ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS

WEB USAGE PATTERN DISCOVERY USING DATA MINING AND STATISTICAL ANALYSIS: The Case of AAU Official Web Site

By:

Mekonnen Tsegaye Belihu

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.

January 2009

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS

WEB USAGE PATTERN DISCOVERY USING DATA MINING AND STATISTICAL ANALYSIS: The Case of AAU Official Web Site

By:

Mekonnen Tsegaye Belihu

Name and Signature of the Examining Board

Mr. Ermias Abebe, Chairman	Signature	Date
Dr. Manoj V.N.V, Advisor	Signature	Date
Dr. Dejene Ejigu, External Examiner	Signature	Date

Acknowledgement

First and for most I would like to express my heartfelt gratitude to my advisor Dr. Manoj V.N.V for his constructive comments and encouragement. I would like my gratitude to go to my advisor not only for his guidance throughout the research but also for his trust in me and willingness to accept me as an advisee when I was looking for an advisor.

I would also like to thank Ato Lemma Lessa, head of the Information Science Department, AAU, for his unreserved effort in search of advisors and his hospitality whenever I visited his office.

My thanks also goes to Ato Dagne Minda and Hiwot Mustefa, staff of ICT development office of AAU, and their office-mates, for providing me with the required information and their kind hospitality whenever I visited their offices.

Last, but not least, I would like to extend my thanks to my mother Abebech Sisay and my Siblings Hanna, Teddy, Azeb with her family, Seble, Fiker, Million, and Yisem, for their material and moral support.

Mekonnen Tsegaye

Table of Contents

List of Tables

List of Figures

List of Equations

List of Appendices

Abbreviations and Acronyms

3W: (see WWW)

AAU: Addis Ababa University

ARFF: Attribute Relation File Format

ASCII: American Standard Code for Information Exchange

FP: Frequent Pattern

GIF: Graphics Interchange Format

HTML: Hypertext Markup language

HTTP: Hypertext Transfer Protocol

ICT: Information and Communication Technology

IP: Internet Protocol

ISP: Internet Service Provider

JPEG: Joint Pictures Expert Group

KDD: Knowledge Discovery from Database

OLAP: On-Line Analytical Processing

OLTP: On-Line Transaction Processing

RDBMS: Relational Database Management System

URL: Uniform Resource Locator

W3C: World Wide Web Consortium

WEKA: Waikato Environment for Knowledge Analysis

WUMPrep4WEKA: Web Usage Mining Preprocessor for WEKA

WWW (or 3W): World Wide Web

XML: eXtensible Mark-up Language

Abstract

The focus of this study is to analyze a Web usage pattern of AAU official Web site based on both statistical and data mining approaches thereby to see the possibility of this approach for Web usage pattern discovery. The statistical analysis is applied to generate basic statistical reports whereas the data mining technique is applied primarily to get the association rules that can reveal the Web pages/ URLs that are most frequently accessed together by the users.

It is believed that the statistical analysis is deemed insufficient to give a full picture of how a given Web site is being used. In fact, the figures like which page is most accessed; when is month of a year, day of a month, and time of a day on which most users accessed the given page; etc. can be useful for decision making for instance to schedule the Web maintenance time. Data mining, on the other hand, can be used for getting some hidden patterns such as the pages accessed together by many users, etc. thereby to be used for improving both the Web design and services.

For the research, Web access log data of three months of the year 2007 have been collected as sample of the one year log data. Then, preprocessing tasks are carried out for making the log file ready for data mining. The preprocessing task has been accomplished using WEKA Plug-in tool and some codes that have been written with a Python programming language.

Then, *Mach5 Analyzer* and *WEKA* have been used for statistical analysis and association rule generation, respectively. Data mining experiment has been conducted following the statistical analysis. From the week days and weekends, minimum number of users has been recorded on Monday and Saturday, respectively, the Web site has maximum users around 3:00PM everyday; the */index.php* and the */webmail.php* pages have been found the top-two most

requested pages; and more than half of the users have entered into the Web site directly through the home page and 67% of them have left the site without visiting the other pages; and */search/* is the top directory that encounters error. From the data mining output, the */academics/index.php* page has been found as center of interest to which many users tend to forward requests from other pages. Interesting association rules among other pages have been also generated.

From the research, it has been found that statistical analysis alone can not be good enough to get an insight about a Web site's usage pattern. But, using both statistical and data mining approaches can help to get a better understanding how a Web site is being used. Conclusions have been drawn in order to give some recommendations, accordingly. Finally, a model for Web usage analysis has been proposed for those who wish to do further research in the area.

