ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE PATTERN DISCOVERY: THE CASE OF ADDIS ABABA UNIVERSITY OFFICIAL WEB SITE

TADELE ASITATIKIE DAGNE

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE PATTERN DISCOVERY: THE CASE OF ADDIS ABABA UNIVERSITY OFFICIAL WEB SITE

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

TADELE ASITATIKIE DAGNE

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

WEB USAGE PATTERN DISCOVERY: THE CASE OF ADDIS ABABA UNIVERSITY OFFICIAL WEB SITE

By

Tadele Asitatikie Dagne

Name and signature of Members of the Examining Board

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

DEDICATION

This research work is dedicated to my sister Yigardu.

ACKNOWLEDGEMENT

First of all, I thank Almighty God for His everlasting shepherd.

Next, I would like to speak out my thanks to my advisor Dr. Rahel Bekele for her continuous follow up and guidance throughout the life of the research work. I have been strongly impressed by her constructive comments and guidance.

Then, I am happy to acknowledge staffs of AAU ICT office for their cooperation in data collection and data analysis of the research work.

I am glad to pass my heartfelt thanks to Temesgen Dagne for his unreserved and countless support in my tertiary education life as a whole.

I could hardly enumerate Mebratu's supports for my success. He has devoted himself for my achievement.

My thanks also go to Tena and Melese Tesfaye who have been initiating me to finalize my thesis work as soon as possible.

I would like to extend my thanks to all of my family members who have been supporting me in either finance, moral or *pray.*

To all others who helped me in completing this study, I am equally grateful.

Tadele Asitatikie

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ACRONYMS

AAU=Addis Ababa University

CGI= Common Gateway Interface

CRISP=CRoss-Industry for Standard Process

CRM=Customer Relation Management

DNS=Domain Name Service

FPG=Frequent Pattern Growth

FTP=File Transfer Protocol

HTTP=HyperText Transfer Protocol

ICT=Information Communication Technology

ID= Identification

IP= Internet Protocol

IT=Information Technology

JPEG/JPG= Joint Photographic Experts Group

KDD=Knowledge Discovery of Databases

MFR= Maximum Forward Reference

n.d= no date, i.e. the date is not available

PC=Personal Computer

PNG=Portable Network Graphics

RL= Reference Length

SDDPC= Systems Design and Data Processing Center

SSL= Secured Sockets Layer

URL=Uniform Resource Location

Weka = Waikato Environment for Knowledge Analysis

XML=eXtended Markup Language

ABSTRACT

Web sites are one of the well known communication media where very huge information could be disseminated for visitors' access. Since most of the organizations in the world use web sites to communicate their business, smooth running and better service of the web sites is fundamental point for the benefit of users and/or the organization.

In order to develop improved web site services, the extent to which users are satisfied on the web site should be computed to evaluate the site. On the first case, the site designers could bring amendments on the site by their expectation of users' desire. Secondly, users of the site may be communicated through interview, questionnaire or observation to gather the users' trend of the web site usage. Thirdly, researchers could undergo web usage analysis on web server log files of the users' requests.

In this research, server log of Addis Ababa University web site was utilized to discover web usage pattern using statistical analyzer and data mining applications. Access log file was a target data source as it is expected to show users' usage trends.

The methodology proposed by Srivastava et al. (2000) was used in the research work which comprised of raw log preparation, pattern discovery and pattern analysis. The well known association mining algorithm, Apriori, was used to extract the common directories that were accessed together. Meanwhile, PredictiveApriori was used in mining common navigational sequences in the web site. Then, the discovered rules were analyzed in consultation with the site map of the web site and the concerned site designers.

The research found that there are directories which are frequently visited in the web site. Other findings are also discussed based on which recommendations are forwarded.

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Addis Ababa University (AAU) is one of the accredited governmental higher education institutions in Ethiopia. It was founded in 1950 by the name of University College of Addis Ababa. Then, the university was renamed to Haile Selassie I University in 1962. The current name of AAU was given in 1975. Currently, the university delivers programs in undergraduate, graduate and postgraduate level in different fields of studies (Teshome 1990).

Formerly, a Systems Design and Data Processing Center (SDDPC), which was established in 1970, was used to offer batch-oriented data processing services to the central administration in such application areas like payroll and general ledger (Tesfaye n.d). The center also used to provide services to student/staff researchers in analyzing survey results using statistical packages. However, dissatisfaction on the effectiveness of the center was manifested from the user community by the survey of Tesfaye (n.d). Despite SDDPC's long existence in the institution, no administrative unit was provided with on-line access facilities to any data maintained by the center.

The same writers claimed that certain developments have been observed in the widespread penetration of IT in various academic and research units of AAU through time. A fairly large proportion of the introduced IT equipments, however, were used as stand-alone system to provide access to word processing applications and some other related office applications.

As an initiation of the quality information access, networking of computers was also to a certain extent exercised in AAU (Tesfaye n.d). Accordingly, some units have installed rudimentary Local Area Networks (LANs) to connect their personal computers (PCs) with a view of setting up computing centers to support teaching and research in their

respective units. Besides, the use and application of IT as a delivery vehicle for the curriculum were also emerging. This might be some indicator which could lead to the conclusion that networking was gaining a foothold in AAU (Tesfaye n.d).

Accordingly, large scale networking was desired for the intention of initiating and implementing Information Communication Technology (ICT) related projects and activities. So, a project called AAUNet was established in 2000 which initiated the opening of the ICT Development Office at AAU.

ICT Development Office is one of the offices in AAU that serves the university community. Among the duties and responsibilities of the office, developing and maintaining of the AAU official web site is the primary one. The web site was published in 2001 and it is hosted at the main campus of the university, Sidist Kilo. It is meant to serve both the university community and outsiders who want to surf the university's service(s).

Discussions with current ICT staffs of the university disclosed some facts about the development of the web site (A list of discussants is attached in Appendix G). Accordingly, individual academic and administrative units have been engaged in designing and submitting their respective pages after which the ICT Development Office publishes the pages to the home page of the official site. There have been cases that the site did not contain all information about the university services. Even the available pages are not consistent enough to attract customers. Users are used to feel less comfort on these aspects.

1.2. Statement of the Problem

Web sites provide information of the web on various aspects of users' or organizations' need. The information may include graphics, text, audio, video, or other multimedia data forms which could be accessed at real time of any distance. They are becoming very popular by the user community. This, in turn, tends to increase the web sites' growth in politics, academic, economic and other aspects. Ratheke and Schreiweis (2003) claimed

that information available in web pages is different from the information which is available in other media. First of all, users of web pages could satisfy their information need by interactively directing themselves to the information source they want. Secondly, a web user can skip to another source in few mouse clicks following the link structure. Thirdly, a user can view multiple sources simultaneously. Fourthly, content of most web sites is dynamic and varies from time to time to give users up-to-date information.

Until 15th of December 2010, about 125,620,840 active domains are registered globally (DomainTool web site 2010). In contrast to their dynamic growth, most web sites are not useful for most of the users. Han and Kamber (2006) claimed that 99% of the web sites are not useful for 99% of the users. There are several factors that contribute for this problem. As one of the factors, Lydon and Fennell (2003) argued that a large number of web sites are poorly designed because user requirements are not often incorporated into the web design process. The researchers added that web sites focus more on quantity rather than content quality that could suit the users' requirements. Thus, the effectiveness of a certain web site should be evaluated regularly so as to assure that it fits the need of users (Myra 2000).

One of the techniques used to evaluate web sites is web mining (Srivastava et al. 2000). It is used to discover and retrieve useful and interesting patterns from web dataset such as web log, web site content, users' registration and web site links. All of web data can be mined mainly in three different dimensions namely web content mining, web structure mining and web usage mining (Suneethe et al. 2009, Srivastava et al. 2000, Borges et al. 1999).

According to Cooley et al. (1997), each of the web mining approaches has its own application area. In this regard, web content mining involves efficient extraction of useful and relevant information from web sites and databases. It concerns about distinguishing documents which could better meet requirement of a certain individual. Web structure mining involves the techniques used to study web pages schema of a collection of hyper-links. It is commonly used to discover authoritative and hub pages of a certain web site.

Web usage mining is the application of data mining techniques to web log data in order to discover user access patterns.

Among the three dimensions of web mining, mining for usage pattern discovery demands web usage mining. In this approach, interesting patterns are discovered about web site users' access pattern. Evaluation of the existing web sites could then be accomplished on top of the discovered patterns. The analysis is expected to help web site designers to come up with some sort of design change that could better fit the users' need such as web pre-fetching, link prediction, site reorganization and web personalization (Srivastava et al. 2000).

From a preliminary investigation, users of AAU official web site are not satisfied by the site and the web site is being blamed in its services. As to the researcher's interview with communities of the university, complexities in using the site have been claimed. Discussion with ICT Development staffs also informed the researcher that there are users who have been claiming on the usage difficulty of the web site and failure to get the services of the web site. Moreover, broken page links have been claimed from the site users.

The site has been redesigned repeatedly since its establishment. As the researcher interviewed the web site design team members of the ICT Development Office, the redesigning task of the web site has not ever taken a solid input except simple discussions and recommendations internally. Indeed, the web site designers browsing expectation and the actual users' navigation patterns have not been taken into consideration. But, such inputs could not be fully trusted to be the right directions of designing better web site. This, in turn, could not better fit the site users and is costly practice on unreasonable frequent site changes. So, the usage patterns of the web site need to be analyzed for effective web site evaluation and recommendations for the web site redesign (Spiliopoulou 2000).

To address the problem, Mekonnen (2009) used web log analyzer tool and tried to analyze hit statistics, most requested pages, most visited directories, most frequent entry

and exit pages, users' visiting time and common errors encountered of the web site on a monthly and day of the week basis. Besides, he used Apriori based association rule mining to come up with some of the frequently accessed group of pages. Finally, he recommended to use combination of statistical analyzer tools and data mining approaches for discovery of better usage pattern.

However, his findings did not show common navigational sequences in which web pages are visited. Unlike association rule mining of market-basket analysis, a server session of usage log is an ordered sequence of pages requested by a user (Srivastava et al. 2000). So, interesting rules about common navigational sequences should be extracted for exploration of the appropriate usage trends which is the main interest of this work.

This research, therefore, focused on the usage pattern of the official web site of AAU using combinations of statistical analysis and data mining approaches.

1.3. Research Questions

Mekonnen (2009) claimed that most of the users of AAU official web site start and exit their navigation at the home page of the web site. The entrance through which web site users join a certain web site is informative about users' usage behavior toward the site. In case most of the web site users join a site via the home page, this could inform the researcher that either the site users use the exact URL to open the site or the query term that is used to search in search engine returns the home page of the site. Otherwise, the researcher could conclude that there are many alternative routes toward the web site.

Shen et al. (1999) and Batista and Silva (2002) revealed that discovery of sets of pages that are commonly accessed together is very interesting usage patterns of web sites. Also, Mekonnen (2009) showed that discovery of sets of pages that are frequently visited by users tends to reveal interesting usage pattern for the usage of AAU official web site.

Hu et al. (n.d) claimed that sequential pattern mining of web log data is one of the important areas in web usage analysis researches to discover common sequences of URL requests. Accordingly, the pages that are accessed in certain sequence frequently could be

rearranged in the discovered order so that the users' visiting time is saved. Thus, the web site could be easily visited by the users as its contents are arranged in a situation that meets the majority users' behavior.

As Srivastava et al. (2000) claimed, one of the goals of web usage pattern discovery is to compare the web site designers' expectation of web site navigation and the actual users' visit trends.

Thus, the research is expected to answer the following questions:

- Do most of the users of AAU official web site start navigation at the home page of the site or are there many arbitrary routes to its pages?
- Which pages are the most common pages on which users give up their navigation?
- Are there pages that are accessed together more frequently?
- Are the pages that are accessed together more frequently directly linked to each other in the web site?
- Are there frequent navigations in sequence but dispersedly located in the web site?
- What is the extent of match between web site designers' expectation of page navigations and actual users' navigation?

1.4. Objective of the Study

1.4.1. General Objective

The general objective of this study is to apply web data mining techniques on access log files of AAU official website to discover interesting usage patterns of the website that could be an input for evaluation of the website.

1.4.2. Specific Objective

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

- To discover frequent entry pages to AAU official web site;
- To find out more frequent exit pages of the web site;
- To distinguish pages that are visited together more frequently;
- To discover common navigational sequences in the web site;
- To analysis the site designers' page navigation expectation and actual users' navigation
- To forward appropriate recommendations as future research directions.

1.5. Scope and Limitations of the Study

In the study, access log file of AAU official web site was used as a potential data source. Besides, previous usage statistics was used in the log filtering tasks and site map of the web site was used in the pattern analysis. In the pattern discovery, association mining and sequential access mining techniques were employed.

The study was conducted on access log files of six months duration only.

1.6. Research Methodology

In order to discover better web usage patterns for a web site, there are a number of steps that need to be followed (Srivastava et al. 2000). Accordingly, the following steps were followed in this research.

a. Review of Related Literature

A review of relevant literature has 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 sites, particularly of official web site of AAU.

b. Identifying Available Data Sources

The potential source of data for this research was the web server of AAU official web site. The server log file contains different types of files and access log file was the focus in this research. The access log file contains records of each user requests that contains remote 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. Besides, the site map of the web site was used.

c. Data Preparation

The access log file is not suitable for direct input to the pattern discovery tool selected by the researcher, Waikato Environment for Knowledge Analysis (Weka). Thus, the raw log file was prepared step by step. Firstly, Wekaprep tool was used and the log file was filtered upon removal of non-human requests or automatically downloaded graphic files. Secondly, user session file was created by the same tool. Then, python code was written to develop transaction file of the user sessions and to transform the transaction file to input that are compatible to the respective algorithms.

d. Usage pattern discovery

Mach5 statistical analyzer tool was applied on the filtered log file to extract facts and figures that show general statistics on the site usage. Weka's association mining algorithm, Apriori, was used to discover the common directories that have been accessed together. Besides, PredictiveApriori algorithm was to discover the common navigation sequence in the web site.

e. Analysis of usage patterns

The discovered usage patterns were evaluated in consideration of the site map of the web site and the web site designers' navigation expectation.

1.7. Organization of the Thesis

The thesis is organized in five chapters. The first chapter illuminated on the problems that form the basis of this study. The general and specific objectives of the study as well as research questions were discussed.

In the second chapter, literature on data mining and web mining and their application in different areas are reviewed. Moreover, some researches related to web usage analysis were briefly presented.

The third chapter provided detailed process model of the research at hand. Each of the steps that were contained in the web usage analysis methodology proposed by Srivastava et al. (2000) was discussed.

The experimentation part of this research was discussed in chapter four. The different steps that are encompassed in the adopted methodology were described. The web usage patterns were discovered by integrating statistical analysis and data mining applications. Finally, interpretation of the experiments' results was given.

In the fifth chapter, 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

The natural evolution of information technology and phenomenal rate of data growth tend to bring wide availability of huge amount of data (David et al. 2001). Meanwhile, users need more sophisticated information and knowledge rather than accumulated data. This, in turn, demands the application of Knowledge Discovery in Databases-KDD (Fayyad et al. 1996).

KDD is the overall process of discovering useful knowledge from data. It encompasses lots of tasks from data selection for the mining task to interpret the discovered patterns. Data mining is one of the steps in the KDD process that is concerned with enumeration of patterns over target data via data analysis and discovery algorithms (Fayyad et al. 1996, Han and Kamber 2006, Hand et al. 2001). Data mining is different from accessing info in a database because the query is not well formed or precisely stated. Moreover, the data needs to be preprocessed before mining, and the output is new knowledge that may not be a subset of the database.

2.1.1. Data Mining Activities

There are various kinds of data mining techniques and algorithms depending on the type of knowledge to be mined out of the data. Fayyad et al. (1996) claimed that the two high-level primary goals of data mining in practice tend to be prediction and description. According to these researchers, prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. The goals of prediction and description can be achieved using a variety of data mining methods. In response to this, data mining research involves a combination of different activities which together solves a business problem.

Various problems of intellectual, economic, or business interests can be articulated in terms of different data mining activities such as data description, clustering (segmentation), concept description, classification, prediction (regression) or association analysis (Han and Kamber 2006, Fayyad et al. 1996). Some of these activities are briefed below.

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns are patterns that occur frequently in a data of consideration. According to Han and Kamber (2006), the three well known frequent patterns are itemsets, subsequences, and substructures. Accordingly, a frequent itemset typically refers to a set of items that frequently appear together in a transactional data set. This pattern is claimed to be the simplest form of frequent pattern mining. Sequence pattern, on the other hand, refers a frequently occurring subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card. The third frequent pattern, frequent substructure, refers to frequently occurring structural forms, such as graphs, trees, or lattices which may be combined with itemsets or subsequences. With the principles of frequent patterns mining, interesting associations and correlations within data are discovered.

Interestingness of association mining rules is commonly measured using support and confidence. Let $I = \{I_1, I_2 \ldots I_m\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \{\}$. The rule $A \rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., the *union* of sets A and B, or say, both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \rightarrow B$ has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(B|A)$. Mathematically, it is depicted as follow.

$support\ (A \rightarrow B) = P(A \cup B)$

$confidence\ (A \rightarrow B) = P(B \mid A)$

Support indicates the 'usefulness' of the rules whereas confidence indicates the certainty of the rule. Rules that satisfy both a minimum support threshold (*min sup*) and a minimum confidence threshold (*min conf*) are called strong.

For a certain transaction database of items and specified minimum support and minimum confidence threshold values, associations among items could be developed with an appropriate algorithm employing the following two generalized association steps.

- **Find all frequent itemsets**: At this step of the association mining, all frequent itemsets of all possible item size are distinguished. Then, each of these itemsets will occur at least as frequent as the predetermined minimum support count, *min sup*.
- **Generate strong association rules from the frequent itemsets**: Once all the frequent itemsets are known, associations of items that satisfy the minimum confidence threshold are identified. Then, rules are developed for the respective item associations. So, the items are claimed to satisfy both minimum support and minimum confidence thresholds.

Given an association rule: buys (X, "computer")→ buys (X,"antivirus_software") with its threshold values:

[Support=2%, confidence=60%], it is interpreted as follows.

A support of 2% means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

Apriori Algorithm

The Apriori algorithm has emerged as one of the best association rule mining algorithms (Han and Kamber 2006). The algorithm follows a complete, depth-first search to enumerate all frequent itemsets, i.e., it works based on data passes. It starts by scanning all transactions in the database and computing the frequent 1-itemsets. Next, a set of potentially frequent candidate 2-itemsets is formed from these frequent items. Then, another database scan is accomplished so as to calculate support values for each of the frequent candidate 2-itemsets out of which frequent 2-itemsets are distinguished. The frequent 2-itemsets are retained for the next pass, and the process is repeated until all frequent itemsets have been enumerated.

The algorithm has three main steps:

- Generate candidates of length k from the frequent $(k-1)$ length itemsets, by a self-join on F_k–1. For example, for $F2 = \{AC, AT, AW, CD, CT, CW, DW, TW\}$, we get $C3 = \{ACT, ACW, ATW, CDT, CDW, CTW\}$.
- Prune any candidate that has at least one infrequent subset. For example, *CDT* will be pruned because *DT* is not frequent.
- Scan all transactions to obtain candidate supports.

Apriori stores the candidates in a hash tree for fast support counting. In a hash tree, itemsets are stored in the leaves; internal nodes contain hash tables (hashed by items) to direct the search for a candidate. It prunes those candidates that have an infrequent subset before counting their supports. This optimization becomes possible because Best First Search (BFS) ensures that the support values of all subsets of a candidate are known in advance.

Despite the fact that the size of candidate frequent itemsets is reduced using "Apriori property", the Apriori algorithm is challenged by two nontrivial computationally expensive processes, i.e. so many database scans and huge number of candidates' generation.

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 thereof.

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.

- 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-itemsets 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.

Mining Sequential Patterns

Sequential pattern mining was first introduced by Agrawal and Srikant (1995) based on their study of customer purchase sequences:Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of a set of items, and given a user-specified minimum support threshold of min sup, sequential pattern mining finds all frequent subsequences, that is, the subsequences whose occurrence frequency in the set of sequences is no less than min sup.

Let I = $\{I_1, I_2 \ldots I_p\}$ be *the total* number of items in a transaction. An itemset is an unordered list of items from I. Let $e=(x_1x_2x_3\ldots x_n)$ be an event (itemset) containing n number of items. Then, $s=<e_1e_2e_3\ldots e_q>$ could be a sequence of q number of events (an q-

sequence) where each of the e_is is an element (itemset) of the sequence. The sequence database of all the sequences could be put as S=<SID, s> where SID is sequence identifier (ID) and s is the respective sequence.

Sequential pattern mining has gradually become an important data mining task, with broad applications in market and customer analysis, web log analysis, and mining Extended Markup Language (XML) query access patterns (Hu et al. n.d).

