ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

LOG DATA ANALYSIS TO DISCOVER WEB USER NAVIGATIONAL BEHAVIOR: THE CASE OF ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

BY

AMARE MULATIE DEHNAW

OCTOBER, 2015

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

LOG DATA ANALYSIS TO DISCOVER WEB USER NAVIGATIONAL BEHAVIOR: THE CASE OF ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

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

AMARE MULATIE DEHNAW

OCTOBER, 2015

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

LOG DATA ANALYSIS TO DISCOVER USER NAVIGATIONAL BEHAVIOR: THE CASE OF ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY WEB USERS

BY

AMARE MULATIE DEHNAW

OCTOBER, 2015

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Dr. Tibebe Beshah	Chair Person,	_____	_____
Dr. Tibebe Beshah	Advisor,	_____	_____
Dr.Million Meshesha	Examiner,	_____	_____
Dr. Workshet Lemenew	Examiner,	_____	_____

DEDICATION

This research work is dedicated to my wife Hana.

ABLE OF CONTENTS

ACKNOWLEDGEMENT

First of all, I would like to thanks God almighty who has been giving me everything to accomplish this thesis: patience, health, wisdom.

Next, I would like to speak out my thanks to my advisor, Dr. Tibebe Beshah, for his excellent guidance, thoughtful, patience, and providing me with an excellent atmosphere for doing research. I have been strongly impressed by his constructive comments and guidance.
Next, I would like to speak out my thanks to my examiner, Dr. Million Meshesha, for his excellent guidance, thoughtful, patience, and providing me with an excellent constructive comment to refine my work. I have been strongly impressed by his constructive comments and feedback.

Then, I would like to express my sincere gratitude to staffs of Adama Science and Technology University ICT administrators especially Yosef and Daniel for their cooperation in data collection and data analysis of the research work.

Finally, I would like to extend my thanks to all of my family members. They were always supporting me and encouraging me with their best wishes.

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ACRONYMS

ASTU	Adama Science and Technology University
CLF	Common Log Format
CRM	Customer Relation Management
CSV	Comma Separated Values
ECLS	Extended Common Log Format
ECSU	Ethiopian Civil service University
ERCA	Ethiopian Revenue and customer authority
FPG	Frequent Pattern Growth
FTP	File Transfer Protocol
HTTP	Hyper Text Transfer Protocol
ICT	Information communication technology
ID	Identification
IIS	Internet Information Services
IP	Internet Protocol
KDD	Knowledge Discovery of Databases
MOI	Minister of Industry
MOST	Minister of Science and Technology
NBE	National Bank of Ethiopia
NCTTE	Nazareth College of Technical Teacher Education
NTC	Nazareth Technical College
URL	Uniform Resource Location
VLAN	Virtual Local Area Network
Weka	Waikato Environment for Knowledge Analysis
WUM	Web Usage Mining
WWW	World Wide Web

ABSTRACT

The Web has become an exceptional world-wide repository of knowledge. It contains valuable information for all types of knowledge workers; yet, the Web is dynamic and noisy. As of the popularity of WWW by web users, and due to the alarming rate at which the WWW is growing in both the sheer volume of traffic and the complexity of different websites, this growth of the World Wide Web has led to the development of different client side and server side tools that mine the information resources to extract knowledge. Analyzing this data will help the organizations to realize the lifetime value of their clients, and provide them with a more sophisticated structure of the web site and services. A massive amount of data is gathered by Web servers in the form of Web access logs. This is a rich source of information for understanding Web user surfing behavior. As a result of this, exploring user navigation behavior is expected to redesign the web accessing policy based on user requirement and experience.

Based on the above expression, to realize web users' navigational behaviors of Adama Science and Technology University web server log data is used to conduct the current study to describe web user navigational behaviors by applying web usage mining. Web Usage Mining is the process of applying statistical analysis and data mining techniques to discover interesting usage navigation patterns of web users.

To explore usage patterns of the Adama Science and Technology University web users the researcher adopted hybrid knowledge discovery approach. Such approach consists of steps, such as problem understanding, data understanding, data preparation, mining user behaviors, evaluation and use of the discovered knowledge. The web log data prepared by using log file viewer tool, to clean irrelevant record from the log data, to categorization, and formatting, using datapreparator-1.7 tool preprocessed log record to converted into the form appropriate for pattern discovery tool by using MS- excel statements.

After preprocessing of log file experiments conducted using statistical analysis with datapreparator-1.7 tool and weka 3.7.4 for generating association rule using Apriori and FP Growth algorithm. The result of statistical analysis and data mining techniques shows that social media and entertainment sites are the most frequently accessed once by the web users' of Adama Science and Technology University. The major challenges that involved in this study are preprocessing of log file due to its large, noisy, and complex nature of log record, and identifying rules and patterns that are potentially interesting. Finally recommendations were done for decision makers ASTU ICT workers, and further researchers to improve the website.

CHAPTER ONE

INTRODUCTION

1.1 Background

World Wide Web continues to grow at an astounding rate in both information and users perspective. The scale of information on the internet is growing at a comprehensible rate, similar to the mystifying size of planets and stars [1].

In line the above suggestion internet has become a place where a massive amount of information and data is being generated every day. Every Minute YouTube users upload 48 hours of video, Face book users share 684,748 pieces of content, integral users share 3600 pictures and Tumbler shares 27,778 new posting [1]. Over the last decade with the continued increase in the usage of WWW, Web mining has been established as an important area of research. Web mining is used to analyze users using WWW who leave abundant information in web log, which is structurally complex and incremental in nature [1].

This growth of the World Wide Web has led to the development of different client side and server side tools that mine the information resources to extract knowledge. Analyzing this data will help the organizations to realize the lifetime value of their clients, and provide appropriate web resources based on users interest [2].

The Web has become an unprecedented world-wide repository of knowledge. Unfortunately, to most companies; web is nothing more than a place where transactions take place [2] . They did not realize that as millions of visitors interact daily with Web services around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts [2] .

Web server creates and maintains log files for the purpose of getting feedback about activity, performance of the server and the problems occurring in the web server. Such log files plays very important role in pattern recognition as analyses of log files helps in identifying relationships and patterns between messages request from the user [3] .

Pattern recognition is the task of finding useful information from web server logs applying various techniques such as filtering, grouping. This extracted knowledge plays a very important role in formulation of important rules (decisions) regarding organization website structure, making marketing and advertising more fruitful and effective [3] .

Web usage mining is the process of discovering and interpreting patterns of user access to web information systems by mining the data collected from user interactions with the system. A typical Web usage mining system consists of tier of tracking, in which user interactions are captured and acquired, and a tier of analysis, in which user access patterns are discovered and interpreted by applying typical data-mining techniques to the acquired data [3] .

Those discovered Knowledge of user access patterns is useful in numerous applications like supporting website design decisions such as content and structure justifications [3] , optimizing systems by enhancing caching schemes and load-balancing, making websites adaptive), supporting business intelligence and marketing decisions [3].

As the researcher had shown on the above different literature about the authoritative feature conducting web usage behaviors of users during this era, within astounding growth of websites over the World Wide Web it can open a window of opportunities for organizations to analyze the lifetime value of their customers, and also improve their cross marketing strategies.

As more and more organizations rely on the WWW to conduct business, the traditional strategies and techniques for market analysis needs to be revisited. In today's world, the traditional strategies and techniques of market analyses have been revisited by analyses of a large collection of unstructured data. As of this conducting research on web data is as part and parcel of the current technology movement.

A Log data is a record that records everything that goes in and out of a particular server. Analyzing such data will yield knowledge but pre-processing of that data is required before analyzing it. Once analyzing the Log data, it provides the activities of users over a potentially long period of time.

Log data can be collected from web server, proxy server and Web client. These logs when mined properly provide useful information for decision making. Log file contain information such as username, IP Address, timestamp, bytes transferred, referred URL, User agent [1].

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Based on the different emphasis and different ways to obtain information, web mining can be divided into three major parts: Web Contents Mining, web structure mining and Web Usage Mining [1].

Web content mining is a process of extracting information from texts, images and other contents. On the other hand, Web Structure Mining is a process of extracting information from linkages of web pages. Further, Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site.

Web usage mining refers to the automatic discovery and associated data collected or generated as a result of user interactions with Web resources on one or more Websites. The goal is to capture, model, and analyze the behavioral patterns and the profiles of users interacting with Web sites [4].

The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests [4].

In essence, web usage mining allows the web-based organizations to gather interesting information about the users navigational behavior which can be later used to perform activities such as personalizing the web content, enhancing the system performance, understanding the nature of web traffic, determining effective marketing strategies, identifying potential customers for E-commerce related applications, developing adaptive websites, and to improve the web server performance in terms of content and bandwidth which motive by web users.

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server [1]. Common getaway interface scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access [1].

Usage mining allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information

of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data can also be useful for developing marketing skills that will out-sell the competitors and promote the company's services or product on a higher level [1].

Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the important information from customers visiting the site. This enables an in-depth log to complete analysis of a company's productivity flow. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service [1].

This web mining also enables Web based businesses to provide the best access routes to services or other advertisements. When a company advertises for services provided by other companies, the usage mining data allows for the most effective access paths to these portals [1].

1.2 Overview of Adama Science and Technology University (ASTU)

Adama Science and Technology University (ASTU) were first established in 1993 as Nazareth Technical College (NTC), for offering degree and diploma level education in technology fields [5] . Later, the institution was renamed as Nazareth College of Technical Teacher Education (NCTTE), a self-explanatory label that describes what the institution used to train back then: candidates who would become technical teachers for TVET colleges/Schools across the country.

In 2003, a new addition to NCTTE came about introduction of business education [5] . Nonetheless, the new entries were solely meant for similar purposes: these graduates were also expected to help overcome the existing dearth of educators in vocational institutions [5] .

Although it is an institution with a history of only two decades, ASTU is known for its dynamic past.

It has always been responsive to the realization of national policies: training of technologists at its infant stage, and later shifting to training of technical trainers, as well as business educators, to fill the gap in TVETs.

Following its inauguration in May 2006 as Adama University, the full-fledged university started opening other academic programs in other areas, an extension to its original mission [5] .

However, it was not until it was nominated by the Ministry of Education as Center of Excellence in Technology in 2008 that it opened various programs in applied engineering and technology. For its realization, it became a university modeled after the German paradigm: it not only became the only technical university in the nation, but also the only one led by a German professor.

Notwithstanding closure of some disciplines as per the new vision and mission, the ensuing three years saw flourishing of graduate programs, of which some (like a few in the undergraduate program) were exceptional to our university. The same period saw pioneering of the university in introducing PhD by Research and MA/MSC by Research programs [5] .

Starting from three years before was stratified into faculties, and nowadays, ASTU's reach was limited to its only campus in Adama town which is concerned only with Science and Technology beside Addis Ababa Science and Technology University under Minister of Technology [5]. Those two Universities are concerned only on producing educated manpower in science and technology discipline through selecting best performer students in preparatory all over the country and using entrance exam [5]. Other faculties already moved to a new university, called Arsi University. The university has now extended its reach to Asella, where two of the total seven schools are located. The faculties at the main campus include: School of Business, School of Engineering and Information Technologies, School of Humanities and Law, School of Natural Sciences, and School of Educational Science and Technology Teachers Education. On the other hand, the two schools in Asella are the School of Agriculture and School of Health and Hospital [5] .

In addition to its main concern (academics), ASTU is also host of research institutes and enterprises. In the main campus, apart from the Institute of Continuing and Distance Education (ICDE), there exist two others: the Further Training Institute (better known as FTI) and Adama Institute of Sustainable Energy. The sister town where the two schools are located, Asella, is also host to the Artificial Insemination Institute and Asella model Agricultural Enterprise.

Following its renaming by the Council of Ministers as Adama Science and Technology University in May 2011, the university has started working towards the attainment of becoming a center of excellence in science and technology, thereby allowing for the realization of goals set in the Growth and Transformation Plan (GTP).

To this end, a South Korean has been appointed as President of the University. Currently, ASTU is setting up a Research Park, in collaboration with stakeholders and other concerned bodies: one of a kind in the Ethiopian context. The university is also venturing out to the wider community

and is currently engaged in various joint undertakings [5] . Even though the official web site of ASTU was modified depending of the technology advancement the first time established in 2007 [6].

1.3 Statement Of the Problem

Nowadays higher education institutions to precede learning and teaching process efficiently, the institution highly depended on World Wide Web resources. Due to the alarming rate at which the World Wide Web (WWW) is growing in both the sheer volume of traffic and the complexity of websites every day, it has become very important to analyze this web traffic and the usage of the web sites by the users.

Since Adama Science and Technology University one of higher education institution and the institution to proceed learning teaching progress the University vastly rely on using WWW. As the researcher as try to investigate the current practice ASTU on using the WWW services, even if the institution highly depend on using WWW resources, the institution cannot realize the interest of users. ASTU to describe the user web navigational behaviors try to check through integrating open sources statically analyzer software with their web server to cache users' web navigational behaviors statistics [6] . But that open source statistical analyzer tool for one month try cache the user web navigational behaviors as daily incomplete reports [6] .

Ultimately as the system administrator of ASTU [6] state that the institution as try prohibit different social media and entertainment sites like Facebook and YouTube for one week through applying other organization experience, since Addis Ababa University, Jimma University, Ethio telecom are prohibit different social media like Facebook and YouTube by using firewall through generalizing as all employee try spoil organization working hours by using the entire time for using social media unwisely without any research.

But Adama Science and Technology University ICT release those social media after one week until other empirical study evidence formed to revise the ICT policy [6] .

On the stated literature the researcher had try to show the importance of researching on area of web usage mining as internationally, and the researcher had seen related works in particular within our country as stated in, [7], [8] and [9] web usage pattern discovery the case of Addis Ababa University official web site. Those of researchers, [7], [8] and [9] had conducted web usage mining

in the case of Addis Ababa University official web site to discover interesting usage patterns of the website that could be an input for evaluation of the website in order to optimize. All the above three researchers are focused the optimization and re-designing of a single official web site. But web usage mining had different application in addition to site modification.

So the researcher based on the above researchers suggestion and the current practice of ASTU had motivated to conduct research on Adama Science and Technology University to describe user web navigational behaviors.

In general the researcher had conducted research on the area in particular Adama Science and Technology University web users to overcome two main problems.

The first problem, because of the user interest on web resources were not described, if it is described that could be an input for decision making, for providing the intentional web resources by web users.

The second problem, ASTU is on the way to apply others organization experiences, since several organization like Addis Ababa University, Jimma University, Ethio-telecom are prohibit different social media like Face book and YouTube by using firewall since they generalize as all employee try spoil organization working hours through using the entire time for using social media unwisely without any research.

1.4 Research questions

Through conducting this research the following questions are addressed:

- How to describe user navigational behaviors web resources on Adama Science and Technology University web server log data?
- Which log file attribute and algorithm are appropriate for describing web user navigational behaviors?
- What interesting rule are identified that could be an input for Adama Science and Technology University ICT policies revision?

1.5 Objective of the Study

1.5.1 General Objective

The general objective of this study is to investigate user's web navigational behaviors through conducting log data analysis in Adama Science and Technology University web users.

1.5.2 Specific Objectives

To achieve the abovementioned general objective, the specific objectives of the research are the following:

- To conduct critical literature review for understanding the problem and identify suitable approaches and algorithms in web usage mining.
- To collect and prepare log data of ASTU web users for web usage mining.
- To select appropriate Web usage mining techniques for pattern discovery.
- To identify an imperative attributes of log data for determining user navigational behaviors.
- To explore suitable data mining algorithms for discovering usage pattern.
- To conduct experiment and evaluation on user web navigational behaviors.

1.6 Scope and Limitation of the Study

The scope of this study is limited to explore the usage patterns on web log data of Adama Science and Technology University web users. Due to WWW significantly and rapidly growing in terms of amount of information stored and number of users access the organization needs to realize user interest. But the user interest on web in different organization is not identified. To recognize the web user interest web usage mining important. Since web usage mining allows the web-based organizations to gather interesting information about the users navigational behavior which can be later used to perform activities such as personalizing the web content, enhancing the system performance, understanding the nature of web traffic, identifying potential users, developing adaptive websites, and to improve the web server performance in terms of content and bandwidth which motive by web users. Because of such above problem drown the researcher is inspired conduct log data analysis in Adama Science and Technology University web users to discover web navigational behaviors.

The source of data for this study is Web server log data of Adama Science and Technology University web user. The researcher applied statistical analysis technique and descriptive data mining technique of association rule mining algorithm such as apriori algorithm and FP-growth algorithm to discover web user navigational pattern.

The log data is prepared to make suitable data for Web usage mining using statistical analysis and association rule mining pattern discovery techniques.

Even though the useful information available in log data, the log data suffer so many limitations, creating challenges for use. In this study the limitations of web log data are higher, since the web server records both human request and non-human request. Due to the massiveness of the data the researcher encountered the problem of unable to open using any preprocessing tool to preprocess. To handle such problem the researcher consume much amount of time to get and use appropriate prepressing tools. Some of the data that are logged are incomplete, such as visit duration, referrer site, user ID, accessed URL. Due to this, the preprocessing task of this research is challenging, because the web log contain noisy, huge and complex records that needed to clean. The other challenging task, transforming the data set of statistical analysis experiment into association rule mining technique experiment by crosschecking the relationship of URL by using accessed time.

1.7 Significance of the study

The web servers in the form of server or access logs generally automatically store this information. Different organization can make use of this data by analyzing for respective purposes.

The analysis of the server logs can provide with information on how to better structure the web site for effective use and in benefit of the organization. For organizations working with intranet technologies, such analysis can lead to important information about managing the workgroup communication and the infrastructure of the organization.

For advertising organizations, such analysis can help them in targeting a specific customer group after analyzing the user access patterns.

The Web Usage Mining involves determining the frequency of the page access by the clients and then finding the common traversal paths of the users.

Some of the significance of this study

Since web usage mining prepares groundwork for developing policies for Web caching, load balancing, data distribution, or intrusion, fraud, and break-ins detection the system administrator can design for of Adama Science and Technology University ICT police based on user need.

Since user satisfaction is an important criterion for Web sites. For instance, it is crucial to e-commerce. So the of Adama Science and Technology University site designer can design based on detailed feedback for site user and the site can be designed right to the user's needs.

Web usage mining focuses on techniques that could predict user behavior while the user behaviors are predicted the system administrator can provide the appropriate web service for the web user based on their need.

Mining of web usage patterns can help in the study of how browsers are used and the user's interaction with a browser interface. Based on this system administrator able to plan the appreciate training required for web user.

1.8 Organization of the Thesis

This thesis is organized into six chapters. The first chapter provides general introduction about the study such as background, statement of the problem, objective, scope, and methodology of the study high lightly were discussed.

In the second chapter, literature on data mining and web mining particularly process, tools, and techniques of Web usage mining and its application that are necessary to understand the methods and terms that are introduced in the study are reviewed. Finally, a few researches which address problems similar to this study were reviewed.

In the third chapter methods algorithms and hybrid methodology were discussed. Chapter four is about data preprocessing activities such as data cleaning, categorization and data conversion process of the study.

Chapter five presents experimentation part of this research with different steps that are incorporated in the adopted methodology were described. The web usage patterns were discovered by integrating statistical analysis and association rule mining. Interesting rules and patterns of the experiments 'results identified. In the six chapters, concluding remarks and recommendations were made. Finally, lists of references and appendices were presented at the end of this paper.

CHAPTER TWO

LITERATURE REVIEW

2.1 Data Mining

Presently, the amount of data stored in databases is increasing at a tremendous speed. This gives rise to a need for new techniques and tools to aid humans in automatically and intelligently analyzing huge data sets to gather useful information. This growing need gives birth to a new research field called Knowledge Discovery in Databases (KDD) or Data Mining, which has

attracted attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization [10] .

According to [10], Data Mining and Knowledge Discovery in Databases (KDD) is defined as the process of automatic extraction of implicit, novel, useful, and understandable patterns in large databases. There are many steps involved in the data mining process, which include data cleaning and preprocessing, data integration, data selection, data transformation and reduction, data-mining task and algorithm selection, and lastly post processing and interpretation of discovered knowledge.

Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [10] .

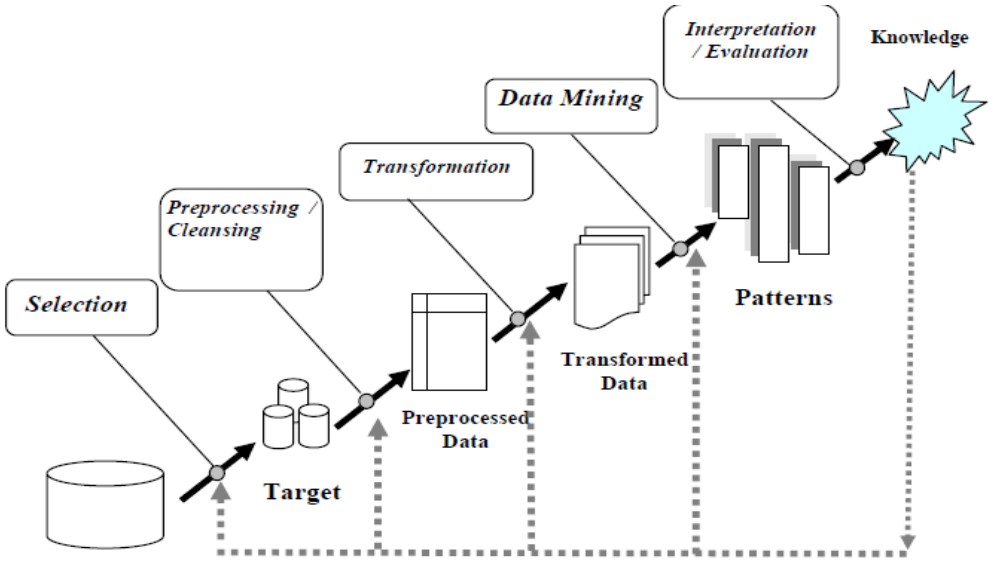

Figure 2.1 Process of knowledge discovery [10]

It is an interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Data Mining has various application areas such as banking, education, and e-commerce.

On the other hand, the new data mining applications such as World Wide Web, spatial data, multimedia data.

World Wide Web is one of the largest and most widely known data source. Today, WWW contains billions of documents edited by millions of people. The total size of the whole documents can be

interpreted in many terabytes. World Wide Web is growing at a very large rate in size of the traffic, the amount of the documents and the complexity of websites. Because, different organizations, individuals or societies provide their public information such as news, markets, company advertisements. The World Wide Web serves to a broad diversity of user communities through web. Due to this, the demand for extracting valuable information from this huge amount of data source is increasing every day. This leads to new area called Web Mining [10] , which is the application of data mining techniques to World Wide Web.

2.2 Web Mining

One of the applications areas of data mining is World Wide Web (WWW), which serves as a huge, widely distributed, global information service center for every kind of information such as news, advertisements, consumer information, financial management, education, government, e-commerce, health services, and many other information services. With the rapid growth of the WWW, it becomes more important to find the useful information from these huge amounts of data. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing sources for data mining. The Web poses great challenges for effective knowledge discovery and data mining applications [11].

According to [11], web mining is defined as "the discovery and analysis of useful information from the World Wide Web "Such data can be the content presented to users of the web sites such as hypertext markup language (HTML) files, images, text, audio or video. Also the psychical structure of the web sites or the server logs that keep track of user accesses to the resources mentioned above can be targets of web mining techniques.

According to [12], Web mining is a step in the Knowledge Discovery from Web process and it aims to analyze data and discover knowledge from the Web. The Web data include all kinds of Web documents, hyperlinks among Web pages, and Web usage logs.

