ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS THE CASE OF THE OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY.

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS THE CASE OF THE OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY.

By

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS THE CASE OF THE OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY.

By

AWET FESSEHA

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor(s),	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner,	_____	_____

Declaration

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

_____

Date

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

_____

Advisor

Acknowledgements

There are many people that I need to thank, for making this long journey so memorable. First, I would like to thank my advisor, Ato Workshet Lemnew, for his firm support of this research. I had a great fortune to study under his supervision, and I am very grateful for his guidance and encouragement.

I would like to thank to my wife Selmawit G/kidan for her all support, specially taking care of my little child while I was busy with thesis.

My thanks Professor Bettina Berendt for her borderless supports. I would also like to thank the members of my roommate, namely, Luel, Gedfaw, Yonas, Gere, for their support in various ways.

Finally, I come to the ones I thank the most for their constant love, support, and

Encouragement, for those who I did not mentioned their name, thanks for all supports.

Table of Contents

Abstract

Websites are becoming among the most important media for communicating with the stakeholders. Now a days, many organizations realized that the need to investigate the behavior of their website users are crucial to meet their objectives through undertake a research. One such a research undertakings is using web usage mining.

Web Usage Mining (WUM), refers to the process of knowledge discovery from databases (KDD) applied to the Web data. It comprises three main stages: the preprocessing of raw data, the discovery of schemas and the analysis (or interpretation) of results.

The redesigning task of the website has not ever taken a solid input except simple discussion and implementation internally. Indeed, the website designers, browsing expectation, and the actual user's behavior navigation pattern were not considered.

In order to develop improved website services, there are many options left for web administrator to evaluate the website. Firstly, the site designer could bring amendments on the website based on their expectations of users. Secondly, the web administrator or web master could collect information from users, using the different data collection methods like interview, discussion, or observation. Thirdly, researcher could use the web usage analysis to understand their web usage behaviors.

In this research paper, the web server log files of the official web server were utilized to study the navigational behavior of users using the web utilization miner tool. Access log file was a target data source as can be shown in the navigational behavior of users.

Based on the methodology proposed in Jaideep, et al (2000) was used in this research work comprised of log raw data, pattern discovery and pattern analysis. For discovery of navigational behavior, the Web Utilization Miner WUM was employed, a mining system for the discovery of interesting navigation patterns. The researcher using WUM's mining language MINT dynamically specifies the interesting criteria for navigation patterns. The researcher found that there are directories, which are frequently visited in the website. Other findings are also discussed based on which recommendation is forwarded.

Web Terminology and Definition

According to the world wide Consortium's (W3C) work on Web characterization terminology Magdalini, P. (2006) based on that the definition are as follows:

- **A Web server**

  Server provides access to the Web resources.

- **A Web resource**

  A Resource accessible through any version of the HTTP protocol,(for Example, HTTP 1.1 or HTTP-NG).

- **A Web page**

  The set of data constituting one or several web resources that can be identified by an URI.

- **Page View**

  It occurs at a specific moment in time, when a web page is displayed in a web browser.

- **User Session**

  A delimited number of user's web requests (embedded or user-input, also called clicks), across one or more web servers.

- **Visit**

  A subset of consecutive page views from a user session occurring closely enough (by means of a time threshold or a semantically distance between pages).

- **Web Request**

  A request made by a web client for a web resource. It can be explicit (initiated by the user), or implicit (initiated by the Web client). Another differentiation is: embedded web request (a request made following a link) or user-input web request (a request manually initiated by the user, e.g. by typing the address in the address bar, selecting the address from the bookmarks, history, etc.).

- **Web Browser or Web Client**

  Client or software, which is capable of sending web requests, handling the responses and displaying the requested URIs.

- **Session**

  We refer to a session as a set of web resources requested during a website visit. It is hard to define session accurately. When a website visitor browses through a website, and then makes a pause and returns, her/his visit may be considered as one or two sessions.

Acronyms

Some of the abbreviations and acronyms used throughout this thesis are listed below:

AAU	Addis Ababa University
CERN	Center for European Nuclear Research
CLF	Common Log Format
CRM	Customer Relationship Management
DNS	Domain Naming System
ECLF	Extended Common Log Format
ETC	Ethiopian Telecommunication Corporation
FQDN	Fully Qualified Domain Name
GMT	Greenwich Mean Time
GSP	Generalized Sequence Pattern
HTTP	Hypertext Transfer Protocol
ICT	Information Communication and Technology
KDD	Knowledge Discovery in Data
LODAP	Log Data Preprocessor
NCSA	National Computer Security Association
OLAP	Online Analytical Process
URL	Uniform Resource Locator
VPN	Virtual Private Network

WAN	Wide Area Network
WWW	World Wide Web
WUM	Web Utilization Miner
WUM	Web Usage Mining
WUMprep	Web mining pre-processing
WUMprep4Weka	Web mining pre-processing for Weka
W3C	World Wide Web Corporation

List of Table

List of Figures

CHAPTER ONE: INTRODUCTION

1.1.Background

Initially, Internet was designed for exchanging mails between users; later it became trendy for use of WWW. The www or 3W is now popular service from among Internet services. There are a number of service providers (ISP) for the use of the Internet across the world. In Africa, the number of the internet users is increasing from time to time. 5.6 % of the world internet users are from Africa, Ethiopia has 0.4 % share among African internet users. This, however, seems insignificant when compared with the rest of the world. Generally speaking the number of the internet users across the world is increasing in a dramatic way[1].

According to Moges, (2005), Addis Ababa University is, one of the oldest higher education institutions in Africa with current enrollment of over 40,000 students in its regular and continuing education programs. The various faculties of the university are distributed over eight major campuses and eight minor campuses, all within the capital, except one that is 45 km south of the capital.

Further explained in Moges, (2005), four major campuses (Main Campus, Business Campus, Technology Campus, and Science Campus) form the core network and connected via fiber network. The remaining campuses are connected with virtual private network (VPN) provided by the national service provider the Ethiopian Telecommunication Corporation (ETC). Addis Ababa University (AAU) has adopted information and communication technology (ICT) resources as strategic tools in advancing its mission of learning, teaching, and public service.

As such, the proper integration, use, and management of ICT resources have become vital to the success of the university. Proper integration, use, and management of AAU's ICT resources entails, among others, equitable sharing of their limited capacity, protection of sensitive information to which they provide access, prevention of abusive practices enabled by their use, and ensuring their manageability through technology standardization[2].

[1] http://www.internetworldstats.com/stats1.htm#africa
[2] www.aau.edu.et/administration/DRAFT ICT POLICY AT AAU

The official web services of AAU are an organized collection of Web pages. Information is presented in various formats, ranging from research papers, and educational content, to multimedia content, and blogs. As the result, the web pages are serving as a bridge between information providers and the information seekers.

1.2.ICT Development in AAU[3]

The ICT Development Office was established around the summer of 1996 through visionary leadership. The newly formed office initiated a project named AAUNet that has resulted in a wide area network (WAN) whose first phase of construction completed in November of 2001.

The network, which connects all the 14 widely distributed campuses of the university, has been growing since. The services delivered through the infrastructure have also been increasing. Despite the pioneering role, AAU has played in the deployment and use of ICT and the fact that it now has a relatively sophisticated infrastructure; however, it is still far from a point where, it served adequately by ICT. At the same time, AAU's need for and dependence on effective ICT support is now greater than ever.

The national attention given to the expansion and improvement of higher education as critical factors in the country's development has explicit and implied requirements for the use of ICT in realizing the objectives. AAU's role as a major contributor to these expansion and enhancement efforts, along with the imperatives contained in its own ambitious strategic plan, call for the speedy improvement of the efficiency and quality of its academic and administrative functions. This is hard, if not impossible, to accomplish without adequate ICT support. There are currently various initiatives underway, both at the ICT development office and various quarters around the university, to meet the growing demand for and address the ICT support needs of the university.

[3] www.aau.edu.et/administration/ICT

1.3.Nature and Content of Addis Ababa University website

Generally, the website designed is to support the objective of the university. The AAU website has both static and dynamic nature. There are few websites that are static in nature. For instance, those pages are not interactive with their users' but the majority of the web pages are dynamic in nature, where support the MYSQL database incorporate with JOOMLA packages help users to interact with websites users.

When we come to Website contents, it posts numerous information regarding to the objective of university. The web presents information on several topics and issues, like links to international journals, each page has information regarding to the objective of the university.

1.4.Statement of the Problem

During the past few years, the World Wide Web has become the biggest and most popular way of communication and information dissemination. Every day, the web grows by roughly a million electronic pages, adding to the hundreds of millions pages already on-line.

As statistics shows the number of websites published every day is increasing quickly. There are now about 184 million registered domain names worldwide, a 9 % increase over the same period of last year4. As those trends become stronger and stronger, there is much need to study web-user behaviors to better serve the users and increase the value of institutions or enterprises.

The importance of the study web users, further explained by Marya, et al, According to her, most websites are set up with little knowledge on the navigational behavior of the users accessing them; Feedback on the occurring navigation patterns can notably aid site owners in efficiently organizing the website they present to their visitors.

Website design is currently based on thorough investigations about the interests of website visitors and on less investigated assumptions about their exact behavior. In Lukas, C., (n, d), concrete knowledge on the way visitors navigate in a website could prevent disorientation and help owners in placing important information exactly where the visitors look for it.

[4] http://news.softpedia.com/news/Domain-Name-Registration-Slows-Down-122419.shtml

Today, understanding the interests of users is becoming a fundamental need for Websites owners in order to better serve their visitors by making adaptive the content and usage, structure of the site to their preferences. The analysis of web log files permits to identify useful patterns of the browsing behavior of users, which it exploited in the process of navigational behavior.

Web usage mining is application of data mining techniques to discover user access patterns from web data. Web usage data captures web-browsing behavior of users from a website. Web usage mining can be classified according to kinds of usage data examined, Further explained in Sulu, G.,(2003), web usage mining is the process of identifying representative trends and browsing patterns describing the activity in the website, by analyzing the users' behavior. Academic institutions are good examples that develop website. One such institution of the education sector which is involved in many activities is the Addis Ababa University.

Based on the preliminary investigation of AAU Official website conducted by the researcher, users are not satisfied by the website; even the users' blame for AAU website services. As to the researcher's interview with communities of the university, the site is further blamed for its complexity.

Discussion with the ICT development staffs also confirmed that users who have been claiming on the usage difficulty of the website was due to the failure to get the services of the website that resulted from broken page links. The website was redesigned repeatedly since it has been launched. As the researcher interviewed the website design team members of the ICT development office, the redesigning task of the website has not ever taken a solid input except simple discussion and implementation internally. Indeed, the website designers, browsing expectation, and the actual user's behavior navigation pattern were not considered.

However, such inputs were not trusted to be the right directions of designing better website. This in turn, could not fit the website users and it is costly to practice on unreasonable frequent site changes, so the usage patterns of the website need to be analyzed for effective website evaluation and recommendations for the website redesign Marya ,(2005).

To address the problem, Mokeonnen,T,. (2009) used tool, web log analyzer and tried to analyze hit statistics, most requested pages, most visited directories, most frequent entry and exit pages, users' visiting time and common errors encountered of the website on monthly and day of the week basis. Besides, he used Apriori based association rule mining to come up with some of the frequently accessed group pages. Finally, he recommended the usage of combination of statistical analyzer tools and data mining approaches for discovery of better usage patterns.

However, his findings did not show common navigational behavior of users in generalized manner. Unlike association rule mining of market–basket analysis, so interesting rules about common navigational sequences should be extracted for exploration of the appropriate usage trends, which is the main interest of this work.

This research, therefore, focused on the navigational behavior of users the official website of AAU using the web utilization miner, which bases on generalized sequences, appropriate tool for web navigational behavior of users.

1.5. Justification

The Introduction of the Internet paved a road to the creation of World Wide Web (www). Now WWW has become the most popular services among other services that the Internet provides. The number of users as well as the number of website has been increasing dramatically in the recent years. Consequently, this created an occasion for the existence of many statistical tools. Those statistical tools give many options to the web master (web administrator) in analyzing the interaction among the users and the website in a statically manner, but those tools are not good enough in providing the efficient and effective user behaviors.

One important data source for the study is the web-log data that traces the users' web browsing, just for each second, gigabytes of data, or even more, are created by the World Wide Web, and even automatically collected and stored by the World Wide Web server.

Based on the description of Kosala et al, (2000), the web log creates an opportunity and encouragement for all data mining researchers. That's why many consider it as the largest data warehouse in the world. Website administrators can then use this information to redesign or customize the website according to the interests and behavior of its visitors, or improve the performance of their systems.

The web administrators in AAU use different tools in order to understand the website usage by the users. None of those tools seem to give enough insight on how users use the website because the tools that are employed do not take into consideration on how users behave. Furthermore, it is important to realize users' behavior. As a result, it can help evaluate a website based on users' behavior during navigation of the website.

1.6.Research questions

Based on the problem statement, attempts were made to answer the questions below.

- Do visitors spend even amount of time in website pages?
- Where do most visitors spend their time after visiting the home page?
- Do most of the users of AAU official website start navigation at the home page or are there many arbitrary routes to its page?
- Which pages are the most common pages on which users give up their navigation?
- What are the extent of match between website designer's expectation of page navigation and actual users' navigation?
- Which pages are accessed together more frequently?
- Do most visitors know where they go after the search engine?
- Are there directories that are frequently accessed than others?

1.7. Scope and Limitation of the Research

Generally, web mining has different branches: such as web content mining, web structure and web usage mining. The focus of this research is on web mining usage pattern of AAU official website.

There are three types of web related log files, namely web access log, error log and proxy log files. However, in this research, web access log records is used as dataset because literatures and previous researches justify that web access log files is the typical source of navigational behavior.

The limitation in this paper is, to investigate the navigational behavior of users', behaviors are expressed using the MINT query. Those queries describes the behavior of users in generalized manner, consecutively the researcher of this paper uses some query that can describe the behavior of the users, it is obvious that the result of any research would be better if more query were used.

1.8. Objectives

1.8.1. General objective

The general objective of the research is to apply web-mining techniques for discovering navigational behavior of AAU official function usage to reveal previously unknown interesting patterns extracted in order to recommend possible measures for further improvement of the official website of AAU.

1.8.2. Specific objectives

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

- To review literature in the area, in order to put concrete background and justification for the research.
- To identify the type of web data, and collect the data from web server.
- To prepare those data set using different preprocessing techniques.
- To analyze, users the navigational behavior.
- To analyze the sequence of the website i.e. based on the user navigational behavior.
- To interpret the interesting pattern to discover new knowledge i.e. finding of the research.

- To draw conclusion based on the findings and possible application of both techniques for web usage pattern or navigational behavior of users.
- To make some appropriate recommendations based on the conclusions.

1.9. Data Collection for the Study

In this study, the data collected from the official web server of the AAU, which stores normally secondary data source in view of the fact that web log keeps every activity of the user regarding to visit of the website.

1.10. Data Selection

At present, towards web usage mining technology, the main data origin has three kinds: server data, client data, and middle data (proxy). The researcher of the paper, uses server data stored in the official web server of AAU, those data are kind of log format that are stored in extended log format. Those formats supported by most Apache server such as AAU web server. See in Appendix I, for the detail of the extended log format.

1.11. Data pre-processing or data cleaning

The rationale of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining areas. For the purposes of different mining applications, irrelevant records in web access log eliminated during data cleaning.

According to olfa, et al, (n.d), most log files are full of dusts that are insufficient, inconsistent, and including noise so the data pretreatment is to carry on a unification transformation to appropriate sets. To have kind of datasets, there are some data cleaning phases that are essential to implement. Those concepts will be described in (the Chapter 3).

1.12. Data analysis

To address the objective of research, different data mining approaches have been performed and some statistical analysis on the data set to get penetration about the web usage trends and to reveal interesting navigational patterns from the web log records. For data analysis, web utilization miner deployed.

1.13. Tools for Experiment

There are commercial and free available tools are exists for web mining. According to Castellano et al, (2007), WUMprep is one of the freely available tools for preparations of web log data. It consist a set of Perl scripts for cleaning the web log file of irrelevant, automatic requests and creating sessions. It designated function for educational purpose, furthermore navigation pattern discovery are performed on the portion of the web server log that contains the sessions. The justification for why these tools selected will be address on chapter THREE.

1.14. Interpretation and reporting results

After excluding least interesting patterns from the analysis, those patterns that are interesting and actionable ones has been interpreted and reported to be used for reaching a conclusion and for further appropriate recommendations.

1.15. Organization of the Thesis

This thesis organized as five chapters, the first chapter deals with the general introduction beginning with background. It also discussed on statement of the problem, data collection, data preparation, scope, and limitation of the study and objective of the study; etc.

In Chapter 2, presents the literature review, based on two main areas, data mining, and web usage mining.

Chapter 3 presents the methodology and algorithms. The researcher points further discussions on employed tools.

Chapter 4, presents the experiment conducted and its finding.

Chapter 5 presents like conclusions, recommendation, and further work in this research area.

CHAPTER TWO: LITERATURE REVIEW

2. Data Mining and Web Usage Mining

2.1.Data Mining

In According to, Lita, et al (2004), define data mining "is the process of extracting previously unknown information from (usually large quantities of) data, which can, in the right context, lead to knowledge, in other words; the concept of data mining in refers to the entire knowledge discovery in databases process (KDD)."

This knowledge is not arbitrary; it relates to a problem, the problem we want to solve. That is why performing data mining to optimize the performance of a web server. In reference of Lukas (n, d), the use of data mining to discover which products are being purchased together or to identify whether the site is being used as expected.

2.1.1. Data Mining Approaches

Data mining have two approaches according to (Brendit, 2011), the approaches are between undirected, and directed data mining, further describe it like this:"There are two styles of data mining. Directed data mining is a top-down approach, used when we know what we are looking for. This often takes the form of predictive modeling, where we know exactly what we want to predict. Undirected data mining is a bottom-up approach that lets the data speak for itself. Undirected data mining finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important."

However, there are no generally applicable rules on how data mining, should be performed.

- Decision trees as a technique for prediction.
- Neural networks as a technique for prediction.
- Navigation patterns in WUM as a query-directed technique for pattern detection.

