ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS USING GENERALIZED SEQUENCE PATTERN
A CASE ON OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY

BY

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

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 and foremost, I would like to thank my advisor, Ato Workshet lemnaw, 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.

Of course, my thanks to Professor Bettina Berendt for her borderless support in giving directions on this work ,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.

" fekri Belibi"

Abstract

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years. Academic researchers have developed an extensive array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular web site. Performing this kind of investigation on AAU web site can provide information that can be used to better accommodate the user's needs.

The Web Use Mining (WUM) , it corresponds to the process of knowledge discovery from databases (KDD) applied to the Web usage data. It comprises three main stages: the preprocessing of raw data, the discovery of schemas and the analysis (or interpretation) of results. A WUM process extracts behavioral patterns from the Web usage.

In this thesis, we find out the navigational behavior of the user of official web site of Addis Ababa University web server recorded in web server for two months (November and December), those recorded are raw data that are full of junks, noises and irrelevant data contents .In this paper present a preprocessing tool WUMprep that uses to filter those unnecessary data, such as irrelevant records, noise data, and it crates the sessions based on specific thresholds.

For discovery of navigational behavior, here presents the Web Utilization Miner WUM, a mining system for the discovery of interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the researcher using WUM's mining language MINT, using those descriptor it can be describe the general behavior of users instead of single users behavior using the most appropriate algorithms known (Generalized sequence pattern) which implemented in WUM.

The General behavior of users constructed by GSP algorithms those behaviors are descried using the MINT query. Those MINT query are intermediate between the users and pages.

The researcher of this paper also recommend that to get a better result by combining the web usage mining with content mining techniques of web usage. Of course without any doubt it could give a better result in terms of efficiency and effectiveness results.

Table of Content

List of Table

List of Figures

Web Terminology and Definition

In accordance with 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.

Abbreviation

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
IBM	International Business machine
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

CHAPTER ONE: INTRODUCTION

1.1.Background

In 1990 the internet was initially designed for exchange mails between users later it becomes trendy for use of WWW. The www or 3w in now popular services among almost any other services the internet provides. There are number of services providers (ISP) for the use of the internet across the world. In Africa, the number of the internet users increasing and increasing from time to time. 5.6% of the world internet users are from Africa, further explained, it shows 2,357.3 %growth from the year 2000-2010 similarly, Ethiopia has 0.4 % share among African internet users .Even if this seems insignificant when it compared with the rest of the world, generally speaking the number of the internet across the world getting increasing and increasing in dramatic way thorough out worldwide1.one of the various reasons for the development of the internet in Ethiopia causes by huge amount of investment in infrastructure like in education ,telecommunication and development in others sectors.

Addis Ababa University, one of the oldest higher education institutes 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.

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

[1] http://www.internetworldstats.com/stats1.htm#africa

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]

There are number of services provided by the Addis Ababa University, one of the popular services are the WWW(world wide web) among other services like teleconference ,data service ,those web services are divide in two as the official web site (internet) and intranet which is not able to be accessed outside the university which uses for local uses. The official web site accessed through the public Ip address offered by the ETC.

The official web services of AAU an organized collection of Web pages information is presented in various formats , ranging from research papers, and educational content, to multimedia content, blogs .that's why the getting information from the official web site is the matter of click-streams in the internet of course if there is connectivity. 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 a few individuals who realized that the AAU would be wise to join the information age by adopting the technology that has been transforming the world. The newly formed office initiated a project named AAUNet that has resulted in a wide area network (WAN) whose first phase of construction was 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 is adequately served by ICT. At the same time, AAU's need for and dependence on effective ICT support is now greater than ever.

[2] www.aau.edu.et/administration/DRAFT ICT POLICY AT AAU
[3] www.aau.edu.et/administration/ICT

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.

1.3. The AAU Official web site

The Addis Ababa university official web site was published around some seven years ago .As the ICT development office of AAU (which have mentioned in previous section) is engaging in ICT related works ,the official web page develop and maintain by this office. the web site is hosted on AAU's own server which is located in main campus of the university (6 kilo), The official web site have the domain of www.aau.edu.et and have statistical IP address.

1.4. Purpose and User Community

The official web site being in work to deliver information both the university activity, in general and about academic and administrative units, in particular, it also delivery information about news, items and its own advertisement for both vacancies and student admission and other, of course it has also some external links to other web and other sites such as collaborative organizations in research activity donor agencies, etc.

1.5. Nature and Content

Generally the web sites designed bear in mind to support the objective of the university. In sections try to discover the nature and content of the web site. The AAU web site has both static and dynamic nature .there are few web sites that are static in nature those pages are not interactively with its users but the majority of the web pages are dynamic in nature which are support the MYSQL database incorporate with JOOMLA packages helps users to interact with web sites users.

When we came to Web site content it posts numerous information regarding to the objective of university which presenting information on several topics and issues, each page have information regarding to the objective of the pages .there are few page which are under construction(content not yet update), but there are advertisement and notice on several pages.

1.6.AAU Web Structure

In the following hieratical graphs displays web site structures of the official web site.

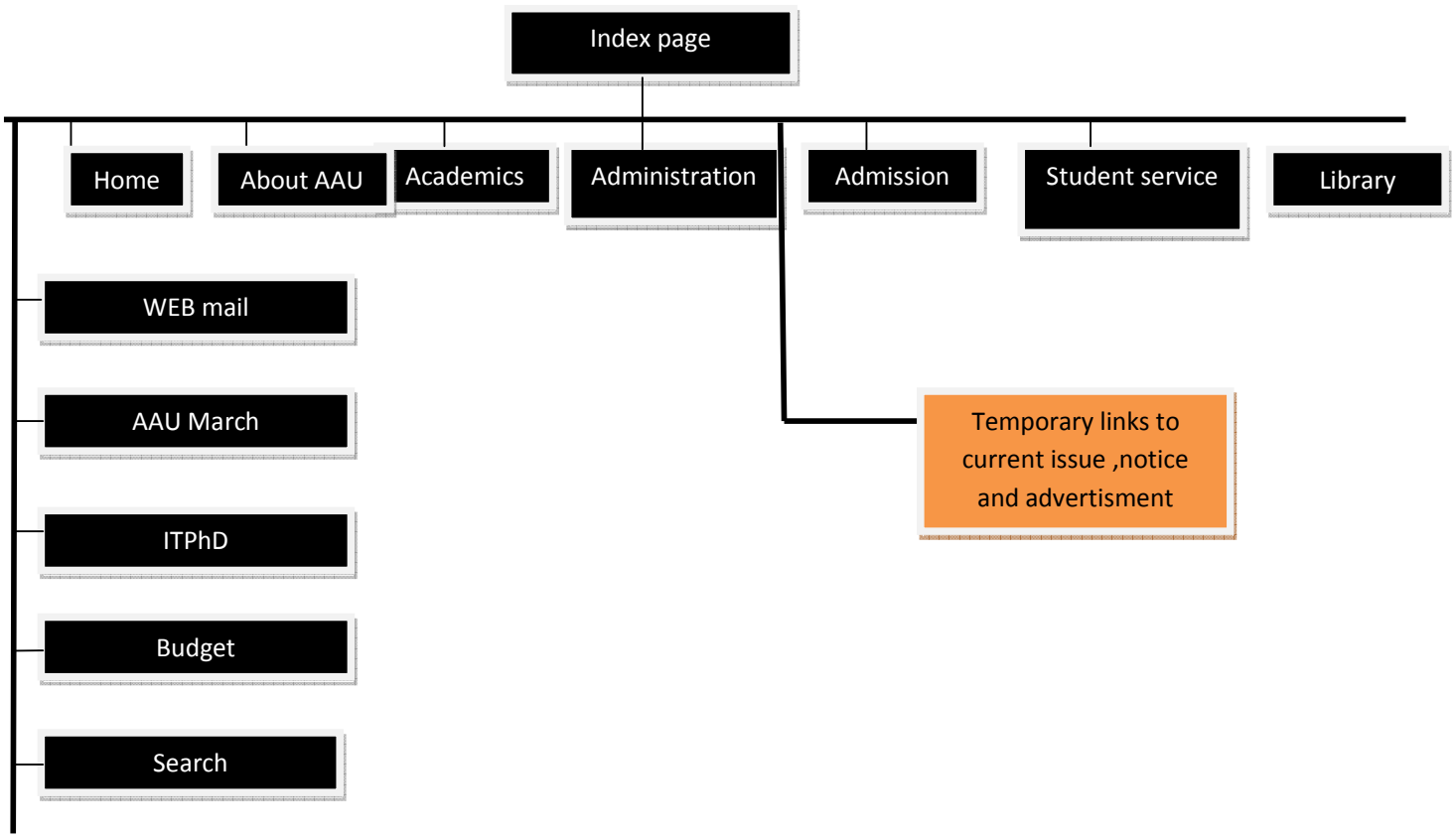


Figure 1 the structure of the official web site of AAU.

There are some other web sites that are accessed to gather with the official web sites like AAU march, ITPhD, college of education, IES (Institute Ethiopian Study), virtual accessed using the main web site.

1.7.Statement of the Problem

Rise of the Internet gave many companies an access to the 'gold' channel. Trading, putting gigabytes of information and communicating online has become one of the sources for understanding of the web users. 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.

As statics shows the number of web sites published every day is increasing quickly still, there are now 184 million registered domain names worldwide, a 9% increase over the same period last year[4].

On the other hand, the education sector is rapidly evolving and the need for web information Places that anticipate the needs of their information seekers are more than ever evident. The need of placement information is not easily imaginable we have to explore where should be places some information in a given web site, in this case of the official web site of AAU. It is important to know the navigational behavior of the users based on the study of the behavior. the need of study of any behaviors scaled up from the taxonomy of animals ,plants and others , in general, further explained that animals classified in to mammals ,vertebrates based up on the whole group behaviors.

According to Mokenen (2001) who were working on web usage mining of the official web site of AAU using the tools of wumprep4weka, for preprocessing or cleaning the data and Weka tool for data mining of the interesting pattern using the aprior algorithms finds out the most frequent access that do not based up the sequence, based on his study he did not truck the general behavior of users.

Like it discussed earlier uses the sequence (generalized sequence pattern) can tell the general behavior of users on navigational behavior of the user of official web site of Addis Ababa University, and not work have been done yet on the topic as to the knowledge of the author.

Web site design is currently based on thorough investigations about the interests of web site visitors and on less investigated assumptions about their exact behavior. In Lukas, C., (n, d) Concrete knowledge on the way visitors navigate in a web site could

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

prevent disorientation and help owners in placing important information exactly where the visitors look for it.

1.8. Scope and Limitation of the Research

Web mining has different branches: web content mining, web structure and web usage mining .the focus of this research is on mining usage pattern of AAU official web site .usually, three types of web related log files, namely web access log, error log and proxy log files. however, in this research work, web access log records is used as dataset because many literature and previous research justify that web access log files is the typical source of navigational behavior.

The limitation in this paper is the lack of manual on how to operate the web mining tools (WUM) and besides to that the web access log stored in Addis Ababa university are erased at the end of every months that's why it is difficult to get a enough data for the research, besides to that the web mining tools need to have a higher capacity (memory) to process the whole log files as batch.

1.9. Justification of the Research

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.

The importance of the study web users further explained by Marya, et al, according to him ,most web sites 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 web site they present to their visitors.

One important data source for the study is the web-log data that traces the user's 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, the importance of www further explained in Kosala et al, (2000), the web log creates an opportunity and encouragement for all Data mining researchers, consider it as the largest data warehouse in the world.

In accordance with 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's why performing data mining to optimize the performance of a Web server. In ref of Lukas, C., (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.

In accordance with Narendra, et al., (2003), Web mining is defined "*as the use of data mining techniques to automatically discover and extract information from web document and services. "*
Furthermore, there is also a widely accepted definition, According to Zalane, et al ,(1998).

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 web site is being used as expected*".

From the above the definitions web mining attempt to get the information (knowledge) or to extract the pattern, for the purposes to have an intended knowledge, so some the techniques should be applied to different web resources to overcome the problems, in ref with Mobasher et al, (1996), web mining is a common term for three knowledge discovery domains that are concerned with mining different parts of the web: web structure mining, web content mining, and web usage mining.

In general, User behavior has two aspects, one concerning the interests of the users and the information they access, the other concerning the way of accessing this information. The first aspect is addressed by techniques for the establishment of user profiles and is not peculiar to web usage. For instance, student profiles are considered in intelligent tutoring systems, the second aspect is addressed by techniques analyzing web server logs.

For example, consider a user that explores the links in a web site to find every bit of information of potential interest and a user that prefers keyword search. Those two users need fundamentally different support, even if both of them are interested in solar energy collectors, chess and medieval sculpture, In this study, concentrate on the second aspect of user support, namely on the analysis of user navigational behavior, because web users is characterized by her/his interests and by her/his navigational behavior.

1.10. Objectives

1.10.1. General objective

The general objective of the research is to apply web mining techniques for discovering of navigational behavior of AAU official web site usage of to reveal previously unknown the interesting, and actionable patterns based on the web access log file in order to recommend possible measures for further r improvement of the official web site of AAU.

1.10.2. Specific objectives

To achieve the general objective of the research, there are specific objective should be addressed, the specific objectives of the research are:

- To review literature review in the area in order to put concrete background and justification for the research.
- To identify and collect the data
- To prepare those data set using different preprocessing techniques.
- To analyze the navigational behavior of the users.
- To analyze the sequence of the web site 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.

1.11. Research Methods

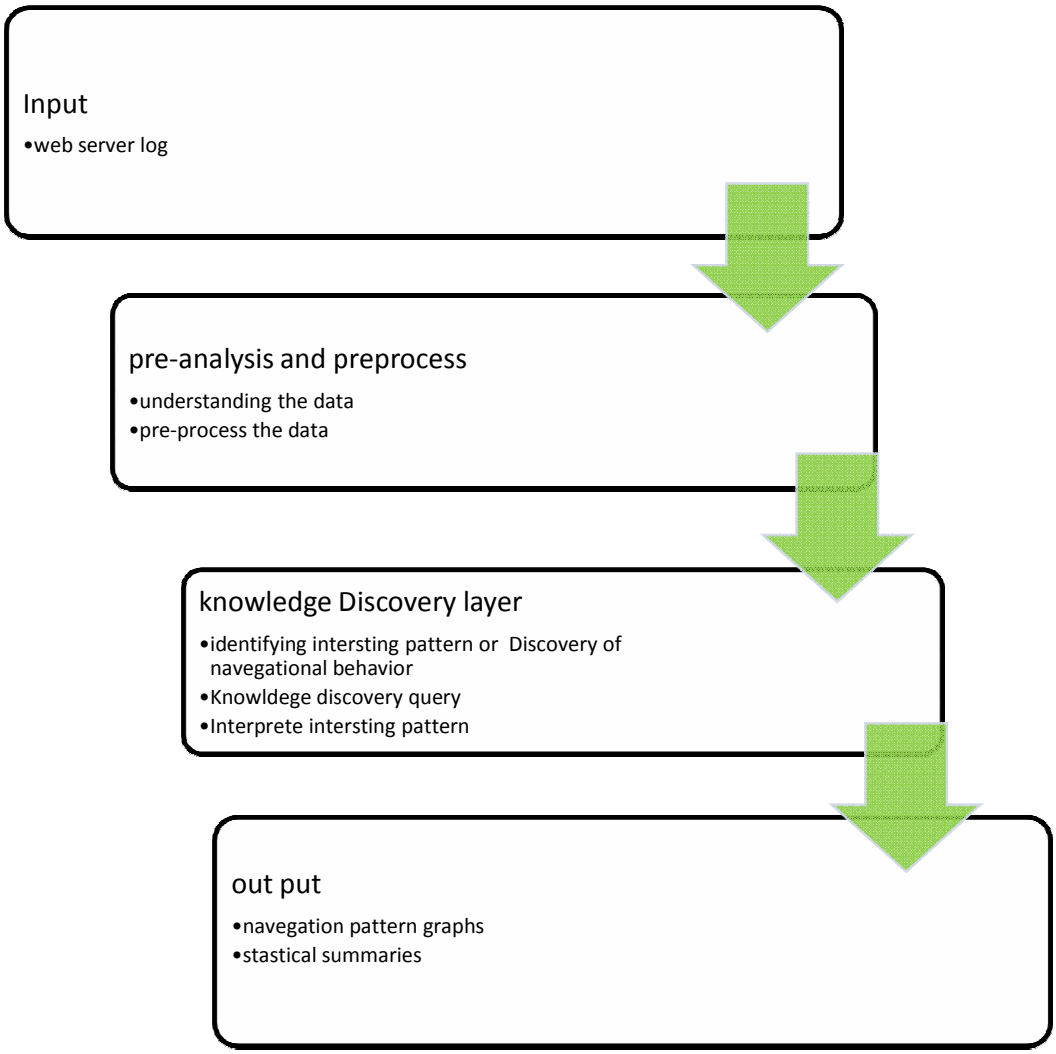


Figure 2:Research method flow

1.12. Data Collection for the Study

In this study the data has been collected from the official web site of the AAU, which is normally secondary data source since web log records every activity of the user regarding to visit of the web site.

1.13. 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 author of the paper, uses server data that are kept in the official web site of AAU in the format of extended log format, which is most apache server supports it.

1.14. Data preprocessing

According to olfa,et al, (n.d) , most log files are full of junks that are insufficient, inconsistent and including noise so the data pretreatment is to carry on a unification transformation to appropriate sets ; to have those sets there are some data cleaning phases are important to implement.

1.15. Data Cleaning

In ref olfa,et al,(n.d), the purpose 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 accordant to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

In addition to the above those also include some phases like, removing robot requests (filtering out spiders or crawlers which are known), removing duplicate requests (removing “dust”), and Filtering relevant status.(those concepts will be described in the Chapter Three).

.

1.16. Data analysis

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

1.17. Tools for Experiment

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, and Anália, et al., (2003),WUM (web utilization miner), Its primary purpose is to analyze the navigational behavior of users in a web site, furthermore ,Navigation pattern discovery is performed on the portion of the web server log that contains the sessions.

The justification for why these tools are selected is given in the chapter FOUR.

1.18. Interpret and report result

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

1.19. Application of results

The hidden unknown information in log formats are important in understanding of users navigational behaviors even if it is not possible to know what will be the results but some knowledge will be revealed by understanding of the general behavior of web site users of AAU .it can be used for improving the web site and it shows some way for further study.

1.20. Organization of the Thesis

This thesis organized as Six chapters ,the first chapter deals with the general introduction to the research of the area in this case the AAU, including the background of the Addis Ababa University in general, it also looks on development of ICT, and how looks like the structure of the official web site, what are their main purposes and later discusses statement of the problem, data collection ,data preparation with other subtopics like, scope and limitation of the study; objective of the study; research methods; etc.

The rest of this thesis is organized as follows. Chapter 2 presents two main areas, Literature review and related works regarding to Data mining and web usage mining.

Chapter 3 this chapter mainly deals with web usage and navigational behavior based on extended of the above chapter in terms of concepts.

Chapter 4 this chapter provides with methodology, in this presents the researcher points why select the tools for preprocessing and the tool for navigational behaviors in general, research process how to achieve the objective.

Chapter 5 in this chapter the experiment conducted and discussed which are based up on the methodology in the previous chapter.

Chapter 6 the last chapter, based on the experiment done in the previous chapter, the conclusions have been reached and recommendation and what it should be done for the future or further work in this research area.

CHAPTER TWO: LITERATURE REVIEW

2. Introduction

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), according to 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

2.1. 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. In Accordance with Berkan, y., (2002), the difference between them is that the former does not store Referrer and Agent information of the requests.

According to Srikant, et al, only few fields are available for navigational patterns discovery, which If are added to the CLF make up the so called Extended combined log format (supported by Apache Web Server).

2.2.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.

I. 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
```

II. 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)"
```

III. 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"
```

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

IV. 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, -,
```

2.3.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)"
```

The details of the fields in the entry are given in the following section.

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:

-

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

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 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)

2.4.Overview and 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 data base 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), refers as Data mining is 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 they can be 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 is often referred to as "secondary" data analysis.

2.5.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 is 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 is scheduled to depart, is 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).

2.6.Data Mining Approaches

It have mentioned earlier that the web usage mining is the application of data mining .those Data mining have two approaches according to (brendit,2011), the approaches is 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."

But, 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.

2.7.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 ref 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.

2.8. Taxonomy of Web Mining

In ref 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 2.8.1.

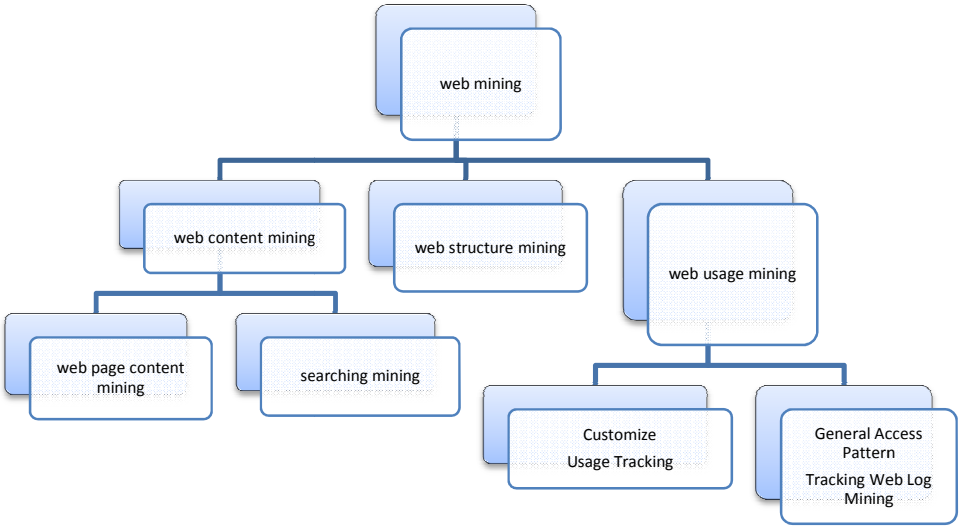


Figure 3: Taxonomy of Web mining,

2.8.1. Web Usage Mining: WUM

Web usage mining can also be defined as the application of data mining techniques to discover user web navigation patterns from web access Zalane et al, (1998), in addition to that, generalized definition accordance 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 to 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.

2.8.2. 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 him, 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.

2.8.3. 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.

2.9.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 web site. 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),this kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, like Most visited pages, average daily hits, etc.

In reference with Bamshad. et al ,(n.d), generalized as this kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, 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 ref 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, we can 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.

According to Bamshad et al ,(n.d)), generalized as 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 data base, 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.

In ref with Bamshad et al, (n.d)), generalized 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 ref (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 web site is needed .the second method computes clusters of page references based on how often they occur together.

In ref 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.

2.10. Related works

Data mining techniques are not easily applicable to Web data due to problems both related with the technology underlying the Web and the lack of standards in the design and implementation of Web pages. Web usage mining is a research field that focuses on the development of techniques and tools to study users' web navigation behavior.

2.10.1. Related Works on the Tools

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.

The other related works on tools on SpeedTracer, 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 web site 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.

Advanced mining algorithms uncover users' movement through a web site. The end result is collections of valuable browsing patterns that help Webmaster better understand user behavior. Further explained in the paper that generates three types of statistics: user-based, path-based and group-based. User-based statistics point reference counts by user and durations of access. Path-based statistics identify frequent traversal paths in web presentations. Group-based statistics provide information on groups of web site pages most frequently visited.

2.10.2. Navigation Pattern Discovery Tools

There are some web usage miner tools which can be used to the navigational pattern discovery for web user behavior of the web site, 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.

2.10.3. Related works in Advances Web Usage Mining

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 on the basis of user demographics and interests, as well as analysis of a site's success.

In Abhishek et al (2011), correlated but not linked web pages are discovered by clustering pages requested together by the site's visitors. This approach can be used to construct dynamic web pages automatically that provide links to pages considered relevant by earlier visitors Pierre, B., et al, (1996).

In the SurfAID project, a warehouse over web usage data is established and time series analysis is combined with association rules to discover unexpectedly evolving correlations among products (Abhishek, et al, 2011) propose the establishment of a warehouse, in which web usage data are combined with customer data, concept hierarchies on page contents and user demographics, as well as enterprise knowledge, e.g. in the form of previously discovered rules Myra,S., & Lukas C. (n.d). . Although user activities form the basis of these types of analysis, the issue of improving the site itself is not addressed.

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 web site 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, we present 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 web site, 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 web site association rules. The “Foot prints “ tool of records the footprints left behind by web site visitors and accumulates 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.

CHAPTER THREE: WEB USAGE MINING AND NAVEGATIONAL PATTERN

3. Introduction

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 web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is Access logs on server side, which keeps information about user navigation. Further explained in Sulu, G.,(2003), Web usage mining is the process of identifying representative trends and browsing patterns describing the activity in the web site, by analyzing the users' behavior. Web site administrators can then use this information to redesign or customize the web site according to the interests and behavior of its visitors, or improve the performance of their systems.

3.1. The General Process of Web Usage Mining

Today, understanding the interests of users is becoming a fundamental need for Web sites 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 can be exploited in the process of navigational behavior.

As it have mentioned earlier , Web Usage Mining (WUM) is the process of knowledge discovery and analysis of Knowledge from World Wide Web, represents a rather recent research field devoted to discover behavioral patterns from Web usage data.

As in Zalane et al (1998), the general processes of WUM distinguish three main steps: data preprocessing, pattern discovery and pattern analysis.

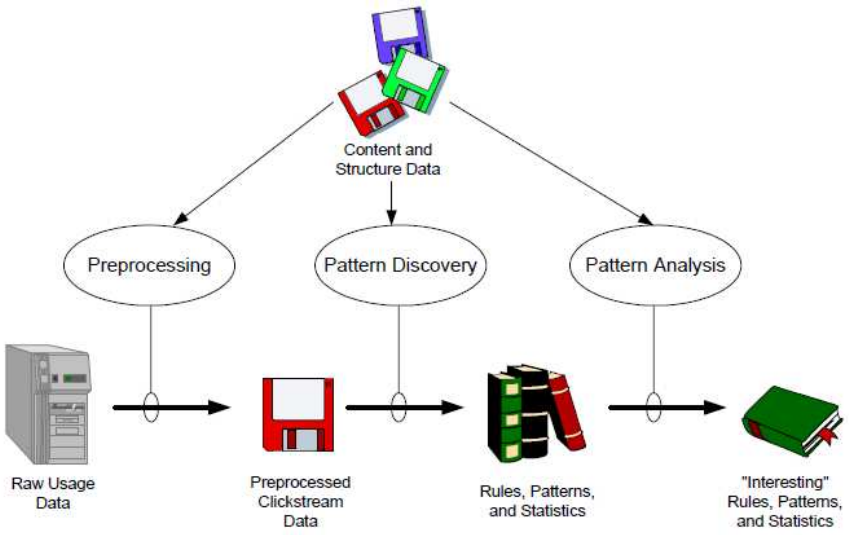


Figure 4: High Level Web Usage Mining Process Jaideep, et al (n.d), page 4

3.2.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.

Besides to the major sources of the data which have mentioned above but, there are also some other resources for web usage mining. According to Castellano, et al (2007) the following can be the source of the data.

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

A different researchers uses different collections over a time for web usage analysis in accordance with Berkan, y.,(2002), were collected for a period of two weeks for Logs Preprocessing and Sequential Pattern Extraction with Low Support.

3.3.Data pre-processing

In ref 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 ref 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.

3.4.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 ref 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 also their Performance is not enough for huge volume of data.

Researchers have shown that the log files contain critical and valuable information that must be taken out. It makes web usage mining a popular research area for many applications in the recent years.

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).

According to 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 Web site and outputs 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.

3.5. Data Cleaning

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

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

3.6. Removing Unnecessary Records

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; In ref 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. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

3.7. 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 s. 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. 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.

In ref 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 Web sites by following every hyperlink on every page within the site in order to update the index of search engine. Requests created by Web robots are not considered usage data and, consequently, have to be removed. To identify web robots' requests, the data cleaning module implements two different heuristics.

Firstly, 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.

3.8. 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 ref (Rajni, P., et al 2009), User's identification is, to identify who access web site and which pages are accessed.

But this task is not easy because few web sites that uses authentication to access the resource so the web records, only records the visitor's host and user agent. Further explained by 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 ref 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. A session is a series of web pages user browse in a single access, or according to Castellano et al, (2007), A 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.

3.9. 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 web site 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; or in e-commerce sites.

- 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 web sites. 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 web sites from the users' viewpoint. According to Federico et al (2000), Exploits,Web Usage mining techniques to suggest proper modifications to web site. Adaptive Web sites represents a further step. In this case, the content and the structure of the web site 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. in ref with (Sulu, G.,(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.

3.10. Navigational Pattern and Sequence

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 we instruct the miner to find only maximal frequent sequences, we obtain fewer and more compact results.

In the ref 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: When we investigate the causes of manufacturing errors, we only want the sequences containing error and cause, not the many events in between. The same is true when we search for operating system signals.

Comparison of GSP and AprioriAll

According to Murat et al (n.b)),On the synthetic datasets, GSP was between 30% to 5 times faster than AprioriAll, with the performance gap often increasing at low levels of minimum support. The results were similar on the three customer datasets, with GSP running 2 to 20 times faster than AprioriAll. There are two main reasons why GSP does better than AprioriAll.

- GSP counts fewer candidates than AprioriAll.
- AprioriAll has to first find which frequent item sets are present in each element of a data-sequence during the data transformation, and then find which candidate sequences are present in it. This is typically somewhat slower than directly finding the candidate sequences.

GSP, a new algorithm that discovers these generalized sequential patterns and has the following advantages for example.

- Empirical evaluation using synthetic and real-life data indicates that GSP is much faster than the Apriori.
- All algorithms presented in GSP scales linearly with the number of data sequences, and have very good scale up properties with respect to the average data-sequence size.

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 5: The mining Algorithms of WUM

3.11. Navigation Patterns and Important to Discover

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)), we 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. In WUM, Mary et al, (2000), an “interestingness descriptor" is a query in our mining language, MINT.

3.12. Knowledge Discovery Queries

Similarly to Lukas, (n, d), we believe 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.

3.13. Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process in accordance with, challenge of pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users.

The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL or MINT query. 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. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match

CHAPTER FOUR: METHODOLOGY

4. Overview of the methodology process

According to Dipa,(2010), web usage mining have three main process in order to discover a knowledge from the data ware house, author of paper use for his work according to this researcher, described above, it is necessary to perform three steps, see fig 5,but the detail of those how to accomplish those main process are described below.

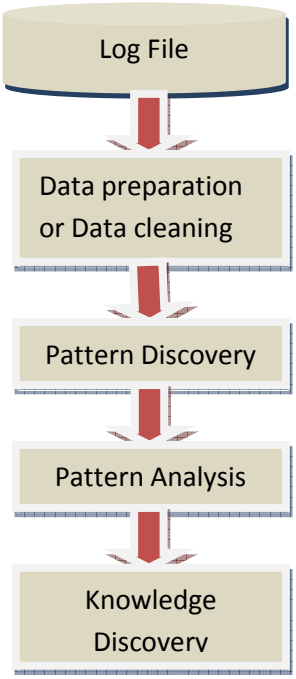


Figure 6: web mining usage main process to discover knowledge.

4.1. Tools Selections for Preprocessing

As stated earlier in chapter ,there a lot of tools uses for preparing a dataset for the intending purpose but the selection of those tools is not easy since every tool have designed for specific purpose but none of them cannot give a good output unless they combine each other in order to meet efficient output. The author 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.

The author 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 author found no need to implement his data preparation into navigational discovery software, besides to that even if the WUM have some capabilities of preprocessing, but does not support the main preprocess phases such as removing robot hosts and etc.

Extended log format of AAU

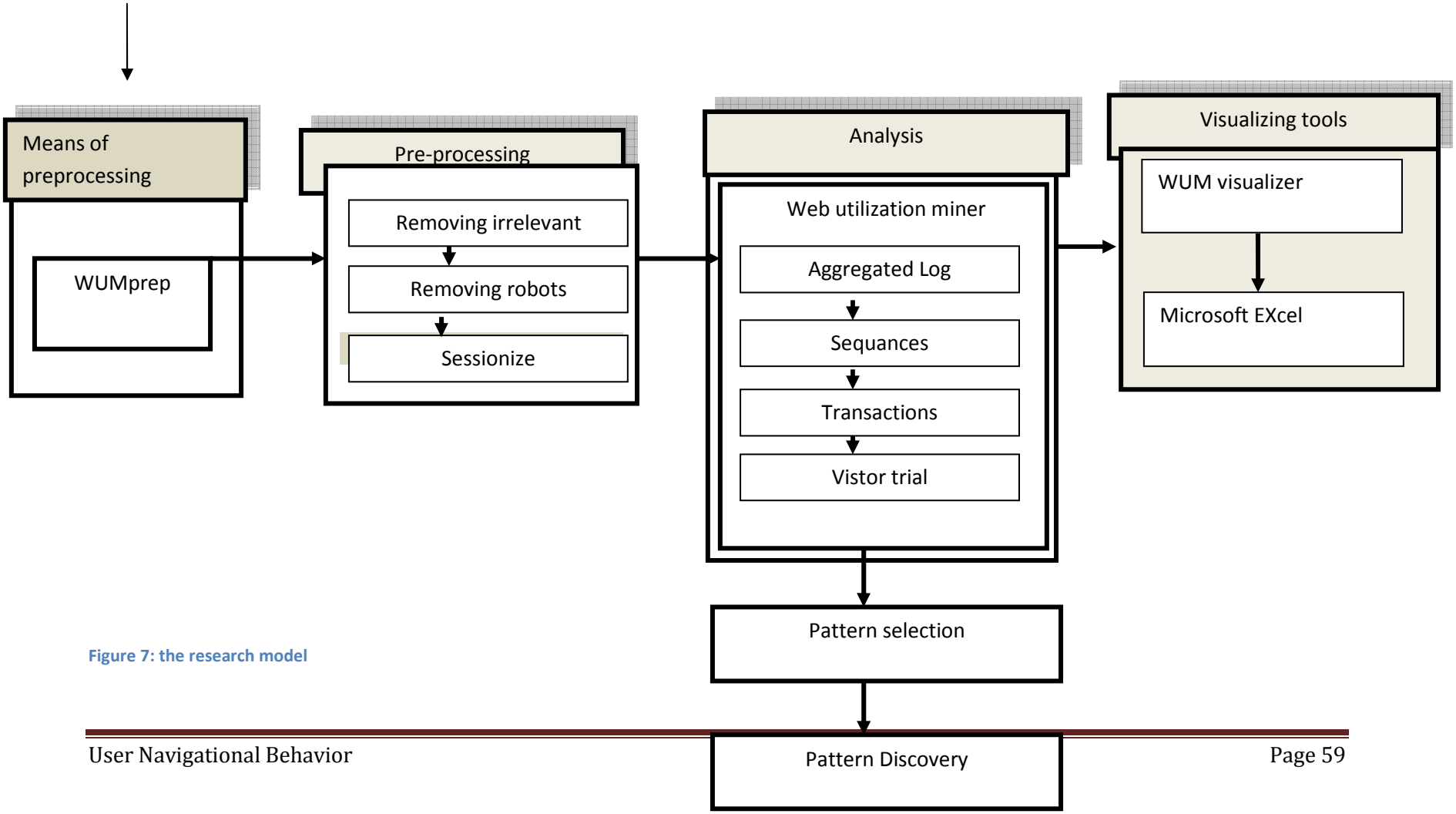


Figure 7: the research model

As it have been mentioned in the above figure 6, which shows how the objective could be achieve, even if the preprocessing done using the WUMprep, the researcher regulate the configuration of the tool to meet the objective . The data cleaning is done based on the following criteria.

4.2. Removing Irrelevant Records and Status

The removing of irrelevant records are significant as it have mentioned in the chapter three , as these requested log files are not only contain requests to the pages comprising the Web site, but also requests of images, scripts etc. embedded in these pages.

The author of this paper uses to remove those embedded extension of files should be removed because these “secondary" requests are not needed for the analysis and thus irrelevant (they must be removed from the logs before mining).Those requests are in the following table with their definitions:

\.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
\.JPG	The most common image compression format used by digital cameras.

Table 3: Irrelevant list of requests

Beside to that, the author only interested on request which only have the status 200 series, because concern only successful requests which mainly shows the users who get what they want and the other requests' are not need any more .

In general, according to the researcher these requests do not represent the effective browser activity of the user visiting the site; hence they are deemed redundant and should be removed.

4.3. Removing Robots

The author of this paper strongly believes to distinguish between human users and hosts that are robots, there exist several heuristics as it have mentioned above in chapter, section three. They are implemented in the script. Firstly, all records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed from the original log files.

4.3.1. Removing Duplicate requests

If a network connection is slow or a server's respond time is low, a visitor might issue several a successive clicks on the same link before the requested page is finally showed in his browser. Those duplicate requests are noise in the date and should be removed. The author of this paper uses, the most widely accepted threshold for of 2 seconds between two consecutive requests the entries that corresponds to robots can be eliminated.

4.3.2. Sessionize

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, the author uses a Session which is computed by taking any URL time stamp ,to achieve theses the researcher uses the most accepted time threshold which is 1800 sec or 30 min to identify the sessions using the these timestamp.

4.4. Divide log format

The preprocessed data needs to be dividing into manageable size before feed into WUM because it takes long time to process the data, so the researcher writes a python code to prepare the processed data for the WUM.

4.5. Tool Selection for Navigational Behavior

The transformation of the web server log into a log of sessions appropriate for mining and the process of navigation pattern discovery are performed in the framework of the Web Utilization Miner WUM, according to Anália et al., (2003), WUM (web utilization miner), Its primary purpose is to analyze the navigational behavior of users in a web site, furthermore, Navigation pattern discovery is 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 used as a basis to analyze the sessions in the rest of the log, comprising the sessions of the active investigators that did not become customers.
The architecture of Web Utilization Miner, There is two major modules: the Aggregation Service prepares the web log data for mining and the MINT-Processor does the mining.

In ref Bettina et al (1999), The Aggregation Service extracts information on the activities of the users visiting the web site 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. We call this the “support” of the node.

In accordance with Marya, et al (n.d), The MINT-Processor mines the aggregated data according to the directives of the human expert. “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.

In ref 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’s designed to work with The WUMprep module (which is responsible for the pre-process phase ;)
- Its free and open source tool (not commercial)
- 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.(is also open source and free)
- WUM uses for the discovery of navigation patterns and visualization of interesting Patterns.
- It’s a sequence miner and support GSP algorithms.
- 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.

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 web site 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.

4.6. General Methodology

The overall pictures of the methodology can be described as the following figure 7, the WUMprep scripts does the preparation steps(in the above illustration) , which is the input for the WUM tools discovers the navigation pattern and mining patterns and visualize the result using WUM visualize based on the miner interests.

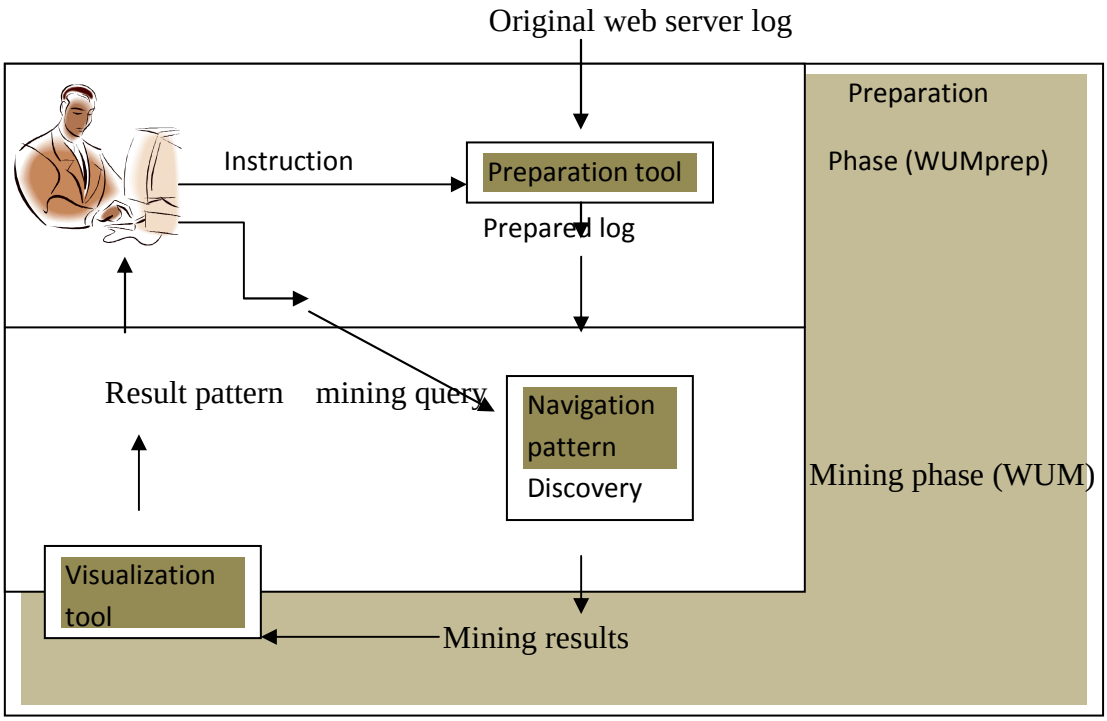


Figure 8: navigational process of WUM

CHAPTER FIVE: EXPERIMENT

5. Over view of Experiment setup

The experiment has been conducted on the following setup

- **Computer Type:** *personal computer (X32-based PC)*
- **Operating system:** *OS Name 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 the web log into manageable size

5.1. Data Collection and Selection

The data for this study is a web access log data of AAU official web site .As mentioned in the chapter one, a web log data is favored by many for web usage analysis. Two months web access logs have collocated for this study, for December and November.

5.2. Data Cleaning

The data collected from the AAU web server logs are full of junks that are not cleaned and should pass through some data cleaning phases (see the figure below) ,it is important steps to truck down the exact behavior of the user of the official web site unless they removed it is difficult to achieve the objective of this paper. Those phases must be undertaken to have cleaned data for further uses (process). The sample Log data are collected from AAU before preprocessing.