Chapter One

Introduction

1.1. Introductory Note

In the human history, a number of inventions had been realized and changed the way human being lives. Scientist, in different times, made several inventions that would not be imaginable and seem impossible for the ordinary people. Among other things, the Internet is one of the remarkable achievements in the communication realm. At the birth of the Internet, in 1969 [13], the intention was to use it for simple mail exchange. However, now the Internet provides much more services which are far beyond what was interned at its birth. World Wide Web service, WWW or 3W, is now a popular service among almost all Internet users. Millions of users all over the world are now using the Internet and thousands of Internet Service Providers (ISPs) are providing access to the Internet. There are more than 1,407,724,920 Internet users in the world and, out of these; some 51 million users are in Africa until 31st of March, 2008. Despite the fact that Africa contributes 3.6% of the world Internet users, it shows 1,030.20 % growth from the year 2000 - 2008. Similarly, Ethiopia has 0.6% share among African Internet users. Even if this seems insignificant compared to the rest of the countries, the growth of Internet users is encouraging. The number of Internet users that was 10,000 in the year 2000 has increased to 291,000 in the year 2008 latest data, which shows a dramatic growth i.e. 2,910.00 % [14] [1].

[1] *The www.internetworldstats.com website presents frequent updates of the statistics*

The WWW services is an organized collection of Web sites by which information can be presented to users in various formats, ranging from a simple plain text to audio/video- that can convey information on different issues, which in turn ranges from a daily cooking guide to genetic engineering; from 'how to built a tent' to space science; etc. Thus, these days, it is becoming a matter of a mouse click to get any information on any topic provided that there is Internet connectivity. Web pages are, therefore, serving as a bridge between the information providers and the information seekers.

1.2. Statement of the Problem

Being the means of information presentation, Web pages are becoming popular by the user community. This contributes for the accelerated growth of Web sites world wide. As statistics shows, the number of Web sites published every day is increasing quickly and until 15th of June 2008 about *103,949,056* active domains are registered globally [12]. In addition to the traditional one, now a new dimension is created for business firms and other organizations to gain competitive advantage by providing their service online via Web sites.

Information available through Web pages is different from the information which is available on other media. This is because, in the former case, users are able to satisfy their need for information by interactively directing themselves to the information source they like to have [Ratheke & Schreiweis, 2003]. In addition to this, unlike the traditional media, for example, printed books, a Web user can skip to another source in few mouse clicks following the link structure. Users interact with a Web browser when they follow a link or specify search options [ibid]. The user can also view multiple sources in almost the same time i.e. simultaneously. In

addition, most Web sites are dynamic and therefore their content and structure vary from time to time.

In contrast with their dramatic growth, most Web sites are not useful for the most of the users. As a study shows, 99% of the Web sites are not useful for 99% of the users [Han & Kamber, 2006]. There are several factors contribute for such situation. Lydon and Fennell [2003] argued that "a large number of websites are poorly designed, because user requirements are often not incorporated into the Web design process ". They focus more on the quantity rather than the content quality. But, the fact is that "...if a consumer encounters a positive experience on a Web site, it is likely that it will increase their time spent at the site" [ibid].

1.3. Scope and Limitation of the Research

Web mining has three different branches: Web Content mining, Web Structure mining, and Web usage mining. The focus of this research is on mining usage patterns of the Addis Ababa University (AAU) official Web site. Usually, there are three types of Web-related log files, namely Web access log, error log, and proxy log files. However, to accomplish this research work, the Web access log records is used as dataset because many literatures and previous researches justify that Web access log file is the typical source of information for discovering Web usage patters. The mining task, in fact, is also accompanied by statistical analysis to get more insight about the Web site's usage pattern [Cooley et al., 1999].

The limitation in this research is lack of latest data hence the year 2007 data has been used instead.

1.4. Justification of the Research

Whether one likes or not, the effect of globalization touches every nation. In this matter, Internet plays an imperative role by bringing the people of the world into a common virtual space. As it is stated earlier, most Web sites are less useful for the majority of the user community. Therefore, continues researches in this area play an important role to improve user satisfaction and Web usability as well.

A user stays for a longer time on a Web site if only he/she satisfies with the Web site content and structure. In this regard, the degree of Web usability is considered as a key factor likely to affect users' interest to use the Web frequently for prolonged time [Lydon & Fennell, 2003]. The need for Web usability is stated as follows:

> *"If website is not useful to users, it will never be used. In order for a website to be successful, users must visit the site to find information or accomplish tasks. No matter what objectives have been set for the website, it must carefully balance the needs of users and the needs of the organization. If users don't find the website helpful, they will not use it, which will, in turn, prevent the owner from meeting the organization's objectives"*[15].