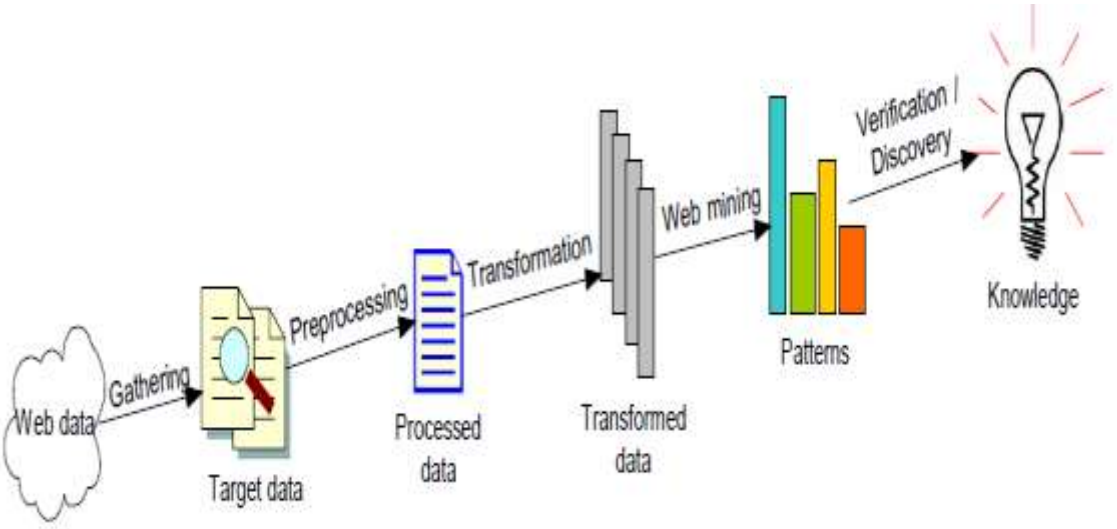

Figure 2.2 Knowledge Discovery Process from Web [12]

Web mining as a sub category of data mining is fairly recent compared to other areas since the introduction of internet and its widespread usage itself is also recent. However, the incentive to mine the data available on the internet is quite strong. Both the number of users around the world accessing online data and the volume of the data itself motivate the stakeholders of the web sites to consider analyzing the data and user behavior.

2.2.1 Taxonomy of Web Mining

According to [13], explanation web mining is mainly categorized into three subsets namely web content mining, web structure mining and web usage mining. While the content mining approaches focus on the content of single web pages, web usage mining uses server logs that detail the past accesses to the web site data made available to public. Usually the physical structure of the web site itself which is a graph representation of all web pages in the web site is used as a part of either method.

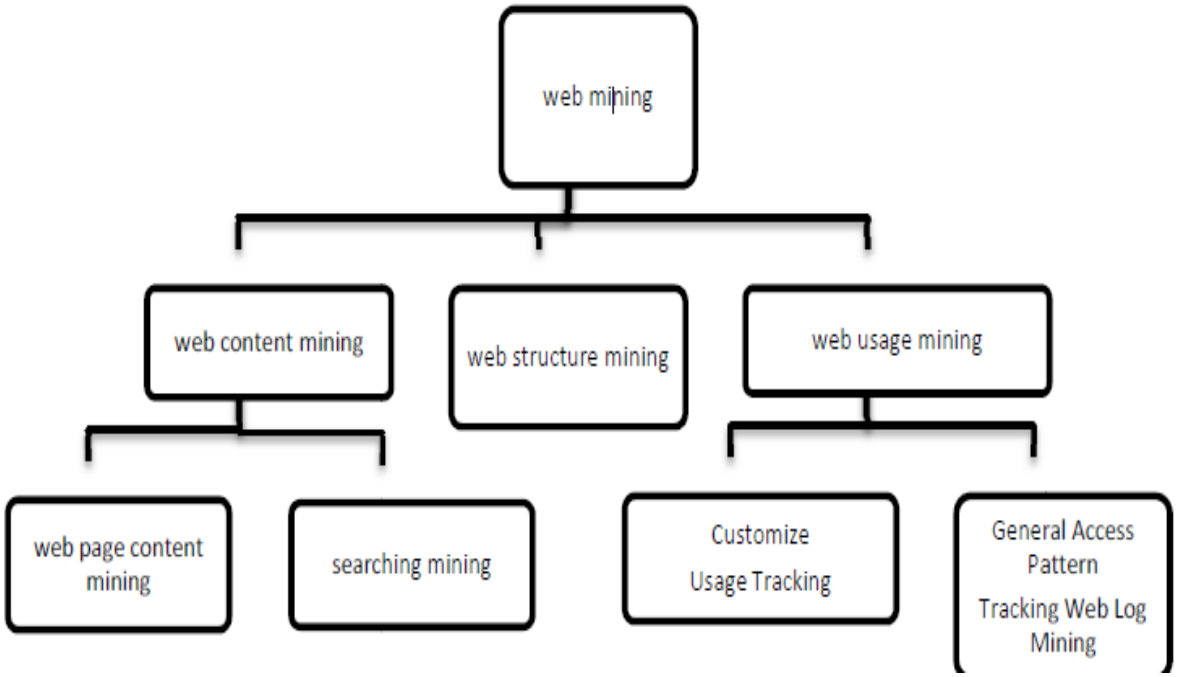

Figure 2.3 Taxonomy of web mining [13]

2.2.1.1 Web Content Mining

Bayir [13], states that " web content mining describes the automatic search of information resources available on-line." The focus is on the content of web pages themselves. Categorizes content mining as agent-based approaches; where intelligent web agents such as crawlers autonomously crawl the web and classify data and database approaches; where information retrieval tasks are employed to store web data in databases where data mining process can take place.

As noted by Yilmaz [14] , most web content mining studies have focused on textual and graphical data since the early years of internet mostly featured textual or graphical information. Recent studies started to focus on visual and aural data such as sound and video content too.

2.2.1.2 Web Structure Mining

According to [13] , one of the most well-known algorithms, Page Rank measure and hubs and authorities are based on the links between pages. Web structure mining focuses on the links rather than the content of the pages, their usage or semantics.

2.2.1.3 Web Usage Mining

The main topic of this thesis is the web usage mining. Usage mining as the name implies focus on how the users of websites interact with web site, the web pages visited, the order of visit, timestamps of visits and durations of them.

While Web content and structure mining utilize real or primary data on the Web, Web usage mining works on the secondary data such as Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, and bookmark data. Web usage mining refers to the application of data mining techniques to discover usage patterns from these secondary data, in order to understand and better serve the needs of Web-based applications. The usage data collected at different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, and single-site browsing behavior to multi-user, multi-site access patterns [15] .

According to [16] , Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data.

The main source of data for the web usage mining is the server logs which log each visit to each web page with possibly IP, referrer, time, browser and accessed page link. Although many areas and applications can be cited where usage mining is useful, it can be said the main idea behind web usage mining is to let users of a web site to use it with ease efficiently, predict and recommend parts of the web site to user based on their and previous user's actions on the web site.

According to [17] , Web usage mining is defined as the study of "user interaction with the web". It can also be understood as "usage patterns discovered from web data by using data mining techniques". To obtain usage patterns of users of our site of interest, we must have an appropriate data set. For that purpose, we require web logs of the website.

A typical example of web log record is as follows: 192.168.10.11 -- frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326.

The fields of the above HTTP log record are as follows: Host -- client IP address - (192.168.10.11), User id – person requesting the web document – (frank), Request time -- the time at which request was made. – (10/Oct /2000 13:55:36 -700), Request Type – the type of client request made - (GET), Resource name – name of the resource which was requested - (/apache_pb.gif), Protocol – the protocol used for the request - (HTTP/1.0), Status code – status code of the web document - (200) and Size -- size of the web document - (2236).

2.3 Source of Web Log Data

According to [18], expression any type of Web usage mining requires having an accurate picture of the WWW traffic. So this section explores the available data sources and their properties. As shown in Figure2.4 below the data sets commonly used for Web usage mining are collected from three target area such as server-level, proxy-level or client-level. Each data source differs in terms of format, accuracy, scope and method of implementation.

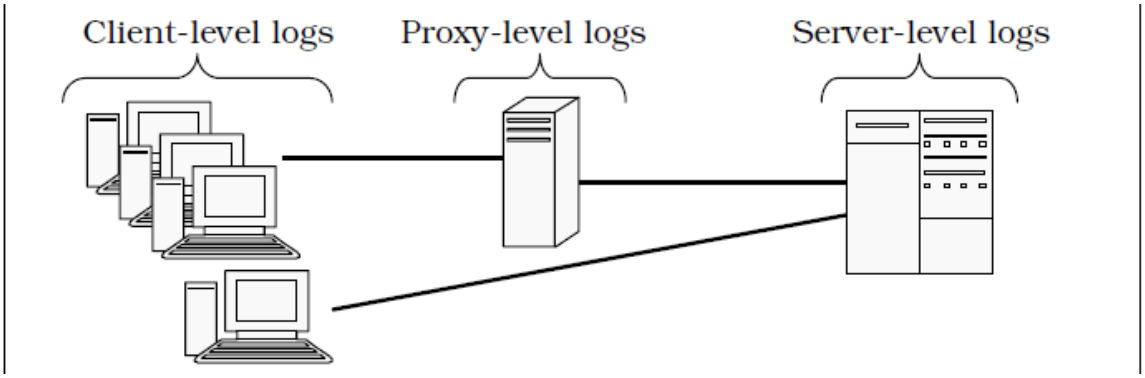

Figure 2.4 log data storage area [18]

Client Side (client level logs): Most of the users have tendency to open several pages simultaneously and in between, use some non-browsing applications such as MS-word, Excel etc. for their own personal work, in such cases data recorded in server log only shows the requested time of the web pages and cannot help us to find out which web page and for how long has been really browsed on client machine. Usage data can be tracked on the client side by using JavaScript, java applets, or even modified browsers.

These techniques avoid the problems of user's session's identification and the problems caused by caching (like the use of the back button). However, these approaches rely heavily on the users' cooperation and rise many issues concerning the privacy laws, which are quite strict.

Proxy Side (proxy level logs): A Web proxy acts as an intermediate level of caching between client browsers and Web servers. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of group of users accessing huge groups of web servers.

Server Side (server level logs):- Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information such as host Internet Protocol (IP) address, remote login name of the user, authentication user name, date, request Uniform Resource Locator (URL), status code, bytes size, referrer URL and user agent, and time spent on each requested page.

2.3.1 Types of Web log File

Web server log files comprise access logs, referrer logs, agent logs and error logs [18] . Access Logs provides the bulk of the Web server data, including the date, time, users IP address, and user action. The following is some of the information that can be obtained from an access log: The IP (Internet Protocol) address of the computer making the request for a document the time stamp (user access date and time) the user's request (e.g., html document or image requested, or data posted).

Referrer Logs provide information on what Web pages, from both the site itself and other sites, contain links to documents stored on the server. The log provides information such as the URLs of sites and pages on sites that referred visitors to a particular page. For example, users may often arrive at a particular Website through a search engine, and the referring search engine along with the keywords used in the originating query, can be obtained from the Referrer log.

Agent Logs: supply data on the browser, browser version, and operating system of the user who accessing the web.

Error Logs contain information on specific events such as file not found, document contains no data; the time, user domain name, and the page on which a user received the error is recorded, providing a server administrator with information on problematic and erroneous links on the server.

2.3.1.1 Types of Web log file formats

According to [18] , there are three kinds of log file formats to record log files. They are: Common Log Format (CLF), Extended Common Log Format (ECLF), and Microsoft IIS (Internet Information Services).

Common Log Format: This log format is supported by a variety of web server applications and includes the following seven fields: Remote host field, Identification field, Auto user field, Date/time field HTTP request, Status code field and Transfer volume field.

Extended Common Log Format, ECLF is a variation of the common log format, formed by appending two additional fields onto the end of the record, the referrer field, and the user agent field. ECLF Log File Format have two additional fields they are referrer and user agent.

Microsoft IIS Log files Format: The IIS format records more fields than the other formats. The Microsoft IIS log format includes the following fields: Client IP address, User name, Date, Time, Service and instance, Server name, Server IP, Elapsed time, Client bytes sent, Server bytes sent, Service status code, Windows status code, Request type, Target of operation, and Parameters.

2.4 Web Usage Mining Process

According to [19] , there are three stages of the Web mining process, which follows the general data mining process. As the authors expression those web mining process are collection and pre-processing, pattern discovery, and pattern analysis.

From the above web mining process collection and pre-processing is the first step involving not only the collection of suitable target data, like access logs and server logs, but also the cleaning and partitioning of these raw data [19] . Although this is the most difficult and time consuming stage in the process, there is no doubt that the result in this step is very critical to the success of the application. It is the crucial precondition, and the final result greatly relies on this task.

Again from the above web mining process according to [19] , expression Pattern discovery is in the pattern discovery stage, data mining, machine learning, and statistical operations are performed to obtain hidden patterns that reflect the typical behavior of users. The users are automatically segmented and classified based on their similar behavior, and then the adaptive user model is developed.

This type of model represents a collection of personal data associated with specific users, such as preferences, interests, and skills. In this research, descriptive statistics were used to describe the features of the logs and user groups; association rules were employed to discover the frequency and patterns of the users; and auto-classification categorized the users even without logged information.

Again from the above web mining process according to [19] , expression Pattern analysis is in the last stage of the process, the discovered patterns and statistics are further processed and filtered in order to meet the different representation requirements.

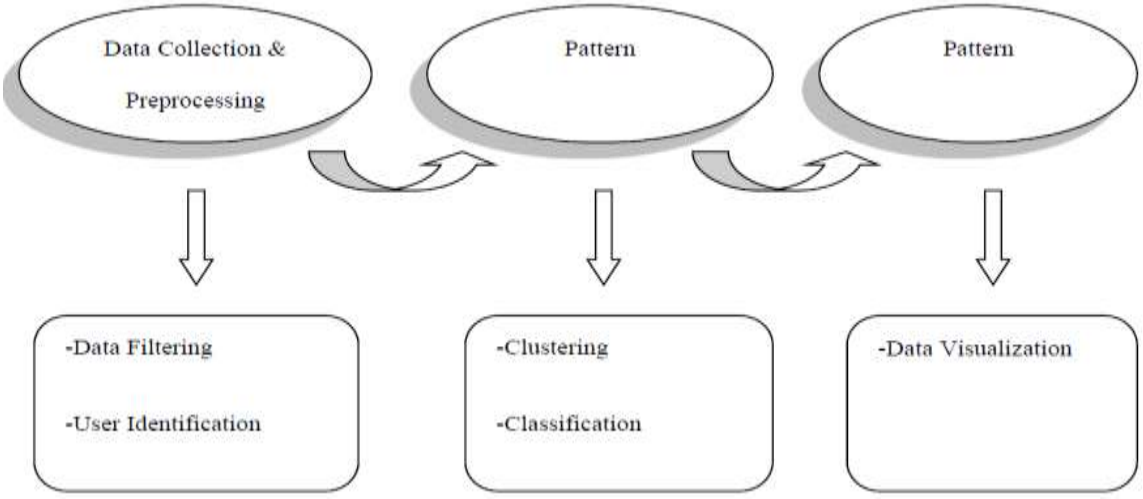

Figure 2.5 Web Mining Processes [19]

Sharma [20], conducted Pattern Discovery and Pattern Analysis on the Rit Web Data, through conducting experiment using WUM (Web Utilization Miner) tool, WUM which is basically a software tool that preprocesses log file data and then performs the analysis. The main purpose of using this tool is to obtain a statistical report of the access log in the HTML format. The visitor sessions are created with a threshold of 30 minutes.

This helps us in getting rid of the unnecessary data. After creating the visitor sessions, the observations can be made on the log data. A comprehensive HTML report can be created that consists of a list of observations and results from the input log file.

This report consists of information like page accesses (page views), average page accesses per day, visitor sessions, average visitor session length, unique visitors, most requested pages, most requested directories, least requested pages, least requested directories, top visitors, most active top-level domains, top entry pages, top exit pages, top referrer pages, top referrer sites, most used browsers and single access pages.

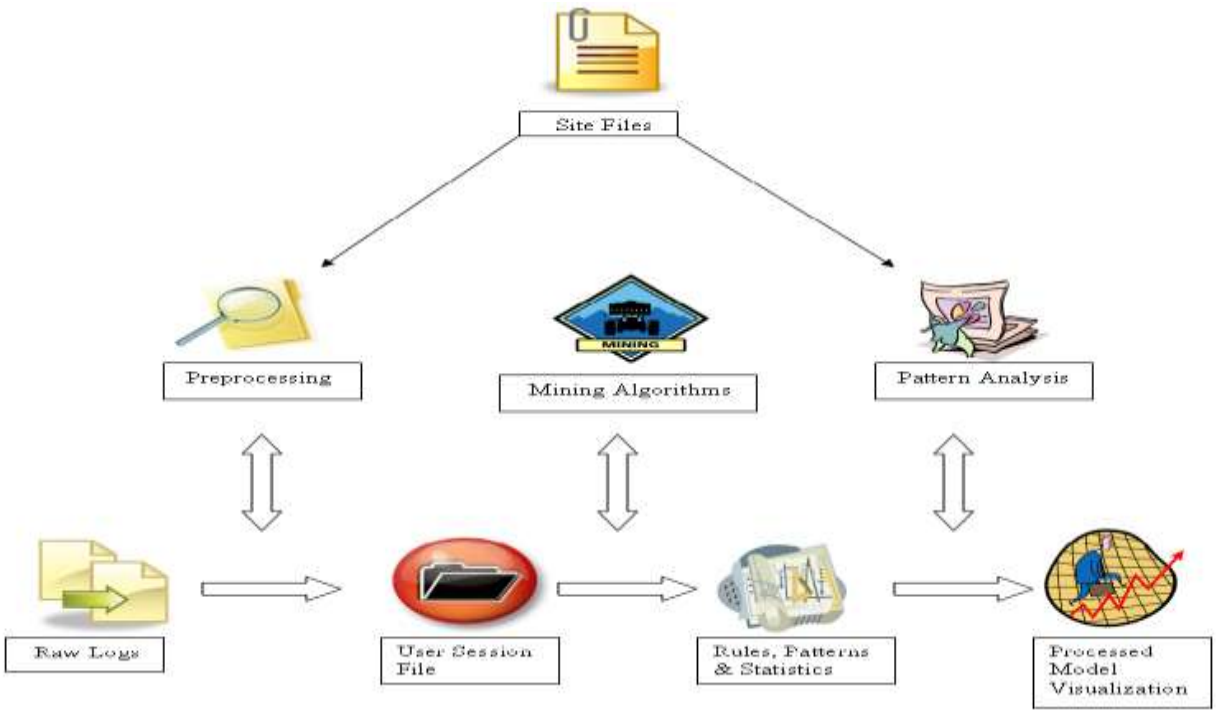

Figure 2.6 Web Usage Mining processes [20]

As explained in [20] , the most important task of the Web Usage Mining process is data preparation. This process is diagrammatically represented in above Figure 2.6.

[20], Had introduced various data preprocessing techniques that help in the data cleaning process and make it ready for effective and correct results. The section below describes these pre-processing tasks in detail.

2.5 Preprocessing

Preprocessing is preliminary and essential step in web usage mining. Because, a web usage data (Web log file) is generally diverse, incomplete, inconsistent, noisy and difficult to be used directly for pattern discovery [20].

A Web log file also have different format from the database or data warehouse data which has a good data structured.

All this has made the work of pretreatment face many technical problems. Due to this, data preprocessing has become the most difficult task in the Web usage mining. This preprocessing method is used to process the actual web logs before the real usage mining process.

2.5.1 Web Log Data Preprocessing Steps

Web log data pre-processing is a complex process. It can take up to 80% of the time in knowledge discovery process [20] . The aim of data preprocessing is to select essential features, to clean irrelevant records and finally transform raw data into sessions. To achieve its goal Web log data preprocessing have the following steps: data cleaning, user identification, and session identification, path completion and transaction identification [20] .

2.5.1.1 Data Cleaning

The purpose of data cleaning is to remove irrelevant items stored in the log files that may not be useful for analysis purposes. When a user accesses a HTML document, the embedded images, if any, are also automatically downloaded and stored in the server log. For example, log entries with file name suffixes such as gif, jpeg, GIF, JPEG, jpg and JPG can be removed. Since the main objective of data preprocessing is to obtain only the usage data, file requests that the user did not explicitly request needs to be eliminated. This can be done by checking the suffix of the URL name.

In addition to this, erroneous files can be removed by checking the status of the request (such as status code 404). The cleaned log represents the user's accesses to the Website.

2.5.1.2 User Identification

Identification of users who access a website is an important step in web usage mining.

The simplest method is to assign different user ID to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. If the IP address of a user is same as previous entry and user agent is different than the user is assumed as a new user.
If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address.

2.5.1.3 Session Identification

User session is a delimited set of web pages visited by a particular user in single visit to the website. Identification of user sessions has also received significant attention as it reveals the navigational behavior of users, which forms the foundation of personalization system. A user may have a single or multiple sessions during a particular time period. Various heuristic methods have been used for identifying user sessions. [18], divides these methods into time-based and context-based. In time-based approach, a page viewing time is defined, and a single user session consists of all those web pages, which are requested by a particular user within the page viewing time. Context-based approach is not very strict, and depends on the users' perspective.
A web page, which is a navigational page (that contain primarily hyperlinks to other web pages and are used just for browsing purpose) for one user might be a content page (that contain the actual information of user's interest) for the other.

2.5.1.4 Path Completion

Determining the missing important web page access due to the proxy server and the browser is essential for mining information. This is accomplished by path identification process. If the requested page is not linked to the previous accessed page by the unique user, then from which page request came is identified using the referrer log file. If the page is available in the user's history, then it is assumed that the user pressed back button. Hence each and every session reflects the complete path, including the web pages that have been backtracked.
At the end of path completion the user session file gives the paths consisting of a group of page references including repeated page accesses made by a user. At the end of this stage the user session file is ready for transaction identification process.

2.5.1.5 Transaction Identification

In order to group individual Web page references into meaningful transactions for the discovery of patterns such as association rules, an underlying model of the user's browsing behavior is needed [20]. As a result, further preprocessing step is required, which is transaction identification. Transaction identification used to prepare data in the format appropriate for the specific data mining algorithm to be used. According to [20], for this process two kinds of transaction are identified, travel path transactions and content only transactions. The travel path is a combination of auxiliary and content pages accessed by a user.

Auxiliary pages are pages that used to facilitate the browsing of a user while searching for information. The content only transactions are only content pages which are user interest.

User session in a user session file transaction identified thought either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. Therefore, the task of identifying transactions is one of either merging small transactions into larger transaction or dividing a large transaction into multiple smaller ones or in order to create transactions [16]. Transaction identification approach defined as ether a merge or a divide approach. Merge approach is based on identify transaction consisting of all of the page reference for a given user session.

The divide approach such as transaction identification by reference length is depending upon the time taken a user spends on a page correlates to whether the page should be classified as auxiliary or content pages for that user.

2.5.1.6 Data Fusion

Data fusion basically involves joining the log files from different servers and sources. During this process, it is important to add the name of the server and the time clock to all the files belonging to the respective servers. All of these files are then anonym zed for privacy reasons. In this study, no need have to worry about this step because the access logs belong to only one server [18].

2.6 Pattern Discovery

When data preprocessing is completed, the next phase of web usage mining is pattern discovery. There are several methods that can be used to discover patterns, and these methods are rooted from fields such as data mining, statistics, and machine learning [20] .

2.6.1 Statistical Analysis

Statistical techniques are the most common method to extract useful information about visitors to the web resources [20] . Statistical analysis can be performed on the session file variables such as page views, viewing time and length of navigational paths.

The output of applying statistical methods could be determining the most frequently accessed pages, average viewing time of a page, average length of navigation paths to a specific page, or the most common invalid URI.

Despite its lack of depth, the output of statistical analysis can occasionally help in reorganizing web content, making better marketing decisions and enhancing system performance, facilitating the site modification task and provide support for marketing decision. There different statistical analysis tools that helps to discover descriptive pattern from the web log file, such as Web log storming, Weblog expert, google analytics and AWStats [21] .

Data-preparator-1.7 tool which were assist the researcher with exploring and preparing data in various ways prior to data analysis or data mining. It includes operators for cleaning, discretization, numeration, scaling, attribute selection, missing values, outliers, statistics, visualization, balancing, sampling, row selection, and several other tasks [22].

Data-preparator-1.7 is written in Java and requires Java Runtime Environment (JRE) to be installed on individual machine.

Datapreparator-1.7 discovers detailed website statistics with interactive graphs and reports. Very complete detailed log analysis of activity from every visitor to the website is only a mouse-click away. In addition to this the other statistical analysis tool can process limited dataset but data-preparator-1.7 tool can process massive dataset. Due to this strength, Data-preparator-1.7 tool selected for statistical pattern discovery of web usage mining in this study [22] .

Web Log Storming: is an interactive, desktop-based Web Log Analyzer for Windows (www.weblogstorminig.com). It's easy to track sessions, hits, page views, downloads, or whatever metric is most important to each user. Website behavior, from the top entry and exit pages, to the paths that users follow, can be analyzed. We can learn which countries and cities visitors came from, and which operating systems and browsers they use [23] .