2.1.2. Motivation of Data Mining

Data mining according Sulu, (2003), has emerged, as one of the most is exciting and dynamic fields in computer science and software engineering. The term “data mining “and “knowledge discovery in data base “or KDD are often used synonymously. Knowledge discovery in database is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns models in data.

Data mining is a step in, knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or model in data. Simply stated, data mining refers to the process of extracting previously unknown, valid and potentially useful knowledge from data. Similar to the above definition, according to Ian (2005), data mining defined as the process of discovering patterns in data.

Another definition is that data mining is a variety of techniques used to identify valuable of information or decision-making knowledge in bodies of data, and extracting these in such a way that can put to use in areas such as decision support, prediction, forecasting; and estimation. The data is often voluminous but, as it stands, of low value as no direct can be made of it; it is the hidden information in the data that is useful. For this reason, data mining often referred as “secondary” data analysis.

2.1.3. Limitations of Data Mining

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related.

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to “real world” circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large

pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation, according to Brendit, (2011) of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight scheduled to depart, related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables.

In fact, the Individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations).

2.2. Web usage mining

Web mining" is the use of data mining techniques to extract useful patterns from the web. Those extracted patterns are used to improve the structure of websites, improve the availability of the information in the websites and the way those pieces of information are introduced to the website user, and to improve data retrieval and the quality of automatic search of information resources available in the website is being used as expected". According to Narendra, et al., (2003), Web mining defined "as the use of data mining techniques to automatically discover and extract information from web document and services".

Web usage Terminology

There are various definitions regarding to the use of most common terminology in web usage mining besides what it have been described in the beginning of thesis (terminology and definition), depends on the field of study the same "terminology" can have different meanings.

In general, According to Lavoie, B., et al (1999) there are different meanings by authors in the WUM literature and W3C's web characterization authority (W3C's WCA). The summarize definitions are as follows.

Term	W3C's WCA	WUM Literature
User	Person using a browser	Login or cookie or IP or (IP, User Agent)
User session	Delimited user requests over multiple servers	Delimited user requests on one server
Visit	Server session	-
Episode	Related user requests	Related user requests

Table 1 : Terminology comparison table

2.2.1. Web Log Information

Since the thesis is about user navigational on web access using web usage mining that is based on web server logs, it is important to understand what information web server logs contain and types of log format.

A web log is a file to which the web server writes information each time a user requests a resource from that particular site. Most logs use the format of the common log format Cooley et al., (1997a) furthermore, those are confirmed by (Lavoie, et al (1999)the most popular log file formats (developed by the CERN and the NCSA) are the common log format (CLF) and an extended version of the CLF, combined log format, known as ECLF. According to Berkan, y., (2002), the difference between them is that the former does not store referrer and agent information of the requests.

Types of Log Format

Besides the above, the types of log formats can be categorized [5]into four; those are common, extended, cookie, and MS-IIS.

1. Common: The common log contains the requested resource and a few other pieces of information, but does not contain referral, user agent, or cookie information. The information is contained in a single file. The example is as follows:

```
picasso.wiwi.hu-berlin.de - - [10/Dec/1999:23:06:31 +0200]
"GET /index.html HTTP/1.0" 200 3540
```

2. Extended: An extended combined log format is an extension of the common log format. The combined format contains the same information as the common log format plus three (optional) additional fields: the referral field, the user agent field, and the cookie field. Examples are as follows:

```
picasso.wiwi.hu-berlin.de - - [10/Dec/1999:23:06:31 +0200] "GET
/index.html HTTP/1.0" 200 3540 "http://www.berlin.de/"
"Mozilla/3.01 (Win95; I)"
```

3. Cookie: Cookies take the form KEY = VALUE. Multiple cookie key-value pairs are delineated by semicolons **(;).**

```
picasso.wiwi.hu-berlin.de - - [10/Dec/1999:23:06:31 +0200] "GET
/index.html HTTP/1.0" 200 3540 "http://www.berlin.de/"
"Mozilla/3.01 (Win95; I)" "VisitorID=10001; SessionID=20001"
```

4. MS-IIS: Kind of log format stores at server side of the Microsoft web server which normally known as MS-IIS.

```
picasso.wiwi.hu-berlin.de, -, 10.12.99, 23:06:31, W3SVC2, WWW,
100.100.100.100, 547, 444, 0, 200, 0, GET, /index.html, -,
```

[5] http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format

Contents of Log Format

Most apache formats are NCSA[6] combined log format , here are a single format example entry of the log file , is shown in an entry is stored as one long line of ASCII text, separated by tabs and spaces, based on, (Berkan, y.,2002) (Cooley et al., 1997a).

```
66.249.67.111--[12/Dec/2010:04:26:46+0300]"GET
/index.php/component/events/view_week/1995/04/03 HTTP/1.1" 200
28776 "-""Mozilla/5.0(compatible;Googlebot/2.1;
+http://www.google.com/bot.html)"
```

Address

66.249.67.111

This is the address of the computer making the HTTP request. The server records the IP and then, if configured, will look up the Domain Name Server (DNS) for its FQDN.

RFC931 (Or Identification) :

-

Rarely used, the field was designed to identify the requestor. If this information is not recorded; a hyphen (-) holds the column in the log.

Authuser:

-

List the authenticated user, if required for access. This authentication is sent via clear text, so it is not really intended for security. This field is usually filled by a hyphen **-.**

Time Stamp:

[12/Dec/2010:04:26:46 +0300] [01/Nov/2001:21:56:52 +0200]

The date, time, and offset from Greenwich Mean Time (GMT x 100) are recorded for each hit. The date and time format is: DD/Mon/YYYY HH:MM: SS.

The example above shows that the transaction was recorded at `04:26:46` on `12/Dec/2010` at a location 3 hours forward GMT. By comparing time stamps

[6] http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format

between entries, it can also determine how long a visitor spent on a given page that is also used as a heuristic in determining sessions.

Target:

"GET /index.php/component/events/view_week/1995/04/03 HTTP/1.1"

One of three types of HTTP requests is recorded in the log. GET is the standard request for a document or program. POST tells the server that data is following. HEAD is used by link checking programs, not browsers, and downloads just the information in the HEAD tag information. The specific level of HTTP protocol is also recorded.

Status Code:

200

There are four classes of codes regarding to

1. Success (200 series).
2. Redirect (300 series).
3. Failure (400 series).
4. Server Error (500 series).

Transfer Volume:

1749

For GET HTTP transactions, the last field is the number of bytes transferred. For other commands this field will be a hyphen (-) or a zero (0).

The transfer volume statistic marks the end of the common log file. The remaining fields make up the referrer and agent logs, added to the common log format to create the "extended" log file format. Let's look at these fields.

Referrer URL:

http://www.cs.bilkent.edu.tr/guvenir

The referrer URL indicates the page where the visitor was located when making the next request.

User Agent:

Mozilla/4.0 (compatible; MSIE 5.5; Windows 95)

The user agent stores information about the browser, version, and operating system of the reader. The general format is: browser name/ version (operating system).

2.2.2. Taxonomy of Web Mining

In reference Bamshad et al(n.d),web mining are classified in three main areas, namely web content mining, web structure mining and web usage mining ,the detail of those will be discussed in the following section 2.2.3.1

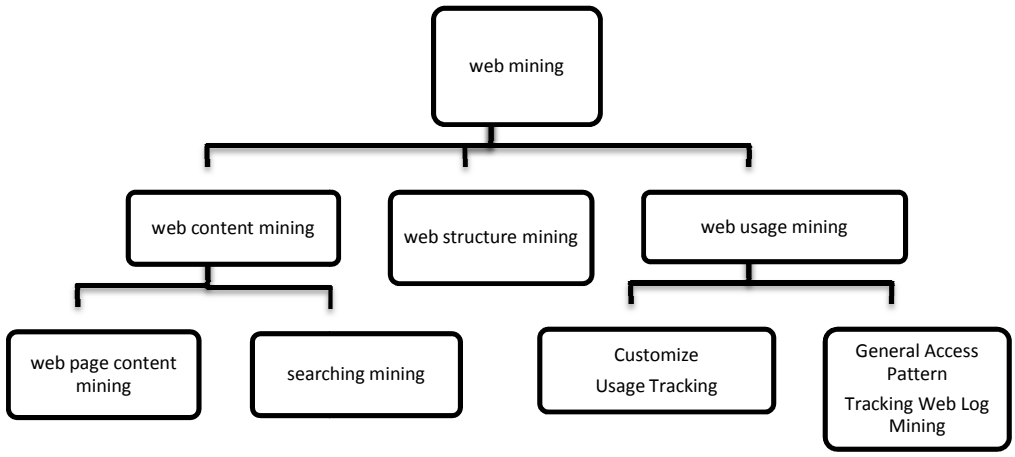


Figure 1: Taxonomy of Web mining,

2.2.2.1. Web Usage Mining: WUM

Web usage mining can also be defined as the application of data mining techniques to discover users web navigation patterns from web access Zalane et al, (1998) in addition to above, generalized definition according to Berkan,(2002), The aim of a general web usage mining system is to discover general behavior and patterns from the log files by adapting well-known data mining techniques or new approaches proposed.

The sources of the data for web usage mining are secondary data as previously discussed such as web server access logs, browser logs ,user profiles ,registration data, user sessions or transactions and other, unlike of web structure and web content which uses primary data.

Furthermore, It has advantage, according Chu-Hui et al , (2008) to enhance the usability of the web information and apply the technology to the web application.
For instance, pre-fetching and caching, personalization, target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc, in addition there are also more goals Lita,et al (2004), includes ,

- The improvement of site design and structure,
- The generation of dynamic recommendations,
- And improving marketing

Finally, according to Jaideep, et al., (n.d) generalized as web usage mining focuses on techniques to search for patterns in the user behavior when navigating the web.

2.2.2.2. Web Structure Mining: WSM

The category of structure mining, according to Istrate (2000), structure is defined by ”hyperlinks between pages and HTML formatting commands within a page” but further explained by Lita, et al (2004), according to author, structure mining which focuses on link information. It aims to analyze the way in which different web documents are linked together, mining the link structure aims at developing techniques to take advantage of the collective conclusion of web pages’ quality which is available in the form of hyperlinks Henri et al , (2000), where links on the web can be viewed as a mechanism of implicit support.

2.2.2.3. Web Content Mining: WCM

Web content mining is a research field focused on the development of techniques to assist a user in finding web documents that meet a certain criterion. The contents of most of the web pages are texts. According to Istrate,(2000), graphics tables, data blocks and data records are also kind of content a web page can have so that web content mining issues for the of improving the contents of the web pages, improving the way they are introduced to the website user, improving the quality of search results, and extracting interesting web page contents.

2.2.3. Generalized structure of web usage mining

Analysis of a website usage could range from straightforward statistical approach of analysis such as page access frequency to sophisticated form of analysis such as the common traversal paths through a website (chitraa et al 210).

As HU et al .n.d clearly articulated that in their work, web usage mining involves usage data gathering, usage data preparation, navigation pattern discovery, pattern analysis and visualization and pattern application. These tasks could be generalized into the following three important phases: pre-processing, pattern discovery and pattern analysis.

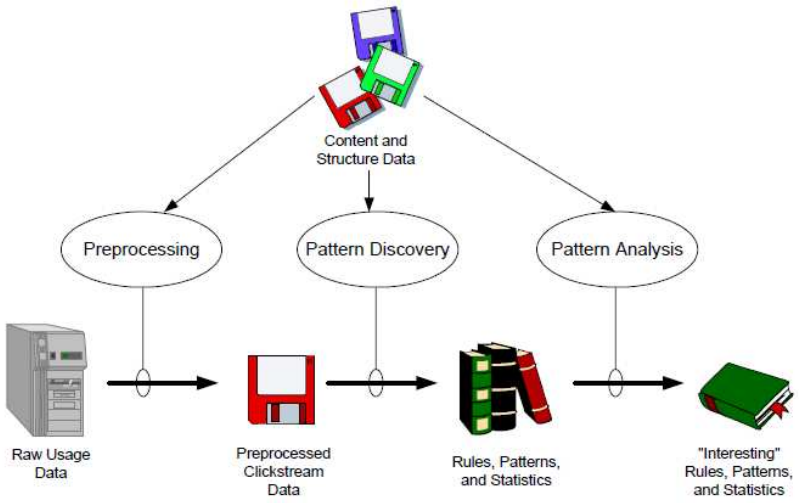


Figure 2: high level web usage mining process Jaideep, et al (n.d), page 4

A. Data collection

Data for web usage mining can be collected at several levels. According to Kerkhofs et al (2001), may be faced with data from a single user or a multitude of them on one hand and a single site or a multitude of sites. The second way of data collection is on the web server level.

These servers explicitly log all user behavior in a more or less standardized fashion. It generates a chronological stream of requests that come from multiple users visiting a specific site, but according to Briand, et al, (2005) can be the collection of the data for web usage mining most commonly from:

- The web usage data includes data from web server access log, proxy server
- Logs, browser logs, user profiles, registration data, cookies, and user queries.

According to Castellano, et al (2007) the following can also be the source of the data.

- E-commerce and product-oriented user events. E.g. shopping cart changes, or product click-through, etc.
- Meta-data, page attributes page content, site structure.

Sources of Data for Web Usage Mining

Data that can be used for web usage mining can be collected at one of these three parts and thus we talk in reference with Berkan, y (2002), of those is:

Server level collection:

- The server stores data regarding requests performed by the client, thus data regard generally just one source;

Client level collection:

- It is the client itself, which sends to a repository information regarding the user's behavior. This can be done either with an ad-hoc browsing application or through client-side applications running on standard browsers.

Proxy level collection:

- Information is stored at the proxy side, thus web data regards several websites, but only users whose web clients pass through the proxy.

B. Data pre-processing

In reference with Dipa, (2010), data pre-processing is an important step in the knowledge discovery process, because quality decisions are based on quality data. more ever, this idea of importance of preprocessing steps discuss in, Haji, et al, (2007), emphasis on fundamental role in achieving meaningful and reliable results from WUM process, without effective preprocessing the results obtained will have negative impact on the next steps of the process pattern discovery and pattern analysis.

It is important to understand that the quality data is a key issue when we are going to mining from it. In reference with Suneetha et al (2009), nearly 80 % of mining efforts often spend to improve the quality of data, furthermore, the attributes that we can look for in quality data includes accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility.

Tools of Preprocessing

Most existing tools provide mechanism for reporting user activity in the servers and various forms of data filtering. By using these tools, determination of the number of accesses to the server and to individual files, most popular pages, the domain name and URL of the users who visited the site can be solved, but not adequate for many applications, furthermore, In reference Cooley et al., (1997a) the administrator of a system has an access to the server log. However, the pattern of site usage cannot be analyzed without the use of a tool. Therefore, data mining method would ease the system administrator to mine the usage patterns of a particular site. These tools have no ability in-depth analysis and their performance is not enough for huge volume of data.

There are commercial and free available tools are exists, according to Castellano, et al (2007), one of the freely available tool for web log data preparation called WUMPrep which consists of a set of Perl scripts for cleaning the web log file of irrelevant and automatic requests and creating sessions in it and its main purpose for educational purpose. According to Dipa, (2010), the other open source preprocessing tools are WUMprep4Weka; those tools are designed to work with WEKA, unlike of WUMprep, which designed to use with WUM (web utilization miner).

Further described in Castellano et al, (2007), there are commercial preprocessing tools but the most common tools on tare LODAP (Log Data Preprocessor) and EasyMiner, the later developed by MINEit software ltd. Both of them designed to understand the most common log file formats. They designed to take input log files related to a Website and output a database containing some statistics about pages visited by users and the identified user sessions. The preprocessing of log files is aimed to the preparation of web data in order to mine significant usage patterns. A key feature of LODAP is the wizard-based interface that guides the user during the preprocessing of the log data.

Data Cleaning

First, irrelevant data should be removed to reduce the search space and to bias the result Space. The intention is to identify user sessions to build up out of page views, not all hits in a log file are necessary. As web log files record all user interactions, they represent a huge and noisy source of data, often comprising a high number of unnecessary records.

According to Castellano et al, (2007), the data cleaning is intended to clean web log data by deleting irrelevant and useless records to retain only usage data that can be effectively exploited to recognize users' navigational behavior.

Removing Unnecessary Records as

According to Enrique et al ,(2000), there are two kinds of records are unnecessary and should be removed: firstly the records of graphics, videos and the format information the records have filename suffixes of GIF,JPEG, CSS, and so on, which can found in the URI field of the every record. Reference Mohd, et al, (2008), For example, by filtering out image requests, the size of web server log files reduced to less than 50 % of their original size.

Secondly, the records with the failed HTTP status code, by examining the status field of every record in the web access log, the records with status codes over 299 or under 200 are removed.

Robots

According Castellano et al, (2007), the term 'robot' to refer to any programmable software agent that does not access a site interactively. Furthermore, explained in the paper, these requests can mislead the analyst, because these sequences do not reflect the way human visitors navigate the site.

Types of Robots

In a number of literatures there, many types of robots but, According to Brendit, (2011), two types of robots can be distinguished (categorized) as"*ethical robots*" and "*unethical robots*". Ethical robots take by the "netiquette (internet rules) for robots" or: before they access any page of a site, they access the file robots.txt in order to see what they are allowed to visit and index, and what not. Furthermore explained in that, ethical robots have two effects: first, they show their "robot identity," and second, they only access pages they are allowed to see. Unethical robots don't do this. They may not even access robots.txt.

There are ways to detect whether it's a robot or not based on requests to the web server, according to Jose et al, (2007); two subsequent requests for the same URL are collapsed into one if the time between the requests did not exceed a threshold, e.g., 5 sec. This threshold can be longer than that for robots because a person needs more time than a program to make a renewed request. But according Rajni et al, (2009) the most widely accepted threshold for of 2 seconds between two consecutive requests the entries that corresponds to robots can be eliminated.

Exclusion of Robots

The most important step of data cleaning was the removal of robot accesses from the log data. In reference Berkan, (2002), requests originated by web robots. Log files may contain a number of records corresponding to requests originated by web robots. Web robots (also known as web crawlers or web spiders) are programs that automatically download complete websites by following every hyperlink on every page within the site in order to update the index of search engine.

Requests created by web robots are not considered usage data and, consequently, have to be removed. To identify web robots' requests, the data-cleaning module implements two different heuristics.

