WEB USAGE PATTERN DISCOVERY AND ANALYSIS FOR WEBSITE OPTIMIZATION: THE CASE OF ETHIO TELECOM OFFICIAL WEBSITE

A thesis submitted to College of Natural Sciences of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Science in Information Science

BY

Senait Mezgebu

ADDIS ABABA UNIVERSITY

MAY, 2015

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

DEPARTMENT OF INFORMATION SCIENCE

WEB USAGE PATTERN DISCOVERY AND ANALYSIS FOR WEBSITE OPTIMIZATION: THE CASE OF ETHIO TELECOM OFFICIAL WEBSITE

BY

SENAIT MEZGEBU TEMERE

APPROVED BY: THE BOARD OF EXAMINERS

_____	_____
Chair person, Departments Graduate committee	Signature
Million Meshesha (PhD) _____	_____
Advisor	Signature
Dereje Teferi (PhD) _____	_____
Internal Examiner 1	Signature
Tibebe Beshah (PhD) _____ __	_____
Internal Examiner 2	Signature

Dedicated

To

My 4 years old Son, Nathanan Abebe

Contents

ACKNOWLEDGMENT

A great debt of gratitude is owed to my advisor Dr. Million Meshesha for his unreserved fruitful advices, comments and discussions.

I am also deeply grateful to my husband Abebe Regassa for his indispensable support and valuable comments.

I would like to express my science gratitude to web master of Ethio telecom official website and web server administrators for their cooperation in providing me the information required for this research.

Last but not least I would like to thank my lovely kids for their patience while I was busy with my Thesis: My daughter, Fenet Abebe and My son, Nathanan Abebe.

Senait Mezgebu

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ADSL Asymmetric digital subscriber line

CLF Common Log Format

CMS Content Management System

CRBT Call ring back tone

CSV Comma-Separated Values

ECLF Extended Common Log Format

ETC Ethiopian Telecommunication Corporation

FP Frequent Pattern

HTTP Hypertext transfer protocol

KDD Knowledge Discovery in Data

SEO Search Engine Optimization

SQL structured query language

SSL Secured shell

URL Uniform Resource Locator

W3C World Wide Web Consortium

WEKA Waikato Environment for Knowledge Analysis

WUM prep web usage mining preprocessing

WUM Web Usage Mining

ABSTRACT

In this ever growing Internet era, websites are becoming among the most important media for communicating with the stakeholders. Nowadays, many organizations realized that the need to investigate the behavior of their website users to meet their objectives through undertaking a research. Ethio telecom being a sole telecom operator and Internet service provider in Ethiopia should continuously assess and monitor its official website in order to analyze customers' usage behavior and restructure the website accordingly. In this study, an attempt is made to discover useful patterns from the server log files of Ethio Telecom Official website using web usage mining. The current research follows, the Web Usage Mining processes model suggested by Sharma [21] which consists of, data collection, data preprocessing, pattern discovery and pattern analysis. Server log data was used for pattern discovery and Google analytics reports were exported for statistically analyzing the website usage.

The access Log files exported from the web server cannot be used directly for web usage mining task as it may consist of large amount of irrelevant information. So preprocessing on web usage data is required to eliminate noisy data and make data effective for further analysis task. The applied preprocessing task includes data cleaning, session identification, feature selection, transaction identification and transformation of the data to Weka understandable format. Finally, a total of 301,580 transactions are used for the experiment.

After the preprocessing was completed, experiment was conducted with the datasets using Weka Software and FP-Growth algorithm to discover interesting patterns. Google Analytics and MS Excel were also employed to yield different useful statistical reports including, visitor's location, top channels, top landing, top exit, and most frequently accessed pages. Some of the behavior identified from the statistical analysis shows the access rate of new visitors exceed the existing visitors, major visitors of the website are located locally and the home page is listed from the top list of landing as well as exit page.

The experimental findings indicate the existence of strong relationship between internet page and business page where all visitors who visited internet page also accessed business page and vice versa. Besides, these visitors who had visited the internet page and business page also visited troubleshooting page which implies that visitors are looking for a set up procedure from trouble shooting page after getting information about the available internet and business products from respective pages. Furthermore, the experimental findings also pointed out discussion pages, i.e.

Forum, general discussion and call conference pages are most frequently accessed together implying these pages need to be directly linked while restructuring the website.
Moreover, promotion is required to make the website more popular especially with respect to referrals and the website global usage. Finally, recommendations are forwarded for further researches and website reconstruction.

Keywords: Web Usage Mining, Pattern Discovery, FP-Growth, Google Analytics, Web log analysis.

CHAPTER ONE

INTRODUCTION

1.1. Background of the study

The World Wide Web consists of billions of web pages. Nowadays, World Wide Web is the most popular means of transferring information. The Web has opened a new way of doing businesses; amazon.com the so-called "on-line Wal-Mart" is one of the examples. The growth of the web has resulted in a huge amount of information. Several kinds of data, such as text, image, audio or video have to be handled and organized over a web so that it can be accessed by many users effectively and efficiently [1].

The users want to have effective search tools to find relevant information easily and precisely. Web service providers want to find ways to predict users' behaviors and personalize information to reduce the traffic load and design the website suited for the different group of users [2].

Data mining is a field of study that is used to discover new and useful knowledge from large data sources such as data stored in large databases, warehouses and other massive information repositories [1]. Web mining is the application of data mining techniques to discover patterns from the Web. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. According to the differences of mining objects, web mining can be divided into three different types [3]: which are Web usage mining, Web content mining and Web structure mining. Web content mining analyzes web content such as text, multimedia data and structured data within web pages or linked across web pages. On the other hand, web structure mining is the process of using graphs and network mining theory and methods to analyze the nodes and connection structures on the web. It extracts patterns from hyperlinks, where a hyperlink is a structural component that connects a web page to another location.

Web usage mining which is the focus of this study is the application of data mining techniques to discover interesting usage patterns from log file, in order to understand and better serve needs of Web-based applications. Web usage mining is the process of finding out what users are looking

for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site [2].

Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – users, Web service providers and business analysts. The users want to have effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the users/consumers' needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, Web mining becomes an active and popular research field [4].

1.2. Background of the company

Ethiopian Telecommunication Corporation (ETC), currently named as Ethio telecom, is the oldest public telecommunications operator in Africa. It has been providing various telecom services such as fixed line telephony, mobile, Internet and value added services to the country since its establishment, 1894 [35].

In the year 2010, the Ethiopian government has decided to transform the telecommunication infrastructure and services to world class standard, considering them as a key lever in the development of Ethiopia. Thus, Ethio telecom was born from this ambition in order to bring about a paradigm shift in the development of the telecom sector to support the steady growth of the country [35].

Since its first launch in 1998, Ethio telecom website has gone through two major renovations. The current website, which was launched in 2014, is developed in-house by gathering requirements from different stakeholders across the organization. It is a bilingual website presenting its contents

in two languages, namely Amharic and English though the home page is displayed in the default English language while accessing the site for the first time with domain name, http://www.ethiotelcom.et/. The website is dynamic and a multipurpose one serving from simple information display to some e-commerce activities [7].

The major contents and services that are available via the website are[7]: Caller Ring Back Tone(CRBT) media selection, Asymmetric digital subscriber line(ADSL) service ,Telephone directory, Vacancy information, BID information, Contact information for support on telecom services ,Forum , Press Release and Product News Management , Feedback Management and Product catalogue details.

Some of the software's and platforms used to implement and deploy the website from client side are ASP. NET, HTML, Java Script, JQuery, CSS and AJAX. And also SQL Server 2000 and Linux are used from server side.

Ethio telecom uses a professional Content Management System (CMS) to manage the website content and structure [7].

As a world class telecom operator and internet service provider, Ethio-telecom has the following future plan regarding the website [7]:

- To develop Mobile Version of the website:
- To improve the look and feel of the website
- To maintain the quality of content.
- To design shop locators using Google map.
- To design a map indicating service coverage using Google map.
- To produce video contents like mobile service configuration procedures.
- To develop Searchable Amharic Telephone directory.
- To integrate the website with ZSMART self-care system and other service platforms.

The current Ethio telecom official website map structure is shown in figure: 1-1 below.

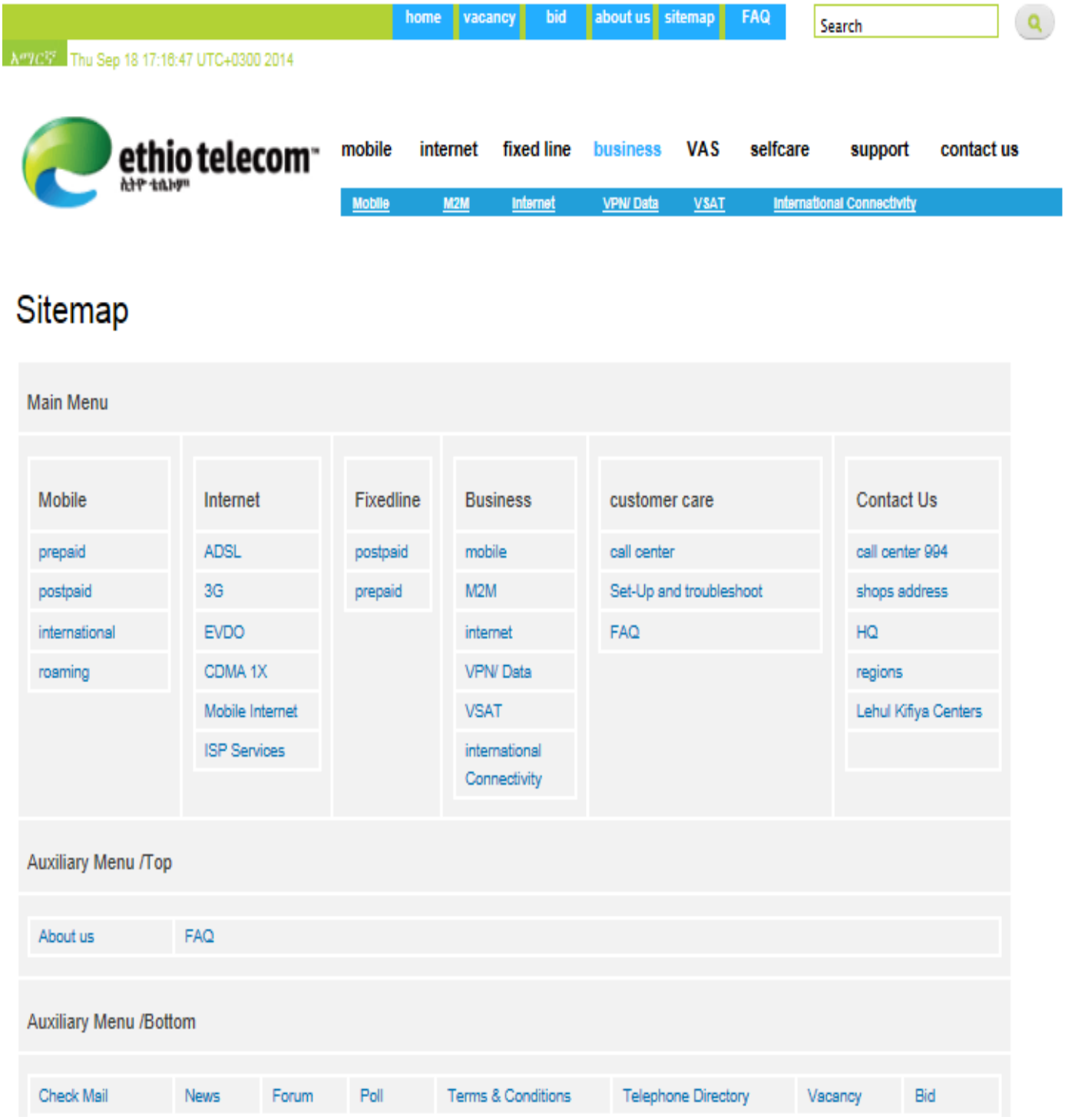


Figure 1-1 Sample English site map structure of Ethio telecom website

1.3. Statement of the Problem

A web site is the most direct link a company has to its current and potential customers. Companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. These rich results yielded by web usage analysis, offer great opportunities for future website and content improvements [8].

Unfortunately, to most companies, web is nothing more than a place where transactions take place. They did not realize that as millions of visitors interact daily with websites around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company for understanding customers' behavior, which can be used for improving customer services and relationship, launching target marketing campaigns and measuring the success of marketing efforts [8].

Most of the existing websites do not satisfy the interest of most users. There are several factors that contribute for this problem. One of the factors that contribute to this is that websites are poorly designed because user requirements are not often incorporated into the web design process. Websites focus more on quantity rather than content quality that could suit the user requirements. Thus, the effectiveness of a certain website should be evaluated regularly so as to assure that it fits the need of users [9].

Web Usage Mining can help organizations to address specific groups of customers' need. Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly. Also, the company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely [8].

Ethio telecom is a sole telecom operator and internet service provider in Ethiopia which can be reached at official website, www.ethiotelecom.et. This web site is one of the main communication means being used to reach customers. In addition to providing information of company's products or services and telephone directory, the website also serves as e-commerce gateway by providing some services like CRBT (call ring back tone) and ADSL (asymmetric digital subscriber line).

In order to realize the company's mission and vision, it is required to address various needs of its customers. One way of understanding customers' needs is through analyzing the behaviors of the website access and streamline the website towards their need. From the feedback of the web master, web administrators and researcher's observations, problems such as broken links, duplicate page titles, not fully optimized search engine, nonspecific page titles, slow response time, and non-intuitiveness of the page link hierarchy are observed. So it is worth study to carry out experimental work on web log data collected from Ethio telecom web server to find out useful browsing patterns.

To the best of the researcher's knowledge this work is the first attempt on analyzing the user access of Ethio telecom official website. More ever the new pattern analysis methodology adopted has contributed for the productivity of the findings and confirm the uniqueness of the study. In this study, therefore an attempt has been made to perform the website access statistics, pattern discovery and analysis so as to explore users browsing behavior of Ethio telecom web site.

To this end, the following research questions are explored and answered in this study:

- How to prepare appropriate dataset with relevant attributes for web usage mining?
- Which tools and algorithms to use for data preprocessing and pattern discovery?
- How the visitors' navigational flow with in Ethio telecom website looks like?
- To what extent does the extracted patterns are interesting to describe user's navigation and access behavior?

1.4. Objective of the Study

1.4.1. General Objective

The general objective of the study is to identify web usage pattern using web usage mining technique so as to optimize Ethio telecom official website.

1.4.2. Specific Objectives

In order to achieve the general objective of the study, the following specific objectives are formulated.

- To review relevant literatures to understand the problem area and web usage mining techniques and algorithms
- To preprocess web log file so as to prepare dataset for web usage mining
- To select appropriate web usage mining tools, techniques and algorithms for data analysis.
- To discover web usage patterns and access frequency.
- To evaluate the discovered association rules for their validity and novelty

1.5. Significance of the Study

Results extracted from this study will play an important role in improving customer experience of the web. It will give an insight into the accessibility and effectiveness of Ethio telecom official website with respect to customers' access. This will give the Company a chance to improve the structure, content and services of the website in order to make it more usable and effective. This in turn, helps the company to address a broad customer base and increase its revenue.

Moreover, web usage information mining could help to engage new customers, maintain current customers and track customers who are leaving web site. Usage information can be extracted to increase web server efficiency by pre-fetching and caching strategies [8].

As improving customer service is one of the major priorities of Ethio telecom, 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 will also be improved.

1.6. Scope and Limitation of the Research

From the three knowledge discovery domains that pertain to web mining, the aim of this study is to apply web usage mining for discovering web usage pattern using association rule mining techniques and describe users' web access statistically in order to explore Ethio telecom website usage.

Though there are three sources of web log files, namely web access log, client log and proxy log files, in this study, only web access log file is used as dataset. The reason is that literatures and previous researches justify, web access log files is a typical source for performing Web Usage Mining [10].

Due to time constraint, the web server log used for the pattern discovery of the study is the year of 2014 as the server log data was exported at the end of the same year when the preprocessing task was started. However, the statistical analysis is made on all available recent data exported as of May 25, 2015.

Descriptive modeling, which identifies patterns or relationship in data, is employed for pattern discovery. Due to the fact FP growth algorithm is complete and faster than Apriori and the rules generated by Apriori are subset of FP growth, FP Growth algorithm has been experimented and Apriori is not used for the analysis.

1.7. Research Methodology

A conceptual and empirical literature review is conducted from secondary sources such as books, journals, articles, conference papers, research papers and integrated in order to thoroughly understand web usage mining concepts and techniques. Data preprocessing techniques and tools have been thoroughly reviewed in order to prepare a clean data for the data mining task.

1.7.1. Research Method

There have been various web usage mining models reviewed such as Hengshan et.al model and found that all are discussed about the same basic concept with difference in detailed steps required for each milestones. Sharma model is selected because it is the most popular and frequently used by most scholars and previous researchers.

An experimental research methodology is used following a four-step web usage mining process model suggested by Sharma [21]. The model consists of Data collection, Data preprocessing, Pattern discovery and Pattern analysis, as depicted in the figure 1-2 below:

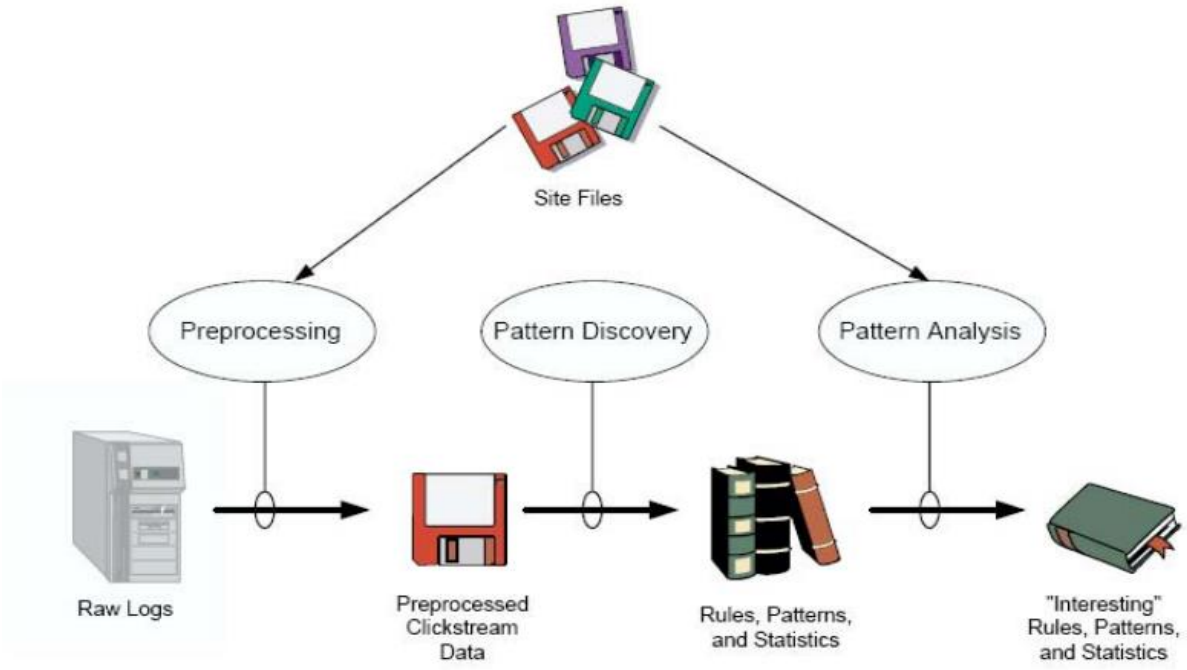


Figure 1-2 High Level Web Usage Mining Process Model [21]

1.7.2. Data Source

In this study, the entire available access log has been retrieved from the website server. As the website server contains the log of one year, log file with a duration from January 2014 to December 2014 is used. This is due to the log data was extracted from the server in the month of January 2015. The researcher selected the duration as it is the possible recent log data available in the website sever. In addition to the server logs, a six month reports from Google analytics repository has been extracted in census. Moreover, the researcher has collected additional qualitative data from secondary documents such as web site project report and proposal and by interviewing the

web master and web server Administrators about the problems of the web site and their future plans regarding the web site.

1.7.3. Data Preprocessing

Once the log data is transferred from the web server, further preprocessing task is carried out for Data Cleaning, Session Identification, Feature Selection, transactions identification and Data Transformation. WUM Prep [11, 13], Weka, EM editor, Oracle 11g release 2 and MS Excel 2010 have been used for data preprocessing. Moreover, Java programming and Perl software are also used for the preprocessing tasks. Google Analytics and MS Excel 2010 [13, 14] have been used for statistical analysis and Weka has been used for pattern and association rule discovery and analysis. The reasons for selecting these preprocessing, statistical analysis and pattern discovery tools are given below.

For the preprocessing task WUM prep is selected because unlike other tools such as Weka preprocessing tool, this tool is specifically designed to preprocess web log files. As a result it is flexible and comprehensive tool for preprocessing web log data.

For further preprocessing of the web log file, Java programming is selected because it is open source and the researcher is familiar with it. Java was used to develop programs that are used for further cleaning of the log data, transaction identification and transformation of the log file to the required data format for weka FP growth algorithm.

Perl programming language is used because it is required to run the scripts in WUMprep. MS excel is used to visualize, further clean and transform data.

1.7.4. Pattern Discovery

Once the web usage data is well prepared, the next task is pattern discovery. Pattern discovery adopts the various data mining techniques for discovering access patterns from the preprocessed usage data [12]. To discover usage patterns several experiments are conducted using data mining and statistical techniques. In this study, weka software is selected and used for pattern and association rule discovery because weka implements the popular algorithms for association mining such as APriori and FP-Growth and it is available for free. FP-Growth algorithm is applied for

pattern discovery. This algorithm is selected because of its efficiency and completeness over other algorithms such as Apriori.[17][24]

Statistical analysis is also conducted to give descriptive analysis of the web usage logs. In conducting the same Google analytics and MS Excel are used for statistical analysis and charting, calculation and summary purpose respectively. Google Analytics is selected because it is free and being utilized by the web master of the Ethio telecom website.

1.7.5. Pattern analysis

Once the access patterns have been identified, there needs to analyze the pattern and determine how that information can be used. Pattern analysis filters out interesting patterns from the set of rules generated during the pattern discovery phase [12]. The statistical reports and patterns discovered are analyzed using different techniques such as Literature Review, Website Structure and content analysis, business knowledge and discussion with domain experts of the website. After analyzing the findings, recommendations are forwarded for future researches as well as to improve the website performance.

1.8.Organization of the thesis

This thesis report is organized in to five chapters. The first chapter is introduction part containing background of the study, background of the company, problem statement, and objectives of the study, scope and limitation of the research, significance of the research and research methodology. The second chapter is literature review part in which different conceptual and empirical concepts regarding web usage mining are discussed. The third chapter is data preparation with sub topics of data collection, data cleaning, session identification, feature selection, transaction identification and data transformation are discussed. In the fourth chapter which is the Experimentation and analysis section different statistical analysis and experimentation of pattern discovery are presented. Finally, the fifth chapter summarizes the research findings and provides conclusion and recommendations based on the findings.