Web Log Expert: is a powerful access log analyzer. It can help to reveal important statistics about website usage: activity of visitors, access statistics, and paths through the site, visitors' browsers, and much more (www.weblogexpert.com). The HTML reports of web log expert include multiple graphical charts to show the number of visitors, and what pages they viewed after arriving [23].

Google Analytics: It is a free utility provided by Google which helps in keeping a track of unique visitors (www.googleanalytics.com). Google Analytics allows to dig down deep into stats to see breakdowns of individual regions, states/provinces, cities and numerous other items to better identify your site visitors [23].

AWStats is a free powerful tool that generates advanced web usage statistics (www.awstats.org). This tool works as a CGI Script or from command line. It displays all sorts of information that the log contains. It uses partial file information to be able to process large log files [23].

2.6.2 Data mining techniques

Data mining tasks can be classified into two categories: descriptive data mining and predictive data mining [10] . Predictive data mining like classification used to constructs one, or a set of, models, performs inference on the available set of data, and attempts to predict the behavior of new data sets.

Descriptive data mining such as clustering, association rule mining, used to describes the data set in a concise and summary manner and presents interesting general properties of the data. Clustering is a technique to group users who exhibit similar browsing patterns, or web pages which exhibit similar contents [11]. According to [11] Web usage mining allows the overlapping of clusters. Clustering is a method of gathering items that have similar characteristics.

In the context of web mining, we can have two distinct cases, user clusters and page clusters. User clustering identifies groups of users that seem to have similar behaviors when browsing through a

website. Page clustering results in groups of pages that are apparently related to each other in terms of user's perception. Such clustering information is then used for personalizing a website.

2.6.2.1 Association Rule Mining

Association is the discovery of association relationships or correlations among a set of items [24] . They are often expressed in the rule form showing attribute-value conditions that occur frequently together in a given set of data.

In the context of Web usage mining, association rule used to discover associations between Web pages based on their co-occurrence in user sessions [11] . It refers to set of pages that are accessed together with a support value more than some specified threshold.

Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests.

2.6.2.1.1 Association rule mining algorithm

There are number of association rule discovery algorithm, however the most widely algorithm for pattern discovery are Apriori and Frequent Pattern Growth (FP-Growth) [24] .

Apriori Algorithm

Apriori algorithm is an efficient algorithm, which is one of the best available methods for discovering frequent patterns [24]. Apriori runs breadth-first search algorithm and uses a hash tree structure to count candidate items at each step. Apriori algorithm searches for large itemsets during its initial database pass and use its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent itemsets and those below are called small itemsets. The algorithm is based on the large itemset property which states. Any subset of a large itemset is large and any subset of frequent item set must be frequent.

The three main steps in Apriori algorithm [24]:

- Generate candidates of length k from the frequent k-1 length item sets, by a self-join on Fk-1.
- Prune any candidate with at least one infrequent subset.
- Scan all transactions to obtain candidate supports

FP-Growth Algorithm

The FP-growth algorithm is frequent items set mining uses the FP-tree structure to attain a divide-and conquer to break down the mining problem into a set of smaller problems [24] . It retains the item set association information and compressed databases are divided into a set of conditional databases, each one associated with a frequent item. It takes the help of prefix tree representation of the given database of transactions (called FP tree), which saves considerable amount of memory for storing the transactions. Every node in the tree represents one item and each path represents the set of transactions that involve with the particular item.

All nodes referring to the same item are linked together in a list, so that all the transactions that containing the same item can be easily found and counted. Performance of FP-growth compared with Apriori in large databases [24] . Frequent patterns from web log data using FP for finding the most frequently access pattern generated. Apriori for association rule mining. Finding frequent sets using candidate set generation. Apriori sets many candidates so the cost is high compared to FP and its getting slow. FP requires less memory, low cost, very fast than Apriori.

The steps involved in FP-Growth Algorithm are [24]:

- ✓ Scan the database once to find frequent 1-itemset (single item pattern)
- ✓ Sort the frequent items of step 1 in frequency descending order, f-list
- ✓ Database again to select only the frequent items from each of the transaction items and construct the FP-tree for the selected items.

2.7 Pattern Analysis

Pattern analysis is about undergoing further interpretation of the discovered patterns before applying the discovered rules to useful application [11] .

It is used to extract the interesting rules, patterns or statistics from the output of the pattern discovery process by filtering the irrelative rules or statistics. Another aim of this analysis is to

obtain some information can offer valuable insights about users‟ navigational behavior. For example we can understand the number of users that started from a page and proceeded through some certain pages and finally visited their goal page. Also, we can obtain some information about page popularity or some pages that contain the most information for a visitor.

2.8 Applications of Web Usage Mining

The general goal of web usage mining is to gather interesting information about users navigation patterns (i.e., to characterize web users). This information can be exploited later to improve the website from the users' viewpoint. The results produced by the mining of web logs can use for various purposes [25] :

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

Personalization of web content

Web usage mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users [25], 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. [26], 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 in the Design of web sites

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

According to [25] , exploits web usage mining techniques to suggest proper modifications to website. Adaptive websites represents a further step in this case; the content and the structure of the website can be dynamically reorganized according to the data mined from the users' behavior.

E-commerce

Mining business intelligence from web usage data is dramatically important for ecommerce web-based companies according to [27]. 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.

Improvement in System Performance

Performance of web services is an important issue for user satisfaction. Web usage mining is an important research area for detecting web traffic behavior, which can be used to develop new policies for increasing the web server performance.

Web caching, load balancing, network transmission or data distribution are the common application areas of web mining for performance improvement [13] .

2.9 Related works

Different aspects of web usage mining have been addressed by various intellectuals in international context and to some extent in national level during the past few years. Each of the researchers had attempted to explore various aspects of the web mining endeavor that ranges from developing a web mining architecture to application of data mining techniques for web mining. From various work in international context on the area of web usage mining the researcher had reviewed some of them that of which are important this research work.

And also the researcher had seen some national level work on the area of web usage mining. Some of those stated blow.

Mekonnen [7], conducted a research on web usage pattern discovery on Addis Ababa University official website. He used combination of statistical analysis and data mining approaches for pattern discovery.

Mekonnen used a Weka plug in called WUMPrep4 to clean the raw log and develop session file and transform the session file to a format that is compatible to Apriori algorithm of Weka. Regarding usage patterns pattern of the research, association mining, i.e. discovery of common pages of the site that are accessed together, was accomplished for the prepared transaction file. Besides, he used Mach5 statistical analyzer tool to discover possible statistical reports of the web log record. Finally, he recommended to use combination of statistical analysis and data mining based pattern discovery for discovery of effective usage patterns of websites.

Tadele [8] , attempts web usage pattern discovery on Addis Ababa University official website. He follow the web usage mining process that is suggested by Srivastava [11]. He also used python code for data preparation, Mach5 statistical analyzer and Weka tool for association rule and sequences mining with Apriori algorithm. His finding shows the daily access trend, top entry and exit page and page that display error response of the website.

He also discovers the correlation between pages that accessed together. His recommendation is mainly the need for reconstruction of Addis Ababa University official website in user friendly manner.

Awet [9], also conducted a research on exploring the navigational behavior of users of Addis Ababa University official website. He used the same web usage mining process as Tadele [8] , used.

He used WUMprep tool for data preparation, Web Utilization Miner for statistical analysis and pattern discovery. . His finding shows most requested pages, top entry and exit pages, referrer page and pages that accessed frequently after the home page of the website. He recommended to go for combining web usage mining such as content mining with web usage mining for efficiency of exploring user's behavior on the website.

Getahun [28], conducted a research on Web usage pattern discovery and analysis by region taking the case of Ethiopian Airline official website.

He follows the Web usage mining process model suggested by Sharma [20] . He used IANA IP assignment dataset to divide the server log dataset into different region based on IP address. He

used WUMprep tool, Java programming, and MS Excel for data preparation, Google Analytics for statistical analysis and Weka tool for association rule mining with FP Growth algorithm.

His finding shows the other page of the website except home page needs optimization and more promotion is required to make the website more accessible with respect to referrers. He also show the navigational behavior of user across region have similarity and difference. Finally, he recommended for feature researcher to include proxy server and client server logs data and proxies and client cookies method of user identification method for better usage pattern discovery.

Among international work [29] , they were conducted study on Analysis of Web Server Log Files to Increase the Effectiveness of the Website Using Web Mining the study were focused on analyzing the web server log files of an Educational Institution's Website that is www.davkota.orgto discover the usage behavior of the Website users the authors they use experimental research method using web log expert tool. The min finding on the researchers work they have analyzed the log files of Web server using smart sync software (www.smsync.com) with the help of weblog Expert tool.

For the experiment seven days duration log files which magnitude 187.55MB data after preprocessing from total of hits 4114. From this experiment the researchers had obtained results that shall definitely help to the Website Maintainers, Website Analysts, Website Designers and Developers to manage their System by determining occurred errors, corrupted and broken links. This work will also increase the effectiveness of the Website.

Wei [30], conducts research on exploring health website users on Clarian Health website in India by web mining. His objective is to examine the navigation behavior of different user groups and to make some suggestions to reconstruct a website for more customized Web service. Wei, used access weblog files from one local health provider's website. He used WUM-prep, a Perl-based tool supporting data preparation for mining Web server log files. Web Utilization Miner (WUM), to discover navigation patterns over the aggregated view of the web log. Rapid Miner, to generate and compare classifiers of naive Bayesian and Support Vector Machine (SVM). His findings show that users are not searching health information as much as was thought.

The top two health topics which patients are concerned about are children's health and occupational health. Patients and doctors have different search strategies when looking for information on the website. Wei recommended, redesigning and improving the website by adding more intuitive portal and more customized links for both user groups.

Anand [31], conducted a research on data mining of Web access log on computer science website of Royal Melbourne Institute of Technology University.

He used data mining techniques to find access patterns hidden inside huge volumes of web access data. His objective is to investigate the access patterns between visitors from within Australia and visitors from outside Australia, visitors from within Royal Melbourne Institute of Technology University and visitors from outside Royal Melbourne Institute of Technology University. Anand used the data mining techniques such as classification, association rules, with three pattern discovery process such as transaction identification and feature extraction, discovery of the access patterns, analysis of the discovered patterns.

He identified long transaction of access log manually. His finding shows, visitors from Australia generally visit the root page while visitors from outside Australia do not. However, some Visitors from outside Australia visit the root page and pages about post graduate programs (such as Master of Technology). Finally, he recommend for further researcher, that data mining techniques with better preprocessing like cleaning Web robots through the web browser information , to improve the discovered interesting patterns.

2.10 Summary of related work

Authors (year)	Objective	Methods/approaches/ techniques	Key findings	Remark

Mekonnen (2009)	To discover interesting usage patterns of the website that could be an input for evaluation of the website.	➢ combination of statistical analysis ➢ data mining approaches ➢ WUMPrep4 were used to clean the raw log ➢ Mach5 statistical analyzer tool were used for statistical analysis	recommended using combination of statistical analysis and data mining based pattern discovery for discovery of effective usage patterns of websites	➢ To explore user navigational behaviors on ASTU web server user. ➢ The frequent accessed site were social media site like Facebook, you tube.
Asitatikie Tadele (2011)	To discover interesting usage patterns of the website that could be an input for evaluation of the website.	➢ Mach5 statistical analyzer and Weka tool	Recommendation is mainly the need for reconstruction of Addis Ababa University official website in user friendly manner.	➢ Datapreparator-1.7 tool for statistical analysis and Weka 3.7.4 for association rule discovery are used. ➢ Web log data of this research preprocessed by using datapreparator-1.7 tool, log file viewer tool and MS- Excel 2013. ➢ During preprocessing step the data categorized into six categories. ➢ Again the experiment were done by six categories to explorer user interest by clustered time interval.
Fesseha Awet (2011)	To discover interesting usage patterns of the website that could be an input for evaluation of the website.	➢ WUMprep tool for data preparation ➢ Web Utilization Miner for statistical analysis and pattern discovery	Finding shows most requested pages, top entry and exit pages, referrer page and pages	
Getahun Negatu (2014)	To discover interesting usage patterns of the website that could be an input for evaluation of the website.	➢ WUMprep tool, Java programming, and MS Excel for data preparation, ➢ Google Analytics for statistical analysis ➢ Weka tool for association rule mining	Recommended for feature researcher to include proxy server and client server logs data and proxies and client cookies method of user identification method for better usage pattern discovery.	

Wei Kong(2003)	To exploring health website users on Clarian Health website in India by web mining.	➢ WUM-prep, a Perl-based tool supporting data preparation ➢ Web Utilization Miner (WUM), to discover navigation patterns over the aggregated view of the web log.	Recommended, redesigning and improving the website by adding more intuitive portal and more customized links for both user groups.	

Table 2.1 Summary of related work

The researchers who are conducted research on the area of web usage mining in Addis Ababa University official web site mainly focused on site modification objective and the methodology they have used is the same; that is why the results obtained by the three researchers converge together.

This study is totally different from the previous researchers work on area of web usage mining since the research objective, scope, and approach used are different.

CHAPTER THREE

METHODS AND ALGORITHMS

3.1 Overview

In this study statistical and association rule techniques are applied to explore user navigational behaviors on Adama Science and technology university web server users. An attempt was made to describe web usage access patterns using datapreparator-1.7 tool for statistical analysis user navigational behaviors and weka 3.7.4 for generating association rule using Apriori and FP Growth algorithm.

3.2 General approach

Research methodology used to understand what research methods are going to be used, for the choice of the study design and a strategy of data collection, organization and analysis. Different literature reviewed from books, journal articles and the internet to identify Web usage mining research techniques, processes and tools.

The general approach used in this research is hybrid knowledge discovery methodology. Since hybrid knowledge discovery methodology can comprise other methodology and in each step contain an opportunity to return back if challenge happened in some where step and fix the gab during progress of representing the user's navigation pattern using recorded server web log data.

In the web usage mining process, the techniques of statistical analysis and data mining are applied so as to discover the tendencies and the patterns in the browsing behaviors on Adama Science and Technology University web server users. There is extraction of the navigation patterns as the browsing patterns could be traced. When it is talked about the browsing nature of the user it deals with frequent access of the web resources. This information extracted from the log data.

Figure 3.1 below shows the steps of hybrid knowledge discovery methodology followed in this research.

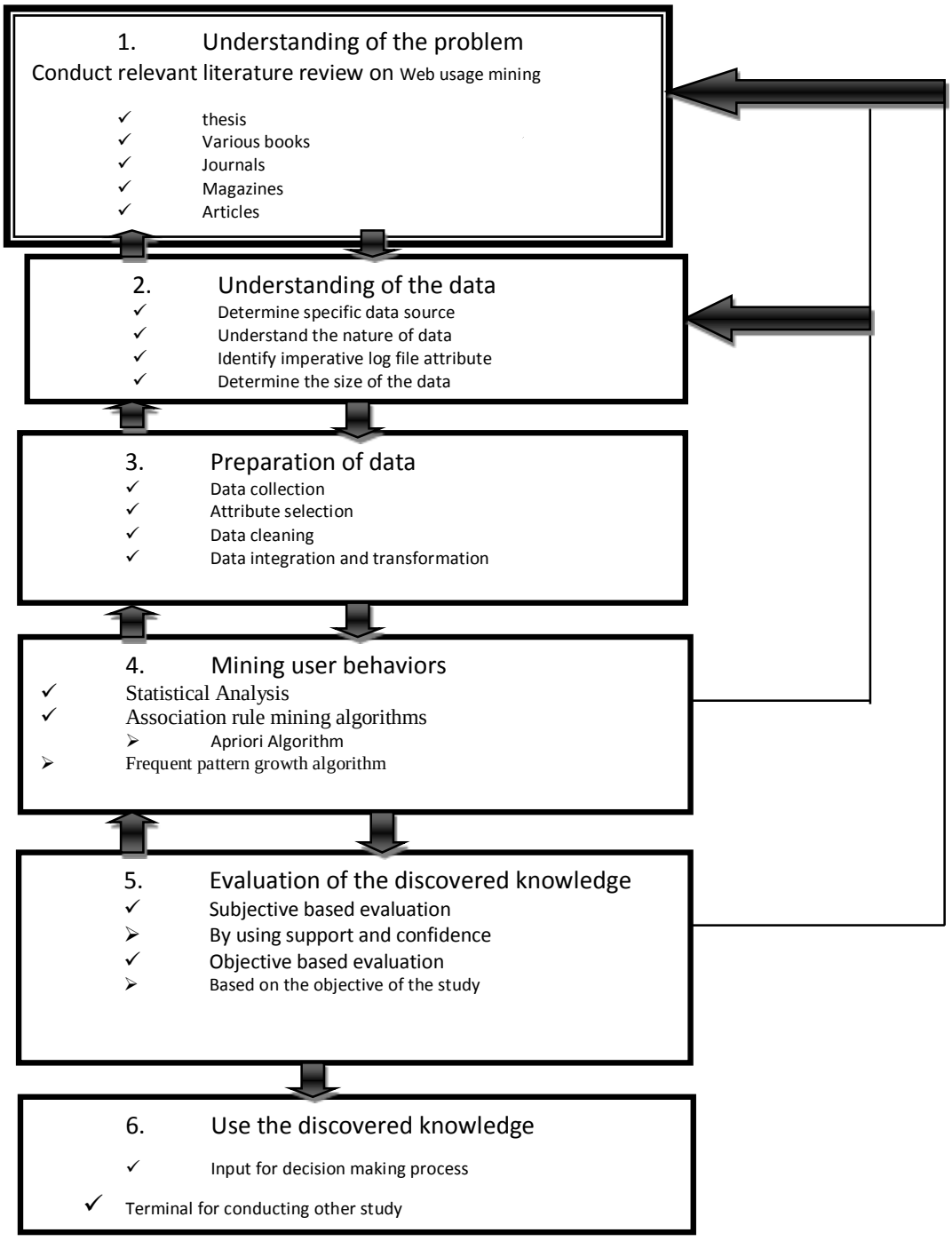

Figure 3.1: Major Steps of Web Usage Mining Process using hybrid knowledge discovery methodology [32].

3.2.1 Understanding of the problem

The researcher had reviewed different relevant literature that have been conducted to assess data mining technology and web usage analysis. Various books, journals, magazines, articles and papers pertaining to these subject areas have been consulted to understand the potential applicability of web usage mining in evaluation of web usage navigational behaviors, particularly on Adama Science and technology University web server users. In addition to this the researcher had communicate domain experts to grasp or understand about the problem.

3.2.2 Understanding of the data

According to [27] , Web server log is an important source for performing Web usage mining because it explicitly records the browsing behavior of site visitors. The server records the time and date of the transaction, the name of the file that was sent and how big that file was, internet address to which the file was sent, if the user goes to a page by clicking a link on some other page, the server records the address of the page with that link, and also records some details about how the file was sent and any errors that may have occurred as well as information about the browser that the user is using. Based on the above researcher explanation the data that recorded in the server logs reflects the (possibly concurrent) access of a Web site by multiple users.

3.2.2.1 Data Collection and acquisition

Data collection is the initial step in web log preprocessing. The user interaction details with the website are recorded in the form of web logs in three different places, (i) Web Server, (ii) Proxy server and (iii) Browser machine. From those three places of log data to get reliable dataset the researcher were select server level log data for this research. In general the important point is that the server side data is an aggregate picture of the usage of a service by all users, while on the client side data is complete picture of usage of all services by a particular client, with the proxy side middle frame.

Data gathered from web servers is placed into special files called logs and can be used for web usage mining. Usually this data is called web log data as all visitors activities are logged into this file [33] .

In this research, the source of data set is web server log data. The dataset source of this study is Adama Science and Technology University web users log data.

The access log data contains records of each user requests that contains remote host IP address (identify who has visited the website), date (access time of the website), visiting path (path taken by the user while visiting the website), path traversed (the path taken by the user within the website), success rate (status code returned by the server), URL (the page that accessed by the user), request type (GET or POST) and user agent (browser type that user use to send request).apart from this derived attribute are added in preprocessing chapter.

The size coverage of the dataset the researcher used for the experiment one month dataset which are recorded from April 07, 2015 to May 06, 2015web users log data. The reason why the researcher used one month dataset to represent the user web navigational behaviors of Adama Science and Technology University web user, because of the massiveness of the data size it is difficult to use and prepare extra dataset. Before deciding the size of the dataset the researcher collected the one year log data from Adama Science and Technology University through collecting from the old and new server since ASTU ICT migrate from the old sever to new one. In addition to the above reason why the researcher used only 30 days dataset and the way how select the specific month the through checking the whole dataset and the researcher had get the data set not complete due to the migration data from old server to new one some day's log data is missed from each month except one month.

To approve the sufficiency data in terms size and time appropriateness to represent one year web users navigational the researcher had communicate the domain experts and the domain experts Adama Science and Technology University said that the magnitude of the log data each month are almost same and also the network traffic when the observe on server almost it show the same user navigating. Based on the above idea the researcher decide the selected size of data enough to represent one year user web navigational behaviors.

As [34], put across that web usage mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more web servers the following information obtained through web usage mining.

Number of Visitors: It is the count of users who navigates to your website and browses one or more pages on your site.

Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

Visitor Referral Website: The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

Number of Hits: This number usually signifies the number of times any resource is accessed in a Website.

Time and Duration: This information in the server logs give the time and duration for how long the Website was accessed by a particular user.

Visitor IP address: This information gives the Internet Protocol (I.P.) address. It is the address of the visitors who visited the website.

Browser Type: This information provides the data of the kind of browser that was used for accessing the web site.

Cookies: A message given to an online browser by an online server. The browser stores the message during a document known as cookie. The message is then sent back to the server whenever the browser requests a page from the server.

Platform: This information provides the kind of operating system. That was accustomed access the web site.

3.2.3 Preparation of data

Kumar and Aggarwal [34], state that one of the important core steps of knowledge discovery is data preprocessing. The main goal of this step is to create minable objects for knowledge discovery despite the presence of ambiguities and incompleteness in data. This step is highly data-source dependent.

The techniques used to overcome these shortcomings may vary greatly from one data source to the other. Therefore, handle such shortcoming the researcher focus on s server-level Web access log files. The access log file is not suitable for direct input to the pattern discovery tool that was special for researcher work. Researcher had performed process to convert the raw data into the data abstraction necessary for the further applying the data mining algorithm.

Since this information needs to be integrated to form a complete data set for data mining. However, before the integration of the data, Web log files need to be cleaned, using techniques like filtering the raw data to eliminate outliers and irrelevant items, grouping individual page accesses into semantic units.

Data preparation (or data preprocessing) in this context means manipulation of data into a form suitable for further analysis and processing.

It is a process that involves many different tasks and which cannot be fully automated. Many of the data preparation activities are routine, tedious, and time consuming. It has been estimated that data preparation accounts much of the time spent on a data mining research.

3.2.3.1 Tools for Log data preparation

Datapreparator-1.7 tool which were assist the researcher with exploring and preparing data in various ways prior to data analysis or data mining. It includes operators for cleaning, discretization, numeration, scaling, attribute selection, missing values, outliers, statistics, visualization, balancing, sampling, row selection, and several other tasks. Datapreparator-1.7 is written in Java and requires Java Runtime Environment (JRE) to be installed on individual machine.

The Web server usually registers all users', spiders and bots access activities of the website as Web server logs. The data present in the log file cannot be used as it is for the mining process. Therefore, data preprocessing is mainly one phase in Web usage mining. Generally, several preprocessing tasks need to be done before performing web mining algorithms on the Web server logs [35] . Data processing includes task such as data collection, data cleaning, user and session identification, attribute selection, transaction identification, and data conversion.

3.2.3.2 Data Cleaning

The data cleaning process involves removing the irrelevant data from the collected log data. The data can be in the form of requests from a non-analyzed source, data with missing attributes or the attributes that are not needed for the research goal.

This step helps in reducing the size of the data to a great extent. This reduction in size also helps in removing any false associations that could have been created because of this data. When a request to a web page is made, there are various attribute loaded in that request.

This includes the image files and graphics that are loaded with the web page because of the HTML tags. Since the researcher is interested only in the data that is requested by the user and not any system generated data, the researcher need to make sure that only the user requested data is present in the server logs.

Therefore, any of the system-generated data should be avoided and removed from the log files. The entries in the log files with the suffix jpg, jpeg, JPEG, class files, .ico files, style sheets and gif files can be removed as these entries do not contribute to the interest of this research.

3.2.3.3 Attribute selection

After dataset cleaned the next step is to identify the best attributes which is important for association algorithm. For data mining techniques after dataset is cleaned from the access log data appropriate attributes selected based on the objective of the study and statistical analysis experiment result.