Mining Prediction Problems[1]

As Han and Kamber (2006) explained, classification and prediction are two forms of data analysis that can be used to extract models describing important data classes and to predict future data trends respectively. In the case of classification, an already established categorical label (discrete and unordered) is associated for a data tuple of investigation whereas prediction is about forecasting of an ordered and continuous *predicted attribute*[2] for a tuple. Both classification and prediction follow two steps: training and testing. However, the training data set could not be used in the testing phase of prediction (Han and Kamber 2006).

Decision tree Induction, Bayesian Networks, Support Vector Machines, Rule-Based Algorithms and Neural Network Learning are some of techniques for classification modeling. Regression analysis[3] is one type of numeric predictions. Besides, some of the classification techniques such as support vector machines may be adapted to work for prediction.

Mining Clustering Analysis

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters (Han and Kamber 2006). The process of grouping a set of physical or abstract objects into classes of similar objects is

[1] Classification and prediction are prediction problems; prediction is synonymous with 'numeric prediction' (Han et al. ,2006)
[2] Predicted attribute refers that a continuous attribute value of a predictor
[3] Regression analysis is a statistical methodology that is most often used for numeric prediction

called clustering. Contrary to classification and prediction, clustering analysis employs unsupervised learning because the clusters have not been set in advance. Clustering is also called data segmentation as clustering partitions large data sets into groups according to their similarity.

There are lots of clustering methods which could be grouped as Partitioning, Hierarchical, Density-based, Grid-based, and Model-based.

2.1.2. Knowledge Discovery Process

The discovery of hidden knowledge from dataset is done following some set of standards. There are lots of methodologies for the realization of this task such as Cross Industry Standard Process for Data Mining (CRISP-DM), Sample, Explore, Modify, Model, and Assess (SEMMA) and KDD.

2.1.2.1. CRISP-DM

This data mining methodology is mostly followed for industry researches. There are six commonly agreed steps in this process.

a. **Business understanding**: This initial phase focuses on understanding the research objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

b. **Data understanding**: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

c. **Data preparation**: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks of data preparation include cleaning, attribute selection as well as transformation and cleaning of data for modeling tools.

d. Modeling: In this phase, various modeling techniques are selected and applied; then, their parameters are calibrated to optimal values. Typically, there are several techniques for a certain data mining problem area. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

e. Evaluation: At this stage of the research, one has to build a model (or models) that appear(s) to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly and review the steps executed to construct the model to be certain that it properly achieves the business objectives. The objective of evaluation is to determine if there are some important business issues that have not been sufficiently addressed. At the end of this phase, a decision on the use of the data mining results could be reached.

f. Deployment: Creation of the model is generally not the end of research works. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision making processes. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across enterprises. In many cases, it is the user-not the data analyst-who carries out the deployment steps. Even if the analyst will not carry out the deployment effort, it is important for the user to understand clearly what actions need to be carried out in order to actually make use of the created models.

The model is shown in the next page diagrammatically.

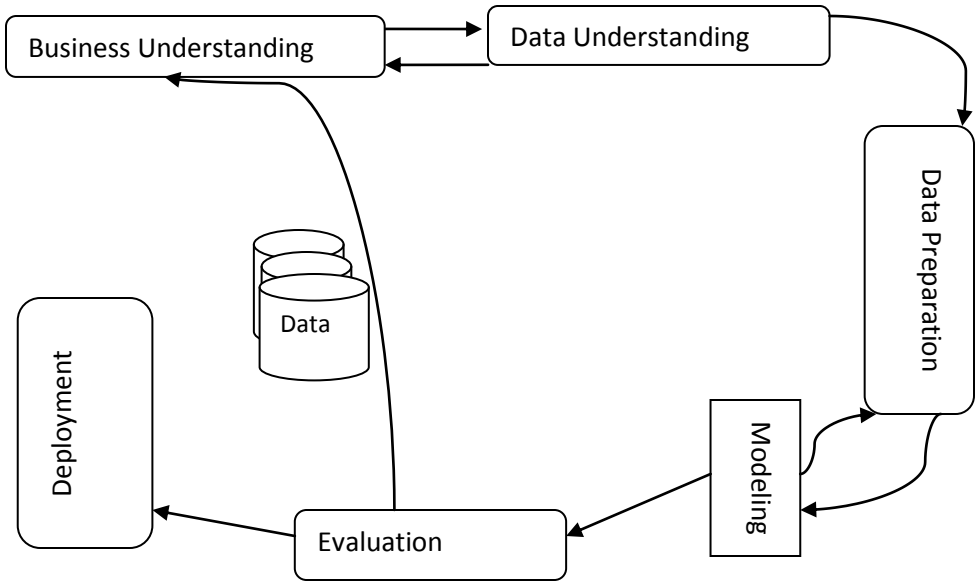


Figure 2. 1: Phases of CRISP-DM Process Model

(Source: CRISP-DM Consortium 2000)

2.1.2.2. SEMMA Process

SEMMA process refers to the process of conducting data mining research with five stages of sampling, exploring, modifying, modeling and assessing.

a. **Sample**: It concerns sampling the data by extracting a portion of a large data set that is big enough to encompass significant information of the dataset, yet small enough to manipulate quickly.
b. **Explore**: It consists of the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
c. **Modify**: It consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
d. **Model**: It concerns about modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
e. **Assess**: It concerns about assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

2.1.2.3. KDD Process

KDD process is comprised of the following five tasks.

a. **Selection**: It involves collecting data from possible sources to get the target data

b. **Preprocessing**: It concerns with cleaning the target data so as to get preprocessed data.

c. **Transformation**: It is all about any sort of rearrangements of the preprocessed data so that it would be easy to apply the data mining tool.

d. **Data mining**: It involves applying appropriate data mining techniques and algorithms onto the transformed data to get interesting patterns of the data.

e. **Interpretation/Evaluation**: At this stage of the KDD process, the discovered pattern is evaluated to give recommendations.

The diagrammatic representation of the KDD process is depicted below.

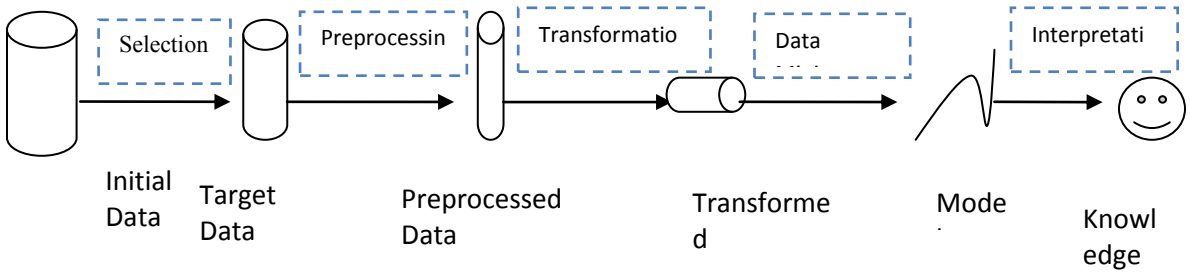


Figure 2.2: Phases of KDD Process

(Source: Dunham 2002)

2.1.3. Application of Data Mining Technology

Data mining has so many applications in reduced costs of doing business, improved profitability, or enhanced quality of service for different fields. Areas in which such benefits have been demonstrated include insurance, direct mail marketing, telecommunications, retail, medicine, customer relationship management and others (Huang 2003). Data mining applications have been shown to be highly effective in addressing many important business problems. Each business is interested in knowing the behavior of its customers on behalf of knowledge gained in data mining.

Customer Relation Management (CRM) is one of the areas where data mining is highly applicable. Henock (2002) attempted to study the application of clustering techniques to support CRM at Ethiopian airlines using CRISP-DM methodology. The researcher mined Ethiopian Airline's frequent flyer programs database to extract important customer information. Then, the customer data was segmented into five clusters which could clearly show the different customer types.

Pattern discovery is another application of data mining technology. Shegaw (2002) researched on the applicability of data mining to predict child mortality based upon community-based epidemiological datasets gathered by the Butajira Rural Health Project. He employed three steps of data mining process iteratively: data collection, data preparation, and model building and testing. Neural networks and Decision tree techniques were employed by the researcher to develop child mortality patterns. Similarly, Helen (2003) tried to identify interesting patterns and relationships out of census and survey database for the intention of understanding the nature of child labor problem in Ethiopia. The researcher developed the model using Apriori algorithm of association rule mining.

A special type of data mining called web mining is also applicable in different areas. Web mining may be practiced in the analysis of content, structure or usage of web sites so that the usability of web sites could be maintained. Web usage mining is the main intent of this research.

2.2. Web Usage Analysis

The World Wide Web grows at an alarming rate in both the volume of traffic and the size and complexity of web sites (Cooley et al. 1999). According to the survey conducted by Hu et al (n.d), in less than a decade, the World Wide Web has become one of the world's three major media, with the other two being print and television. This occurrence has brought challenges for web community. Thus, distinguishing documents that meet a group of web users, web site design, web server design and simple navigation through a web site are found more difficult (Chitraa and Davamani 2010). Accordingly, knowing users' navigation through web is a critical task in developing applications like e-commerce, personalization, web site and recommender systems. This activity of discovering users'

navigation through web sites and extract information from web documents and services is known as web mining (Kosala and Blockeel 2000). Web mining was first invented by Etzioni (1996).

2.2.1. Types of Web Mining

Commonly, web mining has three main research areas namely web content mining, web structure mining and web usage mining. Web usage mining is also known as web log mining as web server log files are commonly used for the mining purpose (Suneetha and Krishnamoorthi 2009, Srivastava et al. 2000, Borges and Levene 1999).

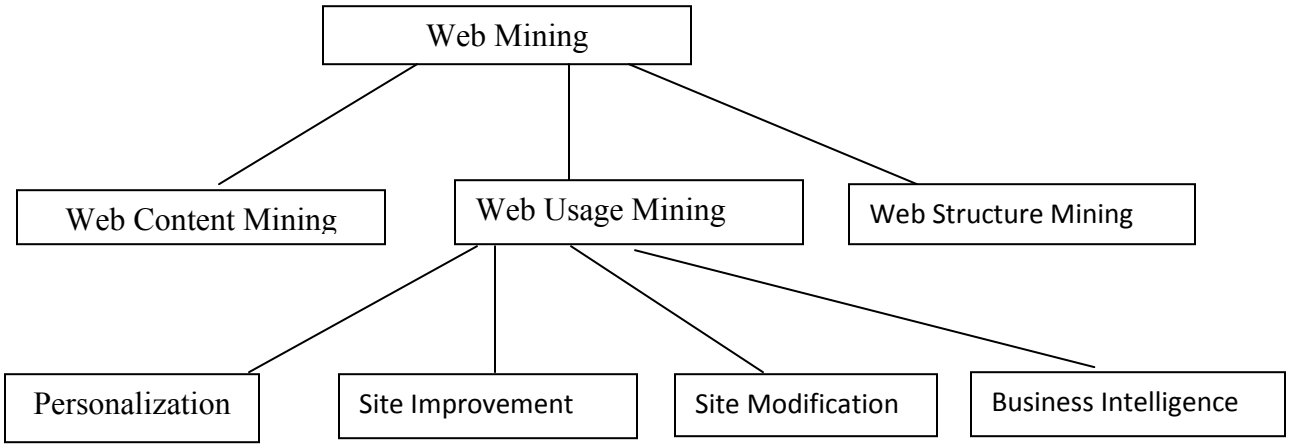


Figure 2. 3: Types of Web Mining
(Source: Srivastava et al. 2000)

Web Content Mining: Web content mining is the task of discovering useful information from the content/documents of the web to provide an efficient mechanism to help the users in finding the information they seek (Chitraa and Davamani 2010). Accordingly, inner-document information is used in web content mining. Content of includes multimedia, structured (i.e. XML documents), semi-structured (i.e. HTML documents) and unstructured (i.e. plain text) data. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing different documents by keywords, categories and contents (Pal et al. 2002, Madria et al. 1999, Chakrabarti 2000).

Web Structure Mining: Madria et al. (1999) described web structure mining as the process of discovering the structure of hyperlinks within the web. Accordingly, it is aimed at exploring useful knowledge from the web structure, hyperlink references and so on. Web structure mining discovers the link structures at the inter-document level for the intention of identifying the authoritative and the hub pages for a given subject. Authoritative pages contain useful information, and are supported by several links pointing to it, which means that these pages are highly-referenced. A page that has a lot of referencing hyperlinks means that the content of the page is useful, preferable and maybe reliable. Hubs are web pages that contain many links to authoritative pages, thus they help in clustering the authorities. Web structure mining can be achieved only in a single portal or on the whole web. Mining the structure of the web supports the task of web content mining. Using the information about the structure of the web, the document retrieval can be made more efficiently, and the reliability and relevance of the found documents can be greater. The graph structure of the web can be exploited by web structure mining in order to improve the performance of the information retrieval and to improve classification of the documents (Chakrabarti et al. 1998, Kleinberg et al. 1999, Hou and Zhang 2003, Han and Elmasri 2004).

Web Usage Mining: Web usage mining attempts to discover useful knowledge from the *secondary data* [4]obtained from the interactions of the users with the web (Srivastava et al. 2000). It aims to discover interesting and frequent user access patterns from web browsing data stored in the log files of web/proxy servers or browsers. The mined patterns can facilitate web recommendations, adaptive web sites, and personalized web search and surfing (Eirinaki and Vazirgiannis 2003), improve web structure and web server performance (Pei et al. 2000). Web usage mining is the target of the research. So, it is dealt in the next sections.

[4] Data from browser logs, user profiles, web server access logs, registration data, user sessions or transactions, cookies, user queries, mouse clicks and any other data that is the result of interaction with the Web.

2.2.2. Application of Web Usage Mining

Web usage data mining has various real world applications. It could help to engage new customers, maintain current customers, track customers who are leaving web site and so on (Abraham 2005). Masseglia et al. (1999) also showed that usage information can be extracted to increase web server efficiency by pre-fetching and caching strategies.

a. Personalization

In web sites, personalization for a user is achieved by keeping track of previously accessed pages. Making dynamic recommendations to a web server on the basis of customers' profiles ,in addition to usage behavior, is very attractive to many applications e.g. cross sales and up sales in E-Commerce (Mobashar et al. 1999).

b. System Improvement

Web-based systems may be prone to network traffic and security issues as long as they are open for access by lots of users from different corners of the world (Fawcett and Prouost 1999). Caching policy is a well known strategy for improving the performance of such systems. Web usage mining can be used for developing policies for web caching, network transmission, load balancing or data distribution (Pirolli et al. 1996, Qiang and Haining 2003).

c. Site Modification

The attractiveness of a web site, in terms of content and structure, is crucial for its usability (Spiliopoulou 2000). Web usage mining could provide knowledge on user navigation behavior. Then, this knowledge could be utilized by web site designers to design the web site for users' ease of navigation. Besides, structure of a website may be set to change automatically on the basis of usage patterns discovered from server logs as in the case of adaptive website (Perkowitiz and Etzioni 1998, Pirolli et al. 1996).

2.2.2. Generalized Structure of Web Usage Mining

Analysis of a web site usage could range from straightforward statistical approach of analysis such as page access frequency to sophisticated form of analysis such as the common traversal paths through a web site (Chitraa and Davamani 2010). As Hu et al. (n.d) clearly articulated in their work, web usage mining involves usage data gathering, usage data preparation, navigation pattern discovery, pattern analysis and visualization and pattern application. These tasks could be generalized into the following three important phases: pre-processing, pattern discovery and pattern analysis (Srivastava et al. 2000). These tasks are depicted below diagrammatically.

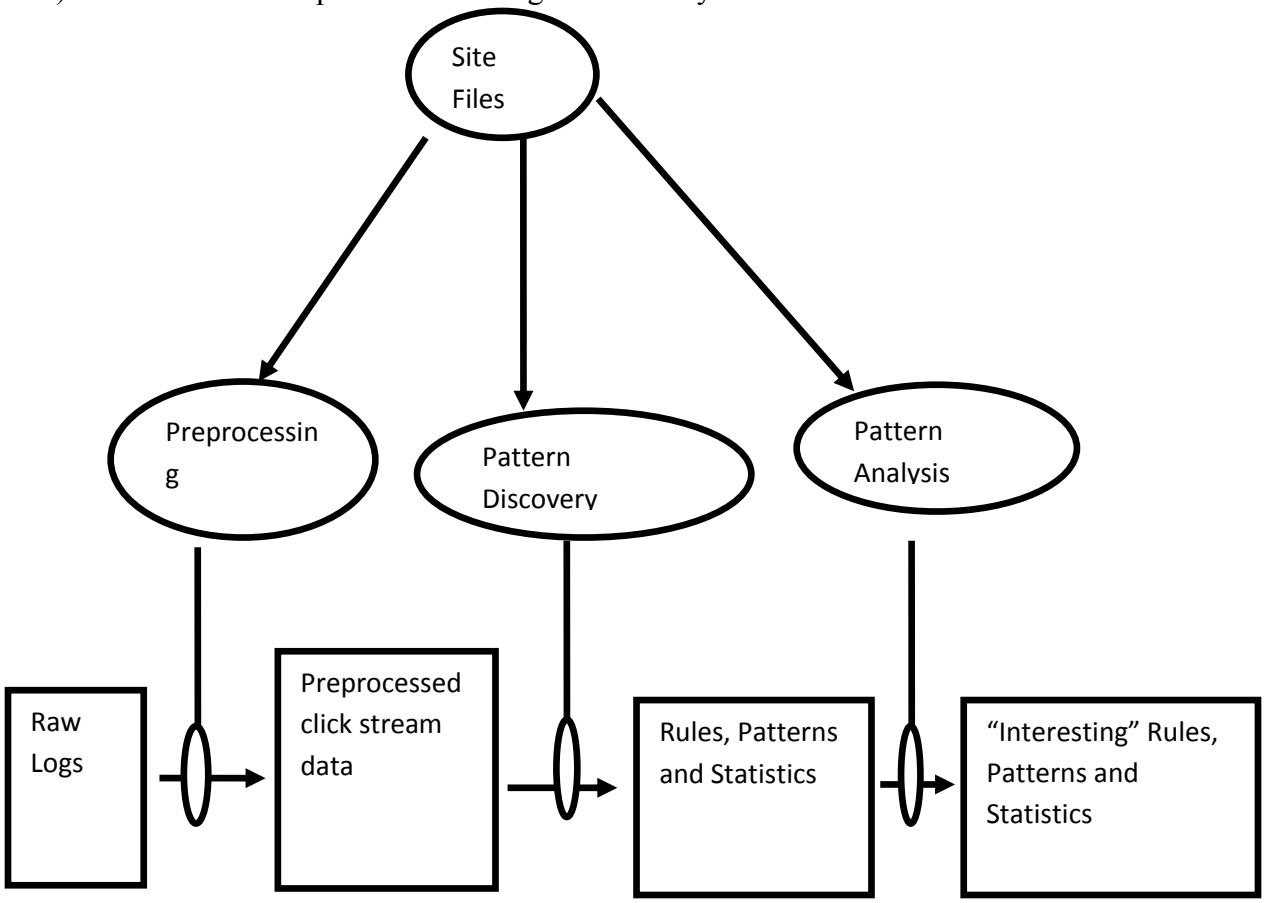


Figure 2. 4: High level Web Usage Mining Process
(Source: Srivastava et al. 2000)

a. Data Collection

Gathering of usage data concerns with acquisition of users' usage data on web sites. There are two possible sources namely trial runs by humans and web log file (Hu et al. n.d). As these researchers claimed, utilization of humans' trial runs is not a recommended data source for the mere fact that this source is costly and prone to biases. Rather, web log data is used as a typical data source for most of the web usage mining system developments. Log file is recorded whenever users of a certain web site send requests to the respective web servers (Srivastava et al. 2000). The log data could be collected from web server logs, web proxy server logs or clients' browser logs. Meanwhile, each of the three web log data sources suffers from two major drawbacks as discussed by Hu et al. (n.d).

Server-Side logs: Server-side logs are logs which show all requests towards a web site for a certain period of time. Generally, these logs supply the most complete and accurate usage data. Its two drawbacks are:

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

Proxy-Side logs: Proxy-side logs are logs which show all requests that have been delivered via the proxy under discussion. A proxy server takes the HTTP requests from users and passes them to a web server; the proxy server then returns the results passed to them to the user.

The two drawbacks associated with proxy-side log are:

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

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

The drawbacks of this approach are:

- The design team must deploy the special software and have the end-users install it.
- This technique makes it hard to achieve compatibility with a range of operating systems and web browsers.

Web Log File

Clicking links in a page of a web site or giving the full URL onto the address bar of the browser would request the browser for the information and the browser requests the information from the web server. Each request is recorded by the server in so-called *log files (*Hu and Cercone n.d*)*. Access logs contain requests for a given period of time. The time interval used is normally an attribute of the web server. In fact, web servers may record others log files such as error log, agent log, or referrer log (Hu and Cercone n.d). There is a log file present for each period and the old ones are archived or erased depending on the usage and importance.

Currently, there are three log file formats which are based on American Standard Code for Information Interchange (ASCII) text format. Each of them is described below.

W3C Extended Log file Format:

It is a customizable format with a variety of different fields that could be included when desired. In this format, time is recorded as UTC (Greenwich Mean Time). Here is an example log of such format.