CHAPTER TWO

LITERATURE REVIEW

Necessity is the mother of invention. Since ancient times, our ancestors have been searching for useful information from data by hand. However, with the rapidly increasing volume of data in modern times where data are being collected and acculturated at a dramatic pace across a wide variety of fields, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery from large data [6].

2.1. Overview of Knowledge Discovery

When the scale of data manipulation, exploration, and inference grows beyond human capacities, people look to computer technology to automate the book keeping. The problem of knowledge extraction from large data repository involves many steps, ranging from data manipulation and retrieval to fundamental mathematical and statistical inference, search, and reasoning [6].

Researchers and practitioners interested in these problems have been meeting since the first KDD Workshop in 1989. Although the problem of extracting knowledge from data (or observations) is not new, automation in the context of large data repository opens up many new unsolved problems [6].

Knowledge discovery is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6]. It is the overall process of discovering useful knowledge from data including how the data is stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualized, and how the overall human-machine interaction can be modeled and supported. As depicted in the below figure 2.1 data mining is a particular step in this knowledge discovery process. [6]

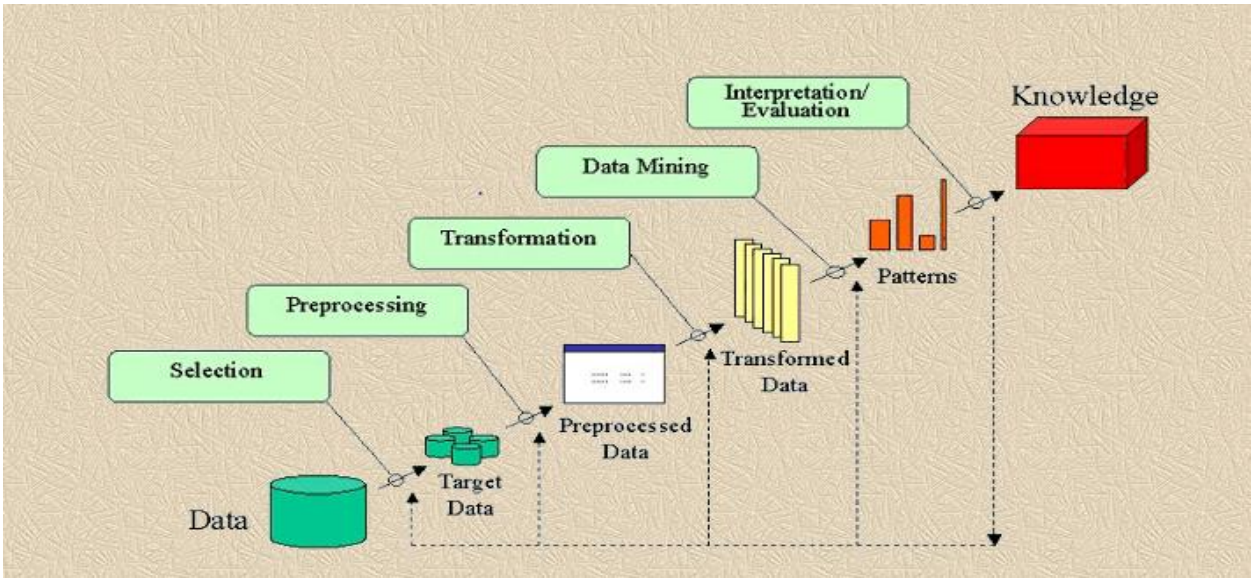


Figure 2-1 Knowledge Discovery process [6]

Data mining commonly defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the Web, etc. The patterns must be valid, potentially useful, and understandable. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization [3].

According to the definition given by Hand, "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [3].

One of the different applications of data mining is web mining.

2.2. Web Mining and its taxonomy

As many believe, it was Oren Etzioni who first proposed the term Web mining in 1996 [11]. As noted by Etzioni Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services.

As depicted in figure 2.2 below, Web Mining has three major sub-disciplines; namely, **Web Content Mining**, **Web Structure Mining** and **Web Usage Mining [2]**.

Web structure mining: Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines or

link-based categorization of web pages, ranking of web pages through a combination of content and structure.
We can also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

Web content mining: Web content mining extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics.
The type of data is mainly of textual in nature and the tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc, for many purposes.
Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not traditional data mining tasks.

Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files, client logs and proxy server logs. The main source of data consists of the textual logs that are collected when users access web servers which are represented in standard formats. "Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications." [20]. Web usage mining differs from web structure mining and web content mining, in that web usage mining reflects the behavior of humans as they interact with the Internet. Because of this, web usage mining is of intense interest for e-marketing and e-commerce professionals. Analysis of user behavior can provide insights leading to customization and personalization of a user's web experience.

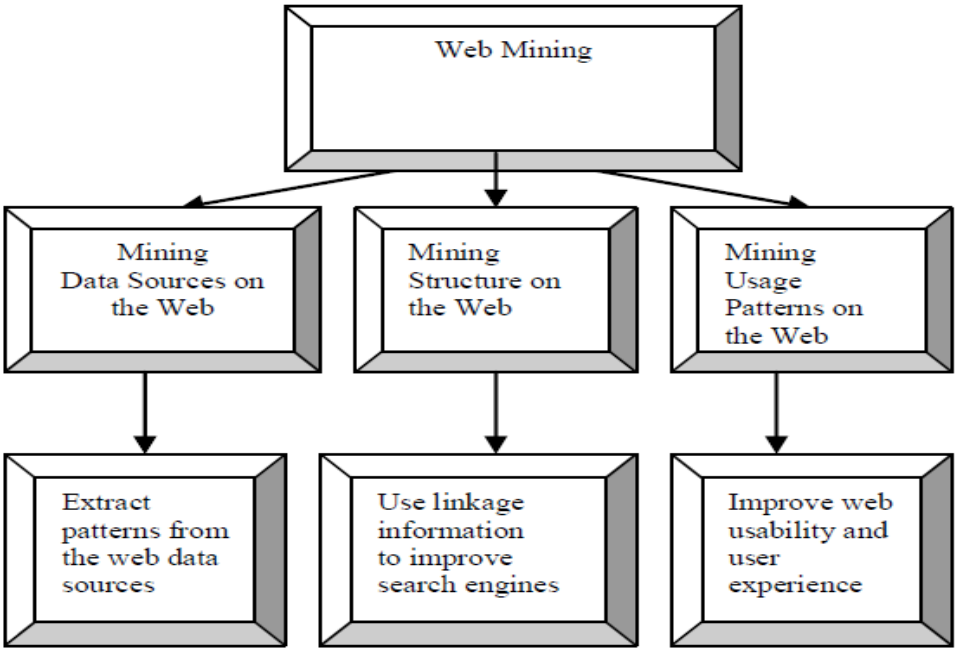


Figure 2-2 Taxonomy of Web Mining

To explore information mining on the Web, it is necessary to know data mining, which has been applied in many Web mining tasks. However, Web mining is not entirely an application of data mining. Due to the richness and diversity of information and other Web specific characteristics, Web mining has developed many of its own algorithms such as Cite Seer and Link analysis [3].

2.3. Challenges in Web mining

The Web has many unique characteristics, which make mining useful information and knowledge fascinating and challenging task. Some of these characteristics are [1]:

- The amount of data/information on the Web is huge and still growing. The coverage of the information is also very wide and diverse. One can find information on almost anything on the Web.
- Data of all types exist on the Web, e.g., structured tables, semi structured Web pages, unstructured texts, and multimedia files (images, audios, and videos).
- Information on the Web is heterogeneous. Due to the diverse authorship of Web pages, multiple pages may present the same or similar information using completely different words and/or formats. This makes integration of information from multiple pages a challenging problem.

- A significant amount of information on the Web is linked. Hyperlinks exist among Web pages within a site and across different sites. Within a site, hyperlinks serve as information organization mechanisms. Across different sites, hyperlinks represent implicit conveyance of authority to the target pages. That is, those pages that are linked (or pointed) to by many other pages are usually high quality pages or authoritative pages simply because many people trust them.
- The information on the Web is noisy. The noise comes from two main sources. First, a typical Web page contains many pieces of information, e.g., the main content of the page, navigation links, advertisements, copyright notices, privacy policies, etc. For a particular application, only part of the information is useful. The rest is considered noise. To perform fine-grain Web information analysis and data mining, the noise should be removed. Second, due to the fact that the Web does not have quality control of information, i.e., one can write almost anything that one likes, a large amount of information on the Web is of low quality, erroneous, or even misleading.
- The Web is also about services. Most commercial Web sites allow people to perform useful operations at their sites, e.g., to purchase products, to pay bills, and to fill in forms.

- The Web is dynamic. Information on the Web changes constantly. Keeping up with the change and monitoring the change are important issues for many applications.

- The Web is a virtual society. The Web is not only about data, information and services, but also about interactions among people, organizations and automated systems. One can communicate with people anywhere in the world easily and instantly, and also express one's views on anything in Internet forums, blogs and review sites.

All these characteristics present both challenges and opportunities for mining and discovery of information and knowledge from the Web.

2.4. Web usage mining process

The Web mining process is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a database or warehouse. For Web mining, data collection can be a substantial task from log files [2].
Once the data is collected, it goes through the same three-step standard data mining process: data pre-processing, pattern discovery and pattern analysis [21].

2.4.1. Data source and web log files

An important task in any data mining application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web usage mining due to the characteristics of click stream data mentioned above and its relationship to other related data collected from multiple sources and across multiple channels. Collectively, this process is referred as data preparation [2].
The data preparation is a process of transforming the raw server logs stored in the various data sources into a suitable data file for Web usage mining. It is often the most time consuming and computationally intensive step in the Web usage mining process, and often requires the use of special algorithms and heuristics not commonly employed in other domains. This process is critical to the successful extraction of useful patterns. The primary source of data in web usage mining is web log file [21].

2.4.1.1. Web Log data

Web log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text [10].

2.4.1.1.1. Location of Web log File

As depicted in figure 2.3, Web log file is located in three different locations, **Web server logs**, **Web proxy server** and **Client browser** [10]. Each type of data collection differs not only in terms of the location of data source, but in kinds of data available, the segment of population from which the data was collected and its method of implementation

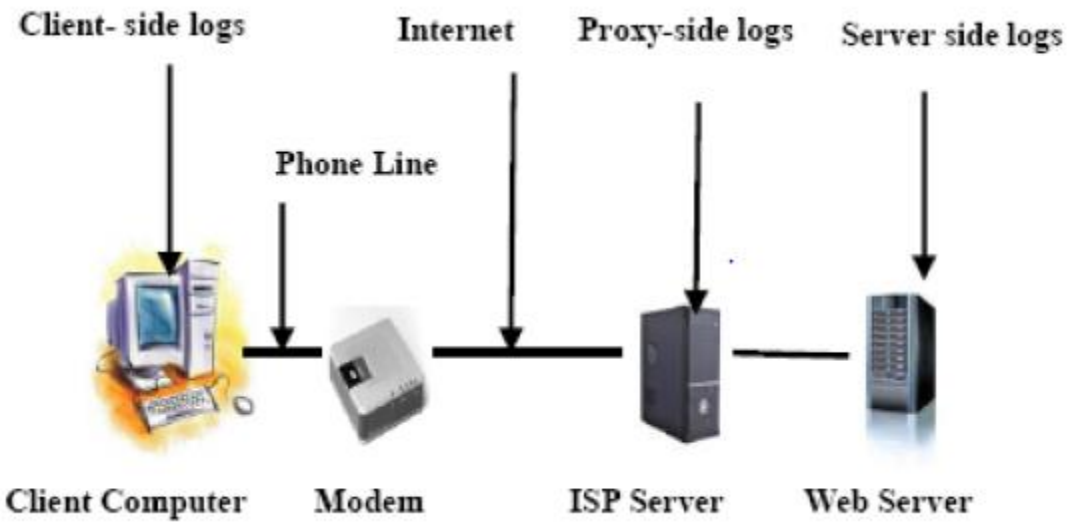


Figure 2-3 various data sources for web usage mining

i. Web server logs or Web log files

A web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. It provides the most accurate and complete usage of data to web server. Web servers are surely the richest and the most common source of data. The log file do not record cached pages visited. Data of log files are sensitive, personal information so web server keeps them closed [10].

These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format e.g.: Common Log Format, Extended Log Format, LogML which will be discussed on the subsections [2].

The method portion for the request is usually either Get, POST or HEAD. GET requests an object from the web server, POST sends information to the WEB server and HEAD requests just the HTTP header for an object.

ii. Web proxy server

Many internet service providers (ISPs) give to their customer Proxy Server services to improve navigation speed through caching. The proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Client send request to web server via proxy server [10].

In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers [2].

iii. Client browser

Log file can reside in client's browser window itself. Client level collection can be implemented by different methods such as using a remote agent (such as java scripts or java applets), modifying the source code of an existing browser to enhance its data collection capabilities, or using HTTP cookies which are pieces of information generated by a web server and stored in user's computer that are ready for future access and so on [10].

These collection techniques avoid the problems of users' sessions identification and the problems caused by caching (like the use of the back button).In addition, they provide detailed information about actual user behaviors.

2.4.1.1.2. Type of Web log file

There are four types of server logs [4]: **Access log file, Error log file, Agent log file** and **Referrer log file.**

i. Access log file

The data from Access Logs provides an extensive view of a Web servers and users. Such analysis enables server administrators and decision makers to characterize the users and usage patterns. Access Logs are also called Transfer Logs. It stores information about which files are requested from web server. The Access log format is shown in table 2.1 below.

Client IP address	User ID	Access time	HTTP request method	Protocol used for the transmission	Status of the request	Amount of data transfer red

Table 2-1 Access log file

Sample Access log entry

123.45.6.78.9 - [11/May/2011:04:05:45 -0500] "GET/HTTP/1.0" 200 3250

This line consists of the following fields. The first field with123.45.6.78 is an IP address of the client. Then a user ID field in our case '-'(hyphen) which represents anonymous user ID, the third field with (11/May/2011:04:05:45-0500) shows an access time of the webpage (including day, month, year, time and time zone). The fourth field is GET/HTTP/1.0 which represents the HTTP request method and protocol used for the transmission. Further information on status code returned by the server is provided. Here, 200 indicate success. Finally information about number of bytes transmitted, 3250

ii. Error log file

The Error Log provides the time, domain name of the user, and page on which a user received the error to a server administrator. These error messages inform server administrators of erroneous links on their servers. It stores information about errors and failed requests of the web server. The Error log format is given below in Table 2.2.

Date	Time	Error	IP Address	Error Message

Table 2-2 Error Log Format

Sample Error log entry

[Wed May 11 17:35:45 2011] [error] [client 132.1.0.1]

client denied by server:/export/home/live/ap/htdocs/testdoc

iii. Referrer log file

It stores information of the URLs of web pages on other sites that link to web pages. That is, if a user gets to one of the server's pages by clicking on a link from another site, the URL of that site will appear in this log. The Referrer log format is given below in Table 2.3.

Date	Time	Time Zone	Referrer URL

Table 2-3 Referrer Log Format

Sample Referrer log entry

The following is an example of a record in a Referrer log:

[Wed May 11 17:35:45 2011+0500] *http://www.ibm.com/index.html*

iv. Agent log file

The Agent log provides information on a user‘s browser including browser version and operating system. It records information about the web clients that sends requests to web server. This is the major information, as the type of browser and the platform determines what a user is able to access on a web site [4]. The Agent log format is given in Table 2.4 below.

Date	Time	Time Zone	Version Number	Plat form

Table 2-4 Agent Log Format

Sample Agent log entry

The following is an example of a record from an Agent log:

[10/Nov/2011:19:15:06+0500] “Microsoft Internet Explorer – 5.0”

2.4.1.1.3. Attributes of log file

The log file stores different information using different attributes of log file which includes some important information [13].

Client IP

Client IP is the IP address of client machine from where users browse the website.

Date and time

Date is recorded when user made access the date is stored in the format YYYY-MM-DD. Time information is stored in the format as HH-MM-SS.

Sever client status

Client status code return by server like 200(success), 404(error).

User agent

User agent records the information about the browser type, version and operating system that is used by the client at the time of accessing the website.

Referrer

Referrer is the previous page from where client jumps to the new web page or website.

Server client bytes

Number of bytes sent by the server to client.

Client server bytes

Number of bytes received by client from server.

2.4.1.1.4. Web log file Format

There are many log formats available to create log files to capture the behavior and activities of users on website. Log text files are stored using.txt extension. These files are stored in ASCII format.log files are used to monitor the behavior of user and takes feedback from client side. The major log file formats are: **W3C Extended log file format**, **NCSA common log file format**, and **IIS log file format** [13][4].

In NCSA and IIS log file format the data logged for each request is fixed however W3C format allows user to choose properties and change the format for each request [10].

i). Common Log Format (CLF)

This is the most common and standardized text format of a web server log file [13][4]. This can be produced by several web servers and read by variety of log analysis programs.

The log file entries produced in CLF appear as follows

host/iprfcnamelogname [DD/MMM/YYYY:HH:MM:SS- 0000]

"METHOD/PATH HTTP/1.0" code bytes

Sample line of a Common Log Format

127.0.0.1 - john [12/nov/2011:12:53:46-0700] "GET/apache_pb.gif HTTP/1.0"200 2326

Each part of this log entry is described below

127.0.0.1(%h) this is the IP address of the client (remote host) which made the request to the server.

-(%1) The "hyphen" in the output indicates that the requested piece of information is not available.

john (%u) This is the user ID of the person requesting the document as determined by HTTP authentication.

[12/nov/2011:12:53:46-0700] The time which the server finishes processing the request.

The format is as follows:

[day/month/year: hour: minute: second zone]

day = 2*digit

month = 3*letter

year = 4*digit

hour = 2*digit

minute = 2*digit

second = 2*digit

zone = (`+' | `-') 4*digit

“GET/apache_pb.gif HTTP/1.0” (\ “%r\”) the request line from the client is given in double quotes. The request line contains a great deal of useful information. First, the method used by the client is GET. Second, the client requested the resource /apache_pb.gif, and third, the client used the protocol HTTP/1.0

200 (% >s) this is the status code that the server sends back to the client.

2326(%b) the last entry indicates the size of the object returned to the client, not including the response headers.

W3C (World Wide Web Consortium) Extended log file format

This is the default log file format used by IIS. It uses ASCII text format and the time recorded as UTC (Greenwich Mean Time). This is the customizable format, can add or remove fields depending on what information is needed to record. Sample lines in a W3C Extended log file format with the following fields: Time, Client IP Address, Method, URI Stem, Protocol Status, and Protocol Version is shown below:

#Software: Microsoft Internet Information Service 5.1

#Version: 1.0

#Date: 2011-11-11 14:35:15

#Fields: time c-ipcs-method cs-uri-stem sc-status cs-version 6:32:15 172.16.255.255 GET/default.htm 200 HTTP/1.0

ii). Microsoft IIS (Internet Information Services) log file format [13][4]

Microsoft IIS log file format is a non-customizable ASCII text based format used to record more information than the NCSA Common format but less than the W3C format. It uses comma to separate fields and uses the local time. It includes the user‘s IP address, user name, request date and time, Service status code and number of bytes received, the elapsed time, the number of bytes sent, the action (for example, a download carried out by a GET command) and the target file. A sample lines in an IIS log file format is shown below:

192.168.114.201, —, 11/25/2011, 9:45:25,

W3SVC2, SALES1, 192.168.114.201, 4504,163, 3223,200, 0, GET, / SalesDeptLogo.gif, —

,172.16.255.255, anonymous, 11/25/2011, 23:58:11, MSFTPSVC, SALES1,192.168.114.201, 60, 275, 0, 0, 0, PASS, /introduction.htm, —,

iii). NCSA Common log file format

NCSA (National Centre for Supercomputing Applications Common format) is a fixed (non-customizable) ASCII text based format [13] [4]. It does not support FTP sites, only available for SMTP and NNTP services. Since the entries are small with this format, the storage space required for logging is less compared to other formats. It logs the basic information about user requests such as remote host name, user name, date, time, request type, HTTP status code, and the number of bytes sent by the server. It records the time by using the local time and fields are separated by spaces. A sample lines in a NCSA log file format with the following fields is shown below:

172.21.13.45- REDMOND\sam [11/11/2011:25:28:06 - 0800] "GET scripts/iisadmin/ism.dll?http/serv HTTP/1.0" 200 3401

2.4.1.1.5. The Status Field (HTTP Status)

Browser requests have different status based on success or failure. The following are summary of status codes held in the HTTP status filed of the web log file [22]:

2xx – Success series

- 200: success
- 201: created
- 202: accepted
- 204: no content

3xx – Redirection series

- 301: moved permanently
- 302: moved temporarily
- 303: not modified
- 304: use cached document

4xx – Client error series

- 400: bad request
- 401: unauthorized

- 403: forbidden
- 404: not found

5xx – server error series

- 500: internal server error
- 501: not implemented
- 502: bad gateway
- 503: service unavailable

2.4.2. Data Preprocessing

The raw web log data is generally diverse, incomplete, inconsistent, noisy and difficult to be used directly for pattern mining. Quality data gives quality output. The attributes of quality data includes accuracy, completeness, consistency, accessibility, and timeliness. In order to obtain quality data we have to preprocess it. Preprocessing of log fie is complex and laborious job and it takes majority of the total time of web usage mining process. The Data preprocessing phase includes Data cleaning, User Identification, Session Identification, Path completion and Transaction identification. Various steps involved in data preprocessing phase are shown in Figure 2.4 below [4].