3.2.3.4 Data categorization

The total of 1126,427 dataset is categorized under 6 different experimental phase (morning office hour web user dataset, lunch time web user dataset, afternoon office hour web user dataset, non-office hour web user dataset, non-weekend web user dataset, and weekend web user dataset).Then, transaction is identified with 34 items (page viewed) selected from varsities of site, which is selected by consultation of domain experts Adama Science and Technology University ICT workers.

3.2.3.5 Data formatting

This process is approached after all of the data cleaning and preprocessing tasks have been performed on the access logs. In this process, the preprocessed data is basically formatted according to the needs of the respective data mining algorithms, which are applied to extract important information from the preprocessed data. The formatting of data differs from the kind of algorithms that are used.

3.2.4 Mining user behaviors

In this study statistical analysis and data mining experiments conducted to discover navigational behaviors of Adama Science and Technology University web server users. To accomplish experiment Data-preparator-1.7 tool for statistical analysis and Weka 3.7.4 for association rule discovery are used.

After data preparations have been completed, the next step is pattern discovery. The pattern discovery phase consists of different techniques derived from various fields such as statistics analysis and association rule discovery are used [11] .

3.2.4.1 Statistical Analysis

Statistical analysis techniques are the most common method to extract descriptive knowledge about user navigational behaviors of web resources.

In this study Datapreparator-1.7 tool used as a tool to analyze and describe user navigational behaviors of Adama Science and Technology University web server users.

Data-preparator-1.7 tool were applied on the filtered log file to extract facts and figures that show general statistics on the site usage. Data-preparator-1.7 had been selected to analyze the log records statistically [22]. This tool is selected as they are recommended by researchers and the reports are found very important and helpful.

3.2.4.2 Association Rule Mining

Association rule mining finds interesting associations rules for a large set of data items. In this study association rules can be used to find correlations between web pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests.

Beside the above tool Weka 3.7.4 is used for the discovery of usage patterns. It is a freely available data mining tool which is widely used in data mining researches and projects. The tool has preprocessing feature and it incorporates clustering, association and classification techniques along with sets of algorithms.

The most common file format for Weka is Attribute-Relation File Format (ARFF) files. Indeed, other file formats such as CSV data files, C45 names files, and Binary serialized instances could be used as input to the tool. ARFF files have three sections: relation, attributes and data.

The researcher selected Weka data mining tool on consideration of the following facts. Firstly, the popular algorithms for association mining such as Apriori, predictive apriori and FP Growth and clustering algorithm are implemented in Weka. Secondly, the tool is available for free. Thirdly, the researcher is well experienced in using of the tool.

Given a set of transactions where each transaction is a set of items (item set), an association rule implies the form $X \Rightarrow Y$, where X and Y are item sets; X and Y are called the body and the head, respectively.

A rule can be evaluated by two measures, called confidence and support. Support for the association rule $X \Rightarrow Y$ is the percentage of transactions that contain both item set X and Y among all transactions. The confidence for the rule $X \Rightarrow Y$ is the percentage of transactions that contain an item set Y among the transactions that contain an item set X. Support (usefulness) can be measured with a minimum support threshold (minsup) [34].

Confidence (certainty) can be measured with a minimum threshold for confidence (minconf). Association rule mining is the task of finding all rules with support S and confidence C such that $S \geq$ minsup and $C \geq$ minconf, where minsup is support threshold and minconf is the confidence threshold [34] .

The rule $X \Rightarrow Y$ holds in the transaction set D with support S. where S is percentage of transaction in D that contain X U Y(i.e. the union of items sets X and Y.or both X and Y).

This is taken to the probability, $P(X \cup Y) = \frac{\#\text{of transaction with the item set X U Y}}{\#\text{ of total transaction}}$

Support show the probability that all the predicates in X and Y fulfill together.

$$\text{Support } (X \Rightarrow Y) = \frac{\#\text{ of tuble containing both X and Y}}{\text{Total number of tuble}}$$

The rule $X \Rightarrow Y$ has the confidence C in the transaction set D where C is the percentage of the transaction in D containing X that also contain Y. this is taken to be the conditional probability.

$$P(X / Y) = \frac{\#\text{ of transaction with the item set X U Y}}{\#\text{ of total transaction with item set X}}$$

Confidence measure how often predicates Y fulfill if predicate X get fulfilled.

$$\text{Confidence } (X / Y) = \frac{\#\text{ of tuples containing both X and Y}}{\#\text{ of tuples containing X}}$$

According to [34] , association rule mining can be viewed as a two-step process. First, find all frequent item sets, by definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, min sup.
Then generate strong association rules from the frequent item sets, by definition, these rules must satisfy minimum support and minimum confidence.

In the context of web usage mining, item sets are sets of web pages accessed and association rule mining is used to discover the set of web pages accessed together in a user session.

Given a set of web pages accessed by the user, other frequently co-occurred pages may be recommended to the user. The most popular algorithm used to discover association rules are Apriori and FP Growth algorithm [34] .

3.2.4.2.1 Apriori Algorithm

Apriori algorithm [34], is an efficient algorithm, which is one of the best available methods for discovering frequent patterns. Apriori is most widely used when working on databases containing transactions. The algorithm serves as a basis for many other pattern discovery methods. Apriori runs breadth-first search algorithm and uses a hash tree structure to count candidate items at each step. Its search is complete and bottom up with a horizontal layout and discovers all frequent item sets.

Apriori is an iterative algorithm that counts item sets of a specific length at each step while going over the database. The initial tasks of the method are scanning all records in the database and finding the first frequent item sets. After this step, using these frequent items, it forms potential frequent candidate 2-itemsets. An additional pass for scanning all transactions in the database is performed for determining supports of these patterns. By observing supports, the infrequent ones are eliminated from candidate 2-itemsets, while the remaining ones will form frequent 2-itemsets. This process is repeated until all frequent item sets have been discovered [34].

There are three main steps in this algorithm [34]:

1. Generate candidates of length k from the frequent k-1 length item sets, by a self-join on Fk-1.

2. Prune any candidate with at least one infrequent subset.

3. Scan all transactions to obtain candidate supports.

Apriori Algorithm Pseudo code

The next gives Apriori algorithm for finding all frequent item sets. The first pass of the algorithm simply counts item occurrences to determine the large 1-itemsets.

A subsequent pass, say pass k, consists of two phases. First, the large item sets Lk-1 found in the (k-1) the pass are used to generate the candidate item sets Ck (Set of candidate item set of size k), using the Apriori candidate generation function (apriori-gen) described below.

Next, the database is scanned and the support of candidates in Ck is counted. For fast counting, an efficient determination if the candidates in Ck that are contained in a given transaction t is needed [34].

According to a discussion in [34] , Apriori algorithm shown below.

```
L1 = {large  1- itmesets};
for (k= 2;Lk-1 ? Ø; k++) do
      Ck= apriori – gen(Lk-1);
      forall transactions  t ? D do
         Ct = subset (Ck, t);
              forall candidates c ? Ct do
                   c.count++;
               end
       Lk = {c ? Ck || c.count = minsup}
 end
return Uk    Lk;
```

3.2.4.2.2 FP-Growth Algorithm

The FP-growth algorithm: mining frequent patterns without candidate generation [36].It compress a large database into a compact Frequent-Pattern tree (FP-tree) structure.

FP-Tree frequent pattern mining is used in the development of association rule mining. FP-Tree algorithm overcomes the problem found in Apriori algorithm. By avoiding the candidate generation process and less passes over the database, FP-Tree was found to be faster than the Apriori algorithm [34]. It adopts a divide and conquer strategy. Firstly it compresses the database representing frequent items into a frequent pattern tree or FP-tree. It retains the item set association information and compressed databases are divided into a set of conditional databases, each one associated with a frequent item.

It takes the help of prefix tree representation of the given database of transactions (called FP tree), which saves considerable amount of memory for storing the transactions.

An FP-Tree is a prefix tree for transactions. Every node in the tree represents one item and each path represents the set of transactions that involve with the particular item.

All nodes referring to the same item are linked together in a list, so that all the transactions that containing the same item can be easily found and counted. Large databases are compressed into compact FP tree structure. FP tree structure stores necessary information about frequent item sets in a database [36].

A frequent-pattern tree (or FP-tree in short) is a tree structure defined below [36]. It consists of one root labeled as "null", a set of item-prefix subtrees as the children of the root, and a frequent-item-header table. Each node in the item-prefix sub-tree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none. Each entry in the frequent-item-header table consists of two fields, (1) item-name and (2) head of node-link (a pointer pointing to the first node in the FP-tree carrying the item-name).

Based on this definition, we have the following FP-tree construction algorithm [36] .

FP-tree construction

- ✓ Input: A transaction database DB and a minimum support threshold
- ✓ Output: FP-tree, the frequent-pattern tree of DB.
- ✓ Method: The FP-tree is constructed as follows.

Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as F- List, the list of frequent items.

Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in DB do the following. Select the frequent items in Trans and sort them according to the order of F-List. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree ([p | P], T).

The function insert tree ([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N's count by 1; else create a new node N, with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree (P, N) recursively .

let P be the single prefix-path part of Tree and Q be the multipath part with the top branching node replaced by a null root, the following algorithm show that mining of frequent patterns with FP-tree by pattern fragment [36],

```
if Tree contains a single prefix path then
        for each combination (donated as B) of the nodes in the path P do;
            generate pattern B _ a with support = minimum support of nodes in B;
            let freq_pattern_set(P) be the set of patterns so generated;
            else let Q be Tree;
                for each item ai in Q Do{
                    generate pattern B= ai_a with support= ai.support;
                    (construct B's conditional pattern-base and then B's conditional FP-tree
TreeB;
                   if TreeB= Type equation here.
                    then call FP-growth(TreeB,B);
                    let freq_pattern_set(Q) be the set of patterns so generated;}
return(freq_pattern_set(P  freq_pattern_set(Q) _ (freq_pattern_set(P* freq_pattern_set(Q)))
```

Frequent Pattern Growth (FPG) Algorithm

FPG algorithm avoids the problem of _candidate generation and test 'which is common in Apriori algorithm. This algorithm follows depth-first search in which different set of combinations with a given single or pair of items. The approach has the philosophy of growing long patterns from short ones using local frequent items only. Thus, FPG algorithm preserves complete information for frequent pattern mining. Moreover, irrelevant information is reduced in advance and compact structure.

Given a transaction database along with a minimum support and a minimum confidence threshold values, the following three steps are involved in FP-tree construction [36] .

- ✓ Scan the database once to find frequent 1-itemset (single item pattern)
- ✓ Sort the frequent items of step 1 in frequency descending order, f-list
- ✓ Scan database again to select only the frequent items from each of the transaction items and construct the FP-tree for the selected items.

Once the FP-tree is constructed, conditional pattern-base for each of the frequent 1- item sets will be extracted from which conditional FP-tree will be identified. Then, interesting rules is developed on consideration of the minimum threshold values and the conditional FP-tree.

3.2.5 Evaluation the discovered knowledge

The discovered usage patterns is evaluated based on research desired objective expectation and through crosschecking the statistical analysis techniques experiment result and association rule mining technique experiment results.

3.2.6 Use the discovered knowledge

Based on the evaluation the discovered knowledge, the researcher as empirical study recommend that interesting rule to be applicable for decision making process for ASTU.

CHAPTER FOUR

DATA PREPARATION

4.1 Overview of Data Preparation

The access log data used in data mining operations is not suitable for direct input to the pattern analysis and pattern discovery tools that was special for researcher work. Researcher had performed process to convert the raw data into the data abstraction necessary for the further applying the data mining algorithm. Since this information needs to be integrated to form a complete data set for data mining. However, before the integration of the data, Web log files need to be cleaned, using techniques like filtering the raw data to eliminate outliers and irrelevant items, grouping individual page accesses into semantic units.

The aims of the preprocessing step in Web usage mining process are approximately to convert the unprocessed log file into a set of transactions (one transaction being the list of pages visited by one user) and to remove noisy requests (e.g. hidden requests or requests made by Web robots).

The main preprocessing tasks include in this research data collection, data cleaning, attribute selection, transaction identification, and data transformation.

4.2 Data Collection

In this research, the source of data set is web server log data. The server log data collected from Adama Science and Technology University web users. The user access log from April 07, 2015 to May 06, 2015 are taken, which are accessed by web users. Before deciding the size of the dataset the researcher collected the one year log data from Adama Science and Technology University through collecting from the old and new server since ASTU ICT migrate from the old sever to new one. In addition to the above reason why the researcher used only 30 days dataset and the way how select the specific month the through checking the whole dataset and the researcher had get the data set not complete due to the migration data from old server to new one some day's log data is missed from each month except one month.

Sample log data extracted from row log data of Adama Science and Technology University web server shown below the figure.

38.111.147.84 - - [08/Apr/2015: 1430272797.008 +1100] "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

66.249.64.142 - - [10/Apr/2015: "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

38.111.147.84 - - [07/Apr/2015: 1430272796.702 +1100] "GET / http://www.youtube.com/ /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

38.111.147.84 - - [08/Apr/2015: 1430272797.008 +1100] "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

66.249.64.142 - - [10/Apr/2015: "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_038.111.147.84 - - [07/Apr/2015: 1430272796.702 +1100] "GET / http://www.youtube.com/ /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

38.111.147.84 - - [08/Apr/2015: 1430272797.008 +1100] "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

66.249.64.142 - - [10/Apr/2015: "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_038.111.147.84 - - [07/Apr/2015: 1430272796.702 +1100] "GET / http://www.youtube.com/ /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

38.111.147.84 - - [08/Apr/2015: 1430272797.008 +1100] "GET / www.addiszefen.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 66.249.64.142 - - [10/Apr/2015: "GET /www.facebook.com /HTTP/1.0" 200 450910 , "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0

Figure 4.1 sample web log data

No	Attribute of log File	Description of Attribute of log File
1	Date	The date on which the activity occurred.
2	Time	The time at which the activity occurred.
3	Server IP Address	The IP address of the server on which server generated the log entry.
4	Client IP Address	The IP address of the client that accessed to the server.
5	Time Stamp	The duration of the action
6	URI	Full page address or request as it came from the client.
7	HTTP Status	The HTTP status code returned to the client, indicating whether the file was successfully retrieved and if not what error message was returned.
8	Bytes	Number of bytes transferred.
9	Agent	Operating system and browser software at the client.
10	Method	Method of request (Get, Post) the action the client was trying to perform.

Table 4.1 Description of web server log file attribute description

4.3 Data Preprocessing

One of the important core steps of knowledge discovery is data preprocessing. The main goal of this step is to create minable objects for knowledge discovery despite the presence of ambiguities and incompleteness in data. Data preparation (or data preprocessing) in this context means manipulation of data into a form suitable for further analysis and processing. It is a process that involves many different tasks and which cannot be fully automated. Many of the data preparation activities are routine, tedious, and time consuming. It has been estimated that data preparation accounts much of the time spent on a data mining research.

The aim of data pre-processing is to clean data from irrelevant records, user and session identification, select essential features and finally transform raw data for mining. All these stages will be analyzed in more detail in order to understand why pre-processing plays an important role in knowledge discovery process complex web log data. The Web log data of this research preprocessed by using data-preparator-1.7 tool, log file viewer tool and MS- Excel 2013.

4.3.1 Data Cleaning

In this phase record contains images, videos, Cascading Style Sheet files (CSS), scripts, and flash animations that are not necessary for statistical analysis and Web usage mining are removed.

To remove the page/files with file contain unwanted for statistical analysis and Web usage mining were selected based on the objective of the study and consultation of domain expert's Unwanted attribute are removed using Log file viewer standard edition and MS- Excel 2013. The screenshot of log file viewer standard edition shown below in figure 4.2.

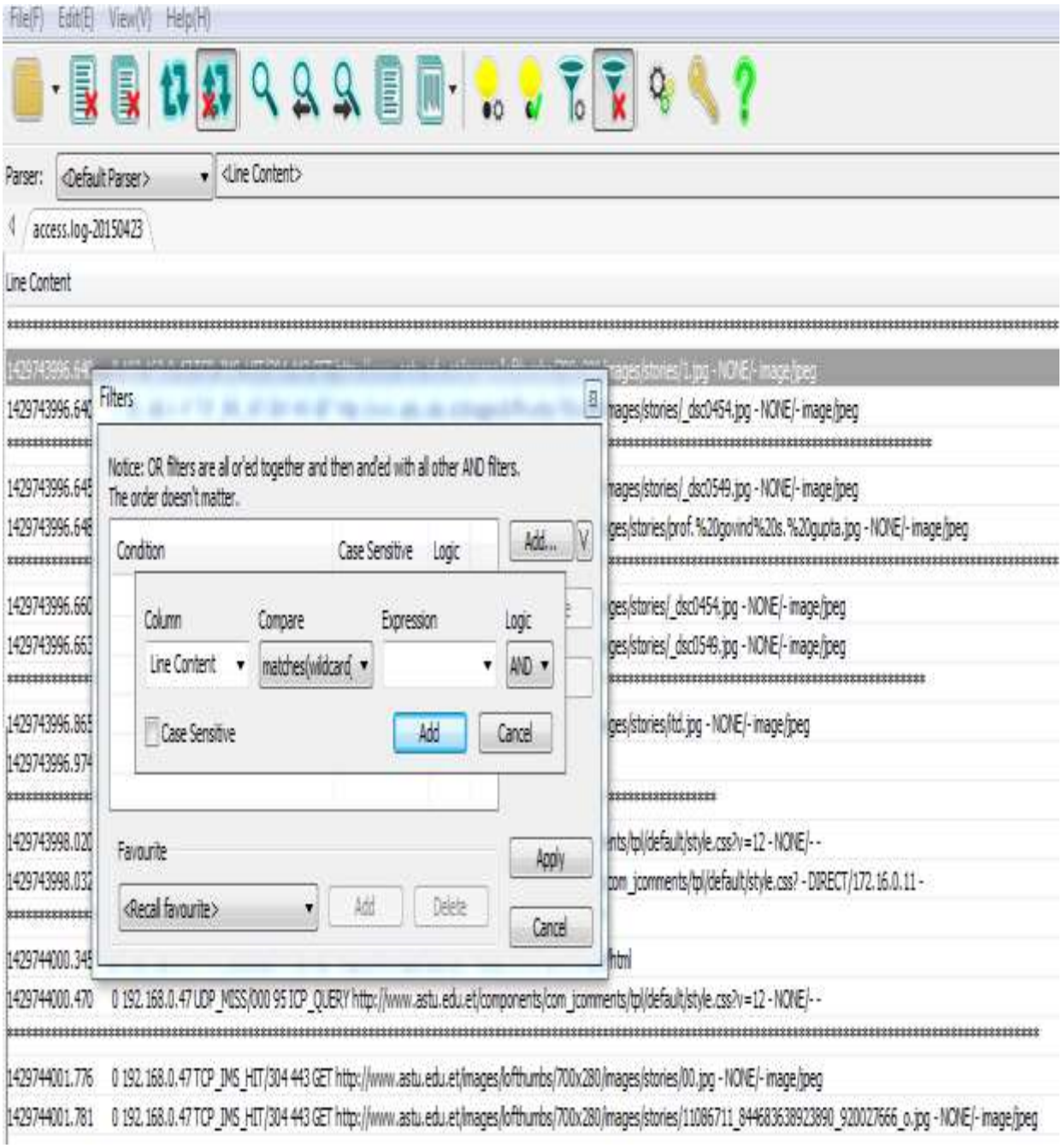

Figure 4.2 data cleaning

After removing unwanted file from each log data record since from selected attribute access time were recorded by linux time format like this 1430272797.008.for further preprocessing Linux time must first converted into normal time format. Because of MS- Excel 2013 doesn't contain built-in functions for working with UNIX dates and time so they must be derived. But MS- Excel 2013 allows you to add formula that can convert the Linux time into normal time.

To do this the reseacher were use EPOCH formal that able to covert linux time in a simple ways using the MS- Excel 2013 function by the formula =CELL/(60*60*24)+"07/04/2015".

Further cleaning done by using MS- Excel 2013. In this cleaning phase the log record is converted to the .log extensions file to MS- Excel 2013 file.

No	Log Date	# of Record Before Cleaning	# of Log Record After Cleaning by log Viewer	# of Log Record After Cleaning by Excel
1	2015-05-06	26,365,536	265,536	2,177
2	2015-05-05	44,554,536	455,536	1,749
3	2015-05-04	6,706,466	706,466	1,490
4	2015-05-03	9,806,065	806,065	7,734
5	2015-05-02	8,904,644	904,644	10,950
6	2015-05-01	89,766,230	9,766,230	6,198
7	2015-04-30	9,646,546	646,546	2,394
8	2015-04-29	19,562,324	62,324	1,510
9	2015-04-28	11,629,456	629,456	1,345
10	2015-04-27	66,235,456	635,456	1,593
11	2015-05-26	62,010,456	610,456	2,445
12	2015-04-25	11,536,564	156,564	1,617
13	2015-04-24	2,435,546	245,546	6,522
14	2015-04-23	1,863,456	186,456	5,122
15	2015-04-22	1,653,445	165,445	1,326
16	2015-04-21	2,626,578	266,578	2,042
17	2015-04-20	3,553,645	355,645	1,992
18	2015-04-19	993,656	99,656	8,052
19	2015-04-18	2,993,654	299,654	5,754
20	2015-04-17	6,222,356	622,356	6,444
21	2015-04-16	7,122,363	712,363	3,106
22	2015-04-15	6,266,646	626,646	1,362
23	2015-04-14	6,236,369	623,369	2,033
24	2015-04-13	8,474,659	847,659	2,885
25	2015-04-12	6,566,456	656,456	7,029
26	2015-04-11	8,005,465	400,465	19,239

27	2015-04-10	6,492,456	649,456	5,216
28	2015-04-09	1,664,958	166,958	2,890
29	2015-04-08	5,462,554	546,554	2,775
30	2015-04-07	6,222,345	622,345	1,436
Total	30 days	162,781,886	23,738,886	126,427

Table 4.2 Summary of preprocessed daily log dataset

4.3.2 Data Categorization

The total of 1126,427 dataset is used for both statistical analysis and data mining techniques. The researcher based on the advice of domain experts categorized the experiment into six categories such as experiment one using morning office hour web user log dataset, experiment two using lunch time web user log dataset, experiment three using afternoon office hour web user log dataset, experiment four using non office hour web user log dataset, experiment five using non weekend web user log dataset, and experiment six using weekend web user log dataset.

The categorization process were done using MS- Excel 2013 finds and filter operation by balancing the objective of the study and the advice of experts on the nature of the log dataset.

No	Categorized web user dataset	Time interval	# of Log Record
1	morning office hour web user log dataset	8:00AM-11:59AM	34,220
2	lunch time web user log dataset	12:00AM – 1:29PM	20,689
3	afternoon office hour web user log dataset	1:30PM- 5:29PM	41,519
4	non office hour web user log dataset	5:30PM - 7:59AM	29,999
5	non weekend web user log dataset	From Monday - Friday	86,226
6	weekend web user log dataset	Saturday and Sunday	40,201

Table 4.3 Categorized in time interval of accessed log dataset

4.3.3 Attribute Selection

For statistical analysis after the dataset is prepared and categorized into six categories from each category the access log data records such as IP address (identify who has visited the website), date (access time of the website), visiting path (path taken by the user while visiting the website), path traversed (the path taken by the user within the website), success rate (status code returned by the server), URL (the page that accessed by the user), request type (GET or POST), Ways of access(type web address directly, use search engine, other) and user agent (browser type that user use to send request) were selected. Apart from this derived attribute time interval value of this attribute (morning office hour, lunch time, afternoon office hour, non-office hour, non-weekend, and weekend), class of site attribute value of this attribute (social media and entertainment site, educational site, governmental site and email site user) were added from the original attribute of log dataset by considering the objective of the study and the advices of domain experts.

For association rule mining techniques experiment appropriate attributes selected from each castigatory of dataset in statistical analysis experiment result by identifying frequent URL.

4.3.4 Data Formatting

After the dataset are categorized based on the objective of the study and the suggestion of domain experts, the dataset is transformed into appropriate extension for both statistical analysis and data mining rule techniques to describe users navigational behaviors of Adama Science and technology university web users.

The whole attributes were in attribute selection phase were prepared and saved as(morning office hour web user log dataset.xls, lunch time web user log dataset.xls, afternoon office hour web user log dataset.xls, non-office hour web user log dataset.xls, non-weekend web user log dataset.xls, and weekend web user log dataset.xls) for statistical analysis techniques experiment.