```
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)"
```

Table 4: A small extract of a Web server log contents

From the original web log see table 4, which can be easily seen a lot of junks, noises as well as robots (spiders, crawlers) those should be removed in order to have clean web logs to have appropriate ,efficient ,effective data logs.

5.2.1. Removing Irrelevant

As a result of removing irrelevant the number of log lines decrease in enormous seize the reason for it , those log files which contains the some extensions (see previous chapter), and those repeated requests that may came from inpatient users will be eliminated that's why the number of records seized in such amount. The original size of the records before were 50701 KB records (KB) and after the log filter preprocessing it became to 12416.KB.

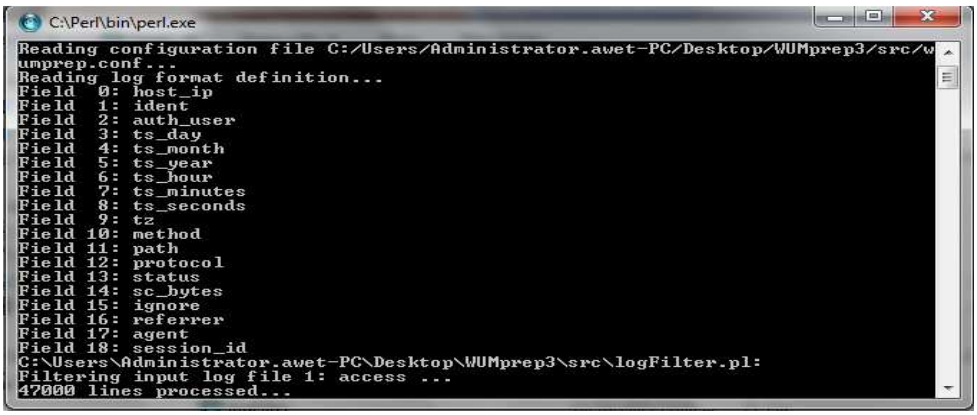


Figure 9: removing irrelevant records sample.

5.2.2. Detect Robots

The process of detect robots are very important to elimiatee the irelevent records whch are caused by the misusers that cames from other resources like (speider ,web crawlers) in other words , web surfing requested that are too fast that ordirnary pepole do not do in such fast ways caused by web crawlers.According to my expertment the number of robots are based on the maximum page view and against“index list“ in the WUMprep.The number of robots that detected from the web seerver logs are shown below,

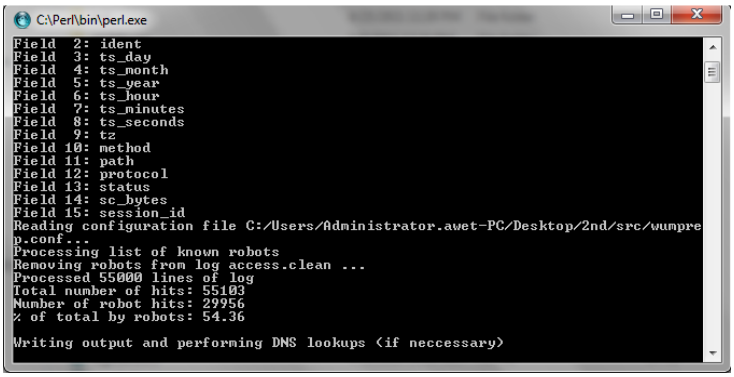


Figure 10: sample removing of robot hits

According to my experiment, for the months of December, the numbers of robots inside the Log format are 54.36 % robots from the total hits, for the months of November the total number robots are against the total hit are 39.68%. Samples of robot log lines that are 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] "GET /robots.txt
HTTP/1.1" 200 --304 "-" "Mozilla/5.0 (compatible; DotBot/1.1;
http://www.dotnetdotcom.org/,crawler@dotnetdotcom.org)"
(robots.txt)
```

Figure 12: Sample robot log files.

From the above results what it can be observable easily that some of the requests that came from same IP address that is (208.115.111.247) within two seconds, those

requests originated from the same IP address within two seconds. ([05/Dec/2010:05:03:20 +0300) and (05/Dec/2010:05:03:21 +0300)

5.2.3. Sessionize

The Sessionize which are resulted after the detection of the robots and give the following results as shown below,

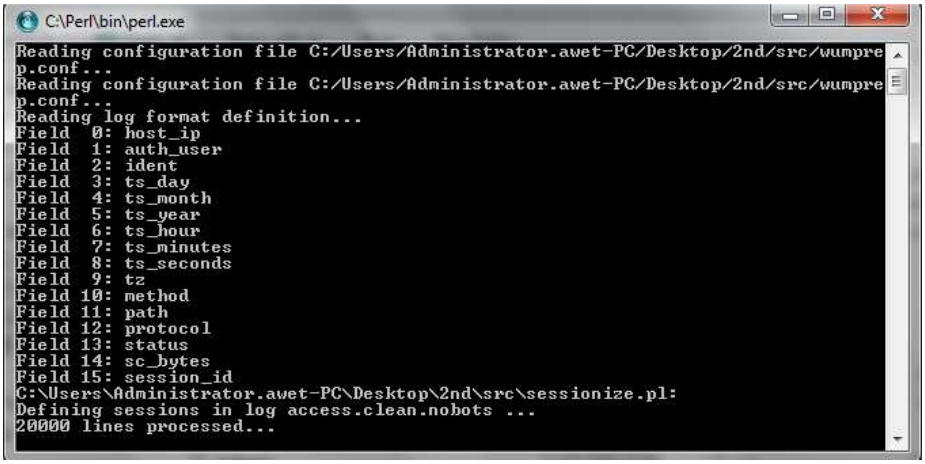


Figure 13: sample sessionaize process

The sesseinalize creates number of sesseionize, according to my experiment the number of session's creates are about 23411.some log lines which are exceed from the threshold i.e. 1800 sec or 30 min are removed .for the detail see in sample of in the appendix.

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

As it can be observable from the above fig 13,that the only status that filter from the web log files are GET and the status of 200 which indicates the successful requests from the web sites users, besides to that the session are identified . The types of log formats are converted from the Extended log format into Common log formats (see chapter Two, types of log formats).

5.3. Generalized Reports on Log Preprocessing

In this section the result of preprocessing will be discussed in general manner, an average user requests per day is 200220 lines. The preprocessing phases undertaken for both months (December and November) gives the following results after undergone through different phases of preprocessing for the months, and summarizes for one week in December the following tables. See for the months of November in appendix.

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
220340	150127	70564	25005	25005
230087	160743	72087	24060	24060
200406	148906	63480	21000	21000
190967	138967	50653	19734	19734
200190	178300	60752	20674	20674
200150	167543	47897	19653	19653
220205	120950	62096	23765	23765

Table 5: A Sample records for the week in December after undertaken the preprocess phases.

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

As it have been mentioned earlier the log files are contains irrelevant data, irrelevant records and noises, that's why we can observe from the above experiment result in the table the size of original log entry records decreased in average of 80%.for the months of November the size of records of original entry decreased in average of 73%.(see in the appendix).

5.4. Navigational Behavior of December

5.4.1. Aggregated LOG tree

The aggregated tree are results from the web miner after the sessions creates based on the above preprocessed tool (WUMprep) and imported to the miner resulted the aggregated tree for the months of December as follows.

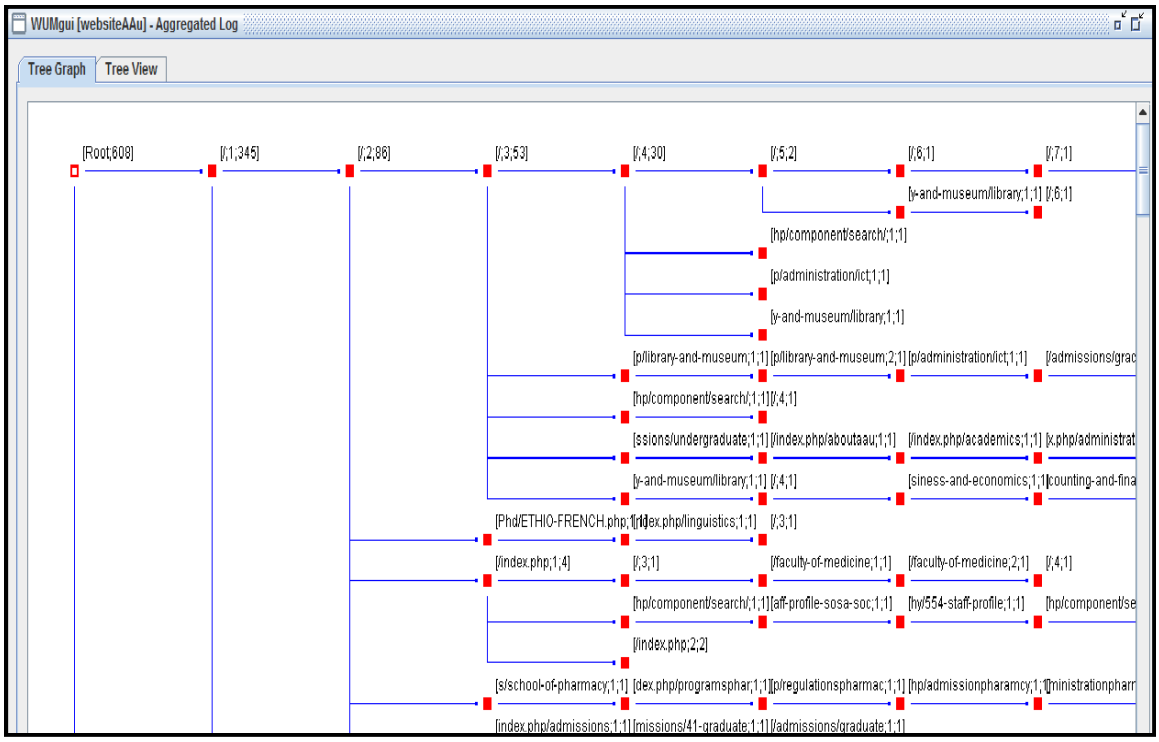


Figure 15: Sample aggregated tree for the month of December

As we can see from the above figure 14 ,an aggregated tree that the total number of nodes or total traverse make by users are 608 for the month of December, based on the aggregated tree the MINT query applied to find interesting pattern or for sequence analysis from it. The researcher chooses the some Examples to find interesting pattern for the month of December. For the month of November see in the appendix.

5.4.2. Sequence and Navigational Discovery of Users

As previously mentioned in chapter three, the generalized sequence pattern describes the behavior of users by filtering out the interesting pattern from the aggregated tree using the MINT query. ,In the following sections, the experiment is undergone using some most interesting patterns using the MINT query to discover the most important issues that should be discovered according to the researchers of interest , like Where do visitors of page Home go after wards?, Where do visitors go after typing the www.aau.edu.et (/)?, To Find out pages that always visit together and look at its pattern, Where do visitors go after search page of AAU (/index.php/component/search)? What is interested in navigation patterns between two pages.

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

Using the MINT (see appendix for syntax of MINT) query the author is interested where users go after the accessing the home pages until the next five pages, using the following query to the MINT to discover users' navigational behavior.

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 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
```

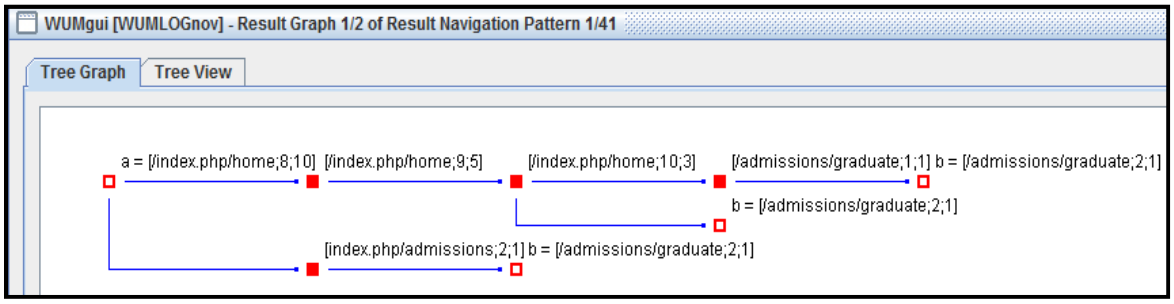
The above query results a following patterns using WUMvisulizer but the author puts some sample results in the following figure.

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

Here, we receive all pages reached within 5 pages after HOME (index.php/home), which has been accessed 100 or more times, provided that those pages have been accessed by at least 50% or 100% of the visitors visiting HOME, but as we can see from the result the most accessed pages is /index.php/library-and-museum users stay 100% visting the content of it, It is also clear that most users who visited the home page also stay in 100% within the page of /index.php/registrar those are the most .of course the other pages like /index.php/admissions/graduate users stay in those pages users stay in the page for average 26%,even if they are the most visited pages after Home pages.

Navigation pattern:



As we can see from the navigation pattern most people are going to the page of /admissions/graduate after visiting the home pages, it's clear to see that most users stay in

the HOME page (/index.php/home) and navigate 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 above MINT 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 receive 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%.

Navigation pattern

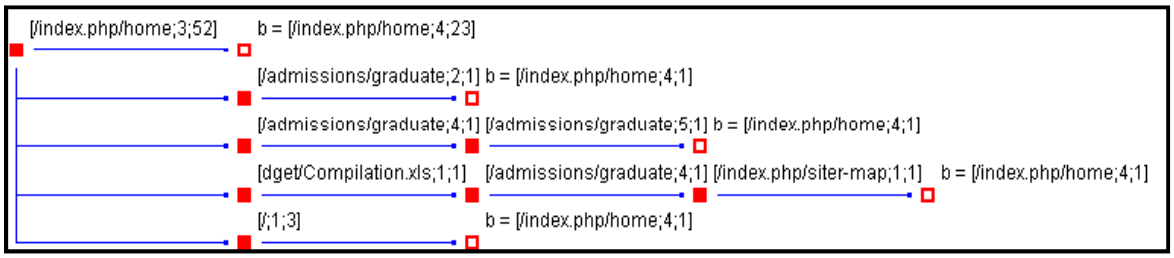


Figure 16 :Navigation pattern

From the figure 15, its easily observable here, we see that when visitor start from looking at **/index.php/home** page, 20% of them will stay within this subject area.

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

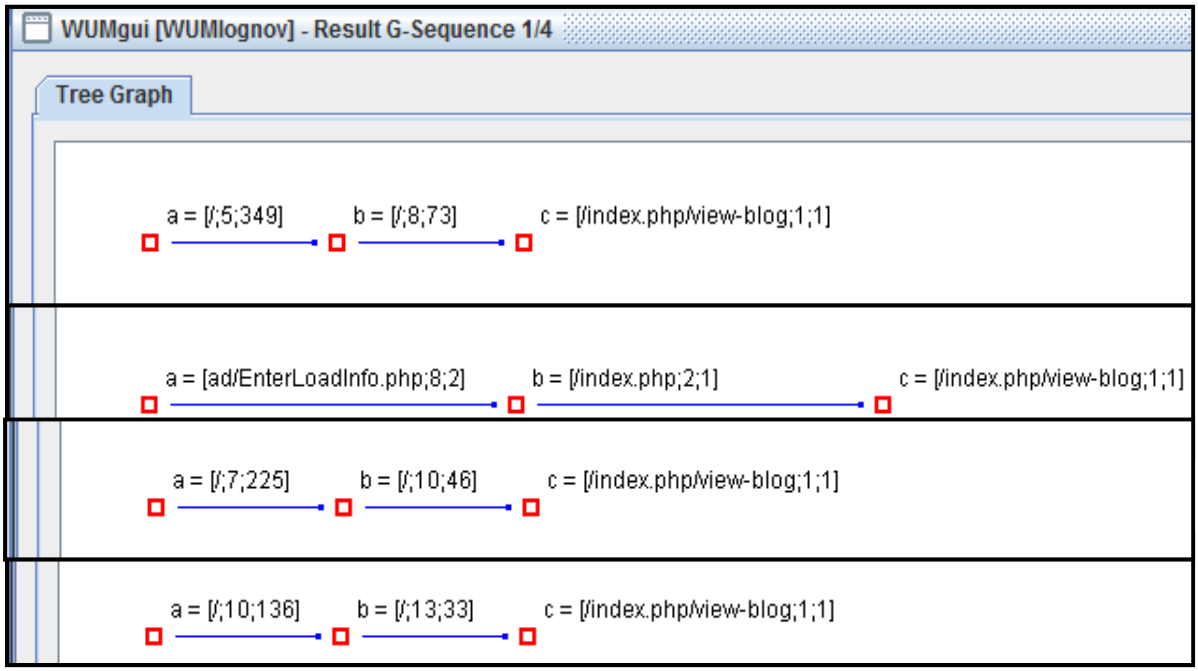
In this query, we specify a template t with two variables a, b, thus seeking for with two pages bound to a and b and at most 5 arbitrary page occurrences in between denotes that "a" should be bound to the first page which is /index.php/home, this page should be visited at least 20 % and b page should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

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 out of the query give us two patterns ,Here we recive 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 stay in the home page before reaching to /index.php/view-blog pages.

G-sequance



the Users do not take a single paths to reach to /index.php/view-blog most of the users take a path from the root pages,and the second most users take 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
```

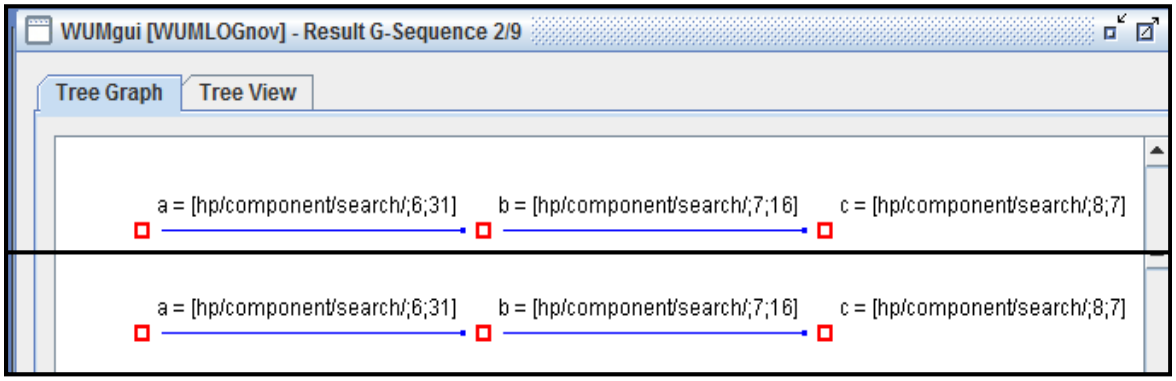
Pattern

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 ten patterns show that user's do know where they are looking for. most of users who stays in search engine 100% and also stay in this page for average of 40% ,they do search function stay within search the page.

G-sequence pattern



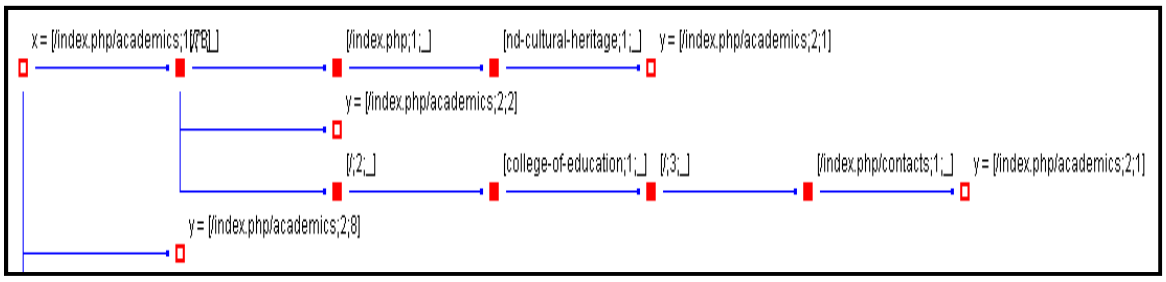
Sample of the above ,the author do not need to put all the g-sequence from the navigation pattern that users stay in the search page as we can see from the above result, that users stay in the search page.