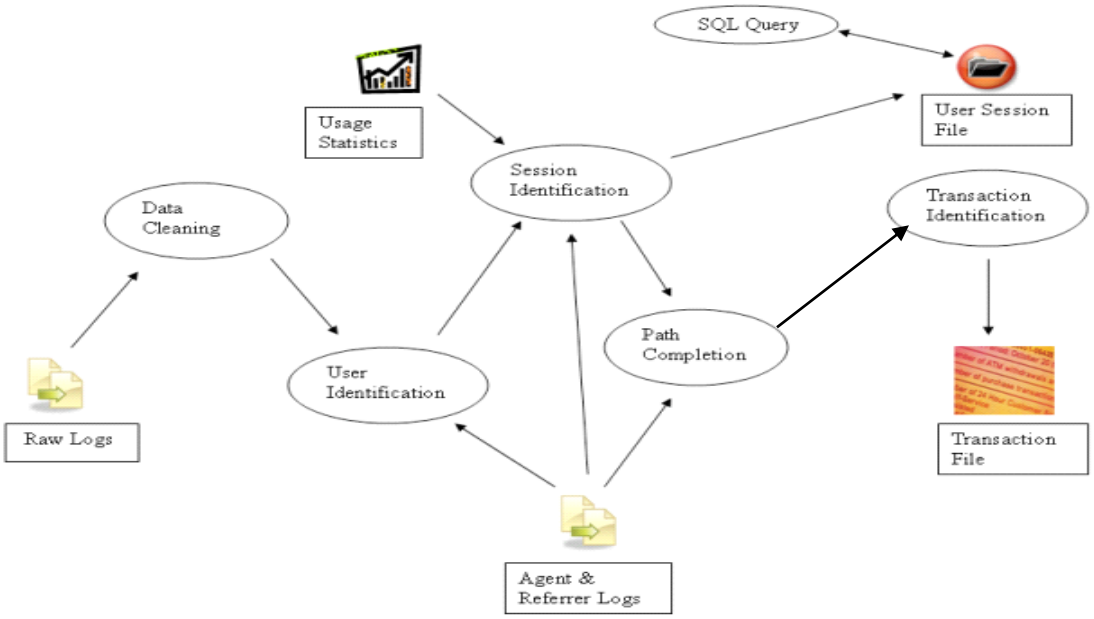


Figure 2-4 Steps in Data preprocessing

2.4.2.1. Data Cleaning

The purpose of data cleaning is to remove irrelevant items stored in the log files that may not be useful for analysis purposes such as robot crawls, graphics, audio, duplicate requests, etc. When a user accesses a HTML document, the embedded images, if any, are also automatically downloaded and stored in the server log. Log entries with file name extensions **.ico**, **.gif**, **.GIF**, **.jpeg**, **.JPEG**, **.jpg**, **.JPG**, **.js**, **.css, .pdf, .txt, .png,** and **.flv** should be removed. Since the main objective of data preprocessing is to obtain only the usage data, file requests that the user did not explicitly request should be eliminated. Data cleaning also involves the removal of references resulting from spider navigations which can be done by maintaining a list of spiders or through heuristic identification of spiders and Web robots. The cleaned log represents only the user's accesses to the Website [21]. The cleaning process may also involve the removal of some of the data fields. The most common failure codes in Web Log are 401 (failed authentication), 403 (Forbidden request to a restrict sub-directory) and 404 (file not found) messages. Such entries are useless for analysis process and therefore they are cleaned form the log files [23].

2.4.2.2. User Identification

Once HTTP log files have been cleaned, next step in the data preprocessing is the identification of users. User identification is to discover who access web site and which pages are accessed. IP address, User agents and referring URL fields of log files are used to identify user [4].
This task is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The data recorded by a Web server are not sufficient for distinguishing among different users and for distinguishing among multiple visits of the same person. The standard Common Log file Format (CLF), as well as Extended CLF (ECLF), only records host or proxy IP from which requests originate, so different visitors sharing the same host or one proxy server cannot be distinguished [24].

One of the main obstacles to user identification is the use of proxy servers. Proxy servers provide an intermediary solution but the difficulty of user identification still persists. All requests coming from a proxy server have the same identifier even though the requests are put forth by multiple users [21].

The other thing that poses problems for obtaining reliable usage data is local caching. To reduce network traffic and improve performance, the pages that are requested are cached by most Web browsers. Hence, when the user navigates backwards by using the "back" button, the repeat page access is not recorded in Web server log [21].

Methods such as cookies, user registration or the remote agent, a cookie-like mechanism make the identification of a visitor possible. The shortcomings of such methods are that they rely on user's cooperation, but user often denies providing such cooperation due to privacy concerns [24].
There are some heuristics for user identification. For instance, two accesses having the same IP but different browser (versions) or operating system, which are both recorded in agent field, are originated from two different users. But this method will render confusion when a visitor uses different platforms [24].

User agent plays an important role in user identification. It refers to the browser used by the client. A change in the browser or the operating system under the same IP address represents a different user heuristically [25].
In general, the following procedure could be used to identify users [1]:
Step 1. Sort the web log file by ID address and then by time stamp.
Step 2. For each distinct ID address, identify each agent as belonging to a different user.
Step 3. For each user identified in step 2, apply path information garnered from the referrer field and the site topology to determine whether this behavior is more likely the result of two or more users.
Step 4. To identify each user, combine the user identification information from steps 1 to 3 with available cookie and registration information.

2.4.2.3. Session Identification

User session is the set of consecutive pages visited by a single user at a defined duration. As long as user is connected to the website, it is called the session of that particular user. Most of the time, 30 minutes time-out was taken as a default session time-out. A session is a set of page references from one source site during one logical period [8]. For logs that span long periods of time, it is very likely that users will visit the Website more than once. The goal of session identification is to

divide the page accesses of each user into individual sessions. Hence the log file, after user identification, may be further divided into sessions for every user. Hence each user's page visits will be split into one or more sessions. Similar to user identification, Cookie and Session mechanisms both can be used in the session identification, yet same problems also exist [24].

To define user session, two criteria are usually considered [24], Upper limit of the session duration as a whole and upper limit on the time spent visiting a page.

The Following rules can also be used to identify users' sessions [11].

i. If there is a new user there is new session.
ii. In one user session, if the referrer page is null, there is a new session.
iii. If the time between page requests exceeds a certain limit (30 minutes). It is assumed that user is starting a new session.

The reference page is estimated by access time of this page and the next one i.e. the reference length of an accessed page equals the difference between the access time of the next and the present page. If this time is few seconds than that page can be considered as an auxiliary page and otherwise that page can be considered as a content page [11].

Mechanisms such as local caches and proxy servers can severely distort the overall picture of user traversals through a Website. If user clicks **backward** to visit a page that has had a copy stored in Cache, browser will get the page directly from Cache. Such a page view will never be trailed in access log, causing the problem of incomplete path, which need mending [24].

To identify each user and session uniquely we can take measures like IP address, operating system, browser, timeout period, etc. Once the data cleaning is performed, all useful data records are saved in database and irrelevant entries are considered to be removed [14]. So, now the remaining process which is user and session identification is performed.

Session Identification Procedure

The session identification procedure may be summarized as follows [1]:

1. For each distinct user identified in the preceding section, assign a unique session ID.
2. Define the timeout threshold t.
3. For each user, perform the following:
    a. Find the time difference between every two consecutive web log entries.

b. If this difference exceeds the threshold t, assign a new session ID to the later entry.

4. Sort the entries by session ID.

Algorithm 2.2: Algorithm for User & Session Identification [14]

```
Input: processed weblog file
Output: identification of user & session.

1. for each record in dataset do
    2. ifcurrentIP is not in ListOfIP then
        addcurrentIP in ListOfIP
        mark whole record as a new user and session
        assign a new sessionID and userID

    3. else if currentOS is not in ListOfOS then
        addcurrentOS in ListOfOS
        mark whole record as a new user and session
        assign a new sessionID and userID

    4. else if currentBrowser is not in ListOfBrowser then
        addcurrentBrowser in ListOfBrowser
    mark whole record as a new user and session
    assign a new sessionID and userID


    5. else if current record timestamp is more than 1800 seconds (30minutes * 60 seconds
)
        mark whole record as a new user and session
        assign a new sessionID and userID
    6. else
        mark current record with existing sessionID and userID
        end if
end of loop
```

Algorithm 2.2 above marks each record in database with respective user and session identified groups which later can be used for further proceedings of web usage mining process. The resulted group of records can be inserted into database and later results of which can be very helpful like total number of users, total number sessions, difference between total number of records before pre-processing and post-preprocessing, etc. [14].

2.4.2.4. Transaction Identification

During each visit session, a user may pursue one, or more than one, information need(s). The goal of transaction identification is to identify, from sessions, pages clicked to fulfill individual information needs during the same visit. If a user pursues only a single information need, a whole session can be viewed as a single transaction. If a user has multiple information needs, and each page in his/her visit session fulfills one of them, a session can be viewed as a set of transactions, each consisting of a single page. In most cases, however, a transaction consists of more than one, but not all pages of a session, and the task of transaction identification is to divide and identify such meaningful transactions. A transaction can be identified using one of the approaches such as maximal forward reference, reference length and time window leading to the creation of a user transaction file [21].

2.4.2.5. Path Completion

Not all page views seen by the user are recorded in the web server log. For example, many people use the "Back" button on their browsers to return to a page viewed previously. When this happens, the browser returns to a page that was previously cached locally rather than accessing the web server again. This leads to "holes," missing pages, in the web server's record of the user's path through the Web site. Knowledge of site topology must be applied to complete these paths, in a process known as path completion [1].

This is important and difficult phase because it involves the use of referring URLs and site topology. Path completion is used to obtain the complete user access path. The incomplete access path of every user session is recognized based on user session identification. There are chances of missing pages after constructing transactions due to proxy servers and caching problems. However by examining the site topology and the referrer field it is possible to rebuild the path followed by the user. At the end of this stage the user session file is ready [1].

Generally Data preprocessing results in the creation of a user session file consisting of a set of users, each user associated with one or more sessions. A user session is a collection of page references made by a user during a single visit to a site. In certain cases, user sessions may be

further divided into semantically meaningful groups of page references referred to as transaction [21].

2.4.3. Pattern discovery

Pattern discovery deals with extracting information from preprocessed data. There are several techniques that are deduced from different fields such as data mining, statistics, machine learning and pattern recognition and applied to web usage data to discover user access patterns of the web. Statistical Analysis tools can be used to give a description of the traffic on a web site for example most visited pages, average daily hits etc. Association Rules consider every URL requested by a user in a visit as item and find out relationships between them with a minimum support level. Sequential Patterns are used to discover time ordered sequence of URL‘s followed by past users in order to predict future ones. Clustering forms meaningful clusters of URL‘s by discovering similar attributes between them according to user behavior [4].

Data mining involves different techniques to uncover hidden patterns form a large dataset. The primary techniques are classification, clustering, and association rule discovery [15].

Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers and Support Vector Machines [2].

Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered [2].

- Usage clusters and
- Page clusters.

Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers. In both applications

permanent and dynamic HTML pages can be created that suggest related hyperlinks to the user according to user's query or past history of information needs. [2].

Unlike classification, which analyzes class-labeled data sets, clustering analyzes data objects without consulting class labels. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together. [16]

Generally there are two clustering approaches: [3]

a. Partitioning clustering approach: Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors

   Typical methods for this approach are: distance-based: K-means clustering and model-based: expectation maximization (EM) clustering.

b. Hierarchical clustering approach: Create a hierarchical decomposition of the set of data (or objects) using some criterion

   Typical methods for this approach are: agglomerative Vs divisive and single link Vs complete link

Association Rule mining is the process of finding interesting correlations, frequent patterns or associations among sets of items in the transaction databases, relational databases or other information repositories [17]. The detail is described in section 2.5

As association rule discovery is very vital in finding interesting correlations among different pages of the website as compared to classification and clustering which mainly focuses on classifying and grouping, association task is the focus of this study.

2.4.4. Pattern analysis

This is the final stage of Web Usage Analysis. The goal of this process is to extract the interesting patterns from the output of the pattern discovery process by eliminating the irrelevant patterns. Pattern Analysis involves the validation and interpretation of the mined patterns. Validation can be used to remove the irrelevant patterns and to extract the interesting patterns from the output of the pattern discovery process. The output of mining algorithms is in mathematic form and not suitable for direct human interpretations. So, Visualization techniques are used to interpret the results. The most general ways of analyzing user access patterns are either by using a knowledge query mechanism on a database such as SQL or data cubes to perform OLAP operations. Visualization techniques, such as graphing patterns are used for an easier interpretation of the results [4].

A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful and novel [16]. A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents new knowledge.

2.5. Association Rule Mining

Mining of association rules is a fundamental data mining task. It is perhaps the most important model invented and extensively studied by the database and data mining community. Its objective is to find all co-occurrence relationships, called associations, among data items. Since it was first introduced in 1993 by Agrawal et al [2], it has attracted a great deal of attention.

Association Rule mining deals with discovering patterns of the form (A->B) from frequent item sets where A is an antecedent and B is a consequent. This means that the occurrence of A implies the occurrence of B with a given support and confidence level.

The following basic steps are involved in association rule discovery [16].

- Generate item frequent k-item sets that satisfy the minimum support threshold value where k is one or multiple items.
- Generate association rules on those frequent k-item sets with the minimum confidence constraint.

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. These pages may not be directly connected to one another via hyperlinks. For example, association rule discovery may reveal a correlation between users who visited a page containing electronic products to those who access a page about sporting equipment. Aside from being applicable for business and marketing applications, the presence or absence of such rules can help Web designers to restructure their Website. The association rules may also serve as a heuristic for pre-fetching documents in order to reduce user-perceived latency when loading a page from a remote site [2].

Let I = {I1, I2…Im} be an item set. Let D, the task-relevant data, be a set of database transactions where each transaction T is a non-empty item set such that T ⊆I. Each transaction is associated with an identifier, called a TID. Let A be a set of items. A transaction T is said to contain A if A⊂T. An association rule is an implication of the form A =>B, where A⊂I, B⊂ I, A≠∅, B ≠∅ and A ∩ B =∅. The rule A =>B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A∪B. This is taken to be the probability P (A∪B).

The rule A =>B has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A) [14].

Support: The support of a rule, *X* ->*Y*, is the percentage of transactions in *T* that contains *X* ->*Y*, and can be seen as an estimate of the probability, Pr(*X*->*Y*). The rule support thus determines how frequent the rule is applicable in the transaction set *T*. Let *n* be the number of transactions in *T*.[2] The support of the rule *X* ->*Y* is computed as follows:

Support (X=> Y) = P (X∪ Y)

Support is a useful measure because if it is too low, the rule may just occur due to chance. Furthermore, in a business environment, a rule covering too few cases (or transactions) may not be useful because it does not make business sense to act on such a rule (not profitable) [2].

Confidence: The confidence of a rule, A ->B, is the percentage of transactions in T that contain A also contain B. It can be seen as an estimate of the conditional probability, Pr (A | B) [2]. It is computed as follows:

$$\text{Confidence } (A \Rightarrow B) = P(B|A) = \frac{\text{Support}(A \cup B)}{Support\ (A)} = \frac{Support{-}count(A \cup B)}{Support{-}count(A)}$$

Confidence thus determines the predictability of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict B from A.A rule with low predictability is of limited use.

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong [16][2].

2.5.1.1. Association Rules Discovery Algorithms

Even though there are a number of techniques for association rules discovery, as noted by Han and Kamber[18], the most widely used algorithms for pattern discovery are Apriori and Frequent Pattern Growth (FP-Growth).

i. Apriori

Apriori employs an iterative approach known as a *level-wise* search, where *k*-item sets are used to explore (*k*+1)-item sets. 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 *L*1. Next, *L*1 is used to find *L*2, the set of frequent 2-itemsets, which is used to find *L*3, and so on, until no more frequent *k*-item sets can be found. The finding of each *Lk* requires one full scans of the database [18].

To improve the efficiency of the level-wise generation of frequent item sets, an important property called the Apriori property is used to reduce the search space. Apriori property states that, all nonempty subsets of a frequent item set must also be frequent [18]

The Apriori algorithm works in two steps: [2]

1. **Generate all frequent item sets:** A frequent item set is an item set that has transaction support above min support.

The Apriori algorithm relies on the ***apriori***or **downward closure** property to efficiently generate all frequent item sets.

Downward Closure Property: If an item set has minimum support, then every non-empty subset of this item set also has minimum support.i.e. Any (*k*-1)-item set that is not frequent cannot be a subset of a frequent *k*-item set. Hence, if any (*k*-1)-subset of a candidate *k*-item set 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 item sets [18].

2. **Generate all confident association rules from the frequent item sets**: A confident association rule is a rule with confidence above minconf.

Limitations with Apriori approach [2]

Apriori algorithm may suffer from the following limitations:

- The main problem with association rule mining is that it often produces a huge number of itemsets (and rules), tens of thousands, or more, which makes it hard for the user to analyze them to find those useful ones. This is called the **interestingness** problem.[2]
- While implementing Apriori property, it requires two nontrivial computationally expensive processes which makes the algorithm costly in both space-wise and time-wise [18] . An efficient implementation of the Apriori algorithm involves sophisticated data structures and programming techniques.
- The Apriori algorithm reduces the size of candidate frequent itemsets by using "Apriori property by which potential sources to associative rules may be lost.

ii. FP-Growth

To overcome the inefficiency of Apriori mentioned above, an algorithm named frequent pattern growth (FP-Growth) was introduced. FP-growth was first proposed by Han et al. [18]. 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 item set 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 [18].

As presented in Algorithm 2.1, 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 [18]. 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 [18].

Algorithm 2.1

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
IncrementN'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 asβ) of the nodes in the path P
generate pattern βUα with support count = minimum support count of nodes in β;
else
for each ai in the header of Tree
generate patternβ = aiU αwith support count = ai:support count;
constructβ's conditional pattern base and thenβ 's conditional FP tree Tree ;
if Tree β≠Ø then
call FP-growth(Treeβ ,β)

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 [18].

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

Algorithm 2.1: The FP-growth algorithm for discovering frequent item sets without candidate generation works as follows: [18].

2.5.1.2. Performance of Apriori vs. FP-Growth Algorithms

According to the experiment by Kaur and Aggarwal [17] FP-Growth Algorithm outperforms Apriori in terms of time complexity.

Moreover, different studies show that the performance of both Algorithms is almost the same for large support values but FP-Growth algorithm outperforms for small support values. Mishra and Choubey [24] conducted an experiment using datasets with different sizes and different minimum support values. Accordingly, they have concluded that FP-Growth algorithm is more efficient than Apriori by several magnitudes especially as the support value gets lower and lower.

2.6. Application of Web Usage Mining

In web usage mining or web log mining, user's behavior or interests are revealed by applying data mining techniques on web log file. The ability to know the patterns of user's habits and interests helps the operational strategies of enterprises. Various applications are built efficiently by knowing users navigation through web [22].As clearly described by El-Yazeed [22], applications of Web Usage Mining include the following:

- **Modification of website design**

In addition to modifications to the linkage structure, identifying common access behaviors can be used to improve the actual design of Web pages and to make other modifications to the site [27]. The ways the users access a website are restricted by the website's link structure. The organization of web pages within the website has great influence on the quality of the web service provided [28].

- **Schema modifications**

Web Usage Mining can help the website designers to redesign the internal structure of the website such as the database schema.

- **Improve website and web server performance**

By analyzing web traffic behavior, the frequent access patterns can be discovered and applied for developing policies of web caching, document pre-fetching, and data distribution [28]. By determining frequent access behavior for users, needed links can be identified to improve the overall performance of future accesses [27].

Search Engine Optimization (SEO) is one and the major way of improving website usability and performance. SEO describes a series of techniques which improve the visibility of a website in search engine result pages. The goal of such optimization is to rank as highly as possible for a certain search query [29].

The internet is incredibly large; currently there are approximately 50 billion indexed web pages and without search engines it would be impossible to find useful information in this clutter. It would be like searching for a needle in a haystack [29].

Search engines bring order to this chaos. By building an index they can show you the most relevant web pages for your search query. But this index changes frequently. New websites are added daily, web pages are redesigned, new files are uploaded, etc. Because the internet is a dynamic entity, search engines need a tool to help the index stay up to date [29]

That's where crawlers come into play. These automated robots scrape web pages for information; they index links, images, videos and other files [29].

In order to show the most relevant results from their index, search engines need to use a ranking system. There are several factors that are taken into consideration and understanding these factors is essential for SEO success [29]. Among several factors that determine website ranking, search keywords play a vital role.

- **Improve web personalization**

Web personalization is the process of customizing the content and structure of a website to the specific and individual needs of users [28]. 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 [27].

- **Recommender Systems**

Web usage patterns can be used to gather business intelligence to improve Customer attraction, Customer retention, sales, marketing and advertisement, cross sales [27].

- **Fraud detection and future prediction**

Web usage mining is useful for detecting intrusion, fraud, and attempted break-ins to the system. 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 [27].

2.7. Related Works

Hofgesang[30] has conducted a study with a goal of analyzing user behavior by mining enriched web access log data on association rules mining algorithm is used for mining frequent page sets and for generating interesting rules. The study also applied the mixture model to build a predictive model of navigational behaviors of users. The major results of the study are content and origin based data enrichment and a tree like visualization of frequent navigational sequences [30]

According to Iváncsy [29], User identification is one of the most challenging tasks in Web Usage Mining. This is because multiple users can use the same computer or a single user can use multiple computers to browse the website. Furthermore, proxy servers can hide relevant information about unique users as multiple computers appear on the internet using the same IP address through the proxy server.

Sharma [21] has conducted a research using Rochester Institute of Technology (RIT) web data, USA. The researcher has done the study by classifying the visitors into two, Inside USA and Outside USA using five days web usage data. In this study, hostname or IP was used to classify the visitors. This was done with the help of GeoIP JAVA API. This study has identified directories

that are frequently accessed by visitors from USA and directories that are frequently accessed by outside USA visitors.

Tadele [9] has conducted a research on a Web Usage Pattern Discovery of Addis Ababa University website. Tadele has utilized Web server log files and followed the steps: raw log preparation, pattern discovery and pattern analysis. He used Apriori algorithm for pattern discovery. Moreover, he utilized Weka preprocessing 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 [15] has conducted his study on exploring the navigational behavior of users of Addis Ababa University (AAU) official website. Awet has utilized Web 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.

Serkaddis [26] has used statistical and pattern discovery techniques to explore the usage of Ethiopian Airlines website. Serkaddis has utilized Web server log files and followed the steps: data preparation, pattern discovery and pattern analysis in her study. She used Apriori algorithm for pattern discovery with 16 common URL attributes. Moreover, she utilized MS Excel and Java

programming for data preparation, Mach5 for statistical analysis and Weka for pattern discovery. She has discovered a couple of important usage association rules. Findings show that the most frequently accessed web pages are the home page, online booking page and ShebaMiles page and suggested to place advertisements on price discount and other marketing promotions on these pages. These pages are also the top exit pages. She recommended the need for further research using algorithms such as FP-Growth for pattern discovery with sufficient amount of log files.

Getahun [31] has conducted a research as an extension to Serkaddis with a focus towards addressing various issues regarding the Ethiopian Airlines web usage status. He used the Web Usage Mining process model suggested by Sharma [12] as methodology and Server log data for pattern discovery in weka. He applied FP-Growth algorithm to discover interesting patterns in different regions and google Analytics and MS Excel as a tool to yield different useful statistical reports including top landing, top exit, and most frequently accessed pages. In his finding, he indicated that the home page of Ethiopian Airlines website accounts for over 50% of entry as well as exit page and other pages need optimization. Moreover he also indicated that there are similarities and variations from one region to another with respect to the accessibility and usage pattern. Finally, he has forwarded a recommendation for further researches and improvement areas in order to improve the website.

As can be observed from the above related works, this study is the first attempt made about web site usage pattern discovery and analysis on Ethio telecom. Like Getahun's work the current study adopted FP growth algorithm on the cleaned access log file. The contribution of this study is discovering and analyzing web usage pattern for web site optimization.

Hereunder in Table 2.5 summary of studies done on web usage mining are presented.