First, all records containing the name "robots.txt" in the requested IADIS international conference applied computing 2007 resource name (URL) are identified and straightly removed.

The Second heuristic is based on the fact that the crawlers retrieve pages in an automatic and exhaustive manner, so they are characterized by a very high browsing speed (intended as total number of pages visited/total time spent to visit those pages). Hence, for each different IP address we calculate the browsing speed and all requests with this value exceeding a threshold (pages/second) are regarded as made by robots and are consequently removed. The value of the threshold is established by analyzing the browser behavior arising from the considered log files.

User and Session Identification

Once the web log file is processed and all the irrelevant entries have been removed, it is necessary to identify the users that visit to the site. The task of user and session identification is found out the different user sessions from the original web access log. In reference (Rajni, P., et al 2009), user's identification is, to identify who access website and which pages are accessed. However, this task is not easy, because few websites that uses authentication to access the resource so the web records, only records the visitor's host and user agent.

Further description in Castellano et al, (2007) the problem to identify the user identification getting worst because different visitors sharing the same host cannot be distinguished. In addition to that, if proxy servers are used, the problem becomes even more sensitive. The only way to identify a user in reference Rajni, (2009) to use cookies or authentication mechanisms make the identification of a visitor possible, but are undesirable due to privacy concerns.

The goal of session identification is to divide the page accesses of each user at a time into individual sessions. According to Castellano et al, (2007), session is made up of all the visited pages by a user, the technique is based on establishing a time threshold, so if two access take more than the fixed time thresholds. It is considered as a new session, most accepted threshold of 30 minutes or 1800sec but according to Jose et al (2007), threshold of most commercial products establish a threshold of 25.5 minutes.

C. **Navigation Patterns and Important to Discover**

Sequence and Navigational Pattern

According to Lukas (n, d), sequence is an ordered list of items, in our case web pages, ordered by time of access. In the pioneering work of sequence mining is defined as follows: “given is a collection of transactions ordered in time, where each transaction contains a set of items”.

The goal is to discover sequences of maximal length that appear more frequently than a given percentage threshold over the whole collection.” A frequent sequence is “maximal,’’ if no sequence containing it is also frequent. If it is instruct the miner to find only maximal frequent sequences, we obtain fewer and more compact results.

In the reference Berendt et al, 2000, the definition of the sequence-mining problem has an implication: The items constituting a frequent sequence did not necessarily occur adjacently. They just appear in many data records in the same order, this is often desirable: to investigate the causes of manufacturing errors, only the sequences containing error and cause, not the many events in between. The same is true when we search for operating system signals.

Generalized sequences

Berendt et al (1999), a sequence depicts the behavior of a single user, as a vector of adjacent page requests. Generalized sequence rather interested in *patterns* that reflect the navigational behavior of many users. Such patterns may be interesting because they are frequent, i.e., match many sequences, or for some other application-specific reason related to their statistics, structure, or content. Some of the important definitions explained by the author are:

Let L be a sequence log over sequences from U* and let g = g1*g2*:::*gn be a g-sequence over elements of U. For each i = 2,…, n and for each j < i, the “confidence of gi towards gj within g” is the ratio of the number of sequences containing g1 * . . .. gi−1 * gi to the number of sequences containing g1 * … gj :

$$confidence_{g1*\ldots g2}=\frac{hits(g1*\ldots gi-l*gi)}{hits(g1*\ldots gi)}$$

Whereby the confidence of $g1$ *within* g *is:*

$$confidence_{(gl,e,g)}=\frac{hits(gl)}{|L|}$$

Input: Template < v1; _; v2; : : : ; vk > and predicates of type A, B, C

Output: A set of navigation patterns.

1. Generate the set of All gSequences by traversing the Aggregated Log:

a) For each order-preserving sequence of nodes < n1; _; : : : ; _; nk > in a branch produce the g-sequence d =< d1; _; : : : ; _; dk >, where di = (ni:page; ni:occurrence).

b) if d is already in All gSequences, then skip it.

c) else if for all i = 1; : : : ; k:

i. The web page referred to in ni satisfies the type A predicates for variable vi.

ii. The position of ni in the sequence is allowed by the template.

iii. The occurrence number in ni is permitted for vi.

then add d to All gSequences.

2. Construct the navigation pattern for each g-sequence d in All gSequences:

a) Compare d with the g-(sub)sequences already in the set Tested gSequences and test if it can be rejected without building the navigation pattern.

b) If d is not rejected, construct the navigation pattern for it:

i. Find all branches of the Aggregated Log that conform to d.

ii. Merge at each element of d.

iii. Compute the supports of the nodes produced by merging.

iv. Test the C predicates against the navigation pattern.

v. If d is rejected

then store the smallest prefix that caused the rejection in the set Tested gSequences, marking it as R(ejected).

else store d in Tested gSequences, marking it as S(uccessful).

c) If d is not rejected, then output its navigation pattern.

Figure 3: WUM_gseqm algorithm, Page 22, Bettina et al 1999.

Navigational Pattern

Navigation pattern can be defined as a graph built according to a pattern descriptor. Obviously, the patterns to be discovered must be described according to more general criteria. In particular, Murat et al (n.b) need a way of specifying the “interestingness" of navigation patterns, as subjectively conceived by the mining expert.

We suggest that, informally, “interestingness" is a specification concerning given an “interestingness descriptor", it must build all conformant navigation patterns by assigning appropriate values to all components of the statement not explicitly specified. WUM, Mary et al, (2000), an “interestingness descriptor" is a query in our mining language, MINT.

Knowledge Discovery Queries

Similarly to Lukas, (n, d), he believes that good mining results require a close interaction of the human expert and the mining tool, in which the expert uses her/his domain knowledge to guide the miner. Therefore, WUM provides a mining query language, with which the expert can specify the subjective characteristics that make a navigation pattern of interest to her/his.

The notion of interestingness based on beliefs is discussed in Dietmar, et al (n.d) a belief is a rule of the form A->B, which is expected to be true. The same study proposes mechanisms for the verification of beliefs and the discovery of belief violations in the context of association rules. To the best of our knowledge, there is no respective formalism for beliefs on sequential patterns. However, MINT allows the specification of beliefs or belief violations as predicates. Predicates can also be used to specify the structure or statistics a navigation pattern should have to be of significance. Thus, besides the classical mining criterion of a support threshold, much more elaborate criteria are supported.

2.2.4. Techniques of Web Usage Mining

It is very difficult to classify a specific technique for web usage mining; techniques are combined together in discovering web usage mining, but In general the techniques applied to web usage can classified according to Bamshad et al ,(n.d), are:

Statistical Analysis

Statistical techniques are the most common methods to extract knowledge about visitors to a website by different kinds of statistical analysis (frequency, median, mean, etc) of the session file. One can extract statistical information such as the most frequently accessed pages, average view time of a page or average length of path through a site. According to Federico et al (2000), many tools, available, perform this kind of analysis also for free, and its aim is to give a description of the traffic on a Website, like most visited pages, average daily hits, etc.

Association Rules

Association rules capture the relationships among items based on their patterns of co-occurrence across transactions. Given a set of transactions, where each transaction is a set of items, an association rule is an expression of the form X=>Y, where X and Y are sets of items such that no item appears more than once in X u Y. The intuitive meaning of such a rule is that transactions in the database which contain the items in X tend to also contain the item in Y. According to Maja (2011), two common numeric quantifies how often the items in X and Y occur together in the same transaction as fraction of the total number of transactions.

In the reference Kobra (n.d), describes the association rules in context of web usage mining, refers to sets of pages that are accessed together with support value exceeding some specified threshold.

Furthermore explained, in Federico et al (2000) it clearly indicates that these pages (sets of pages) may not be directly connected to one another via hyperlinks. For example, using association rule discovery techniques can help to find correlations such as following.

- 40 % of users visit the web page with URL /home/page1 and the web page with URL /home/page2 in same user session.

- 30% of users, who accessed the web page with URL /home/products, also accessed /home/products/computers.

Further generalized in Bamshad et al, (n.d), the main idea is to consider every URL requested by a user in a visit as basket data (item) and to discover relationships with a minimum support level between them.

Sequential Patterns

This discovers frequent subsequences as patterns in a sequence database, in an important data-mining problem with broad applications, including the analysis of customer purchase behavior, web access patterns, scientific experiments, disease treatments and so on. According to Kobra,E.,(n.d), sequential pattern mining finds all of the frequent subsequences, i.e., and the subsequences whose occurrence frequency in the set of sequences is no less than min_support.

In web server logs, a visit of a user is recorded over a period of time. A time stamp can be attached either to the user session or to the individual page requests of user sessions. By analyzing this information with sequential pattern discovery methods, the web mining system can determine temporal relationships among data items such as the following:

- 30 % of users who visited /home/products/dvd/movies, had visited /home/products/games with in the past week.
- 40 % of users request the page with URL /home/products/monitors after visiting the page /home/products/computers.

Further descriptions in Bamshad et al, (n.d), and generalized as; the attempt of this technique is to discover time ordered sequences of URLs followed by past users, in order to predict future ones.

Clustering

According to Kobra (n.d), clustering is a technique to group together a set of items having similar characteristics. In the web usage domain, there are three kinds of interesting clusters to be discovered: 1st session clusters; 2nduser clusters; 3rd page clusters.

Session clustering implementation allows clustering of user sessions in which users have similar access patterns. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. In reference (Castellano, G., et al, 2007), Page clustering can be partitioned into two methods. The first is to cluster pages according to their contents. For this method an analysis of the content of website is needed. The second method computes clusters of page references based on how often they occur together.

In reference with Robert, C., et al, (1997), generalized as meaningful clusters of URLs can be created by discovering similar characteristics between them according to user's behaviors.

Classification

Classification is the task of mapping a data item into one of several predefined classes Robert et al, (1997), in the web domain, and 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 Maja, (2011), supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc. For example, classification on server logs may lead to the discovery of interesting rules such as:

- 30 % of users who placed an online order in /Product/Music are in the 18-25 age groups and live on the west coast.

2.2.5. Applications of Web Usage Mining

The general goal of web usage mining is to gather interesting information about users navigation patterns (i.e., to characterize web users). This information can be exploited later to improve the website from the users' viewpoint. The results produced by the mining of web logs can use for various purposes:

- To personalize the delivery of web content.
- To improve user navigation through prefetching and caching.
- To improve web design.
- To improve the customer satisfaction.

Personalization of web content

Web usage mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users (Federico et al, 2000), personalized site maps are an example of recommendation system for links.

Prefetching and Caching

The results produced by web usage mining can be exploited to improve the performance of web servers and web-based applications. Lukas (n, d), further explained that typically, web usage mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time.

Support to the Design

Usability is one of the major issues in the design and implementation of websites. The results produced by web usage mining techniques can provide guidelines for improving the design of web applications. Uses output to evaluate the organization and the efficiency of websites from the users' viewpoint.

According to Federico et al (2000), exploits web usage mining techniques to suggest proper modifications to website. Adaptive websites represents a further step in this case; the content and the structure of the website can be dynamically reorganized according to the data mined from the users' behavior.

E-commerce

Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies according to Sulu, (2003). Customer relationship management (CRM) can have an effective advantage from the use of web usage mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure.

2.3.Related works

Web usage mining encompasses studies in which knowledge is obtained through the analysis of web usage. This covers correlations among products or web pages, market segmentation based on user demographics and interests, as well as analysis of a site's success.

The discovery of web usage patterns with conventional mining techniques is proposed in Tianyi, (1995), discover frequently accessed paths by applying a methodology similar to the discovery of association rules organize URL requests into user sessions. Bamshad et al ,(n.d) and then apply association rule discovery and sequence mining to extract correlations among pages Berendt, et al,(2000) propose a similar approach for mining frequent traversal paths and groups of most frequently visited pages Masseglia,et al,(n.d),contribute an approach for mining dynamic databases more efficiently for sequences. However, in Carsten et al., (2000) it has been shown that conventional mining algorithms are not appropriate for the discovery of web usage patterns, because

- ✓ Modeling navigation patterns as associations or sequences oversimplifies the problem and
- ✓ Statistical measures like frequency of access are too simple for navigation pattern discovery.

The different conception of navigation patterns between WUM and other sequence miners is due to the fact that they concentrate on patterns that reflect correlations among events (here: page accesses).

WUM focuses rather on depicting and exploiting the navigation behavior of user groups, in order to improve the website accordingly. Our first results have shown that the model of navigation patterns is appropriate in this context Carsten et al (2000), but also that it must be accompanied by a model that measures and improves success and by a procedure for the mining process. In this study, it presents the complete framework of modeling success and navigation behavior and combining the two to improve the success of a site.

Also apply OLAP technology to analyze web usage Myra, (n.d), for e-commerce applications. The data of interest in this context include not only web logs, but also a concept hierarchy, background knowledge of the expert, as well as previously discovered results. The study reveals the importance of electronically capturing and exploiting data from multiple sources in order to perform web usage mining. However, the work presents no results on how those different information assets are combined during analysis.

The miner proposed in Navin, et al (2010) discovers statistically dominant paths using a methodology for the discovery of association rules. However, the assumptions made on building those paths are rather over-restrictive. For instance, visitors of a web page do not usually visit all children of this page, with the exception of certain application domains like electronically available course material.

The association rules target goal that on discovering all frequent patterns among the transactions, the problem originally initiated by (Agrawal et al) and is based on detecting frequent item sets in the market basket, but in the context of web usage mining, association rules refer to set of page that are accessed together. Usually these rules should have a minimum support and confidence to be valid.

Further explained in Enrique et al (2000), The Apriori algorithm is widely accepted to solve this problem. Association rules can be used to re-structure a website, to find shortcuts, an application especially useful for wireless devices or to prefetch web pages to reduce the final latency the data used to obtain frequent patterns in a web

mining problem has a very important characteristic: it is sequential. The user accesses a set of pages in a given order and it is very important to capture this order in the final model obtained. Unfortunately, the two previous methods lack any kind of representation of this order. Clustering identifies groups of pages that are accessed together without storing any information about the sequence.

Association rules indicate the miner proposed in one of the earliest works in this are discovers statistically dominant paths using a methodology for discovers statistically dominant paths using a methodology for the discovery of a website association rules. The “Foot prints “tool of records the footprints left behind by website visitors and accumulate them into frequently accessed paths. The “PageGather” tool of uses a clustering methodology to discover web pages visited together and to place them in the same group.

The “WEBMINER” tool of (Bamshad.m. et al, (n.d)) provides a query language on top of external mining software for association rules and for sequential patterns. However, the expressiveness of the language is restricted by the input parameters acceptable by the miner to the best of our knowledge, current miners do not support generic specifications on the structure of the patterns to be discovered, e.g. page revisits, cycles etc.

According to Ballman, et al (1997), SpeedTracer is a web usage mining and analysis tool which tracks user browsing patterns, generating reports to help webmaster to refine website structure and navigation. SpeedTracer makes use of referrer and agent information in the preprocessing routines to identify users and server sessions in the absence of additional client side information. The application uses innovative inference algorithms to reconstruct user traversal paths and identify user sessions.

There are some web usage miner tools, which can be used to the navigational pattern discovery for web user behavior of the website, according to Bettina, et al (1999), the two most important tools for navigation pattern are, MiDAS, and WUM tools. The main difference between them are MiDAS designed with the demands of e-commerce application in mind and its commercial products whereas, Carsten et al(2000) the WUM are free source web utilization miners, but both of them are equipped with a mining language.

According to Sulu (2003), the query processor is incorporated to the miner in order to specify characteristics of discovered paths that are interesting to the analyst. Incorporating the mining language early in the mining process allows the construction only of patterns that have the desired characteristic while irrelevant pattern are removed. However, no performance studies were reported and the use of query langue to find patterns with predefined characteristics may prevent the user finding unexpected patterns.

The number of tools and their application a lot of works are done because of it is broad research activity and also the extensive use of the WWW, most widely tools are summarized as by Jaideep, et al (n.d) ,follows with their applications namely general , business ,site modification characterization and personalization.

Project	APPLICATION	DATA Source			DATA Type				User		Site	
	FOCUS	Serves	Proxy	Client	Structure	Content	Usage	profile	single	multi	single	multi
WebSIFT	General	X			X	X	X			X	X	
SpeedTracer	General	X					X			X	X	
WUM[7]	General	X			X		X			X	X	
Shahabi	General			X	X		X				X	
Site Helper	Personalization	X				X	X		X		X	
Letizia	Personalization			X		X	X		X			X
Web Watcher	Personalization		X			X	X	X		X		X
Krishnapuram	Personalization	X					X			X	X	
Analog	Personalization	X					X			X	X	
Mobasher	Personalization	X			X		X			X	X	
Tuzhilin	Business	X					X			X	X	
SurfAid	Business	X				X	X			X	X	
Buchner	Business	X					X	X		X	X	
WebTrends,Hitlist ,Accurue,etc	Business	X					X			X	X	
WebLogminer	Business	X					X			X	X	
PageGather,SC	Site Modification	X			X	X	X			X		X

[7] The WUM(web utilization miner) are going to implement for web usage navigational pattern in the paper

ML												
Manley	Characterization	X				X	X			X		X
Arlitt	Characterization	X				X	X			X		X
Pitkow	Characterization	X		X		X	X			X		X
Almedia	Characterization	X					X			X		X
Rexford	System Improve	X	X				X			X	X	
Schecher	System Improve		X				X			X	X	
Aggarwal	System Improve		X				X			X	X	

Table 2: Web usage mining research projects and products.

CHAPTER THREE: METHODOLOGY

3. Overview

According to Jaideep, et al (2000), web usage mining has three main processes to reveal knowledge out of the data warehouse or log file (see figure 4). As discussed in chapter two, there are various tools and algorithms for mining web logs.

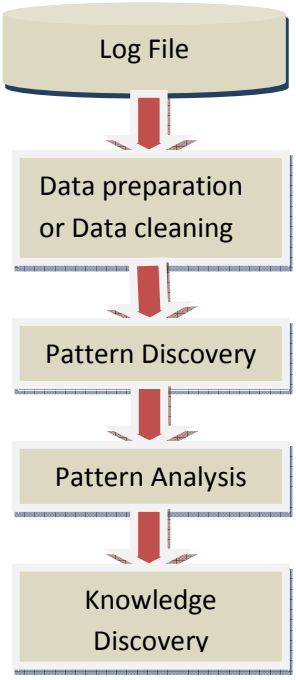


Figure 4: web mining usage main process to discover knowledge.