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"
```

The above query results the following 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 web site.

5.5.Statistical Analysis for the Months of December

The WUM can generate a comprehensive report in terms of simple tables the researcher used other tool (Microsoft Excel) for better visualization. report will be discussed like what are ,most requested pages, most visited pages, most visited directory as well as most referee pages for the month of December will be discussed .For the month of November see in appendix.

5.5.1. Most requested pages

The following table shows the top ten most accessed pages during the months of November .For the rest of the months see in appendix.

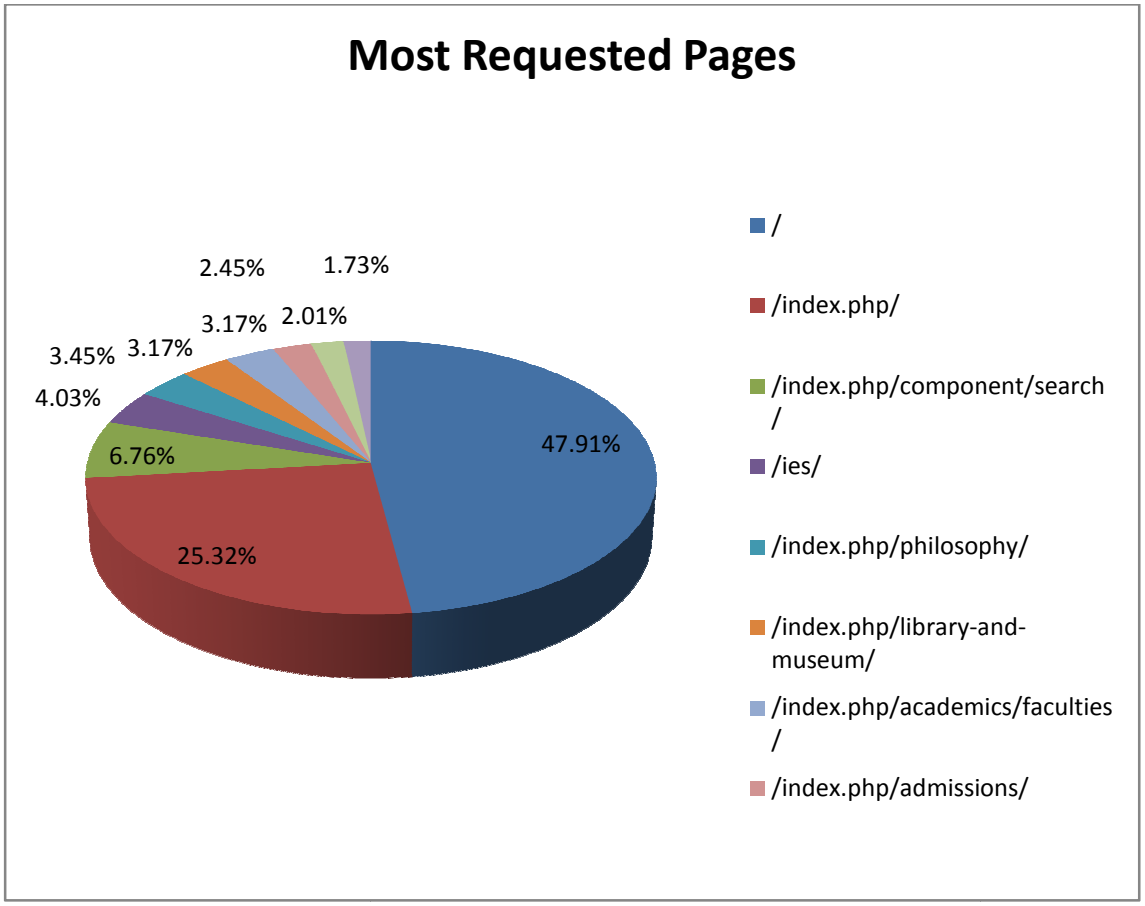


Figure 17: Top 10 most requested pages.

A figure shows the Top ten most pages during the months of December ,As it is shown the most requested pages is the / 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 is most popular by most users in all the three months .in fact ,this shows that most visitors enter into the site directly by typing the web site address as it shown in the above directory. The search engine of the Addis Ababa University the second most accessed pages followed by the /index.php/library-and-museum pages.

5.5.2. Most visited directories

The root directory "/" is the most accessed directory where the root directory in root folder is located .Most users also shows interest on the contents under the **/index.php/** folder. It is also possible to say that from the output those are also important visited pages /index.php/component/search/.

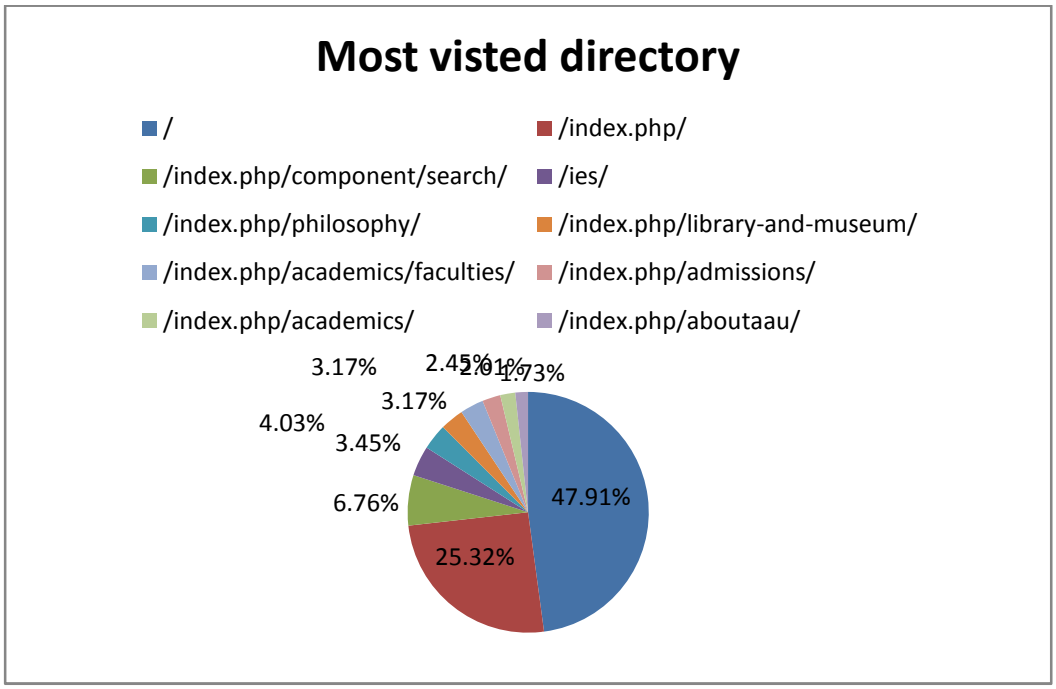


Figure 18: Top ten requested directories

For the rest of the months, most of the directory are requested ,are the same as the above until the top three directory but the others are became familiar in the next pages.

5.5.3. Most Top Entry Pages and Top Exit Pages

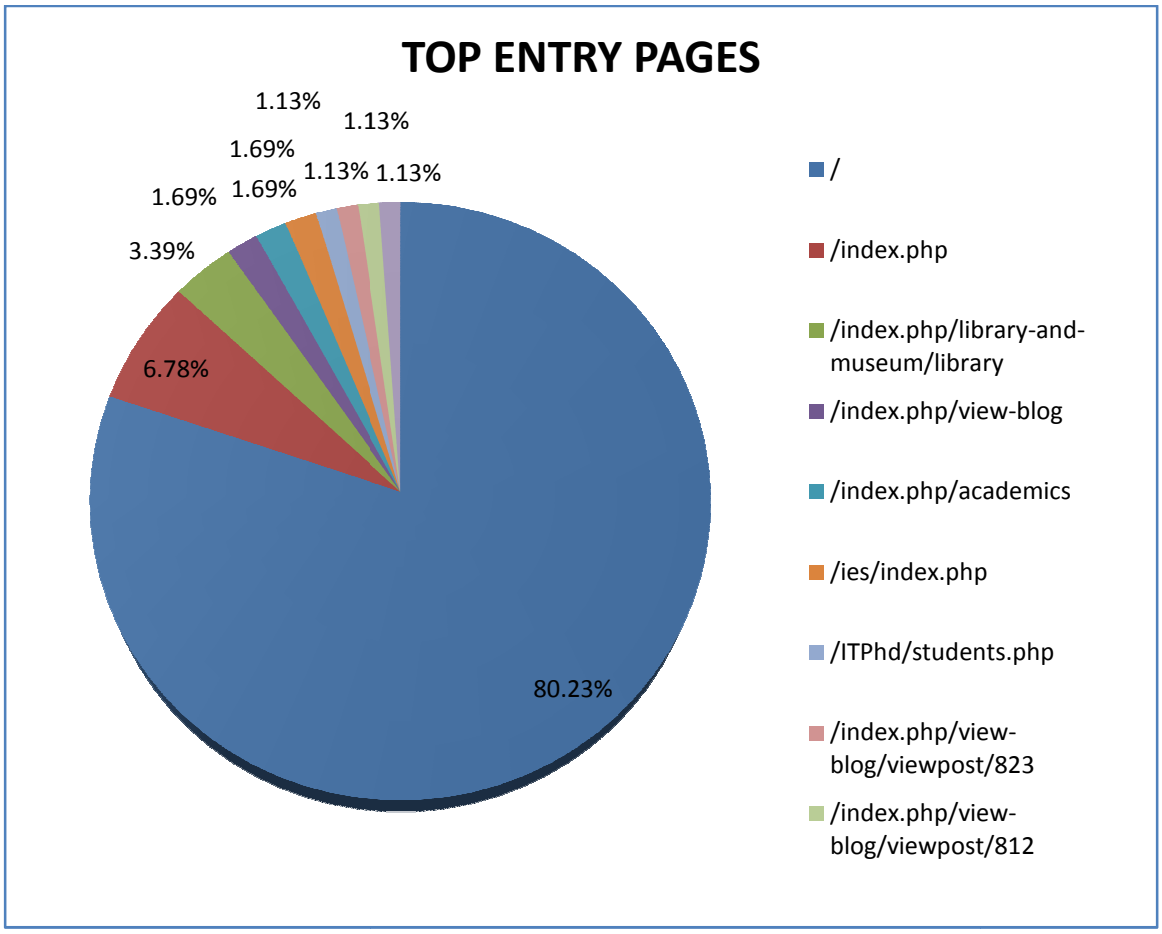


Figure 19:Top ten entry pages

The entry pages are pages that indicated that the web site users first visited where as the top exit pages is the last pages the users visited official web site .From the figure below what we can observe is that the "/" root pages where it is located accessed more than any other pages almost half of the request (80. 23%) and the /index.php the second most top entry page and last not least the /index.php/library-and-museum/library the third most top entry of pages, followed by /index.php/view-blog and /index.php/academics the 4th and 5th top entry pages. For the rest of the months see appendix of December.

For the month of November, as it as shown in the figure 55.68ve % of the visitors have entered into the web site directly through the / and index.php.

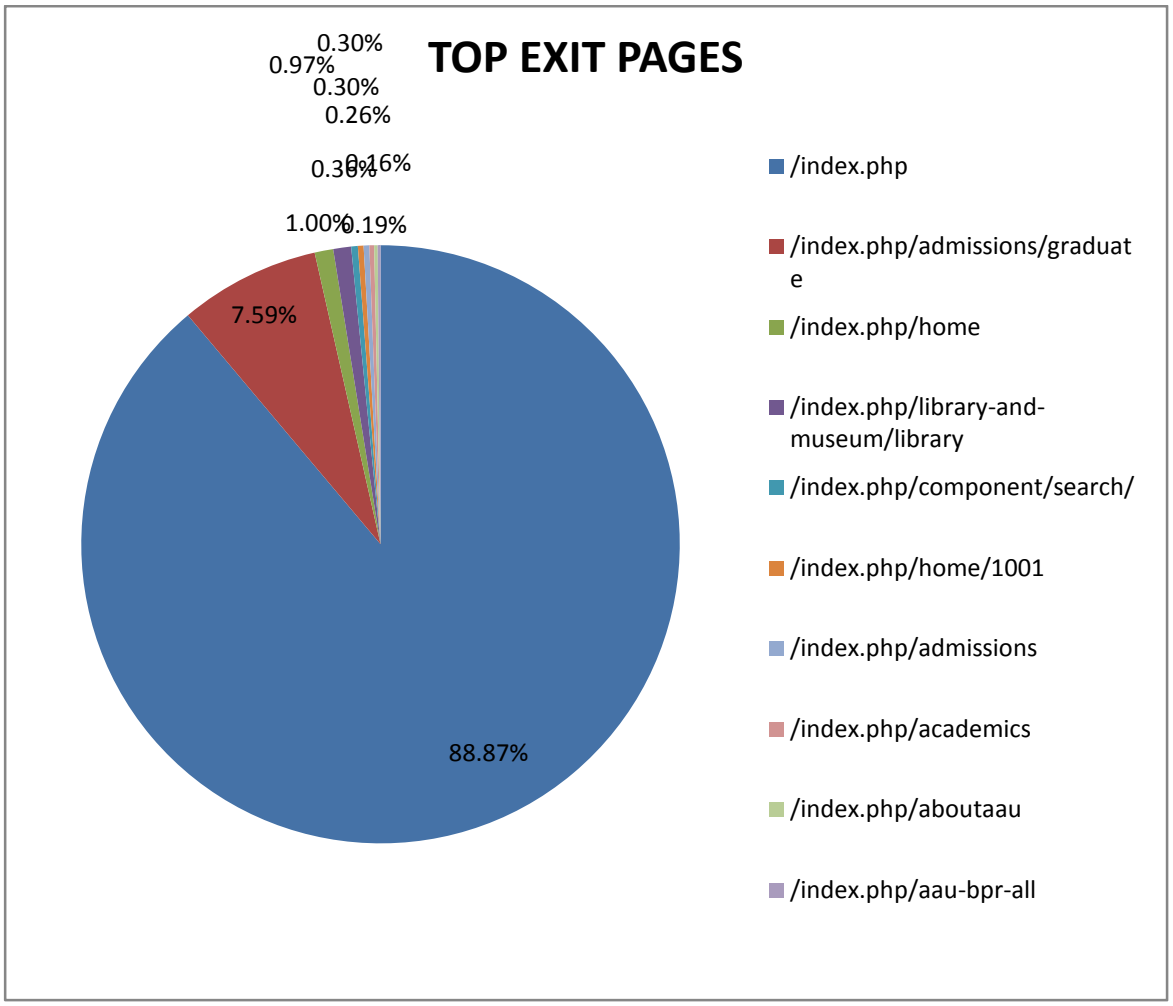


Figure 20: Top most exit pages.

From the above the figure ,we can see that the top exit pages are the "/" or after the user type the web site address and leave the web site 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/.

5.5.4. Top Referrer Pages

The top referee pages are pages where the visitor was located when making the next request with the official web sites.

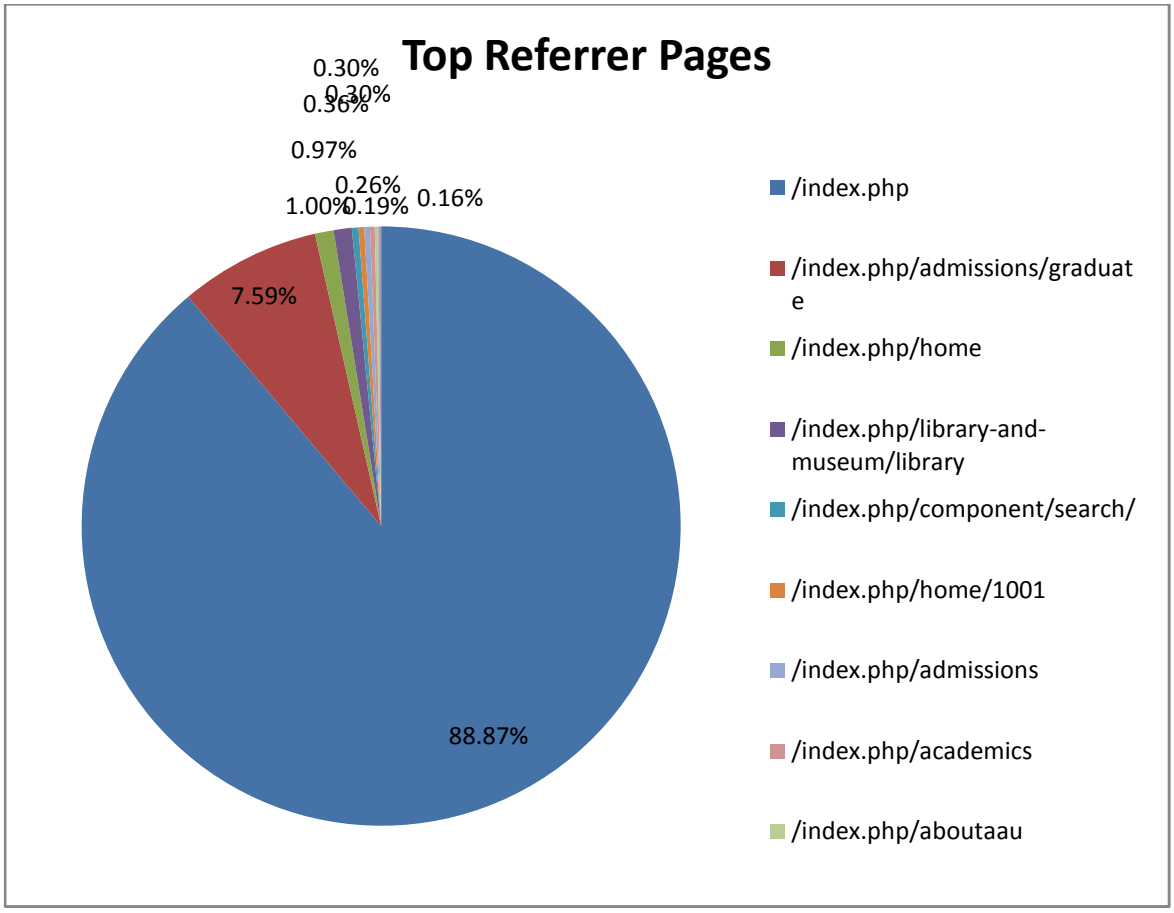


Figure 21: Top Ten referee pages.

From the above figure , what it can be easily observable that most users makes the next request from the page of /index.php which covers in more percentage 88.8 %. the next popular page where users requests the next page are initiated from /index.php/admission/graduate and which covers 7.33%.the third most refree pages are /index.php/home which covers the percentage of 0.96%.

For the months of November are almost the same as the above but the only difference are below three requests for more details see in appendix.

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATION

Conclusion

- From the navigational behavior of users that we can indicate easily users is no single point where users go after home page and can be conclude that users navigate from top of the page (hierarchy) to the lower hierarchy.
- From the navigational behavior search behavior can be conclude that most users use the search engine effectively or know what they are looking for.
- most request pages are requested to the web site by typing the official name of the web site that is www.aau.edu.et,that's why the most request web page becomes the root page of course it clear that the web server is an apache server, when type the official name hit the root directory of the official web site .from the request pages the second top most requested pages are /index.php/component/search pages, it indicates that most users use this page for searching key words with in the pages. What else can be concluding that /index.php/library-and-museum the third request pages, can be conclude that most users are interesting in the content, the reasons could be most journals associated to it.
- Most visited directories are of course the root directory since most of users are typing the name of the official web site 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 in side it.
- most users use enter to the web page using the page of /index.php , /index.php/library-and-museum/library, /index.php/view-blog and most users also leave from those pages that it can be conclude that almost the other pages 1/3 ,most users leave without visiting the web pages.
- It easy to see that most users use the /index.php, /index.php/admissions/graduate, /index.php/home are most users requests from those pages to make for further

requests, so it would be very useful if the administrator can put some urgent notice and advertisement within those sites since they are most accessed web sites.

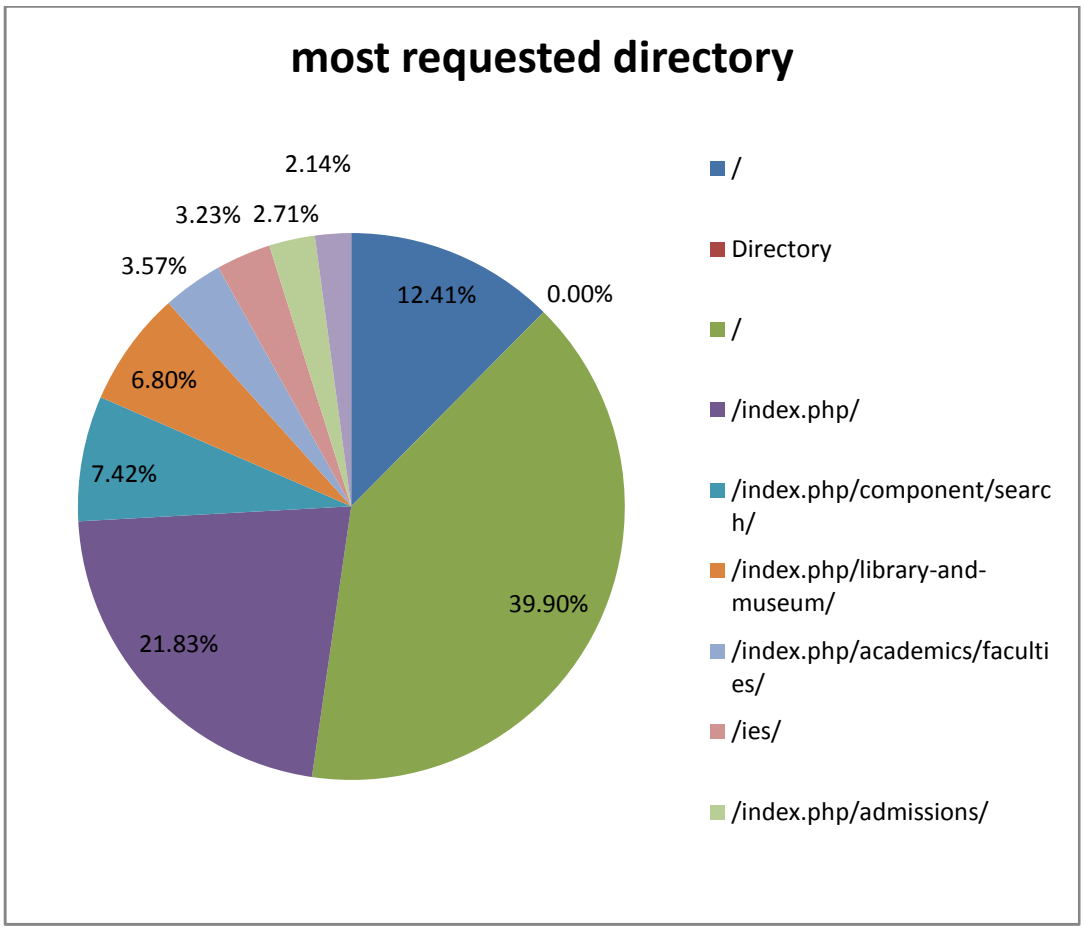
Recommendation

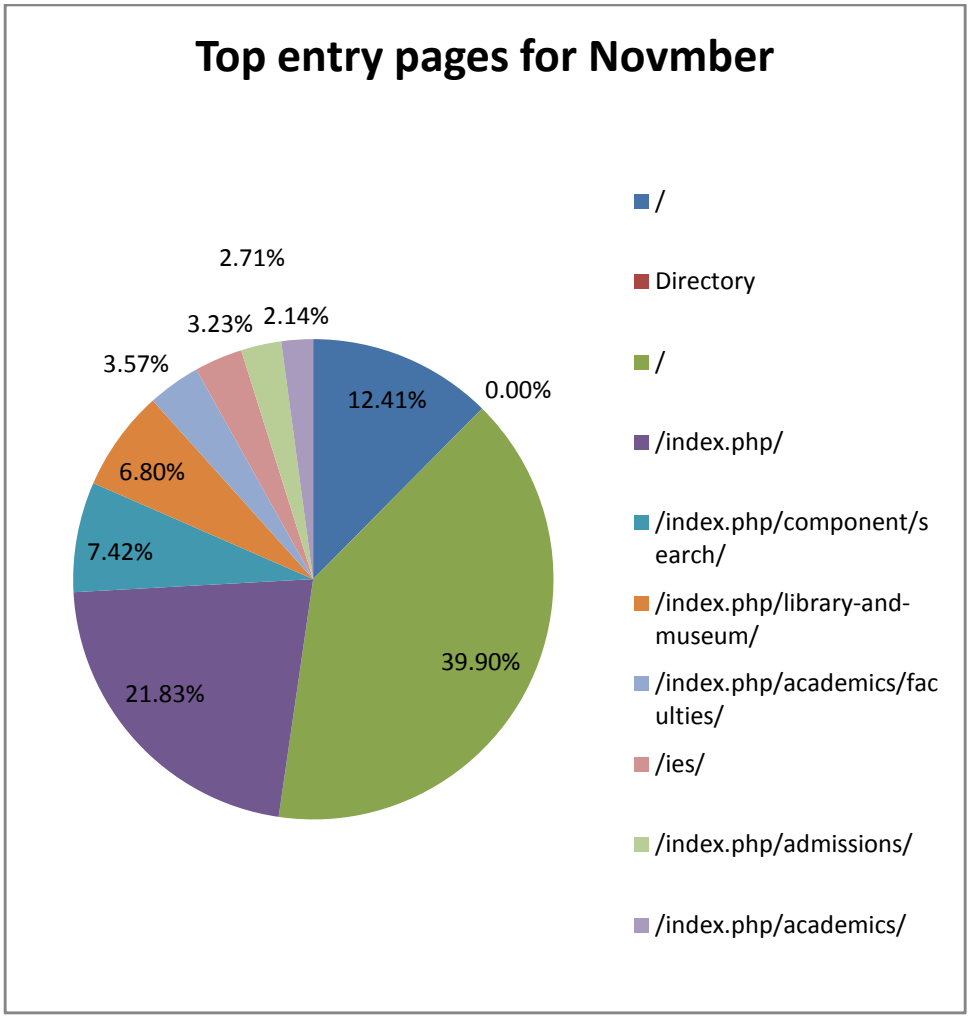
- Most users come to the page web site directly by either typing the name or from the search engine that displays the home page .This could be an indicator the web site has a kind of sickness .the web master therefore should do some kind of assessment on the department index pages make sure that those pages contain those key word for indexing in search pages.
- The most together accessed pages are the home pages is accessed with itself so it is important that ,It also important to recommend that the concerned body that is in charge of AAU official web site design should create quick links from one to other pages for those pages mostly accessed to gather.
- It is also clear that most users left the web site from some pages mainly from the /index.php/home ,/index.php/admission/graduate,/index.php/academics from it, it possible to recommend that the web master should use those page for advertisement and notice and also possible to recommended further it is possible to link to other department links in order to encourage web site users to stay in the web site.
- It is possible to recommend that the web administrators should make the most accessed pages, to be prefetching or cached to prevent the latency of the network bandwidth or prevent delay to access those pages.
- From the navigational behavior 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.
- 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” as a consequence could be better result.
- The other recommendation for further work, divide the web page based up on concept of hierarchy which concept divide pages according to the service they provide, once hierarchal classified the pages it would give better result.

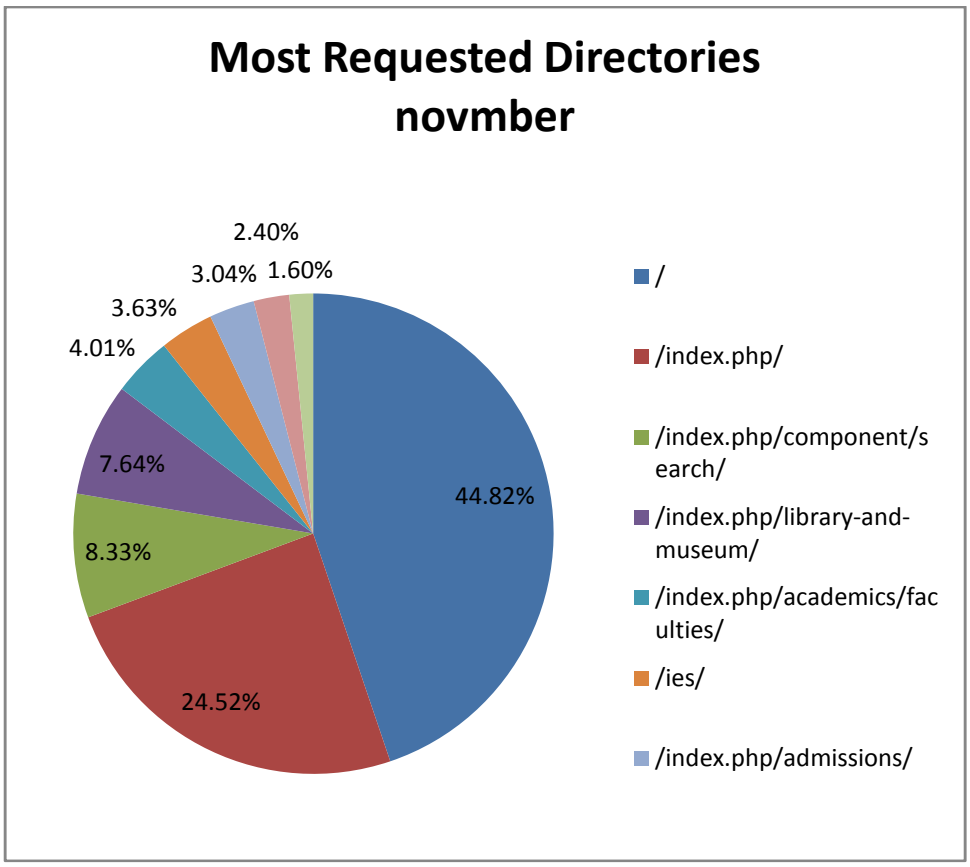
- The last not the least, recommendation for the further work, since by combining different technique of web usage mining such as content mining with web usage mining (work of this thesis) it could give better result in terms of efficiency .

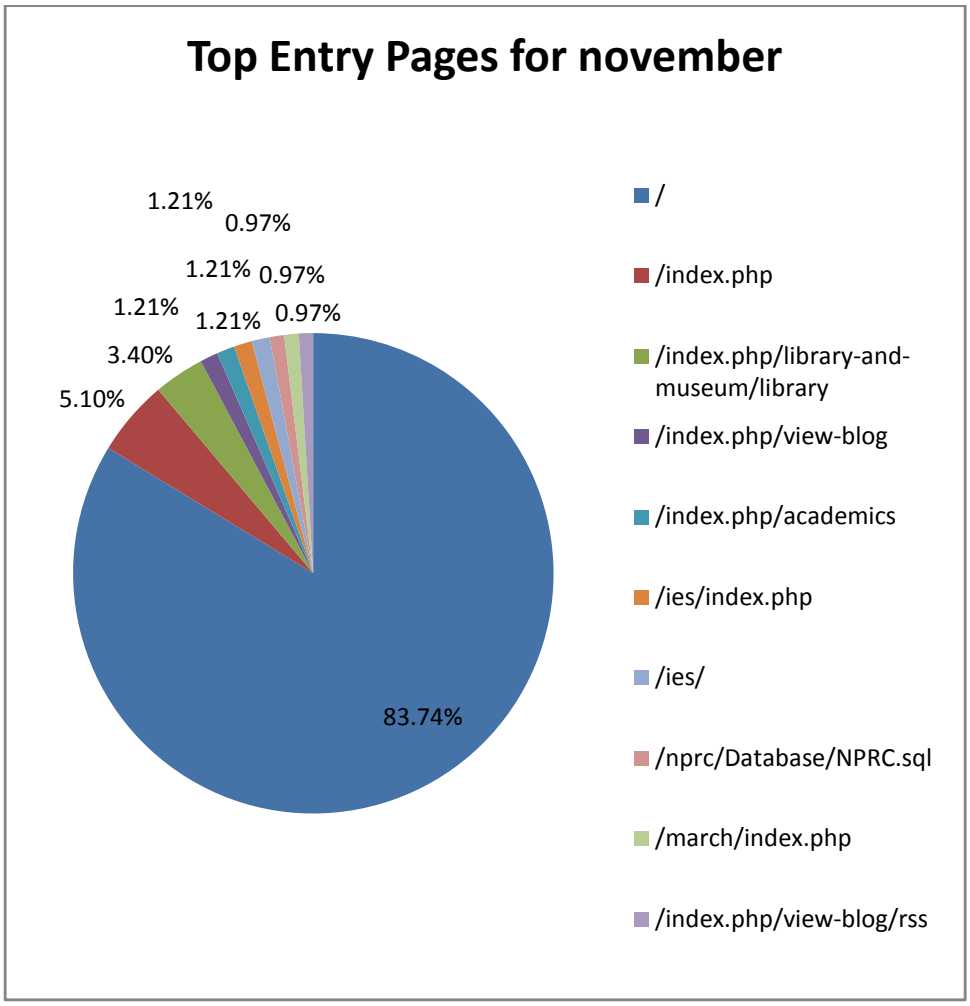
Appendix A: statistical report for the months of November

Appendix for month of November: Most Requested Directories for the months of November




Top entry pages for Novmber
12.41%
0.00%
39.90%
21.83%
7.42%
6.80%
3.57%
3.23%
2.14%
2.71%
/
Directory
/
/index.php/
/index.php/component/search/
/index.php/library-and-museum/
/index.php/academics/faculties/
/ies/
/index.php/admissions/
/index.php/academics/


Most Requested Directories novmber
2.40%
3.04%
1.60%
3.63%
4.01%
7.64%
44.82%
8.33%
24.52%
/
/index.php/
/index.php/component/search/
/index.php/library-and-museum/
/index.php/academics/faculties/
/ies/
/index.php/admissions/



The following are also the sample of one week for the month of November the results of those as explained in chapter 5.

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
210240	140127	69564	20004	20004
240067	160743	72087	24060	24060
203406	148906	63480	21000	21000
200967	138967	50653	19734	19734
200190	178300	60752	20674	20674
200150	167543	47897	19653	19653
220205	120950	62096	23765	23765

Appendix B: 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 C: 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 web sites 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 Web Sites 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 Web Sites. 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 web sites. 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 Web Site 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.

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

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.

Acknowledgements

There are many people that I need to thank for making this long journey so memorable. First and foremost, I would like to thank my advisor, Ato Workshet lemnaw, 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.

Of course, my thanks to Professor Bettina Berendt for her borderless support in giving directions on this work ,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.

" fekri Belibi"

Abstract

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years. Academic researchers have developed an extensive array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular web site. Performing this kind of investigation on AAU web site can provide information that can be used to better accommodate the user's needs.

The Web Use Mining (WUM) , it corresponds to the process of knowledge discovery from databases (KDD) applied to the Web usage data. It comprises three main stages: the preprocessing of raw data, the discovery of schemas and the analysis (or interpretation) of results. A WUM process extracts behavioral patterns from the Web usage.

In this thesis, we find out the navigational behavior of the user of official web site of Addis Ababa University web server recorded in web server for two months (November and December), those recorded are raw data that are full of junks, noises and irrelevant data contents .In this paper present a preprocessing tool WUMprep that uses to filter those unnecessary data, such as irrelevant records, noise data, and it crates the sessions based on specific thresholds.

For discovery of navigational behavior, here presents the Web Utilization Miner WUM, a mining system for the discovery of interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the researcher using WUM's mining language MINT, using those descriptor it can be describe the general behavior of users instead of single users behavior using the most appropriate algorithms known (Generalized sequence pattern) which implemented in WUM.

The General behavior of users constructed by GSP algorithms those behaviors are descried using the MINT query. Those MINT query are intermediate between the users and pages.

The researcher of this paper also recommend that to get a better result by combining the web usage mining with content mining techniques of web usage. Of course without any doubt it could give a better result in terms of efficiency and effectiveness results.

Table of Content

List of Table

List of Figures

Web Terminology and Definition

In accordance with 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.

Abbreviation

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
IBM	International Business machine
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

CHAPTER ONE: INTRODUCTION

1.1.Background

In 1990 the internet was initially designed for exchange mails between users later it becomes trendy for use of WWW. The www or 3w in now popular services among almost any other services the internet provides. There are number of services providers (ISP) for the use of the internet across the world. In Africa, the number of the internet users increasing and increasing from time to time. 5.6% of the world internet users are from Africa, further explained, it shows 2,357.3 %growth from the year 2000-2010 similarly, Ethiopia has 0.4 % share among African internet users .Even if this seems insignificant when it compared with the rest of the world, generally speaking the number of the internet across the world getting increasing and increasing in dramatic way thorough out worldwide1.one of the various reasons for the development of the internet in Ethiopia causes by huge amount of investment in infrastructure like in education ,telecommunication and development in others sectors.

Addis Ababa University, one of the oldest higher education institutes 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.

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

[1] http://www.internetworldstats.com/stats1.htm#africa

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]

There are number of services provided by the Addis Ababa University, one of the popular services are the WWW(world wide web) among other services like teleconference ,data service ,those web services are divide in two as the official web site (internet) and intranet which is not able to be accessed outside the university which uses for local uses. The official web site accessed through the public Ip address offered by the ETC.

The official web services of AAU an organized collection of Web pages information is presented in various formats , ranging from research papers, and educational content, to multimedia content, blogs .that's why the getting information from the official web site is the matter of click-streams in the internet of course if there is connectivity. 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 a few individuals who realized that the AAU would be wise to join the information age by adopting the technology that has been transforming the world. The newly formed office initiated a project named AAUNet that has resulted in a wide area network (WAN) whose first phase of construction was 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 is adequately served by ICT. At the same time, AAU's need for and dependence on effective ICT support is now greater than ever.

[2] www.aau.edu.et/administration/DRAFT ICT POLICY AT AAU
[3] www.aau.edu.et/administration/ICT

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.

1.3. The AAU Official web site

The Addis Ababa university official web site was published around some seven years ago .As the ICT development office of AAU (which have mentioned in previous section) is engaging in ICT related works ,the official web page develop and maintain by this office. the web site is hosted on AAU's own server which is located in main campus of the university (6 kilo), The official web site have the domain of www.aau.edu.et and have statistical IP address.

1.4. Purpose and User Community

The official web site being in work to deliver information both the university activity, in general and about academic and administrative units, in particular, it also delivery information about news, items and its own advertisement for both vacancies and student admission and other, of course it has also some external links to other web and other sites such as collaborative organizations in research activity donor agencies, etc.

1.5. Nature and Content

Generally the web sites designed bear in mind to support the objective of the university. In sections try to discover the nature and content of the web site. The AAU web site has both static and dynamic nature .there are few web sites that are static in nature those pages are not interactively with its users but the majority of the web pages are dynamic in nature which are support the MYSQL database incorporate with JOOMLA packages helps users to interact with web sites users.

When we came to Web site content it posts numerous information regarding to the objective of university which presenting information on several topics and issues, each page have information regarding to the objective of the pages .there are few page which are under construction(content not yet update), but there are advertisement and notice on several pages.

1.6.AAU Web Structure

In the following hieratical graphs displays web site structures of the official web site.

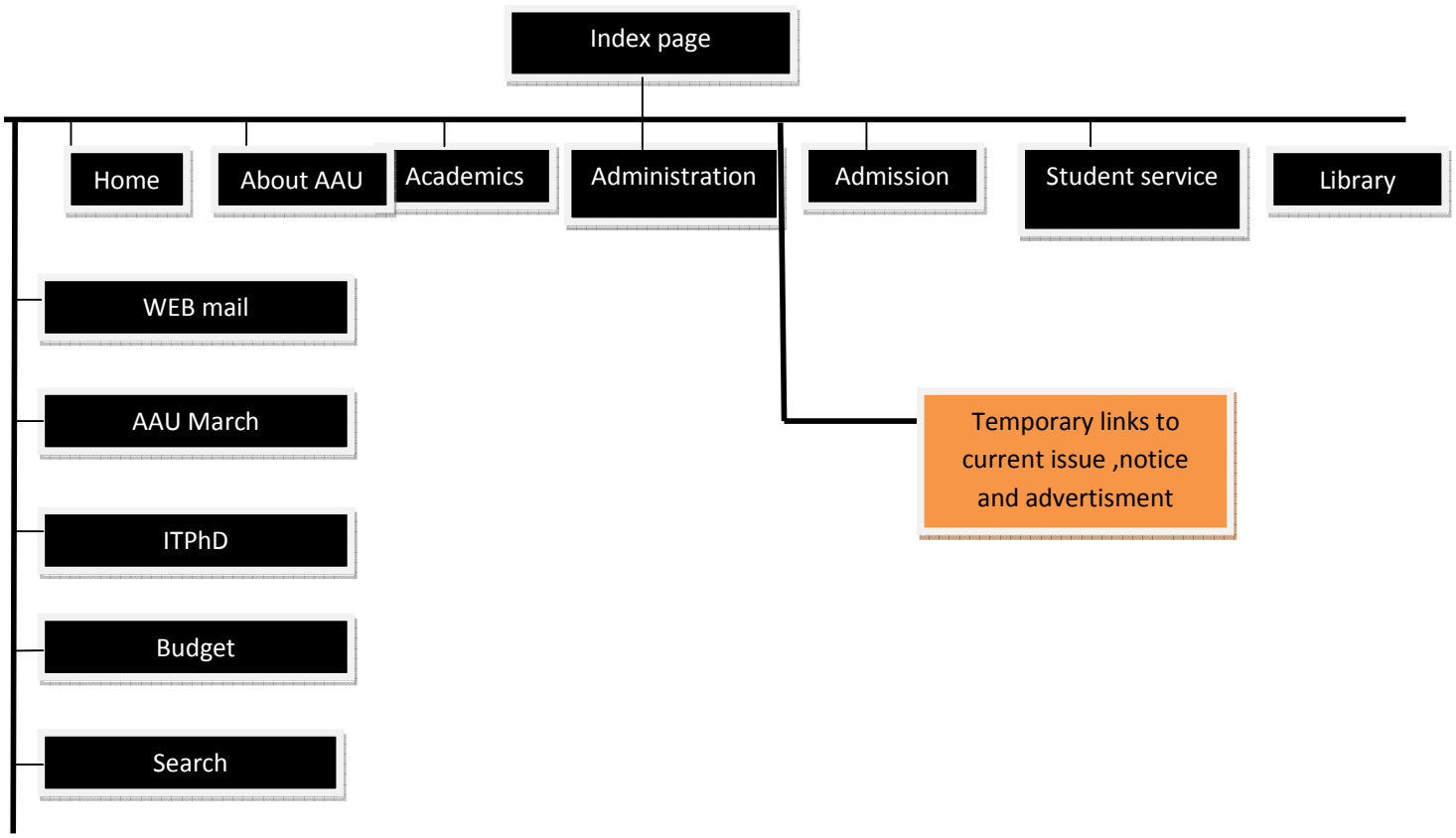


Figure 1 the structure of the official web site of AAU.

There are some other web sites that are accessed to gather with the official web sites like AAU march, ITPhD, college of education, IES (Institute Ethiopian Study), virtual accessed using the main web site.

1.7.Statement of the Problem

Rise of the Internet gave many companies an access to the 'gold' channel. Trading, putting gigabytes of information and communicating online has become one of the sources for understanding of the web users. 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.

As statics shows the number of web sites published every day is increasing quickly still, there are now 184 million registered domain names worldwide, a 9% increase over the same period last year[4].

On the other hand, the education sector is rapidly evolving and the need for web information Places that anticipate the needs of their information seekers are more than ever evident. The need of placement information is not easily imaginable we have to explore where should be places some information in a given web site, in this case of the official web site of AAU. It is important to know the navigational behavior of the users based on the study of the behavior. the need of study of any behaviors scaled up from the taxonomy of animals ,plants and others , in general, further explained that animals classified in to mammals ,vertebrates based up on the whole group behaviors.

According to Mokenen (2001) who were working on web usage mining of the official web site of AAU using the tools of wumprep4weka, for preprocessing or cleaning the data and Weka tool for data mining of the interesting pattern using the aprior algorithms finds out the most frequent access that do not based up the sequence, based on his study he did not truck the general behavior of users.

Like it discussed earlier uses the sequence (generalized sequence pattern) can tell the general behavior of users on navigational behavior of the user of official web site of Addis Ababa University, and not work have been done yet on the topic as to the knowledge of the author.

Web site design is currently based on thorough investigations about the interests of web site visitors and on less investigated assumptions about their exact behavior. In Lukas, C., (n, d) Concrete knowledge on the way visitors navigate in a web site could

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

prevent disorientation and help owners in placing important information exactly where the visitors look for it.

1.8. Scope and Limitation of the Research

Web mining has different branches: web content mining, web structure and web usage mining .the focus of this research is on mining usage pattern of AAU official web site .usually, three types of web related log files, namely web access log, error log and proxy log files. however, in this research work, web access log records is used as dataset because many literature and previous research justify that web access log files is the typical source of navigational behavior.

The limitation in this paper is the lack of manual on how to operate the web mining tools (WUM) and besides to that the web access log stored in Addis Ababa university are erased at the end of every months that's why it is difficult to get a enough data for the research, besides to that the web mining tools need to have a higher capacity (memory) to process the whole log files as batch.

1.9. Justification of the Research

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.

The importance of the study web users further explained by Marya, et al, according to him ,most web sites 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 web site they present to their visitors.

One important data source for the study is the web-log data that traces the user's 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, the importance of www further explained in Kosala et al, (2000), the web log creates an opportunity and encouragement for all Data mining researchers, consider it as the largest data warehouse in the world.

In accordance with 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's why performing data mining to optimize the performance of a Web server. In ref of Lukas, C., (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.

In accordance with Narendra, et al., (2003), Web mining is defined "*as the use of data mining techniques to automatically discover and extract information from web document and services. "*
Furthermore, there is also a widely accepted definition, According to Zalane, et al ,(1998).

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 web site is being used as expected*".

From the above the definitions web mining attempt to get the information (knowledge) or to extract the pattern, for the purposes to have an intended knowledge, so some the techniques should be applied to different web resources to overcome the problems, in ref with Mobasher et al, (1996), web mining is a common term for three knowledge discovery domains that are concerned with mining different parts of the web: web structure mining, web content mining, and web usage mining.

In general, User behavior has two aspects, one concerning the interests of the users and the information they access, the other concerning the way of accessing this information. The first aspect is addressed by techniques for the establishment of user profiles and is not peculiar to web usage. For instance, student profiles are considered in intelligent tutoring systems, the second aspect is addressed by techniques analyzing web server logs.

For example, consider a user that explores the links in a web site to find every bit of information of potential interest and a user that prefers keyword search. Those two users need fundamentally different support, even if both of them are interested in solar energy collectors, chess and medieval sculpture, In this study, concentrate on the second aspect of user support, namely on the analysis of user navigational behavior, because web users is characterized by her/his interests and by her/his navigational behavior.

1.10. Objectives

1.10.1. General objective

The general objective of the research is to apply web mining techniques for discovering of navigational behavior of AAU official web site usage of to reveal previously unknown the interesting, and actionable patterns based on the web access log file in order to recommend possible measures for further r improvement of the official web site of AAU.

1.10.2. Specific objectives

To achieve the general objective of the research, there are specific objective should be addressed, the specific objectives of the research are:

- To review literature review in the area in order to put concrete background and justification for the research.
- To identify and collect the data
- To prepare those data set using different preprocessing techniques.
- To analyze the navigational behavior of the users.
- To analyze the sequence of the web site 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.

1.11. Research Methods

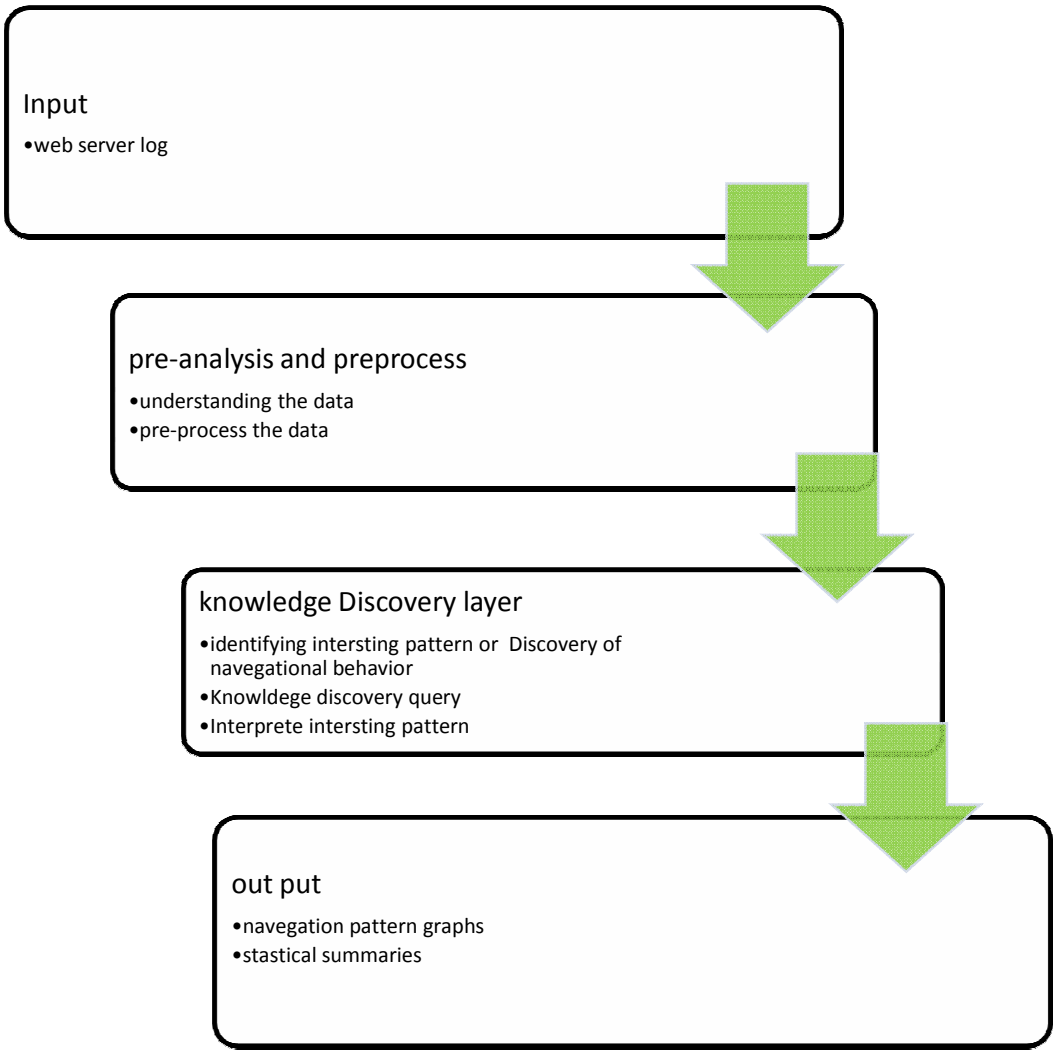


Figure 2:Research method flow

1.12. Data Collection for the Study

In this study the data has been collected from the official web site of the AAU, which is normally secondary data source since web log records every activity of the user regarding to visit of the web site.

1.13. 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 author of the paper, uses server data that are kept in the official web site of AAU in the format of extended log format, which is most apache server supports it.

1.14. Data preprocessing

According to olfa,et al, (n.d) , most log files are full of junks that are insufficient, inconsistent and including noise so the data pretreatment is to carry on a unification transformation to appropriate sets ; to have those sets there are some data cleaning phases are important to implement.

1.15. Data Cleaning

In ref olfa,et al,(n.d), the purpose 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 accordant to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

In addition to the above those also include some phases like, removing robot requests (filtering out spiders or crawlers which are known), removing duplicate requests (removing “dust”), and Filtering relevant status.(those concepts will be described in the Chapter Three).

.

1.16. Data analysis

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

1.17. Tools for Experiment

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, and Anália, et al., (2003),WUM (web utilization miner), Its primary purpose is to analyze the navigational behavior of users in a web site, furthermore ,Navigation pattern discovery is performed on the portion of the web server log that contains the sessions.

The justification for why these tools are selected is given in the chapter FOUR.

1.18. Interpret and report result

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

1.19. Application of results

The hidden unknown information in log formats are important in understanding of users navigational behaviors even if it is not possible to know what will be the results but some knowledge will be revealed by understanding of the general behavior of web site users of AAU .it can be used for improving the web site and it shows some way for further study.

1.20. Organization of the Thesis

This thesis organized as Six chapters ,the first chapter deals with the general introduction to the research of the area in this case the AAU, including the background of the Addis Ababa University in general, it also looks on development of ICT, and how looks like the structure of the official web site, what are their main purposes and later discusses statement of the problem, data collection ,data preparation with other subtopics like, scope and limitation of the study; objective of the study; research methods; etc.

The rest of this thesis is organized as follows. Chapter 2 presents two main areas, Literature review and related works regarding to Data mining and web usage mining.

Chapter 3 this chapter mainly deals with web usage and navigational behavior based on extended of the above chapter in terms of concepts.

Chapter 4 this chapter provides with methodology, in this presents the researcher points why select the tools for preprocessing and the tool for navigational behaviors in general, research process how to achieve the objective.

Chapter 5 in this chapter the experiment conducted and discussed which are based up on the methodology in the previous chapter.

Chapter 6 the last chapter, based on the experiment done in the previous chapter, the conclusions have been reached and recommendation and what it should be done for the future or further work in this research area.

CHAPTER TWO: LITERATURE REVIEW

2. Introduction

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), according to 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

2.1. 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. In Accordance with Berkan, y., (2002), the difference between them is that the former does not store Referrer and Agent information of the requests.

According to Srikant, et al, only few fields are available for navigational patterns discovery, which If are added to the CLF make up the so called Extended combined log format (supported by Apache Web Server).

2.2.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.

I. 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
```