Author	Title	Methods/ Algorithm	Tools	Data source	Findings
Tadele Astatke (2011)	Web usage Pattern Discovery of AAU official website	WUM Process, Apriori	Wekapreprocessing,Python, Mach5, Weka	Web Server Log	Institute of Ethiopian Studies is most frequently accessed, About AAU Academic page, Academic Page Library page, Student Service and Administration pages are frequently accessed together
Awet Fesseha (2011)	Web Usage: Exploring Navigational Behavior of Users, the Case of the Official website of AAU	WUM Process, Apriori	WUMprep, Web Utilization Miner	Web Server Log	Registrar and Graduate Admission pages are most frequently accessed, Home Page Academic page
Serkaddis Adem (2011)	WEB USAGE PATTERN DISCOVERY: The Case of Ethiopian Airlines Website	WUM Process, Apriori	MS Excel, Java, Mach5, Weka	Web Server Log	Price discount information is not accessed as expected, most users enter and leave at the home page
Getahun Negatu(2013)	Webusage pattern discovery and analysis by region: the case of Ethiopian Airlines official website	WUM process, FP Growth	Wekagoogle Analytics and MS Excel	Web Server Log	The home page of Ethiopian Airlines website accounts for over 50% of entry as well as exit page and other pages need optimization. There are similarities and variations from one region to another with respect to the accessibility and usage pattern.

CHAPTER THREE

DATA PREPARATION

An important task in Web Usage Mining is the creation of an effective target data set to which web mining algorithms can be applied. The primary data source used in the process of Web Usage Mining is the server log files [32]. A Web server log is an important source for performing Web usage mining since it explicitly records the browsing behavior of users to the site.

Log files consist of large amount of irrelevant information so data log files cannot be directly used for Web Usage Mining task. Preprocessing on web usage data is required to eliminate noisy data and make data effective for further analysis task. The main preprocessing tasks include data cleaning, user identification, session identification, transaction identification and data transformation [32].

3.1. Data Collection

In this research, datasets are collected from the server of Ethio telecom official web site: WWW.Ethiotelecom.et .The user access logs for the full year of 2014 were transferred from the Web Server. The original size of the data was 9 GB before preprocessing. The server log is in common Log Format containing information about the access logs of the website. Below in figure 3-1 is the sample data extracted from the raw log file.

```
331911:624520|97.74.24.9 - - [05/Feb/2014:09:17:44 +0300] "GET /?q=node/222378 HTTP/1.0" 200 22729↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:45 +0300] "GET /?q=node/595589 HTTP/1.0" 200 22149↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:45 +0300] "HEAD /?q=node/445594 HTTP/1.0" 200 -↓
331911:625078|179.43.138.193 - - [05/Feb/2014:09:17:45 +0300] "GET /?q=user HTTP/1.0" 200 15322↓
331911:625093|108.163.195.198 - - [05/Feb/2014:09:17:46 +0300] "POST /?q=user/register HTTP/1.0" 200 16248↓
331911:624161|216.151.159.52 - - [05/Feb/2014:09:17:46 +0300] "GET /?q=node/28117 HTTP/1.1" 200 19936↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:46 +0300] "HEAD /?q=node/354354 HTTP/1.0" 200 -↓
331911:617886|94.242.255.187 - - [05/Feb/2014:09:17:46 +0300] "GET /?q=mobile-postpaid HTTP/1.1" 200 25882↓
331911:625081|142.91.62.129 - - [05/Feb/2014:09:17:46 +0300] "GET /?q=node/667767 HTTP/1.0" 200 18945↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:46 +0300] "HEAD /?q=node/595589 HTTP/1.0" 200 -↓
331911:625078|179.43.138.193 - - [05/Feb/2014:09:17:46 +0300] "POST /?q=user HTTP/1.0" 200 15650↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:46 +0300] "HEAD /?q=node/239983 HTTP/1.0" 200 -↓
331911:625096|216.152.251.24 - - [05/Feb/2014:09:17:46 +0300] "GET /?q=node/20065 HTTP/1.1" 200 20445↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:47 +0300] "HEAD /?q=node/354354 HTTP/1.0" 200 -↓
331911:625096|216.152.251.24 - - [05/Feb/2014:09:17:47 +0300] "GET /?q=node/9705 HTTP/1.1" 200 19214↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:47 +0300] "HEAD /?q=node/168790 HTTP/1.0" 200 -↓
331911:620191|72.167.131.154 - - [05/Feb/2014:09:17:47 +0300] "GET /?q=node/595589 HTTP/1.0" 200 22149↓
331911:625095|221.233.94.184 - - [05/Feb/2014:09:17:47 +0300] "GET /?q=node/325665 HTTP/1.1" 200 22571↓
```

Figure 3-1sample raw access log

3.2. Data Preprocessing

The inputs for data preprocessing phase are the access logs transferred from the web server and the outputs are user session file, transaction file and transformed csv file which will be used for pattern discovery task. The data preprocessing steps depicted in figure 3-2 are followed in order to clean data, identify session, select feature, identify transaction and transform data. These steps are customized from the basic log file preprocessing steps discussed in the literature review part.

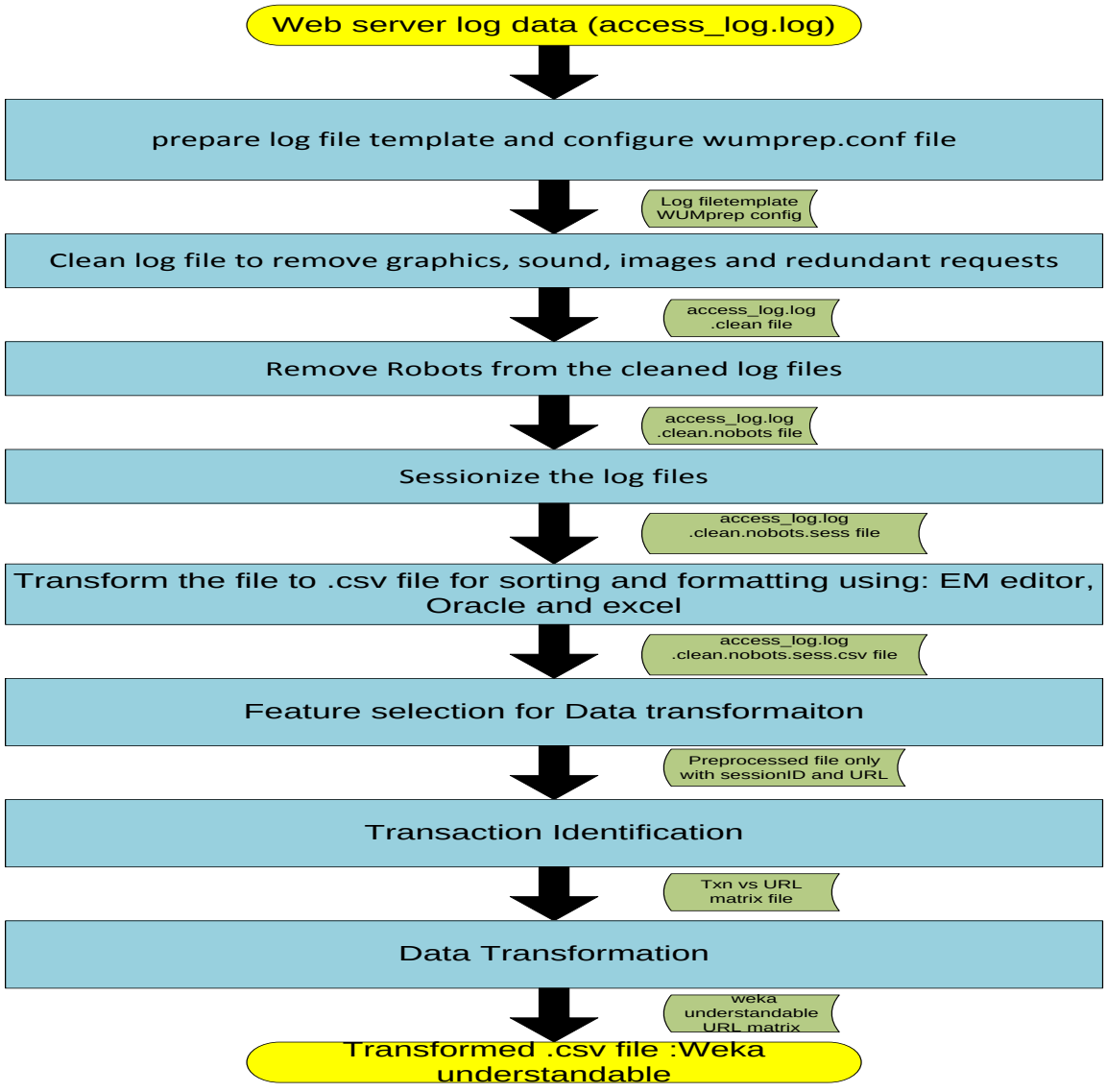


Figure 3-2 Preprocessing Steps followed in preparing server log file for mining

3.2.1. Data cleaning and Session identification

The HTTP protocol requires a separate connection for every file that is requested from a web server. Therefore a user's request to view a particular page often results in several log entries since graphics and scripts are downloaded in addition to the HTML file. Since the main intent of web usage mining is to get a picture of user's behavior, it is unnecessary to include file requests that the user didn't explicitly request [33]. Therefore, removing those secondary requests is significant. For such specific task, the researcher used WUMprepConfigEditor from WUMprep suite which is a graphical; Java based editor for WUMprep configuration files. It presents the configuration options of the different WUMprep scripts conveniently in a tabbed interface, with one tab per log file preparation task. The graphical user interface of this editor is depicted in fig 3-3.

To clean these irrelevant records, the researcher required to configure the editor with the necessary parameters. In the log filter, files with extensions, .ico, .gif, .GIF, .jpeg, .JPEG, .jpg, .JPG, .js, .css, .pdf, .txt, .png, and .flv were entered in the path parameter so that files with these extensions are excluded from the log file. While most of these already exist in the configuration editor, .pdf, .png,.swf and .flv were added by the researcher based on a closer look of the server log data and confirmation of the web master.

Another task on this editor before utilizing the created configuration file is importing the log file to be processed against the specified parameters. Other parameters such as Remove Robots and Sessionize are set to default. For instance, the removeRobots.pl perl script requires specifying the robot database for its lookup while removing robots, crawlers or spiders. In this case, it uses indexers.lst by default. Moreover, sessionize.pl script is set to use the maximum page view time of 1800 seconds (30 minutes) which is also the default setting. The configuration file should be saved in the same directory with the log file to be preprocessed.

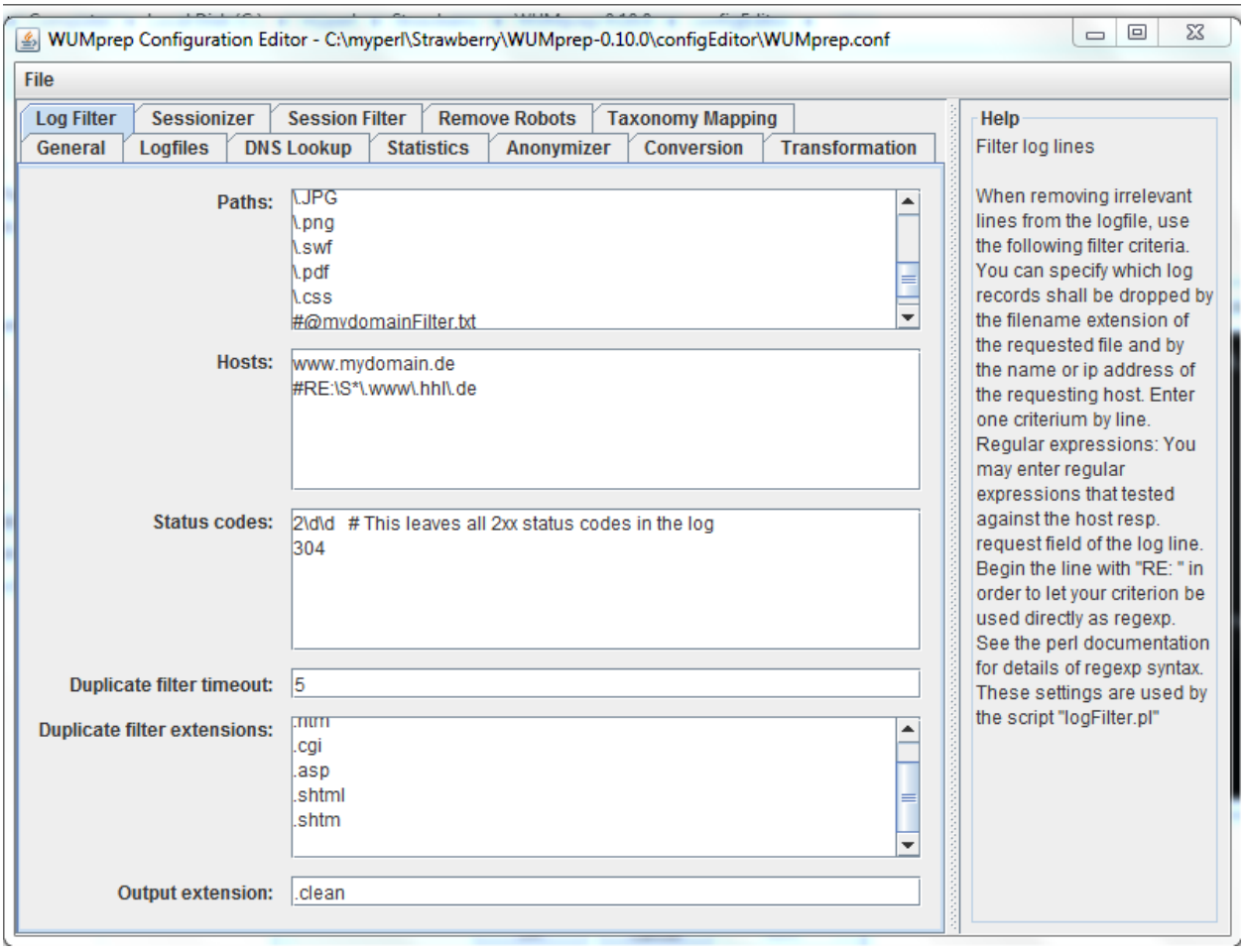


Figure 3-3WUMprep Configuration Screen

Another important setting for the WUMprep tool is specifying the format of the Server log in **log file Template** file. A sample log file template from the WUMprep suit is customized according to the parameters of the access log file used in this study. The log file template customized is shown below:

@host_ip@ @ident@ @auth_user@
[@ts_day@/@ts_month@/@ts_year@:@ts_hour@:@ts_minutes@:@ts_seconds@ @tz@] "@method@ @path@ @protocol@" @status@ @sc_bytes@

Finally wumprep.conf, logfileTemplateand indexers.lst are placed in the same directory/folder as the log file for preprocessing using the preprocessing Perl scripts.

Following this a list of the files comprising the Perl parts of the WUMprep modules of the WUM architecture are executed one-after-the-other as per the procedure discussed on the documentation of the software.

First, **logFilter.pl** was used to remove unnecessary access logs such as graphics, video and audio access logs that are repeated and redundant (see figure 3.4). This script will use the WUMprep configuration

file mentioned above. Then, **removeRobots.pl** was applied to remove crawlers (robots or spiders) from the log file (see figure 3.5). Next, **sessionize.pl** divides the log file into separate user sessions (see figure 3.6). And finally transformLog.pl used to convert the file to .csv format (see figure 3.7)

After running scripts, a cleaned file is created in the same directory as the original log file with '.clean' suffix. Similarly, after running removeRobots.pl, another file with '.nobots' extension is created. Note that the removeRobots.pl script uses the preprocessed file with '.clean' extension. Running the sessionize.pl script separates the given log file into different sessions. It creates a new file suffixed with '.sess' extension. This script uses the preprocessed file with '.nobots' extension. And finally .sess file is transformed to .csv file by running transformLog.pl script which use the default template log .csv file as an input.

The size of the last log file (access_log.clean.nobots.sess.csv) resulted from WUMprep steps, is reduced by 68.5 %(from 89,995,289 hits to 28,316,753 hits)

```
C:\myperl\Strawberry\WUMprep-0.10.0\configEditor>perl "C:\myperl\Strawberry\WUMp
rep-0.10.0\src\logFilter.pl"
Reading configuration file wumprep.conf...
Reading configuration file wumprep.conf...
Reading log format definition...
Field  0: host_ip
Field  1: ident
Field  2: auth_user
Field  3: ts_day
Field  4: ts_month
Field  5: ts_year
Field  6: ts_hour
Field  7: ts_minutes
Field  8: ts_seconds
Field  9: tz
Field 10: method
Field 11: path
Field 12: protocol
Field 13: status
Field 14: sc_bytes
C:\myperl\Strawberry\WUMprep-0.10.0\src\logFilter.pl:
Filtering input log file 1: C:\myperl\Strawberry\WUMprep-0.10.0\configEditor\acc
ess_log.log ...
89995289 lines processed, 57228983 lines removed (63.59%) - finished
```

Figure 3-4 Sample preprocessing screen with '.logFilter.pl' script

```
C:\myperl\Strawberry\WUMprep-0.10.0\configEditor>perl "C:\myperl\Strawberry\WUMp
rep-0.10.0\src\removeRobots.pl"
Reading configuration file wumprep.conf...
Reading log format definition...
Field  0: host_ip
Field  1: ident
Field  2: auth_user
Field  3: ts_day
Field  4: ts_month
Field  5: ts_year
Field  6: ts_hour
Field  7: ts_minutes
Field  8: ts_seconds
Field  9: tz
Field 10: method
Field 11: path
Field 12: protocol
Field 13: status
Field 14: sc_bytes
Reading configuration file wumprep.conf...
Processing list of known robots
Removing robots from log C:\myperl\Strawberry\WUMprep-0.10.0\configEditor\access
_log.log.clean ...
Processed 32733000 lines of log
Total number of hits: 32766306
Number of robot hits: 3780611
% of total by robots: 11.54

Writing output and performing DNS lookups (if neccessary)

Storing robot hosts
```

Figure 3-5 Sample screen of removing robots with '.removeRobots.pl' script

```
C:\myperl\Strawberry\WUMprep-0.10.0\configEditor>perl "C:\myperl\Strawberry\WUMp
rep-0.10.0\src\sessionize.pl"
Reading configuration file wumprep.conf...
Reading configuration file wumprep.conf...
Reading log format definition...
Field  0: host_ip
Field  1: ident
Field  2: auth_user
Field  3: ts_day
Field  4: ts_month
Field  5: ts_year
Field  6: ts_hour
Field  7: ts_minutes
Field  8: ts_seconds
Field  9: tz
Field 10: method
Field 11: path
Field 12: protocol
Field 13: status
Field 14: sc_bytes
C:\myperl\Strawberry\WUMprep-0.10.0\src\sessionize.pl:
Defining sessions in log C:\myperl\Strawberry\WUMprep-0.10.0\configEditor\access
_log.log.clean.nobots ...
28956233 lines processed - finished

C:\myperl\Strawberry\WUMprep-0.10.0\configEditor>
```

Figure 3-6 Sample sessi-onization screen with 'sessionize.pl' script

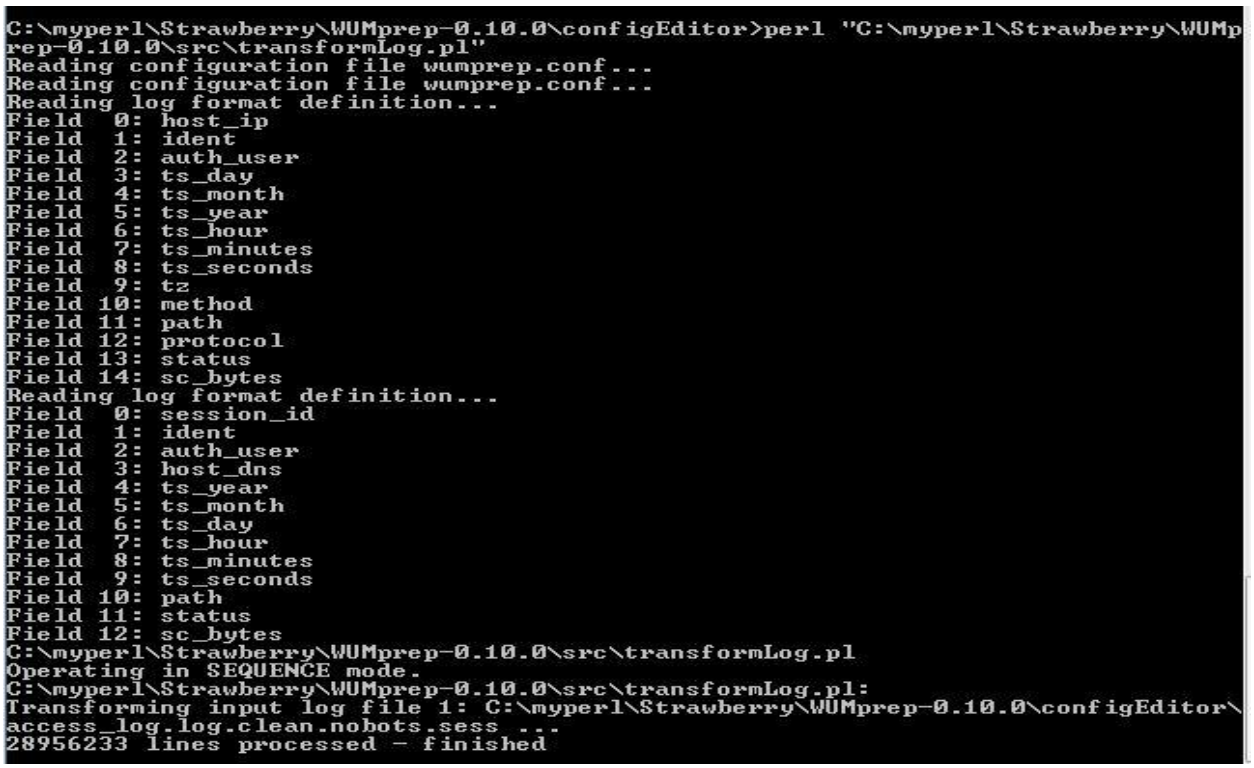


Figure 3-7 Sample transformation screen with 'transformLog.pl' script

In addition to the above, the researcher is only interested in users request for information from server; the records with POST or HEAD method should be removed. Log files should have the records with GET methods. By using Oracle software only GET log files are filtered and used for next steps. In doing so, the size of log files is reduced by 41 % (from 28,316,753 hits to 17,528,156 hits)

The HTTP status code is then considered in the next process of cleaning by examining the status field of every record in the web access log, the records with status code over299 or under 200 are removed because the records with status code between 200 and 299, gives successful response.

At last, after accomplishing the above discussed cleaning tasks, out of the total of 89,995,289hits/records in the log files, only 17, 546, 232hits/records remained. This indicates that majority of the records in the log files (about 81.5%) are removed while preprocessing. Moreover, the size of the log files shrinks to approximately (19.5%) of the original size after applying the data cleaning steps.