After the dataset are categorized based on the objective of the study and the suggestion of domain experts, from the whole attributes are used for statistical analysis techniques experiment.

URL attribute are selected for data mining techniques of experiments to describe users navigational behaviors of Adama Science and technology university web users.

From categorized dataset (morning office hour web user log dataset, lunch time web user log dataset, afternoon office hour web user log dataset, non-office hour web user log dataset, non-weekend web user log dataset, and weekend web user log dataset) single selected URL attribute which contains 34 attributes (textual learning materials, www.aau.edu.et, www.adamacity.gov.et, www.addiszefen.com,www.aljazeera.com,www.amharictube.com,www.amra.gov.et,www.articles.org,www.astu.edu.et,www.catholic.org,www.diretube.com,www.educationaltutorial.com,www.dstv,www.ecsu.edu.et,www.erca.gov.et,www.ertagov.com,www.ethiopianorthodox.org,www.ethiopianreporter.com,www.ethiotelecom.et,www.ethiotube.net,www.facebook.com,www.freee-books.ne,www.gmail.com,www.moin.gov.et,www.most.gov.et,www.nbe.gov.et,www.onlainporno.tv,www.oromia.gov.com,www.scholarships.com,www.skype.com,www.twitter.com,Download software, www.yehabesha.com, and www.youtube.com),values converted into 34 attribute and the values of the those attribute(accessed or ?) (Page viewed) were selected.

To do transformation dataset into dataset that could be compatible to data mining techniques of experiments the reseacher is used in a simple way using the MS- Excel 2013 function by the formula.

When the researcher select attributes that is important to the study totally based on the appropriateness attributes for both technique and research objective.

Before selecting URL attribute values for data mining technique those URL value originated to represent different user's request on the web. Since web server respond to users based on the verities of user request about single site, to converge into representative URL the researcher is done by converging similar feature requests.

For instance if someone browse chemistry educational tutorial in youtube the researcher coverage this to www.educational tutorial.com in order to simplify the hugeness of the URL for different purpose when the web user browsed. After this modification from millions of URL 34 URL were selected to represent the rest converged URL.

The selected attribute were attached in, Appendix (A).

URL	URL Description
URL1	textual learning materials
URL2	www.aau.edu.et
URL3	www.adamacity.gov.et
URL4	www.addiszefen.com
URL5	www.aljazeera.com
URL6	www.amharictube.com
URL7	www.amra.gov.et
URL8	www.articles.org
URL9	www.astu.edu.et
URL10	www.catholic.org
URL11	www.diretube.com
URL12	www.educational tutorial.com
URL13	www.dstv.com
URL14	www.ecsu.edu.et
URL15	www.erca.gov.et
URL16	www.ertagov.com
URL17	www.ethiopianorthodox.org
URL18	www.ethiopianreporter.com
URL19	www.ethiotelecom.et
URL20	www.ethiotube.net
URL21	www.facebook.com
URL22	www.free-ebooks.net
URL23	www.gmail.com
URL24	www.moin.gov.et
URL25	www.most.gov.et
URL26	www.nbe.gov.et
URL27	www.onlain-porno.tv
URL28	www.oromia.gov.com
URL29	www.scholarships.com
URL30	www.skype.com
URL31	www.twitter.com
URL32	Download software
URL33	www.yehabesha.com
URL34	www.youtube.com

Table 4.4 Sample page/data frequently accessed by the web users

Using selected URL attribute that contain 34 other attributes in the second attribute selection phase then transformed into minable manner by using data mining algorithm in a simple way using the MS- Excel 2013 function by the formula and saved as(morning office hour web user log dataset.csv, lunch time web user log dataset.csv, afternoon office hour web user log dataset.csv, non-office hour web user log dataset.csv, non-weekend web user log dataset.csv, and weekend web user log dataset.csv) for data mining techniques experiment.

URL22	URL23	URL26	URL27	URL28	URL29	URL30	URL31	URL32	URL33	URL34
?	?	accessed	accessed	?	accessed	accessed	?	accessed	accessed	accessed
?	?	?	?	accessed	accessed	accessed	accessed	accessed	?	accessed
?	?	?	?	?	?	?	?	?	?	?
?	?	accessed	accessed	?	?	accessed	?	accessed	accessed	?
?	?	?	accessed	accessed	accessed	accessed	accessed	?	accessed	?
?	?	?	?	accessed	?	?	?	?	?	accessed
?	?	?	?	accessed	?	?	?	?	?	accessed
?	?	?	?	accessed	?	?	?	?	?	accessed
?	?	?	?	?	?	?	?	?	?	accessed
?	?	?	?	?	?	?	?	?	?	accessed
?	?	?	?	?	?	?	?	?	?	?
?	?	?	accessed	accessed	accessed	accessed	accessed	accessed	accessed	accessed
?	?	?	accessed		accessed		accessed		accessed	accessed

Figure 4.3 sample transformed selected attribute dataset for association rule discovery.

CHAPTER FIVE

EXPERIMENT AND FINDINGS

After data preparation is completed, the experiment conducted to describe user web navigational behaviors of Adama Science and Technology University web users. In this study experiments conducted using statistical analysis and data mining techniques to discover navigational behaviors of Adama Science and Technology University web users. To accomplish the experiment Data-preparator-1.7 tool for statistical analysis and Weka 3.7.4 for association rule discovery were used. The experiment conducted using the following hardware and software requirements:-

Computer Type: Satellite L855 lap top

Operating system: Microsoft window 7 ultimate edition.

Processor: Intel (R) Core (TM) i3- 3120M CPU @2.50GHZ 2.50GHZ.

RAM: 4.00GB

Data preprocessing tool: web log viewer, data preparator-1.7 tool, MS- Excel 2013

Statistical Analysis Tool: data preparator-1.7 tool

Web mining tool: Weka 3.7.4

5.1 Experiment Setup

The experimental arrangement is done using the total of 126,427 prepared records dataset experiment is categorized into six different experimental class based on the objective of the study and researcher understanding of the dataset occurrences.

In addition to the above statement the base for the researcher categorized the experiment as declared in the preprocessing chapter the researcher based on the nature of the data occurred and the suggestions taken from the domain experts of Adama Science and Technology University ICT workers specially system administrators. Those categorized experiment with dataset limit are listed below.

- ✓ Experiment one using morning office hour web user dataset (8:00AM-11:59AM) with the total of 34,220 prepared log dataset.
- ✓ Experiment two using Lunch time web user dataset (12:00AM – 1:29PM with the total of 20,689 prepared log dataset.

- ✓ Experiment three using afternoon office hour web user dataset (1:30PM- 5:29PM with the total of 41,519prepared log dataset.
- ✓ Experiment four using non office hour web user dataset from 5:30PM - 7:59AM with the total of 29,999 prepared log dataset.
- ✓ Experiment five using non weekend web user dataset (Monday – Friday) with the total of 86,226 prepared log dataset.
- ✓ Experiment six using weekend web user dataset (Saturday and Sunday) with the total of 40,201 prepared log dataset.

Those above six categorized experiment is done for both statistical analysis and data mining techniques, even though for association rule mining experiment the dataset is changed through identifying the most frequent URL based on the result of statistical analysis. For association rule mining to find correlations between web pages accessed together in categorized time interval Apriori and FP-Growth algorithm are used.

5.2 Statistical Analysis

Statistical analysis techniques are the most common method to extract descriptive knowledge about user navigational behaviors of web resources. In this study Data-preparator-1.7 toolis used as a tool to analyze and illustrate user navigational behaviors of Adama Science and Technology University web users.

By using Data-preparator-1.7 tool the summary of web usage report is generated concerning most frequently accessed page (URL), and frequently accessed categories of site(social media and entertainment site, educational site, governmental site and email site user). Statistical analysis using Data-preparator-1.7 tool conducted using 30 (thirty) days web log data which taken from April 07, 2015 to May 06, 2015 in Adama Science and Technology University web users with total of 126,427 prepared Web log records. The detail of the report of the Statistical analysis technique is described below with in each experiment and with the size of prepared dataset.

5.2.1 Experiment one using morning office hour web user dataset

Morning office hour log dataset which encompassed the time interval from 8:00AM-11:59AM user access log data. As expressed in detail in preprocessing phase for experiment one after preprocessing statistical analysis using Data-preparator-1.7 tool conducted using 30 (thirty) days web log data with 34,220 preprocessed web log records. The detail reports of the experiment is described as below in figure 5.1 and 5.2.

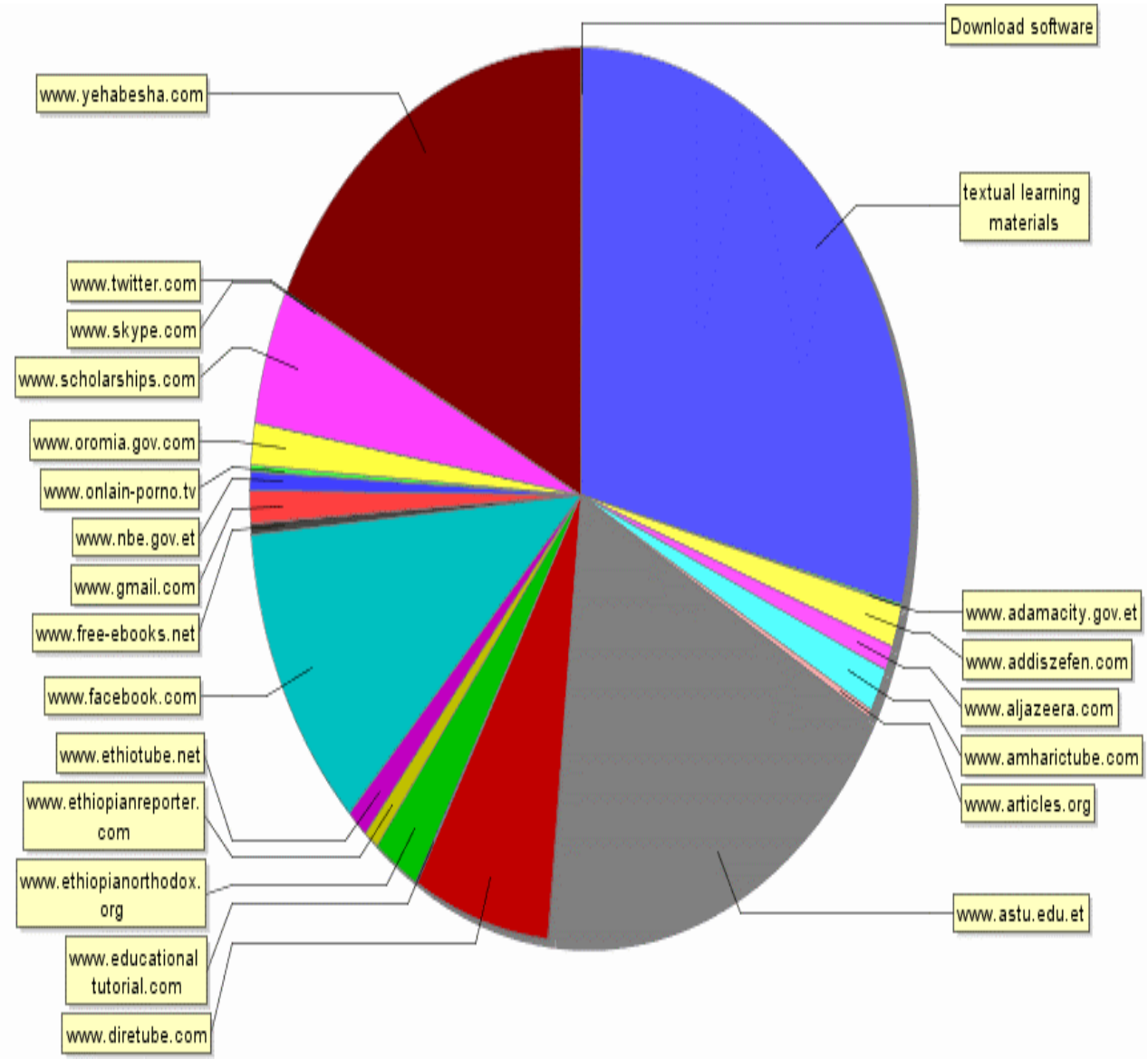

Figure 5.1 Different URL visitors' statistics for morning office hour web users

As shown in the above figure 5.1 top frequently access page (URL) are textual learning materials, www.yehabesha.com, www.astu.edu.et, www.facebook.com and www.diretube.com.

This indicates that even though textual learning materials as educational site and www.astu.edu.et as organizational site are most frequently visited from the morning office hour web users, but based on the total of most frequently accessed URL in the morning office hour web users social media and entertainment site more frequently is visited.

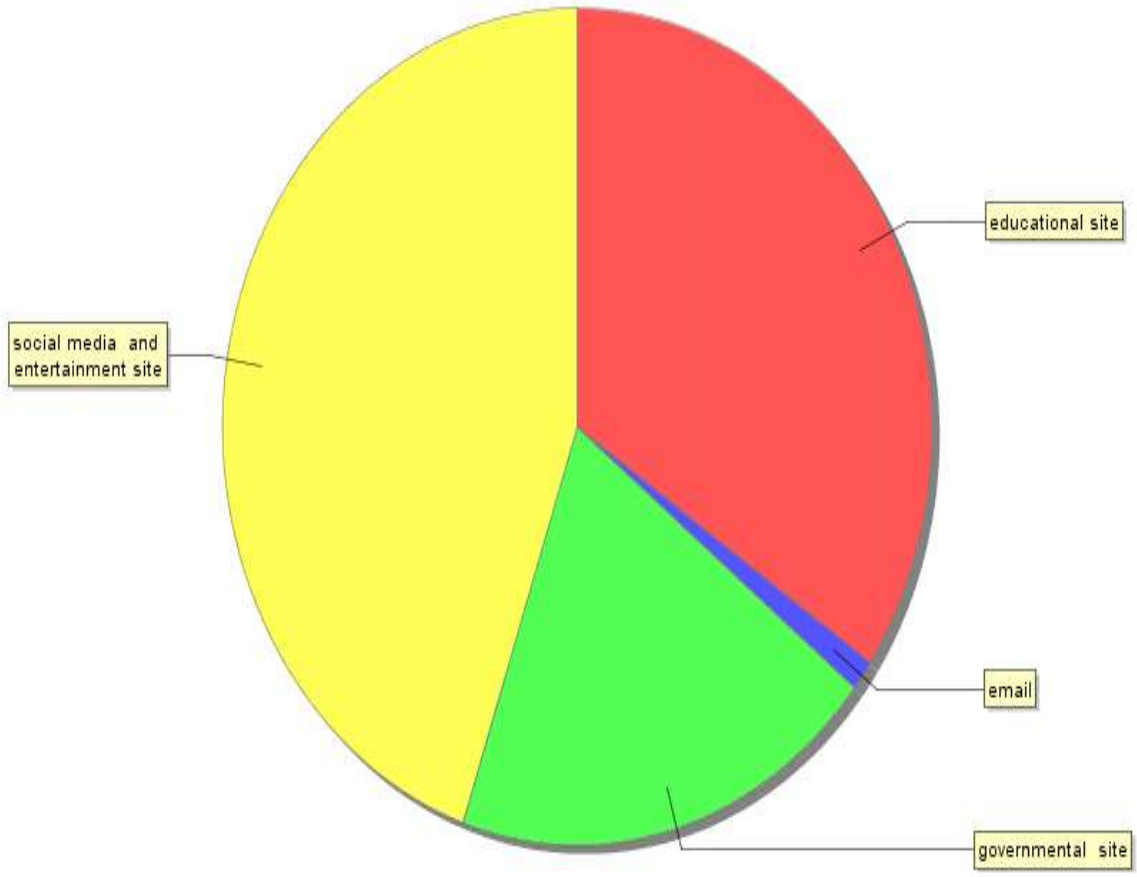

Figure 5.2: The statistics of browsed site by Morning office hour web users

Even if in figure 5.1 is described the most frequently accessed site by sum up the most frequently accessed URL to assure using by using own attribute statistics that shown in the above figure 5.2 in morning office hour web users social media and entertainment site frequently accessed site.

5.2.2 Experiment two using Lunch time web user dataset

Lunch time log dataset which encompassed the time interval from 12:00AM – 1:29PM web log data .As expressed detail in preprocessing phase for experiment two after preprocessing statistical analysis using Data-preparator-1.7 tool conducted using 30 (thirty) days web log data with 20,689 preprocessed web log records. The detail reports of the experiment is described as below in figure 5.3 and 5.4.

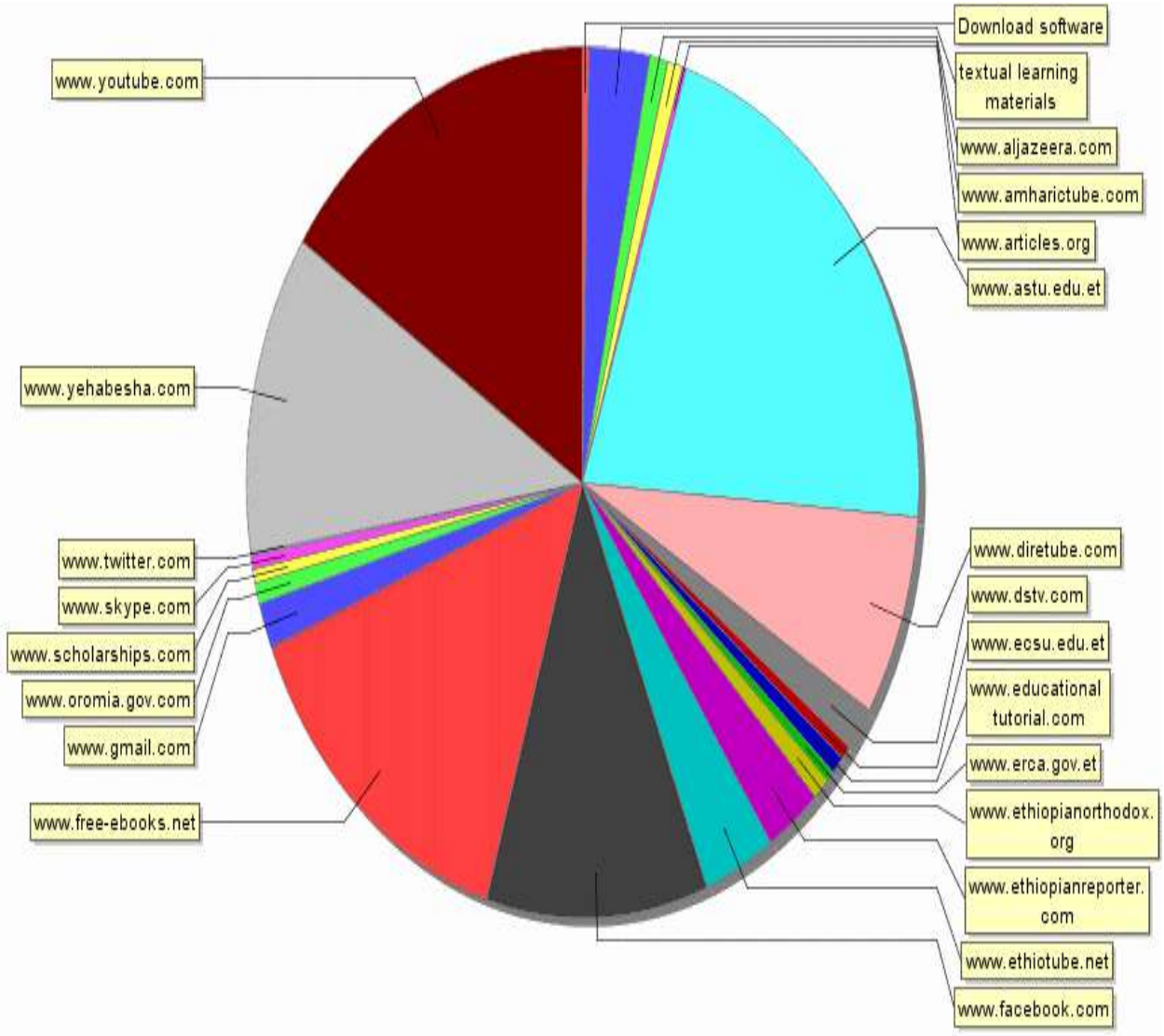

Figure 5.3Different URL visitors' statistics for Lunch time web users

As shown in the above figure 5.3 top frequently access page (URL) are www.astu.edu.et,www.yehabesha.com,www.youtube.com,www.freeebooks.net,www.facebook.com,www.diretube.com.

This indicates that even though www.astu.edu.et as organizational site and www.freeebooks.net as educational site are most frequently visited in the Lunch time web users, but based on the total of most frequently accessed URL in the Lunch time web users social media and entertainment site more frequently is visited.

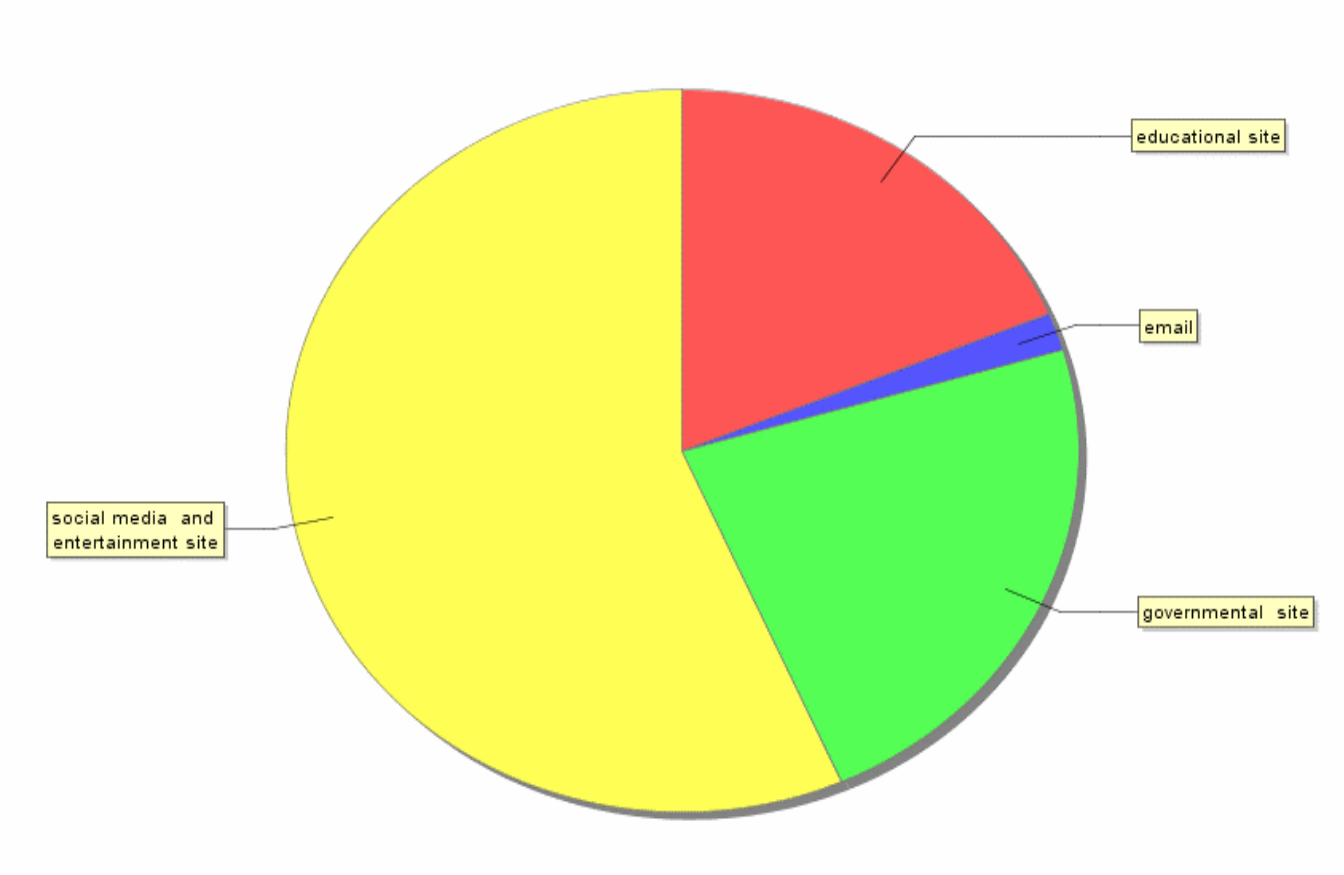

Figure 5.4The statistics of browsed site by Lunch time web users

Even if in figure 5.3 is described the most frequently accessed site by sum up the most frequently accessed URL to assure using by using own attribute statistics that shown in the above figure 5.4 in Lunch time web users social media and entertainment site frequently accessed site.