II. 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)"
```

III. 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"
```

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

IV. 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, -,
```

2.3.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)"
```

The details of the fields in the entry are given in the following section.

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:

-

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

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 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)

2.4.Overview and 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 data base 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), refers as Data mining is 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 they can be 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 is often referred to as "secondary" data analysis.

2.5.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 is 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 is scheduled to depart, is 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).

2.6.Data Mining Approaches

It have mentioned earlier that the web usage mining is the application of data mining .those Data mining have two approaches according to (brendit,2011), the approaches is 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."

But, 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.

2.7.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 ref 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.

2.8. Taxonomy of Web Mining

In ref 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 2.8.1.

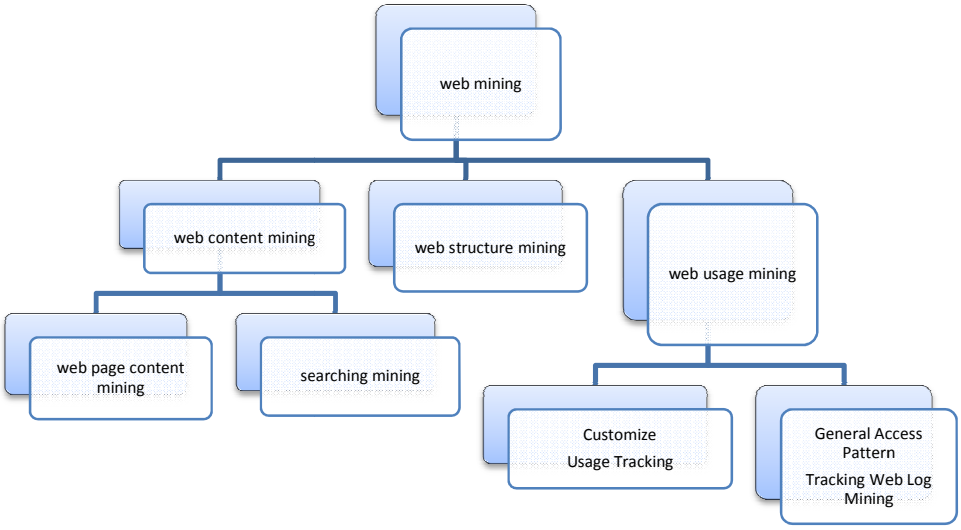


Figure 3: Taxonomy of Web mining,

2.8.1. Web Usage Mining: WUM

Web usage mining can also be defined as the application of data mining techniques to discover user web navigation patterns from web access Zalane et al, (1998), in addition to that, generalized definition accordance 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 to 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.

2.8.2. 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 him, 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.

2.8.3. 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.

2.9.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 web site. 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),this kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, like Most visited pages, average daily hits, etc.

In reference with Bamshad. et al ,(n.d), generalized as this kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, 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 ref 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, we can 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.

According to Bamshad et al ,(n.d)), generalized as 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 data base, 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.

In ref with Bamshad et al, (n.d)), generalized 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 ref (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 web site is needed .the second method computes clusters of page references based on how often they occur together.

In ref 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.

2.10. Related works

Data mining techniques are not easily applicable to Web data due to problems both related with the technology underlying the Web and the lack of standards in the design and implementation of Web pages. Web usage mining is a research field that focuses on the development of techniques and tools to study users' web navigation behavior.

2.10.1. Related Works on the Tools

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.

The other related works on tools on SpeedTracer, 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 web site 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.

Advanced mining algorithms uncover users' movement through a web site. The end result is collections of valuable browsing patterns that help Webmaster better understand user behavior. Further explained in the paper that generates three types of statistics: user-based, path-based and group-based. User-based statistics point reference counts by user and durations of access. Path-based statistics identify frequent traversal paths in web presentations. Group-based statistics provide information on groups of web site pages most frequently visited.

2.10.2. Navigation Pattern Discovery Tools

There are some web usage miner tools which can be used to the navigational pattern discovery for web user behavior of the web site, 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.

2.10.3. Related works in Advances Web Usage Mining

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 on the basis of user demographics and interests, as well as analysis of a site's success.

In Abhishek et al (2011), correlated but not linked web pages are discovered by clustering pages requested together by the site's visitors. This approach can be used to construct dynamic web pages automatically that provide links to pages considered relevant by earlier visitors Pierre, B., et al, (1996).

In the SurfAID project, a warehouse over web usage data is established and time series analysis is combined with association rules to discover unexpectedly evolving correlations among products (Abhishek, et al, 2011) propose the establishment of a warehouse, in which web usage data are combined with customer data, concept hierarchies on page contents and user demographics, as well as enterprise knowledge, e.g. in the form of previously discovered rules Myra,S., & Lukas C. (n.d). . Although user activities form the basis of these types of analysis, the issue of improving the site itself is not addressed.

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 web site 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, we present 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 web site, 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 web site association rules. The “Foot prints “ tool of records the footprints left behind by web site visitors and accumulates 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.

CHAPTER THREE: WEB USAGE MINING AND NAVEGATIONAL PATTERN

3. Introduction

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 web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is Access logs on server side, which keeps information about user navigation. Further explained in Sulu, G.,(2003), Web usage mining is the process of identifying representative trends and browsing patterns describing the activity in the web site, by analyzing the users' behavior. Web site administrators can then use this information to redesign or customize the web site according to the interests and behavior of its visitors, or improve the performance of their systems.

3.1. The General Process of Web Usage Mining

Today, understanding the interests of users is becoming a fundamental need for Web sites 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 can be exploited in the process of navigational behavior.

As it have mentioned earlier , Web Usage Mining (WUM) is the process of knowledge discovery and analysis of Knowledge from World Wide Web, represents a rather recent research field devoted to discover behavioral patterns from Web usage data.

As in Zalane et al (1998), the general processes of WUM distinguish three main steps: data preprocessing, pattern discovery and pattern analysis.

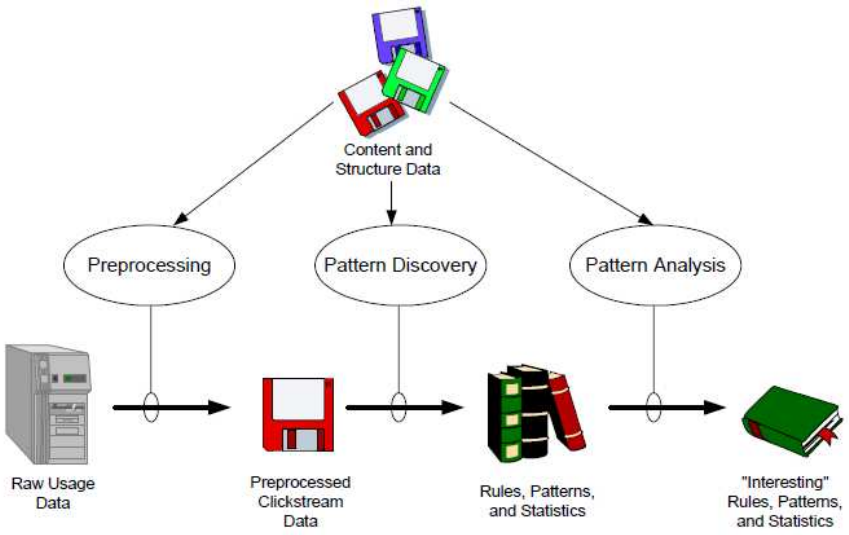


Figure 4: High Level Web Usage Mining Process Jaideep, et al (n.d), page 4

3.2.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.

Besides to the major sources of the data which have mentioned above but, there are also some other resources for web usage mining. According to Castellano, et al (2007) the following can be the source of the data.

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

A different researchers uses different collections over a time for web usage analysis in accordance with Berkan, y.,(2002), were collected for a period of two weeks for Logs Preprocessing and Sequential Pattern Extraction with Low Support.

3.3.Data pre-processing

In ref 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 ref 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.

3.4.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 ref 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 also their Performance is not enough for huge volume of data.

Researchers have shown that the log files contain critical and valuable information that must be taken out. It makes web usage mining a popular research area for many applications in the recent years.

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).

According to 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 Web site and outputs 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.

3.5. Data Cleaning

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

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

3.6. Removing Unnecessary Records

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; In ref 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. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

3.7. 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 s. 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. 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.

In ref 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 Web sites by following every hyperlink on every page within the site in order to update the index of search engine. Requests created by Web robots are not considered usage data and, consequently, have to be removed. To identify web robots' requests, the data cleaning module implements two different heuristics.

Firstly, 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.

3.8. 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 ref (Rajni, P., et al 2009), User's identification is, to identify who access web site and which pages are accessed.

But this task is not easy because few web sites that uses authentication to access the resource so the web records, only records the visitor's host and user agent. Further explained by 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 ref 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. A session is a series of web pages user browse in a single access, or according to Castellano et al, (2007), A 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.

3.9. 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 web site 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; or in e-commerce sites.

- 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 web sites. 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 web sites from the users' viewpoint. According to Federico et al (2000), Exploits,Web Usage mining techniques to suggest proper modifications to web site. Adaptive Web sites represents a further step. In this case, the content and the structure of the web site 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. in ref with (Sulu, G.,(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.

3.10. Navigational Pattern and Sequence

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 we instruct the miner to find only maximal frequent sequences, we obtain fewer and more compact results.

In the ref 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: When we investigate the causes of manufacturing errors, we only want the sequences containing error and cause, not the many events in between. The same is true when we search for operating system signals.

Comparison of GSP and AprioriAll

According to Murat et al (n.b)),On the synthetic datasets, GSP was between 30% to 5 times faster than AprioriAll, with the performance gap often increasing at low levels of minimum support. The results were similar on the three customer datasets, with GSP running 2 to 20 times faster than AprioriAll. There are two main reasons why GSP does better than AprioriAll.

- GSP counts fewer candidates than AprioriAll.
- AprioriAll has to first find which frequent item sets are present in each element of a data-sequence during the data transformation, and then find which candidate sequences are present in it. This is typically somewhat slower than directly finding the candidate sequences.

GSP, a new algorithm that discovers these generalized sequential patterns and has the following advantages for example.

- Empirical evaluation using synthetic and real-life data indicates that GSP is much faster than the Apriori.
- All algorithms presented in GSP scales linearly with the number of data sequences, and have very good scale up properties with respect to the average data-sequence size.

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 5: The mining Algorithms of WUM

3.11. Navigation Patterns and Important to Discover

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)), we 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. In WUM, Mary et al, (2000), an “interestingness descriptor" is a query in our mining language, MINT.

3.12. Knowledge Discovery Queries

Similarly to Lukas, (n, d), we believe 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.

3.13. Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process in accordance with, challenge of pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users.

The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL or MINT query. 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. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match

CHAPTER FOUR: METHODOLOGY

4. Overview of the methodology process

According to Dipa,(2010), web usage mining have three main process in order to discover a knowledge from the data ware house, author of paper use for his work according to this researcher, described above, it is necessary to perform three steps, see fig 5,but the detail of those how to accomplish those main process are described below.

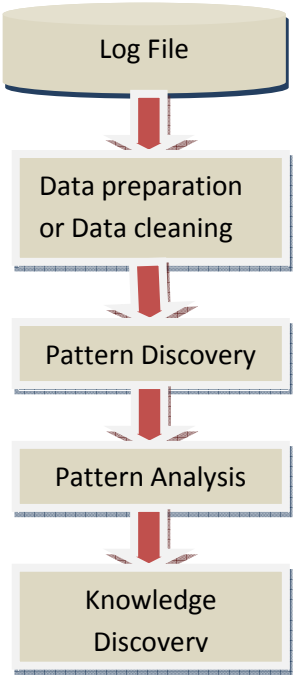


Figure 6: web mining usage main process to discover knowledge.

4.1. Tools Selections for Preprocessing

As stated earlier in chapter ,there a lot of tools uses for preparing a dataset for the intending purpose but the selection of those tools is not easy since every tool have designed for specific purpose but none of them cannot give a good output unless they combine each other in order to meet efficient output. The author 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.

The author 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 author found no need to implement his data preparation into navigational discovery software, besides to that even if the WUM have some capabilities of preprocessing, but does not support the main preprocess phases such as removing robot hosts and etc.

Extended log format of AAU

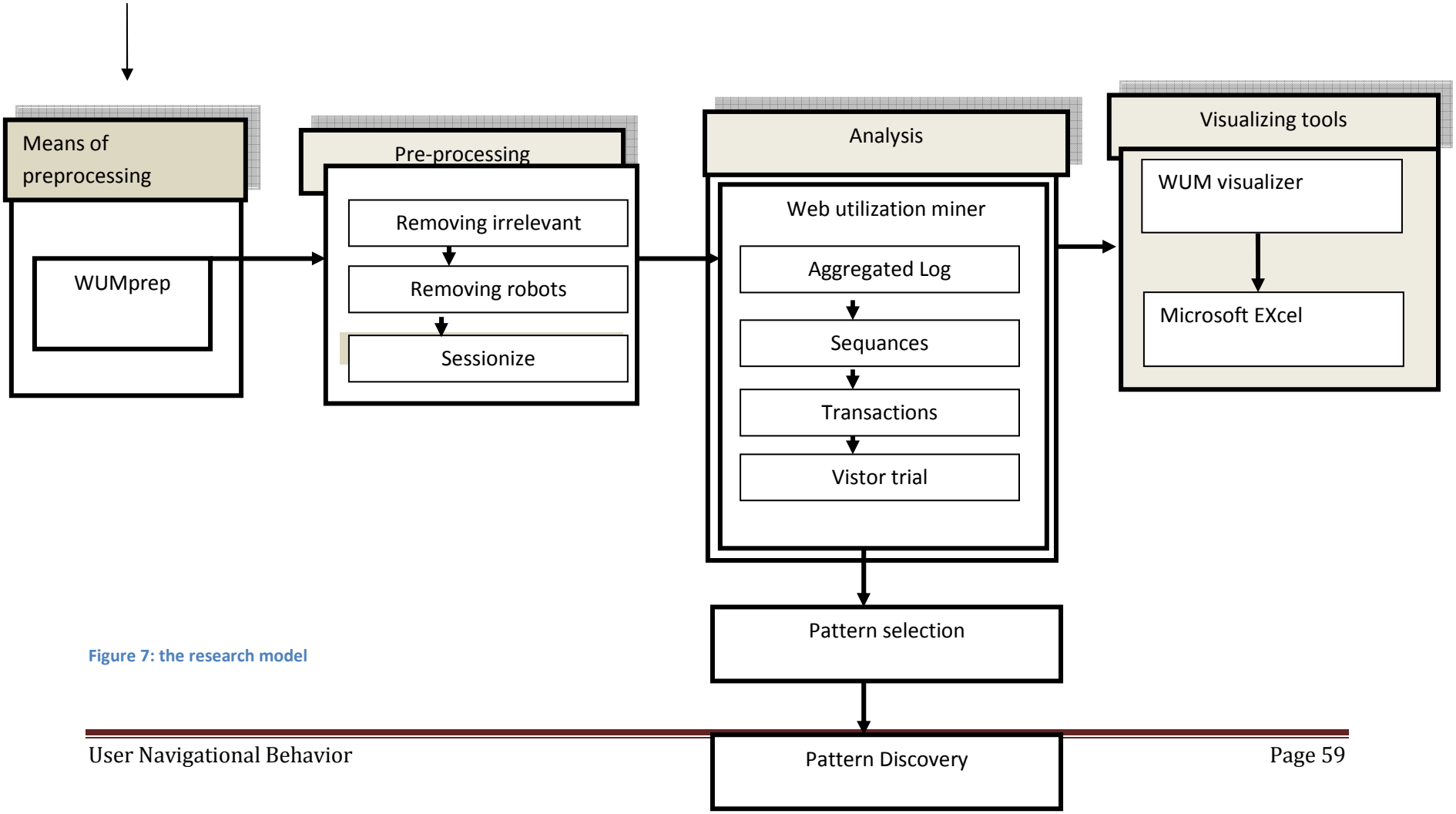


Figure 7: the research model

As it have been mentioned in the above figure 6, which shows how the objective could be achieve, even if the preprocessing done using the WUMprep, the researcher regulate the configuration of the tool to meet the objective . The data cleaning is done based on the following criteria.

4.2. Removing Irrelevant Records and Status

The removing of irrelevant records are significant as it have mentioned in the chapter three , as these requested log files are not only contain requests to the pages comprising the Web site, but also requests of images, scripts etc. embedded in these pages.

The author of this paper uses to remove those embedded extension of files should be removed because these “secondary" requests are not needed for the analysis and thus irrelevant (they must be removed from the logs before mining).Those requests are in the following table with their definitions:

\.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
\.JPG	The most common image compression format used by digital cameras.

Table 3: Irrelevant list of requests

Beside to that, the author only interested on request which only have the status 200 series, because concern only successful requests which mainly shows the users who get what they want and the other requests' are not need any more .

In general, according to the researcher these requests do not represent the effective browser activity of the user visiting the site; hence they are deemed redundant and should be removed.

4.3. Removing Robots

The author of this paper strongly believes to distinguish between human users and hosts that are robots, there exist several heuristics as it have mentioned above in chapter, section three. They are implemented in the script. Firstly, all records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed from the original log files.

4.3.1. Removing Duplicate requests

If a network connection is slow or a server's respond time is low, a visitor might issue several a successive clicks on the same link before the requested page is finally showed in his browser. Those duplicate requests are noise in the date and should be removed. The author of this paper uses, the most widely accepted threshold for of 2 seconds between two consecutive requests the entries that corresponds to robots can be eliminated.

4.3.2. Sessionize

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, the author uses a Session which is computed by taking any URL time stamp ,to achieve theses the researcher uses the most accepted time threshold which is 1800 sec or 30 min to identify the sessions using the these timestamp.

4.4. Divide log format

The preprocessed data needs to be dividing into manageable size before feed into WUM because it takes long time to process the data, so the researcher writes a python code to prepare the processed data for the WUM.

4.5. Tool Selection for Navigational Behavior

The transformation of the web server log into a log of sessions appropriate for mining and the process of navigation pattern discovery are performed in the framework of the Web Utilization Miner WUM, according to Anália et al., (2003), WUM (web utilization miner), Its primary purpose is to analyze the navigational behavior of users in a web site, furthermore, Navigation pattern discovery is 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 used as a basis to analyze the sessions in the rest of the log, comprising the sessions of the active investigators that did not become customers.
The architecture of Web Utilization Miner, There is two major modules: the Aggregation Service prepares the web log data for mining and the MINT-Processor does the mining.

In ref Bettina et al (1999), The Aggregation Service extracts information on the activities of the users visiting the web site 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. We call this the “support” of the node.

In accordance with Marya, et al (n.d), The MINT-Processor mines the aggregated data according to the directives of the human expert. “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.

In ref 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’s designed to work with The WUMprep module (which is responsible for the pre-process phase ;)
- Its free and open source tool (not commercial)
- 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.(is also open source and free)
- WUM uses for the discovery of navigation patterns and visualization of interesting Patterns.
- It’s a sequence miner and support GSP algorithms.
- 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.

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 web site 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.

4.6. General Methodology

The overall pictures of the methodology can be described as the following figure 7, the WUMprep scripts does the preparation steps(in the above illustration) , which is the input for the WUM tools discovers the navigation pattern and mining patterns and visualize the result using WUM visualize based on the miner interests.

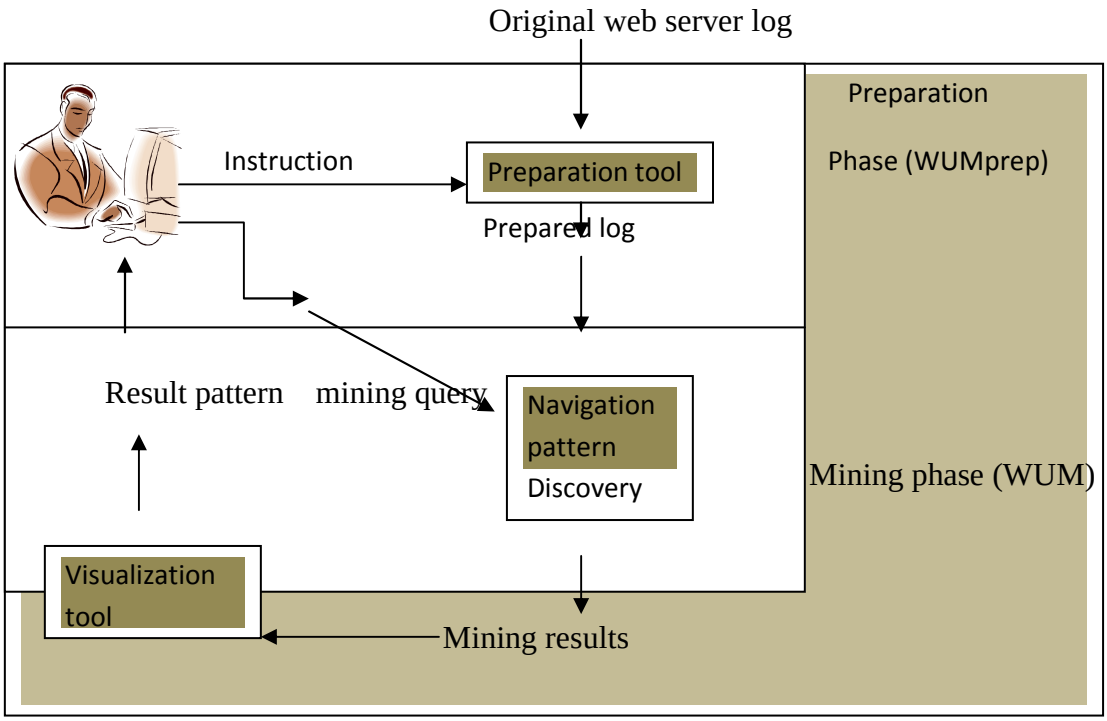


Figure 8: navigational process of WUM

CHAPTER FIVE: EXPERIMENT

5. Over view of Experiment setup

The experiment has been conducted on the following setup

- **Computer Type:** *personal computer (X32-based PC)*
- **Operating system:** *OS Name 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 the web log into manageable size

5.1. Data Collection and Selection

The data for this study is a web access log data of AAU official web site .As mentioned in the chapter one, a web log data is favored by many for web usage analysis. Two months web access logs have collocated for this study, for December and November.

5.2. Data Cleaning

The data collected from the AAU web server logs are full of junks that are not cleaned and should pass through some data cleaning phases (see the figure below) ,it is important steps to truck down the exact behavior of the user of the official web site unless they removed it is difficult to achieve the objective of this paper. Those phases must be undertaken to have cleaned data for further uses (process). The sample Log data are collected from AAU before preprocessing.

```
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)"
```

Table 4: A small extract of a Web server log contents

From the original web log see table 4, which can be easily seen a lot of junks, noises as well as robots (spiders, crawlers) those should be removed in order to have clean web logs to have appropriate ,efficient ,effective data logs.

5.2.1. Removing Irrelevant

As a result of removing irrelevant the number of log lines decrease in enormous seize the reason for it , those log files which contains the some extensions (see previous chapter), and those repeated requests that may came from inpatient users will be eliminated that's why the number of records seized in such amount. The original size of the records before were 50701 KB records (KB) and after the log filter preprocessing it became to 12416.KB.

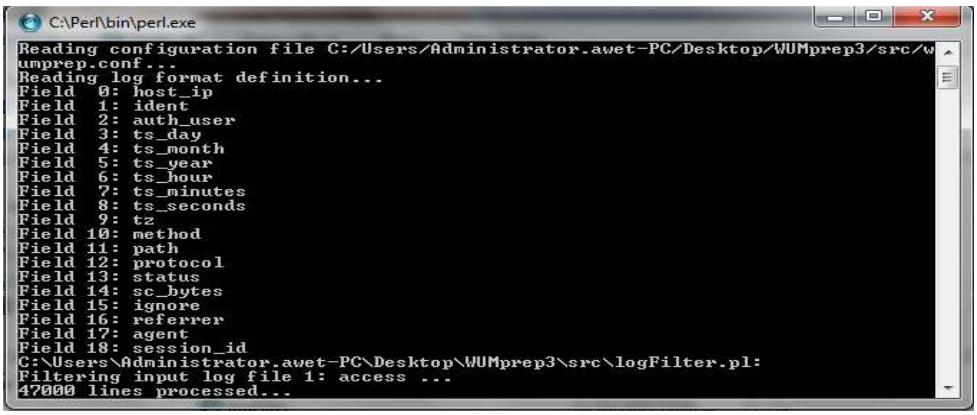


Figure 9: removing irrelevant records sample.