#Software: Microsoft Internet
Information Services 5.1
#Version: 1.0

#Date: 1998-05-02 17:42:15
#Fields: time c-ip cs-method cs-uri-stem
sc-status cs-version
7:42:15 172.16.255.255 GET /default.htm
200 HTTP/1.0

Microsoft IIS Log file Format:

This format is non-customizable which includes basic items such as user's IP address, user name, request data and time, service status code and number of bytes received. Moreover, elapsed time, the number of bytes sent, the action and the target file may also be included in this format. The items are separated by commas which make the format easier to read than other ASCII format that use spaces as separator. The time is recorded as local time. A hyphen (**-**) acts as a placeholder if there is no valid value for a certain field. Here is an example log in this format.

192.168.114.201, —, 03/20/98, 7:55:20,
W3SVC2, SALES1, 192.168.114.201, 4502,
163, 3223, 200, 0, GET, /DeptLogo.gif, —,
172.16.255.255, anonymous, 03/20/98,
23:58:11, MSFTPSVC, SALES1,
192.168.114.201, 60, 275, 0, 0, 0, PASS,
/intro.htm, —,

NCSA Common Log File Format:

This format is non-customizable and available for non-FTP sites. It records remote host name, user name, date, time, request type, HTTP status code and the number of bytes sent by the server. Like W3C Extended Log file Format, items are separated by spaces. Local time is used to record time. Here is an example log in this format.

172.21.13.45 — REDMOND\fred
[08/Apr/1997:17:39:04 -0800] "GET

/scripts/iisadmin/ism.dll?http/serv
HTTP/1.0" 200 3401

NCSA Combined (Extended) Log File Format:

This format is similar to NCSA Common Log File Format except the fact that three additional fields are incorporated here. The additional fields are referrer, user agent and cookie fields.

Most of log files of web servers are stored in a common log file format (CLFF) (Logging Control In W3C Control, 1995) or in an extended log file format (ELFF) (W3C, no year).

Web servers can be configured to write different fields into the log file in different formats. The common fields used by web servers to record the log file are elaborated below. *remotehost, rfc931, authuser, date, request, status, bytes, referrer, user_agent, cookies*. Each of these fields is elaborated underneath.

Field Name	Description of the field (with example)
Remotehost	Remote hostname (IP address if Domain Name Service (DNS) hostname is not available) Example: 10.128.4.240
rfc931	The remote login name of the user. Example: –
Authuser	The user name with which the user has authenticated himself. Example: –
[date]	Date and time of the request with the web server's time zone Example: [20/Jan/2010:10:17:37 + 0100]
"request"	The request line exactly as it came from the client. It consists of three subfields: the request method, the resource to be transferred, and the used protocol. Example: "GET/HTTP/1.1"
Status	The HTTP status code returned to the client. Example: 200
Bytes	The content-length of the document transferred. Example: 12079

"referrer"	**The url the client was on before requesting the url; It would be null in case the user selected a bookmark, typed a URI, chose from a history list or has certain privacy features activated.** **Example. http://www.google.com/?hl=en&q=Addis+Ababa+University**
"user_agent"	**The software the client claim to be using.** **Example:** **"Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)"**
"Cookies"	**Cookies used by sites to collect user information** **Example:** **"USERID=CustomerA; IMPID=01234"**

Table 2. 1: The most commonly used fields of access log file entries by web servers

b. Usage Data Preprocessing

In classical data mining activities, preprocessing is expected to abolish problems in the data due to improper data entry. Despite the fact that human beings are not directly involved in the data entry of web log data, the log data is usually noisy and ambiguous on behalf of requirements of fields for the pattern discovery (Chitraa and Davamani 2010). These researchers claimed that usage data preprocessing is accomplished to get a set of minable objects for a particular web site (or sets of web sites). Accordingly, web log preprocessing consists of converting the usage information contained in the various available data sources into the data abstractions necessary for pattern discovery. Usage data gathering and usage data preparation are encompassed at this phase.

According to Srivastava et al. (2000), usage data preprocessing is arguably the most difficult task in the web usage mining process due to the incompleteness of available data for pattern discovery. The input for this phase is raw usage data from sources such as web server while the output of the preprocessing task is transaction file of user accesses. The detailed preprocessing of a usage data proposed by Srivastava et al. (2000) is shown diagrammatically below. In the diagram, the dashed lines are optional preprocessing paths.

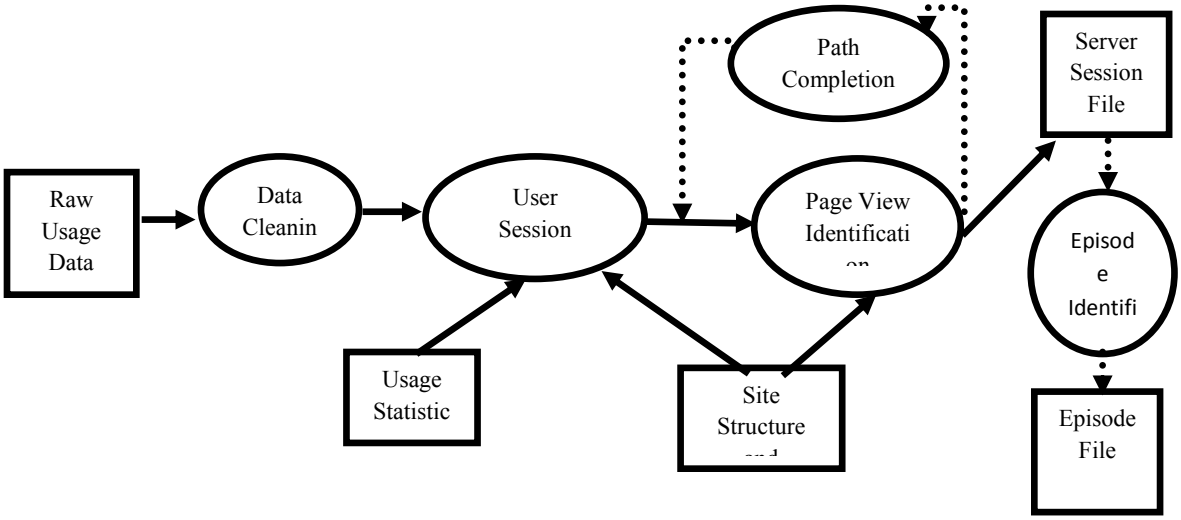


Figure 2. 5: Details of Usage Preprocessing
(Source: Srivastava et al. 2000)

According to Thakare and Gawali (2010), data preparation is the most time consuming task in web usage mining. Moreover, Pyle (1999) claimed that data preparation is critical for the success of pattern discovery. The preparation task depends on both the nature of data and the type of analysis (Hand et al. 2001). Raw log data contains a vast amount of variant request entries. The discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses to the web site (Cooley et al. 1999). Processing of certain types of requests would lead to unfair conclusions (e.g., requests generated by spider engines). The same researcher revealed that stripping the data has a positive effect on processing time and the required storage space. Usage data preparation is, therefore, about restoring the raw log data in a reliable and consistent way. Removing undesirable entries, distinguishing among users, building sessions and restoring the contents of a session are some of the tasks which are discussed here.

Cleaning Access Log Data (Data Cleaning)

Web logs contain user activity information, of which some are not closely relevant to usage mining and can be removed without noticeably affecting the mining. It is a web site specific step as long as tasks to be accomplished depend on the content and volume of the usage log which, in turn, depend on the web site itself (Srivastava et al. 2000). Here are the tasks in removing undesirable entries:

Filtering log image entries (Graphics files): The HTTP protocol requires a separate connection for every file that is requested from the web server (Cooley et al. 1999). Images are automatically downloaded based on the HTML page requested and downloads are recorded in the logs. Thus, log image entries may not help the usage mining and can be removed depending on the web site under consideration. The suffixes could be checked (.GIF, .JPEG) for the removal.

Filtering Dynamic Pages: Dynamic pages cannot be located on the web server as an individual file since these pages are built by a specific engine using several data sources. Thus, dynamic pages cannot be analyzed in a simple way. At the mean time, there is no standard for the structure of URL requests for dynamic pages except that parameters appear after the "?" in the URL which consist of name/value pairs. Therefore, dynamic pages can basically be filtered out by searching for the question mark in "request" fields of log entries.

Filtering Request Methods: HTTP/1.0 allows several methods to be used to indicate the purpose of a request. The most often used methods are get, head and post. Since using the get method is the only way of requesting a document that could be useful for usage pattern discovery, the request method filter ignores any other requests. The filter examines the "request" field of the log entry for the "get" method identifier.

Filtering and Replacing Escape Characters: URL escape characters are special character sequences made up of a leading "%" character and two hexadecimal characters. They substitute special characters in URL requests that could be problematic while transferring requests to different types of servers. Special characters are simply replaced by sequences of standard characters. In most cases the task is only to replace these escape sequences with their representatives, but in certain instances URLs contain corrupted sequences that cannot be interpreted. In these cases the entries should be ignored. Corrupt sequences can be caused by typing errors of the users, automatically generated robot requests, etc.

Filtering Unsuccessful Requests: It is common occurrence to face failed requests in surfing of sites and such requests are also recorded to the log file. The user doesn't use the requested page to navigate through the site since the error page doesn't provide any link to follow. Thus, failed requests need to be removed. The failed requests could be checked out using the status field of the log entries. Here is elaboration of the most important status codes as defined by Grace et al. (2011).

1xx Info
HTTP_INFO – Request received, continuing process
100 Continue – HTTP_CONTINUE
101 Switching Protocols -HTTP_SWITCHING_PROTOCOLS
102 Processing – HTTP_PROCESSING
2xx Success
HTTP_SUCCESS – action successfully received, understood, accepted
200 OK – HTTP_OK
201 Created – HTTP_CREATED
202 Accepted – HTTP_ACCEPTED
3xx Redirect
HTTP_REDIRECT – The client must take additional action to complete the request xx Info
301 Moved Permanently – HTTP_MOVED_PERMANENTLY
302 Found – HTTP_MOVED_TEMPORARILY
304 Not Modified – HTTP_NOT_MODIFIED
4xx Client Error
HTTP_CLIENT_ERROR – The request contains bad syntax or cannot be fulfilled
400 Bad Request – HTTP_BAD_REQUEST
401 Authorization Required –HTTP_UNAUTHORIZED
402 Payment Required – HTTP_PAYMENT_REQUIRED
404 Not Found – HTTP_NOT_FOUND
405 Method Not Allowed – HTTP_METHOD_NOT_ALLOWED
5xx Server Error
HTTP_SERVER_ERROR – The server failed to fulfill an apparently valid request.

500 Internal Server Error – HTTP_INTERNAL_SERVER_ERROR

501 Method Not Implemented – HTTP_NOT_IMPLEMENTED

505 rviceTemporarily Unavailable – HTTP_SERVICE_UNAVAILABLE

505 TP Version Not Supported – HTTP_VERSION_NOT_SUPPORTED

Filtering Robot/Spider transactions/Automated agents: A robot, also known as spider or crawler, is a program that automatically fetches web pages from their respective repositories. It is used to feed pages to search engines or other intended softwares. Tan and Kumar (2002) claimed that requests from robots and other software agents have to be removed from the raw log data as such requests don't reflect human navigation behavior. As robot-access patterns are usually different from human-access patterns, many of the robot accesses can be detected and removed from the logs. The agent field of the web log data could be checked to distinguish the availability of such records in the log.

Filtering Request URLs for a Domain name: A URL of a page request consists of a domain name and the path of the requested document relative to the public directory of the domain. Since the domain name is not ambiguous to the responsible web server, it stores only the relative path of the request in the access log files, without the domain name. However, log file entries tend to contain the whole absolute path in a few cases. This leads to mapping errors during data integration as long as the mapping table contains only relative paths and comparison is based on paths similarity. For these reasons, a URL in the "request" field has to be transformed to the relative format.

Attribute Selection: It is common task to reject some fields which are found irrelevant for pattern discovery. For example, the file size (bytes) field is not important for association mining.

Normalize uniform resource location (URL) (typically for request field**) and parse any common gateway interface (CGI) data (**typically for referrer field**)**: In lots of web servers, request for variants of URL such as www.aau.edu.et, aau.edu.et, aau.edu.et/, www.aau.edu.et/index.php and others return the same result. Therefore, normalization of the URI is a basic need as far as data cleaning is concerned. Besides, a CGI data need to be parsed and the name/value pairs need to be broken into separate fields. For example, the

referrer field http://www.google.com/search?hl=en&q=addis+ababa+university+admission+2011 is a CGI data containing a referrer name and the requested value separated by the symbol"?" So, this must be partitioned into the two separate categories (Srivastava et al. 2000). The request "GET/ HTTP/1.0" missed the request URL. Thus, normalization is expected to fill the missed request.

Distinguishing among users (User Identification)

Hu et al. (n.d) defined a web user as an individual that accesses files from one or more web servers through a browser. Accordingly, a web log records users' activities sequentially according to the time each occurred. Moreover, requests of a user may be distributed in requests of other users which tend to make user identification tough. In order to study the actual user behavior, users in the log must be distinguished.

Registration is the first method to identify users. Contrary to intranet applications where users are required to identify themselves by following the login process, user identification is much more complicated in the case of public domains as internet protocols such as HTTP and TCP/IP do not require user authorization from client applications (e.g., web browser). Thus, this method may be used if the web site requires authentication where users have to follow a login process each time they want to browse web sites.

Cookies are the second methods to distinguish users in web sites. In this method, explicit user authentication is not required. Rather, servers can use so called cookies[5], user specific files stored on client machines. Each time a user visits the same files, the server can obtain user information from stored cookies. Eirinaki and Vazirgiannis (2003) and Huysmans et al. (2003) used cookies to identify users. However, utilization of cookies for identification purpose is prone to two problems. Firstly, the users can lock the use of cookies, so a web server cannot store information locally in the user machine. Secondly, the user can delete the cookies.

As a third method, web users could also be distinguished through their IP address recorded in the log file. This also has two drawbacks. Firstly, different users can be registered with

[5] Cookies are *data sent by the server to the client; data is locally stored in cookies and sent to the server with each request (WCA, 2011).*

identical IP address. Secondly, the same user can be registered with different IP addresses. Fortunately, the first problem could be alleviated with the help of session identification in identifying a user with IP address. For recognition of general users' pattern, the second drawback is not a big deal (Etminani et al. 2009).

Combination of these methods and other features has been suggested for better user identification. Cooley et al. (1999) suggested IP address and browser agent type as a unique user identification pair for accurate identification for usage pattern that is based solely on access log files. This is suggested by the expectation that a user rarely uses combinations of browsers at a time. According to the claim of this method, each unique IP address and browser agent pair shows a unique user. Combination of IP and referrer field is also used as a heuristic of user identification. However, this method fails to work for the situations where the user types the URL directly to the address bar or uses bookmarks. At the mean time, some researchers such as Baglioni et al. (2003) used IP/cookie pairs for user identification.

Session Identification

For logs that span long periods of time, it is very likely that individual users will visit a web site more than once or their browsing may be interrupted. The goal of session identification is to divide the page accesses of each user into individual sessions (Srivastava et al. 2000).

According to the claims by Spiliopoulou et al. (2003), there is no best heuristic for all cases. Session reconstruction could be accomplished using time-oriented heuristics or navigation-oriented heuristics. The time-oriented heuristics is, in turn, viewed in two angles. The first tends to assume an upper bound threshold on the time spent in the entire site during a visit. The second assumes an upper bound on page-stay time to create newer sessions. That means, long time elapses between pages is considered. Navigation-oriented heuristics may use referrer fields or site topology (Spiliopoulou et al. 2003).

Here is an algorithm for session identification:

```
Input: L,|L|, Δt//the log entries, the number of log entries and a time threshold
Output: S, |S|//session
for each Li of L
        if METHODi is 'GET' and URLi is 'WEBPAGE'
                if ɔ Sk Є OPEN_SESSION with IPk==IPi then
                        if ((TIMEi –END_TIMES(Sk)) then
                                Sk= (IPk, PAGEk u URLi)
                        else
                                CLOSE_SESSION (Sk)
                                OPEN_SESSION (IPi, URLi)
                        end if
                else
                        OPEN_SESSION (IPi, URLi)
                end if
        end if
end for
```

Figure 2. 6: A session identification algorithm in WUMprep tool

A time threshold is usually used to identify sessions and it is called *session timeout*. If the time between page requests exceeds a certain time limit, it is assumed that the user starts a new session. Thirty minute is used as a default time out in many commercial web usage mining products.

Path Completion/ Restoring the Contents of a Session

Activities like web caching and using the back button of a browser are common while surfing the web. These activities will cause information discontinuance in log file, i.e. some requests may not be recorded in the log file which could mislead the user session (Cooley et al. 1999). So, important accesses that are not recorded in the access logs need to be added to the user session at this step. This task could be accomplished using referrer field and site topology.

Transaction Identification/ Episode Identification

Given a user session file, it may be considered as a single transaction of many page references or a set of many transactions each consisting of a single page reference. Transaction identification is intended for creating meaningful clusters of references for

each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones-dividing or combining small transactions into fewer larger one-merging. Let L be a set of user session file. For a session entry l € L, the fields l.ip: the client IP address, l.uid: the client's user id, l.url: the URL of the accessed page, and l.time: the access time are found important as elements of the transaction. Hence, a general transaction model is shown below.

$t=<ip_t, uid_t, \{(l^t_1.url, l^t_1.time),\ldots,(l^t_m.url, l^t_m.time)\}>$

Where for $1<= k <= m$, l^t_k € L, $l^t_k.ip=ip_t$, $l^t_k.uid=uid_t$----------1

There are three dividing approaches for transaction identification: reference length, maximum forward reference (Chen et al. 1998) and time window (Cooley et al. 1999).

Reference Length Approach: This approach claims that the amount of time a user spends on a page correlates to whether the page should be classified as auxiliary or content page for the user. Accordingly, a cut off time length, t, could be calculated a maximum likelihood estimate provided that a percentage of auxiliary references is given. The auxiliary references percentage can be estimated based on the structure and content of the site or experience of the data analyst with other server logs. Let γ be the percentage of auxiliary references, λ be reciprocal of observed mean reference length. Then, upon integration of exponential distribution from γ to 0, the cut off length could be given as:

$$t = \frac{-\ln(1-\gamma)}{\lambda} \text{----------2}$$

When using this approach, the generalized transaction model of 1 above will be as follow:

$t=<ip_t, uid_t, \{(l^t_1.url, l^t_1.time, l^t_1.length),\ldots,(l^t_m.url, l^t_m.time, l^t_1.lenght)\}>$

Where for $1<= k <= m$, l^t_k € L, $l^t_k.ip=ip_t$, $l^t_k.uid=uid_t$----------3

The length of each reference is estimated by taking the difference between the time of the next reference and the current reference.

Once the cut off length is computed, the pages will be added to the transaction by checking the following two conditions. In the case of auxiliary pages: $for\ 1 \le k \le (m-1) : l_k^{\ t}.length \le c \ \ and\ k = m : l^t{}_k.length \ge c$, the respective pages are added to transaction. Whereas content-only pages that satisfy the condition $for\ l \le k \le m : l_k^t.length > c$ are added to the transaction list.

Maximum Forward Reference (MFR) Approach: According to MFR approach of transaction identification, the maximal forward reference pages are the content pages and the pages that lead to each maximal forward reference are auxiliary pages .The approach adds page access entries to a session list up to the page before a backward reference is made. Backward reference is defined to be a page that is already contained in the set of pages for the current transaction. In that case, it starts a new session list and goes on with iteration. For example, an access sequence of A B C B D E E E F G would be broken into four transactions, i.e. A B C, B D E, E, and E F G. MFR approach does not require an input parameter that is based on assumptions about a particular set of data. In this aspect, it is preferable to Reference Length (RL) approach. Meanwhile, MFR has two drawbacks. Firstly, it does not consider that some of the "backward" references may provide useful information. Secondly, it may include entries within the same session even if a week elapsed between them. The generalized transaction model in 1 above works for MFR approach. The MFR approach uses all the fields that have been suggested for RL approach. Besides, the referrer page from which a request has been made is useful for the approach.

Time Window Approach: This approach divides page accesses for a user using a time window. This window or time interval is suggested to be approximately 30 minutes (Hay et al. 2003), (Cooley et al 1999), (Xing and Shen 2004), (Catledge and Pitkow 1995).

c. Navigation Pattern Discovery

According to Han and Kamber (2006), pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Moreover, pattern discovery is about looking for interesting usage

patterns out of the prepared log data. This is accomplished using algorithms like sequential pattern generation, association rule mining or clustering (Cooley et al. 1999).

Sequential Pattern Generation: Agrawal and Srikant (1995) described sequential pattern discovery as the problem of discovering sequential patterns to find inter transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. Each of the following three systems uses a variant of sequential pattern generation to find navigation patterns.

Web Utilization Miner (WUM) which was developed by Spiliopoulou and Faulstich (1998) was used to discover navigation patterns using an aggregated materialized view of the web log. This technique offers a mining language that experts can use to specify the types of patterns they are interested in. Using this language, only patterns having the specified characteristics are discovered, while uninteresting patterns are removed early in the process.