Description of the model

A. Log file:

A web server log is an important source for performing web usage mining because it explicitly records the browsing behavior of site visitors. According to Jaideep, S (n,d), the data recorded in server logs reflects the (possibly concurrent) access of a website by multiple users. Those logs files can be stored in various formats such as common log or extended log formats. An example of log Extended log format is given in Appendix I. In

this paper, an extended log formats are implemented i.e. collected from the web server of Addis Abba University that official hosts the web site of Addis Ababa University.

B. Data preparation:

Data preparation or data cleaning is an important step that must be undertaken. The data cleaning is a set phases, those phases have stages such as removing irrelevant requests, removing duplicate requests, identifying users and creating sessions are among those phases. In this paper all those steps will be employed.

C. Pattern Discovery:

In navigation pattern discovery, *all* patterns that have certain properties, such as a minimal frequency within the whole population. In sequence mining, pattern discovery is usually subject to constraints on minimal frequency and maximal length. In particular, the mining language of WUM, MINT, supports the specification of “templates”.

A template is a vector of variables and wildcards, and is accompanied by constraints on the statistics and content of the events (here: page occurrences), to which the variables can be bound during mining. Similarly, the wildcards pose structural restrictions on the g-sequences that match the template.

D. Pattern Analysis:

Pattern analysis is the last step in the overall web usage mining process in, challenge of pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL or MINT query (see in Appendix VII for syntax of MINT). According to Dietmar, et al (n.d) there is another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data.

3.1. Tools Selections

As it has been described in chapter TWO, a lot of tools that exits for preparing a dataset for various purposes, but the selection of tools is not easy, since every tool has designed for specific purpose, however most of tools cannot able to produce quality outputs unless combine various tools in order to yield efficient and effective results.

The researcher of this paper selects the two major tools **(**WUMprep) and WUM (web utilization miner) to meet the objective of the research i.e. navigation behavior of the web users. The explanation of the why those tools are selected, given below.

3.1.1. Tools for Log preparation

The researchers choose the WUMprep tools because data preparation using WUMprep scripts is a straightforward and efficient one time procedure that prepares the data. Its primary purpose is to be used in conjunction with the web usage miner WUM, but WUMprep might also be used standalone or in conjunction with other tools for Web log analysis.

Therefore, the researchers of this paper no need to implement other data preparation mechanisms, besides the navigational tool i.e. WUM (web utilization miner) combative with the data processing tool (WUMprep).

3.2. Data cleaning

3.2.1. Removing Irrelevant Requests and Status

Removing of irrelevant record is significant. As those requested log files not only contain requests to the pages comprising the website, but also requests of images, scripts etc. embedded in these pages. Therefore, removing those secondary requests is significant. The table below shows the list of irrelevant extension requests within the log file.

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

Table 3: Irrelevant requests, (Extension of URL).

```
Read record in database (Web mining base).
For each record in database
Read fields (URI – stem) //URI- stem indicates
The target URL//
If fields = {*.gif,*.jpg,*.css,*.jpeg,*.ico} then
Remove records
Else
Save records
End if
Next record
```

Figure 5: Algorithms for removing irrelevant requests

In addition to the above, the researcher only interested to filter out requests that are only successful requests which mainly show users who got what they want, not what they did not. Because these requests do not indicate effective browser activity of the users who had been visiting the website.

Both “200” status from the log HTTP records and request message which contains “GET” only filtered out. Figure below shows an algorithm that extracts both GET and status 200.

```
Input: Web Server Log File
Output: cleaned file
Step1: Read LogRecord from Web Server Log File
Step2: If(LogRecord.mehod=’GET’) AND
        (LogRecord.Sc-status<> (301,404,500)
        then Insert LogRecord in to new file.
        End of If condition.
Step 3: Repeat the above two steps until eof (Web Server     Log File)
Step 4: Stop the process.
```

Figure 6: Algorithms that extracts GET and status

3.2.2. Removing Robots

To distinguish between human users and hosts that are robots, there exist several heuristics. They are implementing using script. Firstly, all records containing the name against “robots.txt” which contains list of known robots, in the requested resource name (URL) are identified and straightly removed from the original log files.

3.2.3. Removing Duplicate Requests

If a network connection is slow or a server's response time is low, a visitor might issue several successive clicks or requests on the same link before the requested page finally appears in his browser. Those duplicate requests are noise in the date and should be removed. The researcher used the most widely accepted threshold of 2 seconds between two consecutive requests to eliminate redundant requests (Jose et al (2007).

3.2.4.Session

A session is a contiguous series of requests from a single host (in context of web usage mining, a session requested of series pages order in time). Multiple sessions of the same host can be divided by measuring a maximal page view time for a single page. Session which is computed by taking any URL time stamp , in this paper implemented the most accepted time threshold which is 1800 sec or 30 min to identify the sessions using the these timestamp, Sulu, G.,(2003).

L : the set of input logs.
|*L*|: the number of input logs
D*t* : time interval (1800 sec or 30 min)
S : The set of sessions
|*S*| : The number of sessions.
Input: *L*,|*L*|, D*t*
Output: *S*, |*S*|
Function Log-Parser (|*L*|, *L*, D*t*)
For each *Li* of *L*
If METHOD***i*** is 'GET' and URL**i** is 'WEBPAGE' //1.if
If $ *Sk* Î Open_Sessions with IPk = IP***i*** then //2.if
If ((TIMEi-END_TIMES(Sk))< D*t*) then // 3.if
Sk = (IPk, PAGEk ***U***URL***i***)
Else
CLOSE_SESSION(Sk)
OPEN_SESSION(IP**i**,URL**i**)
End if //end of 3.if
Else
OPEN_SESSION(IP**i,**URL**i**)
End if //end of 2.if
End if //end of 1.if
End for

Figure 7**:** Algorithm for session creation

3.3. Divide log format

The preprocessed data needs to be divided into manageable size before introduced into WUM because the size of those log files are large and it takes long time to process the data. The researcher implemented a Python code to prepare the processed data for the WUM.

3.4. Tool Selection for Navigational Behavior

According to Anália et al., (2003), navigation pattern discovery performed on the portion of the web server log that contains the sessions. The discovered patterns reflect the desired behavior of the visitors; these patterns are then uses as a basis to analyze.

According to Bettina et al (1999), web utilization miner, have two major modules: the Aggregation Service prepares the web log data for mining and the MINT-Processor does the mining. Further explained, the Aggregation service extracts information on the activities of the users visiting the website and groups consecutive activities of the same user into a transaction. It then transforms transactions into sequences. Its major task is to merge those sequences into a trie structure, on which aggregated statistical information is retained. According to Marya, et al (n.d), Aggregation Service assumes that accesses from the same host come from the same visitor.

Aggregate Trees: The Aggregation service of WUM extracts the visitor trails from the web log and aggregates them by merging trails with the same prefix into a tree structure, the “aggregate tree.” An aggregate tree is a trie, a node of which corresponds to the occurrence of a page in a trail. Common trail prefixes are identified, and their respective nodes are merged into a trie node. This node is annotated with the number of visitors having reached the node across the same trail prefix. It known as “support” of the node.

The MINT-Processor mines the aggregated data according to the directives of the human expert further described in Marya, et al (n.d), “MINT” is the mining language serving as interface between the user and the miner. The expert uses MINT to instruct the miner on the formulation of the output, and, most importantly, on the interestingness criteria to be satisfied by the desired patterns.

According to Bettina,et al , (1999), generalized description like “The MINT-Processor is responsible for identifying common patterns in the large aggregate tree of the aggregated log, merging them to aggregate graph objects, computing the node supports and evaluating the query predicates”.

Besides to the above, the following points could be taken as a reason why the researcher selected the WUM as tool for navigational tool.

- It has designed to work with the WUMprep module (which is responsible for the pre-process phase).
- It is free open source tool.
- WUM has mining language (MINT query) which serving as interface between the user and the miner for filtering the interestingness pattern to be satisfied by the desired patterns, (It is also open source tool).
- WUM, uses for the discovery of navigation patterns and visualization of interesting Patterns.
- It is a sequence miner Myra,S.,(2000), and it can generate comprehensive statistical report regarding the web log in better way so that it can be used as in put for other tools for better visualization e.g. Microsoft EXCEL.

Generally, WUM is a sequence miner, a mining system for the discovery of interesting navigation patterns. Further explained in Marya et al, (n.d), its purpose to analyze the navigational behavior of users in a website and discover navigation patterns in the form of graphs. It discovers patterns comprised of events that are not necessarily adjacent and satisfying user-specific criteria is a mining system for the discovery of interesting navigation patterns.

CHAPTER FOUR: EXPERIMENT AND FINDINGS

4. Over view of Experiment setup

Based on the methodology discussed in chapter 3, the experiment setup conducted on the following set up;

- **Computer Type:** *personal computer (X32-based PC).*
- **Operating system:** *Microsoft window 7 ultimate edition.*
- **Processor:** *Intel (R) Pentium (R) Dual CPU T3200 @2.00GHZ 2.00GHZ.*
- **Web mining tool:** *web utilization miner (WUM7.0 the latest version).*
- **Supported tools**: Java version 1.5 (WUM java based tool).
- **Programming Language**: *Perl (*WUMprep suit of Perl script).
- **Python code:** To divide clean web log into manageable size.

4.1. Data Collection and Selection

As it has been pointed out earlier in chapters, the web log records on official website of AAU was source of data for research. Web data log were collected from the web server of the university's official website. This web server stores different kind of log formats such Access.log, SSL request logs, and SSL error logs. Access log records for both months of November and December were used for the analysis, as they were available at the ICT office.

4.2. Experiment

4.2.1. Data Cleaning

The data collected from the AAU web server logs like any other logs are full of junks, noises, as well as robots (spiders, crawlers). Those should be removed to have clean web logs to achieve appropriate, efficient, effective data logs. Data cleaning is not a single step, its set of steps or phases; the logs need to be cleaned through some preprocessed steps because it is crucial steps to truck down the users' behavior of the official website. The sample log data that are extracted from AAU before any preprocessing steps taken place.

```
66.249.65.124 - - [28/Nov/2010:04:26:35 +0300] "GET
/index.php/global-text-project HTTP/1.1" 200 22916 "-"
"Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
66.249.65.87 - - [28/Nov/2010:04:26:37 +0300] "GET
/index.php/component/events/view_month/2009/06/01?catids=97
HTTP/1.1" 200 38809 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
66.249.65.104 - - [28/Nov/2010:04:26:43 +0300] "GET
/index.php/component/events/view_week/2011/04/26 HTTP/1.1" 200
28388 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
```

Figure 8: A small extracted raw web server log content from AAU.

In the above figure 8, it can be observed that log file need to be processed and steps for preprocessing will be discussed in the following sections.

4.2.2. Removing Irrelevant requests

Removing irrelevant request lead to diminished the original size of raw log files. Those repeated requests that may come within requests. For example, the original raw log records size were 50,701 KB records and after the log filter script employed the preprocessed data became to 12,416 KB. In WUMprep suit, the script **logFilter.pl** is designed to perform this task for removing irrelevant requests.

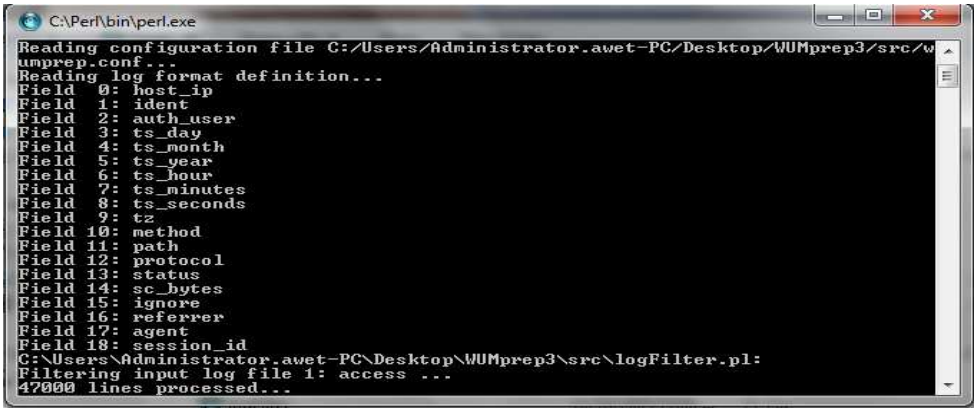


Figure 9: removing irrelevant records sample.

4.2.3. Detection of Robots

The process of detect robots are very important to elimnate the irelevent records whch are caused by misuses that comes from other resources like (speider ,web crawlers). In other words , web surfs that are too fast for ordirnary pepole to do one belived to came from speiders or crawlers. In the expertment, the detection of robots employed after removing irrelevant requests that runs against to “index list“ , which “index list“ contains known list of robots. The number of robots that were detected from the web server logs are shown below,for the detail of robosts see in Appendix VII. In WUMprep suit,they are implemented in the script **removeRobots.pl**,and give the follwing results.

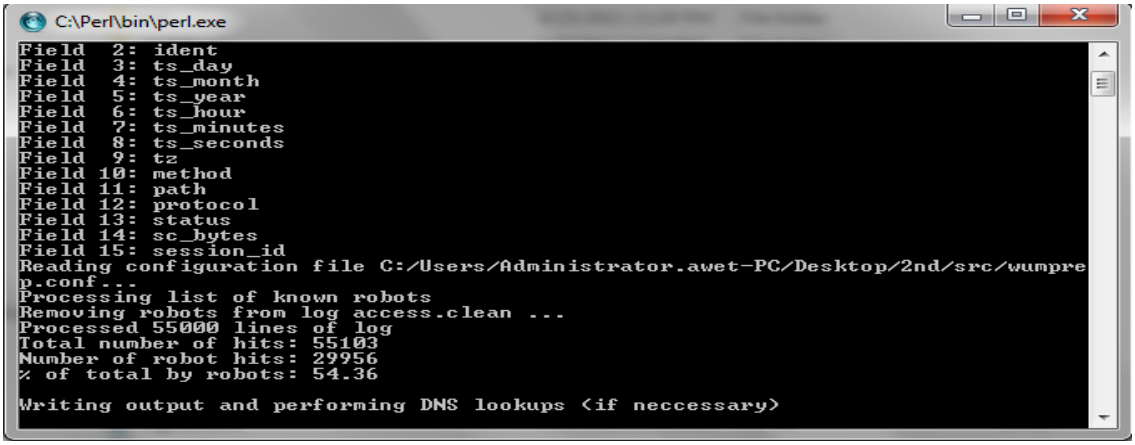


Figure 10: sample removing of robot hits

According to the experiment conducted for one week in December, the numbers of robots inside the log format were 54.36 % of the total hits (see figure 10). For the months of November, the total numbers of robots against the total hit were 39.68 %. Figure 11 shows a Sample of robot log lines that resulted after preprocessed of log filter:

```
208.115.111.247 - - [05/Dec/2010:05:03:20 +0300]
"GET /robots.txt HTTP/1.1" 200 --304 "-"
"Mozilla/5.0 (compatible; DotBot/1.1;
http://www.dotnetdotcom.org/,
crawler@dotnetdotcom.org)" (robots.txt)
208.115.111.247 - - [05/Dec/2010:05:03:21 +0300]
```

Figure 11: Sample robot log files.

From the above results (figure 11), for example, it can be observed requests that came from same IP address that is (`208.115.111.247`) within two seconds, (see the shadow) from the field of date. [05/Dec/2010:05:03:20 +0300 and 05/Dec/2010:05:03:21 +0300), should be removed because those requests do not show an ordinary human behavior.

4.2.4. Session

The session is performed after the detection of the robots and gives the following results as shown below. In the WUMprep suite, **sessionize.pl** is the script that supports this task.

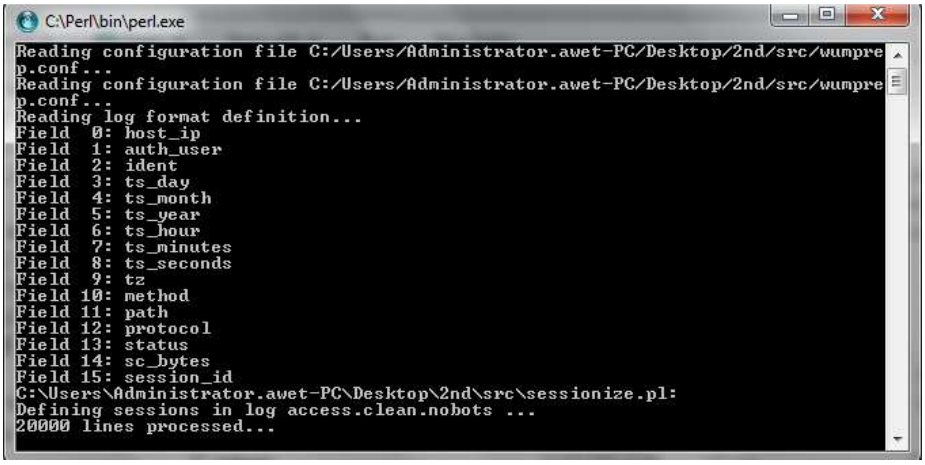


Figure 12: sample-session process

The session.pl creates number of sessions, according to my experiment the number of sessions created were about 23,411. Some log lines which exceed from the time threshold i.e. 1800 sec or 30 min are removed. See the details in Appendix VII.

```
245208:1|10.90.10.28 - - [28/Nov/2010:04:27:21 +0300] "GET /index.php/library-and-museum/library HTTP/1.0" 200

245208:2|10.6.13.66 - - [28/Nov/2010:04:31:19 +0300] "GET / HTTP/1.0" 200

245208:3|207.46.13.93 - - [28/Nov/2010:04:34:39 +0300] "GET /index.php/academics/schools/348-schools?tmpl=component&print=1&page= HTTP/1.1" 200

245208:4|68.52.248.143 - - [28/Nov/2010:04:35:21 +0300] "GET / HTTP/1.1" 200

245208:2|10.6.13.66 - - [28/Nov/2010:04:41:19 +0300] "GET / HTTP/1.0" 200
```

Figure 13: Sample common log format after Session.

From figure above, it can be observed that only a “200” and “GET” filtered out after creating session to indicate that only successful requests from the website users. It can also be noticed that sessions also created.