According to Koutri & Daskalaki [2003], user-centered design effort for Web sites always leads to the construction of usable sites. Giving a due consideration for Web usability is one of the most important steps for effectiveness of the Web services, efficiency of delivering the information, and maximizing user satisfaction in any given context of use and task.

To improve the quality of a Web site, it is indispensable to get some sort of understanding about the usage patterns of the Web site. Web mining is one of the emerging techniques to get some knowledge about the status of the Web site. Web mining is mainly built on usage mining [Aschenbrenner & Rauber, 2006]. Data mining can be performed on Web log records to find association patterns, sequential patters, and trends of web accessing, and so on [Han & Kamber, 2006]. More specifically, usage mining is one of the approaches to address Web usability issues as it can help to discover the users' navigational behavior and other patterns. On the basic level it answers the queries like what parts of a Web site a user visited in a single session; where she/he spent most time; etc. Starting from such basic information, a number of valuable analyses can be conducted for organizational Web sites. Web log analysis is used for characterizing the designated community of an online service [Aschenbrenner & Rauber, 2006]. Based on usage patterns, the structure of the site map can also be improved.

In general, the goal of Web usage mining is to capture and model Web users' behavioral pattern [Dai & Mobasher, 2005]. Such patterns are hidden in vast Web log files [Chen et al., 2005]. Many researchers and literatures show that Web usage mining has a number of benefits. Some of the benefits are: to evaluate the effectiveness of a site in meeting users' expectations; for load balancing and optimization of Web server for better and more efficient user access; for restructuring or customizing a site based on users' predicted needs and interests [Dai & Mobasher, 2005]; to reduce, through pre-fetching and caching, Web latencies that have been perceived by web users year after year; and for administrative personnel, to predict trends of users' need so that they can adjust their product/ service to attract more users [Chen et al., 2005]. In addition, usage mining can also be employed for the implementation of recommender systems [Aschenbrenner & Rauber, 2006].

Therefore, researches on such issues play a vital role for the design and development of more suitable Web sites for users. Moreover, this research is believed to have a significant contribution for the design of Web pages that can address the AAU official Web site users' interest. Beyond this, the research output also indicates possible directions for further related research in this domain because, as far as the writer's knowledge concerned, Web related research activity is unexploited yet in Ethiopia. Having done the experiment, a model for Web usage pattern analysis is also recommended.

In general, because of the growing number of both Web sites and users, such kind of researches have most significant effect for Web design endeavor by considering users' interests [Zilse & Moraes, 2003].

1.5. Objectives

1.5.1. General Objective

The general objective of the research is to apply data mining techniques and statistical analysis for discovering AAU Official Web site usage pattern to reveal previously unknown interesting, noble, and actionable patterns based on the Web access log file in order to recommend possible measures for further improvement of the official Web site of AAU and besides to propose a general model for Web usage pattern discovery by examining the possibility of applying these two techniques for usage pattern analysis.

1.5.2. Specific Objectives

In order to achieve the general objective of the research, there are specific objectives which need to be addressed in the research. Therefore, the specific objectives of the research are:

- To review literatures in the area in order to put concrete background and justification for the research;
- To identify the dataset so as to use them for the knowledge discovery process;
- To prepare the dataset using different preprocessing techniques;
- To analyze the dataset statistically;
- To analyze the dataset using a Web mining software/ tool in order to discover previously unknown association patterns among requested URLs;
- To identify interesting associations from the available patterns;
- To interpret the interesting patterns to discover new knowledge i.e. finding of the research;
- To draw conclusions based on the findings and possible application of both techniques for Web usage pattern analysis.
- To make and forward some appropriate recommendations based on the conclusions; and
- To proposed a model for Web usage pattern discovery.