MiDAS which was developed by Buchner et al. (1999) extends traditional sequence discovery by adding a wide range of web-specific features. In this system, new domain knowledge types in the form of navigational templates, web topologies, syntactic constraints and concept hierarchies have been incorporated.

Chen et al. (1998) proposed a method to convert the original sequence of log data into a set of maximal forward references. Algorithms are then applied to determine the frequent traversal patterns, i.e., large reference sequences, from the maximal forward references obtained.

Association Rule Mining: Association rule discovery can be used to find unordered correlations between items found in a set of database transactions (Agrawal and Srikant 1994). In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold (HU et al. n.d).

OLAP (On-Line Analytical Processing): OLAP is a category of software tools that can be used to analyze data stored in a database. It allows users to analyze different dimensions of multidimensional data. For example, it provides time series and trend

analysis views. According to Zaiane et al. (1998), WebLogMiner uses the OLAP method to analyze the web log data cube which is constructed from a database containing the log data. Data mining methods such as association or classification are then applied to the data cube to predict, classify, and discover interesting patterns and trends. Büchner et al. (1998) also made use of a generic web log data hypercube. Various online analytical web usage data mining techniques are then applied to the hypercube to reveal marketing intelligence.
A model by Borges and Levene (1999) showed navigation records in terms of a hypertext probabilistic grammar, which is a probabilistic regular grammar. For this grammar, each non-terminal symbol corresponds to a web page and a production rule corresponds to a link between pages. The higher probability generated strings of the grammar correspond to the user's preferred trails.

WAP-Tree: Pei et al. (2000) proposed a data structure WAP-tree to store highly compressed, critical information contained in web logs, together with an algorithm WAP-mine that is used to discover access patterns from the WAP-tree.

d. Pattern Analysis and Visualization

Pattern analysis is about undergoing further interpretation and visualization of the discovered pattern (s) before applying the discovered rules to useful application (Srivastava et al. 2000). Accordingly, navigation patterns need further analysis and interpretation before application. The analysis usually requires human intervention. Otherwise, it is distributed to the two other tasks: navigation pattern discovery and pattern applications. Navigation patterns are normally two-dimensional paths that are difficult to perceive if a proper visualization tool is not supported. Srivastava et al. (2000) claimed that a useful visualization tool may provide two functions. Firstly, it displays the discovered navigation patterns clearly. Secondly, essential functions for manipulating navigation patterns such as zooming, rotation and scaling are provided.

Hong and Landay (2001) developed WebQuilt system which allow captured usage traces to be aggregated and visualized in a zooming interface. The visualization also shows the most common paths taken through the web site for a given task, as well as the optimal path for that task as designated by the designers of the site.

e. Pattern Deployment

Pattern deployment is about applying appropriate discovered navigation patterns to an area of interest such as improving web site design, making additional product and topic recommendation, improve web personalization, schema modification, web site and web server performance, fraud detection and future prediction or learning users' behavior.

Web Site/Page Improvement: The most important application of discovered navigation patterns is to improve the web sites/pages by (re)organizing them (Ivory and Hearst 2002). Other than manually (re)organizing the web sites/pages, there are some other automatic ways to achieve this (Ivory and Hearst 2002). Adaptive Web sites, as suggested by Perkowitz and Etzioni (2000), automatically improve their organization and presentation by learning from visitor access patterns. Such sites mine the data buried in web server logs to produce easily navigable web sites. Clustering mining and conceptual clustering mining techniques are applied to synthesize the index pages, which are central to site organization.

Additional Topic or Product Recommendation: E-commerce web sites use recommender systems or collaborative filtering to suggest products to their customers or to provide consumers with information to help them decide which products to purchase. For example, each account owner at Amazon.com is presented with a section of 'Your Recommendations' which suggests additional products based on the owner's previous purchases and browsing behavior. Various technologies have been proposed for recommender systems (Srivastava et al. 2000et al. 2000) and many electronic commerce sites have employed recommender systems in their sites (Schafer et al. 2000). For further studies, the GroupLens (2011) research group at the University of Minnesota is known for its successful projects on various recommender systems.

Web Personalization: Web personalization (re)organizes web sites/pages based on the web experience to fit individual users' needs (Mobasher et al. 2000, Spiliopoulou 2000). It is a broad area that includes adaptive web sites and recommender systems as special cases. The WebPersonalizer system by Mobasher et al. (2002) uses a subset of web log and

session clustering techniques to derive usage profiles, which are then used to generate recommendations. An overview of approaches for incorporating semantic knowledge into the web personalization process was given by Dai and Mobasher (2003).

User Behavior Studies: Knowing the users' purchasing or browsing behavior is a critical factor for the success of e-commerce and other businesses. The 1:1 Pro system by Adomavicius and Tuzhilin (2001) constructs personal profiles based on customers' transactional histories. The system uses data mining techniques to discover a set of rules describing customers' behavior and supports human experts in validating the rules. Chen et al. (2002) proposed an algorithm to cluster web users based on their access patterns, which are organized into sessions representing episodes of interaction between web users and the web server. Using attributed oriented induction, the sessions are then generalized according to the page hierarchy which organizes pages according to their generalities. The generalized sessions are finally clustered using a hierarchical clustering method.

Web Caching: Another application worth mentioning is web caching which is the temporary storage of web objects (such as HTML documents) for later retrieval. There are significant advantages of web caching like reduced bandwidth consumption, reduced server load and reduced latency. Together, they make the web less expensive and improve its performance (Davison 2001). Web caching may, in turn, be enhanced by navigation patterns. Lan et al. (1999) proposed an algorithm to make web servers "pushier." The document to be pre-fetched is determined by a set of association rules mined from a sample of the access log of the web server. Once a rule of the form "Document1 → Document2" has been identified and selected, the web server decides to pre-fetch "Document2" if "Document1" is requested. Web page pre-fetching methods based on access log information are described in Nanopoulos et al. (2001) and Schechter et al. (1998).

2.2.3. Related Works

Different aspects of web usage mining have been addressed by various scholars in the past few years. Each of the researchers 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.

Among many, Shen et al. (1999) have claimed that they have presented a more efficient approach for web access association mining. Their approach consists of three steps namely transform raw web logs to a relation table, convert the relation table to a collection of access transactions and mine the transaction collection to extract association rules. They introduced what they called 'an efficient association mining algorithm' which is, in fact, based on the Apriori algorithm. They used a one hour time duration for a session.

He and Goker (2000) tried to detect the possible session boundaries from a web log data. They underlined the importance of detecting session boundaries as it is essential to establish a common context for various statistics relating to user sessions and frequency of user activities. The researchers discouraged making sessions in web log based on the web log data that has been made available from one user or IP address under the umbrella of one session regardless of the length of time covered by the logs.

Batista and Silva (2002) applied data mining techniques for web usage mining of an online newspaper. Using a web access log file, they generated association rules and clustered the web site's URLs for personalizing the web services based on the reading pattern of the users.

Baglioni et al. (2003) undertook a research on preprocessing and mining of web log data for personalization purpose. They aimed to extract models of the navigational behavior of web site users. They have given an emphasis for the preprocessing step and the importance of domain knowledge for cleaning, correcting and completing the input data to provide ontology of web page semantic.

Mekonnen (2009) undertook a research on web usage pattern discovery of AAU official web site. He used combination of statistical analysis and data mining approaches for discovery of the pattern discovery for the web site. As far as data collection of his research is concerned, he used web logs of three months duration of the year 2007. Then, he 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 by developing his own algorithm. 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 web sites. Meanwhile, his research has not considered sequences in which pages of the site are visited. Indeed, discovery of common navigation sequences is very important in web site redesign more than common pages that are accessed together because common navigation sequence encompasses common pages that are accessed together. Furthermore, the dataset was not large enough to illustrate better usage patterns of the site.

CHAPTER THREE

METHODOLOGY OF THE RESEARCH

3.1. Overview

Srivastava et al. (2000) developed a model for web usage mining. This methodology was employed for the discovery of users' navigation pattern in the official web site of AAU. The tasks in the methodology are:

- Web log preparation
- Pattern discovery
- Pattern analysis

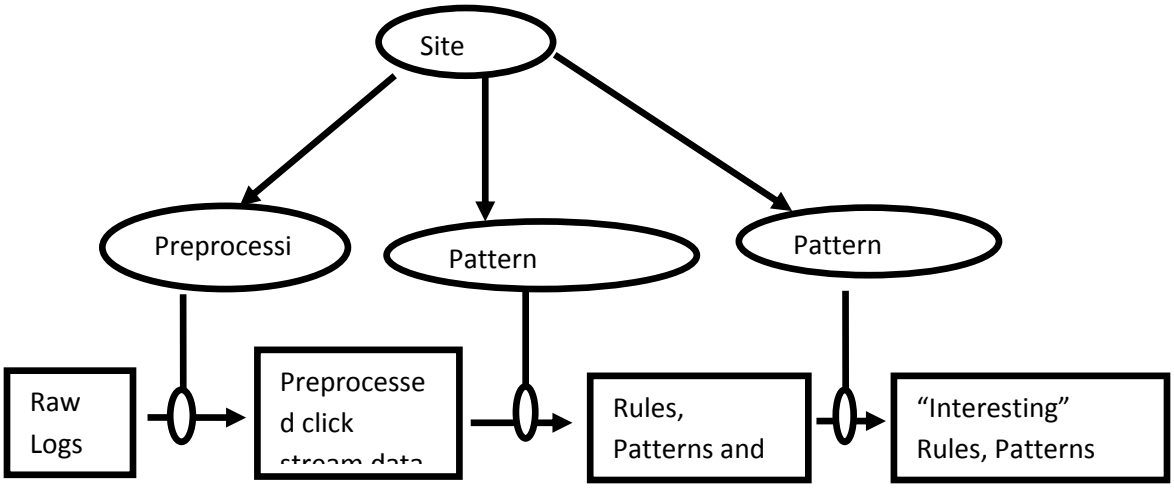


Figure 3. 1: High level Web Usage Mining Process
(Source: Srivastava et al. 2000)

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

3.2. Tools Selection

3.2.1. Tools for Log Preparation

- **WUMPrep:** WUMPrep is a java based open source software. It is used to filter raw logs, generate session files, and remove robot accesses. The tool has a configuration editor in which all sorts of activities to be accomplished are set. Once the configuration is saved at the same directory with the raw log file, perl scripts are available which could be run to perform desired preparation. The Perl scripts include requestFilter.pl, sessionize.pl, sessionFilter and others. The researcher selected the tool regarding the following conditions. Firstly, the tool is very good at detecting and removing non-human requests and automatically downloaded graphics file provided that the user of the tool feed all appropriate data of the log files. Secondly, the tool uses time-based heuristic in developing user sessions of the human requests. Thirdly, it is the only freely available log preparation tool that could be accessed by the researcher.
- **Python 2.6.4**: Once WUMPrep is utilized in preparing the raw web log data, the output is user session file in which requests of the same session ID are scattered across the file. Thus, series of text-based preparation should be accomplished to get a data format that is compatible with the selected Weka's algorithms. Thus, python programming language was selected for the reason that python is very suitable for text processing tasks. Besides, it is open source software that the researcher could get free of charge.

3.2.2. Tool for Usage Pattern Discovery

Weka 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 tool could be used in one of the four modes: Simple Command Line Interface (CLI), Experimenter, Explorer and Knowledge flow.

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

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

3.2.3. Tool for Statistical Analysis

Mach5 Analyzer has been selected to analyze the log records statistically. This tool is selected as it is recommended by researchers and its report is found very important and helpful. The tool reports on rate of page accesses, entry and exit pages in the web site and others.

3.3. Web Log Preparation

The raw log file that was collected from the server of the official web site of AAU was not clean enough to be used in discovery of valid and appropriate usage patterns. Thus, the following preparation tasks were performed.

3.3.1. Data Source Identification

Possible data sources for web usage mining include web server log, site map, and previous usage statistics. This research mainly employed web server log file due to the fact that Suneetha and Krishna (2009) and Cooley et al. (1997) recommended log data for

usage pattern discovery. Indeed, the site map of the web site was utilized one way or another. Moreover, a few usage statistics were used.

The input data for the research was web server log of official web site of AAU. The server log contains access log, error log, SSL access log, SSL error log and SSL request log files individually.

Server Error Logs: It is a file where any sort of errors encountered in processing of requests are recorded. Details of what went wrong and techniques to fix the problem are often contained here. In fact, it is the first place to look when a problem occurs when operating the server. The format of the error log is relatively free-form and descriptive. But there is certain information that is contained in most error log entries. For example, here is a typical error log record:

[Thu Dec 02 09:50:14 2010] [error] [client 66.249.67.111] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 62

From the above error log record, the first item in the log entry is the date and time of the message. The second entry lists the severity of the error being reported. The third entry gives the IP address of the client that generated the error. The fourth item is message of the respective error, which in this case indicates that the server has been configured to deny the client access.

It is not possible to customize the error log by adding or removing information. However, error log entries dealing with particular requests have corresponding entries in the access log. For example, the above example entry corresponds to an access log entry with status code 403.As it is possible to customize the access log, you can obtain more information about error conditions using that log file.

Secure Sockets Layer (SSL) log: This log stores dedicated SSL protocol engine log files. The SSL log may be SSL acess log, SSL error log or SSL request log. Section c, d and e of Appendix D respectively show sample files for these logs.

Server Access Log: It is a type of web log file which records all requests that have been processed by web servers (srivastava et al. 2000). Its location and content could be configured as to the administrators' desire.

Among these log types, access log is employed in this research. In this case, the access log contains remote host IP, remote log in name, user ID, date, time and web server time zone, request URL, status code, bytes, referrer and user agent. Below is a sample access log record from the server log.

```
66.249.67.167 - - [19/Dec/2010:04:36:27 +0300] "GET /index.php/admissions HTTP/1.1" 200 - "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
```

Figure 3. 2: An access log from the web server of AAU official web site

In this log record, the following information could be obtained.

- Client IP address from which a request has been made: 66.249.67.167
- Remote log on ID: unavailable-as shown by dash (-)
- User authentication for the web site: unavailable-as shown by dash (-)
- Date and time at which the request was sent to the web server: December 19, 2010 at 04:36:27 +0300 (hour, minute, second and time zone of the web server respectively)
- Request: GET/index.php/admissions HTTP/1.1 (method of request: GET, request: /index.php/admissions and request protocol: HTTP/ 1.1 respectively)
- Status code: 200 (ok: for Apache Server 1.3)
- Bytes: unavailable-as shown by dash (-)
- Referrer: unavailable-as shown by dash (-)
- User agent: Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)

The raw log data is not suitable for direct application of usage mining algorithms. The data needs to be cleaned and preprocessed. The following tasks were accomplished to prepare the data.

3.3.2. Log Partitioning

Every request for a page is recorded onto a web log record which, in turn, tends to make access web log huge. Consideration of the collected access log from the web server discloses the fact that the access log of a period of one week weighs from 30MB to 100MB. Then, it is very tough task to think of preprocessing all access logs at once. Thus, the access log was partitioned into five separate files. The division is on consideration of the web log access time. Indeed, each log file almost contains logs of four weeks. Each of the following log preparation tasks are applied on the divided log files.

3.3.3. Log Cleaning

As claimed by Chitraa and Davamani (2010), the discovered usage patterns will be valid if and only if human access pages are considered from the web log records. For this reason, non human requests such as automatically downloaded graphic files or robot accesses are removed at this phase. It is known that graphic files are automatically downloaded and recorded onto the log file along with human request for files (Cooley et al. 1999). Requests which are not accessed by humans are distinguished from others. The file extension approach is employed for removal of non-human accesses. On this regard, requests which contain one of the file extensions shown below are considered as non-human request.

S. No	File Extensions	Justifications
1.	.JPEG/JPG, .GIF, .PNG, .CGI, .Js and .Ico	Graphic files that are automatically downloaded without the notice of web users.
2.	.CSS	Cascading style sheets files that are embedded to web pages for the appearance of the page

Table 3. 1: File extensions that show non-human accesses
(Source: Cooley et al. 1999)

Failed requests are also removed. This task was achieved by considering values of the status code as elaborated in chapter two of the paper. Moreover, duplicated request for pages were filtered by using duplicate filter timeout. Furthermore, URLs which are known to return the same page upon request were normalized to one of such URLs to avoid difficulties of getting different statistics for directories which bring the same page. WUMprep uses the following algorithm to clean the raw web log records.

```
Algorithm: Filter only successful human requests
Input: WebLogFile, a file containing raw web log records
Output: CleanedLog, cleaned log record
read log entries from the log file WebLogFile
for each entries in the file
    read fields (URL-stem)//URI-stem indicates the target URL
      if a field does not contain uninteresting extensions
            change the request text part of the entry to small case letters
  if the URL is one of '/index.php' or '/index.php/index.php' or '/'
            assign /index.php/home to the URL part of the entry
        Save the entry
      else
  Remove the log entry
      end if
end for
```

Figure 3. 3: An algorithm for raw web log cleaning

3.3.4. Session Identification

Once the cleaned log is achieved, user sessions need to be developed so that the sequence of requests could be known as transaction using time-based heuristic which is appropriate for the log format in hand.

3.3.5. Feature Selection and Transaction Identification

From the distinguished user session file, requests of the same session ID were collected and the request texts were combined to develop transaction of requests.

3.3.6. Data Integration

The transactions of the divided access log files were merged into one file as warehouse.

3.3.7. Transformation for the data mining tools

The transactions developed so far are not minable for the data mining techniques that the researcher is interested. Firstly, long row of text values would make interpretation vague. Secondly, there are differences in length of transaction which, in turn, implies that there is no fixed number of attributes. On such and other similar cases, the transformation of transaction files was conducted.

The user session contains collection of requests which may impose some sort of difficulties for usage pattern discovery tools. Thus, the users' session was transformed into a format that is minable by Weka. Firstly, unique list of URLs along with given IDs was extracted from the user session. Then, Weka compatible file of the requests is developed using the IDs of the unique URLs distinguished.

3.4. Usage Pattern Discovery

Both straightforward statistical analyzer tool and advanced data mining techniques were applied to discover interesting usage patterns of the web site. The statistical analyzer was used to get generalized statistics on page requests, visiting time of users, and entry and exit pages to and from the web site. The association mining was intended to discover sets of pages that are more frequently accessed by the web site users. Meanwhile, sequence mining pattern discovery was employed to discover common navigation sequences in the dataset.

3.5. Usage Pattern Analysis

The discovered web site usage patterns were evaluated relative to the web site designer's expected navigation patterns. The site map of the web site and the designers of the web site were consulted during the analysis of the discovered patterns.

CHAPTER FOUR

EXPERIMENTATION AND FINDINGS

4.1. Overview

Cooley et al. (1999) elaborated that experimentation is one of the basic tasks that are involved in web usage analysis study. The experiments of this study were conducted following the process model that has been discussed in Chapter Three of the paper. The experimentation was started with collection and cleaning of the raw access log files. The filtered access log was fed into Mach5 Statistical Analyzer tool to discover statistical analysis of the web site usage. The data mining based pattern discovery was, however, preceded by session creation, transaction identification and transformation of the filtered log file. Then, association usage patterns and sequence usage patterns were discovered using the Weka data mining tool. Finally, analyses of the discovered patterns were reported as findings of the study.

4.2. Experiment Setup

The experiment was conducted with the following setup:

- Computer Type: Satellite L505 lap top
- Processor: Pentium(R) Dual-Core CPU T4400@ 2.20GHz
- RAM: 3.00GB
- Operating System: Windows 7 Home Premium
- Data Mining Tool: Weka 3.6
- Statistical Analysis Tool: Mach5 Analyzer 3

4.3. Data Collection and Selection

As it has been pointed out in earlier chapters of the paper, the web log records of AAU official web site was used as source of data for the research work. It is collected from the

web server of the university's official web site. The log file contains access logs, error logs, SSL access logs, SSL request logs and SSL error logs. Sample log files are attached in Appendix D.

The study focused on access logs due to the fact that acknowledged researches in web usage analysis like Srivastava et al. (2000) and Cooley et al. (1999) revealed that access log records show the usage patterns of web sites better.

Access log records of November 01, 2010, 04:31:34 AM to May 08, 2011, 04:43:48 AM were used for the analysis as they were available at the ICT office.

4.4. Web Log Format

The tasks involved in web usage analysis are strongly influenced by the format followed in the log file (Hu and Cercone n.d). The web log format of log under discussion is extended log file format. Sample record from the log file is depicted in Figure 4.1 below.

```
213.55.104.53 - - [22/Nov/2010:16:29:42 +0300] "GET /index.php/sociology-and-social-anthropology HTTP/1.1" 200 30978 "http://www.aau.edu.et/index.php/political-science-and-international-relations/objectivespolitic" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv: 1.9.2.12) Gecko/20101026 Firefox/3.6.12"
```

Figure 4. 1: An access log record

4.5. Data Preprocessing/Web Log Preparation

In fact, there are facts which enforce the researcher to undergo the log preparation before applying the selected data mining tools. These are enumerated below as recommended by Chitraa and Davamani (2010) and Cooley et al. (1999):

- The log file is huge in size and need to be partitioned into manageable size;
- There are requests which are not informative of the usage pattern of AAU web site;

- The web site does not require authentication to start using it. This, in turn, leave authentication ID field value empty in the access log file. Thus, users could not be easily distinguished from the log;
- A certain user may access pages of the web site for longer period of time. Thus, the session need to be constructed.
- Not all fields of the access log are used for the pattern discovery. Thus, features need to be selected.
- As Weka is used by the researcher to discover the usage pattern, some sort of transformation is desired.