5.2.3 Experiment three using afternoon office hour web user dataset

Afternoon office hour log dataset which encompassed the time interval from 1:30PM- 5:29PMuser access log data. As expressed detail in preprocessing phase for experiment three after preprocessing statistical analysis using Data-preparator-1.7 tool conducted using 30 (thirty) days web log data with 41,519 preprocessed web log records. The detail reports of the experiment is described as below in figure 5.5 and 5.6.

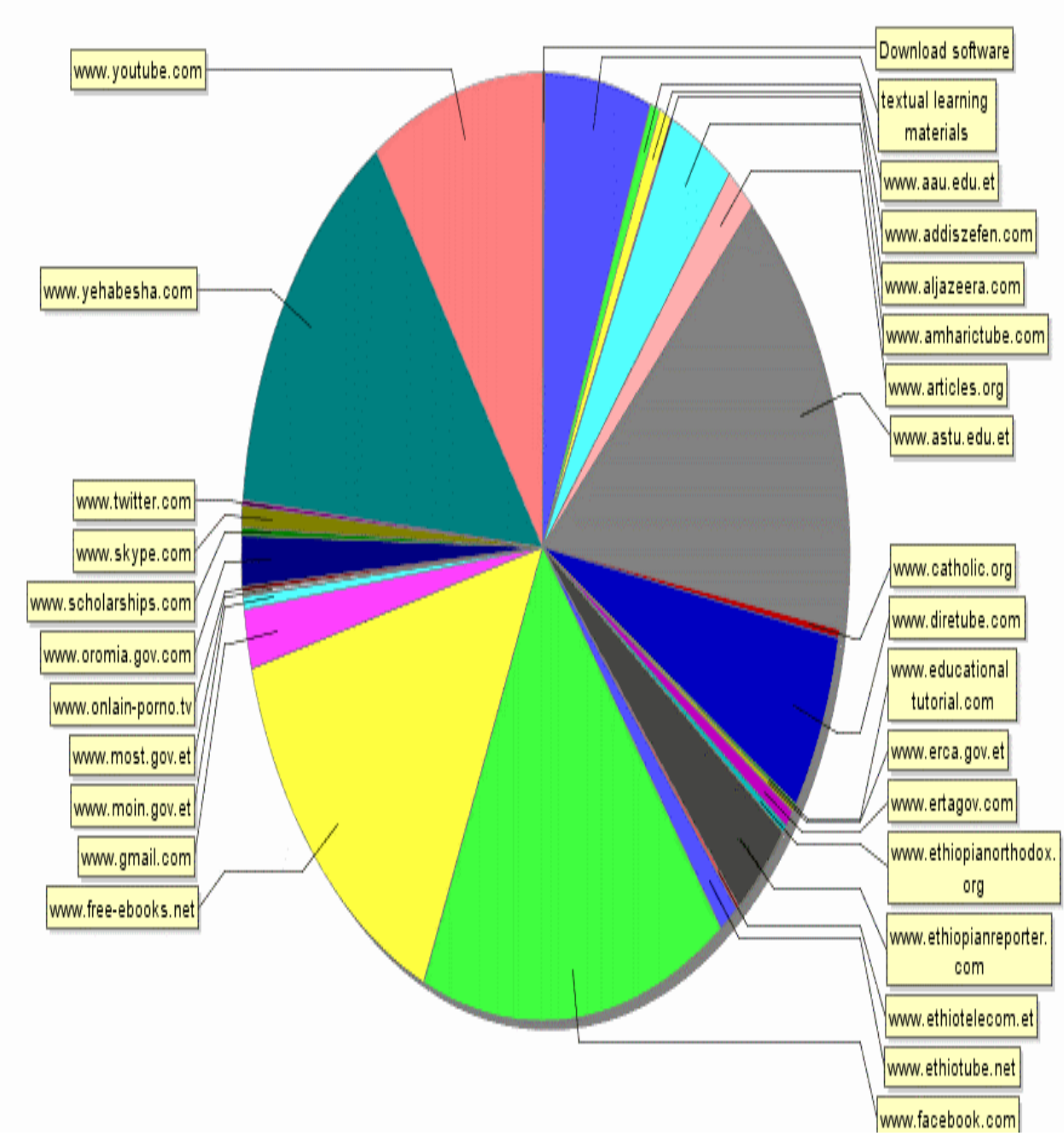

Figure 5.5 Different URL visitors' statistics for Afternoon office hour web users

As shown in the above figure 5.5 top frequently access page (URL) are www.facebook.com, www.astu.edu.et,www.yehabesha.com,www.youtube.com,www.freeebooks.net,www.diretube.com. This indicates that even though www.astu.edu.et as organizational site and www.freeebooks.net as educational site are most frequently visited in the Afternoon office hour web users, but based on the total of most frequently accessed URL in the Afternoon office hour web users social media and entertainment site more frequently is visited.

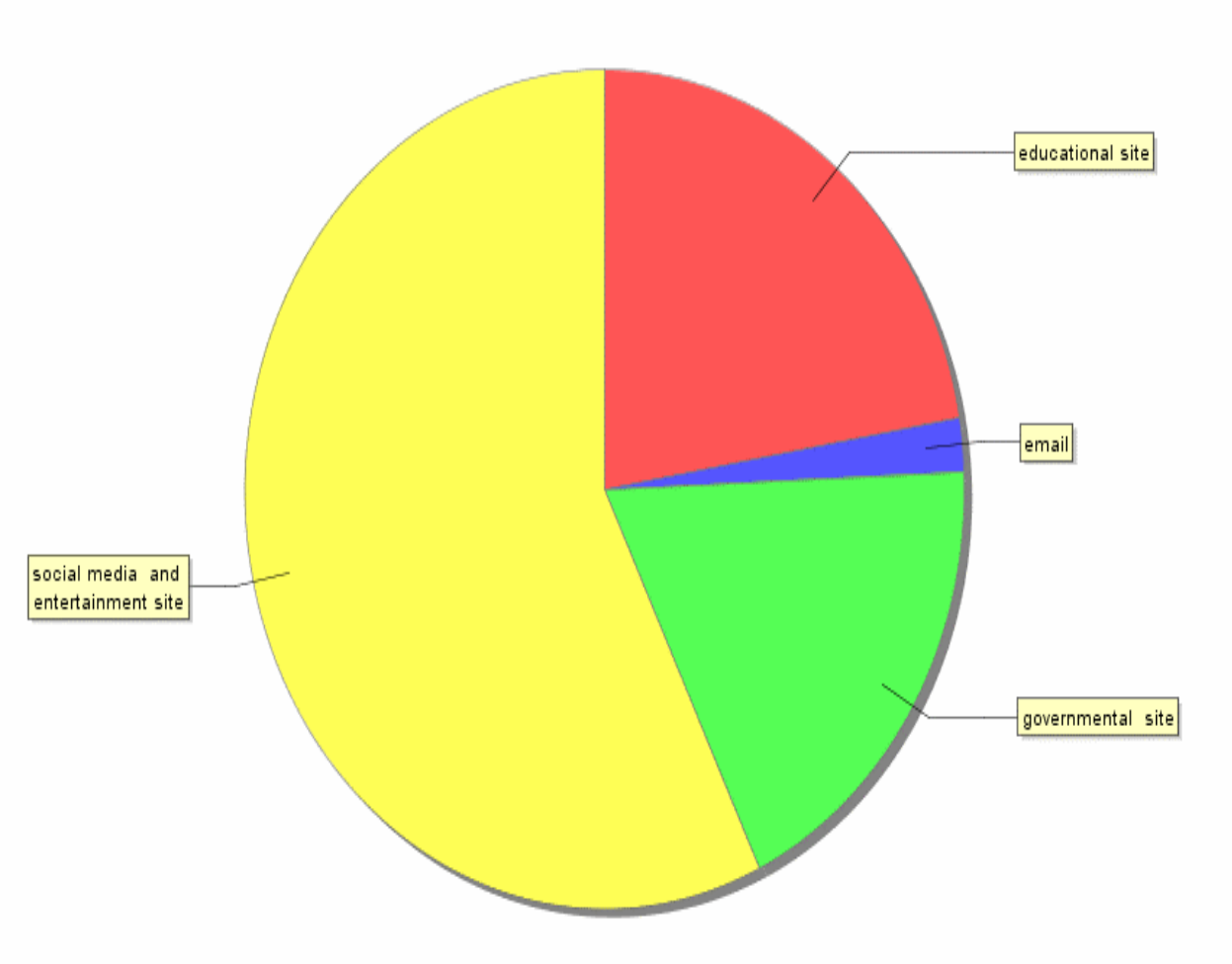

Figure 5.6 The statistics of browsed site by Afternoon office hour web users

Even if in figure 5.5 is described the most frequently accessed site by sum up the most frequently accessed URL to assure using by using own attribute statistics that shown in the above figure 5.6 in Afternoon office hour web users social media and entertainment site frequently accessed site.

5.2.4 Experiment four using Non-office hour web user dataset

Non-office hour log dataset which encompassed the time interval from 5:30PM - 7:59AMuser access log data. As expressed detail in preprocessing phase for experiment four after preprocessing statistical analysis using datapreparator-1.7 tool conducted using 30 (thirty) days web log data with 29,999 preprocessed web log records. The detail reports of the experiment is described as below in figure 5.7 and 5.8.

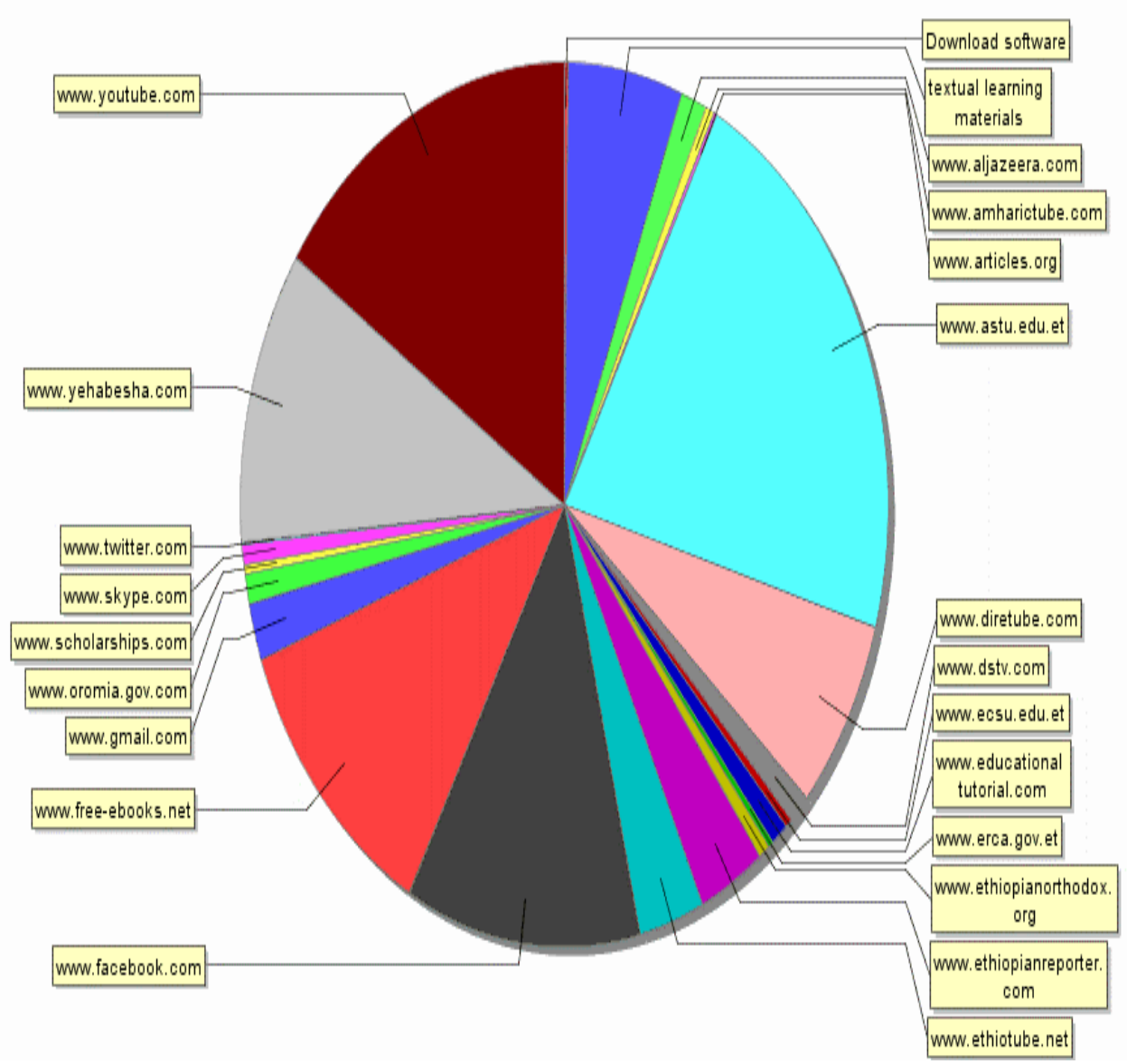

Figure 5.7 Different URL visitors' statistics for Non-office hour web users

As shown in the above figure 5.7 top frequently access page (URL) are www.facebook.com,www.astu.edu.et,www.yehabesha.com,www.youtube.com,www.freeebooks.net, ,www.diretube.com.This indicates that even though www.astu.edu.et as organizational site and www.freeebooks.net as educational site are most frequently visited in the Non office hour web users, but based on the total of most frequently accessed URL in the Non-office hour web users social media and entertainment site more frequently is visited.

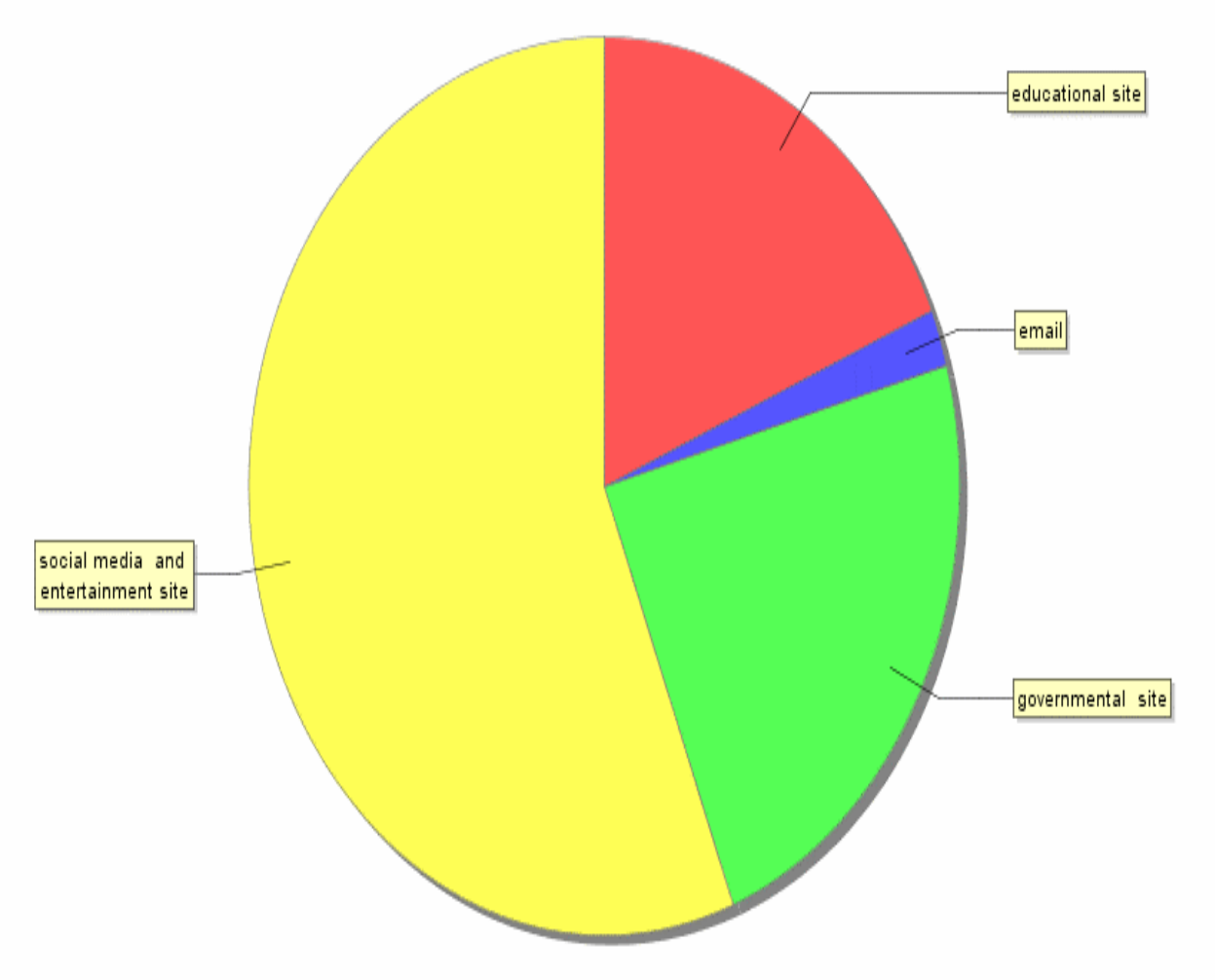

Figure 5.8 The statistics of browsed site by Non-office hour web users

Even if in figure 5.7 is described the most frequently accessed site by sum up the most frequently accessed URL to assure using by using own attribute statistics that shown in the above figure 5.8 in Non-office hour web users social media and entertainment site frequently accessed site.

5.2.5 Experiment five using non weekend web user dataset

Non weekend log dataset which encompassed the time interval from Monday - Friday user access log data. As expressed detail in preprocessing phase for experiment five after preprocessing statistical analysis using datapreparator-1.7 tool conducted using 30 (thirty) days web log data with 86,226 preprocessed web log records. The detail reports of the experiment is described as below in figure 5.9 and 5.10.

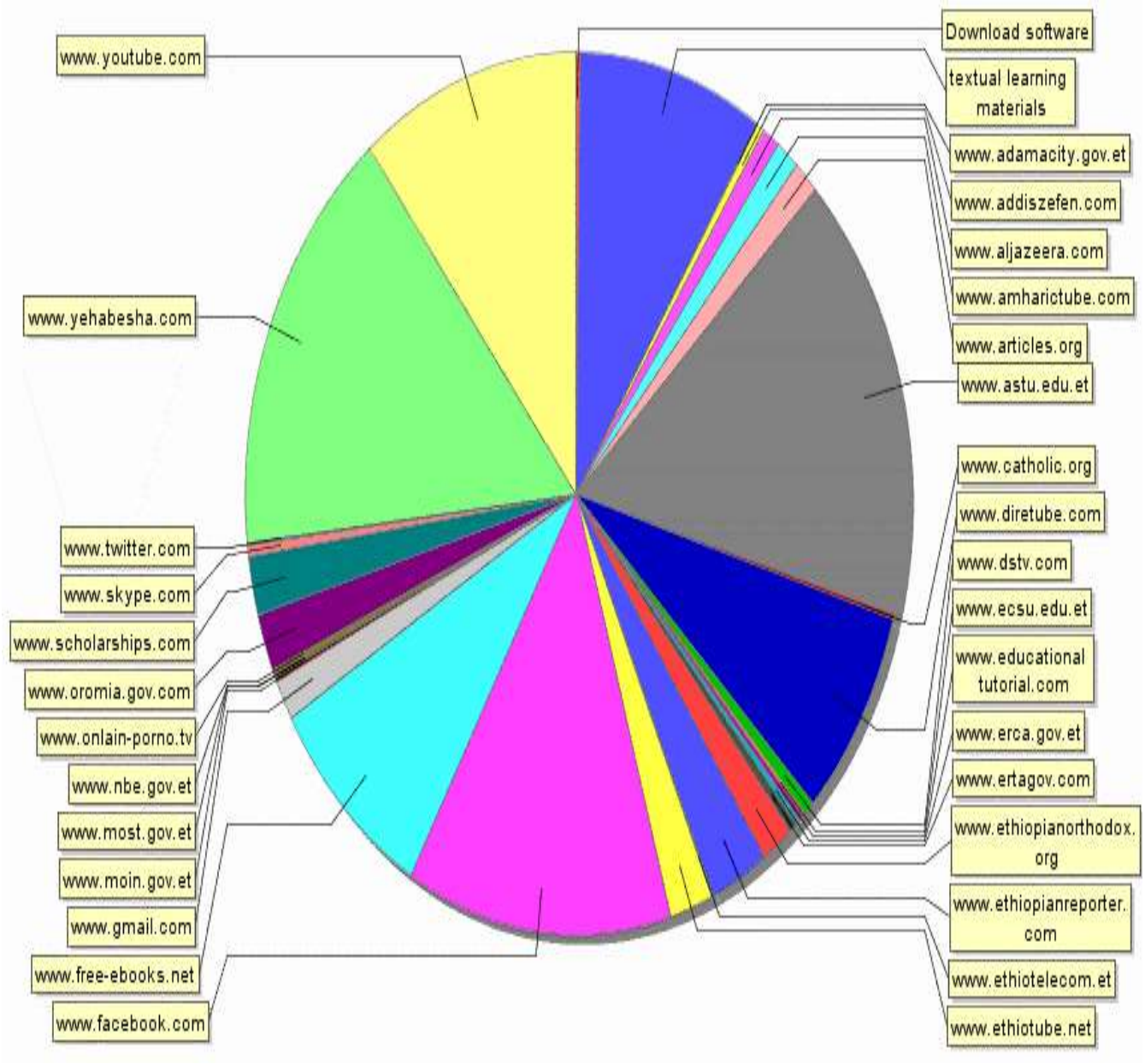

Figure 5.9 Different URL visitors' statistics for Non-weekend web users

As shown in the above figure 5.9 top frequently access page (URL) are www.facebook.com,www.astu.edu.et,www.yehabesha.com,www.youtube.com,www.freeebooks.net, ,www.diretube.com,textual learning materials. This indicates that even though www.astu.edu.et as organizational site and www.freeebooks.net and textual learning materials as educational site are most frequently visited in the Non-weekend web users, but based on the total

of most frequently accessed URL in the Non-weekend web users, social media and entertainment site more frequently is visited.

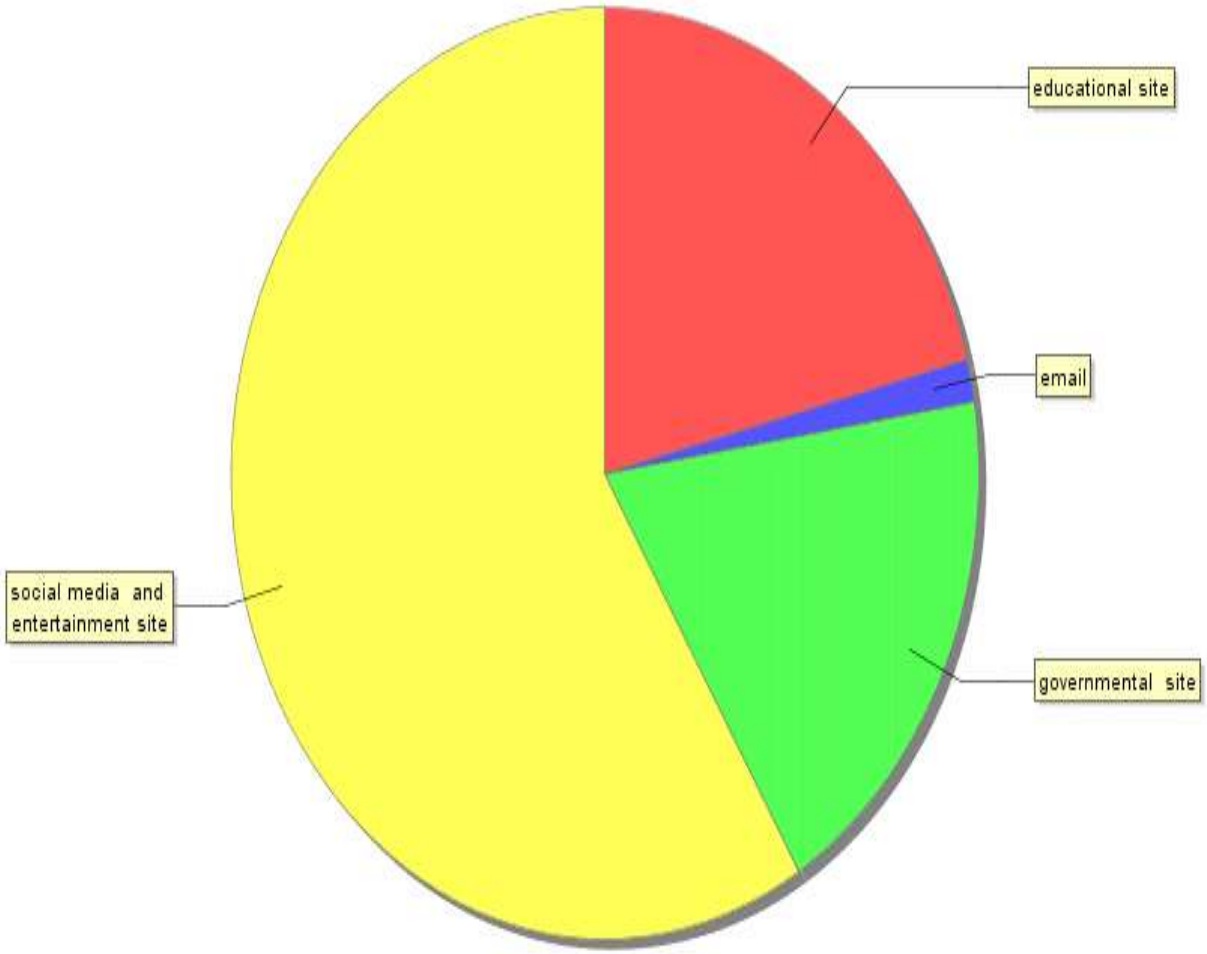

Figure 5.10 The statistics of browsed site by Non-weekend web users

Even if in figure 5.9 is described the most frequently accessed site by sum up the most frequently accessed URL to assure using by using own attribute statistics that shown in the above figure 5.10 in Non-weekend web users social media and entertainment site frequently accessed site.