1.6. Research Methods

Since the goal of Web usage mining is to uncover some hidden but interesting patterns from Web log dataset, it shares the basic procedures with the conventional data mining task. Therefore, the research follows some methodology to come up with some result:

1.6.1. The Research Flowchart

The following flowchart depicts the general research flow:

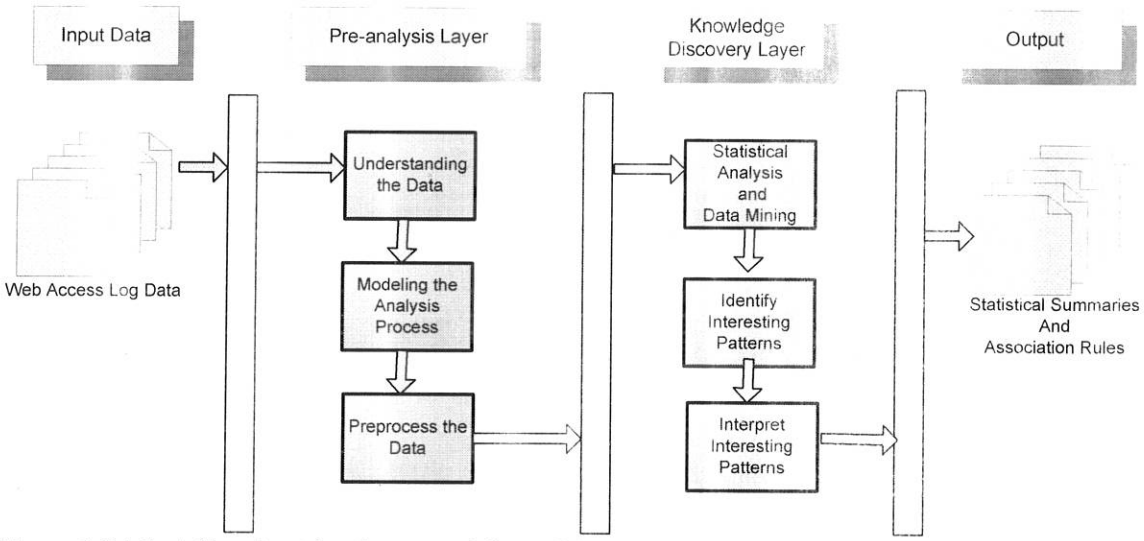


Figure 1.6.1-1: A flowchart for the general flow of the research

1.6.2. Data Collection for the Study

In a conventional data mining activity, the data for the study is derived from a data warehouse or a data mart. In this study the data has been taken from a Web server that hosts the Web site under study. So, Web data of the AAU official Web site has been considered for the study.

1.6.3. Data Selection

There are three types of Web site related data, i.e. data generated due to an interaction between a client computer and a Web server: Web access log, Error log, and Proxy log data. From these three log files, the Web access log file has been selected for the study.

1.6.4. Data Preparation

Data preparation, alternatively called data preprocessing, is used for making the research data ready for analysis because real world data tend to be incomplete, noisy, and inconsistent [Han & Kamber, 2006]. Hence, some preprocessing tasks will be carried out in this step in order to retrieve and analyze significant and useful information. Preprocessing includes data cleaning,

data integration, data transformation, and data reduction. Thus, the following data preparations tasks have been addressed in this study.

- **Data Cleaning**: it is a task that attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. For example requests made from the local machine itself should be excluded from the dataset as it may not be useful to get the (outsider) users Web usage pattern.

  In addition, some other processes need to be made to improve the accuracy of the output and to facilitate the mining process. For example, additional tasks, such as removing certain least useful attributes from the dataset and modifying the column names, in order to make them easily remembered, etc need to be performed.

- **Data Transformation**: in this process the data is transformed to the form appropriate for mining. A Web log file, by its nature, records every request of a give user (i.e. client's IP address) between each fraction of minutes i.e. a given user may have multiple consecutive records for each requests until he/she leaves the Web site. So, these requests need to be transformed to sessions for the respective users. Then, the sessions also should be transformed to a transaction table i.e. putting the pages accessed in the same user session together. Finally, the transaction should be transformed to the format that is suitable for the experiment tool(s).

1.6.5. Data Analysis

To address the objective of this research, different data mining approaches and statistical analysis have been performed on the dataset to get an insight about the Web usage trend and reveal interesting patterns from the Web access log records.

1.6.5.1. Data Analysis and Experiment Tools

There are some freeware and commercial tools on the Internet for Web mining and Web usage statistical analysis purpose. *Mach5 Analyzer* and *WEKA* have been selected for statistical analysis and data mining, respectively. The justification for why these tools are selected is given in the later chapters.

1.6.6. Interpret and Report the Result

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

1.7. Application of Results

Due to the very nature of data mining, one can not guess the kind of knowledge that would be discovered at the end of any knowledge discovery process. Therefore, it is somewhat difficult to explicitly state what can be attained from the Knowledge Discovery process. But, it is believed that some interesting patterns that can reveal useful, actionable, and noble features from the data, as result of the mining process, can be used for improving the Web site under study and the overall research could pave the way for further studies.