Generalized Reports on Preprocessing

Since every activity the official website users are recorded on web server log file. After some significant preprocessing steps, the number of records reduced in substantial manner (see in the following table). Not all those records able to show ordinary behavior of users’. The preprocessing was undertaken for both months of December and November.

Original log entry records per day	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
220,340	150,127	70,564	25,005	25,005
230,087	160,743	72,087	24,060	24,060
200,406	148,906	63,480	21,000	21,000
190,967	138,967	50,653	19,734	19,734
200,190	178,300	60,752	20,674	20,674
200,150	167,543	47,897	19,653	19,653
220,205	120,950	62,096	23,765	23,765

Table 4: A Sample records for the week in December after undertaken the preprocess phases.

Note:*T*he cleaned common log format cannot be directly fed into the WUM they must be dividing for manageable size, using the python code.*

As it has been mentioned earlier in this chapter, the log files are contains irrelevant data, irrelevant records, and noises as consequent it can be discovered from the above experiment table that, the size of original log entry records decreased in average of 80 % for the months of December. For the month of November, the size of records of original entry decreased in average of 73 %. See in the appendix VI.

4.3. Navigational Behavior of December

4.3.1. Aggregated LOG tree

An aggregated tree is created after preprocessing completed on the raw web log. Those log files which are directly feed into WUM after divide in to manageable size using the Python code. The aggregated tree for the months of December as follows.

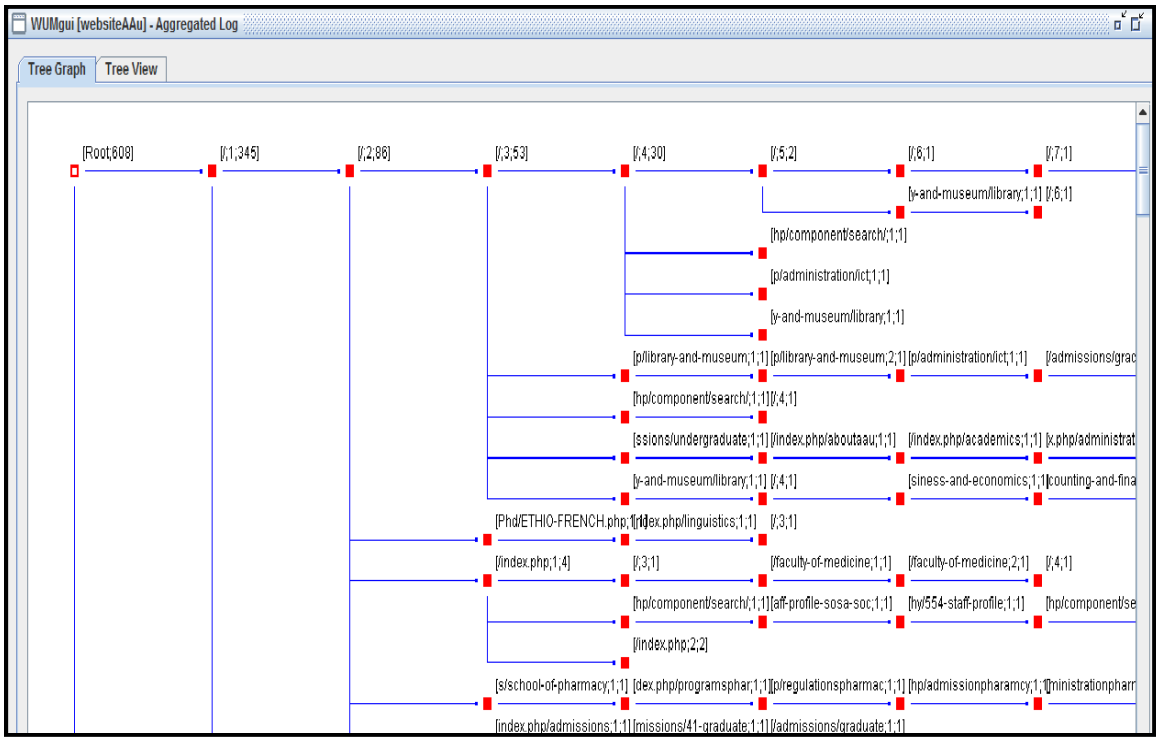


Figure 14: Sample aggregated tree for the month of December

As it can be seen from the above figure 14, an aggregated log tree that, the total numbers of nodes or the total traverse made by users' were 7,225 for one week in the month of December. Based on the aggregated log tree the MINT queries were applied to find interesting patterns.

4.3.2. Sequence and Navigational Discovery of Users

As it has been previously mentioned in chapter three, the generalized sequence pattern describes users the behavior by filtering out the interesting pattern. The experiment is done using most interesting patterns specified using the MINT query.

Sequence analysis 1: Where do visitors of page HOME go afterwards?

Explanation of the query

In this query, it specify a template t with two variables “a”, “b”, thus seeking for two pages bound to “a” and “b” and at most 5 arbitrary page occurrences in between denotes that “a” should be bound to the first page which is /index.php/home and at least visited (confidence) 20 % occurrence in a session.

```
select t
from node as a b, template a [1;5] b as t
where a.url = "/index.php/home"
and (b.support / a.support) > 0.2
```

The above query have resulted a patterns (see in the following figure).

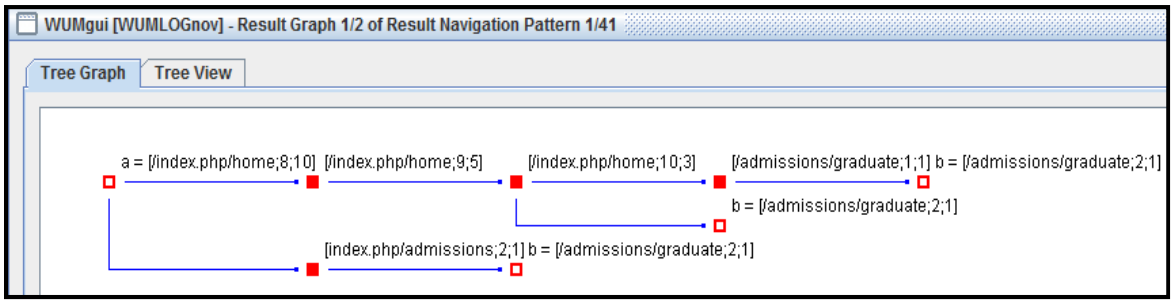
Type of Results: (•) Complete Patterns ( ) Partial Patterns

Pattern	Variable	URL and Occurrence	Abs. Support	Confidence
1	a	/index.php/home; 8	10	1.0
1	b	/index.php/admissions/graduate; 2	3	0.3
2	a	/index.php/home; 13	1	1.0
2	b	/index.php/academics/faculties/faculty-of-medicine; 6	1	1.0
3	a	/index.php/home; 13	1	1.0
3	b	/index.php/registrar; 3	1	1.0
4	a	/index.php/home; 10	4	1.0
4	b	/index.php/admissions/graduate; 2	1	0.25
5	a	/index.php/home; 3	87	1.0
5	b	/index.php/home; 5	27	0.3103448275...
6	a	/index.php/home; 4	49	1.0

Figure 15: navigation pattern

Here, we have received all pages reached within five pages after HOME (index.php/home). Which has been accessed 100 or more times, provided that those pages have been accessed by at least 20 % or 100 % of the visitors visiting HOME, but as it have been discovered most accessed pages is /index.php/academics/faculties/faculty-of-medicine, /index.php/registrar users stay 100 % visiting the content of it. Naturally, users' who have visited 100 % home page also visited the /index.php/admissions/graduate for 30 %, (Berendt et al (1999)).

Navigation pattern:



As it can be seen from the navigation pattern figure, most of users were visited the page of /admissions/graduate after it have been visited the home pages, discovered that most users stay in the HOME page (/index.php/home) and navigated between the home and admission pages, finally to reach the target pages.

Sequence analysis 2: Find out pages that always visit together and look at its pattern.

Explanation of the query

In this query, we specify a template "t" with two variables "a", "b", thus seeking for with two pages bound to "a" and "b" and at most 5 arbitrary page occurrences in between denotes that "a" should be bound to the first page which is /index.php/home. This page should be visited at least 100 % and b page should be at least visited 20 % (confidence) occurrence in a session.

```
select t

from node as a b, template a [1;5] b as t

where a.url = "/index.php/home"

and a.support > 100

and (b.support / a.support) > 0.2
```

The previous issued query results one patterns .

Type of Results: Complete Patterns Partial Patterns

Pattern	Variable	URL and Occurrence	Abs. Support	Confidence
1	a	/index.php/home; 2	170	1.0
1	b	/index.php/home; 4	38	0.2235294117...

Here, we received all pages where "a" is 2nd entry, which has been accessed 100 or more times, provided that "b" has been accessed by at least 22 % of the visitors visiting "a" and "b" has been accessed 22 %, those two URL are accessed together.

Navigation pattern

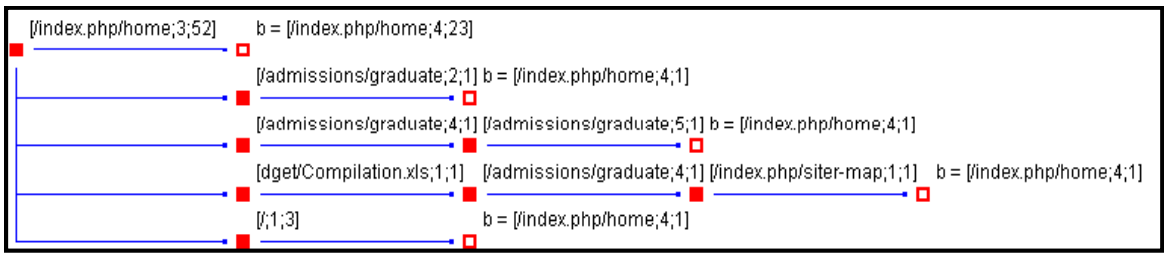


Figure 16 : Navigation pattern

From the above figure 16, we have received that when visitor start at looking at /index.php/home page, 50 % of them stayed within this subject area that is related to the admission and graduate.

GSP analysis 4: Which paths do visitors take to read blogs?

In this query, we specify a template "t" with two variables "a", "b", thus seeking for with two pages bound to "a" and "b" and at most five arbitrary pages. Occurrences in between denotes that "a" should be bound to the first page, which is /index.php/home, page should be visited at least 20 % and "b" page should not be visited in the sessions.

```
select t from node as a b c, template a _ _ b [0;0] c
as t

where c.url = "/index.php/view-blog"

and b.url != "/index.php/view-blog"

and (b.support / a.support) > 0.2
```

Figure 17: Query to identify where users' go after read blogs.

Pattern	Variable	URL and Occurrence	Abs. Support	Confidence
1	a	/; 5	349	1.0
1	b	/; 8	73	0.2091690544...
1	c	/index.php/view-blog; 1	1	0.0028653295...
2	a	/aau_staff_load/EnterLoadInfo.php; 8	2	1.0
2	b	/index.php; 2	1	0.5
2	c	/index.php/view-blog; 1	1	0.5

The output of the above issued query, resulted two patterns ,here we recived most users reaching the page /index.php/view-blog pages after users stay 100% in the page of root page (/) and /aau_staf_load/enterLoadinfo.php ofcourse, some users stay 20% and 50 % respectively stayed in the home page before reached to /index.php/view-blog pages.

G-sequance:

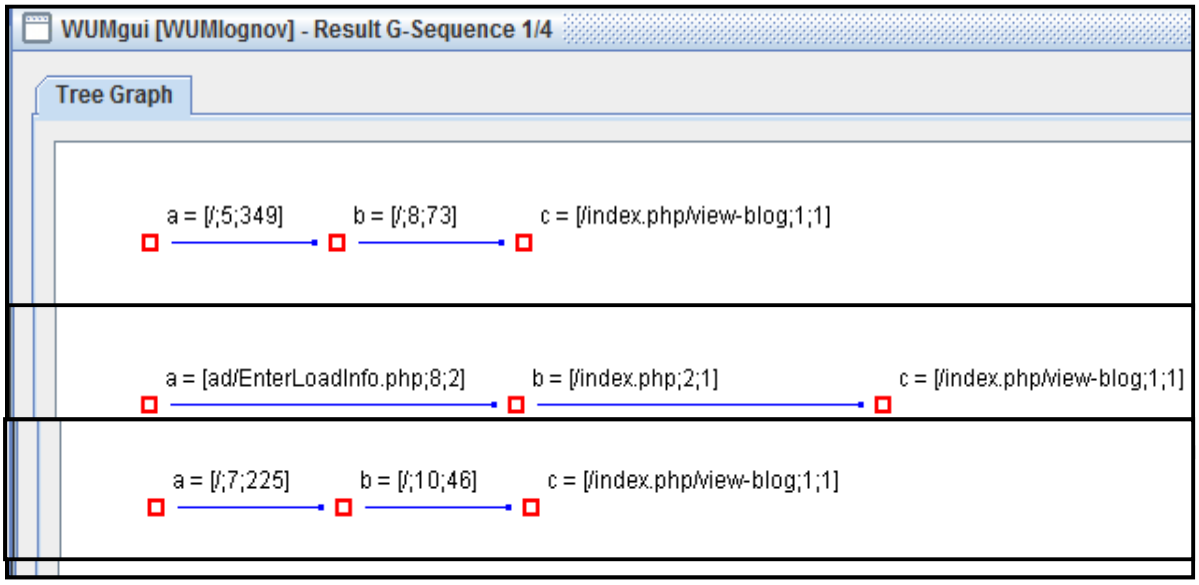


Figure 18: pattern flow

The users did not take a single paths to reach to /index.php/view-blog most of the users took a path from the root pages, and the second most users took to reach using /aau_staff_load/EnterLoadinfo.

GSP analysis 3 :Where do visitors go after search page of AAU pages?

In this query, we specify a template t with three variables a, b, thus seeking for with two pages bound to “a” and “b” occurrences in between denotes that “a” should be bound to the first page which is /index.php/home. “b” page should be at least visited 15% . Page “c” (confidence) occurrence is at least 30% a session.

```
select t
from node as a b c, template
a [0;0] b [0;0] c as t
where a.url = "/index.php/component/search"
and a.support > 10
and (b.support / a.support) > 0.15
and (c.support / b.support) > 0.30
```

Figure 19: Query to issued where visitors go after search engine of AAU

Pattern:

Type of Results: (•) Complete Patterns ( ) Partial Patterns

Pattern	Variable	URL and Occurrence	Abs. Support	Confidence
1	a	/index.php/component/search/; 5	53	1.0
1	b	/index.php/component/search/; 6	21	0.3962264150...
1	c	/index.php/component/search/; 7	12	0.2264150943...
2	a	/index.php/component/search/; 6	31	1.0
2	b	/index.php/component/search/; 7	16	0.5161290322...
2	c	/index.php/component/search/; 8	7	0.2258064516...
3	a	/index.php/component/search/; 1	431	1.0
3	b	/index.php/component/search/; 2	145	0.3364269141...
3	c	/index.php/component/search/; 3	66	0.1531322505...
4	a	/index.php/component/search/; 7	23	1.0
4	b	/index.php/component/search/; 8	10	0.4347826086...
4	c	/index.php/component/search/; 9	7	0.3043478260...

All the above patterns showed that users' do know where they are looking for. Most of users who stayed in search engine 100% and stay in this page for average of 40%, they do search function stay within search the page.

G-sequence pattern

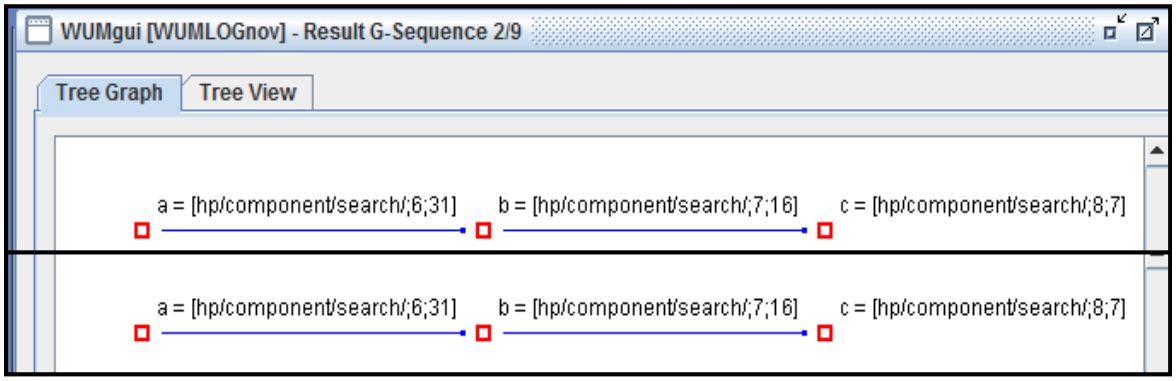


Figure 20: G-sequence result

From the above of the figure, the researcher do not need to put all the g-sequence from the navigation pattern that users stay in the search page as we can see from the above result, that users stay in the search page.

Navigational between two pages

Only patterns starting at a node with support at least 40 % are of interest. One URL is explicitly excluded (index.php). Namely X*Y, shows the second part Y*. Our visualization module currently displays patterns as trees; this is why X*Y is a tree, all leaf nodes of which refer to the same page. This page is the value bound to the variable Y.

```
select t
from node as x y,
template # x * y * as t
where x.url != "/index.php"
and x.support > 40
and y.url = "/index.php/academics"
```

Figure 21: Query issued to find navigational pattern

The above query results the following navigational tree,

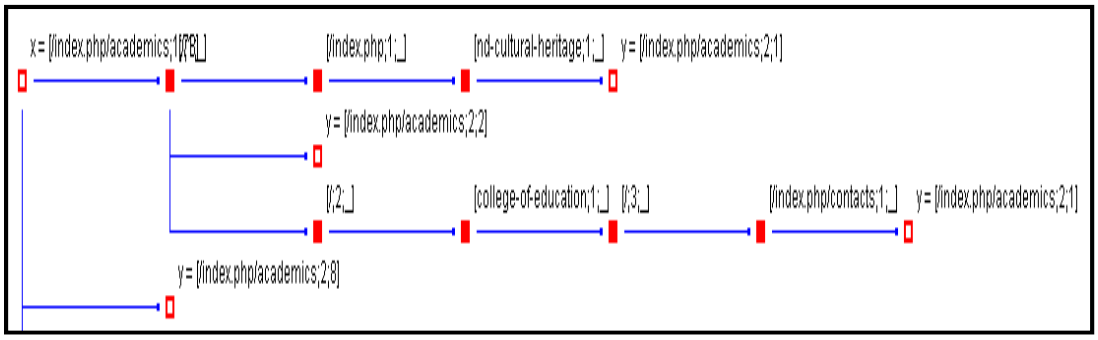


Figure 22: navigational tree

From the above figure that most users who use the academic pages do not leave to other non-academic pages which is not related to their filed, whether stays at this page or leave the website.