3.2.2. Feature Selection

Only the URL attributes are relevant for WUM association rules discovery. Hence, the other attributes are removed from the dataset. But as the number of unique URLs/paths in the log file was too large it was required to select only few important features from them. As a result, the preprocessed and consolidated log file was edited with oracle and grouped by path with frequency so that unique URL in the data with their counts is obtained. Then, the frequency of each URL was calculated. The result was sorted in descending order of URL count so that the most frequently accessed are displayed on top using excel sheet. The researcher then decided to select only those URLs that satisfy minimum count/ frequency. Accordingly, 70 URLs whose frequency is greater than or equal to 5000 were selected and used for further cleaning tasks. This cut-off point is selected on the basis that those URLs that are accessed 5000 times and more are significant.

3.2.3. Transaction identification

In this phase of pre-processing, transactions were identified for the consolidated data. Before running the custom developed java code for identifying the transactions, the log data is further passed through some cleaning steps to remove irrelevant that logs escaped from previous processing activities. The preprocessed log file so far is containing all the access logs with all paths regardless of the frequency of the path. However the paths are filtered in feature selection stage as only top accessed paths are important for the pattern discovery. Similarly the log data should only contain logs with top accessed paths. Accordingly, using oracle the log data is filtered and only logs with accessed paths are left. By this the size becomes 4, 018,098, i.e., Reduced by 77%. Duplication of logs are cleaned which also reduced the log file to 2, 152, 741. This is the final log data which is used for transaction identification.

Once having the cleaned log data, first step was to sort each data in order to arrange the same session IDs together. The resulting file was saved as a separate text file. Then the custom developed Java program was used to identify transaction in each data and the resulting data was saved as a separate file. Note that in this case the first column holds Transaction ID and the next 70columns hold the URLs (URL1, URL2… URL70). As depicted in fig 3.8.Corresponding to each Transaction ID, if a specific URL is accessed, it is indicated with ‘Y’ otherwise, ’?’ is used to signify that the URL does not exist in the transaction. The URLs in the sorted log file was compared with the 70 top accessed URLs to check whether the two URLs match or not. The java code produced a transaction –URL matrix with a total number of 1, 318,248 transactions. Refer to Appendix C for the Java source code used.

```
1.52.230.154 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.54.112.237 ? ? Y ? ? Y ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Y ? ? ? ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.54.8.229 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.57.95.106 ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.59.32.103 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.85.4.142 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.102.188 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.102.52 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.103.70 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.113.140 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.153.80 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.157.86 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.97.143 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
1.9.99.85 ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.1.19.11 ? ? Y ? ? ? ? ? ? ? ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.154.4.5 ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.154.5.3 ? ? Y ? ? Y ? ? Y ? Y ? ? ? Y ? Y ? ? ? ? ? ? ? Y ? ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Y ?
10.32.147.88 ? ? Y ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.35.22.14 ? ? Y ? ? Y ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.40.11.170 ? ? Y ? ? Y ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? Y Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.40.127.34 ? ? Y ? ? Y ? ? Y ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.40.18.234 ? ? Y ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
10.40.2.170 ? ? Y ? ? Y ? ? Y ? ? ? ? ? Y ? Y ? ? ? ? ? ? ? Y ? ? Y ? ? Y Y ? ? ? ? ? ? ? Y ? ? ? ? Y ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
```

Figure 3-8 Sample Transaction

3.2.4. Data format Conversion

The final step in Data Preprocessing is to convert the transactions to a Weka understandable format. A custom developed java code is used for removing the session ID information as it is not required for the weka processing. The resulted file is further edited with MS Excel like converting the text file in to table and labeling column headings as URL1, URL2… URL70. Finally the resulted transaction contains 301,580 numbers of transactions and used for pattern discovery in weka .See fig 3.9.below for sample dataset consisting of URL Matrix and refer to Appendix: C for the Java source code used.

URL1	URL2	URL3	URL4	URL5	URL6	URL7	URL8	URL9	URL10	URL11	URL12	URL13	URL14	URL15	U
Y	?	?	?	?	?	?	?	Y	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
Y	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Y	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Y	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Y	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
Y	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?
?	?	?	?	?	?	?	?	Y	?	Y	?	?	?	?	?

Figure 3-9 Sample Dataset

After dataset preparation, association rule discovery algorithms are used for extracting patterns during experimentation using WEKA knowledge discovery tool and statistical information using Google Analytics.

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS

To discover the pattern of web site visitors and make analysis on the result, an experiment is conducted using the preprocessed data. Some modifications has been done on the dataset in the process of getting interesting patterns whenever required. As shown on the figure 4-1, the experimentation and findings focus on two main aspects: Statistical analysis and pattern discovery.

The computer used for conducting this experiment is Dell Laptop computer having Intel Core i5 processor@2.8 GHz speed and 4 GB RAM size. The Operating System is Windows 7 Ultimate (32-bit). The analysis tools used for discovering the pattern are Weka 3.7.9 knowledge discovery software for pattern discovery and Google Analytics and MS Excel 2010 for Statistical Analysis.

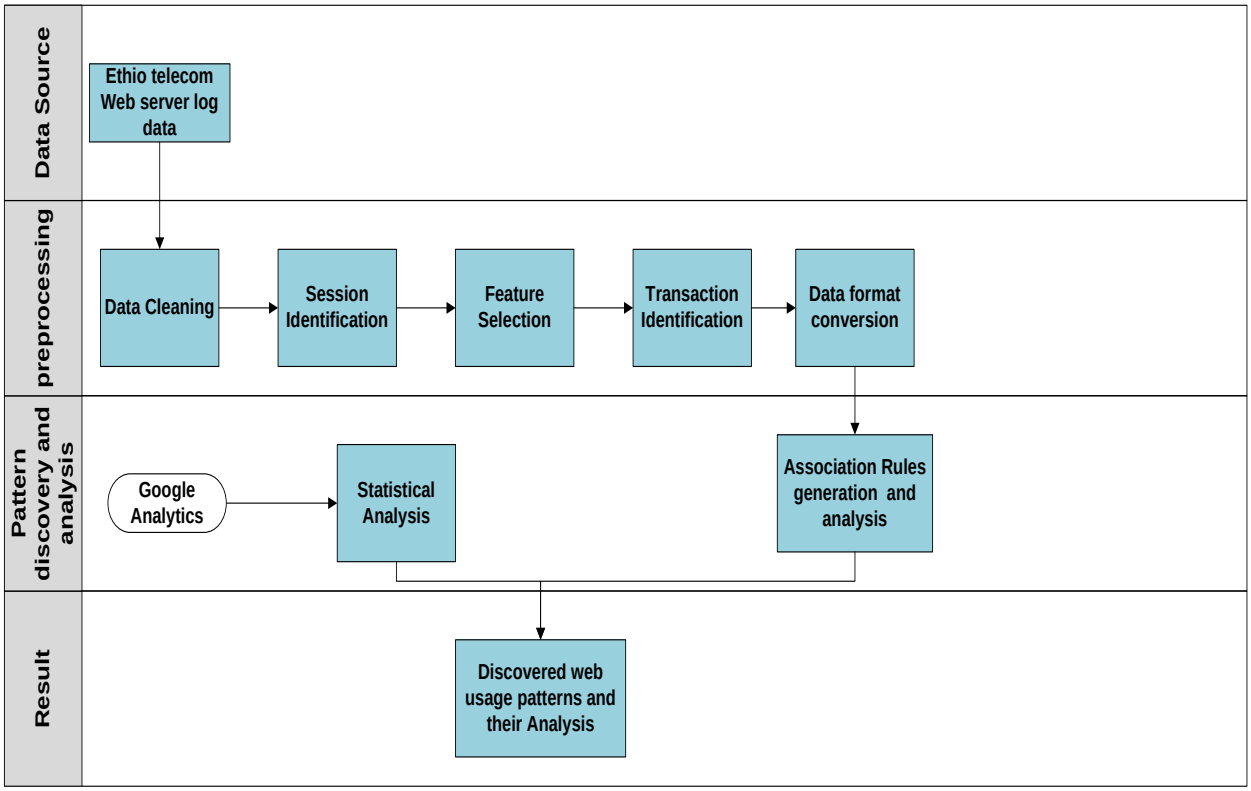


Figure 4-1: Architecture of the current study

The above figure 4-1 shows the architecture of high level tasks performed in this study, these are web server log file exporting from the source, preprocessing , Experimentation and analysis and finally the findings of the study.

4.1. Statistical Analysis

Google analytics, a tool designed for analysis of web usage data, has been used for statistical analysis of the web site visitors. Both standard and customized reports can be generated using this tool. Ethio telecom official website is connected to this well-known web analysis tool since December, 2014. So the statistical analysis is conducted on six months data (from December, 2014- May, 2015) that Google analytics has collected regarding the usage of this web site.

In graphical interface of Google analytics, the reports are organized under various taps based on their contents. The researcher has applied the same organization style for this section by taking only the basic categories of reports from the Google analytics. These are Audience, Acquisition and Behavior.

During the given period from December, 2014 to May, 2015 the website was visited 257,345 times with an average daily visit of 1,414.

The description of the URLs used in the statistical reports is presented in Appendix A.

4.1.1. Audience Reports

This category contains reports about characteristics of the website users such as their geographic location, their usage behavior over multiple visits of the site, what devices they used, and how loyal and engage with the website business. Reports from this category are discussed as follows:

4.1.1.1.Web site visitors by location

This report is the most commonly used one in the Google analytics that shows the geographical location of website visitors. Table 4.1 presents summary of reports by location of visitors to Ethio telecom website.

Country	Acquisition			Behavior		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration
	407,830 % of Total: 100.00% (407,830)	63.14% Avg for View: 63.09% (0.06%)	257,487 % of Total: 100.06% (257,320)	26.30% Avg for View: 26.30% (0.00%)	4.22 Avg for View: 4.22 (0.00%)	00:03:10 Avg for View: 00:03:10 (0.00%)
1. Ethiopia	**220,009** (53.95%)	58.12%	127,873 (49.66%)	21.84%	4.49	00:03:24
2. Indonesia	**113,405** (27.81%)	66.41%	75,307 (29.25%)	32.42%	3.72	00:03:00
3. United States	**16,241** (3.98%)	78.04%	12,674 (4.92%)	33.19%	4.19	00:03:00
4. India	**15,966** (3.91%)	71.31%	11,385 (4.42%)	37.66%	3.51	00:02:31
5. (not set)	**10,221** (2.51%)	72.23%	7,383 (2.87%)	33.21%	3.79	00:02:42
6. Ireland	**5,282** (1.30%)	78.19%	4,130 (1.60%)	43.79%	3.68	00:03:27
7. Germany	**3,222** (0.79%)	74.67%	2,406 (0.93%)	30.57%	4.25	00:02:55
8. Estonia	**2,810** (0.69%)	59.18%	1,663 (0.65%)	25.84%	3.96	00:02:55
9. Netherlands	**2,620** (0.64%)	62.79%	1,645 (0.64%)	21.37%	4.23	00:02:47
10. United Kingdom	**1,452** (0.36%)	74.59%	1,083 (0.42%)	13.57%	6.11	00:02:45

Table 4-1 Summary of Ethio telecom website visitors by location

As shown in the table 4.1 above, the first three listed countries are Ethiopia (53.95%), Indonesia (27.81%) and United States (3.98%). This means that majority of visitors are local visitors. This indicates the website is not popular internationally. And also from the regional perspective, next to Africa (53.95%), Asia (30.72%), North America (3.98%), and Europe (3.78%) are the three continents with high access rate of the Ethio telecom website in 2nd, 3rd and 4thplace which points out the website lacks popularity towards the rest regions of the world.

4.1.1.2.New Vs Regular visitors of the website

This report gives a quick look at the ratio of the first time and repeated visitors. The figure 4-2 below shows the result of new vs returning visitors for the case of Ethio telecom website.

Figure 4-2 New vs Regular visitors of Ethio telecom website.

As shown in the above figure, the percentage of new visitors (63.1%) exceeds the percentage of regular visitors (36.9%). This implies the website is not capable enough to attract visitors for repeated visits.

4.1.1.3. Top Browsers

Understanding the browsers type that visitors used to reach and consume the website is necessary to ensure compatibility with the current versions and to plan upcoming implementations.

Browser	Acquisition			Behavior		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration
	407,830 % of Total: 100.00% (407,830)	63.14% Avg for View: 63.09% (0.06%)	257,487 % of Total: 100.06% (257,320)	26.30% Avg for View: 26.30% (0.00%)	4.22 Avg for View: 4.22 (0.00%)	00:03:10 Avg for View: 00:03:10 (0.00%)
1. Opera Mini	128,857 (31.60%)	66.09%	85,168 (33.08%)	31.93%	3.73	00:02:59
2. Firefox	68,282 (16.74%)	54.63%	37,300 (14.49%)	3.48%	4.87	00:03:27
3. Chrome	67,580 (16.57%)	55.51%	37,515 (14.57%)	17.92%	5.10	00:03:11
4. Safari	36,360 (8.92%)	74.70%	27,162 (10.55%)	45.29%	3.22	00:02:45
5. Android Browser	31,263 (7.67%)	67.47%	21,094 (8.19%)	45.40%	3.61	00:02:58
6. Internet Explorer	24,737 (6.07%)	51.55%	12,752 (4.95%)	6.23%	5.73	00:04:07
7. S40 Ovi Browser	16,237 (3.98%)	78.67%	12,773 (4.96%)	43.82%	3.64	00:03:25
8. UC Browser	14,883 (3.65%)	74.54%	11,094 (4.31%)	40.58%	3.62	00:02:38
9. Opera	9,349 (2.29%)	51.44%	4,809 (1.87%)	25.91%	4.22	00:03:07
10. NetFront	3,197 (0.78%)	75.91%	2,427 (0.94%)	44.92%	3.02	00:04:28

Table 4-2 Top browsers used for accessing Ethio telecom web site.

The three top browsers that accessed Ethio telecom website in the above table 4-2 are Opera mini (31.6 %), Firefox (16.74%) and chrome (16.57%). As the opera mini is mobile browser, the result indicates that mobile users are leading the web users. And also Firefox and chrome are the 2nd and the 3rd most used browsers to access the website. This indicates these browsers are becoming more popular even more than the internet explorer.

Generally the analysis from the selected audience reports shows various characteristics of Ethio telecom website users. Majority of the website visitors comes from local. Most of the users use opera mini, which means that most of the website visitors are mobile users. Also, most of the users are new to the website. That means, users who visit the website are not revisiting again.

4.1.2. Acquisition

In this category, reports about how visitors find the web site, mainly regarding the traffic sources those users are coming from. Hereunder reports about top channels and top website referrals are presented.

4.1.2.1.Top channels

Channels are groups of traffic sources. Table 4-3 shows the top four channels that the website traffics are originated from.

		Acquisition			Behavior		
		Sessions ↓	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration
		407,830	63.09%	257,320	26.30%	4.22	00:03:10
1	Direct	268,948			37.19%		
2	Organic Search	120,097			4.99%		
3	Referral	14,217			6.94%		
4	Social	4,568			6.09%		

Table 4-3 Top group of channels as traffic sources

The above table 4-3 shows, the first most used channel from the list which accounts for 37.19% is direct channel and this channel is used to access the website by typing URL directly on top of the browser. The nice implication of the report is that users can easily remember the URL name or the website is book marked by most users. The second most popular channel for this website which accounts for 4.99% is the organic search where people go to organic search engine like Google and type some key word for searching the website. This is because of the popularity of this search engine.

4.1.2.2. Top Website Referrals

Referrals report gives all of the other websites that are driving traffic in to this web site .In other words; 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. Table 4.4 below summarizes the top ten referrals to Ethio telecom website.

Source	Acquisition			Behavior		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration
	18,785 % of Total: 4.61% (407,830)	60.79% Avg for View: 63.09% (-3.66%)	11,419 % of Total: 4.44% (257,320)	6.73% Avg for View: 26.30% (-74.42%)	4.69 Avg for View: 4.22 (11.19%)	00:02:56 Avg for View: 00:03:10 (-7.54%)
1. search.tb.ask.com	**3,720** (19.80%)	52.88%	1,967 (17.23%)	0.73%	4.90	00:03:48
2. lm.facebook.com	**2,181** (11.61%)	66.71%	1,455 (12.74%)	7.06%	3.36	00:01:52
3. l.facebook.com	**1,526** (8.12%)	55.83%	852 (7.46%)	3.47%	4.66	00:03:23
4. employethiopia.com	**1,255** (6.68%)	47.89%	601 (5.26%)	1.51%	4.33	00:03:19
5. ethionet.et	**1,049** (5.58%)	60.63%	636 (5.57%)	1.81%	6.51	00:03:34
6. en.wikipedia.org	**514** (2.74%)	81.71%	420 (3.68%)	6.61%	5.79	00:02:22
7. facebook.com	**465** (2.48%)	65.16%	303 (2.65%)	4.30%	4.33	00:02:45
8. m.facebook.com	**304** (1.62%)	78.62%	239 (2.09%)	15.13%	2.80	00:01:19
9. africantelecomsnews.com	**303** (1.61%)	79.87%	242 (2.12%)	2.31%	7.63	00:03:20
10. simple-share-buttons.com	**241** (1.28%)	100.00%	241 (2.11%)	100.00%	1.00	00:00:00

Table 4-4: Top web site referrals to Ethio telecom website

Search.tb. ask.com(19.8%) is listed as predominant referral to the website; however, literatures [34] notes that Search.Tb.Ask.com is a browser hijacker, which is promoted via other free downloads, and once installed it will change the browser homepage to home.tb.ask.com, set the default search engine to search.tb.ask.com, and install a toolbar . The next two sources that are listed as second and third most popular referrals to this site are l.facebook.com (11.61) and

lm.facebook.com (8.12%).Literatures discussed that both of these links are new referral sources from Face book which started appearing around the beginning of April 2014[36].

From the above list of referrals, most popular search engines, like Google or yahoo are missing. This means that the website is not well known by the search engine, or it may contain content which is not searchable by search engine.

Generally the acquisition report shows majority of the visitors are accessing the web site by directly writing the domain name information or from book marked list and most popular referrals are not found in the top referrals list of the website.

4.1.3. Behavior Reports

This type of report is about how users interact to the website including what contents they are looking at or how users flow between the pages or screens. Hereunder, reports about top accessed pages, top landing pages, top exit pages and behavior flow are discussed.

4.1.3.1.Top Accessed Pages

All pages report is one of the most interesting reports from behavior reports where most frequently accessed pages can be found listed in decreasing order of their frequency. The top ten frequently accessed pages are depicted on the table 4-5.

	Page Title	Pageviews	Pageviews
		1,719,505 % of Total: 100.00% (1,719,505)	1,719,505 % of Total: 100.00% (1,719,505)
1.	ethiotelecom	451,878	26.28%
2.	Vacancy | ethiotelecom	279,168	16.24%
3.	Manage your 3G/4G mobile internet data usage with these helpful hints. | ethiotelecom	157,479	9.16%
4.	3G Internet Package | ethiotelecom	107,718	6.26%
5.	3G Internet Usage Tutorial | ethiotelecom	53,275	3.10%
6.	Mobile Internet | ethiotelecom	38,583	2.24%
7.	Search | ethiotelecom	30,299	1.76%
8.	Bid | ethiotelecom	26,079	1.52%
9.	Business Internet | ethiotelecom	25,537	1.49%
10.	ADSL Internet | ethiotelecom	24,201	1.41%

Table 4-5: Top accessed Pages in the website

According to this report ‘home page’, ‘vacancy page’ and 3G/4G mobile internet tutorial pages are the three top accessed pages which accounts for 26.28%, 16.24% and 9.18% respectively. So

special attention should be given to these pages in terms of performance optimization and business promotion. The reason for these top three pages is, the home page is the default entry pages, vacancy is used by visitors who are looking for a job and the tutorial page is due to the 3G/4G internet service has been highly promoted by Ethio telecom during the report period.

4.1.3.2. Top Landing Pages

Landing pages are first visited pages from all pages of the website. Table 4-6 below shows the number of visits and new visits, bounce rate, pages per visit and duration of visit for the top landing pages.