The data preprocessing phase of the research took larger portion of the research time. Indeed, it was a very critical part of the web usage pattern discovery for the website under discussion. For the completion of the aforementioned requirements, the log preparation is expected to involve log filtering, session identification, feature selection, transaction identification and transformation of the transaction file into Weka understandable format for the respective pattern discovery techniques. Each of these tasks of the experiment is elaborated below.

4.5.1. Log Partitioning: The raw log records were divided into four manageable portions. In fact, the partitioning was accomplished by keeping log files of five weeks as one.

4.5.2. Log Filtering: The removal of unsuccessful and non-human requests was accomplished by using WUMprep's log filter script. The WUMprep tool has the following parameters in filtering of raw log files.

- **Paths**: If the path of a request matches any of the regular expressions in this parameter, the request is dropped and considered as non-human request. For this experiment, the following expressions were fed line by line: \.Ico, \.GIF, \.JPG, \.JPEG, \.CSS, \.Js, \.GIF, \.PNG, and \.BMP.
- **Hosts**: If some hosts are known to be less important for the usage pattern discovery, those hosts can be specified here. In this experiment, DNS requests are not important for the pattern discovery as long as the user via

the DNSs could not be identified. Thus, some selected DNSs were included for this parameter.

- **Status codes:** This filter is an include filter. Only log records that contain the status code specified here are added as valid requests. This parameter could be left blank if the status code is not a target of research. Regular expressions such as 2XX are also allowed as values here. For the experiment in hand, only successful requests are desired for the usage pattern discovery. So, status code of 200 was used.
- **Duplicate filter timeout**: This parameter is used to decide whether successive requests for the same document from the same host are treated as a single document or not. The value should be given in seconds. For this experiment, the default value, 5 seconds, is used.
- **Output extension**: The extension will be appended to the filtered logfiles names. In this experiment, the default extension, .clean, is used.

Then, the logfilter.pl script was run and the output which is depicted in Table 4.1 was found.

From these, the researcher understood that the filtering task removed approximately 82% of the access log records. Then, the researcher is left with a total of 773,129 human access requests. It should be noted that duplicate requests for a request are excluded from this statistics.

4.5.3. Session Identification

The sessionize.pl script uses the following parameters to develop sessions for the logs.

- **Maximum page view time**: It is to limit the maximum length of a session. If the length of the session exceeds this limit, a new session is begun. The value should be given in seconds. In this experiment, the default maximum page view time, 1800 seconds, is utilized as this time is used by lots of web usage analysis researches worldwide.

Web log partition	Week number	Total requests processed	Percentage of requests removed (%)	Human access requests ready for session identification
Web log 1	1	140,364	81.03	26,621
	2	233,776	80.99	44,445
	3	199,244	80.84	38,174
	4	239,342	80.96	45,561
	5	219,361	78.60	46,938
	Total filtered requests			201,739
Web Log 2	1	192, 091	82.22	34, 145
	2	245,240	82.47	42, 980
	3	224,629	82.67	38, 931
	4	216,591	82.29	38,363
	5	283, 879	82.72	49,055
	Total filtered lines			203,474
Web Log 3	1	113,539	82.7	19,635
	2	324, 059	84.7	49,536
	3	280,860	84.24	44,263
	4	238,965	84.13	37,924
	5	272,363	84.03	43,483
	Total Filtered Requests			194,841
Web Log 4	1	98	78.57	21
	2	241,860	84.14	38,366
	3	365,124	86.32	49,939
	4	273,433	83.92	43,955
	5	205,802	80.18	40,794
	Total Filtered requests			173,075
Total Number of human-access requests				773,129

Table 4. 1: Raw log files and filtered log records

- **Session-ID separator**: Session-ID separator is a character to separate the session ID from the rest of the log line. Normally, it is added at the beginning of each log record. Here also, the default separator, |, is used to separate the session ID from the rest of the log fields during session development.
- **ID cookie**: If the web server uses cookies to identify sessions,this parameter can be used by the sessionizer. In that case, the name of the identifying cookie needs to be entered here. In this experimentation, this parameter is left blank due to the fact that the AAU official web site server does not use cookies to identify sessions.
- **Output extension**: file name extension for the output files of the sessionize script. The default extension of the tool, .sess, is used in this experiment.

Based on this experiment, a total of 95,905 user sessions were developed.

```
084777:1|204.236.197.29 - - [21/Nov/2010:04:26:09 +0300] "GET /index.php?format=feed&amp;type=rss HTTP/1.0" 200 13389 "-" "gMxvwektdqMvMdafosuhflvq"

084777:1|204.236.197.29 - - [21/Nov/2010:04:26:10 +0300] "GET /index.php?format=feed&amp;type=atom HTTP/1.0" 200 13389 "-" "gymhmqtwnUvn moqyawqneiwlsydhnqU"

084777:1|204.236.197.29 - - [21/Nov/2010:04:26:09 +0300] "GET / HTTP/1.0" 200 38762 "-" "nusbnwvfcknkyuvjmauw11"

084777:2|72.14.192.3 - - [21/Nov/2010:04:26:13 +0300] "GET /index.php/academics/faculties/faculty-of-medicine?start=1 HTTP/1.1" 200 37747 "http://www.google.com/search?hl=en&q=addis+ababa+university+admission+2010" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

084777:2|72.14.192.3 - - [21/Nov/2010:04:26:24 +0300] "GET /templates/javanya/images/arrow.png HTTP/1.1" 200 246 "http://smtpgw01.aau.edu.et/index.php/academics/faculties/faculty-of-medicine?start=1" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

084777:3|66.249.65.104 - - [21/Nov/2010:04:26:25 +0300] "GET /index.php/component/e
```

Figure 4. 2: Sample user session file

4.5.4. Feature Selection and Transaction Identification

From the session file, records that have the same session ID are accessed together. At this step, what is important was the sequence of the accesses of the same session ID. Thus, feature selection is one of the targets at this phase. As the tool that is used for pattern discovery could not identify transactions from the session file, a separate python code has been developed to identify the transactions. The algorithm is described as follows.

```
Algorithm: identify transactions of the user sessions
Input: ses, sessionized log data file
Output: trans, transaction file
lines=[] #request records of the sessionized log data
session_text=" "
uniqueIDs=[] # a list to store all possible unique session IDs
open ses
lines= read lines in ses
for (x=0; x< len(lines); x++)# x stands for possible values in lines
        parts=split lines[x] with space character
        sess_ID_IP=split parts[0] with '|' character
        sess_ID=sess_ID_IP[0]
        append each unique sess_ID to uniqueIDs
for (y=0; y<len (uniqueIDs); y++)
        assign session_text to null string
        for (z=0; z <len(lines); z++)
                list=split lines[z] with space character
                dirs=list[6]# the 6th index of the list-the request URL
                ID_IP=list [0].split ('|') # the list with the first index is splitted with '|'
                IDD=ID_IP [0]
                If uniqueIDs[y] equals IDD:
                        Append the request to the session_text
        Write the session_text to trans file
Close the trans file
```

Figure 4. 3: The Algorithm for transaction identification

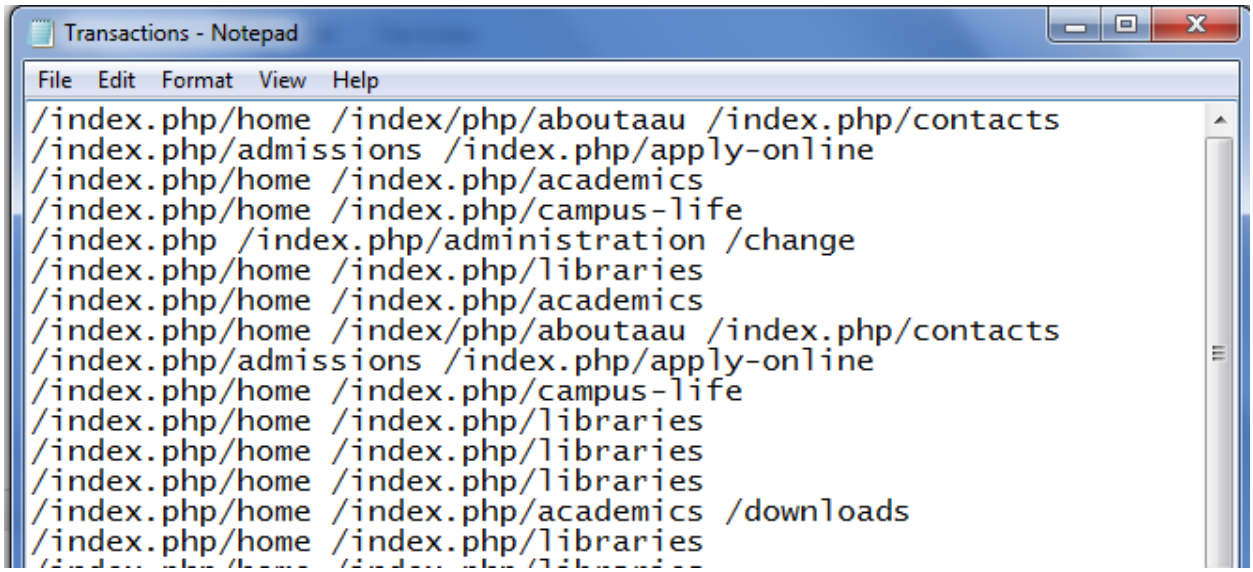


Figure 4. 4: Sample transaction files output of the transaction identification algorithm

4.5.5. Dataset Transformation

What has been accomplished so far is identification of access transaction for each user. In fact, some sort of transformation of the transaction file is desired for the ease of pattern discovery using Weka data mining tool. The tasks involved in transformation and the final outputs of transformation are different for association mining and page access sequence discovery.

a. Preparing the Transaction File for Association Mining

Association rule mining requires each data column to stand for a specific attribute. However, the transaction is just a sequence of accesses in which there is no practice of specific attribute for a data column. The transactions may contain one request or more. The number of attributes should have been equal if the transaction could be used as it is. So, the size of the attributes needs to be fixed by setting the distinguished URL names as attribute name of the transformed file.

As it is known, there are huge numbers of unique URLs in the log file. However, we cannot use all unique URLs for the pattern discovery as it is very tough for the tool of

pattern discovery. Therefore, URLs that fulfill the specified minimum threshold of occurrence are used for the pattern discovery.

The following algorithm was used to distinguish the URLs that met the minimum threshold that could enable the URLs to be an attribute of the new session-URL matrix.

```
Algorithm: Identify URLs that fulfill minimum threshold in the transaction
Input: tr, the transaction file ; th, threshold value of requests
Output: selectedURLs, a list of URLs each in a line
uniqueURL =[]
threshold=m
open tr
aTran=read the lines of tr
close tr
open selectedURLs file  in write mode
for (i=0; i< len(aTran); i++)
        striped= remove new line character at the end of aTran [i]
        parts= split striped with space character
        for (j=0; j<len(parts); j++)
                if uniqueURL does not contain parts[j]
                        append parts[j] to uniqueURL
for (i=0; i<len(uniqueURL); i++)
        count=0
        for (j=0; j<len(aTran); j++)
                striped=remove new line character from aTran[j]
                parts=split striped with space character
                for (k=0; k<len(parts); k++)
                        if parts[k] equals uniqueURL[i]
                                increment count by 1
        if count exceeded th
                session-URL=uniqueURL[i]
                write session-URL to selectedURL file
                write new line character to selectedURL file
close selectedURL file
```

Figure 4. 5: The algorithm employed to extract more frequent request URLs from transaction file

Using the algorithm presented in Figure 4.5, 41 attributes were selected from the available set of URLs. These URLs are attached in Appendix C. Once the number of attributes is specified, the transaction file is read and session-URL matrix is developed. Here, number '1' is used to mark the presence of the request in a transaction line and '?' is used to show the absence of the request in a transaction. The researcher developed the following algorithm to prepare the transaction dataset for input to Weka tool.

```
Algorithm: Session-URL Creator
Input: URLListFile, a file that contains list of unique URLs that met the threshold value
       TransFile, the transaction file
Output: SesURLMatrix: a session-URL matrix
URLs = []
Trans= []
Open URLListFile  file
URLs= read lines of URLListFile
Close URLListFile
Open TransFile
Trans= read lines of TransFile
Close TransFile
for (i=0; i<len(trans); i++)
        assign empty string to session_text
        stripped=remove new line character from trans[i]
        parts=split the fields of stripped with space character
        for (i=0; i<len(URLs); i++)
                if parts contain the URLs[i]
                        if length(session_text) >0
                                append 1 to session_text
                        else if len(session_text) =0
                                assign 1 to session_text
                        append 1 to session_text
                else
                        if len(session_text)>0
                                append '?' to session_text
                        else
                                assign '?' to session_text
        write session_text to SesURLMatrix file
        write new line character to SesURLMatrix file
        close SesURLMatrix file
```

Figure 4. 6: An algorithm to create session-URL matrix of the transaction file

The outputs of the algorithm shown in Figure 4.6 is a space separated list of 1 or '?' for each of the transactions. Thus, it could be easily imported to Microsoft Excel after which it would be converted into CSV file format so that the dataset is presented for Weka data mining tool association mining algorithms.

2. Preparing the Transaction File for Sequence Mining

Contrary to the dataset preparation performed for association rule mining, the preparation for sequence mining does not need extraction of some attributes. Transformation of the transaction dataset for sequence mining was accomplished by using all of the first n URLs of each transaction where n refers the maximum number of URLs to be considered by the experiment. The algorithm used for the transformation of the transaction is presented below.

Algorithm: creates ARFF file for sequence mining from transaction file
Input: tr, access transaction file
Output: arff_File, an arff file format of the transaction file
Raw=" "// a string variable to store the session record
NoAttributes=0// The number of selected attributes for the dataset
associatedURLs={}# the diction of URLs that are found important in the association mining
open arff_File, the file onto which the arff format file is to be written to
write "@relation SequenceDiscovery"
write new line character
while not end of attributes
    write "@attribute <attribute name> {attribute values}
associatedURLs=take the URLs that fulfill the minimum support
URLs={read the URLs' full name}
IDs={read the URLs' ID}
Open tr, the transaction file
lines= {read each of the lines in the transaction}
Write to arff_File file "@data\n"
for (line=0; line <len(lines);lien++)
    empty raw
    parts=lines[line]
    stripped={remove new line character from parts}

```
        parts={split stripped with space character}
        if len(parts)<noAttributes
                for (x=0; x<len(parts); x++)
        if URLs contains parts[x]
                if len(raw) >0
                        raw=raw+','+str (associatedURLs[parts[x]])
                else if len(raw)==0:
                        raw=str(associatedURLs[parts[x]])
        else if URLs does not contain parts[x]
                if len(raw)>0
                        raw=raw+','+'?'
                else if len(raw)==0
                        raw='?'
                for (y=len(parts); y<noAttributes; y++)
                        raw=raw+','+'?'
        else if len (parts)>noAttributes
                for (x=0; x<noAttributes; x++)
                        if URLs contains parts[x]
                                if len(raw)>0
                                        raw=raw+','+str (associatedURLs[parts[x]])
                                else if len (raw)==0
                                        raw=str(associatedURLs[part[x]])
                        else if URLs does not contain parts[x]
                                if len (raw)>0
                                        raw=raw+','+'?'
                                else if len (raw)==0
                                        raw='?'
        if raw is not empty
                write raw's content to arff_File file
                write new line character to arff_File file
close arff_File file
```

Figure 4. 7: An algorithm to develop arff format file for sequence mining

4.6. Usage Pattern Discovery

The usage pattern was discovered from the respective file using statistical analyzer, association mining and sequence mining.

4.6.1. Mach5 Analyzer Statistics

The raw log was filtered to get the human-access requests. Then, the filtered log records were fed into Mach5 statistical analyzer. Some of the report's statistics that are crucial for analysis of usage pattern are selected and depicted below.

a. General Statistics

The following table shows the general statistics that have been extracted from the log file.

Item	Value
Time Period	November 01, 2010, 04:31:34 AM to May 08, 2011, 04:43:48 AM
Report generated on	May 14, 2011 at 06:28:47 PM
Average Data Transferred per Hit	75.96 kilobytes
Average Users per Day	635.81
Hits	605600
Total failed requests	(8.76%)
Average Hits per User	5.66
Total Data Transferred	44.09 gigabytes
Hits on Files	436317
Average Data Transferred per User	430.25 kilobytes
Average Hits per Day	3601.12
Incomplete downloads/file requests	35716 (5.87%)
Average Data Transferred per Day	267.15 megabytes
Log spans a period of	189 days
Total Visiting Users	107452

Table 4. 2: General statistics of the log file

b. Most Visited Directories

Among the discoveries of the Mach5 Analyzer is list of most frequently visited directories in the web site. Appendix A shows the most visited directories and their corresponding number of hits and bandwidth.

c. Users' Visiting Time

From the analysis, differences of hits on different days of a week were noticed. The following figure shows this analysis.

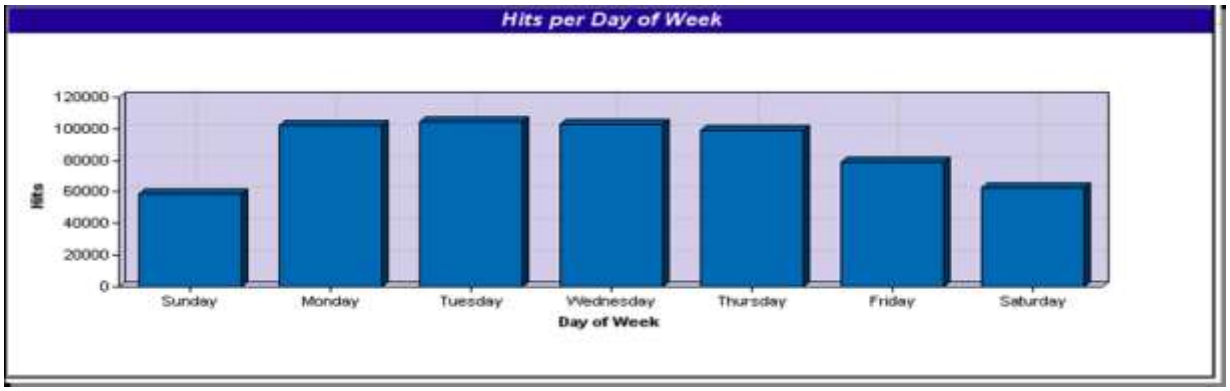


Figure 4. 8: Hits per day of week for the dataset

d. More Frequent Entries to the site

The Mach5 Analyzer showed that some pages are more frequent entry pages. The top 10 pages that are considered as more frequent site entries are presented in the table below.

	Entry Page	Percentage of entry pages
1.	/index.php/home	15%
2.	/ies/index.html	4%
3.	/index.php/academics	3%
4.	/index.php/contacts	3%
5.	/index.php/library-and-museum/library	3%
6.	/index.php/library-and-museum	3%
7.	/index.php/admissions	2%
8.	/index.php/admissions/graduate	2%
9.	/march	2%
10.	/index.php/academics/faculties	2%

Table 4. 3: The frequent entry pages of user sessions

e. More Frequent Site Exits

Like entry pages, the analyzer tool discovered pages through which the site users leave the web site. Table 4.4 below shows the frequent exit pages of the site users.

Exit page	Percentage of the page
/index.php/home	13%
/index.php/component/search/index.html	4%
/ies/index.html	4%
/index.php/library-and-museum/library	3%
/index.php/contacts	2%
/march/index.html	2%
/index.php/admissions/graduate	2%
/libraries/index.html	2%

Table 4. 4: The frequent exit pages of user sessions and their share of the total exit pages

f. Common Errors Encountered

According to the outputs of the analyzer tool, there are errors that users have come across in visiting the web site. Common errors encountered in visiting the official web site of AAU along with the occurrences of the errors are shown at Appendix B of this paper.