4.4.Statistical Analysis for the Months of December

The WUM generated a comprehensive report in web format; the researcher employed other tool like (Microsoft Excel) for better visualization. The report will be discussed in the following sections like, what are, most requested pages, most visited pages, and most visited directory as well as most referee pages for the month of December will be discussed.

4.4.1. Most requested pages

The following table shows the top ten most accessed pages during the months of December. For the rest of the month, see in appendix I.

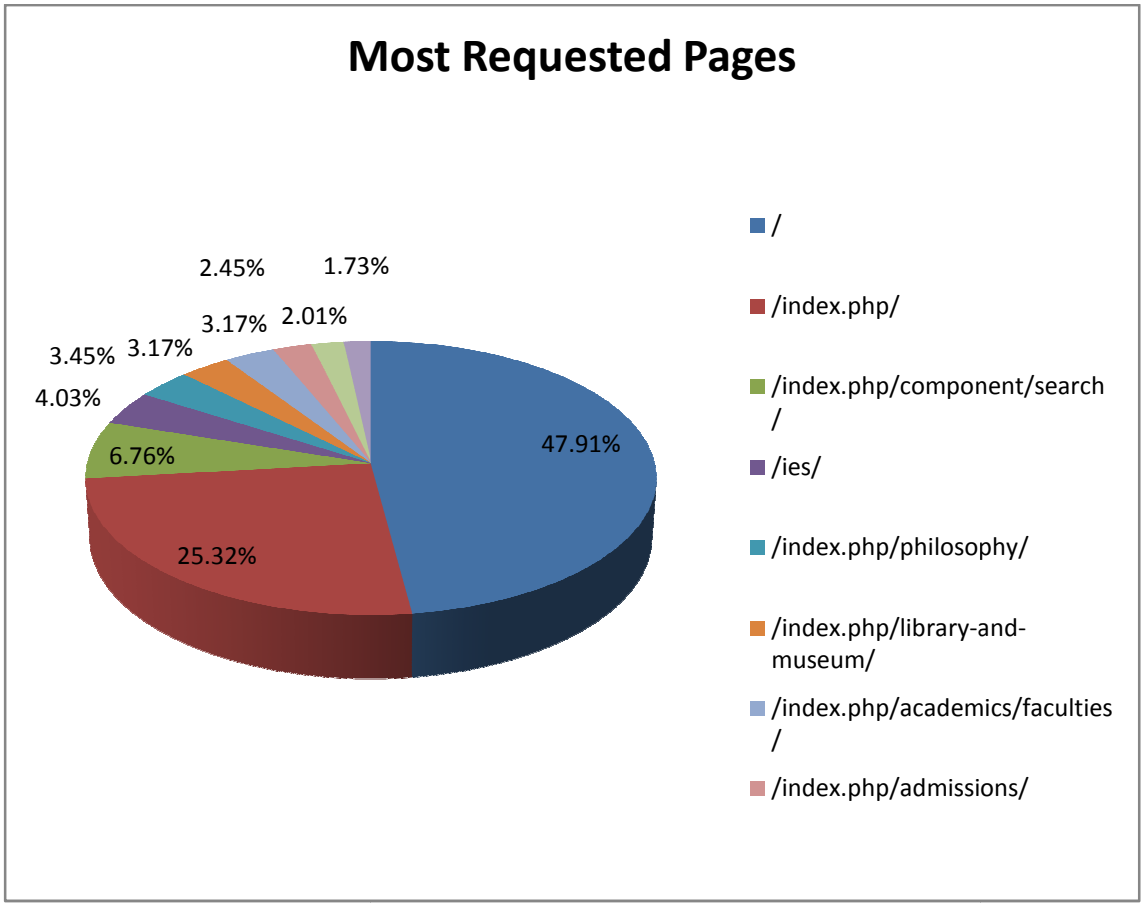


Figure 23: Top 10 most requested pages.

As above figure shows, the top ten most requested pages for the months of December, the most requested pages were root "/" or www.aau.edu.et pages followed by /index.php/component/search/ and the page /index.php/library-and-museum .

This is reflection of that, the /index.php, page was most popular page by most users in all the three months. In fact, this seems to show that most visitors have entered into the site directly or typing the website address. The search engine of the Addis Ababa University page were the second most requested pages which followed by the /index.php/library-and-museum pages.

4.4.2. Most visited directories

The root directory "/" is the most accessed directory where the root directory located in root folder. Most users also interested on the contents under the /index.php/ folder. It is also possible to infer from the result that /index.php/component/search/ the third most visited directory. It would be better if we put the notice and advertisement on those directories (see in figure 24). For the month of November see in Appendix IV.

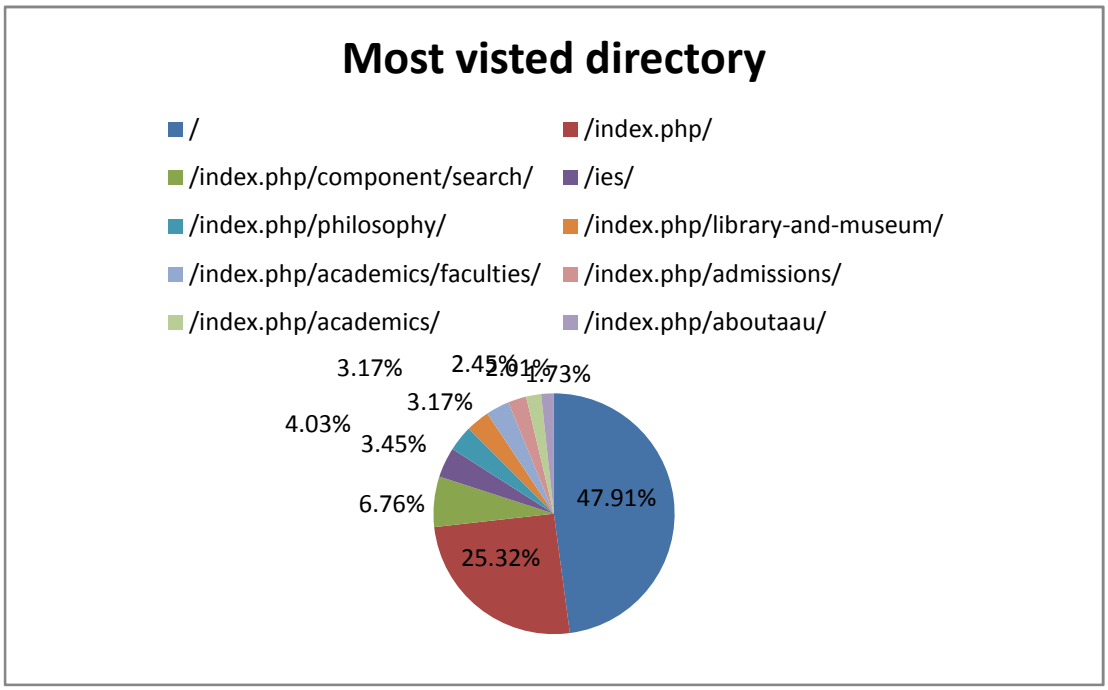


Figure 24: Top ten requested directories

4.4.3. Most Top Entry Pages and Top Exit Pages

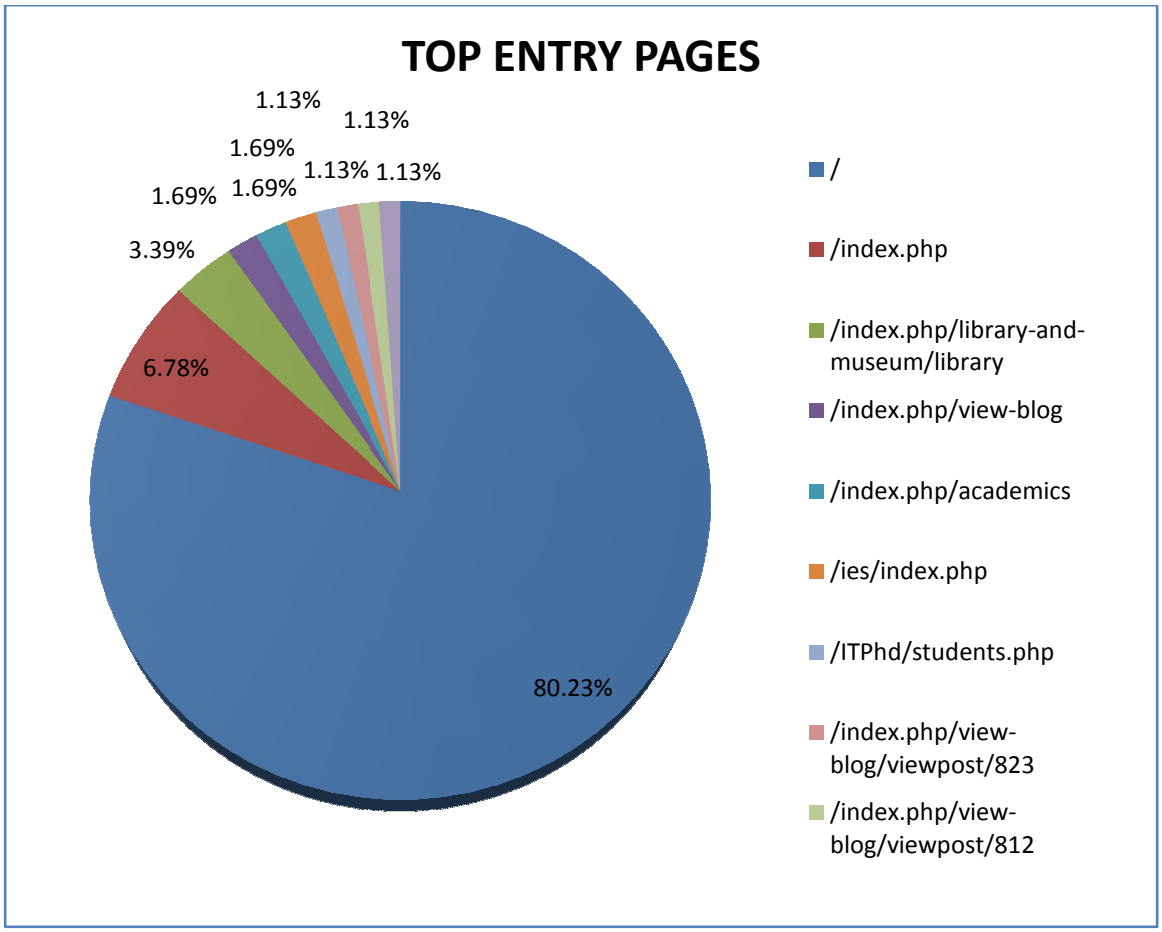


Figure 25: Top ten entry pages

The entry pages are pages that indicated that the website users frequently visited where as the top exit pages is the last visited pages. As shown figure below, what it can observed were the "/" root pages accessed more than any other pages, almost it covered more than half of all request (80.23 %) come from the "/". The second most top entry page /index.php and last not least, the /index.php/library-and-museum/library were third most top entry of pages, followed by /index.php/view-blog and /index.php/academics the 4th and 5th top entry pages respectively. For the month of November see in Appendix III.

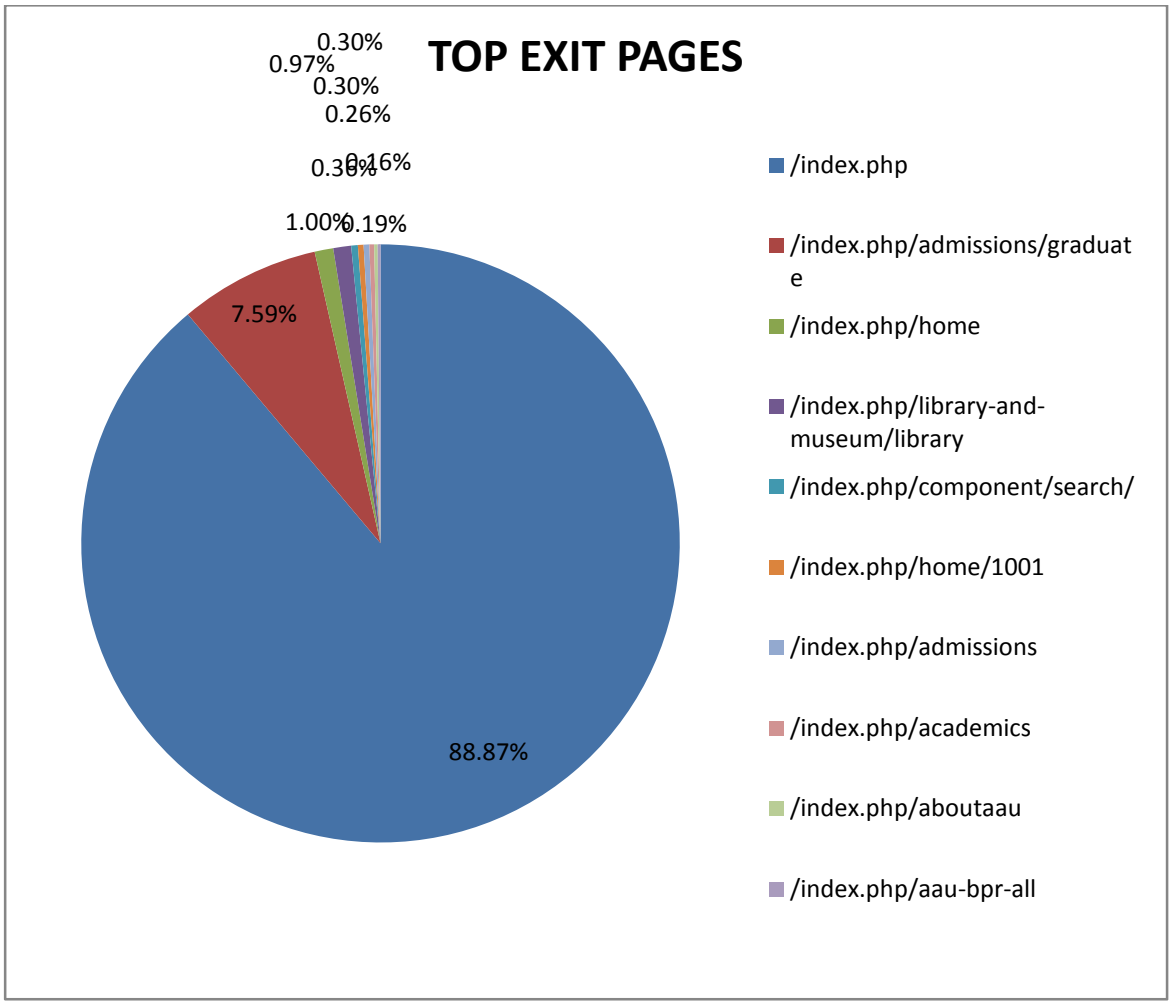


Figure 26: Top most exit pages.

From the above the figure, it can be seen that the top exit pages are the "/" ,it seems that after the user type the website address and leave the website without making any click steams. The second most exit pages are /index.php and last not least, the 3rd most exit page are /index.php/component/search/. For the month of November see in Appendix V.

4.4.4. Top Referrer Pages

The top referee pages are pages where the visitor was last located before making the next request with in the official websites.

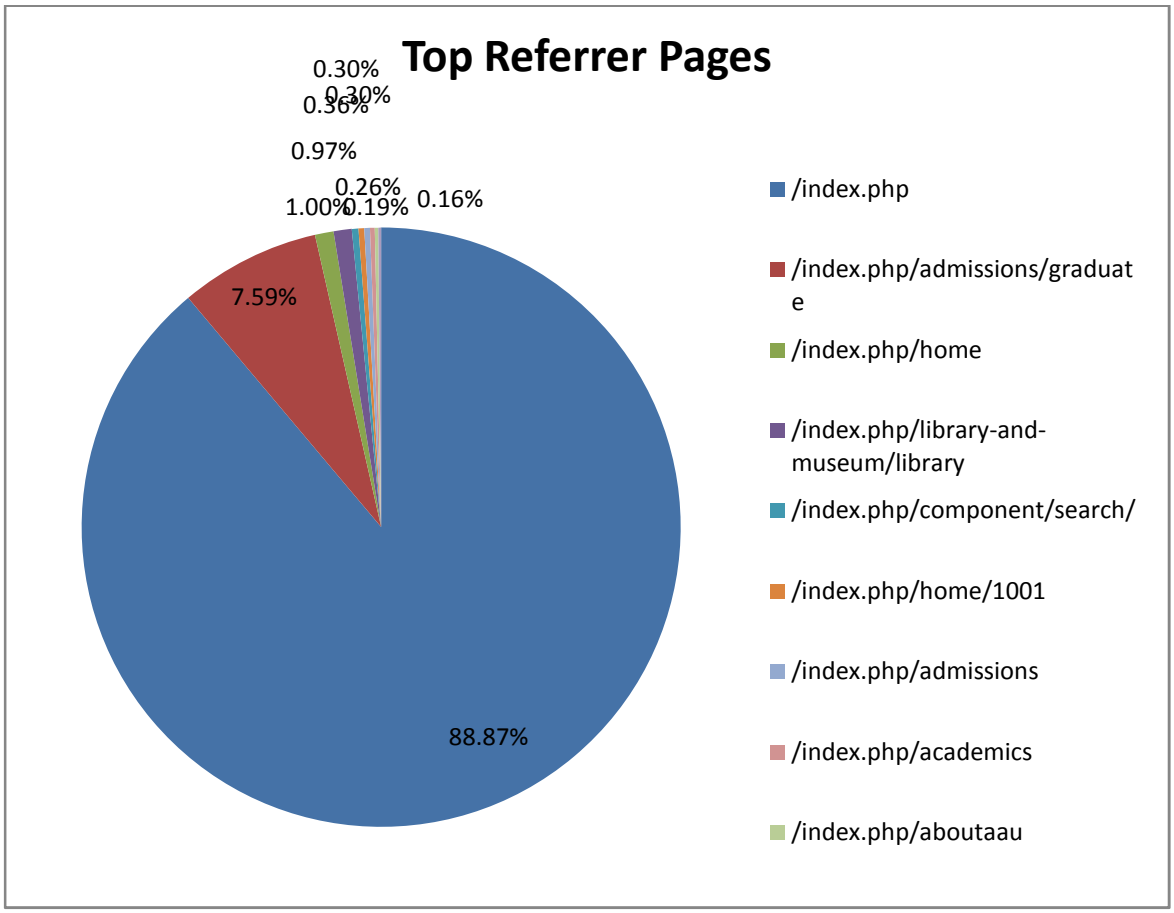


Figure 27: Top Ten referee pages.