5.2.2. Detect Robots

The process of detect robots are very important to elimiatee the irelevent records whch are caused by the misusers that cames from other resources like (speider ,web crawlers) in other words , web surfing requested that are too fast that ordirnary pepole do not do in such fast ways caused by web crawlers.According to my expertment the number of robots are based on the maximum page view and against“index list“ in the WUMprep.The number of robots that detected from the web seerver logs are shown below,

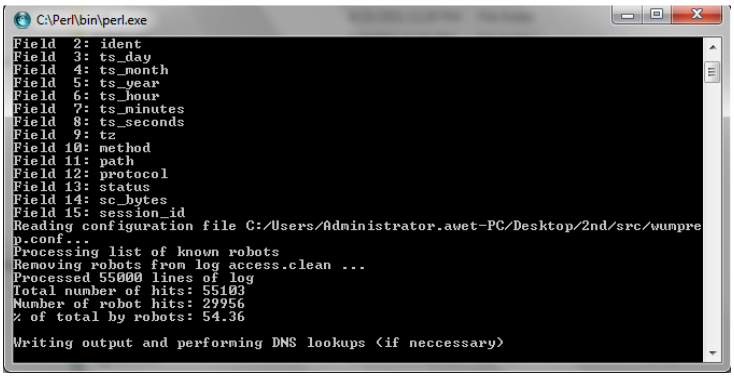


Figure 10: sample removing of robot hits

According to my experiment, for the months of December, the numbers of robots inside the Log format are 54.36 % robots from the total hits, for the months of November the total number robots are against the total hit are 39.68%. Samples of robot log lines that are 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] "GET /robots.txt
HTTP/1.1" 200 --304 "-" "Mozilla/5.0 (compatible; DotBot/1.1;
http://www.dotnetdotcom.org/,crawler@dotnetdotcom.org)"
(robots.txt)
```

Figure 12: Sample robot log files.

From the above results what it can be observable easily that some of the requests that came from same IP address that is (208.115.111.247) within two seconds, those

requests originated from the same IP address within two seconds. ([05/Dec/2010:05:03:20 +0300) and (05/Dec/2010:05:03:21 +0300)

5.2.3. Sessionize

The Sessionize which are resulted after the detection of the robots and give the following results as shown below,

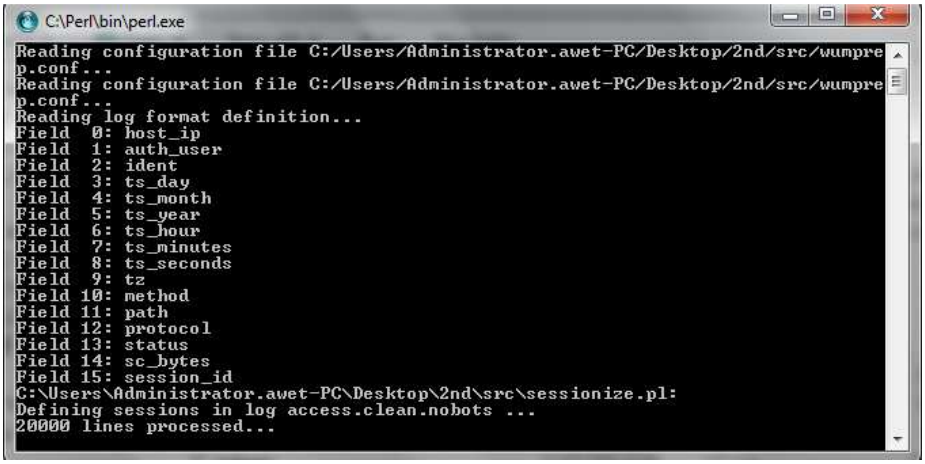


Figure 13: sample sessionaize process

The sesseinalize creates number of sesseionize, according to my experiment the number of session's creates are about 23411.some log lines which are exceed from the threshold i.e. 1800 sec or 30 min are removed .for the detail see in sample of in the appendix.

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

As it can be observable from the above fig 13,that the only status that filter from the web log files are GET and the status of 200 which indicates the successful requests from the web sites users, besides to that the session are identified . The types of log formats are converted from the Extended log format into Common log formats (see chapter Two, types of log formats).

5.3. Generalized Reports on Log Preprocessing

In this section the result of preprocessing will be discussed in general manner, an average user requests per day is 200220 lines. The preprocessing phases undertaken for both months (December and November) gives the following results after undergone through different phases of preprocessing for the months, and summarizes for one week in December the following tables. See for the months of November in appendix.

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
220340	150127	70564	25005	25005
230087	160743	72087	24060	24060
200406	148906	63480	21000	21000
190967	138967	50653	19734	19734
200190	178300	60752	20674	20674
200150	167543	47897	19653	19653
220205	120950	62096	23765	23765

Table 5: A Sample records for the week in December after undertaken the preprocess phases.

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

As it have been mentioned earlier the log files are contains irrelevant data, irrelevant records and noises, that's why we can observe from the above experiment result in the table the size of original log entry records decreased in average of 80%.for the months of November the size of records of original entry decreased in average of 73%.(see in the appendix).

5.4. Navigational Behavior of December

5.4.1. Aggregated LOG tree

The aggregated tree are results from the web miner after the sessions creates based on the above preprocessed tool (WUMprep) and imported to the miner resulted the aggregated tree for the months of December as follows.

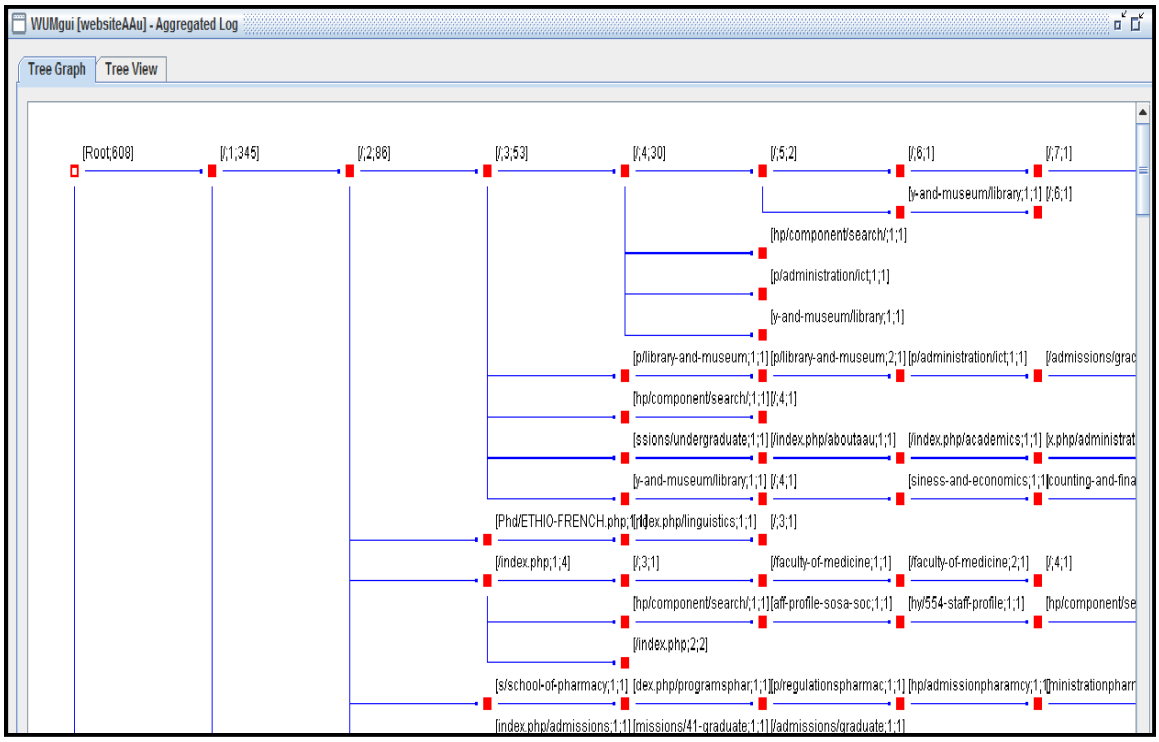


Figure 15: Sample aggregated tree for the month of December

As we can see from the above figure 14 ,an aggregated tree that the total number of nodes or total traverse make by users are 608 for the month of December, based on the aggregated tree the MINT query applied to find interesting pattern or for sequence analysis from it. The researcher chooses the some Examples to find interesting pattern for the month of December. For the month of November see in the appendix.

5.4.2. Sequence and Navigational Discovery of Users

As previously mentioned in chapter three, the generalized sequence pattern describes the behavior of users by filtering out the interesting pattern from the aggregated tree using the MINT query. ,In the following sections, the experiment is undergone using some most interesting patterns using the MINT query to discover the most important issues that should be discovered according to the researchers of interest , like Where do visitors of page Home go after wards?, Where do visitors go after typing the www.aau.edu.et (/)?, To Find out pages that always visit together and look at its pattern, Where do visitors go after search page of AAU (/index.php/component/search)? What is interested in navigation patterns between two pages.

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

Using the MINT (see appendix for syntax of MINT) query the author is interested where users go after the accessing the home pages until the next five pages, using the following query to the MINT to discover users' navigational behavior.

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 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
```

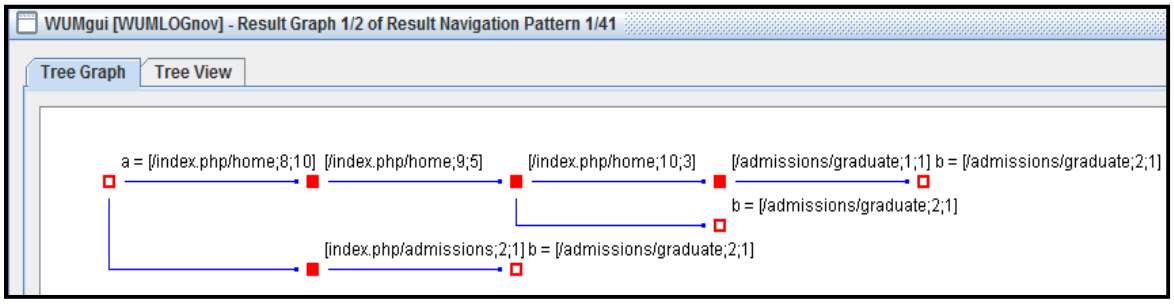
The above query results a following patterns using WUMvisulizer but the author puts some sample results in the following figure.

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

Here, we receive all pages reached within 5 pages after HOME (index.php/home), which has been accessed 100 or more times, provided that those pages have been accessed by at least 50% or 100% of the visitors visiting HOME, but as we can see from the result the most accessed pages is /index.php/library-and-museum users stay 100% visting the content of it, It is also clear that most users who visited the home page also stay in 100% within the page of /index.php/registrar those are the most .of course the other pages like /index.php/admissions/graduate users stay in those pages users stay in the page for average 26%,even if they are the most visited pages after Home pages.

Navigation pattern:



As we can see from the navigation pattern most people are going to the page of /admissions/graduate after visiting the home pages, it's clear to see that most users stay in

the HOME page (/index.php/home) and navigate 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 above MINT 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 receive 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%.

Navigation pattern

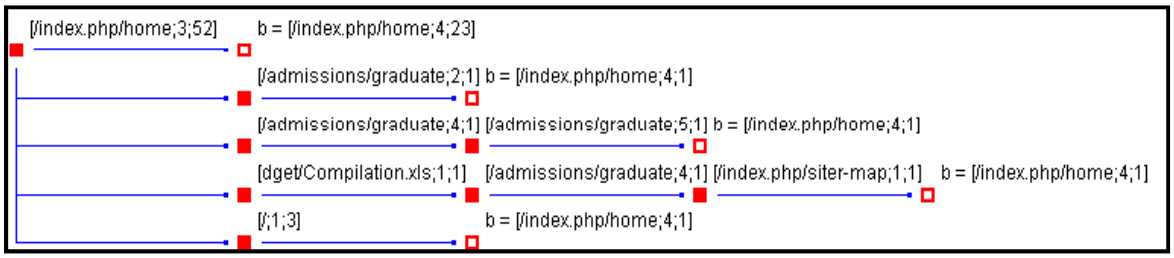


Figure 16 :Navigation pattern

From the figure 15, its easily observable here, we see that when visitor start from looking at **/index.php/home** page, 20% of them will stay within this subject area.

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

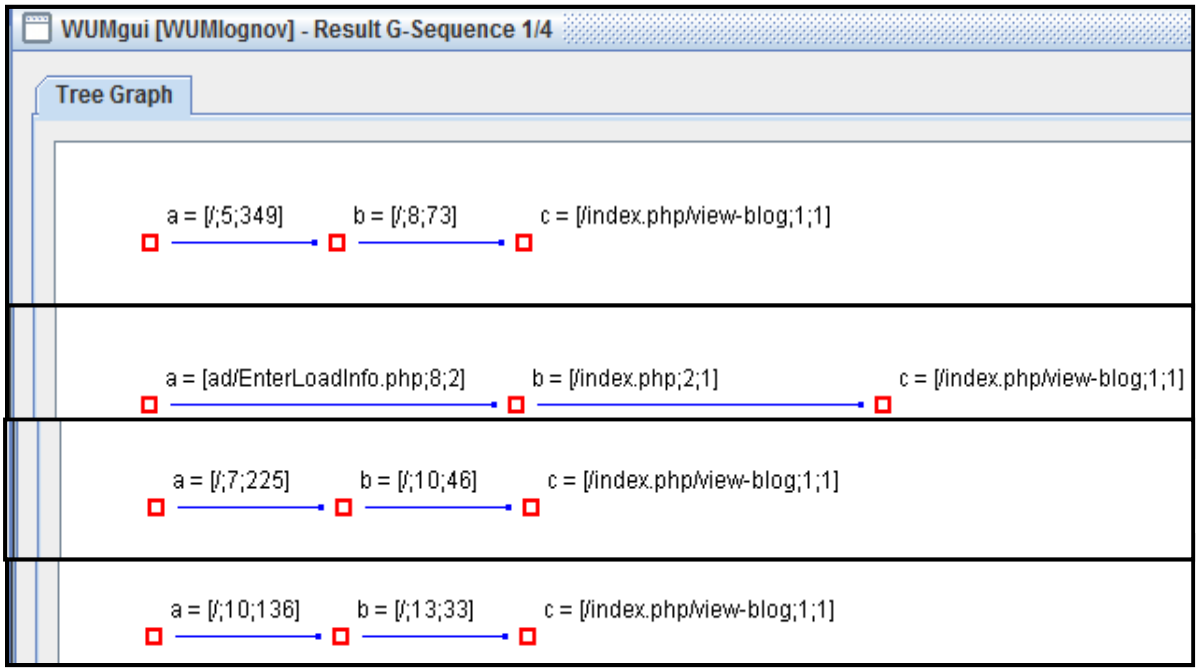
In this query, we specify a template t with two variables a, b, thus seeking for with two pages bound to a and b and at most 5 arbitrary page occurrences in between denotes that "a" should be bound to the first page which is /index.php/home, this page should be visited at least 20 % and b page should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

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 out of the query give us two patterns ,Here we recive 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 stay in the home page before reaching to /index.php/view-blog pages.

G-sequance



the Users do not take a single paths to reach to /index.php/view-blog most of the users take a path from the root pages,and the second most users take 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
```

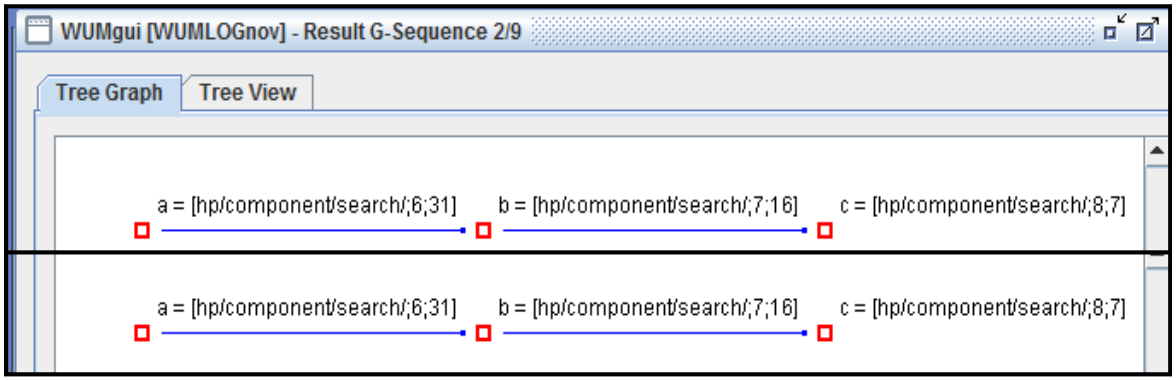
Pattern

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 ten patterns show that user's do know where they are looking for. most of users who stays in search engine 100% and also stay in this page for average of 40% ,they do search function stay within search the page.

G-sequence pattern



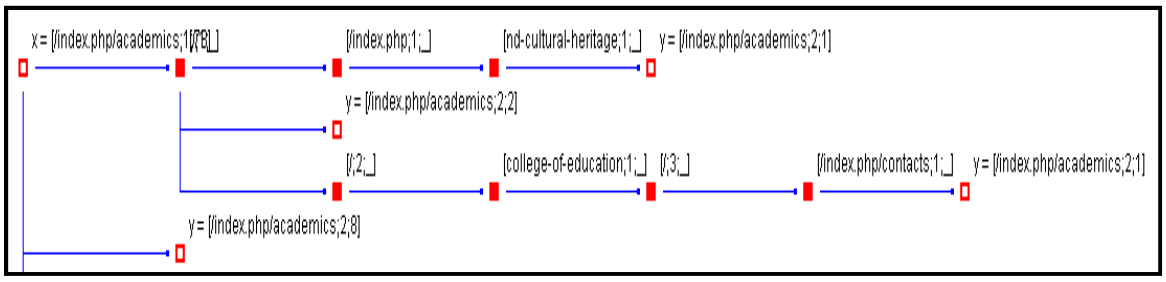
Sample of the above ,the author do not need to put all the g-sequence from the navigation pattern that users stay in the search page as we can see from the above result, that users stay in the search page.

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"
```

The above query results the following 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 web site.

5.5.Statistical Analysis for the Months of December

The WUM can generate a comprehensive report in terms of simple tables the researcher used other tool (Microsoft Excel) for better visualization. report will be discussed like what are ,most requested pages, most visited pages, most visited directory as well as most referee pages for the month of December will be discussed .For the month of November see in appendix.

5.5.1. Most requested pages

The following table shows the top ten most accessed pages during the months of November .For the rest of the months see in appendix.

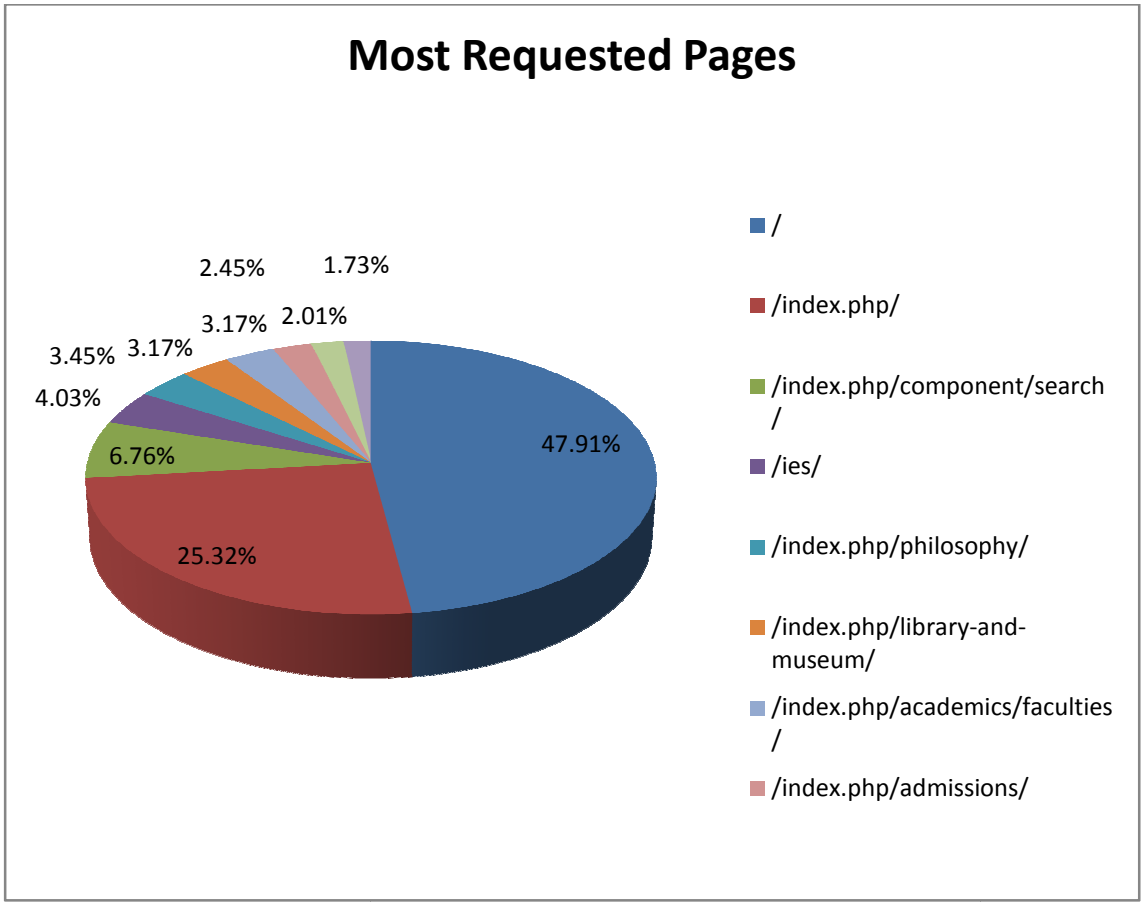


Figure 17: Top 10 most requested pages.

A figure shows the Top ten most pages during the months of December ,As it is shown the most requested pages is the / 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 is most popular by most users in all the three months .in fact ,this shows that most visitors enter into the site directly by typing the web site address as it shown in the above directory. The search engine of the Addis Ababa University the second most accessed pages followed by the /index.php/library-and-museum pages.

5.5.2. Most visited directories

The root directory "/" is the most accessed directory where the root directory in root folder is located .Most users also shows interest on the contents under the **/index.php/** folder. It is also possible to say that from the output those are also important visited pages /index.php/component/search/.

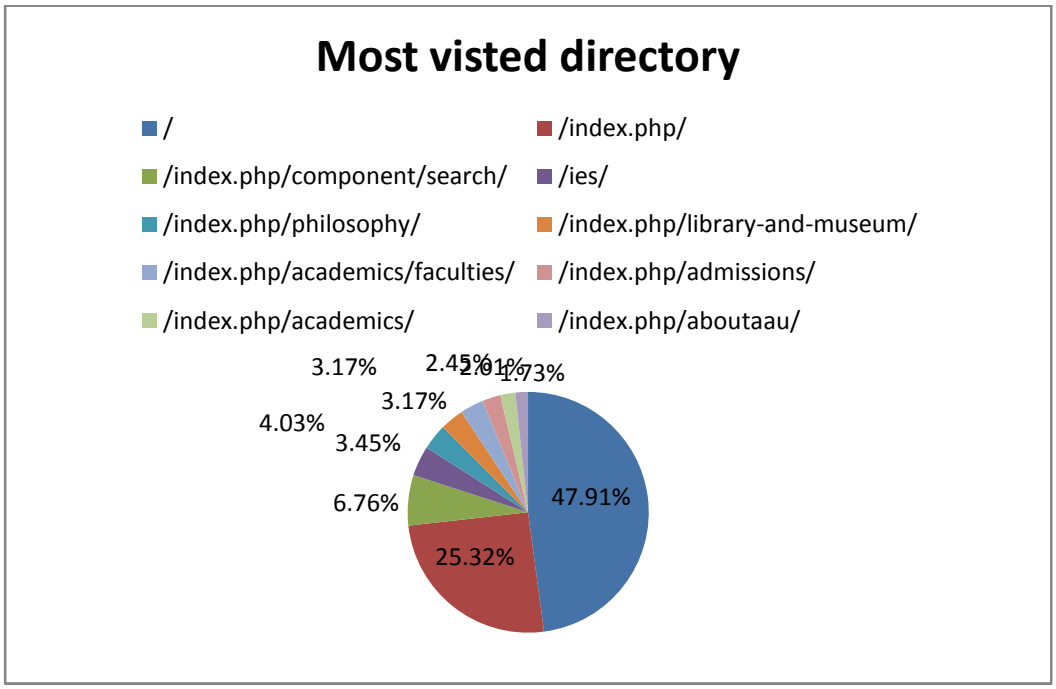


Figure 18: Top ten requested directories

For the rest of the months, most of the directory are requested ,are the same as the above until the top three directory but the others are became familiar in the next pages.

5.5.3. Most Top Entry Pages and Top Exit Pages

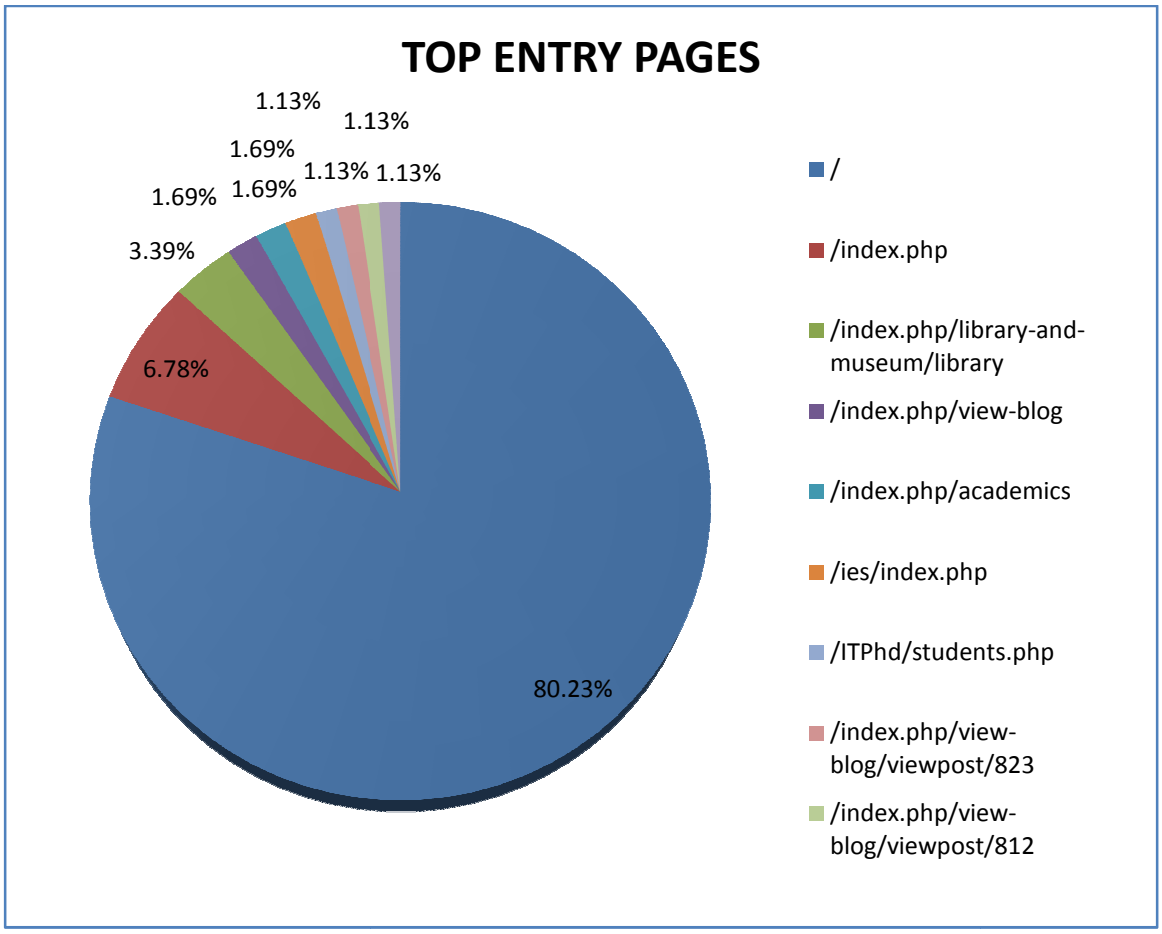


Figure 19:Top ten entry pages

The entry pages are pages that indicated that the web site users first visited where as the top exit pages is the last pages the users visited official web site .From the figure below what we can observe is that the "/" root pages where it is located accessed more than any other pages almost half of the request (80. 23%) and the /index.php the second most top entry page and last not least the /index.php/library-and-museum/library the third most top entry of pages, followed by /index.php/view-blog and /index.php/academics the 4th and 5th top entry pages. For the rest of the months see appendix of December.

For the month of November, as it as shown in the figure 55.68ve % of the visitors have entered into the web site directly through the / and index.php.

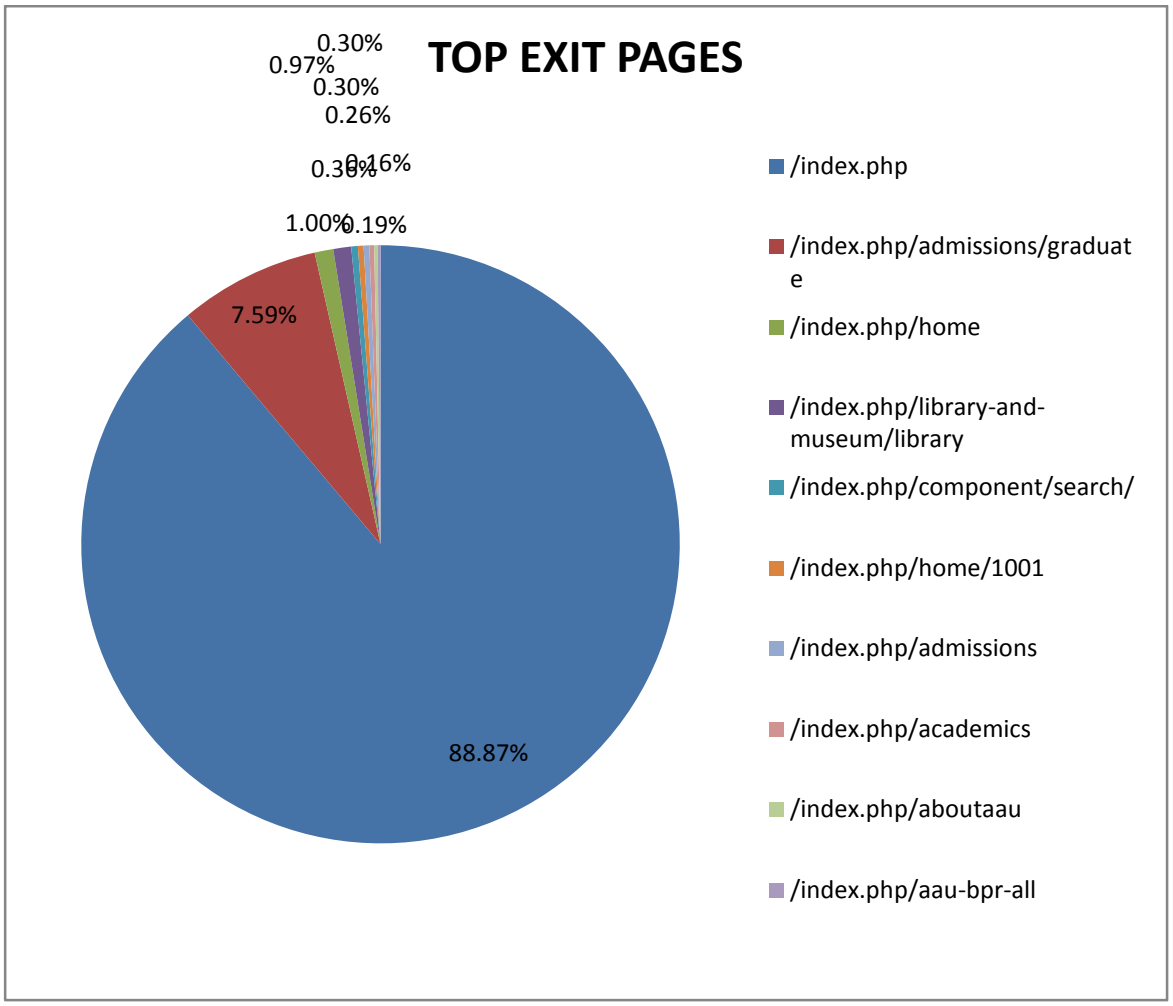


Figure 20: Top most exit pages.

From the above the figure ,we can see that the top exit pages are the "/" or after the user type the web site address and leave the web site 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/.

5.5.4. Top Referrer Pages

The top referee pages are pages where the visitor was located when making the next request with the official web sites.

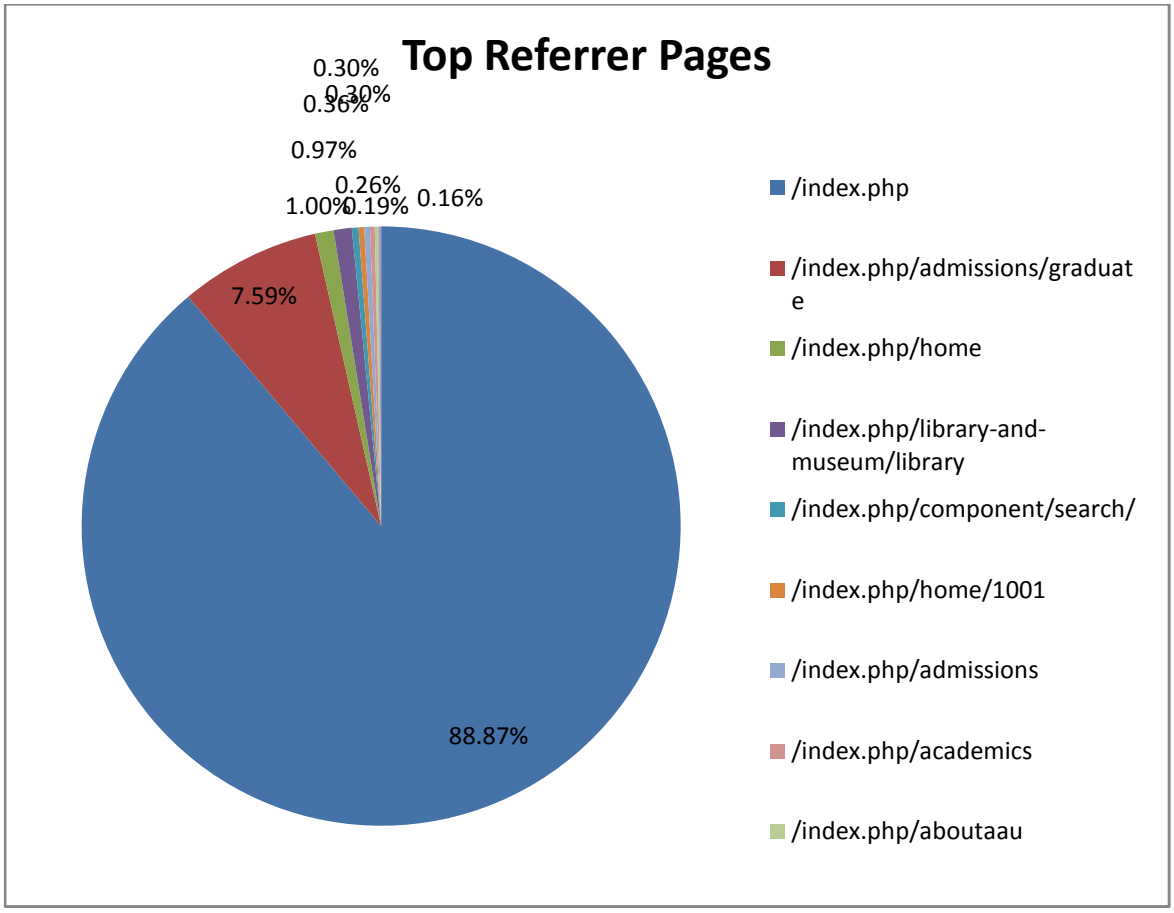


Figure 21: Top Ten referee pages.

From the above figure , what it can be easily observable that most users makes the next request from the page of /index.php which covers in more percentage 88.8 %. the next popular page where users requests the next page are initiated from /index.php/admission/graduate and which covers 7.33%.the third most refree pages are /index.php/home which covers the percentage of 0.96%.

For the months of November are almost the same as the above but the only difference are below three requests for more details see in appendix.

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATION

Conclusion

- From the navigational behavior of users that we can indicate easily users is no single point where users go after home page and can be conclude that users navigate from top of the page (hierarchy) to the lower hierarchy.
- From the navigational behavior search behavior can be conclude that most users use the search engine effectively or know what they are looking for.
- most request pages are requested to the web site by typing the official name of the web site that is www.aau.edu.et,that's why the most request web page becomes the root page of course it clear that the web server is an apache server, when type the official name hit the root directory of the official web site .from the request pages the second top most requested pages are /index.php/component/search pages, it indicates that most users use this page for searching key words with in the pages. What else can be concluding that /index.php/library-and-museum the third request pages, can be conclude that most users are interesting in the content, the reasons could be most journals associated to it.
- Most visited directories are of course the root directory since most of users are typing the name of the official web site 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 in side it.
- most users use enter to the web page using the page of /index.php , /index.php/library-and-museum/library, /index.php/view-blog and most users also leave from those pages that it can be conclude that almost the other pages 1/3 ,most users leave without visiting the web pages.
- It easy to see that most users use the /index.php, /index.php/admissions/graduate, /index.php/home are most users requests from those pages to make for further

requests, so it would be very useful if the administrator can put some urgent notice and advertisement within those sites since they are most accessed web sites.