Landing Page	Acquisition			Behavior			Conversions		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Goal Conversion Rate	Goal Completions	Goal Value
	58,663 % of Total: 100.00% (58,663)	57.94% Avg for View: 57.91% (0.05%)	33,990 % of Total: 100.05% (33,974)	17.79% Avg for View: 17.79% (0.00%)	4.44 Avg for View: 4.44 (0.00%)	00:03:08 Avg for View: 00:03:08 (0.00%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)	$0.00 % of Total: 0.00% ($0.00)
1. /	26,304 (44.84%)	57.04%	15,003 (44.14%)	1.38%	5.57	00:03:36	0.00%	0 (0.00%)	$0.00 (0.00%)
2. /?q=Manage3GInternet	14,564 (24.83%)	72.73%	10,592 (31.16%)	60.90%	2.87	00:03:00	0.00%	0 (0.00%)	$0.00 (0.00%)
3. /?q=vacancy	8,510 (14.51%)	37.29%	3,173 (9.34%)	0.38%	3.40	00:02:12	0.00%	0 (0.00%)	$0.00 (0.00%)
4. /?q=internet-3gpackages	1,459 (2.49%)	69.02%	1,007 (2.96%)	0.21%	5.42	00:02:57	0.00%	0 (0.00%)	$0.00 (0.00%)
5. /?q=bid	747 (1.27%)	26.64%	199 (0.59%)	0.80%	3.04	00:02:18	0.00%	0 (0.00%)	$0.00 (0.00%)
6. /?q=huaweievdocdma	457 (0.78%)	81.84%	374 (1.10%)	0.00%	3.45	00:01:46	0.00%	0 (0.00%)	$0.00 (0.00%)
7. /?q=internet-mobileinternet	334 (0.57%)	65.27%	218 (0.64%)	1.50%	6.84	00:04:29	0.00%	0 (0.00%)	$0.00 (0.00%)
8. /?q=faq	288 (0.49%)	71.18%	205 (0.60%)	0.00%	4.60	00:03:00	0.00%	0 (0.00%)	$0.00 (0.00%)
9. /?q=internet-adsl	264 (0.45%)	65.53%	173 (0.51%)	0.00%	5.44	00:03:33	0.00%	0 (0.00%)	$0.00 (0.00%)
10. /3gtips	229 (0.39%)	75.98%	174 (0.51%)	37.55%	4.92	00:03:39	0.00%	0 (0.00%)	$0.00 (0.00%)

Table 4-6 Top landing pages in the website

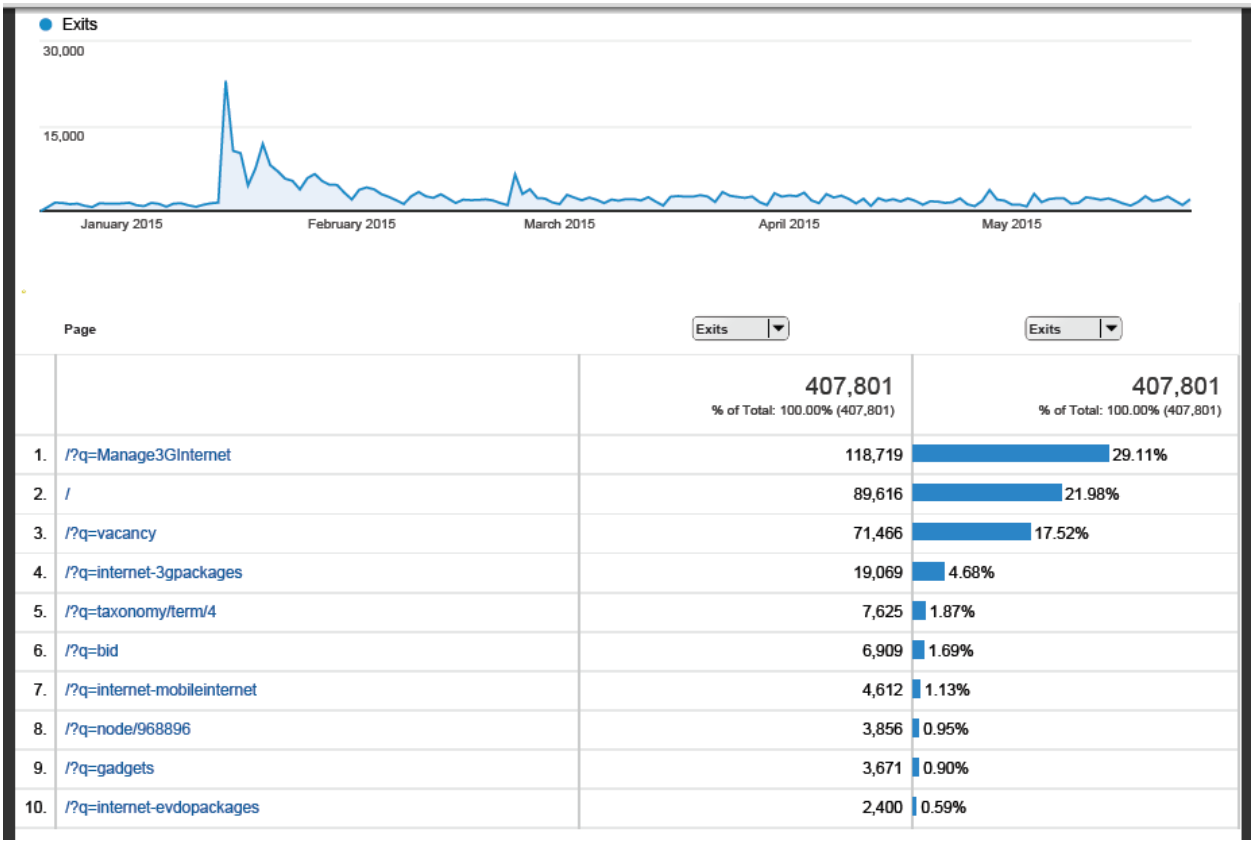
The first and the most frequently used as landing page is the home page of Ethio telecom home page with 44.84%. The second landing page is 3G tutorial with 24.83% and the fourth is internet 3g with 2.49%. These pages got the top place due to the fact that 3G product got introduced and promoted at that time. The vacancy page with 14.51% listed in the 3rd place and the bid page with

1.24% at 5th place are pages which are normally the top landing pages next to home page. Therefore, it is required to make these pages resourceful and more interesting.

4.1.3.3.Top Exit Pages

Top exit pages specify the pages where the visitor last visited the website. These are the exit points where most visitors actually leave the website. So the content of these pages should be reviewed as to make sure that page includes the information that visitors need before they go.

In some cases, a page can be both landing and exit page. Figures 4-3 below shows the top ten exit pages of Ethio telecom website for the specified period.



	Page	Exits	Exits
		407,801 % of Total: 100.00% (407,801)	407,801 % of Total: 100.00% (407,801)
1.	/?q=Manage3GInternet	118,719	29.11%
2.	/	89,616	21.98%
3.	/?q=vacancy	71,466	17.52%
4.	/?q=internet-3gpackages	19,069	4.68%
5.	/?q=taxonomy/term/4	7,625	1.87%
6.	/?q=bid	6,909	1.69%
7.	/?q=internet-mobileinternet	4,612	1.13%
8.	/?q=node/968896	3,856	0.95%
9.	/?q=gadgets	3,671	0.90%
10.	/?q=internet-evdopackages	2,400	0.59%

Figure 4-3 Top Exit pages

As shown in the figure 4-3, the 3G internet tutorial page is the top exit page with 29.11%. This is due to the page is where visitors read or download the manual for 3G product. The home page (21.98%) and vacancy page (17.52%) are the 2nd and the 3rd inlist respectively. This indicates visitors are not encouraged to navigate further after accessing these pages.

4.1.3.4. Behavior flow

This report visualize the path that visitors travel from one page or event to the next and it helps to discover what content keeps visitors engaged with the site. As shown in figure 4-4, the boxes or nodes represent the pages to which traffic flows and the connection between the nodes represent the path from one node to another in the navigation panel. It is possible to get the navigational flow through the specific page selected.

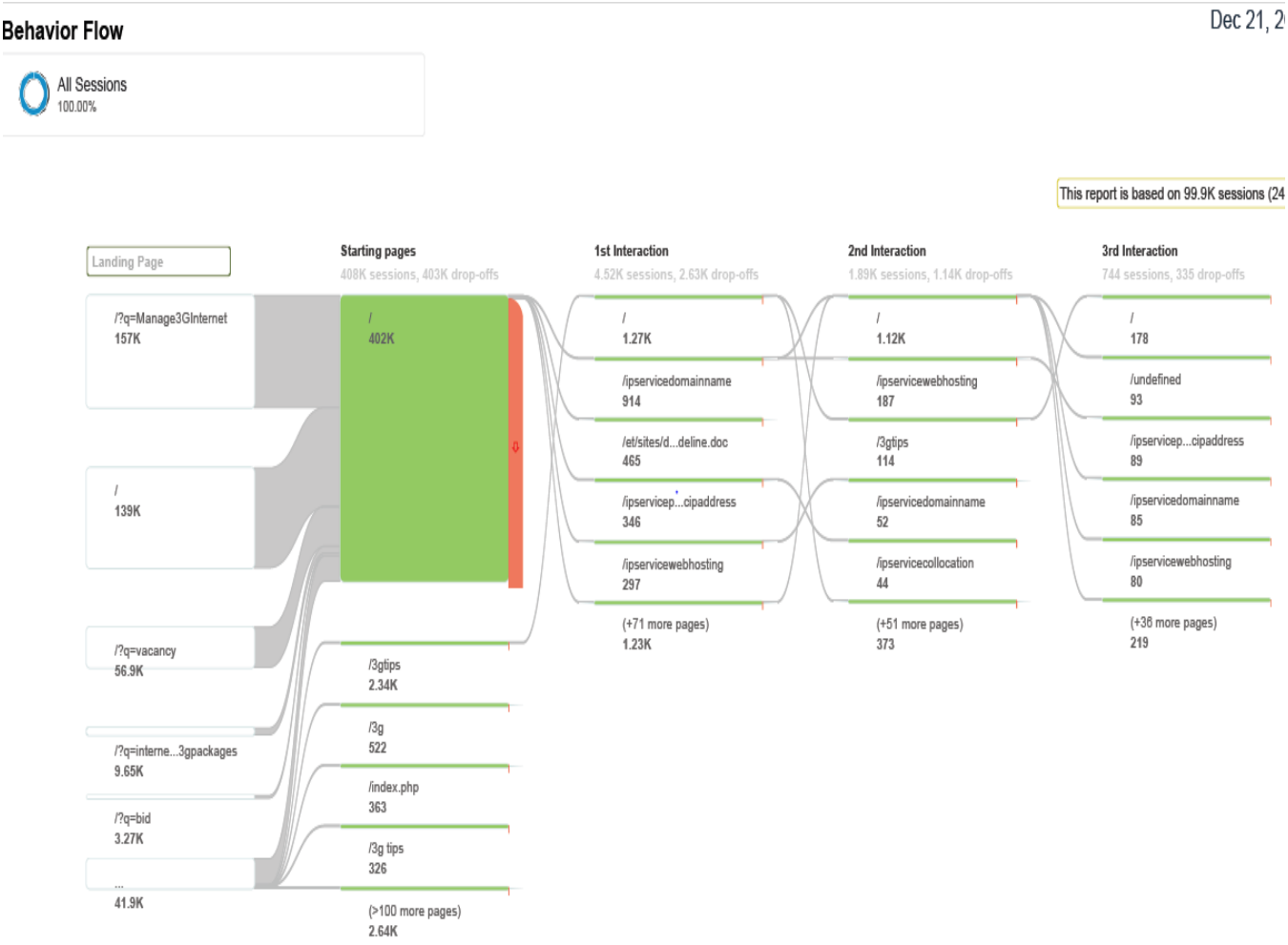


Figure 4-4 Behavior flow within the website of Ethio telecom

The figure 4-4 shows most of the time ‘home page’ is the starting page before taking the next interactions and it is as well the exit page indicated by red line.

4.1.4. Trend Analysis by Month

This report is about the trend of web site access over a year where the access statistics are calculated for each month.

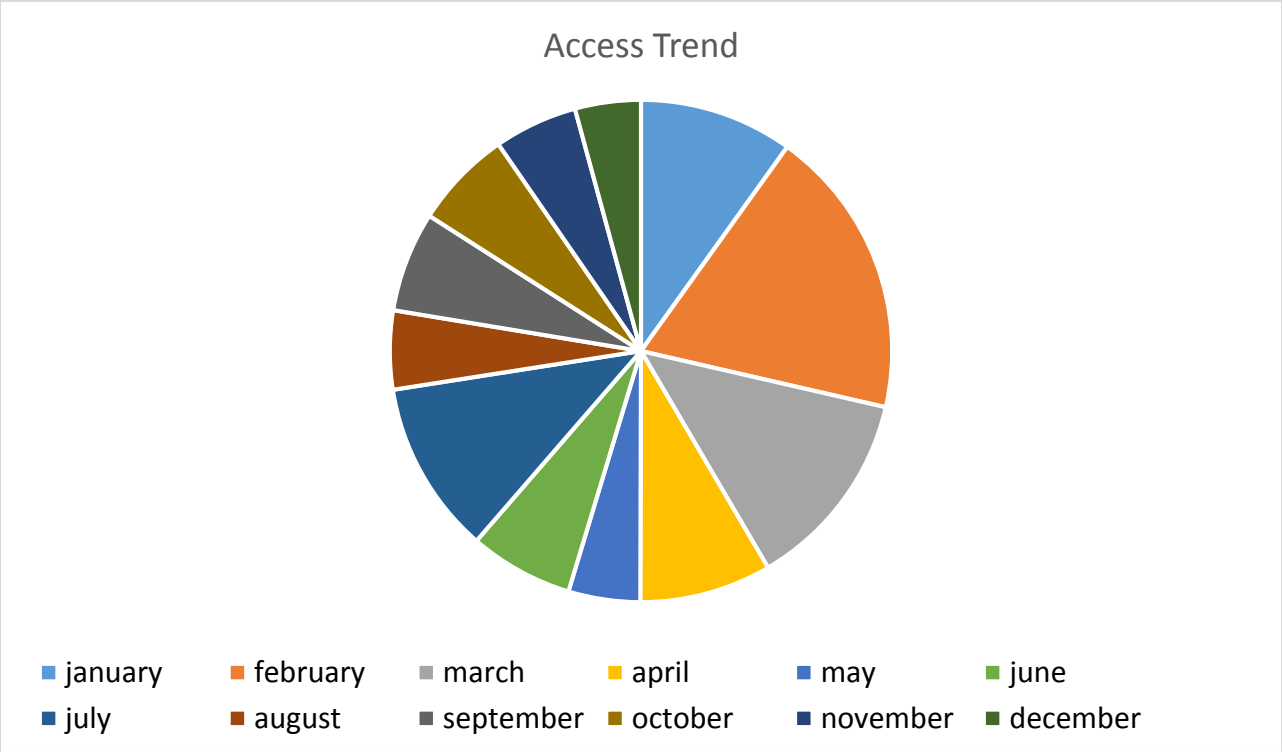


Figure 4-5 Access trend over 12 months

As depicted in the figure 4.5 above, the highest access to the website was recorded in February and the lowest in December. The reason is in the company trend new services are launched around New Year and obviously the first month is used for service disbursement and the next month which is February is most probably the time that users start using the service resulting high traffic to the official website as users are expecting news about the newly introduced product.

4.2. Pattern Discovery and Analysis

In this study, an experiment is conducted using weka software to discover patterns about Ethio telecom web usage. The data set used for this experiment is an access log dataset of Ethio telecom official website which passed through various preprocessing stages and finally converted to .csv format required by the FP growth algorithm in weka using custom developed java code.

The end result of weka after applying FP growth algorithm is a list of association rules. These rules were analyzed to interpret the interesting rules. As all strong patterns are not necessarily interesting, the researcher first identified which of the strong rules are interesting and which are not. Then the interpretation of the interesting rules is made.

The dataset used for the pattern discovery task has initially 60 attributes and 301,580 transactions. After conducting several experiments with this dataset, the researcher in consultation of web master noticed that most of the generated rules were not interesting like rules stating the 'home page' is the gate to others page. This was mainly due to the fact that the derived associations were either implied by the design or they were merely trivial relationships which do not convey new knowledge. Besides, there appeared also some relations that show access to internal functions which are used to pass parameter to other applications or the contents\sub contents within the page.

Due to this, by applying various rearrangements on the dataset like merging child attributes to their respective parent category, the weka result has been analyzed if interesting rules have come.

Moreover, a separate experiment is conducted by removing the main dominant pages visited by 90% and above visitors like the product page as most of the associations generated were overwhelmed by these pages and some resulted rules are considered in the analysis.

Finally an experiment has been conducted on Amharic pages separately so as to check whether interesting patterns have come out however it resulted overlapping rules with those rules generated during aggregate data. The result is discussed in experiment 1 below.

In each experiment the threshold values for minimum support and confidence are selected by heuristics after conducting several experiments with different values.

Experiment 1. Pattern Discovery and Analysis using the Aggregate Dataset

Summary of experimental result on aggregate dataset including all attributes is discussed below. For detail Weka Output, see Appendix B.

To extract association rules in aggregate dataset, a minimum Support = 0.01 (1%) and Minimum Confidence = 0.75 (75%) is used. These threshold values are selected because more interesting association rules are generated at these support and confidence level after trying several experiments with plus or minus of these threshold values.

In this case a total of 96 rules are generated, the researcher with the consultation of the web master founds that 10 of them are interesting and its analysis is given below by grouping related rules together whenever appropriate for the discussion.

Group 1: Association Rules about group discussion/forum sites and home page

The top 20 listed rules, in decreasing order of confidence level have showed the pattern existed among four sites: the home page, call conference page, general discussion page and forum page. This indicates that the pattern of these pages is the dominant one from all patterns existed in the dataset.

In common, all these rules indicate that the registration page (URL6) has been visited before visiting any one of those group discussion sites (call conference (URL12), general discussion page (URL8) and forum page (URL3)). This is mainly to get more privilege to the group discussion; visitors got first registered for the same.

The other common implication that can be derived from this grouping is that Home Page is visited by majority of visitors who have visited the discussion pages where the implication goes to one of the other discussion site. This is because the root/home page is usually the gate to the website contents.

The following are interesting association rules selected in consultation with domain experts.

Rule 1. [URL1=Y, URL8=Y, URL12=Y]: 20077 ==> [URL3=Y]: 20071 <conf:(1)> lift:(5.19) lev:(0.05) conv:(2315.49)

This rule stated that all visitors who have visited the three pages: home, discussion and call conference pages also visited the forum page. The reason can be to participate in all ongoing discussions taken place on the various pages

Rule 2. [URL8=Y, URL12=Y]: 20438 ==> [URL1=Y, URL3=Y]: 20071 <conf:(0.98)> lift:(6.35) lev:(0.06) conv:(46.95)

This rule stated that 98% of visitors who have visited the discussion page (URL8) and call conference page (URL 12) also visited home (URL 1) and the forum page (URL3). This is due to in case of issues not closed in the discussion page or over the call conference will pass to the forum page from home page.

Group 2: Association rules about business-internet page and Fixed Broad Band Unlimited-ADSL-Fiber page involving home page

The researcher in consultation with the domain experts found that the rules grouped under this grouping interesting with strong relationship. All these rules in common implied that there is strong relationship between the two products: Business Internet and Fixed Broad Band Unlimited ADSL Fiber. And in this relationship Fixed Broad Band Unlimited ADSL Fiber product is the driving force for accessing Business Internet product as majority of the visitors visiting Fixed Broad Band Unlimited ADSL Fiber page first have also visited Business Internet page.

The following are interesting association rules that are generated concerning about the relationship between Fixed Broad Band Unlimited ADSL Fiber page (URL 54) and Business Internet page (URL 18).

Rule 3. [URL1=Y, URL54=Y]: 3914 ==> [URL18=Y]: 3786 <conf:(0.97)> lift:(14.15) lev:(0.01) conv:(28.27)

This rule states that 97% of visitors who visited Fixed Broad Band Unlimited ADSL Fiber page (URL 54) and home page together had also visited Business Internet page (URL 18).

Rule 4. [URL54=Y]: 4746 ==> [URL18=Y]: 4433 <conf:(0.93)> lift:(13.66) lev:(0.01) conv:(14.08)

This rule states that 93% of visitors who visited Fixed Broad Band Unlimited ADSL Fiber page (URL 54) had also visited Business Internet page page (URL 18).

Rule 5. [URL18=Y, URL54=Y]: 4433 ==> [URL1=Y]: 3786 <conf:(0.85)> lift:(0.99) lev:(0) conv:(0.92)

This rule states that 85% of visitors who visited Fixed Broad Band Unlimited ADSL Fiber page (URL 54) and Business Internet page (URL 18) together had also visited home page.

Rule 6. [URL54=Y]: 4746 ==> [URL1=Y, URL18=Y]: 3786 <conf:(0.8)> lift:(13.51) lev:(0.01) conv:(4.65)

This rule states that 80% of visitors who visited Fixed Broad Band Unlimited ADSL Fiber page (URL 54) had also visited Business Internet page (URL 18) and home page together

Group 3: Association rules about pattern existed between the Auxiliary and Home Page

The following rules describe the navigation patterns of users when accessing FAQ, bid and about us pages.

Rule 7. [URL20=Y, URL23=Y]: 3598 ==> [URL1=Y]: 3198 <conf:(0.89)> lift:(1.03) lev:(0) conv:(1.21)

This rule states that 89% of visitors who visited FAQ (URL20) and about us page (URL23) also accessed the home page.

Rule 8. [URL16=Y, URL20=Y]: 3458 ==> [URL1=Y]: 3041 <conf:(0.88)> lift:(1.02) lev:(0) conv:(1.11)

This rule states that 88% of visitors who visited bid (URL16) and FAQ (URL20) also accessed the home page.

Rule 9. [URL16=Y, URL23=Y]: 3641 ==> [URL1=Y]: 3192 <conf:(0.88)> lift:(1.01) lev:(0) conv:(1.09)

This rule states that 88% of visitors who visited bid (URL16) and about us page (URL23) also accessed the home page.

Therefore from the above rules, it can be concluded that visitors usually access home page after visiting the auxiliary pages.

Group 4: Association rules about pattern showing the behavior of users in accessing vacancy page.

Hereunder rules that are related to navigational patterns of users accessing vacancy page.

Rule 10. [URL5=Y, URL20=Y]: 4230 ==> [URL1=Y]: 3699 <conf:(0.87)> lift:(1.01) lev:(0) conv:(1.07)

This rule states that 87% of visitors who visited vacancy page (URL5) and FAQ (URL20) also accessed the home page.

This shows applicants also accessed the FAQ page together with the vacancy page, one reason can be to be prepared for the interview.

Rule 11. [URL5=Y, URL23=Y]: 5093 ==> [URL1=Y]: 4434 <conf:(0.87)> lift:(1.01) lev:(0) conv:(1.04)

This rule states that 87% of visitors who visited vacancy page (URL5) and about us page (URL23) also accessed the home page. The reason for going 'about us' page after checking 'vacancy page' could be to get information about the location for submitting application

Group 5: Association rules about pattern that existed among Amharic pages with home

In this grouping, the rules showing the pattern existed among the Amharic pages are summarized. These rules are listed in decreasing order of their confidence level as follows:

Rule 12. [URL5=Y, URL11=Y]: 3955 ==> [URL9=Y]: 3858 <conf:(0.98)> lift:(6.74) lev:(0.01) conv:(34.52)

This rule dictates that 98% of visitors who visited the vacancy page (URL5) and Amharic advertisement page (URL 11) had also visited Amharic home page (URL 9).

The reason for this navigation is to get more information about the vacancy from the Amharic page.

Rule 13. [URL1=Y, URL11=Y]: 14960 ==> [URL9=Y]: 14558 <conf:(0.97)> lift:(6.73) lev:(0.04) conv:(31.75)

This rule dictates that 97% of visitors who visited home page (URL 1) and Amharic advertisement page (URL 11) had also visited Amharic home page.

This implies that majority of visitors had changed the default home page to the Amharic home page where the advertisement is impeded in it.

Rule 14. [URL11=Y, URL28=Y]: 3733 ==> [URL9=Y]: 3379 <conf:(0.91)> lift:(6.26) lev:(0.01) conv:(8.99)

This rule dictates that 91% of visitors who visited Amharic advertisement page (URL 11) and Amharic vacancy page (28) had also visited Amharic home page (URL 9).

These rule tells us after checking Amharic vacancies, visitors moved to the home page to get more information.

Rule 15. [URL1=Y, URL28=Y]: 5514 ==> [URL9=Y]: 4574 <conf :(0.83)> lift :(5.73) lev:(0.01) conv:(5.01)

This rule dictates that 83% of visitors who visited home page (URL1) and Amharic vacancy page (28) had also visited Amharic home page. The same reason of the above rule can be the cause for such pattern.

Experiment 2: Pattern Discovery and Analysis using modified Dataset

The objective of this case is to discover the pattern existed among the top level pages so as to show the behaviors of the users towards the main products. Summary of experimental result on the modified dataset where the child pages are merged as one parent category is given. The top level pages of the website are named by the five main products of the company where the product catalogues are listed as child pages. In this section, the pattern existed among these sites is discussed with selected rules.

To extract association rules in modified dataset, a minimum Support = 0.01 (1%) and Minimum Confidence = 0.75 (75%) is used. These threshold values are selected because more interesting association rules are generated at these support and confidence level after trying several experiments with plus or minus of these threshold values.

In this case a total of 50 rules are generated, the researcher with web master consultancy have Selected 8 rules and its analysis is given below by grouping related rules together whenever appropriate for the discussion. For detail Weka Output, see Appendix C

The top level navigation contents for each main product type are listed in the table 4-7 below. The table shows the top six pages visited by users.