4.6.2. Discovering Association Rules

Apriori algorithm was used for the discovery of association rules. Here is explanation of some of the parameters of the algorithm.

car: It is a parameter to opt for general association rules or class association rules (CAR). In this regard, if car parameter is set to true class association rules are mined and general association rules are mined if car parameter is set to false.

classIndex: This parameter is used to specify the index of the class attribute among the available set of attributes in a dataset. If the parameter is set to -1, the last attribute is taken as class attribute.

delta: The value of this parameter is a factor by which the support is iteratively decreased until minimum support is reached or required number of rules has been generated.

lowerBoundMinSupport: This is a parameter to specify lower bound value for minimum support

metricType: This is used to set the metric type by which discovered rules are ranked. The parameter values are confidence, lift, leverage and conviction. Confidence measures the proportion of the items covered by the premise that are also covered by the consequence. It is known that confidence is the only metric type to mine class association rules. m refers the ratio of confidence to the proportion of all items that are covered by the consequence. Indeed, this metric type is a measure of the importance of the association independent of support. Leverage, on the other hand, is a metric type which is computed as a proportion of additional items covered both the premise and consequence above those expected if the premise and consequence were independent of each other. Conviction is another measure of departure from independence.

minMetric: This parameter is used to set a minimum metric score on which only rules with scores higher than the score are considered.

numRules: The parameter that is used to limit the maximum number of rules to find from the discovery.

outputItemSets: The parameter that is used to opt for the visualization of the itemsets or not. If the parameter is enabled, the itemsets are output.

removeAllMissingCols: It is a parameter to choose for removal of columns of which all values are missing.

significanceLevel: A parameter to test the significance for confidence metric only.

upperBoundMinSupport: A parameter to set an upper bound for minimum support which iteratively decreases.

verbose: A parameter to opt for running the algorithm in verbose mode.

The algorithm was run for a number of times by varying the lowerBoundMinSupport parameter values. Then, two runs were selected as list of rules for better support value. The runs are attached at Appendix E.

4.6.3. Discovering Sequences

PredictiveApriori algorithm was used for discovering of the frequent access sequences. The CAR, classIndex and numberOfRules parameters are used in the algorithm.

For this experiment, parameter CAR is enabled and classIndex is set to -1 so that class association rule is developed for a combination of requests in the sequence.

Consideration of the records in the transaction showed that six is the average number of request URLs in a user session. Thus, predictiveApriori algorithm was applied on the dataset for the first six requests of the transaction.

As it has been shown using the statistical analyzer, most of the user sessions start with /index.php/home, the home page of the site. Moreover, most of the users leave the site there. So, the researcher is not interested on attribute 1 of the dataset. The running output is attached in Appendix F.

4.7. Pattern Analysis

The usage analysis was done by combining statistical analysis and data mining pattern discovery results. Each of these is elaborated in the next subsections.

4.7.1. Statistical Analysis

a. **General Statistics:** The web server accepted 605, 600 hits from 107,452 users. Here, user stands for an ID for a user session-a transaction rather than actual user. On average, each user has been expected to have 6 hits to the web site. Moreover, about 638 users used the web site per day.

b. **Frequently visited directories:** The /index.php page received the largest hits. Indeed, it is the home page of the web site and not astonishment. Downloads, admissions, AAU staff load, academics/faculties and library & museum respectively held the second, third, fourth, fifth and six ranks of the hit statistics. The statistics of school of information science and administrator pages show frequent hits for the pages. However, these two statistics are not interesting for the research. Firstly, most of the hits on SIS are expected from the page designers as long as only few weeks elapsed after the page was published. Secondly, the administrator page is used by the site designers and/or site maintenance users only.

c. **Users' visiting time:** For the log records considered, around 100,000 hits have been requested on Mondays, Tuesdays, Wednesdays and Thursdays individually. On Fridays, there is some decrements and about 80,000 hits have been requested. At the weekends, there is further decrease in request size and about 60,000 hits have been registered.

d. **Frequent site entry pages**: 40% of the entry pages are occupied by only ten pages. Larger portion of the user session records, i.e. 15% start with /index.php page. This page is the home page of the site and may infer that most of the site visitors arrive at the site directly through the home page by either typing the full URL of the site onto the address bar or get the home page of the official site on search engines. Next to the home page of the site, /ies/index.html-the Institute of Ethiopian Studies took 4% of the entry pages. Meanwhile, /index.php/academics, /index.php/contacts, and /index.php/library-and-museum/library, and /index.php/library-and-museum got 3% of the site entry each. Then, /index.php/admissions, /index.php/admissions/graduate, /March and /index.php/academics/faculties took 2% of the site entry each.

e. **Frequent site exit pages:** 33% of the exit pages of the site visit are held by 8 pages. 13% of the user sessions end with /index.php page. This implies that most of visitors of the site stop their visit before moving further from the home page of the site. /index.php/component/search/index.html, /ies/index.html and /index.php/library-and-museum/library cover 4%,4% and 3% of the exit pages

respectively. At the mean time, each of /index.php/contacts, /march/index.html, /index/php/admissions/graduate and /libraries/index.html took 2% of the exit pages

f. **Common Errors Encountered:** Users of AAU official web site have been facing errors of unavailability of requested pages while surfing the site. List of common errors is attached at appendix 1.

From the statistical report, /index.php/administrator/index.php shows the maximum error occurrence. But, administrators are special users and this statistics could not show the average user request. Thus, this could not be considered as a valid usage pattern. The same is true for the pages of /march/media/system/js/opened, /_vti_bin/shtnl.exe/_vti_rpc, /_vti_inf.html.

Page of /index.php/aboutaau/presidents-office, /ictglossary/amharic/a_index.php, /index.php/aboutaau/presidents-office/-communications-team/contact-us, /faculties/fac_phar/his_phar.php, and /index.php/aboutaau/presidents-office/-communications-team/university-partnerships occupied the top five error requests. The error is more likely due to broken links or unintentionally page blocking.

4.7.2. Analysis of the Association Mining Based Discovered Patterns

- It was found out that 96% of the users who requested /index.php/aboutaau page and /index.php/aboutaau/graduate-studies-and-research-office also requested /index.php/admissions page: Most of the site users visited graduate studies and research office page of the about AAU page and admission pages together. This is very interesting for the fact that admission pages are visited along with the graduate and research office page for verification of admission requirements and other specifications to pursue a graduate level study.
- The finding revealed that 94% of the users who visited /index.php/campus-life and /index.php/campus-life/housing-a-dining also visited /index.php/administration page: A good number of users of the site accessed housing and dining page of the student

services page along with administration page. This rule could be interesting by the expectation that student services of the university is one of the administration tasks in the university In this regard, the site users visit administration page so as to fulfill their information need .in these aspects.

- With regard to the discovered patterns, 98% of the site visitors who accessed /index.php/academics and /index.php/admissions pages also visited /index.php/admissions/apply-online: Users who access academic pages and admissions pages are more likely prospective student of the university. Thus, those students more likely visit the online application page of the web site. Therefore, this rule seems very interesting.
- It was discovered that 98% of the site users who visited /index.php/aboutaau and /index.php/aboutaau/campus-guide also visited /index.php/vacancy-announcements page: visitors who access campus guide page of the about AAU will more likely search for vacancies and announcements in the university as long as campus guide provides guidelines on the campus life. Thus, this rule may be very interesting.
- It was found out that 90% of users of the site who accessed /index.php/campus-life also accessed /index.php/library-and-museum page: Library service is one of the services to students in the university. Thus, the site visitors may expect to get library services inside student services page. So, the visit of student services along with museum and library pages must be an interesting rule.
- From the pattern discovery outputs, 92% of the users who requested /index.php/library-and-museum/library also requested /index.php/evaluation-feedback page: The evaluation feedback form is a form through which the preparatory courses of the newly arranged graduate study programs are evaluated by those who took the courses in pre-enrollment of a certain graduate level study in the university. Thus, the form has nothing to do with the library page of the site. It has been requested along with the library pages for the fact that the form was published in inappropriate pages. Therefore, this rule is less interesting.
- The findings revealed that 90% of the site users who visited /index.php/admissions also accessed /index.php/evaluation-feedback page: The evaluation feedback form is expected to be visited by graduate level students who have been enrolling in the

preparatory courses upon admission to a program. So, there is less probability of visiting admissions page and evaluation pages of the site. Therefore, this is less interesting rule.

- According to the findings of the pattern discovery experimentation, 89% of the users who accessed /index.php/academics also accessed /index.php/contacts page: The contacts page contains list of contact details of the available faculties/schools/institutes. Meanwhile, the academics page is comprised of pages of faculties/schools/institutes detail description but contact. So, visitors of the academics page may be expected to visit contacts page for any sort of communication to the respective programs. On this aspect, academics page and contacts page are more likely to be visited together and the rule is interesting.
- It was also found out that 99% of the site users who accessed /index.php/aboutaau and /index.php/aau-bpr-all also accessed /index.php/administration page: BPR (Business Process Reengineering) may be one of the current issues to be dealt in about AAU page. Besides, more information about BPR is expected to be available at the policies and procedures page of administration page. So, it is expected to visit about AAU, AAU BPR and administration pages of the site together. However, BPR may be a short term topic in the site which makes validation of the rule somewhat difficult. Rather, it would be an interesting rule if the BPR is substituted by any topics which need some control from the higher administration departments.
- It was discovered that 98% of the site users who accessed /index.php/aboutaau and /index.php/campus-life also request for /march page:March Project is an project in the university which works in reduction of HIV/AIDS epidemic to bring behavioral changes in the university community. It is very interesting to get a rule that entails the coming of such project visit in students' services pages and about AAU pages. Therefore, the rule is very interesting.
- From the outputs of pattern discovery, 88% of the site visitors that accessed /index.php/aboutaau and /index.php/view-blog also accessed /index.php/administration: Site users that access about AAU page and president's blog are more likely surfing for some administrative tasks/achievements. Thus, visit

for administration pages is expected to augment the visitors' satisfaction on the site. Therefore, the rule is interesting.

- It was also revealed that 86% of the site users that accessed /index.php/siter-map also visited /index.php/academics: The site map of the site is expected to contain higher level pages in hierarchy. User of the site may want to directly go to their page of interest from the site map's links to such pages. In case the site map does not fulfill this feature, visitors are enforced to choose other path of the site. In this rule, the site map of the site is frequently accessed along with academics page. On the occasion of the aforementioned cases, the rule is interesting.

4.7.3. Analysis of the Sequence Mining Based Discovered Patterns

The sequence mining discovered rules are analyzed here under.

On the run that involves second and third requests:

- It is 97% confident that the second request in a session is /index.php/campus-life and the third request is /index.php/.
- It is 96% confident that the second request in a session is /index.php/aboutaau and the third request is /index.php/administration.
- It is 94% confident that the second request in a session is /index.php/library-and-museum and the third request is /index.php/library-and-museum/library

On the run that involves the third and fourth requests:

- It is 99.8% confident that the third request is /index.php/admissions page and the fourth one is /index.php/admissions/apply-online
- It is 94.6% confident that the third request is /index.php/academics page and the fourth one is /index.php/admissions
- It is 94% confident that the third request is /index.php/campus-life page and the fourth one is /index.php/admissions
- It is 94% confident that the third request is /index.php/view-blog page and the fourth one is /index.php/administration

On the run that involves the fourth and fifth requests:

- It is 97.6% confident that the fourth request is /index.php/admissions page and the fifth one is /index.php/admissions/apply-online.
- It is 96.9% confident that the fourth request is /index.php/administration page and the fifth one is /index.php/academics.

On the run that involves the fifth and sixth requests:

- It is 90.6% confident that the fifth request is /index.php/academics and the sixth request is /index.php/registrar

Combining the above: rules, here are the common navigation sequences.

/index.php/academics →/index.php/admissions →/index.php/admissions/apply-online Lots of the site users who visited one of the academics pages visited admissions page soon followed by online application page;

/index.php/aboutaau →/index.php/view-blog →/index.php/administration: The presidents blog followed by administration pages are visited after the about AAU page is visited.

/index.php/campus-life →/index.php/administration: More frequently, administration pages are visited once the students' services pages are visited.

/index.php/library-and-museum →/index.php/library-and-museum/library: The library section of the library and museum is well visited.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The research attempted to apply web usage mining techniques on access log files of AAU official web site and to evaluate the web site. The study was conducted in three phases, namely raw log collection and log preparation, pattern discovery and pattern analysis.

The raw log data preparation was major task of the research owing to the fact that the raw log contains non-human requests, automatically downloaded graphics or failed requests. Moreover, the unprepared log could not be used for the pattern discovery phase which, in turn, consumed sequence of transformation as per the pattern discovery in hand. Next, Mach5 statistical analyzer was applied on the filtered log to explore general statistics, most visited directories, users' visiting time, more frequent entry and exit pages and common errors encountered. Apriori algorithm of Weka was used to discover the available association rules amongst the URLs of the site. Then, sequential mining was performed on behalf of the level wise rule generation of predictiveApriori algorithm of Weka. Then, each of the rules was analyzed in communication with the web site designers to develop appropriate rules that could be utilized for future components of the site.

According to the statistical analyzer tool employed, most of the users of Addis AAU official web site start navigation at the home page of the web site which fits the claims of Mekonnen (2009). That is, most of the users either directly write the full URL of the home page of the site onto the address bar or get the home page of the site from search engines display. Most of these users stop their visit at the same page.

Page of Institute of Ethiopian Studies program has been visited more frequently than other program pages. It is one of the more frequent pages through which users of the web site start their navigation. This is more probably due to the fact that the name of the

institution in the content of the web site contains words that are more frequently used by search engine users.

Regarding the daily users' visit trend, relatively higher hits are considered on Mondays through Thursdays. Meanwhile, the least hits are recorded at the weekends and mild hits are recorded on Fridays.

A number of requests for the pages of about AAU return the error of unavailability of some pages. This includes president's office, president's office/communication team, president's office/reform team and president's office/resource mobilization team. Similarly, administration and human resource and policies and procedures pages of the administration page return a response of unavailability of the pages. These imply that either the content is blocked unintentionally or the link to such pages is broken.

According to the association mining experiments, most of the users of the official web site who visited academics page would more probable visit library page. This is due to search of materials that could be suitable for a certain academic program in hand. Similarly, academics page and admission page have been accessed together followed by online applications for programs. This is due to the fact that users are interested to know the admission criteria, date, payments and other issues as per academics programs and try the online application.

Most of the site users who visit the about AAU page also visit academic pages. This is due to interest to visit of a certain academic program listed in the about AAU page. Furthermore, vacancies and announcements are visited along with about AAU page. This is due to the expectation of site users to get announcements and vacancies being at about AAU page.

Student service and administration pages are requested together by many of the site users. This is owing to the fact that student service is expected as one of the administration tasks in the institution.

The evaluation form of the newly arranged preparatory courses of the graduate study has been frequently accessed with lots of pages that have nothing to do with the courses assessment. This is due to the inclusion of the link in unnecessary places.

5.2. Recommendations

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

Improvement of the web site:

- For ease of users who visit deeper pages than just the home page of the web site, the depth complexity and readability of the pages' links need to be assessed. Besides, the user-interface of the site should be checked by the designers of the site in case the site has features that made it less user-friendly. The user-friendliness of the interface might need a separate detailed study.
- Institute of Ethiopian Studies and AAU March Project have been visited by lots of the site users. There must be something good behind this statistics. Thus, the site's designers are recommended to consider these pages for any sort of smart components in these pages so that the same could be done to other departments' pages.
- Searching of electronic library items returns problems much of the time. Thus, the site designers are recommended to verify the availability of valid hyperlinks or legible texts of hyperlinks at this spot.
- It is advisable to put contact details of department as part of the departments' page rather than a list of contacts at a certain point of the web site.
- It is recommended to create a link from specific academics to e-resources (library services) of a department.
- There are pages that are expected to be accessed but not accessed in the real users. Therefore, lots of amendments are expected from web site designers to alleviate such problems. Also the same if some pages need to be removed away from the site.
- The site designers are recommended to create links to announcement and vacancies pages from about AAU page.

- It is advisable to make academic lists in about AAU page clickable so as to arrive at desired academic programs upon need while in about AAU page.
- The researcher recommends incorporating students' services as a component in the administration page or having links to students' service from the administration page.

Further web usage analysis research

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

- In this work, only access log records of the web server log are utilized. Better usage analysis could be made upon inclusion of other logs such as error log and SSL logs in addition to access log.
- The previous web usage analysis studies have brought completely different usage patterns for different raw log data. Thus, future researchers are recommended to develop an adaptive and dynamic web usage analysis study to alleviate such problems.
- Time-stamp heuristic is used for the development of user session files. Future researchers could use site map of the web site along with the web log files to make use of other heuristics like referrer page approach so that more appropriate patterns could be discovered.
- The thesis used a six month duration web log files. If future researchers use log files of longer period of time, trusted and more preferable usage patterns could be discovered that would perfectly show the actual users' usage pattern.

REFERENCES

Abraham, A. 2005. Natural computation for business intelligence from web usage mining. *Proceeding of Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (*SYNAC2005*).*

Addis Ababa University. 2008. Official web site of Addis Ababa University. Web site: http://www.aau.edu.et; accessed on December 20, 2010.

Adomavicius, G. and A. Tuzhilin. 2001. Using data mining methods to build customer profiles. *IEEE Computer* 34, no. 2.

Agrawal, R. and R. Srikant. 1995. Mining sequential patterns. *In Proceedings of the 11th International Conference Data Engineering (ICDE'95*), Taiwan.

Agrawal, R. and R. Srikant.1994. Fast Algorithms for mining association rules. In *Proceeding of the 20th Very Large DataBases Conference* (*VLDB*), Chile.

Araya, S., M. Silva and R. Weber. 2004. A methodology for web usage mining and its application to target group identification. *Fuzzy Sets and Systems* 148: 139-152.

Baglioni, M, U. Ferrara, A. Romei, S. Ruggieri and F. Turini. 2003. Preprocessing and mining web log data for web personalization. 8th Italian Conference on Artificial Intelligence 2829.

Batista, P. and J. Silva. 2002. Mining web access logs of an on-line newspaper. 12th International Meetings of the Euro Working Group on Decision Support Systems.

Borges, J. and M. Levene. 1999. Data mining of user navigation patterns. In *Proceedings of the Workshop on Web Usage Analysis and User Profiling* (WEBKDD), California.

Büchner, A. and M. Mulvenna. 1998. Discovering internet marketing intelligence through online analytical web usage mining. *ACM SIGMOD Record* 27, no. 4.

Büchner, A., M. Baumgarten, S. Anand, M. Mulvenna and J. Hughes. 1999. Navigation pattern discovery from internet data. In *Proceedings of the Workshop on Web Usage Analysis and User Profiling* (*WEBKDD*), California.

Catledge, L. and J. Pitkow. 1995. Characterizing browsing behaviors on the World Wide Web. *Computer Networks and ISDN Systems* 27, no. 6.

Chakrabarti, S. 2000. Data Mining for Hypertext: A Tutorial Survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on* Knowledge *Discovery and Data Mining, ACM* 1, no. 2.

Chakrabarti, S., B. Dom, and P. Indyk. 1998. Enhanced hypertext categorization using hyperlinks. SIGMOD ’98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York: ACM Press.

Chen, M., J. Park and P. Yu. 1998. Efficient data mining for path traversal patterns. IEEE Transactions on Knowledge and Data Engineering 10, no. 2.

Chen, Z., A. Fu, and F. Tong. 2002. Optimal algorithms for finding user access sessions from very large web logs. *Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Taipei.

Chitraa, V. and A. Davamani. 2010. A survey on preprocessing methods for web usage data. *International Journal of Computer Science and Information Security* 7, no. 3: 78-83

Cooley, R., B. Mobasher, and J. Srivastava. 1997. Web mining: information and pattern discovery on the World Wide Web. *Proceedings of the ninth IEEE International Conference tools with AI (ICTAI ’97*), 558-567.

Cooley, R., B. Mobasher and J. Srivastava. 1999. Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems* 1, no. 1, 5-32.

CRISP-DM Consortium. 2000. Cross Industry Standard Process for Data Mining. Web site: http://www.crisp-dm.org; accessed on 22nd of December 2010.

Dai, H. and B. Mobasher. 2003. A road map to more effective web personalization: integrating domain knowledge with web usage mining. *In Proceedings of the International Conference on Internet Computing (IC),* Nevada.

David, H., H. Mannila and P. Smyth. 2001. Principle of data mining. Massachusetts: *MIT Press*.

Davison, B. D. 2001. A web caching primer. *IEEE Internet Computing* 5, no. 4.

Deshpande, S. and V. Thakare. 2010. Data mining system and application: a review. *International Journal of Distributed and Parallel systems (IJDPS)* , no.1:32-44.

DomainTools. 2010. http://www.domaintools.com ; accessed on 16th December 2010.

Dunham, M. 2002. Data mining, introductory and advanced topics. Prenice Hall.

Eirinaki M. and M. Vazirgiannis. 2003. Web mining for web personalization. ACM Transactions on Internet Technology 3, no. 1: 1-27.

Etminani, K., M. Akbarzadeh and N. Yanehsari. 2009. Web usage mining: users' navigational patterns extraction from web logs using ant-based clustering method. IFSA-EUSFLAT.

Etzioni, O. 1996. The World Wide Web: quagmire or gold mine. *Communications of the ACM* 39, no. 11:65-68.

Fawcett, T. and F. Prouost 1999. Activity monitoring: noticing interesting changes in behavior. *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, CA.

Fayyad, U., P. Gregory and S. Padhraic. 1996. From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*.

Grace, J., V. Maheswari and D. Nagamalai. 2011. Analysis of web logs and web user in web mining. *International Journal of Network Security and Its Application 3, no. 1*.