5.2.6 Experiment six using weekend web user dataset

Weekend log dataset which encompassed the time interval from Saturday and Sunday user access log data. As expressed detail in preprocessing phase for experiment six after preprocessing statistical analysis using datapreparator-1.7 tool conducted using 30 (thirty) days web log data with

40,201 preprocessed web log records. The detail reports of the experiment is described as below in figure 5.11 and 5.12.

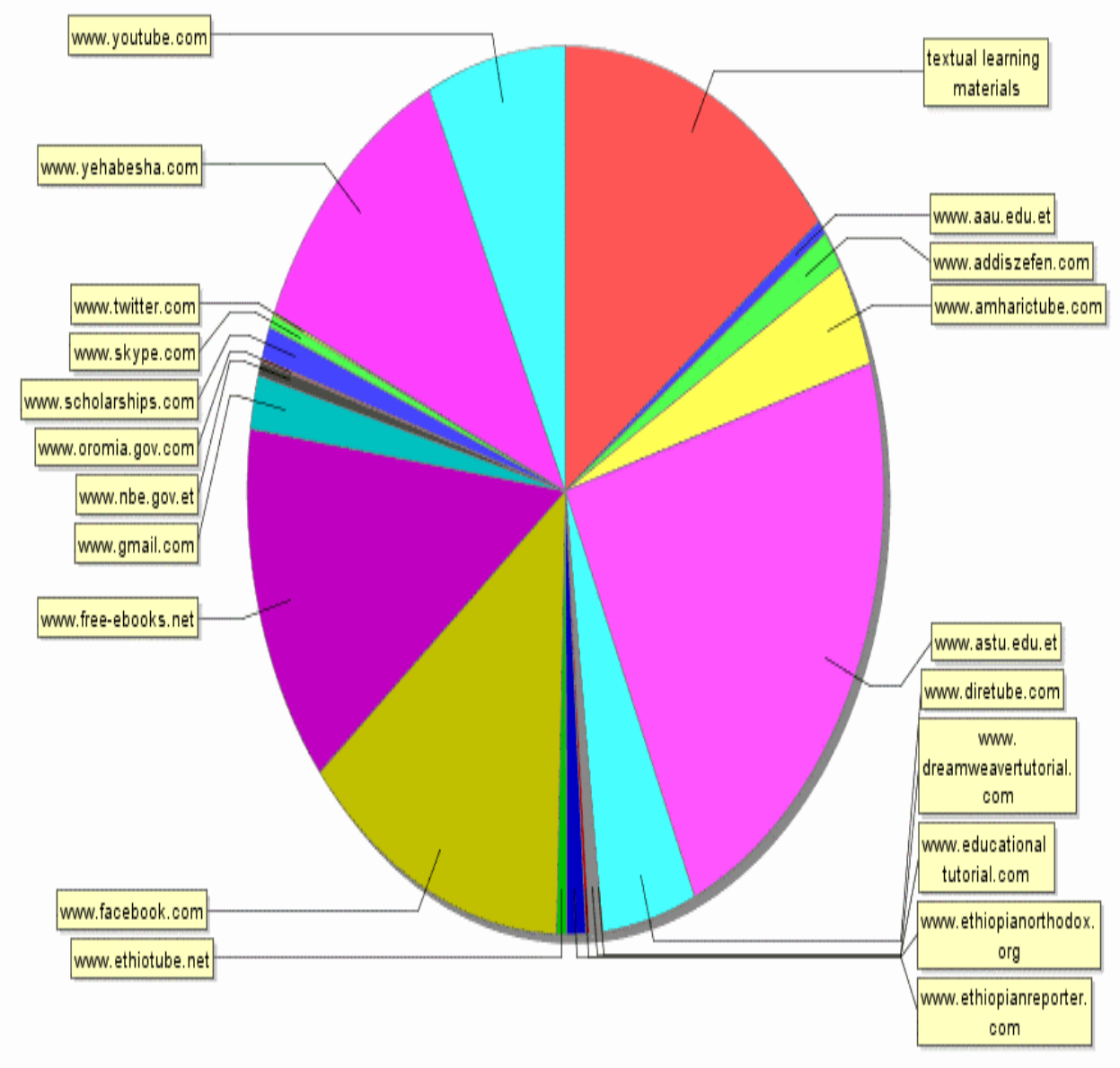

Figure 5.11 Different URL visitors' statistics for Weekend web users

As shown in the above figure 5.11 top frequently access page (URL) are www.facebook.com, www.astu.edu.et, www.yehabesha.com, www.youtube.com, textual learning materials, www.freeebooks.net, www.diretube.com. This indicates that even though www.astu.edu.et as organizational site and www.freeebooks.net and textual learning materials as educational site are

most frequently visited in the Weekend web users, but based on the total of most frequently accessed URL in the Weekend web users, social media and entertainment site more frequently is visited.

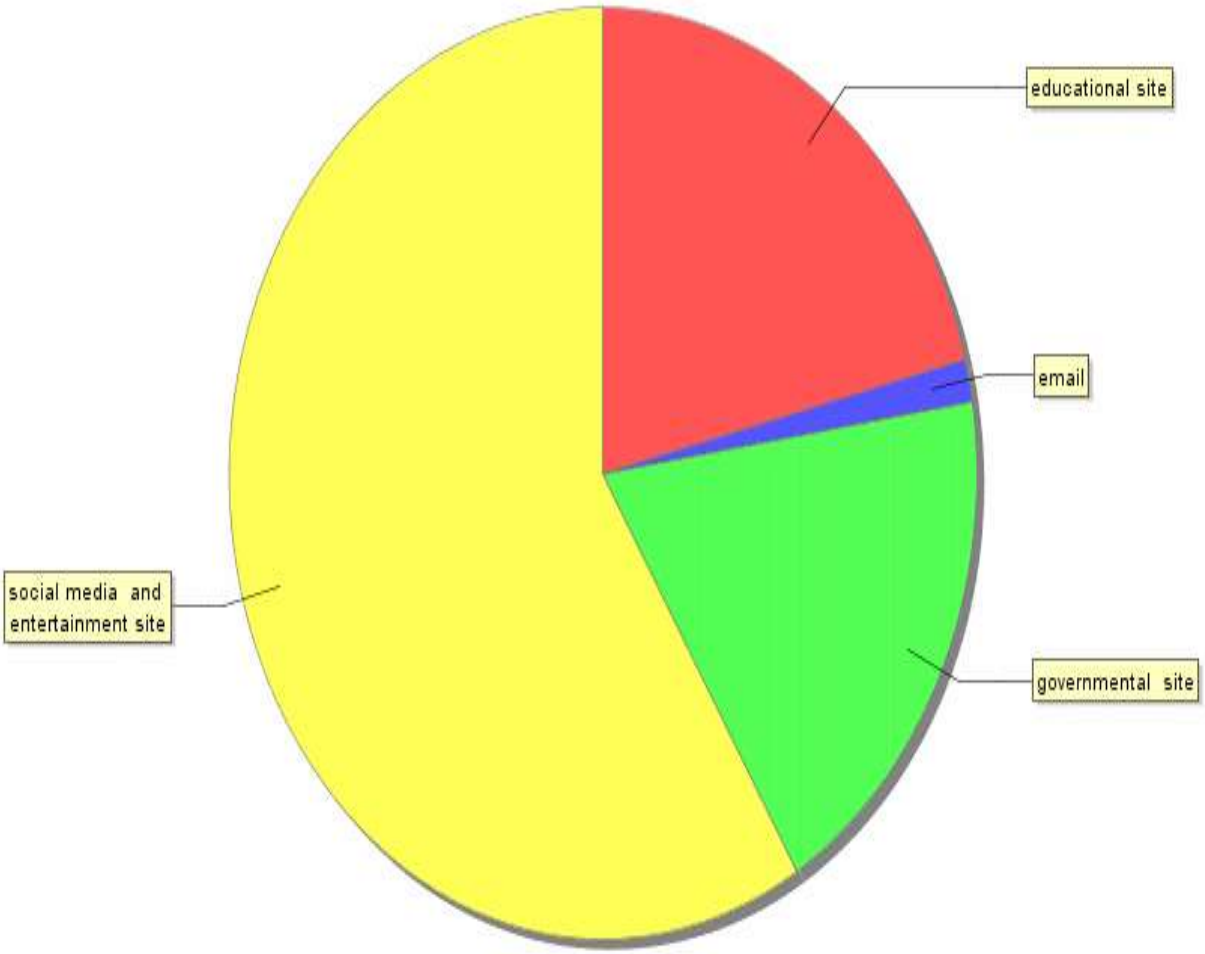

Figure 5.12 The statistics of browsed site by Weekend web users

Even if in figure 5.11 is described the most frequently accessed site by sum up the most frequently accessed URL to assure using by using own attribute statistics that shown in the above figure 5.10 in Weekend web users social media and entertainment site frequently accessed site.

5.3 Association Rule Discovery and Analysis

In this study the main purpose of Web usage mining is to discover an interesting rule from Adama Science and Technology University web users. Pattern discovery is performed after the statistical analysis technique experimental result.

Since the researcher used statistical analysis technique to identify the frequently accessed URL and attribute selection that used transform the data for association rule discovery. The selection of interesting patterns and analysis done through cross checking of the research objective and the results of statistical analysis technique experiment. Interesting rule that discovered from the experiment helps for Adama Science and Technology University ICT to perform different businesses decisions over the web like creating ICT policy. The technique used for this web usage mining process were data mining algorithms. From data mining algorithms association rule mining technique were performed on the preprocessed datasets.

5.3.1 Experimental setup

Association rule mining finds interesting associations from a large set of data items. In this study association rules were used to find correlations between web pages accessed together that categorized by time interval in preprocessing phase. Such rules indicate the possible relationship between pages that are often accessed within the same time, and can reveal associations between groups of users with specific interests. In order to conduct experiments, Weka implementation were used as a tool with Apriori and FP-Growth association rule finding algorithm.

The researcher used explorer environment in Weka, which allows to load experiment data set of .csv format, and to conduct experiment. For experimental purposes, the preprocessed log file issued containing information about a user requests to the web resources. Each line in the web usage log data contains attribute that originally taken from web server and derived attribute by consulting domain experts such like access year, access month, access date, access day, access time, access time class, visitor IP address, status code of protocol, transferred file size, method used, visited URL, site name category, ways of access, server IP address, search engine, file type, platform used and browser used. This preprocessed log data categorized into six categories of experimental once for the experiment statistical analysis technique. The original preprocessed dataset (dataset before categorized into parts) of web log contain 126427. To conduct association rule mining experiment also used the previous categorized six different experimental phase such

as (morning office hour web user dataset, lunch time web user dataset, afternoon office hour web user dataset, non-office hour web user dataset, non-weekend web user dataset, and weekend web user dataset) for statistical analysis technique.

But the researcher using statistical analysis technique results that used to identify the frequently accessed URL and attribute selection, then each six group dataset transformed to Weka understandable format with 34 URL request web resource dataset MS- Excel 2013 function formula by cross checking time and URL. List of each selected URL request for association rule mining are shown in preprocessing phase. While conducting the experiments, different parameter values (car, classindex, lower bound minsupport, metrictype, minmetric, numrules, and upperbound support) are given.

The subsequent experiment result demonstrates interesting patterns discovered from all class of dataset of Weka 3.7.4 experiment using Apriori algorithm. For detail Weka output, see Appendix (B).

5.3.2 Apriori Algorithm Experiment

5.3.2.1 Experiment one using morning office hour web user dataset

Rule: URL9=accessed URL11=accessed URL33=accessed 4809 ==>URL1=accessed 4809 <conf:(1)> lift:(3.47) lev:(0.1) [3421] conv:(3421.67

According to above association rule, if www.astu.edu.et, www.diretube.com and www.yehabesha.com URL are browsed together in the morning office hour, textual educational materials URL had a probability of 100% to be browsed together with the previous URL. This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.2.2 Experiment two using lunch time web user dataset

Rule: URL22=accessed URL33=accessed URL34=accessed 6011 ==> URL21=accessed 6011 <conf:(1)> lift:(2.51) lev:(0.17) [3618] conv:(3618.11)

According to above association rule, if www.freeebooks.net, www.yehabesha.com and www.youtube.com, URL are browsed together during lunch time, www.facebook.com URL had a probability of 100% to be browsed together. This indicate social media and entertainment site

had a great probability to be accessed in the morning office hour. Totally the result of association rule encourage the result of statistical analysis.

5.3.2.3 Experiment three using afternoon office hour web user dataset

Rule: URL21=accessed URL33=accessed 8834 ==> URL9=accessed 8834 <conf:(1)> lift:(2.71) lev:(0.13) [5570] conv:(5570.32)

According to above association rule, if www.facebook.com and www.yehabesha.com URL are browsed together during afternoon office hour, www.astu.edu.etURL had a probability of 100% to be browsed together. This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.2.4 Experiment four using non office hour web user dataset

Rule: URL11=accessed URL21=accessed 10175 ==> URL33=accessed 10175 <conf:(1)> lift:(2.92) lev:(0.22) [6696] conv:(6696.05)

According to above association rule, if www.diretube.com and www.facebook.com URL are browsed together during non-office hour, www.yehabesha.comURL had a probability of 100% to be browsed together.

This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.2.5 Experiment five using non weekend web user dataset

Rule: URL11=accessed URL22=accessed URL34=accessed 11053 ==> URL21=accessed 11053 <conf:(1)> lift:(3.36) lev:(0.12) [7759] conv:(7759.78)

According to above association rule, if www.diretube.com, www.freeebooks.net and www.youtube.com URL are browsed together during non-weekend time, www.facebook.com URL had a probability of 100% to be browsed together. This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule reassure the result of statistical analysis.

5.3.2.6 Experiment six using weekend web user dataset

Rule: URL1=accessed URL9=accessed 12962 ==> URL34=accessed 12962 <conf :(1)>lift :(2.82) lev :(0.21) [8362] conv :(8362.22)

According to above association rule, if textual educational materials and www.astu.edu.et URL are browsed together during weekend time, www.youtube.com entertainment site had a probability of 100% to be browsed together.

This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.3 FP-Growth Algorithm Experiment

The subsequent experiment result shows interesting patterns discovered from all class of dataset of Weka 3.7.4 experiment using FP-Growth algorithm. For detail Weka output, see Appendix (C).

5.3.3.1 Experiment one using morning office hour web user dataset

Rule: [URL33=accessed, URL11=accessed]: 4850 ==> [URL1=accessed]: 4850<conf:(1)> lift:(3.47) lev:(0.1) conv:(3450.84)

According to above association rule, if www.yehabesha.com and www.diretube.com URL are browsed together during morning office hour, textual educational materials educational site had a probability of 100% to be browsed together. This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.3.2 Experiment two using lunch time web user dataset

Rule:[URL22=accessed, URL33=accessed]: 6011 ==> [URL21=accessed]: 6011 <conf:(1)> lift:(2.51) lev:(0.17) conv:(3618.11)

According to above association rule, if www.freeebooks.net and www.yehabesha.com URL are browsed together during lunch time, www.facebook.com entertainment site had a probability of 100% to be browsed together. This indicate social media and entertainment site had a great

probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.3.3 Experiment three using afternoon office hour web user dataset

Rule:[URL21=accessed, URL33=accessed]: 8834 ==> [URL9=accessed]: 8834 <conf:(1)> lift:(2.71) lev:(0.13) conv:(5570.32)

According to above association rule, if www.facebook.com and www.yehabesha.com URL are browsed together during afternoon office hour, www.astu.edu.et organizational site had a probability of 100% to be browsed together.

This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.3.4 Experiment four using non office hour web user dataset

Rule: URL21=accessed, URL11=accessed]: 10175 ==> [URL9=accessed]: 10175 <conf:(1)> lift:(1.8) lev:(0.15) conv:(4516.83)

According to above association rule, if www.facebook.com and www.diretube.com URL are browsed together during non-office hour,www.astu.edu.et organizational site had a probability of 100% to be browsed together. This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.3.5 Experiment five using non weekend web user dataset

Rule: [URL33=accessed, URL34=accessed]: 10171 ==> [URL21=accessed]: 10171 <conf:(1)> lift:(3.36) lev:(0.11) conv:(7140.57)

According to above association rule, if www.yehabesha.com and www.youtube.com URL are browsed together during non-weekend time, www.facebook.com entertainment site had a probability of 100% to be browsed together.

This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.3.3.6 Experiment six using weekend web user dataset

Rule:[URL9=accessed, URL21=accessed]: 9653 ==> [URL1=accessed]: 9653 <conf:(1)> lift:(2.38) lev:(0.14) conv:(5605.08)

According to above association rule, if www.astu.edu.et and www.facebook.com URL are browsed together during weekend time, textual learning materials educational site had a probability of 100% to be browsed together. This indicate social media and entertainment site had a great probability to be accessed in the morning office hour. Totally the result of association rule assure the result of statistical analysis.

5.4 Discussion and explanation

This section gives a concise discussion about the results that acquired from both statistical analysis and data mining technique experiments and how those experiments are done.

In statistical analysis technique experiment within six categories of experiment the figure show that social media and entertainment site are most frequently are browsed with six class of experiments. Even though most of the users are accessed social media and entertainment site with in all class of experiment, education sit are also browsed with by least figure next to social media and entertainment site within morning and afternoon office hour class of experiment.

As association rule mining techniques experiment show that in each six class of experiment social media and entertainment site are had a great probability to be browsed together relative to educational and organizational sites. Even if the correlation of social media and entertainments sites are frequented together more than other sites, educational sites are also associated with the social media and entertainments within morning and afternoon office hour class of experiments. In both result of statistical analysis and data mining techniques shows that most of the web user of Adama Science and Technology University most frequently accessed social media and entertainment sites that encompasses Facebook, dire tube, habesha.com, and entertainment YouTube URLs.

This study were answered the research question as for the first research question since the question asks using way how describe user web navigational behaviors which answered from the way understand the problem up to experimental result.

For the second research question the imperative attribute to discover user web navigational behaviors URL (the page that accessed by the user), access time of the URL, and derived attributes time interval value of this attribute (morning office hour, lunch time, afternoon office hour, non-office hour, non-weekend, and weekend) are identified. In addition to this for second research question two association rule mining algorithm such as Apriori and FP-Growth algorithm were selected.

As stated by [10] data mining tasks can be classified into descriptive data mining and predictive data mining. Predictive data mining like classification used to constructs one, or a set of, models, performs inference on the available set of data, and attempts to predict the behavior of new data sets. Based on this previous explanation classification rule mining is not important for the researcher work, since the researcher objective is not concerned on predictive data mining techniques. Descriptive data mining such as clustering, association rule mining, used to describes the data set in a concise and summary manner and presents interesting general properties of the data. Since the researcher objective focused descriptive data mining technique the researcher select association rule mining technique based describe the frequent page with classified time interval.
Then in this study, two association rule mining algorithm such as Apriori and FP-Growth algorithm to tested with preprocessed web log data of Adama Science and Technology University web users. Since predictive priori and filtered associated algorithm unable generate any rule, to discover association rule within all class of datasets within all experiments the researcher used Apriori algorithm and FP-Growth algorithm. Both Apriori algorithm and FP-Growth algorithm are generate better interesting rule to describe user web navigational behaviors of ASTU web users by encouraging the results of statistical analysis technique experiment. But Apriori algorithm takes long running time and display memory heap size error.
On the other hand FP-Growth algorithm is have better performance in terms of running time and small size requirement than Apriori algorithm. Due to its better performance capacity in terms of running time and error freeness FP-Growth algorithm is selected for association rule discovery in this thesis.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

In this study, the researcher attempted to apply statistical and data mining techniques on Adama Science and Technology University web users to describe users' web navigational behavior. The general approach used in this research is hybrid knowledge discovery methodology, which encompass six phases, such as understanding of the problem, understanding of the data, data preparation, mining user behaviors, evaluation the discovered knowledge, and use the discovered knowledge.

The raw log data preparation was a challenging and time intensive task of the research. Because of Web log files contain non-human requests such as robots, indexers, cookies and spiders, requests such as images, icons, style sheets which are not significant for analysis purpose. For the intention of describe user navigational behavior web resources by finding rules or patterns in web usage data, the researcher only interested in the URL and at what time the resource were accessed. This step deals with were performing accuracy check; transforming the data into categorized phases of dataset that were ready made for the experiment. After preprocessing had completed, several categories of experiments were conducted using statistical and data mining technique in the intention of extract interesting rules and patterns from the user web log record of Adama Science and Technology University web users.

Datapreparator-1.7 tool were applied on preprocessed log dataset to perform statistical analysis in order to explore general statistics, about most frequently visited URL, more frequent class of site used more frequent transaction day, Ways of access , most operating system used, most frequent browser used, more web resource were accessed by time by day. Apriori and FP Growth association rule algorithm of data mining technique using Weka 3.7.4 tools was used to describe user navigational behavior web resources.

For both statistical and data mining techniques the researcher based on the advice of domain experts categorized the experiment into six categories such as experiment one using morning office hour web user log data, experiment two using lunch time web user log data, experiment three using afternoon office hour web user log data, experiment four using non office hour web user log data, experiment five using non weekend web user log data, experiment six using weekend web user log data.

This research, attempted to answer the stated research questions, address the problem that stated in the statement of the problem and had achieved the objectives of the study. In this research, the web server log data that is used for pattern discovery prepared effectively by using different tools such as log file viewer to remove irrelevant record which have file extension such as .jpg, .pang, .gif, .axd, .svc, and the rest preparation were done by data preparatory tool and MS- excel by removing, record that is not accessed by human such as spiders, crawlers, robots cleaned.

Generally in both result of statistical analysis and data mining techniques shows that most of the web user of Adama Science and Technology University frequently were browsed site were social media and entertainment that comprises face book, dire tube, habesha.com, and entertainment you tube are the dominate site.

6.2 **Recommendation**

Adama Science and Technology University aims to produces educated manpower for the country who are equipped by technological experiences, on researching environmental problems and opportunity. To achieve this, it is important to understand problem for education quality.

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

- ✓ Adama Science and Technology University ICT should have to design policy for the way how the University community use web resources.
- ✓ Before prohibiting the different entertainment and social media site like Facebook and YouTube, yehabesha.com, Adama Science and Technology University ICT should have to do best awareness creation using researchers, training sephsium on the area.

The following points are recommended for interested future researchers in web usage mining area.

- ✓ This study is done by concentrating what (URL) is accessed and at what time the web resources is accessed. But for further researcher if depend on what (URL) is accessed, time web resource is accessed and addition who are accessed it (specific userIP Address) of Adama Science and Technology University web users by using Virtual Local Area Network(VLAN) Adama Science and Technology University ICT office can generate more efficient result.
- ✓ Identifying users by proxies and client cookies is one of the potential research area that improve the performance of the web usage mining study. Hence, feature researcher better to use this technique.