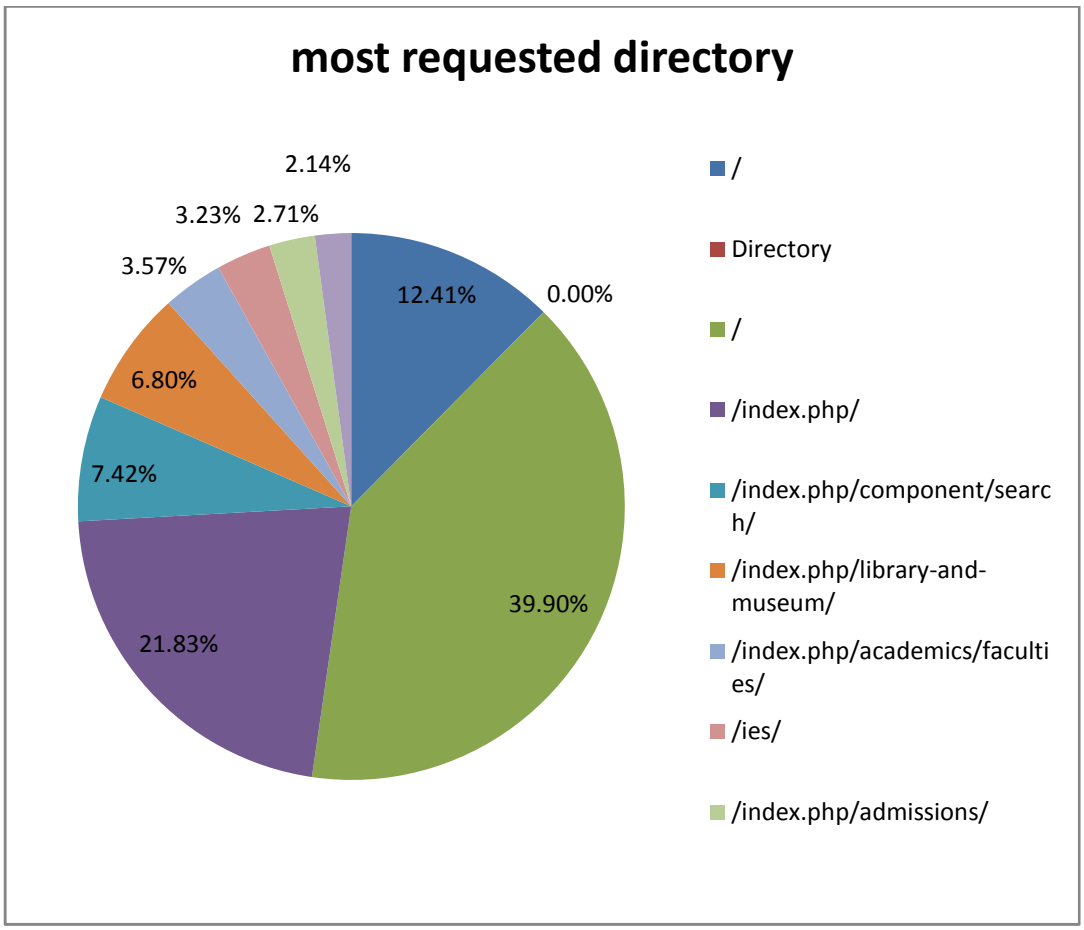
Recommendation

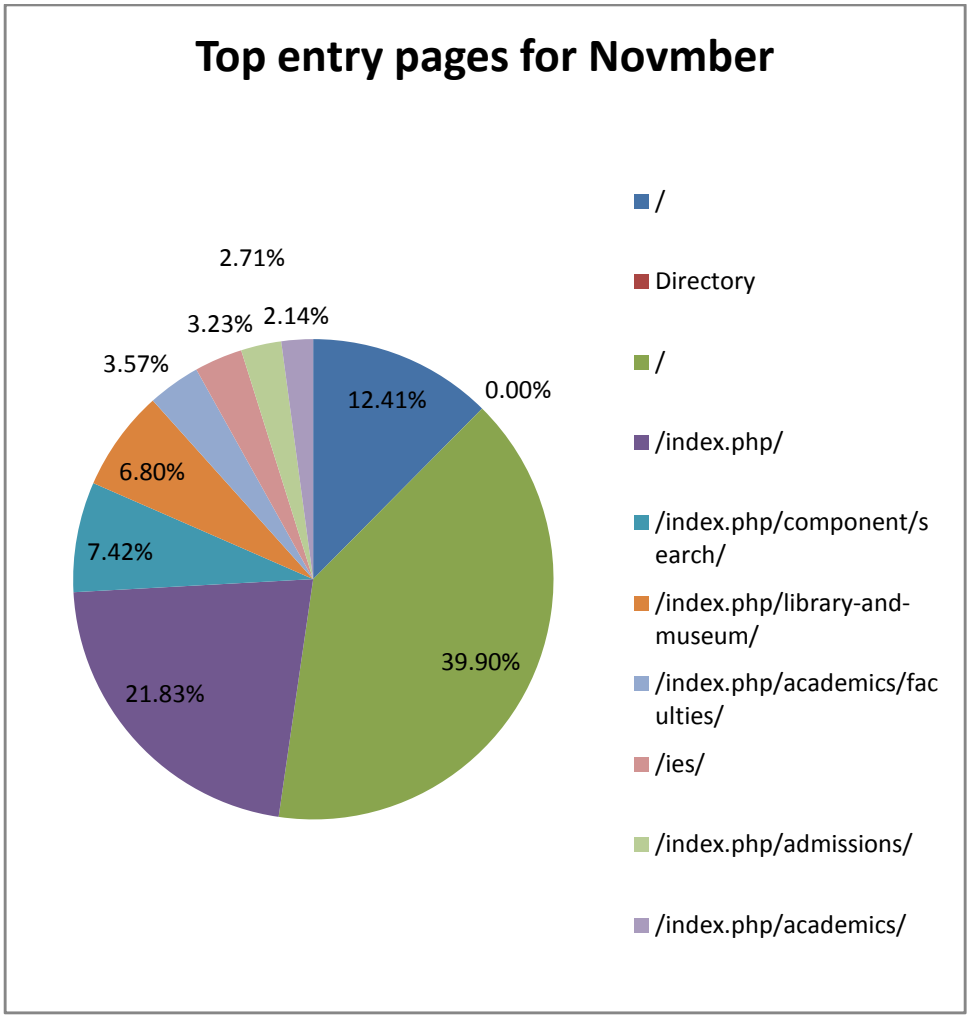
- Most users come to the page web site directly by either typing the name or from the search engine that displays the home page .This could be an indicator the web site has a kind of sickness .the web master therefore should do some kind of assessment on the department index pages make sure that those pages contain those key word for indexing in search pages.
- The most together accessed pages are the home pages is accessed with itself so it is important that ,It also important to recommend that the concerned body that is in charge of AAU official web site design should create quick links from one to other pages for those pages mostly accessed to gather.
- It is also clear that most users left the web site from some pages mainly from the /index.php/home ,/index.php/admission/graduate,/index.php/academics from it, it possible to recommend that the web master should use those page for advertisement and notice and also possible to recommended further it is possible to link to other department links in order to encourage web site users to stay in the web site.
- It is possible to recommend that the web administrators should make the most accessed pages, to be prefetching or cached to prevent the latency of the network bandwidth or prevent delay to access those pages.
- From the navigational behavior 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.
- 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” as a consequence could be better result.
- The other recommendation for further work, divide the web page based up on concept of hierarchy which concept divide pages according to the service they provide, once hierarchal classified the pages it would give better result.

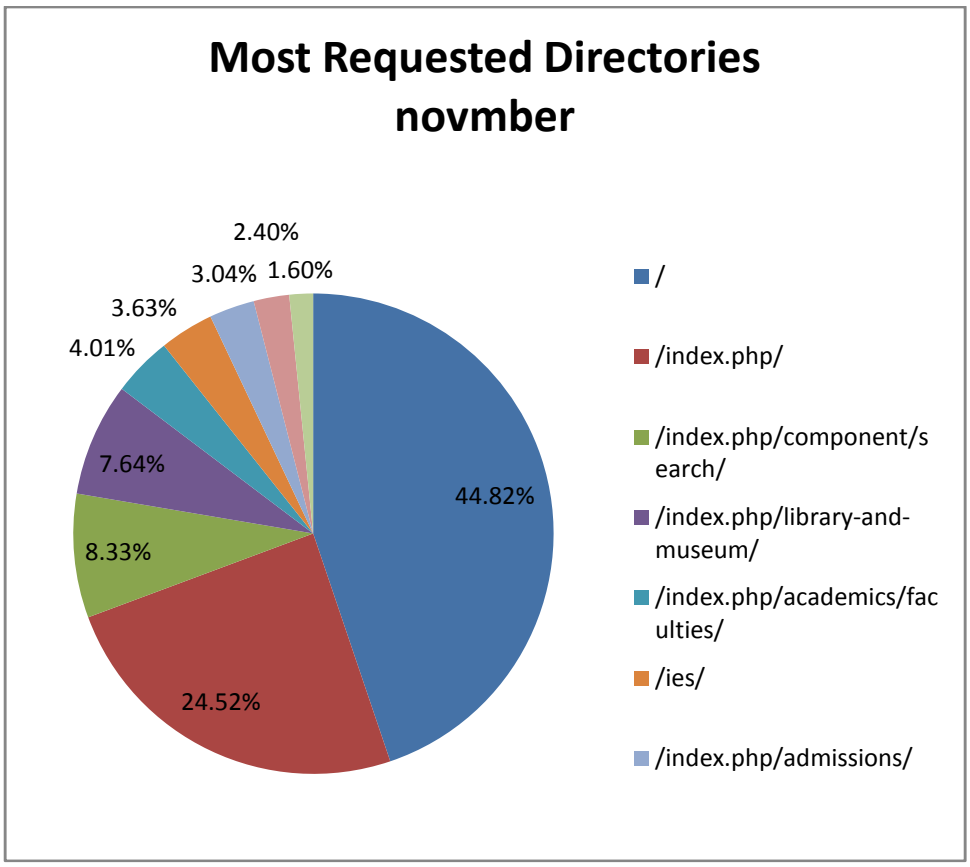
- The last not the least, recommendation for the further work, since by combining different technique of web usage mining such as content mining with web usage mining (work of this thesis) it could give better result in terms of efficiency .

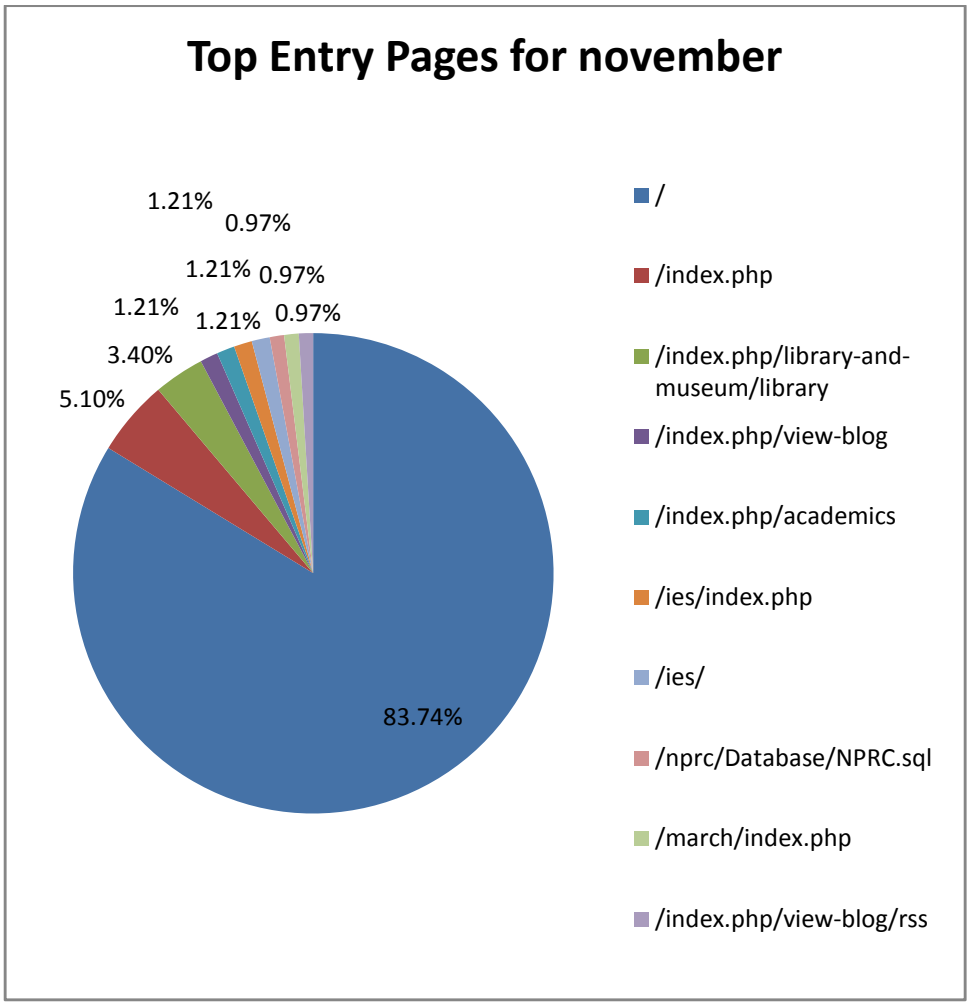
Appendix A: statistical report for the months of November

Appendix for month of November: Most Requested Directories for the months of November




Top entry pages for Novmber
12.41%
0.00%
39.90%
21.83%
7.42%
6.80%
3.57%
3.23%
2.14%
2.71%
/
Directory
/
/index.php/
/index.php/component/search/
/index.php/library-and-museum/
/index.php/academics/faculties/
/ies/
/index.php/admissions/
/index.php/academics/


Most Requested Directories novmber
2.40%
3.04%
1.60%
3.63%
4.01%
7.64%
44.82%
8.33%
24.52%
/
/index.php/
/index.php/component/search/
/index.php/library-and-museum/
/index.php/academics/faculties/
/ies/
/index.php/admissions/



The following are also the sample of one week for the month of November the results of those as explained in chapter 5.

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
210240	140127	69564	20004	20004
240067	160743	72087	24060	24060
203406	148906	63480	21000	21000
200967	138967	50653	19734	19734
200190	178300	60752	20674	20674
200150	167543	47897	19653	19653
220205	120950	62096	23765	23765

Appendix B: 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 C: 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 web sites 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 Web Sites 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 Web Sites. 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 web sites. 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 Web Site 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.

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

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.

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS USING GENERALIZED SEQUENCE PATTERN
A CASE ON OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY

BY

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

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

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS USING GENERALIZED SEQUENCE PATTERN
A CASE ON OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY

BY

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

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 and foremost, I would like to thank my advisor, Ato Workshet lemnaw, 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.

Of course, my thanks to Professor Bettina Berendt for her borderless support in giving directions on this work ,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.

" fekri Belibi"

Abstract

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years. Academic researchers have developed an extensive array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular web site. Performing this kind of investigation on AAU web site can provide information that can be used to better accommodate the user's needs.

The Web Use Mining (WUM) , it corresponds to the process of knowledge discovery from databases (KDD) applied to the Web usage data. It comprises three main stages: the preprocessing of raw data, the discovery of schemas and the analysis (or interpretation) of results. A WUM process extracts behavioral patterns from the Web usage.

In this thesis, we find out the navigational behavior of the user of official web site of Addis Ababa University web server recorded in web server for two months (November and December), those recorded are raw data that are full of junks, noises and irrelevant data contents .In this paper present a preprocessing tool WUMprep that uses to filter those unnecessary data, such as irrelevant records, noise data, and it crates the sessions based on specific thresholds.

For discovery of navigational behavior, here presents the Web Utilization Miner WUM, a mining system for the discovery of interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the researcher using WUM's mining language MINT, using those descriptor it can be describe the general behavior of users instead of single users behavior using the most appropriate algorithms known (Generalized sequence pattern) which implemented in WUM.

The General behavior of users constructed by GSP algorithms those behaviors are descried using the MINT query. Those MINT query are intermediate between the users and pages.

The researcher of this paper also recommend that to get a better result by combining the web usage mining with content mining techniques of web usage. Of course without any doubt it could give a better result in terms of efficiency and effectiveness results.

Table of Content

List of Table

List of Figures

Web Terminology and Definition

In accordance with 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.

Abbreviation

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
IBM	International Business machine
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

CHAPTER ONE: INTRODUCTION

1.1.Background

In 1990 the internet was initially designed for exchange mails between users later it becomes trendy for use of WWW. The www or 3w in now popular services among almost any other services the internet provides. There are number of services providers (ISP) for the use of the internet across the world. In Africa, the number of the internet users increasing and increasing from time to time. 5.6% of the world internet users are from Africa, further explained, it shows 2,357.3 %growth from the year 2000-2010 similarly, Ethiopia has 0.4 % share among African internet users .Even if this seems insignificant when it compared with the rest of the world, generally speaking the number of the internet across the world getting increasing and increasing in dramatic way thorough out worldwide1.one of the various reasons for the development of the internet in Ethiopia causes by huge amount of investment in infrastructure like in education ,telecommunication and development in others sectors.

Addis Ababa University, one of the oldest higher education institutes 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.

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

[1] http://www.internetworldstats.com/stats1.htm#africa

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]

There are number of services provided by the Addis Ababa University, one of the popular services are the WWW(world wide web) among other services like teleconference ,data service ,those web services are divide in two as the official web site (internet) and intranet which is not able to be accessed outside the university which uses for local uses. The official web site accessed through the public Ip address offered by the ETC.

The official web services of AAU an organized collection of Web pages information is presented in various formats , ranging from research papers, and educational content, to multimedia content, blogs .that's why the getting information from the official web site is the matter of click-streams in the internet of course if there is connectivity. 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 a few individuals who realized that the AAU would be wise to join the information age by adopting the technology that has been transforming the world. The newly formed office initiated a project named AAUNet that has resulted in a wide area network (WAN) whose first phase of construction was 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 is adequately served by ICT. At the same time, AAU's need for and dependence on effective ICT support is now greater than ever.

[2] www.aau.edu.et/administration/DRAFT ICT POLICY AT AAU
[3] www.aau.edu.et/administration/ICT

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.

1.3. The AAU Official web site

The Addis Ababa university official web site was published around some seven years ago .As the ICT development office of AAU (which have mentioned in previous section) is engaging in ICT related works ,the official web page develop and maintain by this office. the web site is hosted on AAU's own server which is located in main campus of the university (6 kilo), The official web site have the domain of www.aau.edu.et and have statistical IP address.

1.4. Purpose and User Community

The official web site being in work to deliver information both the university activity, in general and about academic and administrative units, in particular, it also delivery information about news, items and its own advertisement for both vacancies and student admission and other, of course it has also some external links to other web and other sites such as collaborative organizations in research activity donor agencies, etc.

1.5. Nature and Content

Generally the web sites designed bear in mind to support the objective of the university. In sections try to discover the nature and content of the web site. The AAU web site has both static and dynamic nature .there are few web sites that are static in nature those pages are not interactively with its users but the majority of the web pages are dynamic in nature which are support the MYSQL database incorporate with JOOMLA packages helps users to interact with web sites users.

When we came to Web site content it posts numerous information regarding to the objective of university which presenting information on several topics and issues, each page have information regarding to the objective of the pages .there are few page which are under construction(content not yet update), but there are advertisement and notice on several pages.

1.6.AAU Web Structure

In the following hieratical graphs displays web site structures of the official web site.

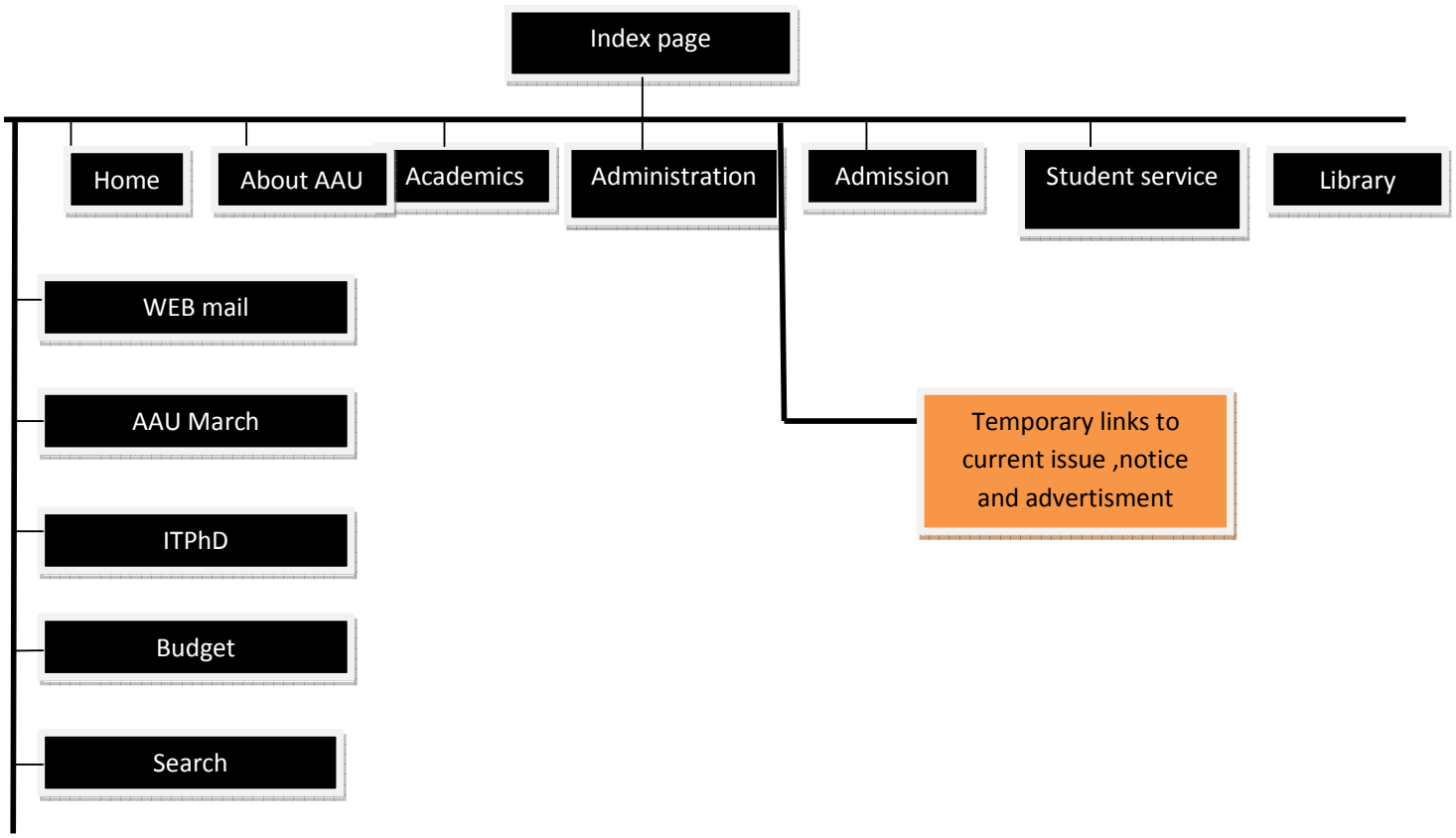


Figure 1 the structure of the official web site of AAU.

There are some other web sites that are accessed to gather with the official web sites like AAU march, ITPhD, college of education, IES (Institute Ethiopian Study), virtual accessed using the main web site.

1.7.Statement of the Problem

Rise of the Internet gave many companies an access to the 'gold' channel. Trading, putting gigabytes of information and communicating online has become one of the sources for understanding of the web users. 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.

As statics shows the number of web sites published every day is increasing quickly still, there are now 184 million registered domain names worldwide, a 9% increase over the same period last year[4].

On the other hand, the education sector is rapidly evolving and the need for web information Places that anticipate the needs of their information seekers are more than ever evident. The need of placement information is not easily imaginable we have to explore where should be places some information in a given web site, in this case of the official web site of AAU. It is important to know the navigational behavior of the users based on the study of the behavior. the need of study of any behaviors scaled up from the taxonomy of animals ,plants and others , in general, further explained that animals classified in to mammals ,vertebrates based up on the whole group behaviors.

According to Mokenen (2001) who were working on web usage mining of the official web site of AAU using the tools of wumprep4weka, for preprocessing or cleaning the data and Weka tool for data mining of the interesting pattern using the aprior algorithms finds out the most frequent access that do not based up the sequence, based on his study he did not truck the general behavior of users.

Like it discussed earlier uses the sequence (generalized sequence pattern) can tell the general behavior of users on navigational behavior of the user of official web site of Addis Ababa University, and not work have been done yet on the topic as to the knowledge of the author.

Web site design is currently based on thorough investigations about the interests of web site visitors and on less investigated assumptions about their exact behavior. In Lukas, C., (n, d) Concrete knowledge on the way visitors navigate in a web site could

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

prevent disorientation and help owners in placing important information exactly where the visitors look for it.

1.8. Scope and Limitation of the Research

Web mining has different branches: web content mining, web structure and web usage mining .the focus of this research is on mining usage pattern of AAU official web site .usually, three types of web related log files, namely web access log, error log and proxy log files. however, in this research work, web access log records is used as dataset because many literature and previous research justify that web access log files is the typical source of navigational behavior.

The limitation in this paper is the lack of manual on how to operate the web mining tools (WUM) and besides to that the web access log stored in Addis Ababa university are erased at the end of every months that's why it is difficult to get a enough data for the research, besides to that the web mining tools need to have a higher capacity (memory) to process the whole log files as batch.

1.9. Justification of the Research

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.

The importance of the study web users further explained by Marya, et al, according to him ,most web sites 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 web site they present to their visitors.

One important data source for the study is the web-log data that traces the user's 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, the importance of www further explained in Kosala et al, (2000), the web log creates an opportunity and encouragement for all Data mining researchers, consider it as the largest data warehouse in the world.

In accordance with 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's why performing data mining to optimize the performance of a Web server. In ref of Lukas, C., (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.

In accordance with Narendra, et al., (2003), Web mining is defined "*as the use of data mining techniques to automatically discover and extract information from web document and services. "*
Furthermore, there is also a widely accepted definition, According to Zalane, et al ,(1998).

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 web site is being used as expected*".

From the above the definitions web mining attempt to get the information (knowledge) or to extract the pattern, for the purposes to have an intended knowledge, so some the techniques should be applied to different web resources to overcome the problems, in ref with Mobasher et al, (1996), web mining is a common term for three knowledge discovery domains that are concerned with mining different parts of the web: web structure mining, web content mining, and web usage mining.

In general, User behavior has two aspects, one concerning the interests of the users and the information they access, the other concerning the way of accessing this information. The first aspect is addressed by techniques for the establishment of user profiles and is not peculiar to web usage. For instance, student profiles are considered in intelligent tutoring systems, the second aspect is addressed by techniques analyzing web server logs.

For example, consider a user that explores the links in a web site to find every bit of information of potential interest and a user that prefers keyword search. Those two users need fundamentally different support, even if both of them are interested in solar energy collectors, chess and medieval sculpture, In this study, concentrate on the second aspect of user support, namely on the analysis of user navigational behavior, because web users is characterized by her/his interests and by her/his navigational behavior.

1.10. Objectives

1.10.1. General objective

The general objective of the research is to apply web mining techniques for discovering of navigational behavior of AAU official web site usage of to reveal previously unknown the interesting, and actionable patterns based on the web access log file in order to recommend possible measures for further r improvement of the official web site of AAU.

1.10.2. Specific objectives

To achieve the general objective of the research, there are specific objective should be addressed, the specific objectives of the research are:

- To review literature review in the area in order to put concrete background and justification for the research.
- To identify and collect the data
- To prepare those data set using different preprocessing techniques.
- To analyze the navigational behavior of the users.
- To analyze the sequence of the web site 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.

1.11. Research Methods

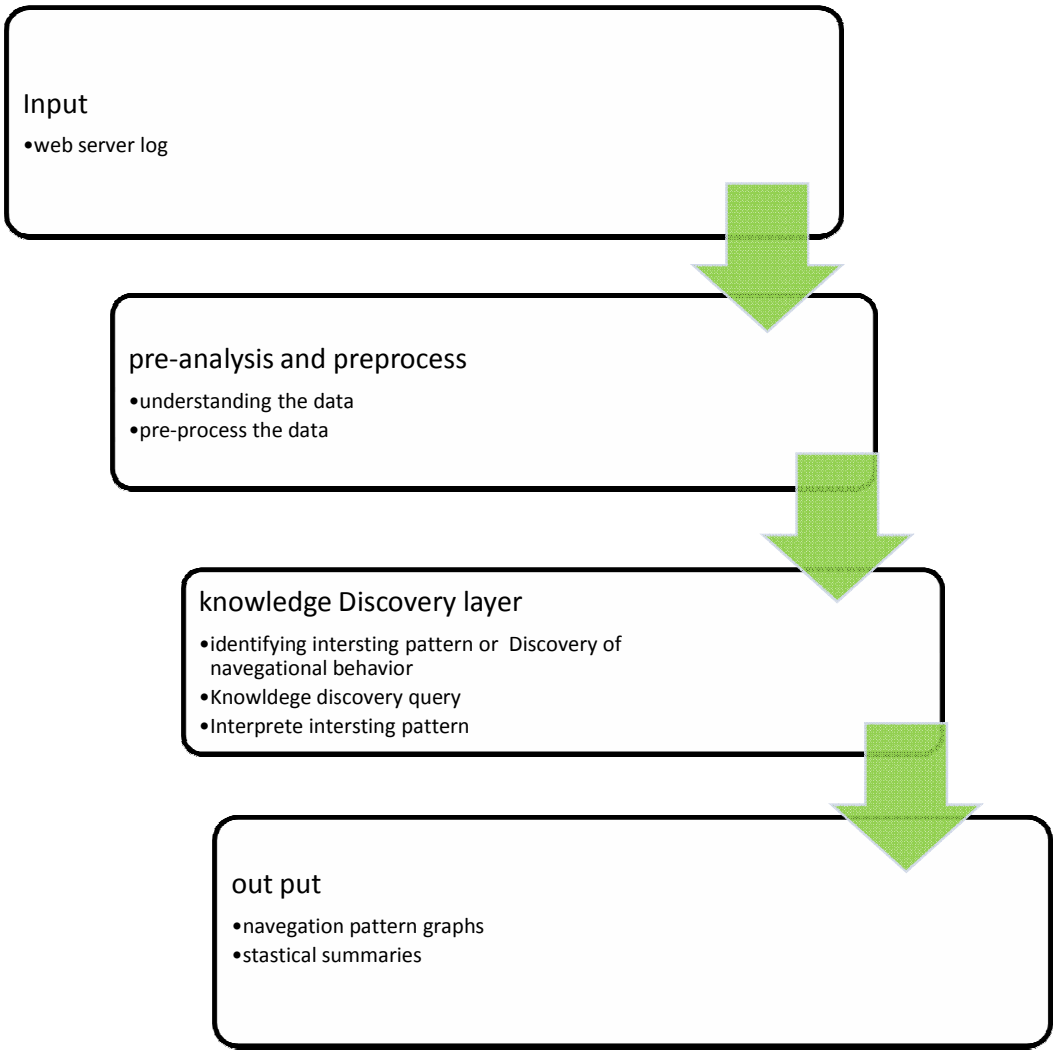


Figure 2:Research method flow

1.12. Data Collection for the Study

In this study the data has been collected from the official web site of the AAU, which is normally secondary data source since web log records every activity of the user regarding to visit of the web site.

1.13. 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 author of the paper, uses server data that are kept in the official web site of AAU in the format of extended log format, which is most apache server supports it.

1.14. Data preprocessing

According to olfa,et al, (n.d) , most log files are full of junks that are insufficient, inconsistent and including noise so the data pretreatment is to carry on a unification transformation to appropriate sets ; to have those sets there are some data cleaning phases are important to implement.

1.15. Data Cleaning

In ref olfa,et al,(n.d), the purpose 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 accordant to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

In addition to the above those also include some phases like, removing robot requests (filtering out spiders or crawlers which are known), removing duplicate requests (removing “dust”), and Filtering relevant status.(those concepts will be described in the Chapter Three).

.

1.16. Data analysis

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

1.17. Tools for Experiment

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, and Anália, et al., (2003),WUM (web utilization miner), Its primary purpose is to analyze the navigational behavior of users in a web site, furthermore ,Navigation pattern discovery is performed on the portion of the web server log that contains the sessions.

The justification for why these tools are selected is given in the chapter FOUR.

1.18. Interpret and report result

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

1.19. Application of results

The hidden unknown information in log formats are important in understanding of users navigational behaviors even if it is not possible to know what will be the results but some knowledge will be revealed by understanding of the general behavior of web site users of AAU .it can be used for improving the web site and it shows some way for further study.

1.20. Organization of the Thesis

This thesis organized as Six chapters ,the first chapter deals with the general introduction to the research of the area in this case the AAU, including the background of the Addis Ababa University in general, it also looks on development of ICT, and how looks like the structure of the official web site, what are their main purposes and later discusses statement of the problem, data collection ,data preparation with other subtopics like, scope and limitation of the study; objective of the study; research methods; etc.

The rest of this thesis is organized as follows. Chapter 2 presents two main areas, Literature review and related works regarding to Data mining and web usage mining.

Chapter 3 this chapter mainly deals with web usage and navigational behavior based on extended of the above chapter in terms of concepts.

Chapter 4 this chapter provides with methodology, in this presents the researcher points why select the tools for preprocessing and the tool for navigational behaviors in general, research process how to achieve the objective.

Chapter 5 in this chapter the experiment conducted and discussed which are based up on the methodology in the previous chapter.

Chapter 6 the last chapter, based on the experiment done in the previous chapter, the conclusions have been reached and recommendation and what it should be done for the future or further work in this research area.

CHAPTER TWO: LITERATURE REVIEW

2. Introduction

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), according to 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

2.1. 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. In Accordance with Berkan, y., (2002), the difference between them is that the former does not store Referrer and Agent information of the requests.

According to Srikant, et al, only few fields are available for navigational patterns discovery, which If are added to the CLF make up the so called Extended combined log format (supported by Apache Web Server).

2.2.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.