From the above figure, what it can be observed that most users made the next request from the page of /index.php, which it covered in more percentage 88.8 %. The next popular pages, where users request the next page are initiated from /index.php/admission/graduate and it covered 7.33 %. The third most referee pages are /index.php/home, which covers the percentage of 0.97 %.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

Conclusion

In this research it was attempted to study the web usage on the user's behavior of the official website of AAU from log data.

The raw log data preparation was a major task of the research owing to the fact that the raw log contains non-human requests, automatically downloaded graphics, or failed requests, moreover, the unprepared log could not be used for navigational behavior pattern discovery in hand .Next the log, WUMprep was applied in those raw log files to filter appropriates fields from the log file. Then WUM (web utilization minter) was employed on the cleaned log data then interesting pattern could be identified using the MINT query.

The result show that, most of the users of Addis Ababa university official website start navigation at the home page of the function i.e. most of the users directly write the full URL of the home page into the address bar or access the home page from the search engine. Most of the users stop their visit at the same page .It infer that the website is not user-friendly for the website and the users opt to stop their visit .

From the navigational behavior of users, most users spend their time on the home page naturally; there are also other pages where users spent the same amount of time as of the homepage like medicine, registrar pages. This may be became most users are interested in getting the registrar pages information. Noticeably, users also leave the home page without spending time. Users do not also spend much time on the graduate pages. This may attributed to the months with in which they accessed they accessed the pages.

From the navigational behavior of users, there are page that most users accesses to gather is home page. This may be users stay and leave without spending much time in other pages.

From the navigational behavior of users, most users do know where they go after visiting the search engine of the Addis Ababa. This may be since most users are educated and know what they are looking for.

From the statistical report from WUM, Most visited directories are naturally the root directory since most of users are typing the name of the official website and most hits are from the root directory, the next most directory are /index.php/ which hosts other sub directory inside it like /index.php/home or other directory inside it. This is because of the web server are apache servers and most of the hits are register as root. Besides to that, there are directories that are frequently accessed than other directory. This might be users are interested in the content of those directory for example users who are interested in the library pages, may be its contents are journals and articles.

Most of the site users who have visited the home page also interested in academic pages this seems due to interest to visit of a certain pages academic program listed in the home page. Most users of the official website make further requests from the page of /index.php, /index.php/admissions/graduate, /index.php/home. It seems may be users could not get what they want from those sites as result they initiate them self for further search within the pages.

Recommendation

The researcher made the following recommendations based on the findings of the study.

Improvement of the website:

For the ease of users who visit deeper pages than just the home page of the website, the depth complexity and readability of the page links need to be assessed. Besides the user-interface as well as links of the pages should be assessed since most users leave from the home page. It might need to separate study in detailed whether site is user friendly or not.

Most users left the website from some pages mainly from the /index.php/home, /index.php/admission/graduate, /index.php/academics from them. It possible to recommend that, the web master should use those pages for advertisement and notice because it creates an opportunities to stay before they are leaving the website. Moreover, it is possible further to recommend to encourage the website users by putting links to other department within those pages to stay further in the website.

Most users stay in the home page and spent less time in visiting other web pages so it is possible to recommend the web administrator should make other pages link with most accessed pages with in the home page. Furthermore, it is advisable to make the links in library such as links to journal and articles make to be clickable in the home page so as users could spend more time.

There are pages that are expected to be accessed, but not accessed in the real users therefore, lots of amendments are expected from the web administrator or web master to alleviate such problem. In addition, the same if some pages need to be removed away from the site.

Further web usage analysis research

- For further work can be recommended that, since the list of robots in “robot.txt” may be out dated over long time or difficult to get to the latest updates. It is possible to identify the normal (non-robot) hosts by merging log files; widely accepted log files for purpose are “agent log file” with “access log file” consequently could be better result.
- The other recommendation for further work, divide the web page based up on concept of hierarchy according to the service they provide, once hierarchal classified the pages it would give better result.
- In this research paper the query applied are few if we are able to add more queries it would have been reflect the behavior of users in detail manner.
- The last not the least, recommendation for the further work, since by combining different technique of web usage mining such as content mining with web usage mining (work of this thesis). It could give better result in terms of efficiency because of; we are able to measure not only from the end users of the page but also from the content of the pages.

APPENDIX I

Sample log file	The Fields Records	Description
The host name if available	@Host_dns@	Host name (or IP address if it could not be resolved)
128.101.35.92	@Host_ip@	Host IP address
-	@Ident@	Ident code
-	@Auth_user@	User authentication code
09	@Ts_day@	Timestamp of the request(Day of visited the page)
/Mar/	@Ts_month@	Timestamp of the request(Month visited the page)
2002	@Ts_year@	Timestamp of the request(year visited the page)
01	@Ts_hour@	Timestamp of the request(Hour visited the page)
03	@Ts_minutes@	Timestamp of the request(minutes visited the page)
18	@Ts_seconds@	Timestamp of the request(Seconds visited the page)
0600	@Tz@	Time zone (e.g., +0200)
GET	@Method@	Request method (GET, PUT…)
/~/index.php/	@Path@	Path/URL of the requested document
HTTP/1.0	@Protocol@	Protocol used for the request
200	@Status@	Server response code
3014	@Sc_bytes@	Number of bytes sent by server to client
http://www.aau.edu.et/Mozilla/4.7	@Referrer@	Referrer information (URL)
(X11; I; SunOS 5.8 sun4u)	@Agent@	User agent information

Table 5: Sample of log format.

Statistical report for the months of November:

Appendix II

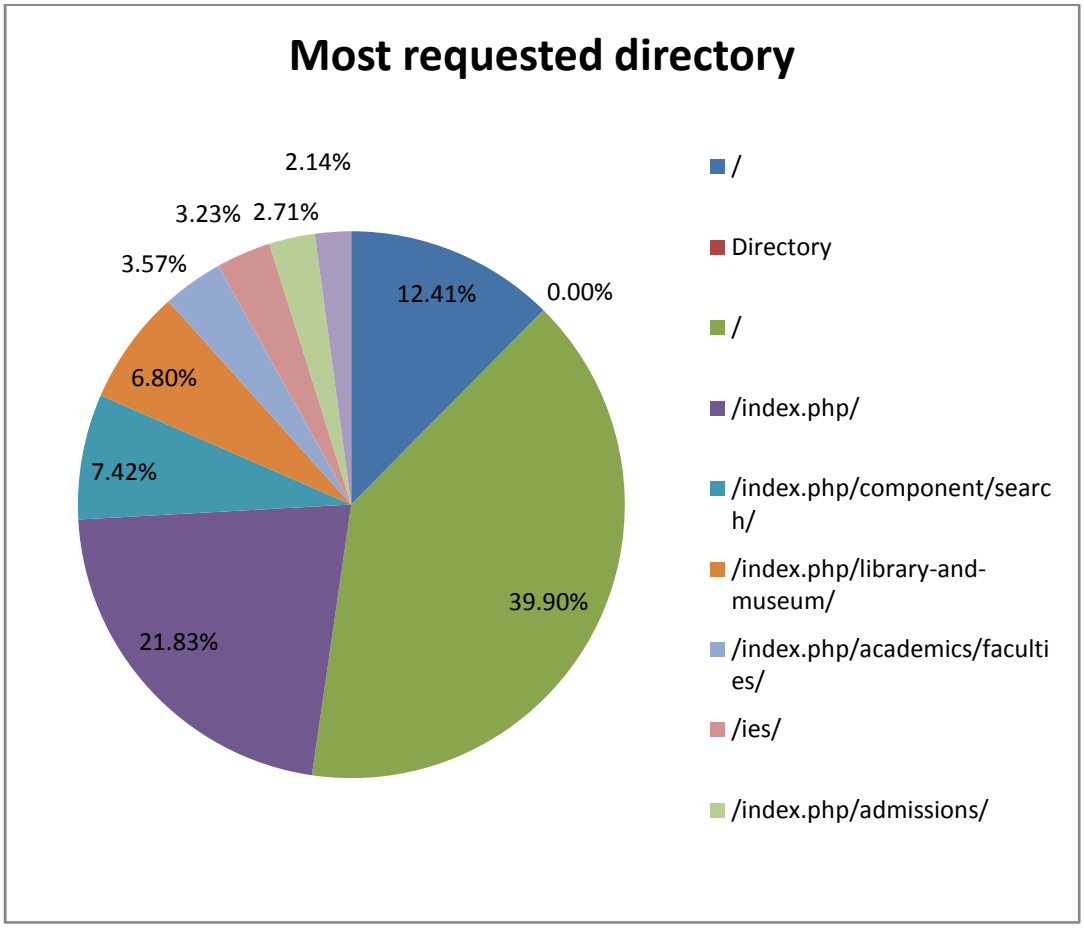


Figure 28: Most requested directory for the month of November.

Appendix III:

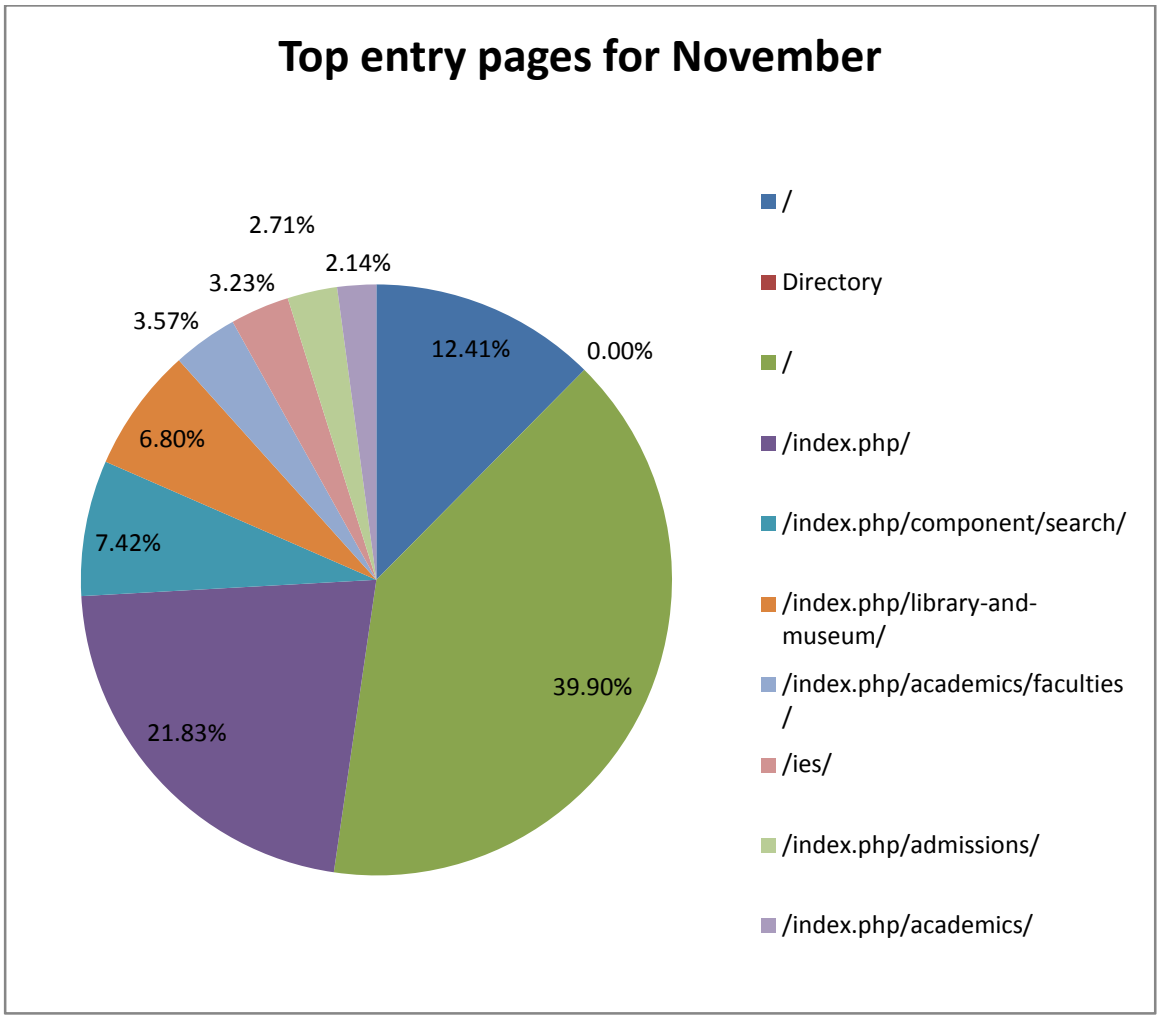


Figure 29: Top entry pages for the month of November.

Appendix IV

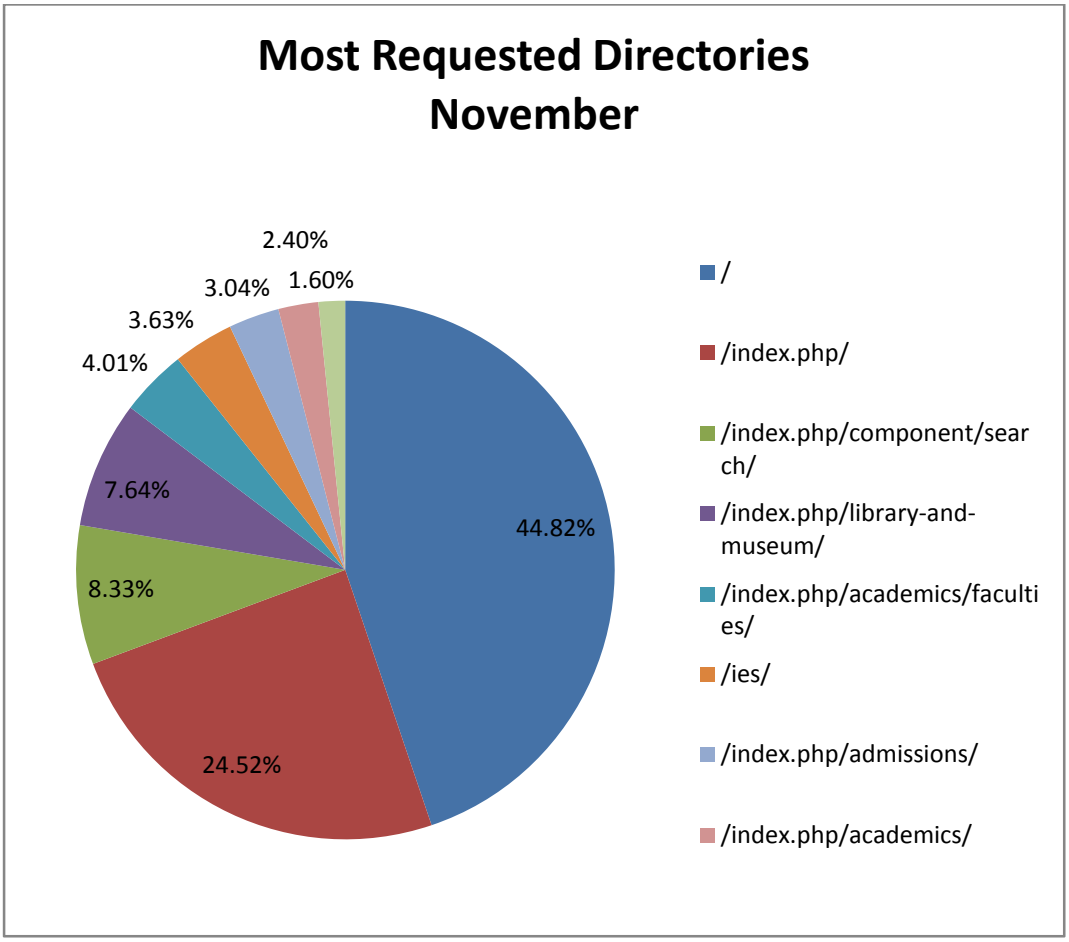


Figure 30: Most requested directories for the month of November.

