (re.compile(r".http://www\.ranks\.nl/tools/spider\.html.*"), "Ranks
robot"),
]

orig_stdout = sys.stdout
```

```
f = file(’merkato1.txt’, ’w’)
sys.stdout = f
g = open(’merkato.txt’, ’r’)
for line in g:
    match_info = COMBINED_LOGLINE_PAT.match(line)
    if not match_info:
        print  ("Unable to parse log line\n")
        continue
    isbot = False
    for pat, botname in BOT_TRACES:
        if pat.match(match_info.group(’client’)):
            isbot = True
            break
    if not isbot:
        sys.stdout.write(line)

sys.stdout = orig_stdout
f.close()
```

Appendix IV

Weka association rule discovery sample outputs

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.01

Relation: 1.Final Weka Data Set For Association

Instances: 42153

Attributes: 6

ARTICLE

COMPANY_PROFILE

NEWS

TENDER

USER

INDEX

=== Associator model (full training set) ===

FPGrowth found 8 rules (displaying top 8)

1. [USER=A]: 6984 ==> [TENDER=A]: 6286 <conf:(0.9)> lift:(1.39) lev:(0.04) conv:(3.54)

2. [INDEX=A, USER=A]: 6123 ==> [TENDER=A]: 5486 <conf:(0.9)> lift:(1.39) lev:(0.04) conv:(3.4)

3. [USER=A]: 6984 ==> [INDEX=A]: 6123 <conf:(0.88)> lift:(1.14) lev:(0.02) conv:(1.88)

4. [TENDER=A, USER=A]: 6286 ==> [INDEX=A]: 5486 <conf:(0.87)> lift:(1.14) lev:(0.02) conv:(1.82)

5. [USER=A]: 6984 ==> [INDEX=A, TENDER=A]: 5486 <conf:(0.79)> lift:(1.8) lev:(0.06) conv:(2.63)

6. [NEWS=A, USER=A]: 677 ==> [TENDER=A]: 527 <conf:(0.78)> lift:(1.2) lev:(0) conv:(1.59)

7.[NEWS=A, COMPANY_PROFILE=A]: 816 ==> [TENDER=A]: 619 <conf:(0.76)> lift:(1.17) lev:(0) conv:(1.46)

8. [NEWS=A, ARTICLE=A]: 593 ==> [INDEX=A]: 448 <conf:(0.76)> lift:(0.98) lev:(0) conv:(0.94)

Appendix V

Questioners

Part 1: Demographic profile of the respondent

Answer the following questions by putting the (√) symbol on the following boxes or write in the space provided.

1. Specify your gender?

   ☐ Male ☐ Female

2. Your age?

   ☐ 25–34 ☐ 35–44 ☐ 45–54 ☐ 55–64

   Others _____

3. Your educational level?

   ☐ Diploma ☐ Master's Degree

   ☐ Bachelor's Degree ☐ PhD (Doctorate Degree)

   Others _____

4. Working experience in the organization?

   ☐ < 5 ☐ 5-9 ☐ 10-14 ☐ 15-19 ☐ > 20

   Others _____

5. Job Title?

   ☐ Web Designer ☐ Database administrator ☐ Customer manager

   ☐ Web administrator and master ☐ Marketing expert

   Others _____

Part 2: Evaluation user profile

	Please indicate the extent to which you agree or disagree with the following statements by putting a tick (√) mark in the appropriate box.	**Strongly Agree**	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**
1	User profile is helping visitors to quickly find relevant information on the web.					
2	User profile and companies user database are same.					
3	User profile is useful for user satisfaction.					
4	User profile Improving website design and usability.					
5	User profile can improving customer relation and loyalty.					
6	User profile can increase cross-reference by recommending pages related to the ones					
7	User profile can making results of information retrieval/search more aware of the context and user interest.					
8	User profile is an adaptive and easy to use.					

DECLARATION

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

This thesis has been submitted for examination with my approval as university advisor.

_____

Advisor