I. 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
```

II. 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)"
```

III. 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"
```

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

IV. 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, -,
```

2.3.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)"
```

The details of the fields in the entry are given in the following section.

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:

-

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

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 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)

2.4.Overview and 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 data base 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), refers as Data mining is 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 they can be 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 is often referred to as "secondary" data analysis.

2.5.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 is 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 is scheduled to depart, is 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).

2.6.Data Mining Approaches

It have mentioned earlier that the web usage mining is the application of data mining .those Data mining have two approaches according to (brendit,2011), the approaches is 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."

But, 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.

2.7.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 ref 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.

2.8. Taxonomy of Web Mining

In ref 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 2.8.1.

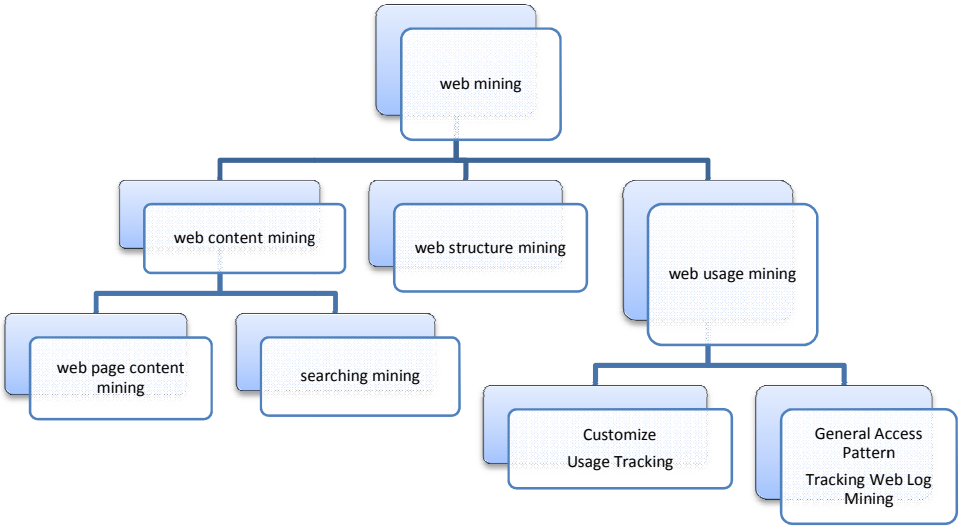


Figure 3: Taxonomy of Web mining,

2.8.1. Web Usage Mining: WUM

Web usage mining can also be defined as the application of data mining techniques to discover user web navigation patterns from web access Zalane et al, (1998), in addition to that, generalized definition accordance 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 to 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.

2.8.2. 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 him, 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.

2.8.3. 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.

2.9.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 web site. 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),this kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, like Most visited pages, average daily hits, etc.

In reference with Bamshad. et al ,(n.d), generalized as this kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, 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 ref 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, we can 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.

According to Bamshad et al ,(n.d)), generalized as 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 data base, 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.

In ref with Bamshad et al, (n.d)), generalized 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 ref (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 web site is needed .the second method computes clusters of page references based on how often they occur together.

In ref 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.

2.10. Related works

Data mining techniques are not easily applicable to Web data due to problems both related with the technology underlying the Web and the lack of standards in the design and implementation of Web pages. Web usage mining is a research field that focuses on the development of techniques and tools to study users' web navigation behavior.

2.10.1. Related Works on the Tools

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.

The other related works on tools on SpeedTracer, 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 web site 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.

Advanced mining algorithms uncover users' movement through a web site. The end result is collections of valuable browsing patterns that help Webmaster better understand user behavior. Further explained in the paper that generates three types of statistics: user-based, path-based and group-based. User-based statistics point reference counts by user and durations of access. Path-based statistics identify frequent traversal paths in web presentations. Group-based statistics provide information on groups of web site pages most frequently visited.

2.10.2. Navigation Pattern Discovery Tools

There are some web usage miner tools which can be used to the navigational pattern discovery for web user behavior of the web site, 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.

2.10.3. Related works in Advances Web Usage Mining

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 on the basis of user demographics and interests, as well as analysis of a site's success.

In Abhishek et al (2011), correlated but not linked web pages are discovered by clustering pages requested together by the site's visitors. This approach can be used to construct dynamic web pages automatically that provide links to pages considered relevant by earlier visitors Pierre, B., et al, (1996).

In the SurfAID project, a warehouse over web usage data is established and time series analysis is combined with association rules to discover unexpectedly evolving correlations among products (Abhishek, et al, 2011) propose the establishment of a warehouse, in which web usage data are combined with customer data, concept hierarchies on page contents and user demographics, as well as enterprise knowledge, e.g. in the form of previously discovered rules Myra,S., & Lukas C. (n.d). . Although user activities form the basis of these types of analysis, the issue of improving the site itself is not addressed.

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 web site 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, we present 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 web site, 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 web site association rules. The “Foot prints “ tool of records the footprints left behind by web site visitors and accumulates 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.

CHAPTER THREE: WEB USAGE MINING AND NAVEGATIONAL PATTERN

3. Introduction

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 web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is Access logs on server side, which keeps information about user navigation. Further explained in Sulu, G.,(2003), Web usage mining is the process of identifying representative trends and browsing patterns describing the activity in the web site, by analyzing the users' behavior. Web site administrators can then use this information to redesign or customize the web site according to the interests and behavior of its visitors, or improve the performance of their systems.

3.1. The General Process of Web Usage Mining

Today, understanding the interests of users is becoming a fundamental need for Web sites 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 can be exploited in the process of navigational behavior.

As it have mentioned earlier , Web Usage Mining (WUM) is the process of knowledge discovery and analysis of Knowledge from World Wide Web, represents a rather recent research field devoted to discover behavioral patterns from Web usage data.

As in Zalane et al (1998), the general processes of WUM distinguish three main steps: data preprocessing, pattern discovery and pattern analysis.

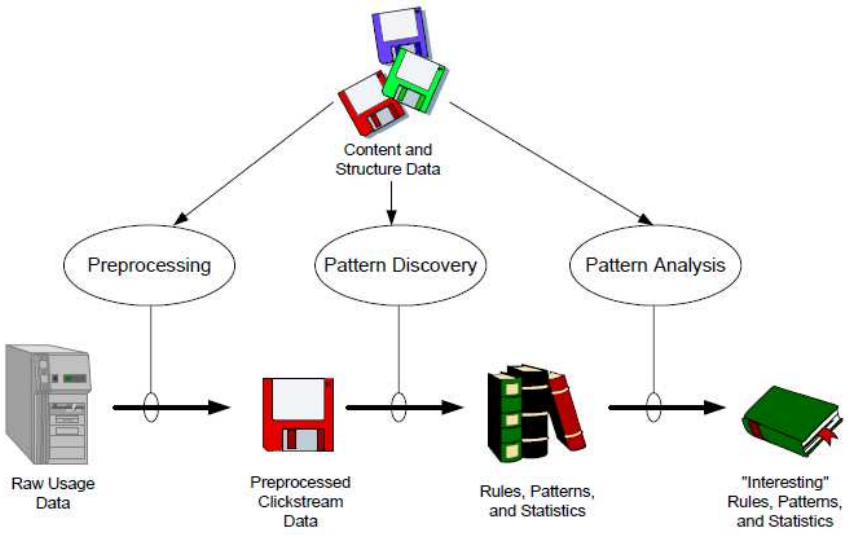


Figure 4: High Level Web Usage Mining Process Jaideep, et al (n.d), page 4

3.2.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.

Besides to the major sources of the data which have mentioned above but, there are also some other resources for web usage mining. According to Castellano, et al (2007) the following can be the source of the data.

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

A different researchers uses different collections over a time for web usage analysis in accordance with Berkan, y.,(2002), were collected for a period of two weeks for Logs Preprocessing and Sequential Pattern Extraction with Low Support.

3.3.Data pre-processing

In ref 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 ref 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.

3.4.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 ref 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 also their Performance is not enough for huge volume of data.

Researchers have shown that the log files contain critical and valuable information that must be taken out. It makes web usage mining a popular research area for many applications in the recent years.

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).

According to 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 Web site and outputs 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.

3.5. Data Cleaning

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

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

3.6. Removing Unnecessary Records

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; In ref 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. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

3.7. 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 s. 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. 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.

In ref 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 Web sites by following every hyperlink on every page within the site in order to update the index of search engine. Requests created by Web robots are not considered usage data and, consequently, have to be removed. To identify web robots' requests, the data cleaning module implements two different heuristics.

Firstly, 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.

3.8. 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 ref (Rajni, P., et al 2009), User's identification is, to identify who access web site and which pages are accessed.

But this task is not easy because few web sites that uses authentication to access the resource so the web records, only records the visitor's host and user agent. Further explained by 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 ref 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. A session is a series of web pages user browse in a single access, or according to Castellano et al, (2007), A 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.

3.9. 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 web site 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; or in e-commerce sites.

- 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 web sites. 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 web sites from the users' viewpoint. According to Federico et al (2000), Exploits,Web Usage mining techniques to suggest proper modifications to web site. Adaptive Web sites represents a further step. In this case, the content and the structure of the web site 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. in ref with (Sulu, G.,(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.

3.10. Navigational Pattern and Sequence

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 we instruct the miner to find only maximal frequent sequences, we obtain fewer and more compact results.

In the ref 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: When we investigate the causes of manufacturing errors, we only want the sequences containing error and cause, not the many events in between. The same is true when we search for operating system signals.

Comparison of GSP and AprioriAll

According to Murat et al (n.b)),On the synthetic datasets, GSP was between 30% to 5 times faster than AprioriAll, with the performance gap often increasing at low levels of minimum support. The results were similar on the three customer datasets, with GSP running 2 to 20 times faster than AprioriAll. There are two main reasons why GSP does better than AprioriAll.

- GSP counts fewer candidates than AprioriAll.
- AprioriAll has to first find which frequent item sets are present in each element of a data-sequence during the data transformation, and then find which candidate sequences are present in it. This is typically somewhat slower than directly finding the candidate sequences.

GSP, a new algorithm that discovers these generalized sequential patterns and has the following advantages for example.

- Empirical evaluation using synthetic and real-life data indicates that GSP is much faster than the Apriori.
- All algorithms presented in GSP scales linearly with the number of data sequences, and have very good scale up properties with respect to the average data-sequence size.

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 5: The mining Algorithms of WUM

3.11. Navigation Patterns and Important to Discover

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)), we 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. In WUM, Mary et al, (2000), an “interestingness descriptor" is a query in our mining language, MINT.

3.12. Knowledge Discovery Queries

Similarly to Lukas, (n, d), we believe 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.

3.13. Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process in accordance with, challenge of pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users.

The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL or MINT query. 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. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match

CHAPTER FOUR: METHODOLOGY

4. Overview of the methodology process

According to Dipa,(2010), web usage mining have three main process in order to discover a knowledge from the data ware house, author of paper use for his work according to this researcher, described above, it is necessary to perform three steps, see fig 5,but the detail of those how to accomplish those main process are described below.

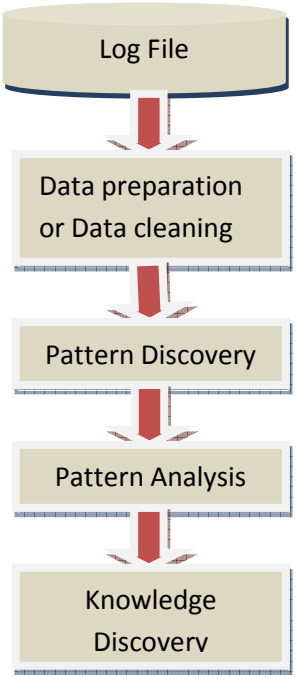


Figure 6: web mining usage main process to discover knowledge.

4.1. Tools Selections for Preprocessing

As stated earlier in chapter ,there a lot of tools uses for preparing a dataset for the intending purpose but the selection of those tools is not easy since every tool have designed for specific purpose but none of them cannot give a good output unless they combine each other in order to meet efficient output. The author 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.

The author 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 author found no need to implement his data preparation into navigational discovery software, besides to that even if the WUM have some capabilities of preprocessing, but does not support the main preprocess phases such as removing robot hosts and etc.

Extended log format of AAU

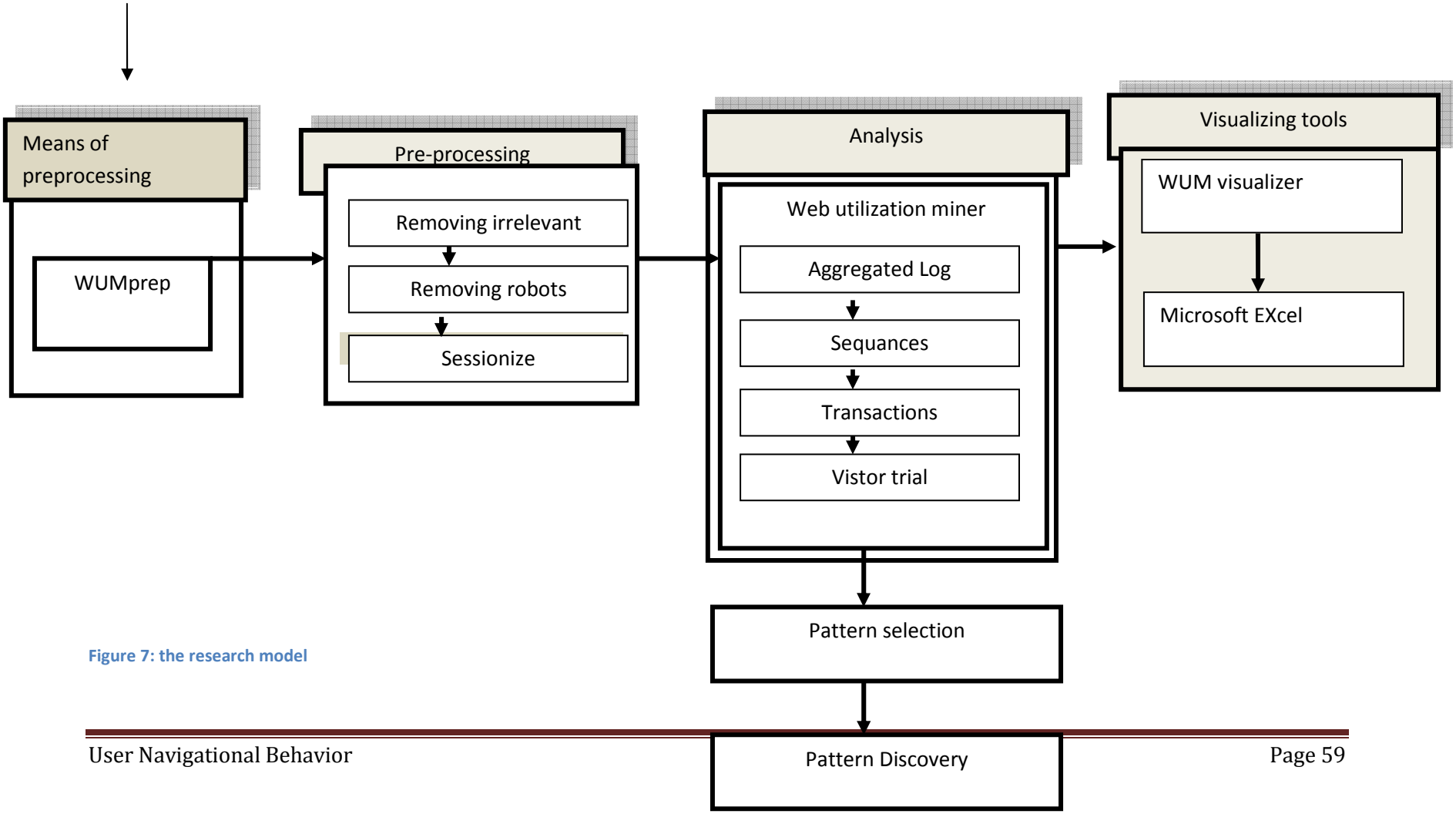


Figure 7: the research model

As it have been mentioned in the above figure 6, which shows how the objective could be achieve, even if the preprocessing done using the WUMprep, the researcher regulate the configuration of the tool to meet the objective . The data cleaning is done based on the following criteria.

4.2. Removing Irrelevant Records and Status

The removing of irrelevant records are significant as it have mentioned in the chapter three , as these requested log files are not only contain requests to the pages comprising the Web site, but also requests of images, scripts etc. embedded in these pages.

The author of this paper uses to remove those embedded extension of files should be removed because these “secondary" requests are not needed for the analysis and thus irrelevant (they must be removed from the logs before mining).Those requests are in the following table with their definitions:

\.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
\.JPG	The most common image compression format used by digital cameras.

Table 3: Irrelevant list of requests

Beside to that, the author only interested on request which only have the status 200 series, because concern only successful requests which mainly shows the users who get what they want and the other requests' are not need any more .

In general, according to the researcher these requests do not represent the effective browser activity of the user visiting the site; hence they are deemed redundant and should be removed.

4.3. Removing Robots

The author of this paper strongly believes to distinguish between human users and hosts that are robots, there exist several heuristics as it have mentioned above in chapter, section three. They are implemented in the script. Firstly, all records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed from the original log files.

4.3.1. Removing Duplicate requests

If a network connection is slow or a server's respond time is low, a visitor might issue several a successive clicks on the same link before the requested page is finally showed in his browser. Those duplicate requests are noise in the date and should be removed. The author of this paper uses, the most widely accepted threshold for of 2 seconds between two consecutive requests the entries that corresponds to robots can be eliminated.

4.3.2. Sessionize

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, the author uses a Session which is computed by taking any URL time stamp ,to achieve theses the researcher uses the most accepted time threshold which is 1800 sec or 30 min to identify the sessions using the these timestamp.

4.4. Divide log format

The preprocessed data needs to be dividing into manageable size before feed into WUM because it takes long time to process the data, so the researcher writes a python code to prepare the processed data for the WUM.

4.5. Tool Selection for Navigational Behavior

The transformation of the web server log into a log of sessions appropriate for mining and the process of navigation pattern discovery are performed in the framework of the Web Utilization Miner WUM, according to Anália et al., (2003), WUM (web utilization miner), Its primary purpose is to analyze the navigational behavior of users in a web site, furthermore, Navigation pattern discovery is 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 used as a basis to analyze the sessions in the rest of the log, comprising the sessions of the active investigators that did not become customers.
The architecture of Web Utilization Miner, There is two major modules: the Aggregation Service prepares the web log data for mining and the MINT-Processor does the mining.

In ref Bettina et al (1999), The Aggregation Service extracts information on the activities of the users visiting the web site 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. We call this the “support” of the node.

In accordance with Marya, et al (n.d), The MINT-Processor mines the aggregated data according to the directives of the human expert. “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.

In ref 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’s designed to work with The WUMprep module (which is responsible for the pre-process phase ;)
- Its free and open source tool (not commercial)
- 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.(is also open source and free)
- WUM uses for the discovery of navigation patterns and visualization of interesting Patterns.
- It’s a sequence miner and support GSP algorithms.
- 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.

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 web site 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.

4.6. General Methodology

The overall pictures of the methodology can be described as the following figure 7, the WUMprep scripts does the preparation steps(in the above illustration) , which is the input for the WUM tools discovers the navigation pattern and mining patterns and visualize the result using WUM visualize based on the miner interests.

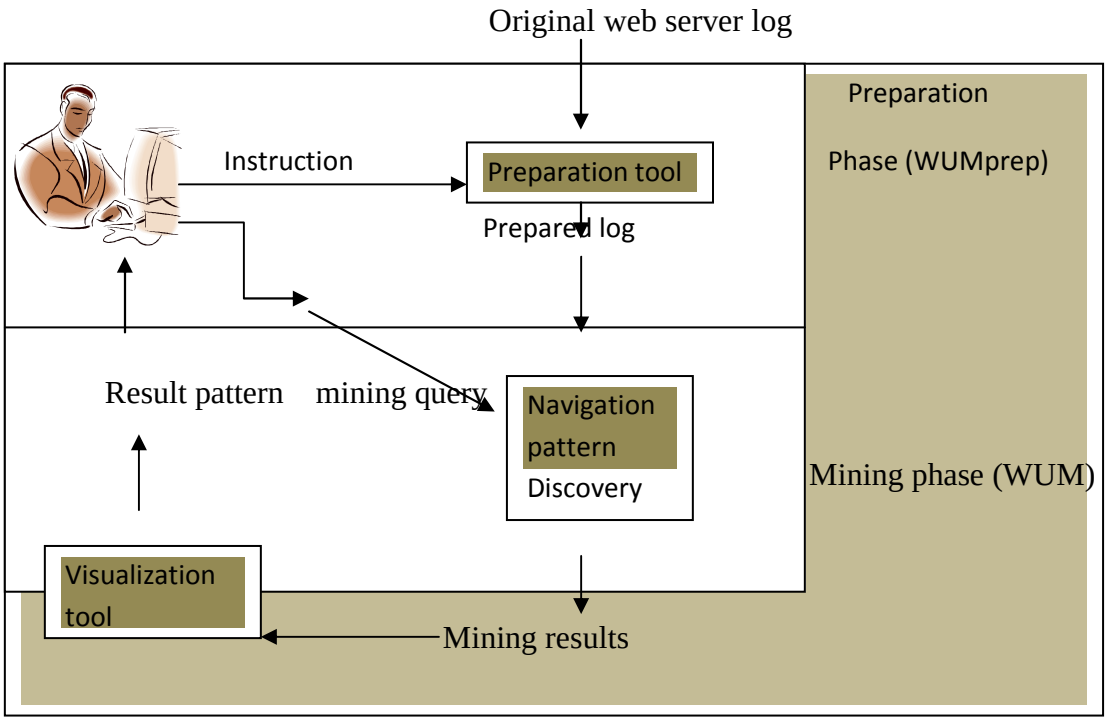


Figure 8: navigational process of WUM

CHAPTER FIVE: EXPERIMENT

5. Over view of Experiment setup

The experiment has been conducted on the following setup

- **Computer Type:** *personal computer (X32-based PC)*
- **Operating system:** *OS Name 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 the web log into manageable size

5.1. Data Collection and Selection

The data for this study is a web access log data of AAU official web site .As mentioned in the chapter one, a web log data is favored by many for web usage analysis. Two months web access logs have collocated for this study, for December and November.

5.2. Data Cleaning

The data collected from the AAU web server logs are full of junks that are not cleaned and should pass through some data cleaning phases (see the figure below) ,it is important steps to truck down the exact behavior of the user of the official web site unless they removed it is difficult to achieve the objective of this paper. Those phases must be undertaken to have cleaned data for further uses (process). The sample Log data are collected from AAU before preprocessing.

```
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)"
```

Table 4: A small extract of a Web server log contents

From the original web log see table 4, which can be easily seen a lot of junks, noises as well as robots (spiders, crawlers) those should be removed in order to have clean web logs to have appropriate ,efficient ,effective data logs.

5.2.1. Removing Irrelevant

As a result of removing irrelevant the number of log lines decrease in enormous seize the reason for it , those log files which contains the some extensions (see previous chapter), and those repeated requests that may came from inpatient users will be eliminated that's why the number of records seized in such amount. The original size of the records before were 50701 KB records (KB) and after the log filter preprocessing it became to 12416.KB.

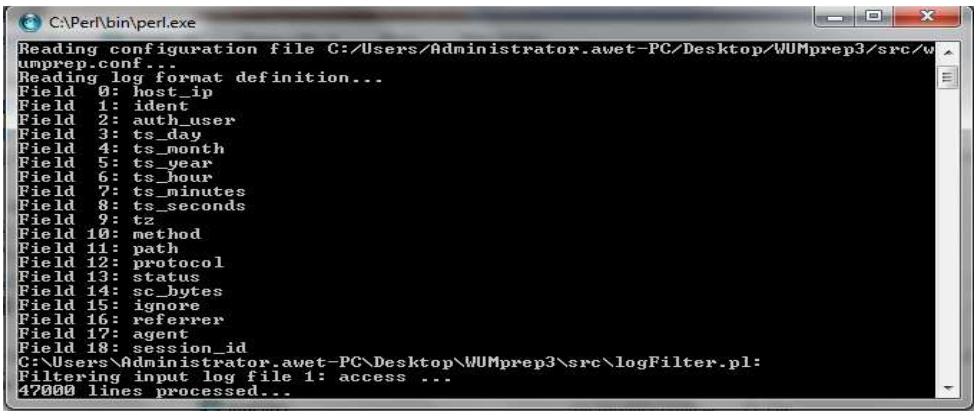


Figure 9: removing irrelevant records sample.

5.2.2. Detect Robots

The process of detect robots are very important to elimiatee the irelevent records whch are caused by the misusers that cames from other resources like (speider ,web crawlers) in other words , web surfing requested that are too fast that ordirnary pepole do not do in such fast ways caused by web crawlers.According to my expertment the number of robots are based on the maximum page view and against“index list“ in the WUMprep.The number of robots that detected from the web seerver logs are shown below,

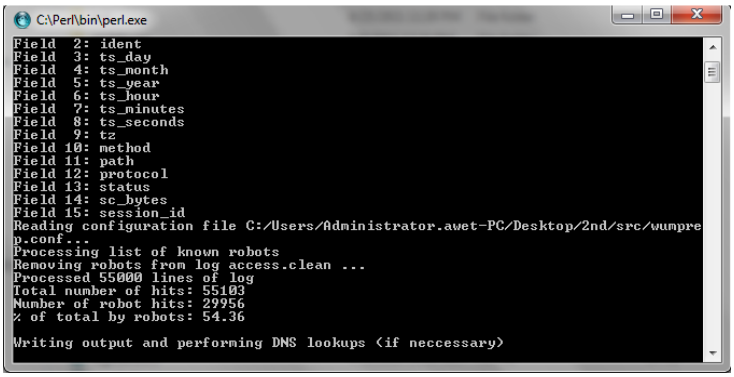


Figure 10: sample removing of robot hits

According to my experiment, for the months of December, the numbers of robots inside the Log format are 54.36 % robots from the total hits, for the months of November the total number robots are against the total hit are 39.68%. Samples of robot log lines that are 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] "GET /robots.txt
HTTP/1.1" 200 --304 "-" "Mozilla/5.0 (compatible; DotBot/1.1;
http://www.dotnetdotcom.org/,crawler@dotnetdotcom.org)"
(robots.txt)
```

Figure 12: Sample robot log files.

From the above results what it can be observable easily that some of the requests that came from same IP address that is (208.115.111.247) within two seconds, those

requests originated from the same IP address within two seconds. ([05/Dec/2010:05:03:20 +0300) and (05/Dec/2010:05:03:21 +0300)

5.2.3. Sessionize

The Sessionize which are resulted after the detection of the robots and give the following results as shown below,

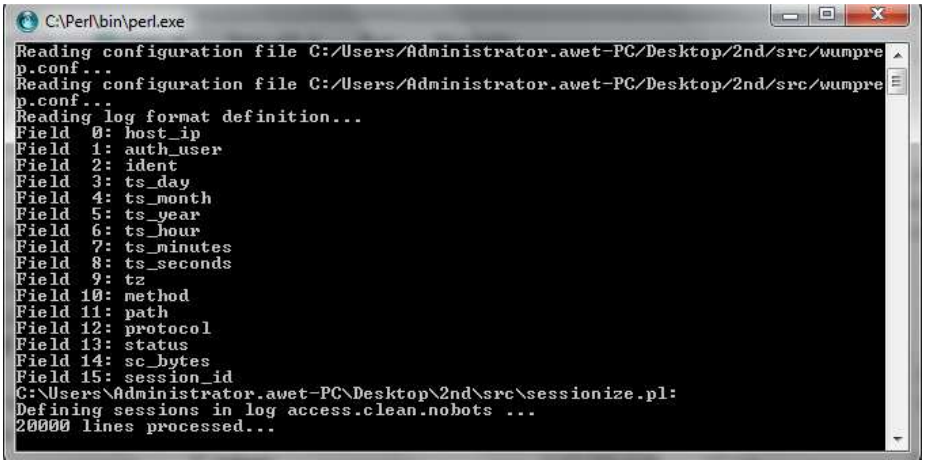


Figure 13: sample sessionaize process

The sesseinalize creates number of sesseionize, according to my experiment the number of session's creates are about 23411.some log lines which are exceed from the threshold i.e. 1800 sec or 30 min are removed .for the detail see in sample of in the appendix.

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

As it can be observable from the above fig 13,that the only status that filter from the web log files are GET and the status of 200 which indicates the successful requests from the web sites users, besides to that the session are identified . The types of log formats are converted from the Extended log format into Common log formats (see chapter Two, types of log formats).

5.3. Generalized Reports on Log Preprocessing

In this section the result of preprocessing will be discussed in general manner, an average user requests per day is 200220 lines. The preprocessing phases undertaken for both months (December and November) gives the following results after undergone through different phases of preprocessing for the months, and summarizes for one week in December the following tables. See for the months of November in appendix.

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
220340	150127	70564	25005	25005
230087	160743	72087	24060	24060
200406	148906	63480	21000	21000
190967	138967	50653	19734	19734
200190	178300	60752	20674	20674
200150	167543	47897	19653	19653
220205	120950	62096	23765	23765

Table 5: A Sample records for the week in December after undertaken the preprocess phases.

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

As it have been mentioned earlier the log files are contains irrelevant data, irrelevant records and noises, that's why we can observe from the above experiment result in the table the size of original log entry records decreased in average of 80%.for the months of November the size of records of original entry decreased in average of 73%.(see in the appendix).

5.4. Navigational Behavior of December

5.4.1. Aggregated LOG tree

The aggregated tree are results from the web miner after the sessions creates based on the above preprocessed tool (WUMprep) and imported to the miner resulted the aggregated tree for the months of December as follows.

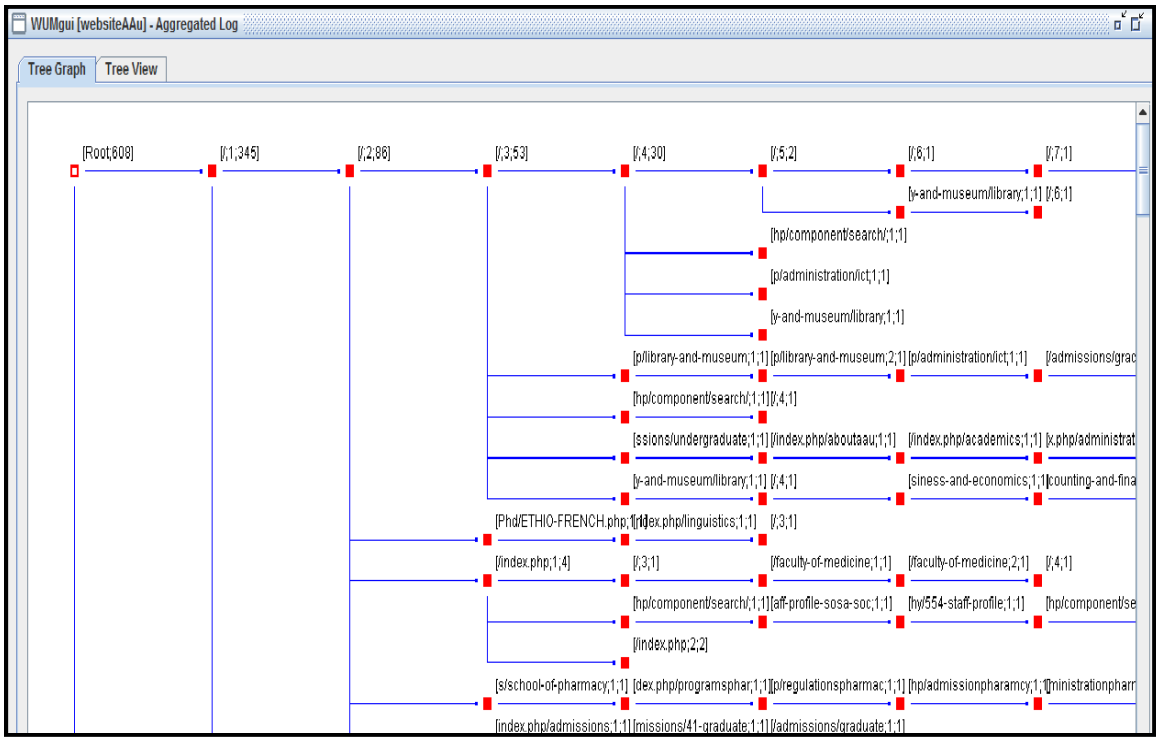


Figure 15: Sample aggregated tree for the month of December

As we can see from the above figure 14 ,an aggregated tree that the total number of nodes or total traverse make by users are 608 for the month of December, based on the aggregated tree the MINT query applied to find interesting pattern or for sequence analysis from it. The researcher chooses the some Examples to find interesting pattern for the month of December. For the month of November see in the appendix.

5.4.2. Sequence and Navigational Discovery of Users

As previously mentioned in chapter three, the generalized sequence pattern describes the behavior of users by filtering out the interesting pattern from the aggregated tree using the MINT query. ,In the following sections, the experiment is undergone using some most interesting patterns using the MINT query to discover the most important issues that should be discovered according to the researchers of interest , like Where do visitors of page Home go after wards?, Where do visitors go after typing the www.aau.edu.et (/)?, To Find out pages that always visit together and look at its pattern, Where do visitors go after search page of AAU (/index.php/component/search)? What is interested in navigation patterns between two pages.

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

Using the MINT (see appendix for syntax of MINT) query the author is interested where users go after the accessing the home pages until the next five pages, using the following query to the MINT to discover users' navigational behavior.

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 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
```

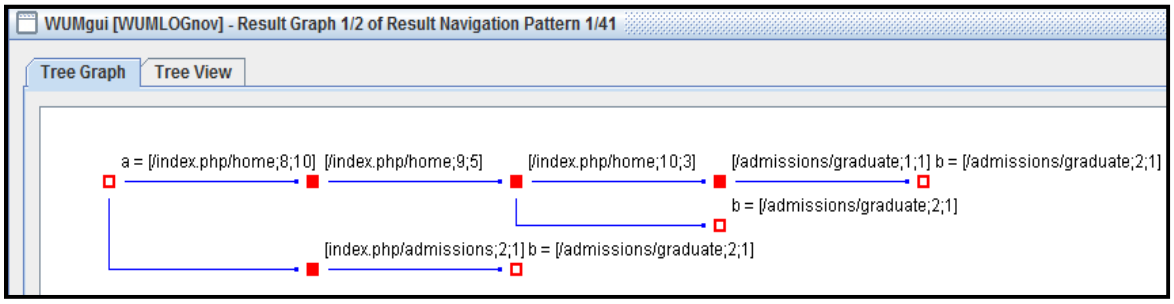
The above query results a following patterns using WUMvisulizer but the author puts some sample results in the following figure.

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

Here, we receive all pages reached within 5 pages after HOME (index.php/home), which has been accessed 100 or more times, provided that those pages have been accessed by at least 50% or 100% of the visitors visiting HOME, but as we can see from the result the most accessed pages is /index.php/library-and-museum users stay 100% visting the content of it, It is also clear that most users who visited the home page also stay in 100% within the page of /index.php/registrar those are the most .of course the other pages like /index.php/admissions/graduate users stay in those pages users stay in the page for average 26%,even if they are the most visited pages after Home pages.

Navigation pattern:



As we can see from the navigation pattern most people are going to the page of /admissions/graduate after visiting the home pages, it's clear to see that most users stay in

the HOME page (/index.php/home) and navigate 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 above MINT 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 receive 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%.

Navigation pattern

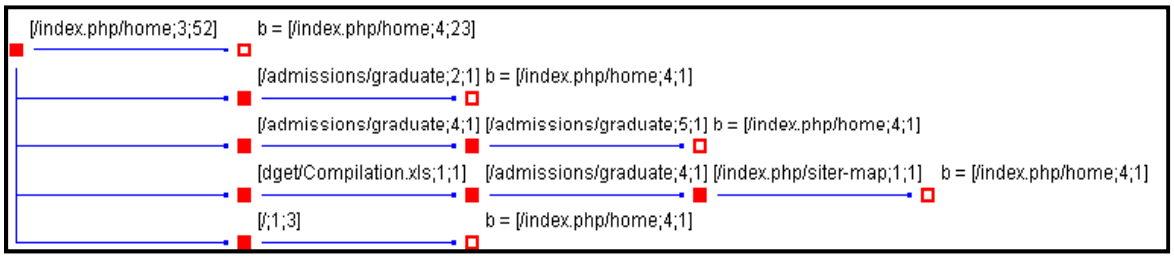


Figure 16 :Navigation pattern

From the figure 15, its easily observable here, we see that when visitor start from looking at **/index.php/home** page, 20% of them will stay within this subject area.

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

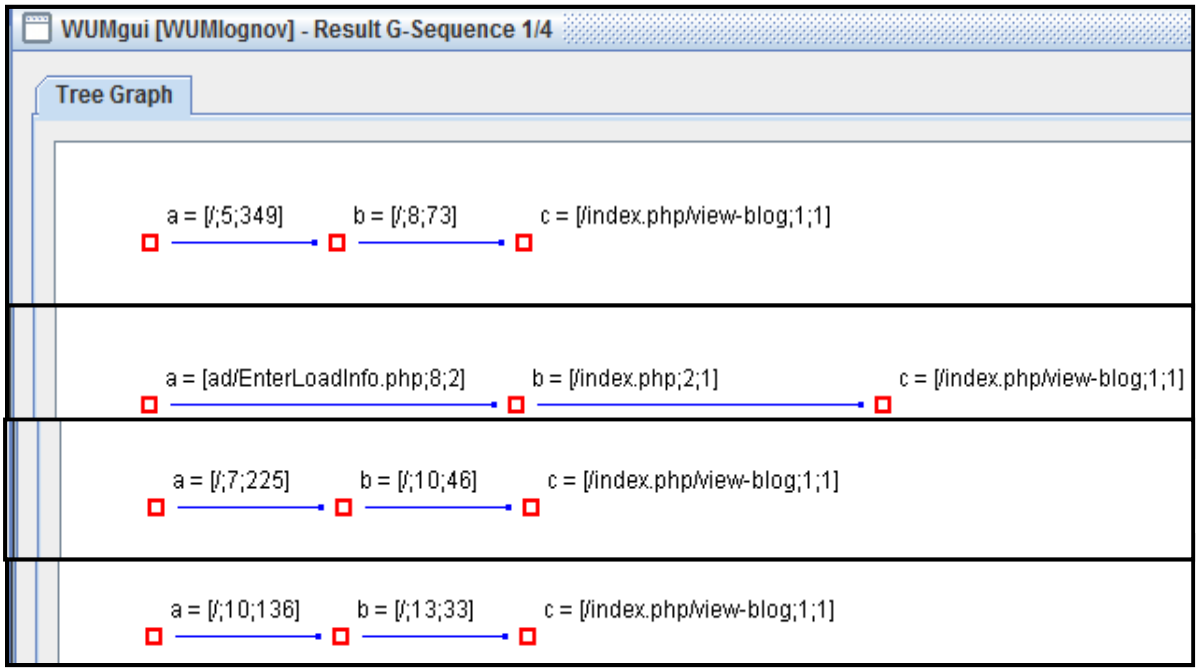
In this query, we specify a template t with two variables a, b, thus seeking for with two pages bound to a and b and at most 5 arbitrary page occurrences in between denotes that "a" should be bound to the first page which is /index.php/home, this page should be visited at least 20 % and b page should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

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 out of the query give us two patterns ,Here we recive 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 stay in the home page before reaching to /index.php/view-blog pages.

G-sequance



the Users do not take a single paths to reach to /index.php/view-blog most of the users take a path from the root pages,and the second most users take 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
```

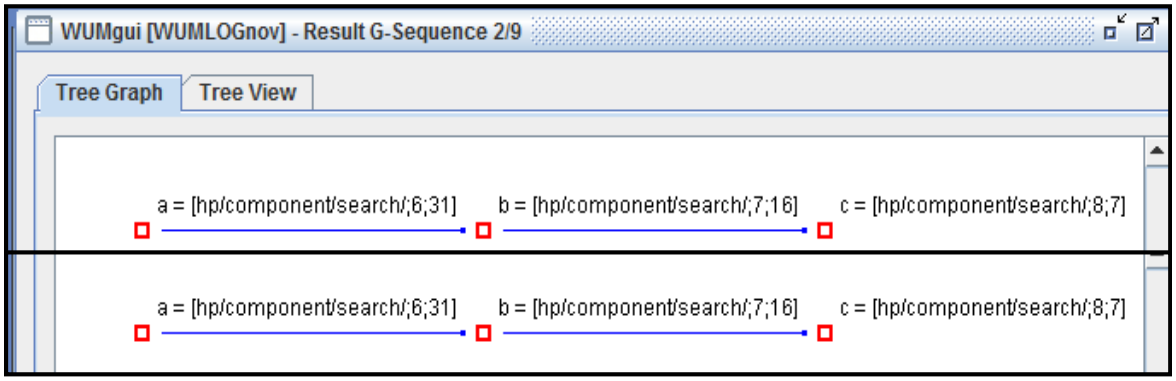
Pattern

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 ten patterns show that user's do know where they are looking for. most of users who stays in search engine 100% and also stay in this page for average of 40% ,they do search function stay within search the page.