NO	URLs	Page name	Number of visits	Visitors in percent
1	/	Home page	260982	86%
2	/?q=internet	internet	249320	82%
3	/?q=business	Business	256751	81%
4	/?q=vas	Value added service	246568	81%
5	/?q=mobile	mobile	155746	52%
6	/?q=landline	Fixed line	6383	2%

Table 4-7 Top accessed contents

Group 6: The pattern discovered over Business page (URL 26) and Internet page (URL33)

A common pattern discovered from the rules containing the business page (URL 26) and internet page (URL 33) is the driving factor for the others and vice-versa. In some rules all visitors who visited the business page had also visited the internet page and in other rules the vice-versa occurred. The implication of this pattern shows the existence of strong relationship between business pages and internet pages as visitors are going from one page to the other

Sample rules for the same are analyzed as below:

Rule 1. [URL33=y, URL1=Y]: 260964 ==> [URL26=y]: 260964 <conf:(1)> lift:(1) lev:(0) conv:(23.36)

This means that all visitors who have visited internet page (URL 33) and home, had also visited business page (URL 26)

Rule 2. [URL26=y]: 301552 ==> [URL33=y]: 301547 <conf:(1)> lift:(1) lev:(0) conv:(5.33)

This means that all visitors who have visited business page (URL 26), had also visited internet page (URL 33).

Rule 3. [URL26=y, URL1=Y]: 260969 ==> [URL33=y]: 260964 <conf:(1)> lift:(1) lev:(0) conv:(4.62)

This means that all visitors who have visited business page (URL 26) and home, had also visited internet page (URL 33).

Group 7: The pattern discovered containing troubleshoot pages (URL2)

This grouping is made based on the pattern observed among the rules containing troubleshooting page (URL2).

As can be seen from the sample rules listed below, one pattern that can be drown dictates, there is high navigational flow by visitors who first visited the trouble shoot pages with home page or internet page have also visited business pages. There is also high navigational flow by visitors who first visited the trouble shoot pages with home page or business have also visited internet page.

Rule 4 . [URL33=y, URL2=Y]: 62924 ==> [URL26=y]: 62924 <conf:(1)> lift:(1) lev:(0) conv:(5.63)

This rule states that from 100% visitors visited internet page (URL 33) and trouble shoot page (URL2) had also visited business page (URL 26).

Rule 5. [URL26=y, URL2=Y]: 62926 ==> [URL33=y]: 62924 <conf:(1)> lift:(1) lev:(0) conv:(2.23)

This rule states that from 100% visitors visited business page (URL 26) and trouble shoot page (URL2) had also visited internet page (URL 33).

Rule 6. [URL26=y, URL1=Y, URL2=Y]: 61600 ==> [URL33=y]: 61598 <conf:(1)> lift:(1) lev:(0) conv:(2.18)

This rule states that from 100% visitors visited business page (URL 26), home page and trouble shoot page (URL2) had also visited internet page (URL 33) .

Rule 7. [URL2=Y]: 62935 ==> [URL33=y]: 62924 <conf:(1)> lift:(1) lev:(0) conv:(0.56)

This rule states that from 100% visitors visited trouble shoot page (URL2) had also visited internet page (URL 33)

Rule 8. [URL2=Y]: 62935 ==> [URL26=y, URL33=y]: 62924 <conf:(1)> lift:(1) lev:(0) conv:(0.56)

This rule states that from 100% visitors visited trouble shoot page (URL2) had also visited internet page (URL 33) and business page (URL26)

4.3. Summary of the Findings

Generally from the findings extracted from the statistical analysis the major ones are: majority of the visitors are local, the number of new visitors exceeds the number of returning visitors, most popular referrals like google or yahoo are missing from referral list, home page is the first top accessed page, top landing page and also the second top listed exit pages.

In addition ,the findings interpreted from the generated association rules include discussion pages are most frequently accessed together ,93% of visitors who visited business internet do so after visiting Fixed Broad Band Unlimited ADSL Fiber page, more than 88% of visitors who visited home page after the bid and vacancy pages, there existed a strong relationship between internet and business pages as one is accessed after the second and vice-versa and the last is those visitors who had visited the internet page and business page also visited troubleshooting page .

As a summary, the findings explored strengths of the website like the top traffic source of the website is the direct channels indicating the domain name of the web site is familiar to the users. On the other hand, weakness of the web site like majority of the visitors are new, popular search engines are missed from website referrals, visitors are leaving from the landed pages and discussion pages like forum page, general discussion page and call conference page are co-accessed are revealed from these findings.

In relating with the reviewed related works done so far, the findings of this study disagrees with the findings of the researches conducted following Apriori as an algorithm but agrees with those researches conducted using FP growth as an algorithm.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

Recently Web usage mining has been gaining a lot of attention because of its potential commercial benefits. The ability to observe potential users as they browse through a virtual store promises to raise business intelligence to a new level. While demographic data can always be added to the web usage mining process, this thesis has shown an interesting pattern can be discovered with just click stream data. On top of the pattern discovery made based on users access log of Ethio telecom website, a statistical analysis also conducted on six months Google analytics repository of the website.

The log files were thoroughly preprocessed in order to filter out irrelevant data which may deceive 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 in order to discover interesting patterns in the website. Accordingly useful results were found.

As shown in the statistical analysis part, the location report revealed that majority of the traffics is originated locally. This indicates that international services are not matured enough to attract visitors across different countries worldwide. In addition to this, the overview of audience report also shown that majority of the visitors, more than 60% are new users which implies the web site is not potential enough to retain the existing visitors

The statistical analysis also revealed that, the first top landing page is the home page which accounts for over 40% of the landing pages and the top traffic source channel is the direct channel. This shows that most users first access the home page by directly typing the URL address, namely, www.ethiotelcom.et which further implies that other pages of the website are not optimized for search engines and the website is not popular enough to be referred by other sites.

The home page is also one of the top exit pages. This could mean, two things; either the users have satisfied their information need from this page alone or may be moving from page to page is not user friendly.
Another finding from the statistical analysis revealed also Google or yahoo, which are the most popular search engines, are missed from the top referrals list. This clearly indicates that the website requires search engine optimization and further advertising.

The access trend analysis has shown that some events like introducing new product are triggering more access usage of the website.

Association Rule Discovery with aggregate dataset revealed that a strong relationship existed among the discussion sites and also between business internet page and fixed broad band unlimited ADSL Fiber page. This indicates that these pages are giving similar purpose to the users in the case of the discussion sites or the existence of some similarity between the two products (business internet page and fixed Broad Band Unlimited ADSL Fiber).

The pattern mined from the aggregate dataset explored that most of the times 'about us' page and 'home' page are accessed next to Bid or vacancy pages which implies users need more information about the company to take next action of whatever is announced on vacancy or bid pages.

And more the pattern discovered from the aggregate data indicates trouble shooting page is mostly accessed with internet or business pages indicating users are looking for setup procedure about these product categories i.e. business and internet.

The pattern discovered from the modified dataset confirms a strong relationship existed between business and internet pages at high level implying there is something common among these products.

This study has attempted to extract a number of interesting user navigational pattern that is important for Ethio telecom to improve its web site. However the pattern discovery task for the current research is limited to successful server log for the year 2014 which enable us to conduct the generic analysis of the users pattern discovery where the client side logs, proxy side logs and error logs are not considered where those logs contains more detailed information that support the user identification task.

Moreover the current study have analyzed the web usage pattern of the association rules without considering the sequential discovery hence the association rules discovered are far too many in number and with low comprehensibility.

5.2. Recommendations

Based on the findings of this study, the following recommendations are forwarded as a future research direction and Website improvement:

5.2.1. Future Work

- In this research, the Web Server log is used to discover patterns. 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.
- In this research, sequential pattern discovery was not considered. Hence, future researches should consider sequential pattern discovery in order to clearly show the time sequences of access; that is, the order in which the pages are accessed.
- Integrating the Server log files with the data warehouse for Online Analytical Mining can help the organization to track its customers' usage behavior on real-time basis. Hence, further research should be carried out in this area.
- Identifying users by way of log servers, proxies and client cookies is one of the potential research areas. Hence, future researches can consider this as a very useful application.
- In this research, the web access pattern is discovered with unified access log data without segregating across the regions. Hence, future researcher can study the regional segregations and its implications.

5.2.1. Website Improvement

- Even though the majority of the customers are local, efforts should be exerted to improve the usage of the website outside the country. This can be achieved through website advertisement and organizing and participating in different partnership campaigns.

- Assessment is required to discover the relationship between the two products internet and business specifically business internet page and Fixed Broad Band Unlimited ADSL Fiber page.
- The company should design a strategy on ways of improving the web site popularity such as reconstructing the website, creating a link to Social Medias, attending and sponsoring website conferences.
- As it can be seen from the top referrals section, even the top search engine Google is not in the list. It is required to optimize the important pages for search engines.
- The top browsers report revealed that majority of the visitors used their mobile hand set, implying a mobile version of the website shall be designed so that it can be easily and conveniently accessed by smart phones.
- Web restructuring is crucial in the Ethio telecom website. The pages in each category should be arranged in the order of their importance (frequency of access) for ease of access.
- From the general observation of the website, there are some pages like self-care with no contents posted, so the website should be reviewed as to identifying those blank pages and making them resourceful.
- With the vast complicated telecom services available in the technology, further study should be conducted to enhance the web site services by adding more web services like self-care provisioning, ecommerce and trouble ticket tracking functionalities on the web site.

REFERENCES,

[1] Daniel T. Larose, *Discovering knowledge in data: An Introduction to Data Mining*, USA: A John Wiley & Sons, INC, publication, 2005.

[2] Bing Liu, *Web data mining: Exploring Hyperlinks, Contents, and usage data*, German: Springer-Verlag Berlin Heidelberg, 2007.

[3] Zdravko Markov & Daniel T. Larose, *Data Mining The Web: Uncovering patterns in web Content, Structure and Usage*, USA: A John Wiley & Sons, INC, publication, 2007.

[4] Naga Lakshmi, Raja Sekhara Rao & Sai Satyanarayana Redd. "*An Overview of Preprocessing on Web Log Data for Web Usage Analysis*." International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, 2013.

[5] Robert Cooley, BamshadMobasher and Jaideep Srivastava ,*Data Preparation for Mining World Wide Web Browsing patterns*, German: Springer-Verlag, *1999.*

[6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, *from Data Mining to Knowledge discovery in databases*,American Association for Artificial Intelligence, 1996.

[7] Ethio telecom web master, *New ethiotelecom website project report*, 2014.

[8] Sanjay BapuThakare& Prof. Sangram Z. Gawali. "*Effective and Complete Preprocessing for Web Usage Mining."* International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 03, 848-851, 2010.

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

[10] Shaily Langhnoja, Mehul Barot and Darshak Mehta. "*Pre-Processing: Procedure on Web Log File for Web Usage Mining.*" International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 12, Gandhinagar, India, 2012 .

[11]Vijayashri Losarwar& Dr. Madhuri Joshi, *Data Preprocessing in Web Usage Mining*, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) , Singapor, 2012.

[12]C.P.Sumathi, R.Padmajavalli and T.Santhanam. "An *overview of preprocessing of weblog files for web usage mining*." Journal of Theoretical and Applied Information Technology. Vol. 34 No.2: JATIT & LLS , 31st December 2011.

[13]Harmit kaurand Hardeepsingh *A. "Survey of Preprocessing Method for Web Usage Mining Process."* International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 2– Mar 2014.

[14]Shaily Langhnoja, Mehul Barot&Darshak Mehta. *"Pre-Processing: Procedure on Web Log FileforWeb Usage Mining."* ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 2, Issue 12, December 2012(IJETAE).

[15]Awet Fesseha. "*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.

[16] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques*, Third Edition, Elsevier.

[17] Gagandeep Kaur and Shruti Aggarwal. *Performance Analysis of Association Rule Mining Algorithms*.‖ Research Paper, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India, 2013.

[18] Jiawei Han and MichelineKamber, *Data Mining: Concepts and Techniques*, Second Edition, Elsevier.

[19] Rahul Mishra and AbhaChoubey. "*Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining*." International Journal of Advanced Research in Computer Science and Software Engineering**,** Volume 2, Issue 9, Computer Science & CSVTU, India, 2012 .

[20] Federico Michele Facca and Pier Luca Lanzi, *Recent Developments in Web Usage Mining*, Research.‖ Journal Article, Artificial Intelligence and Robotics Laboratory Dipartimento di Elettronica e InformazionePolitecnico di Milano, Italy,2007.

[21] Anand Sharmal. "Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT web Data", MS Project, Rochester Intstitute of Technology, Rochester, NY, USA, 2008.

.

[22] N. M. Abo El-Yazeed. *An overview of preprocessing of web log files for web usage mining*,‖ Demonstrator at High Institute for Management and Computer, Port Said University, Port Said, Egypt,2011.

[23] Sanjay Bapu Thakare and Sangrarn and Z.Gawali. "*An Effective and Complete Preprocessing for Web Usage Mining*." International Journal on Computer Science and Engineering, Volume 2, Issue 3, Pune-43, Maharashtra, India, 2010.

[24] KeYiping. *A Survey on Preprocessing Techniques in Web Usage Mining*.‖ Computer Science Department, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, 2003 .

[25] C.P. Sumathi, R. Padmajavalli and T.Santhanam. " *An overview of preprocessing o f web log files for web usage mining*." Journal of Theoretical and Applied Information Technology Volume 34, Issue.2, Department of Computer Science, Tamil Nadu, India, 2011.

[26] SerkaddisAdem. "*Web Usage Pattern Discovery: The Case of Ethiopian Airlines website"* MSc Thesis, Adama University, Adama, Ethiopia, 201.

[27] NareshBarsagade. **"***Web Usage Mining and Pattern Discovery*." A Survey Paper, 2003.

[28] Zhou Baoyao. "*Intelligent Web Usage Mining*, PhD Admission Requirement Report." Nanyang Technological University, Division of Information Systems, School of Computer Engineering, Singapore, 2004.

[29] RenátaIváncsy and SándorJuhász. "*Analysis of Web User Identification Methods."* International Journal of Electrical and Computer Engineering.‖ Volume 2, Issue 3, 2007.

[30] Peter I. Hofgesang." *Web Usage Mining Structuring semantically enriched click stream data."* MSC Thesis, VrijeUniversiteit Amsterdam, Netherlands, 2004.

[31] Getahun Nigatu. "*Web Usage pattern discover and analysis by region: The Case of Ethiopian Airlines website*." MSc Thesis, Addis Ababa University, 2013.

[32] Dharmendra Patel, Dr. Kalpesh Parikh and Atul Patel , *Sessionization :A Vital Stage in Data Preprocessing of Web Usage Mining-A Survey*, IJERA, 2012.

[33] Robert Cooley, BamshadMobasher and Jaideep Srivastava ,*Data Preparation for Mining World Wide Web Browsing patterns*,German: Springer-Verlag, *1999.*

[34] Enigma Software Group USA, LLC, 2013, http://malwaretips.com/blogs/search-tb-ask-virus. [Accessed May 27, 2015]

[35] Worku Bogale. "*A background paper on Telecom and Telecom statistics in Ethiopia*." Addis Ababa, Ethiopia 2005.

[36] Malwaretips: your security advisor, 2013, http: //malwaretips.com/blogs/search-tb-ask-virus. [Accessed May 27, 2015.]

APPENDICES

Appendix A: List of common URLsused for Statistical Analysis

URL	Description
/	Root directory or home page
/?q=vacancy	Vacancy page
/?q=Manage3GInternet	mobile internet tutorial page
/?q=internetmobileinternet	Internet mobile internet page
/?q=bid	Bid page
/?q=internetadsl	Internet ADSL page
/?q=businessinternet	Business Internet page
/?q=businessmobile	Business mobile page
/?q=mobileprepaid	Mobile prepaid page
/?q=internetevdopackages	Internet EVDO packages page
/?q=aboutus	About us page
/?q=mobilepostpaid	Mobile postpaid page
/?q=faq	FAQ page
/?q=internetcdma1x	Internet CDMA 1x page
/?q=mobileinternational	Mobile International page
/?q=customercare	Customer care page
/?q=postpaidpackage	Postpaid package
/?q=troubleshoot	Troubleshoot page
/?q=callcenter	Call center page
/?q=gprspackage	GPRS package page
/?q=unlimitedadsl	Unlimited ADSL page
/?q=internetispservices	Internet ISP services page
/?q=voicepackage	Voice package page
/?q=businessvpndata	Business VPN data page

/?q=voicemail	Voice mail page
/?q=limitedadsl	Limited ADSL page
/?q=contactusshopsaddress	Contact us shops address page

Appendix B: List of selected URLs for Association Rule Discovery

URL code	URL address
URL1	/
URL2	/?q=troubleshoot
URL3	/?q=forum
URL4	/?url.access.method=GET&url=http%3A%2F%2Fwww.ethionet.et%2F
URL5	/?q=vacancy
URL6	/?q=taxonomy/term/1/feed
URL7	/?q=Manage3GInternet
URL8	/?q=taxonomy/term/1
URL9	/etamh/
URL10	/?q=filter/tips
URL11	/etamh/?q=simpleads/load/2/1
URL12	/?q=conferencecall
URL13	/?q=rss.xml
URL14	/?q=internet-3gpackages
URL15	/?q=node/
URL16	/?q=bid
URL17	/?q=forum/1
URL18	/?q=business-internet
URL19	/?q=search/node/Search
URL20	/?q=faq
URL21	/?q=mobile-prepaid
URL22	/etamh/?q=rss.xml
URL23	/?q=aboutus
URL24	/?q=internet-adsl
URL25	/?q=internet-evdopackages
URL26	/?q=business-mobile
URL27	/vacancy/index.php
URL28	/etamh/?q=vacancy
URL29	/?q=user/login&destination=node/158%23comment-form
URL30	/index.php?s=

URL31	/?q=taxonomy/term/4
URL32	/?q=sitemap
URL33	/?q=internet-cdma1x
URL34	/?q=mobile-international
URL35	/?q=mobile-postpaid
URL36	/?q=news
URL37	/?q=contactus-callcenter994
URL38	/?q=customercare
URL39	/?q=contactus-shopsaddress
URL40	/?q=business-internationalconnectivity
URL41	/?q=internet-ispservices
URL42	/etamh/?q=node/143
URL43	/?q=callcenter
URL44	/?q=contactus-lehulukifyacenters
URL45	/?q=business-vpndata
URL46	/etamh/?q=internet-3gpackages
URL47	/?q=landline-postpaid
URL48	/?q=mobile-roaming
URL49	/?q=voicepackage
URL50	/?q=productnews
URL51	/?q=vas
URL52	/?q=contactus-hq
URL53	/?q=contactus-regions
URL54	/?q=fixedbroadbandunlimited-adsl-fiber
URL55	/?q=gprspackage
URL56	/etamh/?q=Manage3GInternet
URL57	/?q=dataroaming
URL58	/?q=business-m2mbusinesssolution
URL59	/?q=huaweievdocdma
URL60	/?q=taxonomy/term/6
URL61	/?q=customercare-faq
URL62	/etamh/?q=business-internet
URL63	/?q=businessmobilewithoutcug
URL64	/etamh/?q=node/24
URL65	/?q=pressrelease
URL66	/?q=postpaidpackage
URL67	/etamh/?q=internet-mobileinternet
URL68	/?q=business-vsat
URL69	/?q=landline-prepaid
URL70	/etamh/?q=bid

Appendix C: Weka Association Rule Discovery sample output

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 100 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.01

Relation: finalds combined

Instances: 301579

Attributes: 60

URL1- URL70

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

FPGrowth found 96 rules (displaying top 96)

1. [URL1=Y, URL8=Y, URL12=Y]: 20077 ==> [URL3=Y]: 20071 <conf:(1)> lift:(5.19) lev:(0.05) conv:(2315.49)

2. [URL8=Y, URL12=Y]: 20438 ==> [URL3=Y]: 20417 <conf:(1)> lift:(5.18) lev:(0.05) conv:(749.99)

3. [URL1=Y, URL3=Y, URL6=Y]: 24370 ==> [URL8=Y]: 24343 <conf:(1)> lift:(5.34) lev:(0.07) conv:(707.69)

4. [URL1=Y, URL8=Y, URL6=Y]: 24374 ==> [URL3=Y]: 24343 <conf:(1)> lift:(5.18) lev:(0.07) conv:(614.92)

5. [URL3=Y, URL6=Y]: 34592 ==> [URL8=Y]: 34509 <conf:(1)> lift:(5.34) lev:(0.09) conv:(334.84)

6. [URL1=Y, URL6=Y]: 24440 ==> [URL8=Y]: 24374 <conf:(1)> lift:(5.34) lev:(0.07) conv:(296.6)

7. [URL1=Y, URL6=Y]: 24440 ==> [URL3=Y]: 24370 <conf:(1)> lift:(5.17) lev:(0.07) conv:(277.9)

8. [URL8=Y, URL6=Y]: 34635 ==> [URL3=Y]: 34509 <conf:(1)> lift:(5.17) lev:(0.09) conv:(220.17)

9. [URL1=Y, URL6=Y]: 24440 ==> [URL3=Y, URL8=Y]: 24343 <conf:(1)> lift:(5.43) lev:(0.07) conv:(203.66)

10. [URL6=Y]: 34808 ==> [URL8=Y]: 34635 <conf:(1)> lift:(5.32) lev:(0.09) conv:(162.66)

11. [URL6=Y]: 34808 ==> [URL3=Y]: 34592 <conf:(0.99)> lift:(5.16) lev:(0.09) conv:(129.5)

12. [URL6=Y]: 34808 ==> [URL3=Y, URL8=Y]: 34509 <conf:(0.99)> lift:(5.41) lev:(0.09) conv:(94.75)

13. [URL1=Y, URL3=Y, URL12=Y]: 20265 ==> [URL8=Y]: 20071 <conf:(0.99)> lift:(5.3) lev:(0.05) conv:(84.5)

14. [URL3=Y, URL12=Y]: 20645 ==> [URL8=Y]: 20417 <conf:(0.99)> lift:(5.29) lev:(0.05) conv:(73.3)

15. [URL3=Y, URL8=Y, URL12=Y]: 20417 ==> [URL1=Y]: 20071 <conf:(0.98)> lift:(1.14) lev:(0.01) conv:(7.92)

16. [URL8=Y, URL12=Y]: 20438 ==> [URL1=Y]: 20077 <conf:(0.98)> lift:(1.14) lev:(0.01) conv:(7.6)

17. [URL8=Y, URL12=Y]: 20438 ==> [URL1=Y, URL3=Y]: 20071 <conf:(0.98)> lift:(6.35) lev:(0.06) conv:(46.95)

18. [URL3=Y, URL12=Y]: 20645 ==> [URL1=Y]: 20265 <conf:(0.98)> lift:(1.13) lev:(0.01) conv:(7.29)

19. [URL1=Y, URL8=Y]: 45507 ==> [URL3=Y]: 44661 <conf:(0.98)> lift:(5.09) lev:(0.12) conv:(43.37)