Appendix V

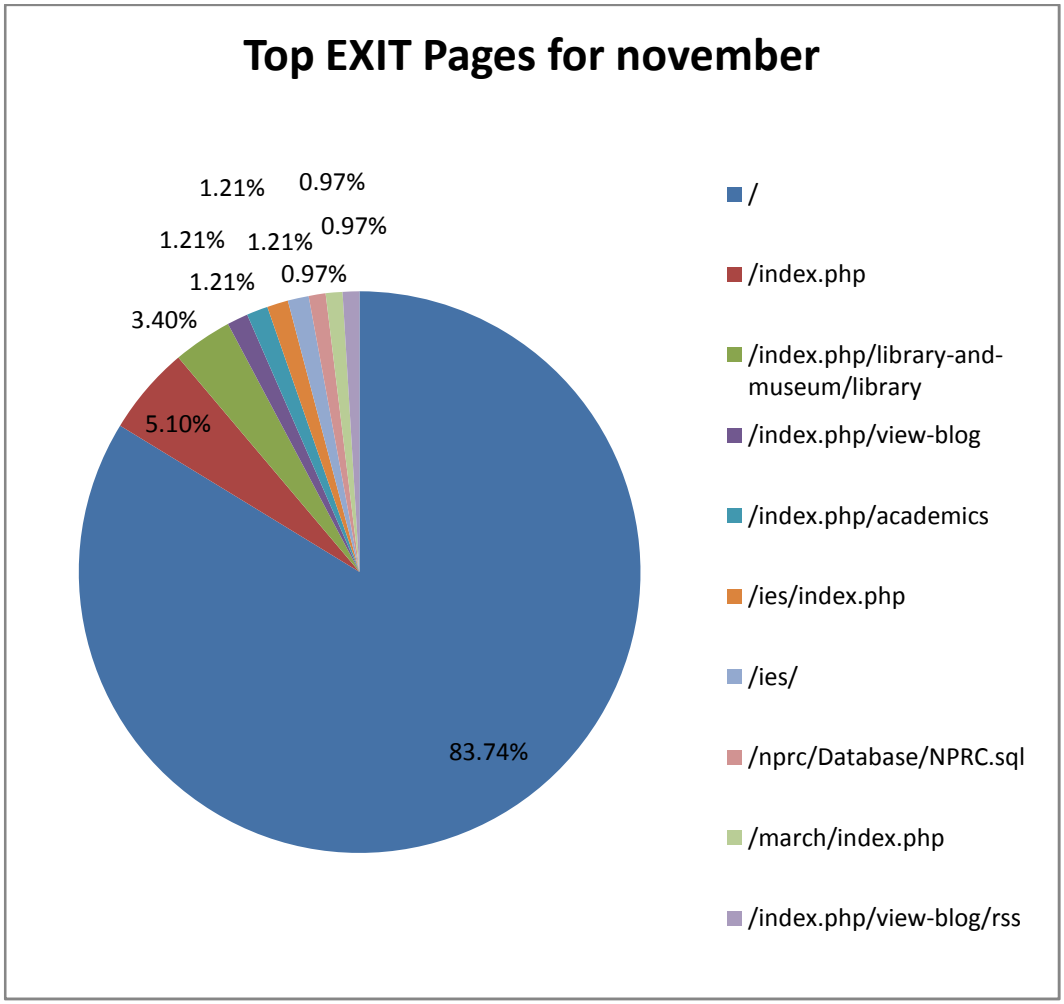


Figure 31: Top exit page for the month of November.

Appendix VI

The following are also the preprocessing for the month of November .

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
210,240	140,127	69,564	20,004	20,004
240,067	160,743	72,087	24,060	24,060
203,406	148,906	63,480	21,000	21,000
200,967	138,967	50,653	19,734	19,734
200,190	178,300	60,752	20,674	20,674
200,150	167,543	47,897	19,653	19,653
220,205	120,950	62,096	23,765	23,765

Appendix VII

Sample removed List of robots:

110.75.173.43 (robots.txt)

119.235.237.16 (robots.txt)

119.235.237.20 (robots.txt)

119.235.237.85 (robots.txt)

119.63.198.11 (robots.txt)

119.63.198.17 (robots.txt)

119.63.198.20 (robots.txt)

119.63.198.21 (robots.txt)

119.63.198.31 (robots.txt)

119.63.198.33 (robots.txt)

119.63.198.35 (robots.txt)

119.63.198.38 (robots.txt)

119.63.198.39 (robots.txt)

119.63.198.41 (robots.txt)

119.63.198.47 (robots.txt)

119.63.198.52 (robots.txt)

119.63.198.54 (robots.txt)

119.63.198.57 (robots.txt)

119.63.198.58 (robots.txt)

123.125.67.227 (robots.txt)

123.125.67.229 (robots.txt)

124.115.6.12 (robots.txt)

130.89.197.30 (robots.txt)

157.55.16.229 (robots.txt)

157.55.16.230 (robots.txt)

174.124.240.38 (robots.txt)

178.154.160.30 (robots.txt)

178.4.31.86 (robots.txt)

178.63.9.74 (robots.txt)

184.154.7.186 (robots.txt)

188.165.226.104 (robots.txt)

193.47.80.48 (robots.txt)

195.215.130.196 (maxViewTime)

202.160.179.85 (robots.txt)

202.180.34.186 (robots.txt)

202.232.133.34 (maxViewTime)

204.236.235.245 (robots.txt)

206.16.59.98 (robots.txt)

206.192.70.55 (maxViewTime)

207.210.81.165 (maxViewTime)

207.241.227.74 (robots.txt)

207.241.228.153 (robots.txt)

207.46.12.236 (robots.txt)

207.46.12.237 (robots.txt)

207.46.12.239 (robots.txt)

207.46.12.240 (robots.txt)

207.46.12.241 (robots.txt)

207.46.13.100 (robots.txt)

207.46.13.101 (robots.txt)

207.46.13.131 (robots.txt)

207.46.13.132 (robots.txt)

207.46.13.133 (robots.txt)

207.46.13.134 (robots.txt)

207.46.13.137 (robots.txt)

207.46.13.138 (robots.txt)

207.46.13.139 (robots.txt)

207.46.13.140 (robots.txt)

207.46.13.142 (robots.txt)

207.46.13.144 (robots.txt)

207.46.13.145 (robots.txt)

207.46.13.146 (robots.txt)

207.46.13.41 (robots.txt)

207.46.13.42 (robots.txt)

207.46.13.44 (robots.txt)

207.46.13.45 (robots.txt)

207.46.13.50 (robots.txt)

207.46.13.52 (robots.txt)

207.46.13.53 (robots.txt)

207.46.13.54 (robots.txt)

207.46.13.85 (robots.txt)

207.46.13.86 (robots.txt)

207.46.13.87 (robots.txt)

207.46.13.88 (robots.txt)

207.46.13.89 (robots.txt)

207.46.13.92 (robots.txt)

207.46.13.93 (robots.txt)

207.46.13.94 (robots.txt)

207.46.13.95 (robots.txt)

207.46.13.97 (robots.txt)

207.46.194.114 (robots.txt)

207.46.194.126 (robots.txt maxViewTime)

207.46.194.137 (robots.txt)

207.46.194.42 (robots.txt)

207.46.194.78 (robots.txt)

207.46.195.105 (robots.txt)

207.46.195.106 (robots.txt)

207.46.195.223 (robots.txt)

207.46.195.224 (robots.txt)

207.46.195.225 (robots.txt)

207.46.195.226 (robots.txt)

207.46.195.227 (robots.txt)

207.46.195.228 (robots.txt)

207.46.195.230 (robots.txt)

207.46.195.231 (robots.txt)

207.46.195.232 (robots.txt)

207.46.195.233 (robots.txt)

207.46.195.242 (robots.txt)

207.46.199.177 (robots.txt)

207.46.199.178 (robots.txt)

207.46.199.179 (robots.txt)

207.46.199.180 (robots.txt)

207.46.199.182 (robots.txt)

207.46.199.183 (robots.txt)

207.46.199.184 (robots.txt)

207.46.199.185 (robots.txt)

207.46.199.191 (robots.txt)

207.46.199.193 (robots.txt)

207.46.199.198 (robots.txt)

207.46.199.199 (robots.txt)

*the shading area show that those which are excdeing the maximuam time (1800 sec) and taken as robots.

APPENDIX VIII:

A the Syntax of MINT

query::=

’SELECT’ selectList

fromClause [whereClause]

[groupClause [havingClause]]

selectList ::=

[’DISTINCT’] derivedColumn

(’,’ derivedColumn)*

derivedColumn ::=

(valueExpr|aggrExpr)

[’AS’ columnName]

aggrExpression ::=

aggrOp ’(’ [’DISTINCT’]

(valueExpr|varName) ’)’

aggrOp ::=

’AVG’ | ’MAX’ | ’MIN’ |

’SUM’ | ’COUNT’ | ’GLUE’

fromClause ::=

’FROM’ tableRef (’,’ tableRef)*

tableRef ::=

’NODE’ ’AS’ nodeVar* |

’TEMPLATE’ template

[’AS’ templateVar]

template ::=

[’*’] (nodeVar [’*’])*

varName ::=

nodeVar|templateVar

whereClause ::=

’WHERE’ condition

(’AND’ condition)*

condition ::=

valueExpr compOp valueExpr

compOp::= ’=’ | ’<’ | ’>’ |

’<=’ | ’>=’ | ’LIKE’

valueExpr ::=

numericExpr | stringExpr

numericExpr ::=

[numericExpr (’+’|’-’)] term

term ::=

[term (’*’|’/’)] factor

factor ::=

[(’+’|’-’)] primary

primary ::=

literal | columnReference |

’(’ valueExpr ’)’

stringExpr ::=

[stringExpr ’||’] primary

columnReference ::=

varName ’.’ columnName

groupClause ::=

’GROUP’ ’BY’ groupExpr

(’,’ groupExpr)*

groupExpr ::=

nodeVar | columnRef

havingClause ::=

’HAVING’ condition

(’AND’ condition)*

References

Abhishek, C.,& Satendra, K.,(2011).A Comprehensive Survey on Frequent Pattern Mining from Web Logs. Computer Applications, SATI, Vidisha, Madhya Pradesh, India. Published in International Journal of Advanced Engineering & Application, Jan 2011.

Agrawal, R. & Srikant, R. (1995). Mining sequential patterns. InICDE, Taipei, Taiwan.

Anália, M.,& Orlando M.,(2003).Assessing web usage profiles. Departamento de Informática, Escola de Engenharia, Universidade do Minho Campus de Gualtar,Braga, Portugal,2003.

Ballman. A., & Yu,S.,(1997). SpeedTracer: A Web Usage Mining and Analysis Tool.Internet Computing, 37(1): 89, 1997.

Bamshad, M., &Robert C, & Jaideep, S(n.d).Data Preparation for Mining World Wide Web Browsing Patterns. Department of Computer Science and Engineering University of Minnesota.

Berendt, B., Myra ,S.,(2000) .Analysis of navigation behaviour in websites integrating multiple information systems. Institute of Pedagogy and Informatics,The VLDB Journal (2000) 9: 56–75.

Berkan, y., (2002). Predicting Next Page Access By Time Length Reference In The Scope Of Effective Use Of Resources.

Bettina, B.,& Myra, S.,(1999).Analysis Of Navigation Behaviour In Websites Integrating.Multiple Information Systems.The VLDB Journal (2000) 9: 56–75.

Briand, H.& Guillet, F. ,(2005).Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support", June.

Brendit,(2011a).Web Mining Usage In E-Commerce.http://vasarely.wiwi.huberlin.de/WebMiningSS02/Session5/index.html#dbs-dataset,[accessed april 13 2011].

Carsten, P.,& Myra,S., (2000).Data Mining to Measure and Improve the Success of Websites. arXiv:cs.LG/0008009 v1 15 Aug 2000 Engineering, Ferdowsi University of Mashhad, Iran.

Castellano, G., & Fanelli, M.,& Torsello. A.,(2007).Log Data Preparation For Mining Web Usage Patterns. Department of Computer Science – University of Bar, IADIS International Conference Applied Computing.

Chu-Hui, L., &Yu-Hsiang, F.,(2008) . Two Levels of Prediction Model for User's Browsing Behavior. Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008 Vol I IMECS 2008, 19-21 March, 2008, Hong Kong.

Cooley, R., Mobasher, B., & Srivastava, J. (1997a). Grouping web page references into transactions for mining world wide web browsing patterns. Technical Report TR 97-021, Dept. of Computer Science, Univ. of Minnesota, Minneapolis, USA.

Dietmar, W., & Peiling, W., & Jin, h., (n.d). Modeling Web Session Behavior Using Cluster Analysis:A Comparison of Three Search Settings. School of Information Studies, University of Wisconsin-Milwaukee.

Dipa, D.,& Kiruthika. M .(2010) .Preprocessing Of Web Logs. International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010, 2447-2452.

Enrique,F.,&Vijay,K.,(2003). A Customizable Behavior Model for Temporal Prediction of Web User Sequences. (Eds.): WEBKDD 2002, LNAI 2703, pp. 66–85, 2003.

Federico,M.,&Pier,L.,(2000). Recent developments inWeb Usage Mining Research. Artificial Intelligence and Robotics Laboratory Dipartimento di Elettronica.

Henri , m., &. Osmar, m ,.(2000). Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support. universite de nice sophia,antipolis.

Ian H.,& Eibe F,p.,(2005). Mining practical machine learning tools and techniques.2nd ed. Department of Computer Science University of Waikato: Diane Cerra.

Istrate,M.,(2000).Web mining in e-commerce.University of Pitești Faculty of Mathematics and Informatics. No1.romaina.

Jaideep, S.,& Robert ,C.,& , Mukund, D., &Pang-Ning,T.,(n.d). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. Department of Computer Science and Engineering University of Minnesota ,Minneapolis.

Jeffrey W. Seifert. (2004). Data Mining: An Overview. December 16, 2004.

John, E.,(1997). Profiling User Responses to commercial websites. Journal of Advertising Research, 37(2):59–66, May-June 1997.

José B., & Mark L.,(n.d) .Mining Users' Web Navigation Patterns and Predicting Their Next Step. School of Computer Science and Information Systems, Birkbeck, University of London.

Jose, M. & Javier, L., (2007).A Tool for Web Usage Mining.8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07), 16-19 December, 2007, Birmingham, UK.

Kerkhofs, J.,& Koen, V., (2001).Web Usage Mining on Proxy Servers: A CaseStudy.Limburg University Centre July 30, 2001.

Kobra,E.,&Mohammad,Akabarzadeh.,&Nooarali,Raeeji.,(n.d).Usage Mining:users' navigational patterns extraction from web logs using Ant-based Clustering Method. . Department of Computer. Iran

Kosala, R. & Blockeel, H.,(2000).Web Mining Research: A Survey. SIGKDD: SIGKDDExplorations. Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, 2(1):1, 15, 2000.

Lavoie, B., & Nielsen, H.,(1999).Web Characterization Terminology & Definitions Sheet. http://www.w3c.org/1999/05/WCA-terms/, May 1999.

Lita, V.,& Lamber ,R.,(2004).Ethical Issues In Web Data Mining. Department Of Philosophy And Ethics Of Technology.Department of Philosophy and Ethics, Faculty of Technology Management, Eindhoven University of Technology, Eindhoven.

Lukas, C.,&Myra, S.,& Karsten, W.,(n.d). A data miner analyzing web navagtioanl behavior of web users. Institut for ur wirtschaftsinformatik, Humboldt-Universit at zu Berlin.

Maja,D., (2011).Web Usage Association Rule Mining System. Interdisciplinary Journal of Information, Knowledge, and Management Volume 6, 2011.

Magdalini,P.(2006). New Approaches To Web Personalization. Athens University Of Economics And Business, Dept. Of Informatics. May 2006.

Myra,S.,(2000). Web Usage Mining For Website Evaluations. Communications of the acm August 2000/Vol. 43, No. 8.

Myra,S., & Lukas C. (n.d). A Web Utilization Miner. Institut f¨ur Wirtschaftsinformatik, HU Berlin.

Murat ,A, &Ismail, H. , Ahmet ,C., (n.d) . A Performance Comparison of Pattern Discovery Methods on Web Log Data. Department of Computer Engineering Middle East Technical University.

Masseglia f.,& poncelet p.,& cicchetti r(n.d). webtool: an integrated framework For data mining, proceedings of the 9th international conference on database.

Mohd ,H.,& Abd, W., &Mohd, N.,& Haji, M.,(2007a).Discovering Web Server Logs Patterns Using Generalized Association Rules Algorithm. Universiti Tun Hussein Onn Malaysia Universiti Utara Malaysia, jan 2007a.

Mohd, H.,& Abd, W.,& Mohd N.,& Haji, M.,& Hafizul, F.,(2008).Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology 4-8 2008.

Moges,D. (2005), Information and Communication Technology at the Addis Ababa University - A Case Study, Paper Presented on the Joint ITU/UNU/CERN Workshop on AFUNET September 25-29, 2005,Geneva

Mobasher b., &Jain n., h., &Srivastava j.,(1996) Web Mining: Pattern Discovery from World Wide Web transactions, t num. TR-96-050, Department of Computer Science, University of Minnesota.

Moknnen Tsegay (2009).Web Usage:Pattern Discovery using data mining and statistical analysis in the case of AAU official website, Addis Ababa University ,Ethiopia

Narendra, K.,& Haresamudram., (n.d). Research & Development in Web Usage Mining conjunction with Information Retrieva:A Survey. GATES Institute of Technology

Navin, K., & Tyagi1, A., & Sanjay, T.,(2010). An Algorithmic Approach To Data Preprocessing In Web Usage Mining. International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283.

Olfa, N.,& Esin, S.,(n.d).Web Usage Mining In Noisy And Ambiguous Environments: Exploring The Role Of Concept Hierarchies, Compression, And Robust User Profiles. Knowledge Discovery & Web Mining Lab, University of Louisville, Louisville, USA http://webmining.spd.louisville.edu

Pierre, B.,& Leyland F., & Richard T.,(1996). The World Wide Web as an Advertising Medium. Journal of Advertising Research, 36(1):43–54, 1996.

Robert,C., &Srivastava,J.,& Mobasher, B.,(1997).Web Mining: Information and Pattern Discovery on the World Wide Web. In Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

Rajni, P.,& Pramila, C., (2009).Web Usage Mining: A Research Area in Web Mining. Department of computer technology, VJTI University, Mumbai.

Sergey ,B.,(2000). Extracting Patterns And Relations From The World Wide Web". Computer Science Department Stanford University.

Sulu, G.,(2003). Recommendation Model For Web Users: User Interest Model And Click Stream Tree., Istanbul technical university, October 2003.

Suneetha, K.,& Krishnamoorthi, R. (2009). Data Preprocessing and Easy Access Retrieval of Data through Data Ware House. Proceedings of the World Congress on Engineering and Computer Science 2009 Vol I WCECS 2009, October 20-22, 2009, San Francisco, USA.

Srikant R., & agrawal R.,(1996).Mining Sequential Patterns: Generalizations and Performance Improvements, Proceedings of the 5th International Conference on Extending Database Technology (EDBT'96), Avignon, France, September 1996, p. 3-17.

Terry, S.,(1997). Reading reader reaction: A proposal for inferential analysis of web server log files. In Proc. of the Web Conference'97, 1997.

Tianyi ,Li.(1995).Web-Document Prediction And Presending Using Asociation Rule Sequential Classifiers , Zhongshan University.

Zalane, O., &.Xin M., & HAN J.,(1998). "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", Proceedings on Advances in Digital Libraries Conference (ADL'98), Santa Barbara, CA, April 1998.