G-sequence pattern



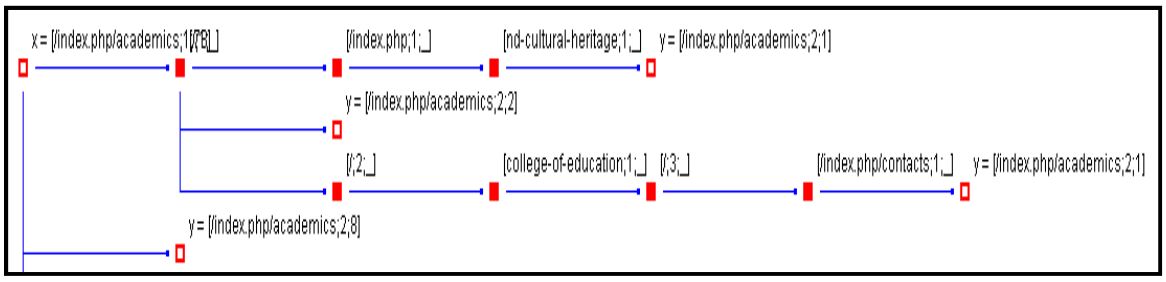
Sample of the above ,the author do not need to put all the g-sequence from the navigation pattern that users stay in the search page as we can see from the above result, that users stay in the search page.

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"
```

The above query results the following 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 web site.

5.5.Statistical Analysis for the Months of December

The WUM can generate a comprehensive report in terms of simple tables the researcher used other tool (Microsoft Excel) for better visualization. report will be discussed like what are ,most requested pages, most visited pages, most visited directory as well as most referee pages for the month of December will be discussed .For the month of November see in appendix.

5.5.1. Most requested pages

The following table shows the top ten most accessed pages during the months of November .For the rest of the months see in appendix.

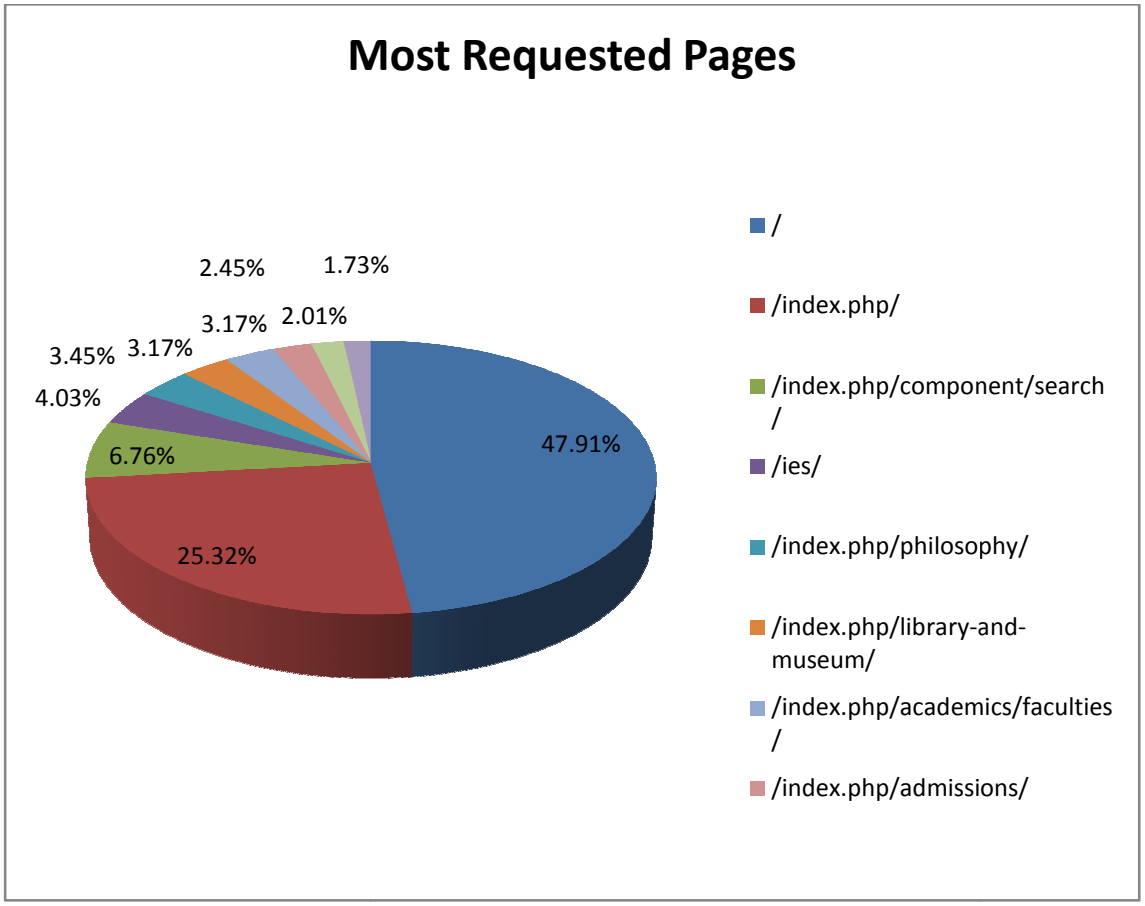


Figure 17: Top 10 most requested pages.

A figure shows the Top ten most pages during the months of December ,As it is shown the most requested pages is the / 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 is most popular by most users in all the three months .in fact ,this shows that most visitors enter into the site directly by typing the web site address as it shown in the above directory. The search engine of the Addis Ababa University the second most accessed pages followed by the /index.php/library-and-museum pages.

5.5.2. Most visited directories

The root directory "/" is the most accessed directory where the root directory in root folder is located .Most users also shows interest on the contents under the **/index.php/** folder. It is also possible to say that from the output those are also important visited pages /index.php/component/search/.

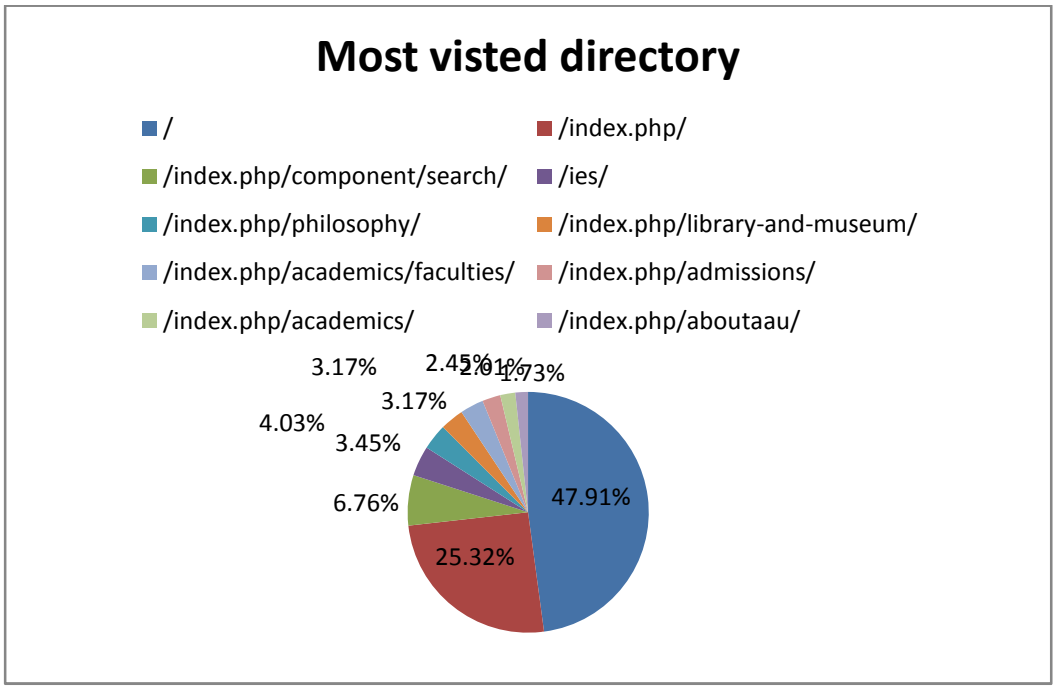


Figure 18: Top ten requested directories

For the rest of the months, most of the directory are requested ,are the same as the above until the top three directory but the others are became familiar in the next pages.

5.5.3. Most Top Entry Pages and Top Exit Pages

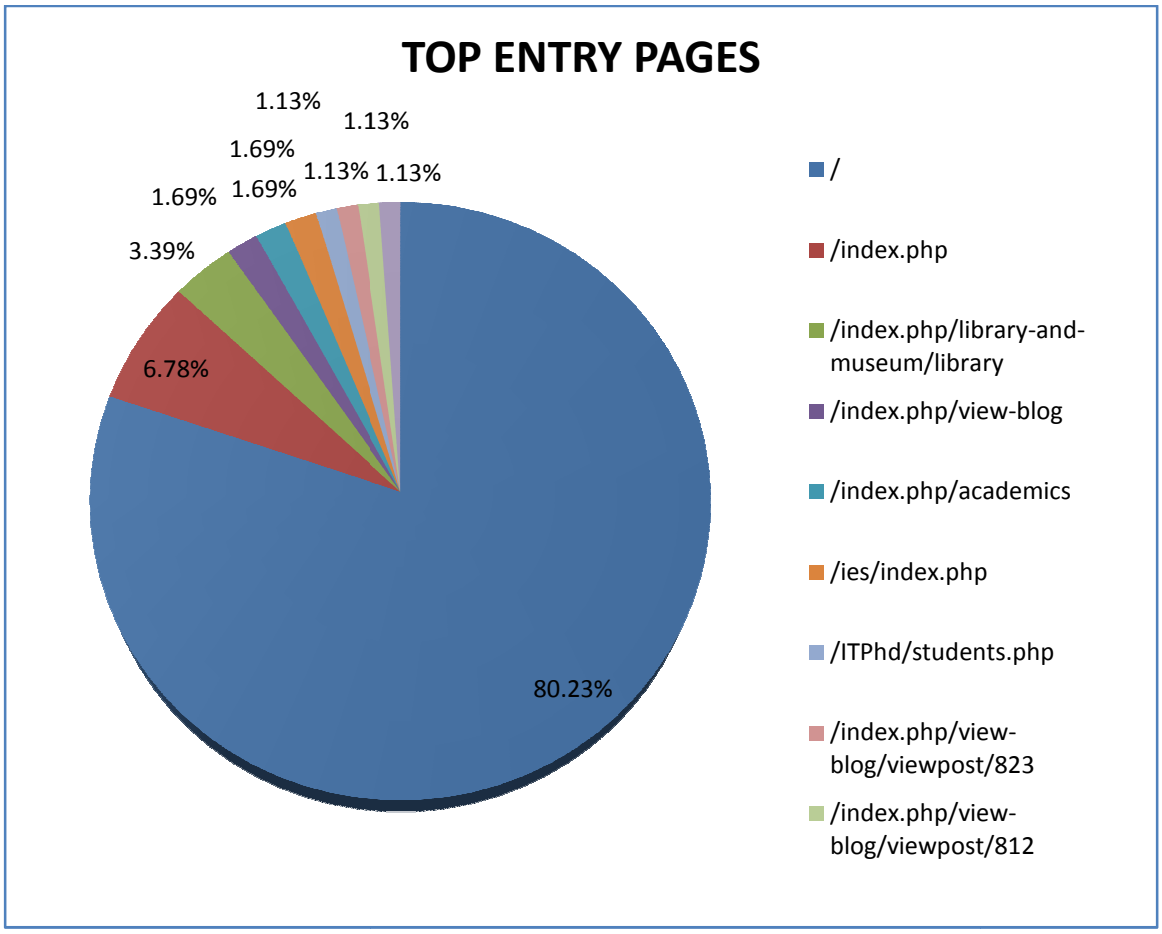


Figure 19:Top ten entry pages

The entry pages are pages that indicated that the web site users first visited where as the top exit pages is the last pages the users visited official web site .From the figure below what we can observe is that the "/" root pages where it is located accessed more than any other pages almost half of the request (80. 23%) and the /index.php the second most top entry page and last not least the /index.php/library-and-museum/library the third most top entry of pages, followed by /index.php/view-blog and /index.php/academics the 4th and 5th top entry pages. For the rest of the months see appendix of December.

For the month of November, as it as shown in the figure 55.68ve % of the visitors have entered into the web site directly through the / and index.php.

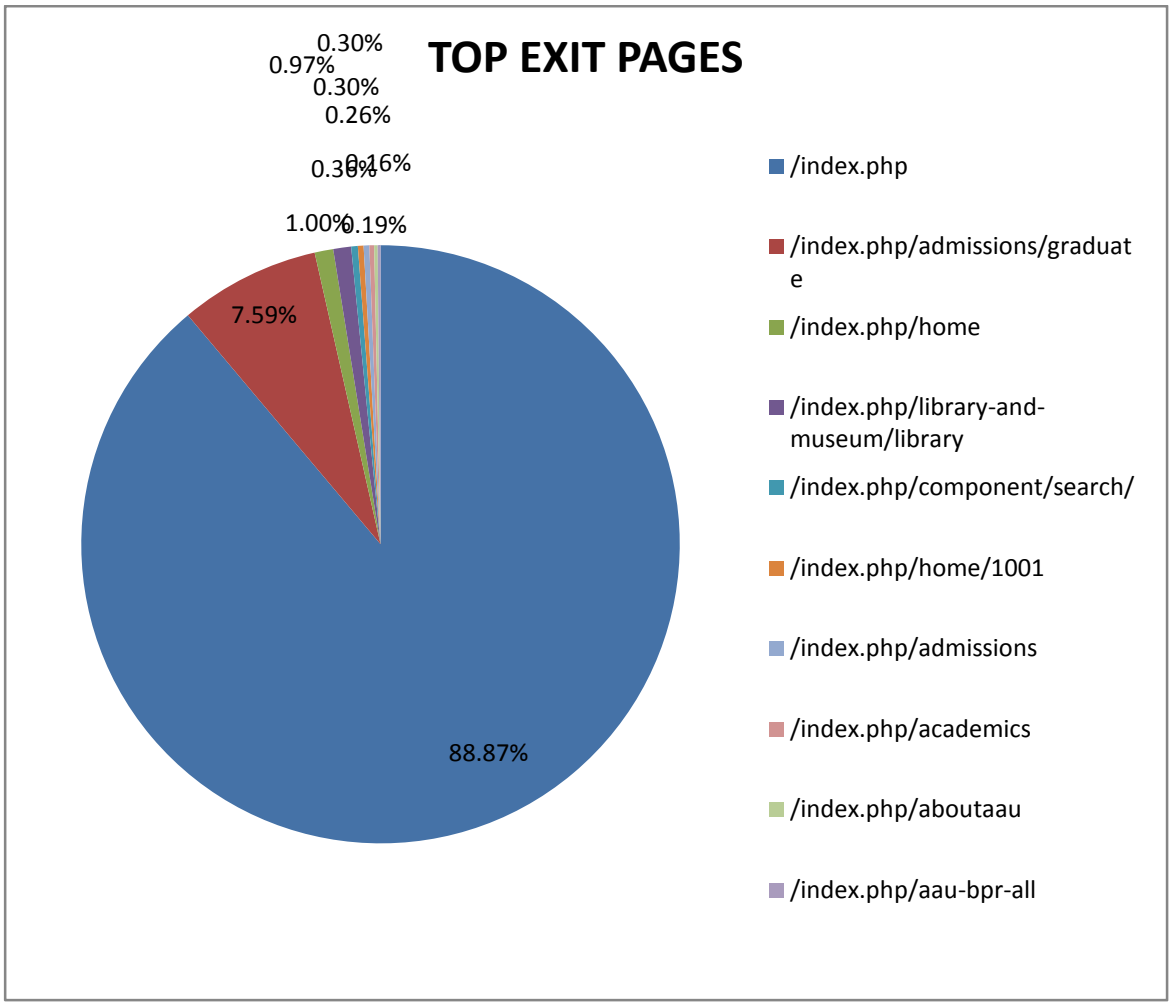


Figure 20: Top most exit pages.

From the above the figure ,we can see that the top exit pages are the "/" or after the user type the web site address and leave the web site 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/.

5.5.4. Top Referrer Pages

The top referee pages are pages where the visitor was located when making the next request with the official web sites.

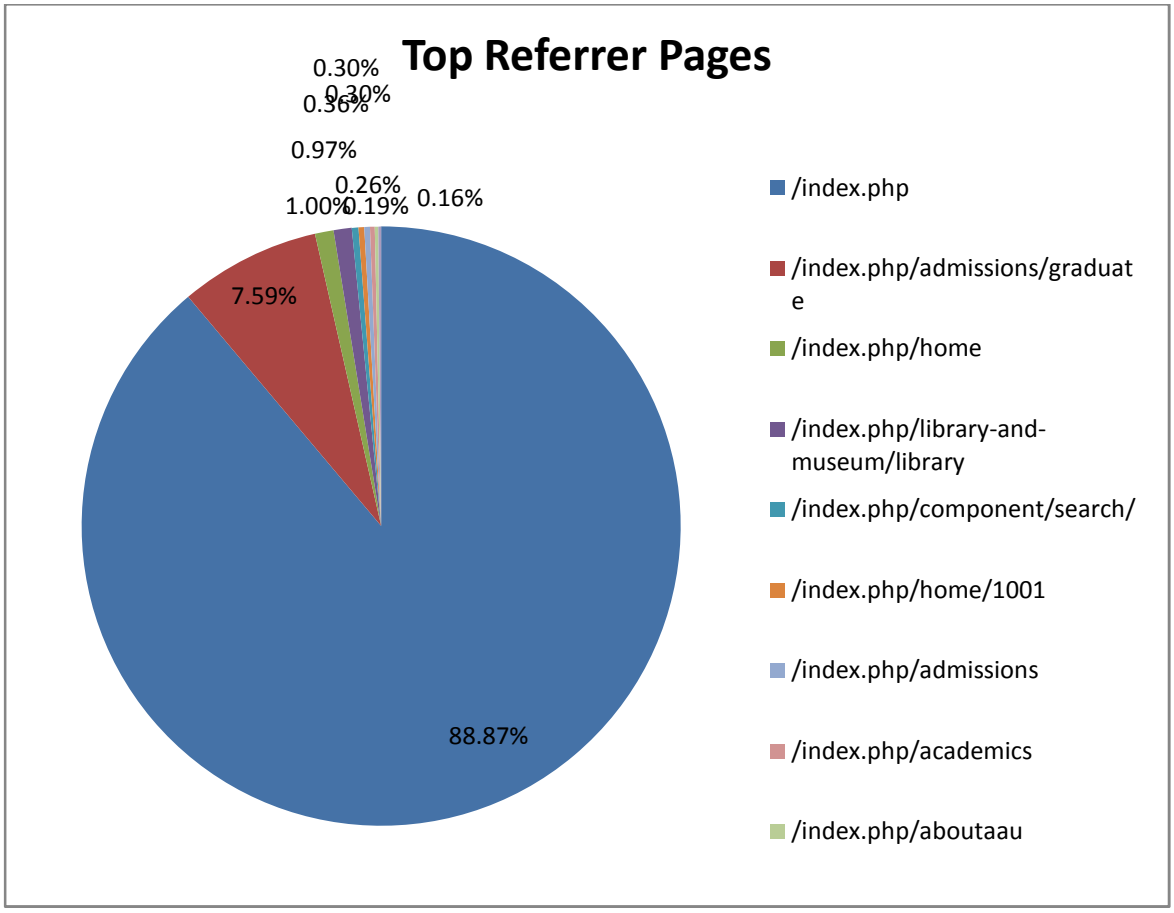


Figure 21: Top Ten referee pages.

From the above figure , what it can be easily observable that most users makes the next request from the page of /index.php which covers in more percentage 88.8 %. the next popular page where users requests the next page are initiated from /index.php/admission/graduate and which covers 7.33%.the third most refree pages are /index.php/home which covers the percentage of 0.96%.

For the months of November are almost the same as the above but the only difference are below three requests for more details see in appendix.

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATION

Conclusion

- From the navigational behavior of users that we can indicate easily users is no single point where users go after home page and can be conclude that users navigate from top of the page (hierarchy) to the lower hierarchy.
- From the navigational behavior search behavior can be conclude that most users use the search engine effectively or know what they are looking for.
- most request pages are requested to the web site by typing the official name of the web site that is www.aau.edu.et,that's why the most request web page becomes the root page of course it clear that the web server is an apache server, when type the official name hit the root directory of the official web site .from the request pages the second top most requested pages are /index.php/component/search pages, it indicates that most users use this page for searching key words with in the pages. What else can be concluding that /index.php/library-and-museum the third request pages, can be conclude that most users are interesting in the content, the reasons could be most journals associated to it.
- Most visited directories are of course the root directory since most of users are typing the name of the official web site 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 in side it.
- most users use enter to the web page using the page of /index.php , /index.php/library-and-museum/library, /index.php/view-blog and most users also leave from those pages that it can be conclude that almost the other pages 1/3 ,most users leave without visiting the web pages.
- It easy to see that most users use the /index.php, /index.php/admissions/graduate, /index.php/home are most users requests from those pages to make for further

requests, so it would be very useful if the administrator can put some urgent notice and advertisement within those sites since they are most accessed web sites.

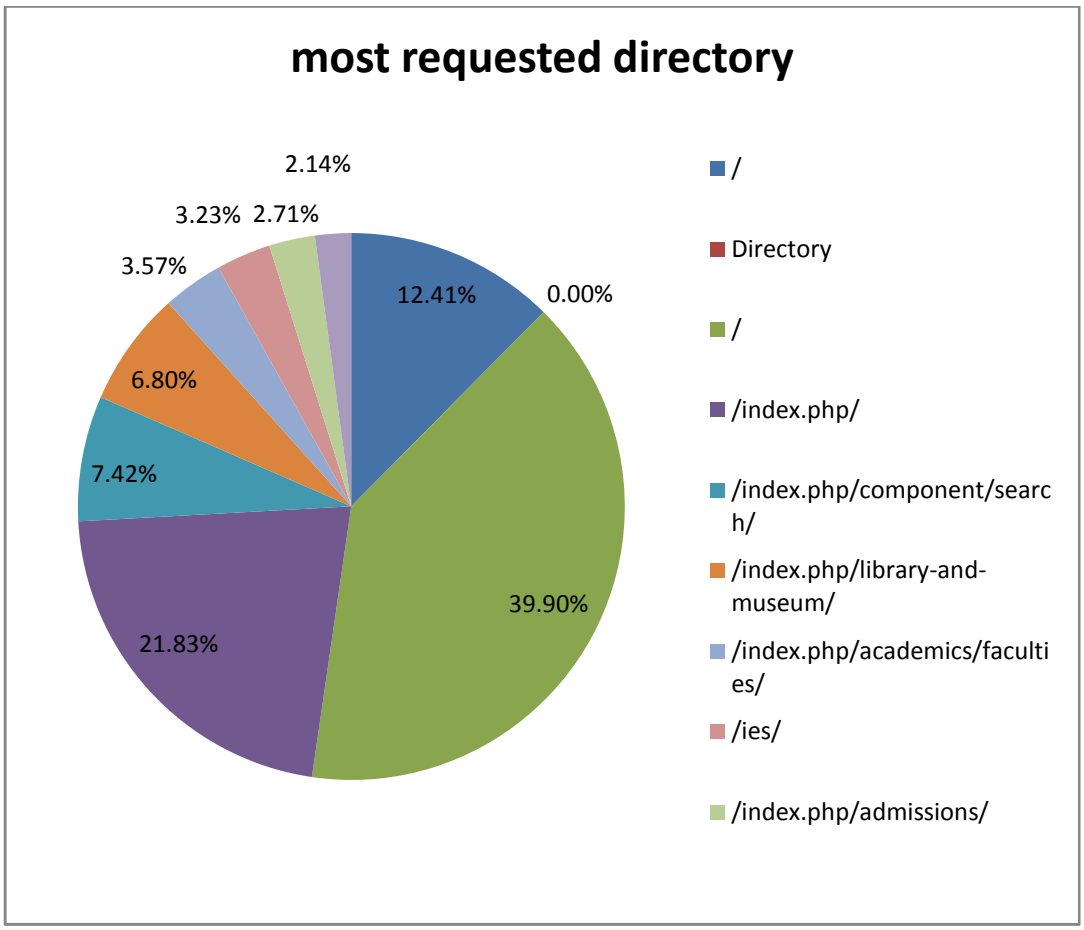
Recommendation

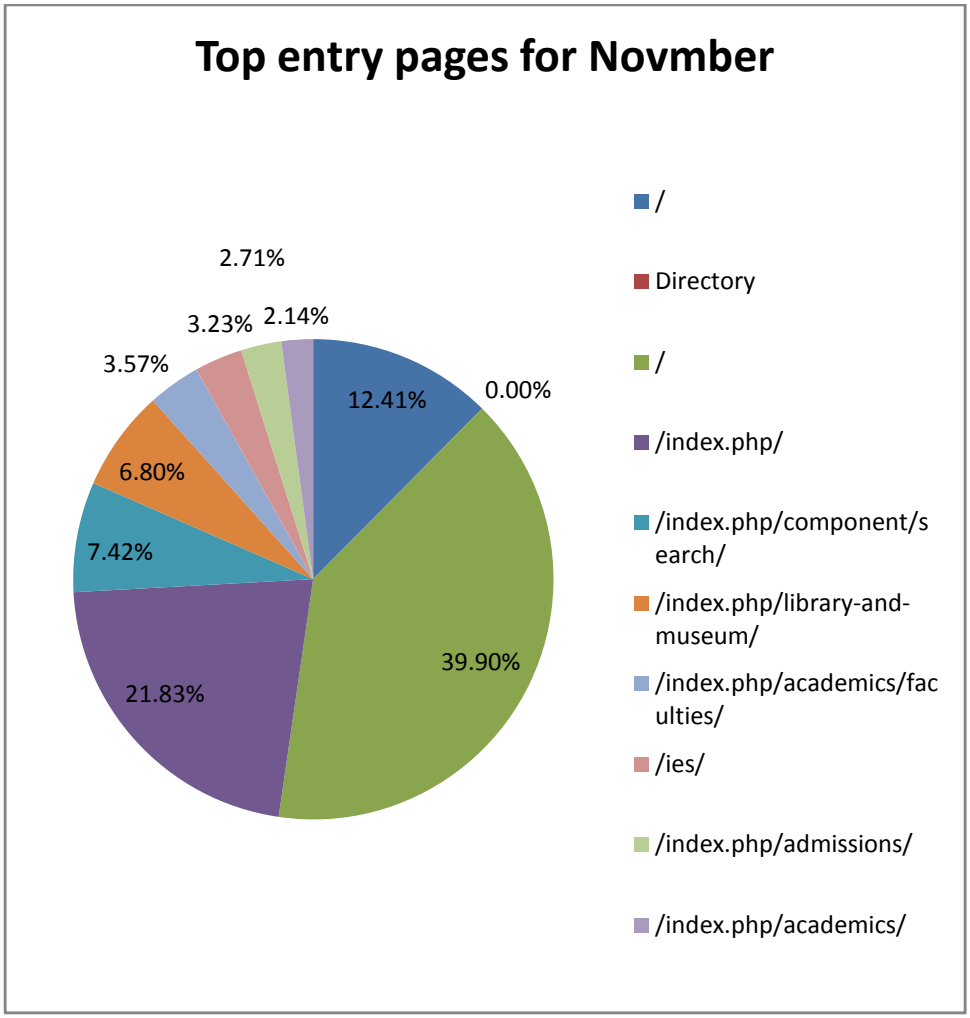
- Most users come to the page web site directly by either typing the name or from the search engine that displays the home page .This could be an indicator the web site has a kind of sickness .the web master therefore should do some kind of assessment on the department index pages make sure that those pages contain those key word for indexing in search pages.
- The most together accessed pages are the home pages is accessed with itself so it is important that ,It also important to recommend that the concerned body that is in charge of AAU official web site design should create quick links from one to other pages for those pages mostly accessed to gather.
- It is also clear that most users left the web site from some pages mainly from the /index.php/home ,/index.php/admission/graduate,/index.php/academics from it, it possible to recommend that the web master should use those page for advertisement and notice and also possible to recommended further it is possible to link to other department links in order to encourage web site users to stay in the web site.
- It is possible to recommend that the web administrators should make the most accessed pages, to be prefetching or cached to prevent the latency of the network bandwidth or prevent delay to access those pages.
- From the navigational behavior 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.
- 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” as a consequence could be better result.
- The other recommendation for further work, divide the web page based up on concept of hierarchy which concept divide pages according to the service they provide, once hierarchal classified the pages it would give better result.

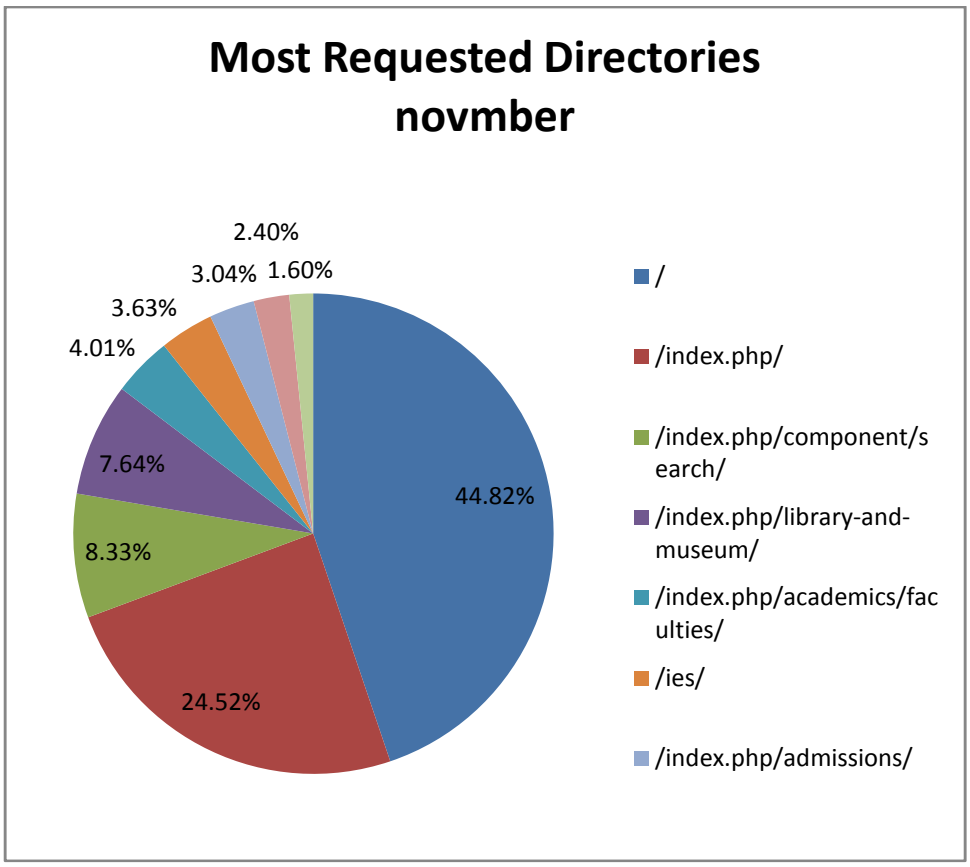
- The last not the least, recommendation for the further work, since by combining different technique of web usage mining such as content mining with web usage mining (work of this thesis) it could give better result in terms of efficiency .

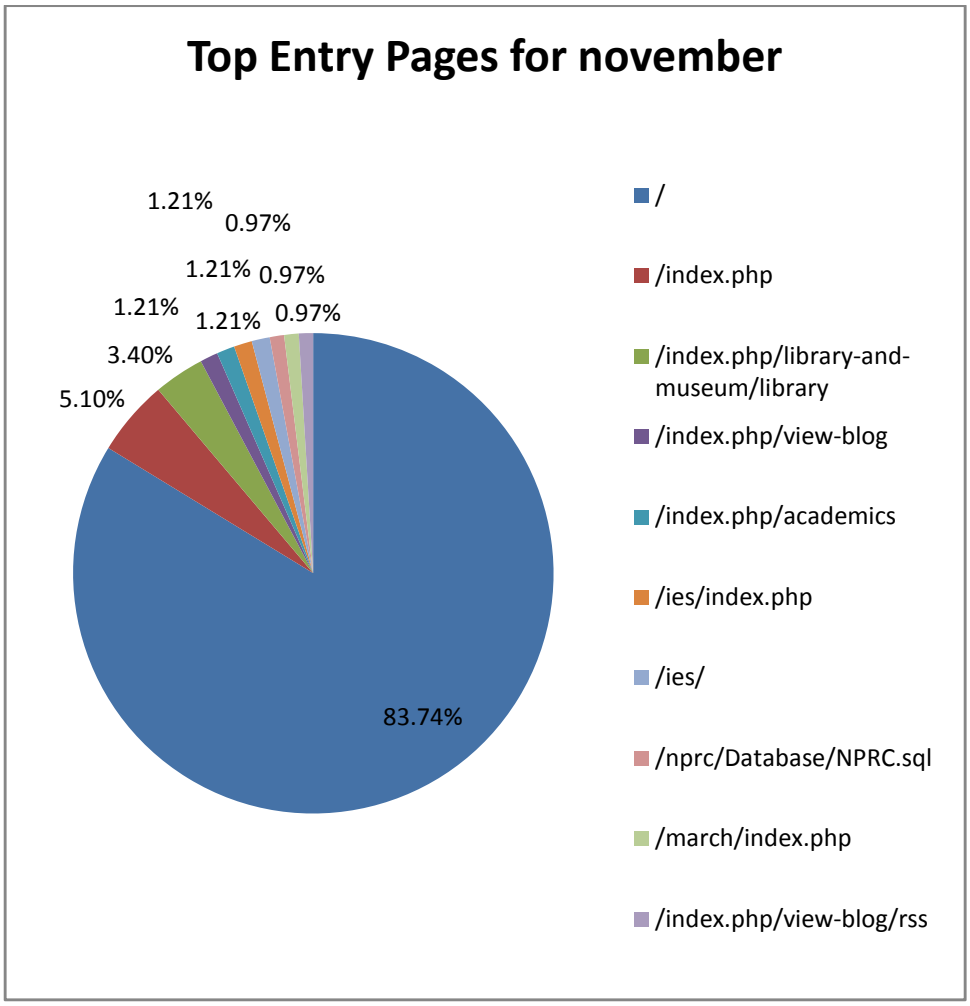
Appendix A: statistical report for the months of November

Appendix for month of November: Most Requested Directories for the months of November




Top entry pages for Novmber
12.41%
0.00%
39.90%
21.83%
7.42%
6.80%
3.57%
3.23%
2.14%
2.71%
/
Directory
/
/index.php/
/index.php/component/search/
/index.php/library-and-museum/
/index.php/academics/faculties/
/ies/
/index.php/admissions/
/index.php/academics/


Most Requested Directories novmber
2.40%
3.04%
1.60%
3.63%
4.01%
7.64%
44.82%
8.33%
24.52%
/
/index.php/
/index.php/component/search/
/index.php/library-and-museum/
/index.php/academics/faculties/
/ies/
/index.php/admissions/



The following are also the sample of one week for the month of November the results of those as explained in chapter 5.

Original log entry records	After removed irrelevant data	After detected robots	After Sessionize	Cleaned data for WUM)*
210240	140127	69564	20004	20004
240067	160743	72087	24060	24060
203406	148906	63480	21000	21000
200967	138967	50653	19734	19734
200190	178300	60752	20674	20674
200150	167543	47897	19653	19653
220205	120950	62096	23765	23765

Appendix B: 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 C: 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 web sites 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 Web Sites 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 Web Sites. 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 web sites. 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 Web Site 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.

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

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.

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE: EXPLORING NAVIGATIONAL BEHAVIOR OF USERS USING GENERALIZED SEQUENCE PATTERN
A CASE ON OFFICIAL WEB SITE OF ADDIS ABABA UNIVERSITY

BY

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

AWET FESSEHA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A TWO STEP APPROACH FOR TIGRIGNA TEXT CATEGORIZATION

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