GroupLens Research. Web site: http://www.cs.umn.edu/Research/GroupLens/; accessed on: February 02, 2011

Han H. and R. Elmasri. 2004. Learning rules for conceptual structure on the web. *Journal of Intelligent Information Systems* 22, no. 3

Han, J. and M. Kamber. 2006. Data mining: concepts and techniques, second edition. San Francisco: Morgan Kaufmann Publishers.

Hay, B., G. Wets and K. Vanhoof. 2003. Segmentation of visiting patterns on web sites using a sequence alignment method. *Journal of Retailing and Consumer Services* 10.

He, D. and A, Goker. 2000. Detecting session boundaries from web user logs. *In proceedings of the BCS-IRSG 22nd Annual Colloquium and Information Retrieval Research*.

Helen, T. 2003. Application of data mining technology to identify significant patterns in census or survey data: the case of 2001 child labor survey in Ethiopia. Thesis paper, Addis Ababa University.

Henock, W. 2002. Application of data mining techniques to support customer relation management at Ethiopian Airline's. Thesis paper, Addis Ababa University.

Hong, J. and J. Landay. 2001. WebQuilt: a framework for capturing and visualizing the web experience. In *Proceedings of the 10th International World Wide Web Conference*, Hong Kong.

Hou, J. and Y. Zhang. 2003. Effectively finding relevant web pages from linkage information. *IEEE Transactions on Knowledge Data Engineering* 15, no. 4.

Hu, W., X. Zong, C. Lee, and J. Yeh. n.d. World Wide Web usage mining systems and techniques. *SYSTEMICS, CYBERNETICS AND INFORMATICS* 1, no. 4.

Hu, X. and N. Cercone. n.d. A data warehouse/OLAP framework for web usage mining and business intelligent reporting.

Huang, Y. 2003. Infrastructure, data cleansing and mining for support of scientific simulation. Department of Computer Science and Engineering, University of Notre Dame, Indiana.

Huysmans, J., B. Baesens and J. Vanthienen. 2003. Web usage mining: a practical study. Katholieke Universiteit Leuven, Department of Applied Economic Sciences.

Ivory, M. and M. Hearst. 2002. Improving Web Site Design. *IEEE Internet Computing* 6, no. 2.

Jochen, H. 2000. Algorithms for association rule mining: a general survey and comparison. Available at: http://www.cs.sfu.ca/coursecentral/884/G2/2002-3/references/high00.pdf; accessed on January 30, 2011.

Kleinberg, J., R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. 1999. The web as a graph: measurements, models and methods. *Lecture Notes in Computer Science* 1627.

Koh, C. and G. Tan. n.d. Data mining applications in healthcare. *Journal of Healthcare Information Management,* 2:64-72.

Kosala, R. and H. Blockeel. 2000. Web Mining Research: A Survey. *SIGKDD Explorations* 2, No. 1.

Lan, B., S. Bressan, and B. Ooi. 1999. Making web servers pushier. *In Proceedings of the Workshop on Web Usage Analysis and User Profiling*, California.

Logging Control in W3C httpd. 1995. The common log file format, accessed at http://www.w3.org/pub/WWW/Daemon/User/Config/Logging.html; accessed on December 05, 2010.

Lydon, B. and T. Fennell. 2003. Web usability: its impact on human factor and consumer search behavior. *Human-Computer Interaction: Theory and Practice(*part I) 1: 793-797

Madria, S., S. Bhowmick, W. Ng, and E. Lim. 1999. Research issues in web data mining. *Data Warehousing and Knowledge Discovery*.

Masseglia, F., P. Poncelet, and R. Cicchetti. 1999. An efficient algorithm for web usage mining. *Networking and Information Systems Journal (NIS)* 2, no.5.

Mekonnen, T. 2009. Web usage pattern discovery using data mining and statistical analysis: the case of AAU official web site, Master Thesis, Addis Ababa University.

Mobashar, B., H. Dai, T. Luo and M. Nakagawa. 2002. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, Kluwer Publishing 6, no. 1.

Mobashar, B., R. Cooley and J. Srivastava. 1999. Creating adaptive web sites through usage based clustering of URLs. *Knowledge and Data Engineering Workshop.*

Mobasher, B., R. Cooley and J. Srivastava. 2000. Automatic personalization based on web usage mining. *Communications of the ACM* 43, no. 8.

Nanopoulos, A. , D. Katsaros and Y. Manolopoulos. 2001. Effective prediction of web-user accesses: a data mining approach. *In Proceedings of the Workshop on Mining Log Data across All Customer Touchpoints (WEBKDD),* California.

Nanopoulos, A., and Y. Manolopoulos. 2000. Finding generalized path patterns for web log data mining. J. Stuller et al. (Eds.): *ADBIS-DASFAA*, LNCS 1884:215-228.

Netcraft. 2010. Netcraft Web Server Survey. Available at http://www.netcraft.com/survey/archive.html; accessed on 16th December, 2010.

Nina, S.., M.Rahaman, I.Bhuiyan, U.Ahmed. 2009. Pattern discovery of web usage mining. *International Conference on Computer Technology and Development*.

Pal, S., V. Talwar and P. Mitra. 2002. Web mining in soft computing framework: relevance, state of the art and future directions. *IEEE Transaction Neural Networks* 13, no. 5.

Pei, J., J. Han, B. Mortazavi-asl, and H. Zhu. 2000. Mining access patterns efficiently from web logs. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD*).

Perkowitz, M. and O. Etzioni. 1998. Adaptive web sites: automatically synthesizing web pages. *In 15th National Conference on Artificial Intelligence,* Madison, WI.

Perkowitz, M. and O. Etzioni. 2000. Towards adaptive web sites: conceptual framework and case study. *Artificial Intelligence*, 118:245-275.

Pirolli, P., J. Pitkow and R. Rao. 1996. Silk from and sow's ear: extracting usable structure from the web. *Conference on Human Factors in Computing Systems – 96*, Vancouver.

Prasanna, C., M. Kishore, P. Rao, L. Sandeep and D. Lakshmi. 2010. An approach for restructuring of web pages. *International Journal of Computer Science and Network Security* 10, no. 2.

Qiang, Y. and H. Henry. 2003. Web-log mining for predictive web caching. *IEEE Transaction on knowledge and data engineering* 15, no. 4: 1050-1053.

Ratheke, C. and V. Schreiweis. 2003. Interaction design elements to improve information presentation on web pages. *Human-Computer Interaction: Theory and Practice* (Part I) 1: 843-846.

Rygielski, A., J. Wang, D. Yen. 2002. Data mining techniques for customer relationship management. *Technology in Society* no. 24: 483–502.

Sarwar, B., G. Karypis, J. Konstan, and J. Riedl. 2000. Analysis of recommender algorithms for e-commerce. *In Proceedings of the ACM Electronic Commerce Conference*, 158-167.

Schafer, J., J. Konstan and J. Riedl. 2000. Electronic commerce recommender applications. *Journal of Data Mining and Knowledge Discovery* 5, no. 1.

Schechter, S.., M. Krishnan and M. Smith. 1998. Using path profiles to predict HTTP requests. *In Proceedings of the 7th International World Wide Web Conference*, Australia.

Shen, L., J. Chneg, F. Ford, and V. Makedon. 1999. Mining the most interesting web access associations. *webNet 2000-World Conference.*

Spiliopoulou, M. and L. Faulstich. 1998. WUM: A tool for web utilization analysis. In *Proceedings of the Workshop on the Web and Databases* (*WEBDB*), Spain.

Spiliopoulou, M., B. Mobasher, B. Berendt and M. Nakagawa. 2003. A framework for the evaluation of session reconstruction heuristics in web usage analysis. *INFORMS JOURNAL ON COMPUTING* 15, no.2: 171-190.

Spiliopoulou, M. 2000. Web usage mining for site evaluation: making a site better fit its users. *Communications of ACM* 43, no. 8: 127-134.

Srivastava, J., R. Cooley and P. Tan. 2000. Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations* 1, no. 2.

Suneetha, K. and D. Krishnamoorthi. 2009. Identifying user behavior by analyzing web server access log file. *International Journal of Computer Science and Network Security* 9, no. 4: 327-332.

Tan, P. and V. Kumar. 2002. Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery 6, 9–35.*

Tesfaye, B. n.d. University wide networking initiatives: AAUNet project development and management, Addis Ababa University.

Teshome, W. 1990. The development of higher education and social change, an Ethiopian experience. East Lansing, Michigan: Michigan State University Press.

Thakare, S and Z. Gawali. 2010. An effective and complete preprocessing for web usage mining. *International Journal on Computer Science and Engineering 02, no: 03:848-851.*

W3C. Extended Log File Format. 2010. *W3C Working Draft WD-logfile-960323*. http://www.w3.org/TR/WDlogfile.html; accessed on 16th December 2010.

W3C. n.d. Logging Control in W3C. http://www.w3.org/Daemon/User/Config/Logging.html; accessed on 16th December, 2010.

WCA. n.d. Web Characterization Terminology & Definitions. Available at http://www.w3.org/1999/05/WCA-terms/; accessed on 6th of February, 2011.

Xing, D. and J. Shen. 2004. Efficient data mining for web navigation patterns. *Information and Software Technology,* 46.

Zaiane, O., M. Xin and J. Han. 1998. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. *In Proceedings of Advances in Digital Libraries (ADL)*, California.

APPENDICES

Appendix A: Most Visited Directories

Directory Name	Hits	Bandwidth
/index.php/home	202510	8 gigabytes
/downloads/	34170	21.43 gigabytes
/index.php/admissions/	27240	688.25 megabytes
/aau_staff_load/	25287	565.76 megabytes
/index.php/academics/faculties/	20668	602.13 megabytes
/index.php/library-and-museum/	17356	482.25 megabytes
/sis/	12545	4.85 megabytes
/index.php/academics/	11504	347.57 megabytes
/index.php/aboutaau/	9106	274.54 megabytes
/index.php/administrator/	9055	0.00 bytes
/index.php/component/search/	8954	1.41 megabytes
/index.php/component/content/article/1-latest-news/	8354	185.23 megabytes
/aau_staff_load/admin/	7849	280.06 megabytes
/proceeding2009/	7407	2.88 gigabytes
/announcement/	6265	233.97 megabytes
/index.php/campus-life/	6103	245.37 megabytes
/index.php/academics/colleges/	5678	158.60 megabytes
/itphd/	5483	31.14 megabytes
/ies/downloads/	5482	530.62 megabytes
/index.php/administration/	5073	129.03 megabytes
/index.php/academics/institutes/	3459	114.03 megabytes

Directory Name	Hits	Bandwidth
/march/	3212	63.43 megabytes
/index.php/linguistics/	3092	196.02 megabytes
/images/stories/	2996	1.04 gigabytes
/index.php/view-blog/viewpost/	2996	247.93 megabytes
/index.php/academics/schools/	2959	70.27 megabytes
/ies/	2889	39.47 megabytes
/index.php/academics/63-faca-institute-college/	2844	85.80 megabytes
/index.php/academics/38-departments-a-programs/	2323	67.95 megabytes
/index.php/home/1-latest-news/	2099	53.84 megabytes
/nprc/	2074	12.29 megabytes
/index.php/chemistry/	2009	67.76 megabytes
/index.php/economics/	1817	84.05 megabytes
/change/	1807	32.16 megabytes
/index.php/linguistics/linguisticsprograms-/	1794	94.45 megabytes
/index.php/sociology-and-social-anthropology/	1732	70.52 megabytes
/energy_center/ec_fot-aau_files/	1724	87.45 megabytes
/index.php/earth-sciences/	1690	42.51 megabytes
/index.php/computer-science/	1658	62.77 megabytes
/workshop/	1628	18.32 megabytes
/index.php/biology/	1545	58.64 megabytes
/index.php/component/user/	1543	31.90 megabytes

Directory Name	**Hits**	Bandwidth
/index.php/political-science-and-international-relations/	1521	98.80 megabytes
/index.php/philosophy/	1467	119.99 megabytes
/administrator/	1435	12.43 megabytes
//feedback_form/	1381	5.38 megabytes
/igs/	1361	199.59 megabytes
/mtadm/html/	1342	9.90 megabytes

Appendix B: Frequent Error Pages

Request	Occurrences
/index.php/administrator/index.php	9055
/march/media/system/js/opened	2913
/index.php/aboutaau/presidents-office	1843
/faculties/sc/index.html	552
/ictglossary/amharic/a_index.php	292
/_vti_bin/shtml.exe/_vti_rpc	291
/_vti_inf.html	291
/index.php/aboutaau/presidents-office/-communications-team/contact-us	268
/faculties/fac_phar/his_phar.php	268
/index.php/aboutaau/presidents-office/-communications-team/university-partnerships	201
/index.php/academics-/faculties/faculty-of-science	196
/index.php/downloads/callforpapers.pdf	188
/faculties/sc/cse/home.html	187
/research/ies/index.htm	186
/faculties/info/cs/index.htm	183
/index.php/academics-/faculties/faculty-of-business-and-economics	182
/index.php/academics-/faculties/faculty-of-technology	180
/research/ies/jes.htm	171
/research/ies/indexn.htm	165
/index.php/academics-/faculty-of-technology/	164

Request	Occurrences
/downloads/addis_ababa_university_icto_2003_action_plan-1.pdf	161
/index.php/aboutaau/presidents-office/-communications-team/international-professors	153
/index.php/academics-/faculties/faculty-of-medicine	153
/downloads/aauictdictstartegicplan.pdf	148
/index.php/aboutaau/presidents-office/-communications-team/team-members	144
/index.php/academics-/faculties/faculty-of-law	141
/contact.php	138
/index.php/academics-/faculties/faculty-of-informatics	135
/index.php/aboutaau/presidents-office/-communications-team/website	132
/faculties/sc/math/math.htm	131
/faculties/fbe/econ/index.htm	126
/index.php/academics-/institute-of-language-studies/department-of-foreign-language-a-literature	125
/index.php/aboutaau/presidents-office/-communications-team/internal-communications	120
/index.php/aboutaau/presidents-office/reform-team/team-members	119
/faculties/linguistics/wocal.htm	118
/faculties/sc/bse/wildlife.html	118
/research/idr/	115
/index.php/academics-/colleges/college-of-commerce	114

Request	Occurrences
/index.php/home/994	114
/index.php/aboutaau/presidents-office/-communications-team	114
/faculties/sc/biology/bioherbarium.htm	111
/index.php/-international-conference-on-educational-research-for-development-13th-15th-may-2009	110
/faculties/index.php	109
/index.php/aboutaau/presidents-office/resource-mobilisation-team/team-members	102
/index.php/academics-/colleges/college-of-social-science	101
/webnews/	101
/index.php/aboutaau/presidents-office/resource-mobilisation-team/5000-in-10/teach-at-aau	101
/crossdomain.xml	99
/index.php/aboutaau/presidents-office/resource-mobilisation-team/friends-of-aau/aims-and-objectives-fri	98
/dspace	98

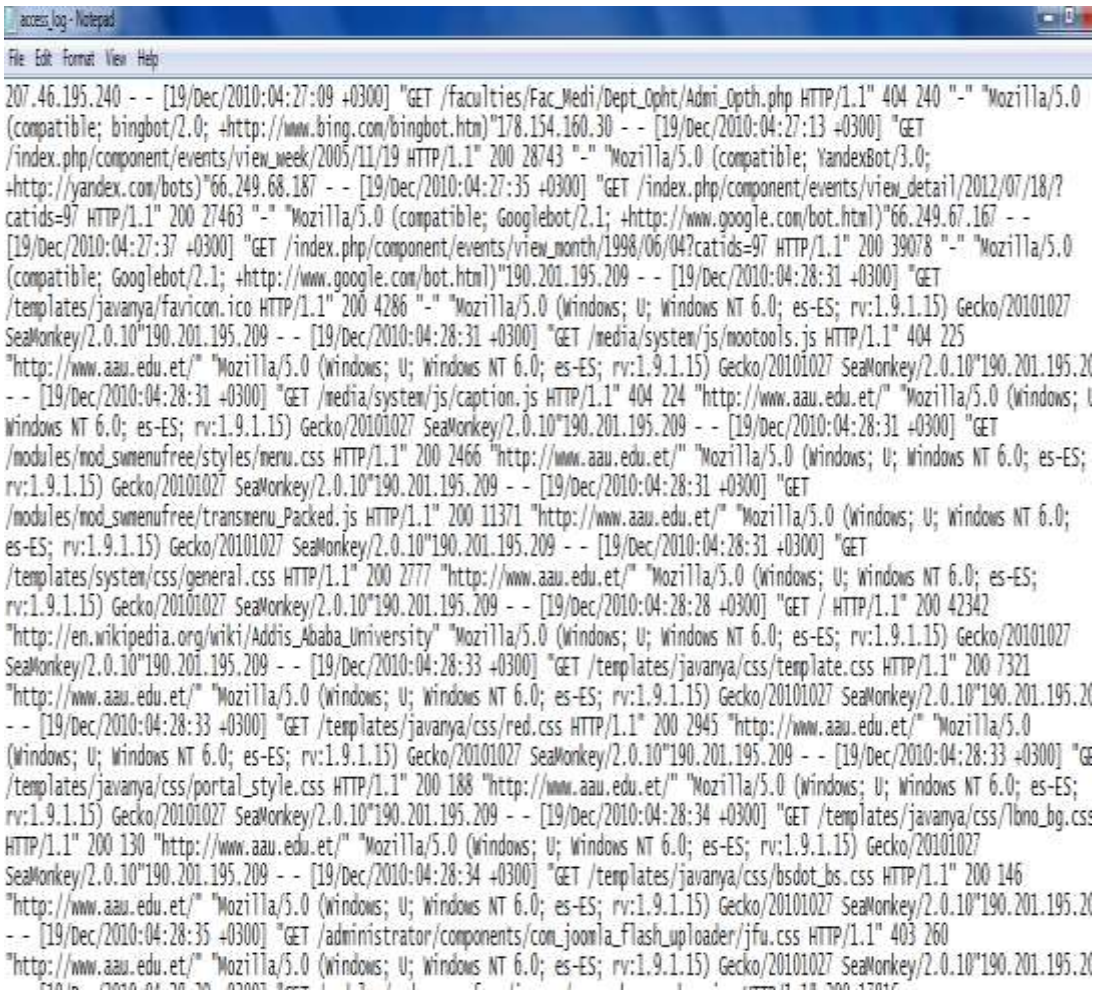
Appendix C: The most common URLs and their Representation

URL …	Full URL
1.	/
2.	/index.php/academics/faculties/faculty-of-informatics
3.	/index.php/academics/faculties/faculty-of-technology
4.	/index.php/electrical-and-computer-engineering
5.	/index.php/academics/colleges/college-of-social-sciene
6.	/index.php/contacts
7.	/index.php/admissions
8.	/index.php/aau-bpr-all
9.	/index.php/administration
10.	/index.php/library-and-museum
11.	/index.php/siter-map
12.	/index.php/aboutaau
13.	/index.php/administration/human-resource
14.	/index.php/campus-life
15.	/index.php/library-and-museum/library
16.	/index.php/registrar
17.	/index.php/academics/faculties/faculty-of-medicine
18.	/ies/
19.	/index.php/civil-engineering
20.	/index.php/academics/63-faca-institute-college
21.	/index.php/view-blog
22.	/index.php/admissions/graduate
23.	/index.php/admissions/continuing-and-distance-education-office
24.	/index.php/campus-life/housing-a-dining
25.	/index.php/admissions/apply-online
26.	/index.php/aboutaau/campus-guide
27.	/index.php/opportunity-for-research-fellowship
28.	/index.php/academics/faculties/faculty-of-business-and-economics
29.	/index.php/admissions/undergraduate
30.	/index.php/aboutaau/graduate-studies-and-research-office
31.	/march/
32.	/index.php/component/search/
33.	/index.php/academics
34.	/index.php/academics/faculties/faculty-of-law
35.	/index.php/academics/faculties/faculty-of-science
36.	/index.php/library-and-museum/46-library-
37.	/index.php/admissions/41-graduate
38.	/index.php/evaluation-feedback
39.	/vacancy-announcement
40.	/web-mail
41.	/ies/downloads