References

[1] C. kumar and P. Bhargavik, "Analysis Of Web Server Log By Web Usage Mining For Extracting Users Patterns," International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR),, vol. III, no. 2, pp. 123-136, june 2013.

[2] L. Jinguang and D. Roopa, "Web Usage Mining: Pattern Discovery and its applications," in IEEE conference, New York, 2014.

[3] S. Nanhay, J. Achin and S. Ram, "Comparison Analysis Of Web Usage Mining Using Pattern Recognition Techniques," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. III, no. 4, pp. 37-147, July 2013.

[4] P. Huiping, "Discovery of Interesting Association Rules Based On Web Usage Mining," in IEEE conference, New York, 2010.

[5] J. Gyu, "Official Web Site of Adama Science and technology university," Adama Science and technology university, 12 June 2013. [Online]. Available: http://www.astu.edu.et/. [Accessed 24 December 2014].

[6] T. Yosef, Interviewee, System Administrtor of Adama Science and Technology University. [Interview]. 28 December 2014.

[7] T. Mekonnen, "Web usage pattern discovery using data mining and statistical analysis: the case of AAU official web site," M.S Thesis, in department of information science, Addis Ababa University, 2009.

[8] T. Asitatikie, "Web usage pattern discovery: the case of Addis Ababa university official web site," M.S thesis, Department of Information Science, Addis Ababa University, 2011.

[9] A. Fesseha, "Web Usage: Exploring Navigational Behavior Of Users The Case Of The Official Web Site Of Addis Ababa University," M.S thesis, Department of Information Science, Addis Ababa University, 2011.

[10] U. Fayyad, P. Gregory and S. Padhraic, "From data mining to knowledge discovery in databases," in American Association for Artificial Intelligence, USA, 1996.

[11] R. Cooley, B. Mobasher and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," in In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, New York, 1997.

[12] M. Zhongming, "Web mining for knowledge discovery," Ph.D dissertation, Department of Business Administration, University of Utah, 2007.

[13] A.Murat, "A New Reactive Method For Processing Web Usage Data," M.S Thesis, Department Of Computer Engineering, Middle East Technical University, 2006.

[14] H. Yilmaz, "Using ontology based web usage mining and object clustering for recommendation," M.S thesis, Department of computer engineering, Middle East Technical University, 2010.

[15] G. Sule, "Recommendation Models For Web Users: User Interest Model And Click-Stream Tree," Ph.D thesis, in department of Computer Engineering, Istanbul Technical University Institute Of Science And Technology, 2003.

[16] S. Kumar and K. Rukmani, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms," in IEEE conference, New York, 2010.

[17] R. Shawkat, "Mining Client Side Para-Data for Adaptive Webpages," M.S thesis, Department of computer engineering, Jordan University of Science and Technology, 2011.

[18] M. Behzad, "Discovering and mining user web page traversal patterns," M.S thesis, in department of computer science, Simon Fraser University, 2001.

[19] W. Robert, "Web usage mining: discovery and application of interesting pattern from web data," Ph.D dissertation, faculty of Graduate School of , University of Minnesota, 2000.

[20] A. Sharma, "Web Usage Mining: Data Preprocessing, Pattern Discovery And Pattern Analysis On The Rit Web Data," M.S thesis, Deparment of Computer engineering, Rochester Institute of Technology Rochester, 2008.

[21] N. Goel and K. Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool," International Journal of Computer Applications, vol. III, no. 6, pp. 145-186, 2013.

[22] n.d, "Data preparator services," [Online]. Available: http://www.datapreparator.com.

[23] S. Padmaja and A. Sheshasaayee, "Web server logs to analyzing user behavior using log analyzer tooL," International Journal of Advance Research In Science And Engineering, vol. III, no. 1, pp. 514-525, 2014.

[24] W. Yan, "Web Mining and Knowledge Discovery of Usage Patterns," in IEEE conference,, New York, 2000.

[25] M. Federico and L. Pier, "Recent developments inWeb Usage Mining Research," in Artificial Intelligence and Robotics Laboratory Dipartimento di Elettronica, USA, 2000.

[26] C. Lukas, S. Myra and W. Karsten, "A data miner analyzing web navagtioanl behavior of web users," in Institut for ur wirtschaftsinformatik, Humboldt-Universit, Berlin, n.d.

[27] G. Sulu, "Recommendation Model For Web Users," in User Interest Model And Click Stream Tree, Istanbul technical university, 2003.

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

[29] k. Arvind and p. Gupta, "Analysis of web server log files to increase the effectiveness of the website using web mining tool," in IEEE conference, Berlin, 2013.

[30] K. Wei, "Exploring Health Website Users by Web Mining," M.S thesis, in department of Health Informatics, Indiana University,, 2012.

[31] S. Anand, "Data Mining of Web Access Logs," M.S thesis, in department of Information Technology, Royal Melbourne Institute of Technology Melbourne, Victoria, Australia, 2003.

[32] K. Cios and W.Pedrycz, "The Knowledge Discovery Process," in IEEE conference, Berlin, 2007.

[33] K. Raymond and B. Hendrik, "Web Mining Research," A Survey," SIGKDD Explorations, vol. II, no. 1, pp. 1-20, 2000.

[34] M. Kumar and N. Aggarwal, "Web usage mining: an analysis," journal of emerging technologies in web intelligence, vol. V, no. 3, pp. 240-246, 2013.

[35] R. Cooley and S. Arvind, "Data preparation for mining world wide web browsing patterns," Journal of Knowledge and Information Systems, vol. I, no. 1, pp. 15-32, 1999.

[36] K. Mukesh and R.Aggarwal, "Mining association between sets of items in massive database," in In: Proc the ACM-SIGMOD International Conference on Management of Data, Berlin, 1993.

Appendices

Appendix (A): list of selected attribute for association rule discovery

URL code	URL Description
URL1	textual learning materials
URL2	www.aau.edu.et
URL3	www.adamacity.gov.et
URL4	www.addiszefen.com
URL5	www.aljazeera.com
URL6	www.amharictube.com
URL7	www.amra.gov.et
URL8	www.articles.org
URL9	www.astu.edu.et
URL10	www.catholic.org
URL11	www.diretube.com
URL12	www.educational tutorial.com
URL13	www.dstv.com
URL14	www.ecsu.edu.et
URL15	www.erca.gov.et
URL16	www.ertagov.com
URL17	www.ethiopianorthodox.org
URL18	www.ethiopianreporter.com
URL19	www.ethiotelecom.et
URL20	www.ethiotube.net
URL21	www.facebook.com
URL22	www.free-ebooks.net
URL23	www.gmail.com
URL24	www.moin.gov.et
URL25	www.most.gov.et
URL26	www.nbe.gov.et
URL27	www.onlain-porno.tv
URL28	www.oromia.gov.com
URL29	www.scholarships.com
URL30	www.skype.com
URL31	www.twitter.com
URL32	Download software
URL33	www.yehabesha.com
URL34	www.youtube.com

Appendix (B): Weka Association rule discovery outputs using Apriori Algorithm
Appendix B (I)
5.3.2.1 Experiment one using morning office hour web user dataset
=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: transformed morning office hour data set
Instances: 34220
Attributes: 24
URL1
URL3
URL4
URL5
URL6
URL8
URL9
URL11
URL12
URL17
URL18
URL20
URL21
URL22
URL23
URL26
URL27
URL28
URL29
URL30
URL31
URL32
URL33
URL34
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.5 (3422 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 6
Size of set of large itemsets L(3): 4
Size of set of large itemsets L(4): 1
Best rules found:
1. URL9=accessed URL33=accessed 7679 ==> URL1=accessed 7679<conf:(1)>lift:(3.47) lev:(0.16) [5463] conv:(5463.71)

2. URL11=accessed URL33=accessed 4850 ==> URL1=accessed 4850 <conf:(1)> lift:(3.47) lev:(0.1) [3450] conv:(3450.84)
3. URL9=accessed URL11=accessed 4811 ==> URL1=accessed 4811 <conf:(1)> lift:(3.47) lev:(0.1) [3423] conv:(3423.09)

Appendix B (II)
5.3.2.2 Experiment two using lunch time web user dataset
=== Run information ===
Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: tranformed lunch time data set
Instances: 20689
Attributes: 23
URL1
URL5
URL6
URL8
URL9
URL11
URL12
URL13
URL14
URL15
URL17
URL18
URL20
URL21
URL22
URL23
URL28
URL29
URL30
URL31
URL32
URL33
URL34
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.5 (5172 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 10
Size of set of large itemsets L(3): 10

Size of set of large itemsets L(4): 5
Size of set of large itemsets L(5): 1
Best rules found:

1. URL33=accessed URL34=accessed 6011 ==> URL21=accessed 6011 <conf:(1)> lift:(2.51) lev:(0.17) [3618] conv:(3618.11)
2. URL21=accessed URL33=accessed 6011 ==> URL34=accessed 6011 <conf:(1)> lift:(3.44) lev:(0.21) [4261] conv:(4261.94)
3. URL33=accessed URL34=accessed 6011 ==> URL22=accessed 6011 <conf:(1)> lift:(2.31) lev:(0.16) [3405] conv:(3405.43)
4. URL22=accessed URL33=accessed 6011 ==> URL34=accessed 6011 <conf:(1)> lift:(3.44) lev:(0.21) [4261] conv:(4261.94)
5. URL22=accessed URL33=accessed URL34=accessed 6011 ==> URL21=accessed 6011 <conf:(1)> lift:(2.51) lev:(0.17) [3618] conv:(3618.11)

Appendix B (III)
5.3.2.3 Experiment three using afternoon office hour web user dataset
=== Run information ===
Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: tranformed afternoon office hour data set
Instances: 41519
Attributes: 29

URL1
URL2
URL4
URL5
URL6
URL8
URL9
URL10
URL11
URL12
URL15
URL16
URL17
URL18
URL19
URL20
URL21
URL22
URL23
URL24
URL25
URL27
URL28
URL29
URL30

URL31
URL32
URL33
URL34

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

Apriori

=======

Minimum support: 0.5(8304 instances)

Minimum metric <confidence>: 0.75

Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 10

Size of set of large itemsets L(4): 5

Size of set of large itemsets L(5): 1

Best rules found:

1. URL21=accessed URL22=accessed 8834 ==> URL9=accessed 8834 <conf:(1)> lift:(2.71) lev:(0.13) [5570] conv:(5570.32)
2. URL9=accessed URL21=accessed 8834 ==> URL22=accessed 8834 <conf:(1)> lift:(2.8) lev:(0.14) [5676] conv:(5676.49)
3. URL21=accessed URL33=accessed 8834 ==> URL9=accessed 8834 <conf:(1)> lift:(2.71) lev:(0.13) [5570] conv:(5570.32)
4. URL9=accessed URL33=accessed 8834 ==> URL21=accessed 8834 <conf:(1)> lift:(2.64) lev:(0.13) [5489] conv:(5489.68)

Appendix B (IV)

5.3.2.4 Experiment four using non office hour web user dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: tranformed non office hour data set

Instances: 29999

Attributes: 24

URL1
URL5
URL6
URL7
URL8
URL9
URL11
URL12
URL13
URL14
URL15
URL17
URL18
URL20

URL21
URL22
URL23
URL28
URL29
URL30
URL31
URL32
URL33
URL34

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

Apriori

=======

Minimum support: 0.5 (9000 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 14
Generated sets of large itemsets:
Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 15
Size of set of large itemsets L(3): 20
Size of set of large itemsets L(4): 15
Size of set of large itemsets L(5): 6
Size of set of large itemsets L(6): 1
Best rules found:

1. URL11=accessed URL21=accessed 10175 ==> URL9=accessed 10175 <conf:(1)> lift:(1.8) lev:(0.15) [4516] conv:(4516.83)
2. URL9=accessed URL21=accessed 10175 ==> URL11=accessed 10175 <conf:(1)> lift:(2.44) lev:(0.2) [6009] conv:(6009.56)
3. URL9=accessed URL11=accessed 10175 ==> URL21=accessed 10175 <conf:(1)> lift:(2.19) lev:(0.18) [5529] conv:(5529.96)
4. URL11=accessed URL33=accessed 10175 ==> URL9=accessed 10175 <conf:(1)> lift:(1.8) lev:(0.15) [4516] conv:(4516.83)
5. URL9=accessed URL11=accessed 10175 ==> URL33=accessed 10175 <conf:(1)> lift:(2.92) lev:(0.22) [6696] conv:(6696.05)

Appendix B (V)

5.3.2.5 Experiment five using non weekend web user dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: tranformed non weeked data set
Instances: 65535
Attributes: 32
URL1
URL3
URL4
URL5

URL6
URL8
URL9
URL10
URL11
URL12
URL13
URL14
URL15
URL16
URL17
URL18
URL19
URL20
URL21
URL22
URL23
URL24
URL25
URL26
URL27
URL28
URL29
URL30
URL31
URL32
URL33
URL34

=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.5 (9830 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 17
Generated sets of large itemsets:
Size of set of large itemsets L(1): 7
Size of set of large itemsets L(2): 14
Size of set of large itemsets L(3): 16

Size of set of large itemsets L(4): 9
Size of set of large itemsets L(5): 2
Best rules found:

1. URL21=accessed URL22=accessed URL34=accessed 11053 ==> URL11=accessed 11053 <conf:(1)> lift:(3.77) lev:(0.12) [8119] conv:(8119.7)
2. URL11=accessed URL22=accessed URL34=accessed 11053 ==> URL21=accessed 11053 <conf:(1)> lift:(3.36) lev:(0.12) [7759] conv:(7759.78)

3. URL11=accessed URL21=accessed URL34=accessed 11053 ==> URL22=accessed 11053 <conf:(1)> lift:(3.59) lev:(0.12) [7970] conv:(7970.94)
4. URL22=accessed URL34=accessed 11053 ==> URL11=accessed URL21=accessed 11053 <conf:(1)> lift:(5.38) lev:(0.14) [8998] conv:(8998.91)
5. URL21=accessed URL34=accessed 11053 ==> URL11=accessed URL22=accessed 11053 <conf:(1)> lift:(5.08) lev:(0.14) [8877] conv:(8877.14)

Appendix B (VI)

5.3.2.6 Experiment six using weekend web user dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: weeked web user dataset
Instances: 40201
Attributes: 20
URL1
URL2
URL4
URL6
URL9
URL11
URL12
URL17
URL18
URL20
URL21
URL22
URL23
URL26
URL28
URL29
URL30
URL31
URL33
URL34

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

Apriori
=======
Minimum support: 0.5 (8040 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 16
Generated sets of large itemsets:
Size of set of large itemsets L(1): 7
Size of set of large itemsets L(2): 20
Size of set of large itemsets L(3): 30
Size of set of large itemsets L(4): 25
Size of set of large itemsets L(5): 11
Size of set of large itemsets L(6): 2

Best rules found:
1. URL1=accessed URL9=accessed 12962 ==> URL34=accessed 12962 <conf:(1)> lift:(2.82) lev:(0.21) [8362] conv:(8362.22)
2. URL9=accessed URL22=accessed 10042 ==> URL1=accessed 10042 <conf:(1)> lift:(2.38) lev:(0.15) [5830] conv:(5830.96)
3. URL1=accessed URL22=accessed 10042 ==> URL9=accessed 10042 <conf:(1)> lift:(2.18) lev:(0.14) [5431] conv:(5431.29)
4. URL22=accessed URL34=accessed 10042 ==> URL1=accessed 10042 <conf:(1)> lift:(2.38) lev:(0.15) [5830] conv:(5830.96)
5. URL1=accessed URL22=accessed 10042 ==> URL34=accessed 10042 <conf:(1)> lift:(2.82) lev:(0.16) [6478] conv:(6478.43)

Appendix (C): Weka Association rule discovery outputs using FP-Growth Algorithm
Appendix C (I)
5.3.3.1 Experiment one using morning office hour data set
== Run information ===
Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1
Relation: transformed morning office hour data set
Instances: 34220
Attributes: 24
URL1
URL3
URL4
URL5
URL6
URL8
URL9
URL11
URL12
URL17
URL18
URL20
URL21
URL22
URL23
URL26
URL27
URL28
URL29
URL30
URL31
URL32
URL33
URL34
=== Associator model (full training set) ===
FPGrowth found 18 rules (displaying top 10)

1. [URL9=accessed, URL33=accessed]: 7679 ==> [URL1=accessed]: 7679 <conf:(1)> lift:(3.47) lev:(0.16) conv:(5463.71)
2. [URL9=accessed, URL11=accessed]: 4811 ==> [URL1=accessed]: 4811 <conf:(1)> lift:(3.47) lev:(0.1) conv:(3423.09)
3. [URL33=accessed, URL11=accessed]: 4850 ==> [URL1=accessed]: 4850 <conf:(1)> lift:(3.47) lev:(0.1) conv:(3450.84)
4. [URL9=accessed, URL33=accessed, URL11=accessed]: 4809 ==> [URL1=accessed]: 4809 <conf:(1)> lift:(3.47) lev:(0.1) conv:(3421.67)

Appendix C (II)
5.3.3.2 Experiment two using lunch time web user dataset
=== Run information ===
Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1
Relation: tranformed lunch time data set
Instances: 20689
Attributes: 23
URL1
URL5
URL6
URL8
URL9
URL11
URL12
URL13
URL14
URL15
URL17
URL18
URL20
URL21
URL22
URL23
URL28
URL29
URL30
URL31
URL32
URL33
URL34
=== Associator model (full training set) ===
FPGrowth found 135 rules (displaying top 20)
1. [URL34=accessed]: 6020 ==> [URL22=accessed]: 6020 <conf:(1)> lift:(2.31) lev:(0.16) conv:(3410.53)
2. [URL11=accessed]: 5993 ==> [URL22=accessed]: 5993 <conf:(1)> lift:(2.31) lev:(0.16) conv:(3395.23)

3. [URL11=accessed]: 5993 ==> [URL33=accessed]: 5993 <conf:(1)> lift:(2.44) lev:(0.17) conv:(3533.98)
4. [URL34=accessed]: 6020 ==> [URL21=accessed]: 6020 <conf:(1)> lift:(2.51) lev:(0.18) conv:(3623.52)
5. [URL11=accessed]: 5993 ==> [URL21=accessed]: 5993 <conf:(1)> lift:(2.51) lev:(0.17) conv:(3607.27)

Appendix C (III)
5.3.3.3 Experiment three using afternoon office hour web user dataset
=== Run information ===
Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1
Relation: tranformed afternoon office hour data set
Instances: 41519
Attributes: 29
URL1
URL2
URL4
URL5
URL6
URL8
URL9
URL10
URL11
URL12
URL15
URL16
URL17
URL18
URL19
URL20
URL21
URL22
URL23
URL24
URL25
URL27
URL28
URL29
URL30
URL31
URL32
URL33
URL34
=== Associator model (full training set) ===
FPGrowth found 135 rules (displaying top 20)

1. [URL33=accessed]: 8834 ==> [URL21=accessed]: 8834 <conf:(1)> lift:(2.64) lev:(0.13) conv:(5489.68)
2. [URL34=accessed]: 8830 ==> [URL21=accessed]: 8830 <conf:(1)> lift:(2.64) lev:(0.13) conv:(5487.19)
    1. [URL21=accessed, URL9=accessed]: 8834 ==> [URL22=accessed]: 8834 <conf:(1)> lift:(2.8) lev:(0.14) conv:(5676.49)
    2. [URL21=accessed, URL22=accessed]: 8834 ==> [URL9=accessed]: 8834 <conf:(1)> lift:(2.71) lev:(0.13) conv:(5570.32)
    3. [URL21=accessed, URL9=accessed]: 8834 ==> [URL33=accessed]: 8834 <conf:(1)> lift:(4.7) lev:(0.17) conv:(6954.39)
    4. [URL33=accessed]: 8834 ==> [URL21=accessed, URL9=accessed]: 8834 <conf:(1)> lift:(4.7) lev:(0.17) conv:(6954.39)
    5. [URL21=accessed, URL33=accessed]: 8834 ==> [URL9=accessed]: 8834 <conf:(1)> lift:(2.71) lev:(0.13) conv:(5570.32)

Appendix C (IV)

5.3.3.4 Experiment four using non office hour web user dataset

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
Relation: tranformed non office hour data set
Instances: 29999
Attributes: 24
URL1
URL5
URL6
URL7
URL8
URL9
URL11
URL12
URL13
URL14
URL15
URL17
URL18
URL20
URL21
URL22
URL23
URL28
URL29
URL30
URL31
URL32
URL33
URL34

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

FPGrowth found 478 rules (displaying top 10)
1. [URL9=accessed, URL22=accessed]: 10117 ==> [URL21=accessed]: 10117 <conf:(1)> lift:(2.19) lev:(0.18) conv:(5498.44)
2. [URL21=accessed, URL22=accessed]: 10117 ==> [URL9=accessed]: 10117 <conf:(1)> lift:(1.8) lev:(0.15) conv:(4491.09)
3. [URL9=accessed, URL21=accessed]: 10175 ==> [URL11=accessed]: 10175 <conf:(1)> lift:(2.44) lev:(0.2) conv:(6009.56)
4. [URL9=accessed, URL11=accessed]: 10175 ==> [URL21=accessed]: 10175 <conf:(1)> lift:(2.19) lev:(0.18) conv:(5529.96)
5. [URL21=accessed, URL11=accessed]: 10175 ==> [URL9=accessed]: 10175 <conf:(1)> lift:(1.8) lev:(0.15) conv:(4516.83)
6. [URL9=accessed, URL21=accessed]: 10175 ==> [URL33=accessed]: 10175 <conf:(1)>

Appendix C (V)

5.3.3.5 Experiment five using non weekend web user dataset

=== Run information ===
Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1
Relation: tranformed non weeked data set
Instances: 65535
Attributes: 32
URL1
URL3
URL4
URL5
URL6
URL8
URL9
URL10
URL11
URL12
URL13
URL14
URL15
URL16
URL17
URL18
URL19
URL20
URL21
URL22
URL23
URL24
URL25
URL26
URL27
URL28

URL29
URL30
URL31
URL32
URL33
URL34

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

FPGrowth found 203 rules (displaying top 10)

1. [URL21=accessed, URL33=accessed]: 10527 ==> [URL22=accessed]: 10527 <conf:(1)> lift:(3.59) lev:(0.12) conv:(7591.62)
2. [URL33=accessed, URL34=accessed]: 10171 ==> [URL21=accessed]: 10171 <conf:(1)> lift:(3.36) lev:(0.11) conv:(7140.57)
3. [URL21=accessed, URL33=accessed]: 10527 ==> [URL11=accessed]: 10527 <conf:(1)> lift:(3.77) lev:(0.12) conv:(7733.29)
4. [URL21=accessed, URL34=accessed]: 11053 ==> [URL22=accessed]: 11053 <conf:(1)> lift:(3.59) lev:(0.12) conv:(7970.94)
5. [URL22=accessed, URL34=accessed]: 11053 ==> [URL21=accessed]: 11053 <conf:(1)> lift:(3.36) lev:(0.12) conv:(7759.78)

Appendix C (VI)

5.3.3.6 Experiment six using weekend web user dataset

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1

Relation: weeked web user dataset

Instances: 40201

Attributes: 20

URL1
URL2
URL4
URL6
URL9
URL11
URL12
URL17
URL18
URL20
URL21
URL22
URL23
URL26
URL28
URL29
URL30
URL31
URL33
URL34

=== Associator model (full training set) ===
FPGrowth found 699 rules (displaying top 10)
1. [URL9=accessed, URL22=accessed]: 10042 ==> [URL1=accessed]: 10042 <conf:(1)> lift:(2.38) lev:(0.15) conv:(5830.96)
2. [URL1=accessed, URL22=accessed]: 10042 ==> [URL9=accessed]: 10042 <conf:(1)> lift:(2.18) lev:(0.14) conv:(5431.29)
3. [URL9=accessed, URL21=accessed]: 9653 ==> [URL1=accessed]: 9653 <conf:(1)> lift:(2.38) lev:(0.14) conv:(5605.08)
4. [URL1=accessed, URL21=accessed]: 9653 ==> [URL9=accessed]: 9653 <conf:(1)> lift:(2.18) lev:(0.13) conv:(5220.89)
5. [URL9=accessed, URL1=accessed]: 12962 ==> [URL34=accessed]: 12962 <conf:(1)> lift:(2.82) lev:(0.21) conv:(8362.22)