20. [URL8=Y]: 56364 ==> [URL3=Y]: 55297 <conf:(0.98)> lift:(5.09) lev:(0.15) conv:(42.61)

21. [URL2=Y]: 62935 ==> [URL1=Y]: 61609 <conf:(0.98)> lift:(1.13) lev:(0.02) conv:(6.38)

22. [URL5=Y, URL11=Y]: 3955 ==> [URL9=Y]: 3858 <conf:(0.98)> lift:(6.74) lev:(0.01) conv:(34.52)

23. [URL1=Y, URL11=Y]: 14960 ==> [URL9=Y]: 14558 <conf:(0.97)> lift:(6.73) lev:(0.04) conv:(31.75)

24. [URL3=Y, URL12=Y]: 20645 ==> [URL1=Y, URL8=Y]: 20071 <conf:(0.97)> lift:(6.44) lev:(0.06) conv:(30.49)

25. [URL1=Y, URL54=Y]: 3914 ==> [URL18=Y]: 3786 <conf:(0.97)> lift:(14.15) lev:(0.01) conv:(28.27)

26. [URL12=Y]: 30165 ==> [URL1=Y]: 29149 <conf:(0.97)> lift:(1.12) lev:(0.01) conv:(3.99)

27. [URL1=Y, URL3=Y]: 46628 ==> [URL8=Y]: 44661 <conf:(0.96)> lift:(5.12) lev:(0.12) conv:(19.26)

28. [URL3=Y]: 58110 ==> [URL8=Y]: 55297 <conf:(0.95)> lift:(5.09) lev:(0.15) conv:(16.79)

29. [URL54=Y]: 4746 ==> [URL18=Y]: 4433 <conf:(0.93)> lift:(13.66) lev:(0.01) conv:(14.08)

30. [URL11=Y]: 22159 ==> [URL9=Y]: 20516 <conf:(0.93)> lift:(6.4) lev:(0.06) conv:(11.53)

31. [URL50=Y]: 6178 ==> [URL1=Y]: 5607 <conf:(0.91)> lift:(1.05) lev:(0) conv:(1.45)

32. [URL11=Y, URL28=Y]: 3733 ==> [URL9=Y]: 3379 <conf:(0.91)> lift:(6.26) lev:(0.01) conv:(8.99)

33. [URL19=Y]: 12809 ==> [URL1=Y]: 11503 <conf:(0.9)> lift:(1.04) lev:(0) conv:(1.32)

34. [URL63=Y]: 3737 ==> [URL1=Y]: 3348 <conf:(0.9)> lift:(1.04) lev:(0) conv:(1.29)

35. [URL20=Y, URL23=Y]: 3598 ==> [URL1=Y]: 3198 <conf:(0.89)> lift:(1.03) lev:(0) conv:(1.21)

36. [URL16=Y]: 17254 ==> [URL1=Y]: 15331 <conf:(0.89)> lift:(1.03) lev:(0) conv:(1.21)

37. [URL65=Y]: 3952 ==> [URL1=Y]: 3511 <conf:(0.89)> lift:(1.03) lev:(0) conv:(1.2)

38. [URL49=Y]: 6153 ==> [URL1=Y]: 5421 <conf:(0.88)> lift:(1.02) lev:(0) conv:(1.13)

39. [URL23=Y]: 12110 ==> [URL1=Y]: 10669 <conf:(0.88)> lift:(1.02) lev:(0) conv:(1.13)

40. [URL16=Y, URL20=Y]: 3458 ==> [URL1=Y]: 3041 <conf:(0.88)> lift:(1.02) lev:(0) conv:(1.11)

41. [URL9=Y, URL20=Y]: 3825 ==> [URL1=Y]: 3358 <conf:(0.88)> lift:(1.01) lev:(0) conv:(1.1)

42. [URL5=Y, URL32=Y]: 3722 ==> [URL1=Y]: 3264 <conf:(0.88)> lift:(1.01) lev:(0) conv:(1.09)

43. [URL16=Y, URL23=Y]: 3641 ==> [URL1=Y]: 3192 <conf:(0.88)> lift:(1.01) lev:(0) conv:(1.09)

44. [URL5=Y, URL20=Y]: 4230 ==> [URL1=Y]: 3699 <conf:(0.87)> lift:(1.01) lev:(0) conv:(1.07)

45. [URL5=Y, URL23=Y]: 5093 ==> [URL1=Y]: 4434 <conf:(0.87)> lift:(1.01) lev:(0) conv:(1.04)

46. [URL66=Y]: 4070 ==> [URL1=Y]: 3516 <conf:(0.86)> lift:(1) lev:(0) conv:(0.99)

47. [URL18=Y]: 20623 ==> [URL1=Y]: 17802 <conf:(0.86)> lift:(1) lev:(0) conv:(0.98)

48. [URL32=Y]: 7691 ==> [URL1=Y]: 6638 <conf:(0.86)> lift:(1) lev:(0) conv:(0.98)

49. [URL57=Y]: 4260 ==> [URL1=Y]: 3647 <conf:(0.86)> lift:(0.99) lev:(0) conv:(0.93)

50. [URL18=Y, URL54=Y]: 4433 ==> [URL1=Y]: 3786 <conf:(0.85)> lift:(0.99) lev:(0) conv:(0.92)

51. [URL5=Y]: 49820 ==> [URL1=Y]: 42540 <conf:(0.85)> lift:(0.99) lev:(0) conv:(0.92)

52. [URL38=Y]: 7885 ==> [URL1=Y]: 6705 <conf:(0.85)> lift:(0.98) lev:(0) conv:(0.9)

53. [URL51=Y]: 6057 ==> [URL1=Y]: 5132 <conf:(0.85)> lift:(0.98) lev:(0) conv:(0.88)

54. [URL26=Y]: 12969 ==> [URL1=Y]: 10980 <conf:(0.85)> lift:(0.98) lev:(0) conv:(0.88)

55. [URL36=Y]: 6407 ==> [URL1=Y]: 5413 <conf:(0.84)> lift:(0.98) lev:(0) conv:(0.87)

56. [URL18=Y, URL24=Y]: 3594 ==> [URL1=Y]: 3030 <conf:(0.84)> lift:(0.97) lev:(0) conv:(0.86)

57. [URL40=Y]: 6997 ==> [URL1=Y]: 5893 <conf:(0.84)> lift:(0.97) lev:(0) conv:(0.85)

58. [URL52=Y]: 4340 ==> [URL1=Y]: 3643 <conf:(0.84)> lift:(0.97) lev:(0) conv:(0.84)

59. [URL20=Y]: 12583 ==> [URL1=Y]: 10557 <conf:(0.84)> lift:(0.97) lev:(0) conv:(0.84)

60. [URL5=Y, URL16=Y]: 5929 ==> [URL1=Y]: 4936 <conf:(0.83)> lift:(0.96) lev:(0) conv:(0.8)

61. [URL18=Y, URL14=Y]: 4744 ==> [URL1=Y]: 3940 <conf:(0.83)> lift:(0.96) lev:(0) conv:(0.79)

62. [URL58=Y]: 4734 ==> [URL1=Y]: 3928 <conf:(0.83)> lift:(0.96) lev:(0) conv:(0.79)

63. [URL1=Y, URL28=Y]: 5514 ==> [URL9=Y]: 4574 <conf:(0.83)> lift:(5.73) lev:(0.01) conv:(5.01)

64. [URL54=Y]: 4746 ==> [URL1=Y]: 3914 <conf:(0.82)> lift:(0.95) lev:(0) conv:(0.77)

65. [URL34=Y]: 7085 ==> [URL1=Y]: 5793 <conf:(0.82)> lift:(0.94) lev:(0) conv:(0.74)

66. [URL53=Y]: 4083 ==> [URL1=Y]: 3338 <conf:(0.82)> lift:(0.94) lev:(0) conv:(0.74)

67. [URL39=Y]: 5091 ==> [URL1=Y]: 4161 <conf:(0.82)> lift:(0.94) lev:(0) conv:(0.74)

68. [URL47=Y]: 4488 ==> [URL1=Y]: 3646 <conf:(0.81)> lift:(0.94) lev:(0) conv:(0.72)

69. [URL44=Y]: 4601 ==> [URL1=Y]: 3731 <conf:(0.81)> lift:(0.94) lev:(0) conv:(0.71)

70. [URL9=Y]: 43635 ==> [URL1=Y]: 35298 <conf:(0.81)> lift:(0.93) lev:(-0.01) conv:(0.7)

71. [URL45=Y]: 5843 ==> [URL1=Y]: 4726 <conf:(0.81)> lift:(0.93) lev:(0) conv:(0.7)

72. [URL3=Y, URL8=Y]: 55297 ==> [URL1=Y]: 44661 <conf:(0.81)> lift:(0.93) lev:(-0.01) conv:(0.7)

73. [URL8=Y]: 56364 ==> [URL1=Y]: 45507 <conf:(0.81)> lift:(0.93) lev:(-0.01) conv:(0.7)

74. [URL41=Y]: 6946 ==> [URL1=Y]: 5586 <conf:(0.8)> lift:(0.93) lev:(0) conv:(0.69)

75. [URL37=Y]: 5990 ==> [URL1=Y]: 4811 <conf:(0.8)> lift:(0.93) lev:(0) conv:(0.68)

76. [URL3=Y]: 58110 ==> [URL1=Y]: 46628 <conf:(0.8)> lift:(0.93) lev:(-0.01) conv:(0.68)

77. [URL21=Y]: 12503 ==> [URL1=Y]: 10004 <conf:(0.8)> lift:(0.92) lev:(0) conv:(0.67)

78. [URL24=Y]: 11818 ==> [URL1=Y]: 9442 <conf:(0.8)> lift:(0.92) lev:(0) conv:(0.67)

79. [URL54=Y]: 4746 ==> [URL1=Y, URL18=Y]: 3786 <conf:(0.8)> lift:(13.51) lev:(0.01) conv:(4.65)

80. [URL48=Y]: 5283 ==> [URL1=Y]: 4212 <conf:(0.8)> lift:(0.92) lev:(0) conv:(0.66)

81. [URL35=Y]: 6771 ==> [URL1=Y]: 5391 <conf:(0.8)> lift:(0.92) lev:(0) conv:(0.66)

82. [URL8=Y]: 56364 ==> [URL1=Y, URL3=Y]: 44661 <conf:(0.79)> lift:(5.12) lev:(0.12) conv:(4.07)

83. [URL14=Y, URL21=Y]: 3884 ==> [URL1=Y]: 3065 <conf:(0.79)> lift:(0.91) lev:(0) conv:(0.64)

84. [URL14=Y, URL33=Y]: 3993 ==> [URL1=Y]: 3133 <conf:(0.78)> lift:(0.91) lev:(0) conv:(0.62)

85. [URL43=Y]: 4790 ==> [URL1=Y]: 3744 <conf:(0.78)> lift:(0.9) lev:(0) conv:(0.62)

86. [URL24=Y, URL25=Y]: 4083 ==> [URL1=Y]: 3188 <conf:(0.78)> lift:(0.9) lev:(0) conv:(0.61)

87. [URL25=Y, URL33=Y]: 3967 ==> [URL1=Y]: 3091 <conf:(0.78)> lift:(0.9) lev:(0) conv:(0.61)

88. [URL61=Y]: 4078 ==> [URL1=Y]: 3175 <conf:(0.78)> lift:(0.9) lev:(0) conv:(0.61)

89. [URL25=Y]: 11151 ==> [URL1=Y]: 8678 <conf:(0.78)> lift:(0.9) lev:(0) conv:(0.61)

90. [URL14=Y, URL24=Y]: 5097 ==> [URL1=Y]: 3938 <conf:(0.77)> lift:(0.89) lev:(0) conv:(0.59)

91. [URL14=Y, URL25=Y]: 5797 ==> [URL1=Y]: 4468 <conf:(0.77)> lift:(0.89) lev:(0) conv:(0.59)

92. [URL14=Y]: 19047 ==> [URL1=Y]: 14660 <conf:(0.77)> lift:(0.89) lev:(-0.01) conv:(0.58)

93. [URL3=Y]: 58110 ==> [URL1=Y, URL8=Y]: 44661 <conf:(0.77)> lift:(5.09) lev:(0.12) conv:(3.67)

94. [URL33=Y]: 8063 ==> [URL1=Y]: 6193 <conf:(0.77)> lift:(0.89) lev:(0) conv:(0.58)

95. [URL9=Y, URL28=Y]: 6077 ==> [URL1=Y]: 4574 <conf:(0.75)> lift:(0.87) lev:(0) conv:(0.54)

96. [URL55=Y]: 4236 ==> [URL1=Y]: 3188 <conf:(0.75)> lift:(0.87) lev:(0) conv:(0.54)

Appendix D: Java source Codes

i. Java Source Code for Feature Selection

```
// This program is used to identify Transactions from preprocessed files
// which is in the form of array of Transaction ID by URL


import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.FileReader;
import java.io.FileWriter;
import java.util.StringTokenizer;
import java.io.*;

/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.
 */
/**
 *
 * @author senait.mezgebu
 */
public class Final {
   public static void main(String[] args)
{
try {
// The foloowing lines of codes open and read the top accessed URLs which
// was already determined by heuristics and assigns them to an array
FileReaderfileReader = new FileReader("C:\\Users\\senait.mezgebu\\Desktop\\my
Thesis\\CH-3 Preprocessing\\prep tools\\java\\Toppath1.txt");
BufferedReaderbr = new BufferedReader(fileReader);
String URLName[]=new String[10000];
String line= br.readLine();
inti=0;
while (line!=null)
{
URLName[i++]=line;
line= br.readLine();
}
```

```
br.close();
// The following lines open sorted preprocessed file of each region for input
FileReader fileReader1 = new FileReader("C:\\Users\\senait.mezgebu\\Desktop\\my
Thesis\\CH-3 Preprocessing\\prep tools\\java\\GETT_ALL 2015-04-06.txt");
BufferedReader br1 = new BufferedReader(fileReader1);
// The following lines open regional file that is transformed to transaction for output
FileWriterfileWriter = new FileWriter("C:\\Users\\senait.mezgebu\\Desktop\\my
Thesis\\CH-3 Preprocessing\\prep tools\\java\\finaloutwzid.txt");
BufferedWriterbw = new BufferedWriter(fileWriter);
String line1=br1.readLine();
StringTokenizer st0= new StringTokenizer(line1);

int n=0;
String Sess0=null;
// Just read a sinle record and assign initial Session ID to Sess0
while (st0.hasMoreTokens())
{
if(n==0)
{
Sess0=st0.nextToken();
//System.out.print("asdasdasd: "+Sess0 +"\n");
br1.close();
break;
}
st0.nextToken();
n++;
}
// Open the sorted regional file again to transform the given file into Transaction
FileReader fileReader2 = new FileReader("C:\\Users\\senait.mezgebu\\Desktop\\my
Thesis\\CH-3 Preprocessing\\prep tools\\java\\GETT_ALL 2015-04-06.txt");
BufferedReader br2 = new BufferedReader(fileReader2);
String line2=br2.readLine();
// Create an array that holds entries for each URL per Transaction ID
// Note that, "?" indicates that the specified URL is not accessed
// in the Transaction and "Y" indicates the URL is accessed in the Transaction
char URL[][]=new char[1][20000];
String pathstr=Sess0+" ";
//System.out.print("Pathr: "+pathstr +"\n");
// Initialize the array of Transaction ID by URL with "?" for the
// initial Transactions
for (int l=0;l<URLName.length;l++)
{
if(URLName[l]!=null)
{
   //System.out.print("Number: "+URLName.length+mk0o;-"\n");
   URL[0][l]='?';
```

```
   }
else
   continue;

}
String Sess="life";
while (line2!=null)
{
StringTokenizerst= new StringTokenizer(line2);

// boolean found=false;
int k=0;

String path;
String col[]=new String[20000];
while(st.hasMoreTokens())
{
col[k++]=st.nextToken(); // assign each token to an array
}
path=col[6]; // Holds URL value
Sess=col[0]; // Holds Session ID
System.out.println("Checking " +Sess0+" "+Sess+" "+path);
if (!Sess.equals(Sess0)) // If session ID is different from previous
{
   //System.out.println("Count "+URLName.length);
for (n=0;n<URLName.length;n++)
{
// Conctenenate an each value of an array of Transaction ID by
// URL matrix value to a string variable
if(URLName[n]!=null)
{
pathstr=pathstr+URL[0][n]+" ";
System.out.println("PATHSTR: "+pathstr+"\n");
}
 else
    continue;
}
// Write the string variable above to output file
bw.write(pathstr);
bw.newLine();
// r++;
Sess0=Sess; // Update Sess0 to next Session ID
//System.out.println("Sess0== :"+Sess0 + "\n");
pathstr=Sess+" "; // Initialize String variable with Session ID
//System.out.println("ASSIGN :"+pathstr + "\n");
// Initialize the array of Transaction ID by URL with "?" for the
```

```
// subsequent Transactions
for (int j=0;j<URLName.length;j++)
{
URL[0][j]='?';
}

}
System.out.println("ASSIGN : "+Sess + "===" + Sess0 +" \n");
// Compare each array of URL with the given path and when match is
// found update the Transaction by URL array value to "Y" and exit
for (int m=0;m<URLName.length;m++)
{
if (path.equals(URLName[m]))
{
//System.out.println("Checking Y \n");
URL[0][m]='Y';
//break;
}
}
line2=br2.readLine();
}
for (n=0;n<URLName.length;n++)
{
// Conctenenate an each value of an array of Transaction ID by
// URL matrix value to a string variable
if(URLName[n]!=null)
{
pathstr=pathstr+URL[0][n]+" ";
System.out.println("PATHSTR: "+pathstr+"\n");
}
 else
    continue;
}
bw.write(pathstr);
bw.newLine();
br2.close();
bw.close();
}
catch (Exception e)
{
System.out.println(e.getMessage().toString());
}
}

}
```

ii. Java source Code for Transaction Identification

```
// This program is used to identify Transactions from preprocessed files
// which is in the form of array of Transaction ID by URL


import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.FileReader;
import java.io.FileWriter;
import java.util.StringTokenizer;
import java.io.*;

/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.
 */
/**
 *
 * @author senait.mezgebu
 */
public class sessionIdremover{
   public static void main(String[] args)
{
try {
// The foloowing lines of codes open and read the top accessed URLs which
// was already determined by heuristics and assigns them to an array
FileReaderfileReader = new FileReader("C:\\Users\\senait.mezgebu\\Desktop\\my Thesis\\CH-3
Preprocessing\\prep tools\\java\\Toppath1.txt");
BufferedReaderbr = new BufferedReader(fileReader);
String URLName[]=new String[10000];
String line= br.readLine();
inti=0;
while (line!=null)
{
URLName[i++]=line;
line= br.readLine();
}
br.close();
```

```
// The following lines open sorted preprocessed file of each region for input
FileReader fileReader1 = new FileReader("C:\\Users\\senait.mezgebu\\Desktop\\my Thesis\\CH-
3 Preprocessing\\prep tools\\java\\PATH_FINAL 2015-05-21_1.txt");
BufferedReader br1 = new BufferedReader(fileReader1);
// The following lines open regional file that is transformed to transaction for output
FileWriterfileWriter = new FileWriter("C:\\Users\\senait.mezgebu\\Desktop\\my Thesis\\CH-3
Preprocessing\\prep tools\\java\\finaloutput111.txt");
BufferedWriterbw = new BufferedWriter(fileWriter);
String line1=br1.readLine();
StringTokenizer st0= new StringTokenizer(line1);

int n=0;
String Sess0=null;
// Just read a sinle record and assign initial Session ID to Sess0
while (st0.hasMoreTokens())
{
if(n==0)
{
Sess0=st0.nextToken();
//System.out.print("asdasdasd: "+Sess0 +"\n");
br1.close();
break;
}
st0.nextToken();
n++;
}
// Open the sorted regional file again to transform the given file into Transaction
FileReader fileReader2 = new FileReader("C:\\Users\\senait.mezgebu\\Desktop\\my Thesis\\CH-
3 Preprocessing\\prep tools\\java\\PATH_FINAL 2015-05-21_1.txt");
BufferedReader br2 = new BufferedReader(fileReader2);
String line2=br2.readLine();
// Create an array that holds entries for each URL per Transaction ID
// Note that, "?" indicates that the specified URL is not accessed
// in the Transaction and "Y" indicates the URL is accessed in the Transaction
char URL[][]=new char[1][100000];
String pathstr=" ";
//System.out.print("Pathr: "+pathstr +"\n");
// Initialize the array of Transaction ID by URL with "?" for the
// initial Transactions
for (int l=0;l<URLName.length;l++)
{
if(URLName[l]!=null)
{
   //System.out.print("Number: "+URLName.length+"\n");
   URL[0][l]='?';
   }
```

```
else
   continue;

}
String Sess="life";
while (line2!=null)
{
StringTokenizerst= new StringTokenizer(line2);

// boolean found=false;
int k=0;

String path;
String col[]=new String[10000];
while(st.hasMoreTokens())
{
col[k++]=st.nextToken(); // assign each token to an array
}
path=col[6]; // Holds URL value
Sess=col[0]; // Holds Session ID
System.out.println("Checking " +Sess0+" "+Sess+" "+path);
if (!Sess.equals(Sess0)) // If session ID is different from previous
{
   //System.out.println("Count "+URLName.length);
for (n=0;n<URLName.length;n++)
{
// Conctenenate an each value of an array of Transaction ID by
// URL matrix value to a string variable
if(URLName[n]!=null)
{
pathstr=pathstr+URL[0][n]+" ";
System.out.println("PATHSTR: "+pathstr+"\n");
}
 else
    continue;
}
// Write the string variable above to output file
bw.write(pathstr);
bw.newLine();
// r++;
Sess0=Sess; // Update Sess0 to next Session ID
//System.out.println("Sess0== :"+Sess0 + "\n");
pathstr=" "; // Initialize String variable with Session ID
//System.out.println("ASSIGN :"+pathstr + "\n");
// Initialize the array of Transaction ID by URL with "?" for the
// subsequent Transactions
```

```
for (int j=0;j<URLName.length;j++)
{
URL[0][j]='?';
}

}
System.out.println("ASSIGN : "+Sess + "===" + Sess0 +" \n");
// Compare each array of URL with the given path and when match is
// found update the Transaction by URL array value to "Y" and exit
for (int m=0;m<URLName.length;m++)
{
if (path.equals(URLName[m]))
{
//System.out.println("Checking Y \n");
URL[0][m]='Y';
//break;
}
}
line2=br2.readLine();
}
for (n=0;n<URLName.length;n++)
{
// Conctenenate an each value of an array of Transaction ID by
// URL matrix value to a string variable
if(URLName[n]!=null)
{
pathstr=pathstr+URL[0][n]+" ";
System.out.println("PATHSTR: "+pathstr+"\n");
}
 else
    continue;
}
bw.write(pathstr);
bw.newLine();
br2.close();
bw.close();
}
catch (Exception e)
{
System.out.println(e.getMessage().toString());
}
}

}
```

Appendix E: Sample Reports