Appendix D: Sample log records from the web server of AAU web site

(a, b, c, d, e)

a. Access Log file

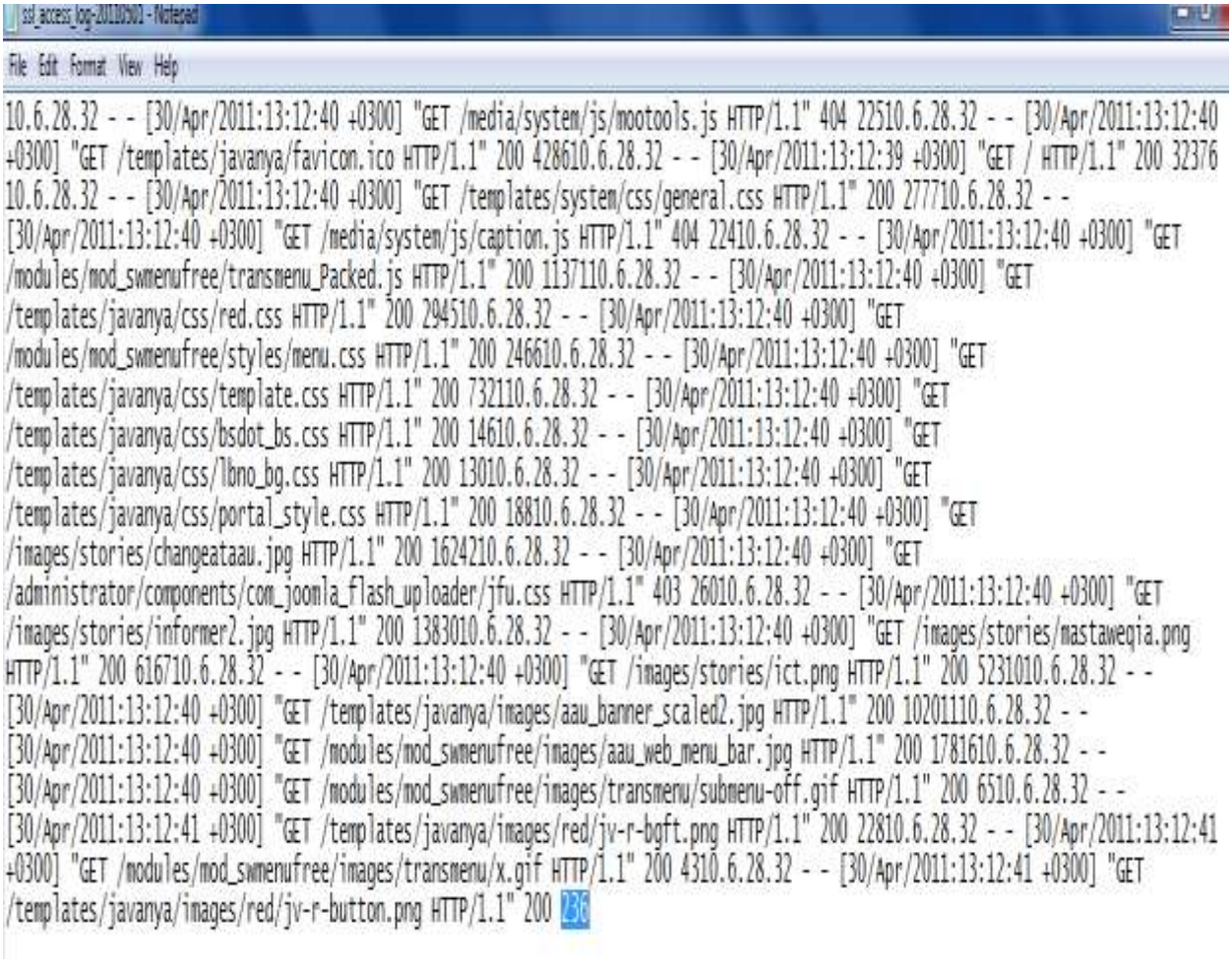


b. Error log file

File Edit Format View Help

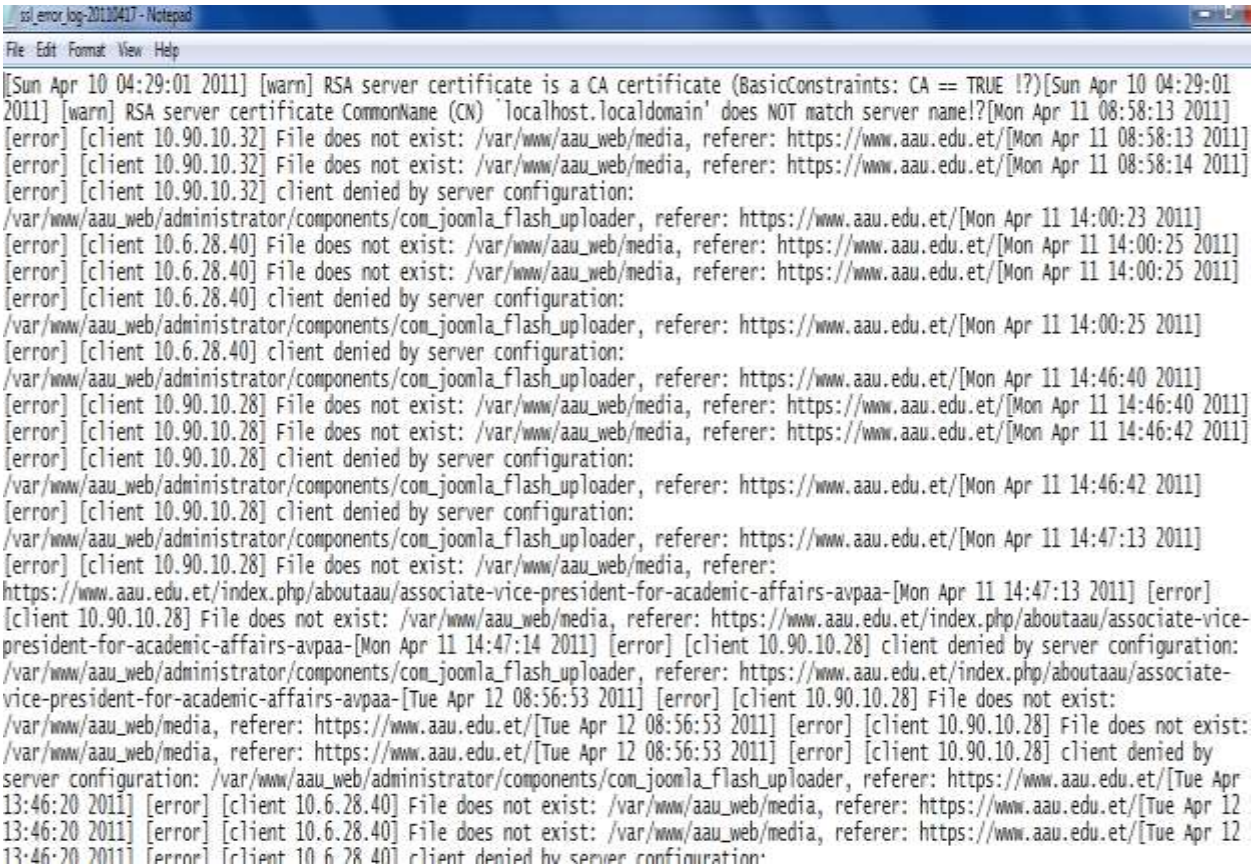
[Sun Dec 19 04:26:48 2010] [notice] Digest: generating secret for digest authentication ...[Sun Dec 19 04:26:48 2010] [notice] Digest: done[Sun Dec 19 04:26:48 2010] [notice] mod_python: Creating 4 session mutexes based on 256 max processes and 0 max threads.[Sun Dec 19 04:26:48 2010] [notice] mod_python: using mutex_directory /tmp [Sun Dec 19 04:26:48 2010] [notice] Apache/2.2.6 (Unix) DAV/2 PHP/5.2.4 mod_python/3.3.1 Python/2.5.1 mod_ssl/2.2.6 OpenSSL/0.9.8b mod_perl/2.0.3 Perl/v5.8.8 configured -- resuming normal operations[Sun Dec 19 04:27:09 2010] [error] [client 207.46.195.240] File does not exist: /var/www/aau_web/faculties[Sun Dec 19 04:27:14 2010] [error] [client 178.154.160.30] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 62[Sun Dec 19 04:27:14 2010] [error] [client 178.154.160.30] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 69[Sun Dec 19 04:27:14 2010] [error] [client 178.154.160.30] PHP Warning: array_merge() [function.array-merge]: Argument #1 is not an array in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 666[Sun Dec 19 04:27:14 2010] [error] [client 178.154.160.30] PHP Notice: Use of undefined constant sitename - assumed 'sitename' in /var/www/aau_web/templates/javanya/index.php on line 55[Sun Dec 19 04:27:35 2010] [error] [client 66.249.68.187] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 62[Sun Dec 19 04:27:35 2010] [error] [client 66.249.68.187] PHP Warning: array_merge() [function.array-merge]: Argument #1 is not an array in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 775[Sun Dec 19 04:27:35 2010] [error] [client 66.249.68.187] PHP Notice: Use of undefined constant sitename - assumed 'sitename' in /var/www/aau_web/templates/javanya/index.php on line 55[Sun Dec 19 04:27:38 2010] [error] [client 66.249.67.167] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 62[Sun Dec 19 04:27:38 2010] [error] [client 66.249.67.167] PHP Notice: Use of undefined constant sitename - assumed 'sitename' in /var/www/aau_web/templates/javanya/index.php on line 55[Sun Dec 19 04:28:31 2010] [error] [client 190.201.195.209] File does not exist: /var/www/aau_web/media, referer: http://www.aau.edu.et/[Sun Dec 19 04:28:31 2010] [error] [client 190.201.195.209] File does not exist: /var/www/aau_web/media, referer: http://www.aau.edu.et/[Sun Dec 19 04:28:35 2010] [error] [client 190.201.195.209] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: http://www.aau.edu.et/[Sun Dec 19 04:28:41 2010] [error] [client 190.201.195.209] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: http://www.aau.edu.et/[Sun Dec 19 04:28:46 2010] [error] [client 178.154.160.30] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 62[Sun Dec 19 04:28:46 2010] [error] [client 178.154.160.30] PHP Notice: Trying to get property of non-object in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 69[Sun Dec 19 04:28:46 2010] [error] [client 178.154.160.30] PHP Warning: array_merge() [function.array-merge]: Argument #1 is not an array in /var/www/aau_web/components/com_events/libraries/datamodel.php on line 666[Sun Dec 19 04:28:46 2010] [error] [client 178.154.160.30] PHP Notice: Use of undefined constant sitename - assumed 'sitename' in /var/www/aau_web/templates/javanya/index.php on line 55[Sun Dec 19 04:28:47 2010] [error] [client 67.195.114.30] PHP Notice: Use of undefined constant sitename - assumed 'sitename' in /var/www/aau_web/templates/javanya/index.php on line 55[Sun Dec 19 04:29:36 2010] [error] [client 66.249.68.187] PHP Notice: Trying to get property of non-object in

c. SSL Access Log



d. **SSL Error**

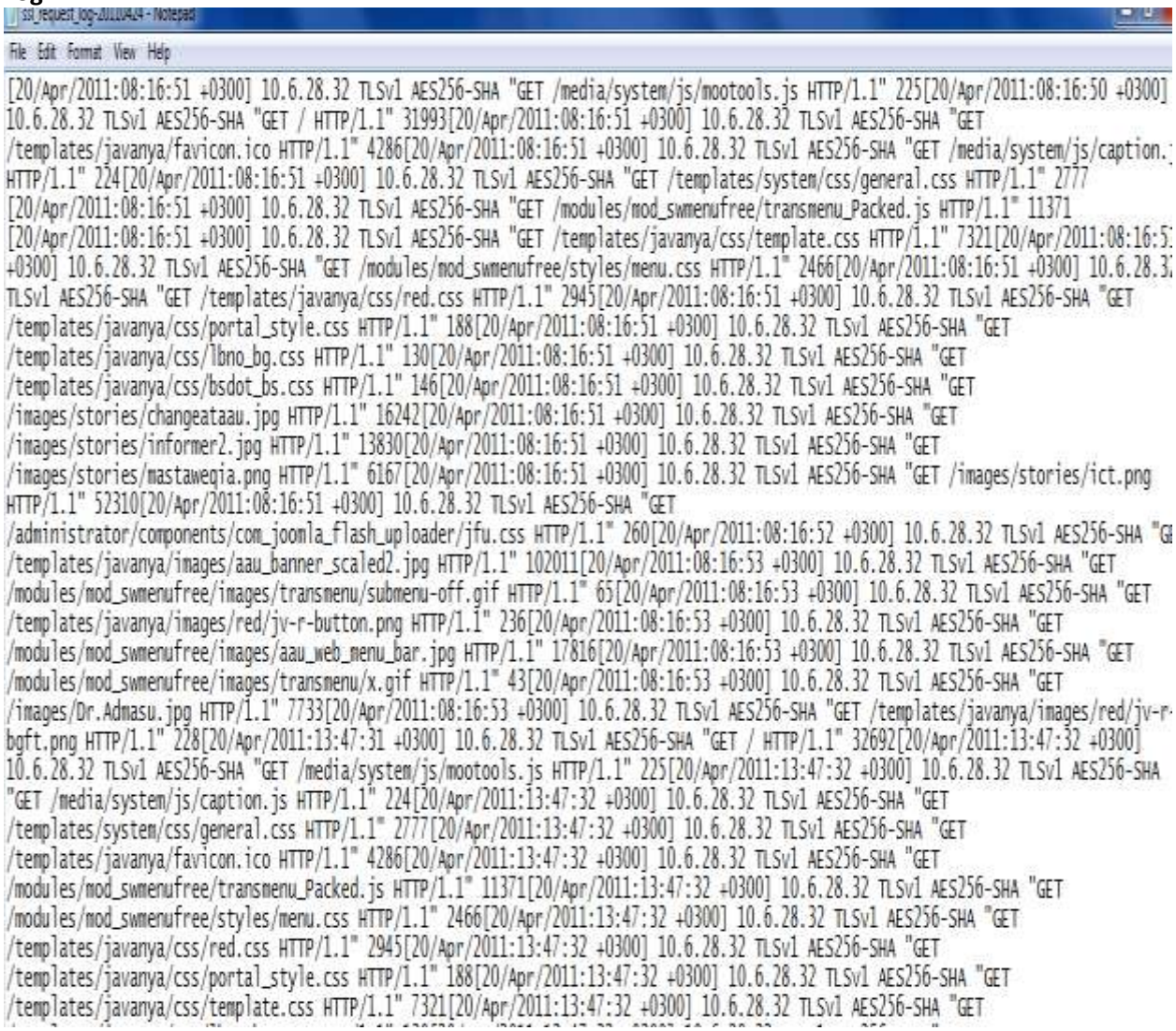
Log


ssl_error_log-20110417 - Notepad

File Edit Format View Help

[Sun Apr 10 04:29:01 2011] [warn] RSA server certificate is a CA certificate (BasicConstraints: CA == TRUE !?)[Sun Apr 10 04:29:01 2011] [warn] RSA server certificate CommonName (CN) `localhost.localdomain' does NOT match server name!?[Mon Apr 11 08:58:13 2011] [error] [client 10.90.10.32] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Mon Apr 11 08:58:13 2011] [error] [client 10.90.10.32] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Mon Apr 11 08:58:14 2011] [error] [client 10.90.10.32] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/[Mon Apr 11 14:00:23 2011] [error] [client 10.6.28.40] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Mon Apr 11 14:00:25 2011] [error] [client 10.6.28.40] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Mon Apr 11 14:00:25 2011] [error] [client 10.6.28.40] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/[Mon Apr 11 14:00:25 2011] [error] [client 10.6.28.40] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/[Mon Apr 11 14:46:40 2011] [error] [client 10.90.10.28] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Mon Apr 11 14:46:40 2011] [error] [client 10.90.10.28] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Mon Apr 11 14:46:42 2011] [error] [client 10.90.10.28] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/[Mon Apr 11 14:46:42 2011] [error] [client 10.90.10.28] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/[Mon Apr 11 14:47:13 2011] [error] [client 10.90.10.28] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/index.php/aboutaau/associate-vice-president-for-academic-affairs-avpaa-[Mon Apr 11 14:47:13 2011] [error] [client 10.90.10.28] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/index.php/aboutaau/associate-vice-president-for-academic-affairs-avpaa-[Mon Apr 11 14:47:14 2011] [error] [client 10.90.10.28] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/index.php/aboutaau/associate-vice-president-for-academic-affairs-avpaa-[Tue Apr 12 08:56:53 2011] [error] [client 10.90.10.28] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Tue Apr 12 08:56:53 2011] [error] [client 10.90.10.28] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Tue Apr 12 08:56:53 2011] [error] [client 10.90.10.28] client denied by server configuration: /var/www/aau_web/administrator/components/com_joomla_flash_uploader, referer: https://www.aau.edu.et/[Tue Apr 12 13:46:20 2011] [error] [client 10.6.28.40] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Tue Apr 12 13:46:20 2011] [error] [client 10.6.28.40] File does not exist: /var/www/aau_web/media, referer: https://www.aau.edu.et/[Tue Apr 12 13:46:20 2011] [error] [client 10.6.28.40] client denied by server configuration:

e. **SSL Request Log**



Appendix E: Run outputs of Apriori algorithm on the AAUCleanedWebLog dataset

Run 1

=== Run information (Edited and Formatted) ===

Scheme: Weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.17 -S -1.0 -c -1

Relation: -AAUCleanedWebLog

Instances: 95905

Attributes: 41

Apriori

=======

Minimum support: 0.17 (15900 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 20

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 1

Best rules found:

1. URL12=1 URL30=1 17346 ==> URL7=1 16652 conf :(0.96)
2. URL14=1 URL24=1 19165 ==>URL9=1 18015 conf :(0.94)
3. URL33=1 URL7=1 25,123 ==>URL25=1 23866 conf :(0.95)
4. URL12=1 URL26=1 17100 ==>URL39=1 15903 conf :(0.93)
5. URL14=38967 ==>URL10=1 35070 conf :(0.9)
6. URL10=1 39152 ==>URL38=1 36019 conf :(0.92)
7. URL7=1 42111 ==>URL38=1 37899 conf :(0.9)
8. URL33=1 50123 ==> URL6=44609 conf :(0.89)

Run 2

=== Run information (Edited and Formatted) ===

Scheme: Weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.03 -S -1.0 -c -1

Relation: -AAUCleanedWebLog

Instances: 95905

Attributes: 41

Apriori

=======

Minimum support: 0.03 (2881 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 45

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 1

Best rules found:

1. URL12=1 URL8=1 2932 ==> URL9=1 2919 conf :(0.99)
2. URL12=1 URL14=1 2951 ==> URL31=1 2900 conf :(0.98)
3. URL33=1 URL7=1 3048==> URL41=1 2998 conf :(0.98)
4. URL9=1 URL33=1 2995==> URL16=1 2883 conf :(0.96)
5. URL12=1 URL2=1 3400 ==> URL9=1 3001 conf :(0.88)
6. URL11=1 3358 ==> URL33=1 2890 conf :(0.86)

Appendix F: Running outputs of common sequence discovery from the sequence transaction file

Run 1

=== Run information (Edited and Formatted) ===

Scheme: Weka.associations.PredictiveApriori -N 25 -A -c -1

Relation: SeqMining

Instances: 95,905

Attributes: 2

Req2

Req3

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

PredictiveApriori

==========

Best rules found:

1. Req2=URL12 35105==> Req3=URL21 34801 acc:(0.97)
2. Req2=URL14 34054==>Req3=URL9 33897 acc :(0.96)
3. Req2=URL10 25678 ==> Req3=URL15 23854 acc :(0.94).

Run 2:

=== Run information (Edited and Formatted) ===

Scheme: Weka.associations.PredictiveApriori -N 25 -A -c -1

Relation: SeqMining

Instances: 95,905

Attributes: 2

Req3

Req4

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

PredictiveApriori

==========

Best rules found:

1. Req3=URL7 41067==> Req4=URL25 39989 acc:(0.998)
2. Req3=URL33 27965==> Req4=URL7 27129 acc:(0.946)
3. Req3=URL14 26876 ==> Req4=URL7 25886 acc:(0.94)
4. Req3=URL21 27534==> Req4=URL9 26543 acc:(0.94)

Run 3

=== Run information (Edited and Formatted) ===

Scheme: Weka.associations.PredictiveApriori -N 25 -A -c -1

Relation: SeqMining

Instances: 95,905

Attributes: 2

Req4

Req5

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

PredictiveApriori

==========

Best rules found:

1. Req4=URL7 21896==> Req5=URL25 21112 acc:(0.976)
2. Req4=URL9 19098==> Req5=URL33 18764 acc:(0.969

Run 4

=== Run information (Edited and Formatted) ===

Scheme: Weka.associations.PredictiveApriori -N 25 -A -c -1

Relation: SeqMining

Instances: 95,905

Attributes: 2

Req5

Req6

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

PredictiveApriori

==========

Best rules found:

1. Req5=URL33 12986==> Req6=URL16 10763 acc:(0.906)

Combination of the above rules reveals that the following are the most common sequences in the users' visit of the web site.

- URL33→URL7→URL25
- URL12→URL21→URL9
- URL14→URL9
- URL10→URL15

Appendix G: List of discussants

No.	Name	Position
1.	Dr. Rahel Bekele	ICT Coordinator
2.	Mr. Abiy Zemede	Member of Networking Team
3.	Mr. Asegid Asfaw	Member of Networking Team
4.	Mr. Bekuma Midekssa	Member of Networking Team
5.	Mr. Dagne Minda	Member of Web Site Design team
6.	Mr. Samson Aklilu	Member of Web Site Design Team
7.	Mr. Zekarias Mintesnot	Member of Networking Team

Declaration

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

_____

Date

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

_____

Advisor