1.8. Organization of the Thesis

The first chapter of the thesis covers a general introduction to the research. It specifies the problem statement and its justification; scope and limitation of the study; objectives of the study; etc.

The second chapter discuses the background of the organization of which the Web mining study is conducted and provides a general overview of data mining technology while the third chapter deals with theoretical background of Web mining in general, Web usage mining in particular.

Since the research area is young and no prior works in local context by local scholars has been done, the writer believes that there is a need to model the research process. The fourth chapter therefore presents the model of the approach used in the research. Then, in the fifth chapter, the experiment and findings of the research are presented and the main activities of the research such as the preprocessing, the experiment/ analysis and the interpretation of the results are covered.

Chapter six, which is the last chapter, presents the conclusions based on the experiment result and appropriate recommendations based up on the conclusions.

Finally, in order to make the main body of the thesis easy on the eye, some figures and other supplementary materials are presented in the appendix section.

Chapter Two

Background and Literature Review

2.1. ICT Development in AAU[2]

Addis Ababa University (AAU) is the oldest higher educational institution in Ethiopia. AAU started its operation in 1950 under the name University College of Addis Ababa. It was renamed Haile Selassie I University in 1962 and then Addis Ababa University in 1975.

AAU runs Diploma, Bachelors, Medical Doctors, Doctor of Veterinary Medicine, Masters, Specialty Certificate and PhD degree programs in various fields of study. It launched its first Master of Science programs in 1979 and its first PhD programs in 1987.

The ICT Development Office at AAU established in January 2003 is a new system that liaisons between the AAU on the one hand and stakeholder such as international donor agencies, collaborating overseas universities and relevant local institutions on the other hand in initiating and implementing ICT-related projects and activities.

The broad duties and responsibilities of the office are developing ICT strategic plan and overseeing its implementation. In carrying out these broad duties and responsibilities, the office would engage in the following major activities:

- Developing short- and long-term ICT strategic plans to introduce and expand application of ICT at AAU;

[2] http://www.aau.edu.et

- Coordinating and overseeing the implementation of existing ICT projects at AAU;
- Developing new ICT project in accordance with the ICT strategic plan and strategies of AAU;
- Engaging and coordinating the fund raising activities to implement new and ongoing ICT projects at AAU;
- Establishing linkages with potential donors, collaborating institution and individuals, that could contribute positively to the success of ICT projects at AAU
- Serving as an information center on ICT-related visions, project and activities of AAU;
- Identifying resources required to implement the projects and facilitating the acquisition of that (human resource, physical resource and infrastructure, and financial resource); and
- Advising the university management on matters related to ICT.

The AAU ICT office has a vision of becoming a center of excellence in utilizing the potential of ICT in learning, research, innovation, and educational environment for national development. It carries a mission to make ICT a critical enabler for AAU in teaching & learning environment, providing ICT community services, and bridging the gap between industry and academia skills through training, research, and consultancy.

The ICT development office works to fulfill the following objectives:

- To provide effective and efficient ICT services to the AAU community;

- To promote e-learning initiative to the AAU and the nation at large;
- To bridge ICT skill gap between academia and industry;
- To promote research on ICT sector application and base line studies;
- To provide ICT community services of development impact;
- To establish collaboration and partnership with public & private higher learning institutions.

2.2. *The AAU Official Web Site*

The Addis Ababa University Official Web site was published around some five years ago. As the ICT development office of AAU is engaging in ICT related activities, the Web site has been developed and being maintained by this office. The Web site is hosted on the AAU's own server, which is located in the main (Sidest Kilo) campus of the university.

2.2.1. Purpose and User Community

The university's Web site is established with the objective of delivering information about the university's activities, in general, and about academic and administrative units, in particular. The web site also delivers news items, its own advertisements for both vacancies and student admission, and many more. It has also external links to other Web site, such as collaborative organizations in research and other activities, donor agencies, etc.

This, the AAU Web site is meant to serve both the university's community and outsiders who are in need of information about and related to the university.

2.2.2. Nature and Content

For presenting information on several topics and issues, the AAU Web site has both dynamic and static nature. As to the content of AAU Web site, so far, the trend has been that individual academic and administrative units have been supposed to design and summit their respective pages (HTML pages). Then the ICT office uploads and links those pages with the home page of the Web site. As it has been told, such trend has been negatively contributing for page layout inconsistencies among the pages in the Web site. This even sometimes has resulted in incompleteness of the content of Web pages and high probability of broken links.

2.2.3. AAU Web Site Structure

The following figure illustrates the structure of the AAU official Web site.

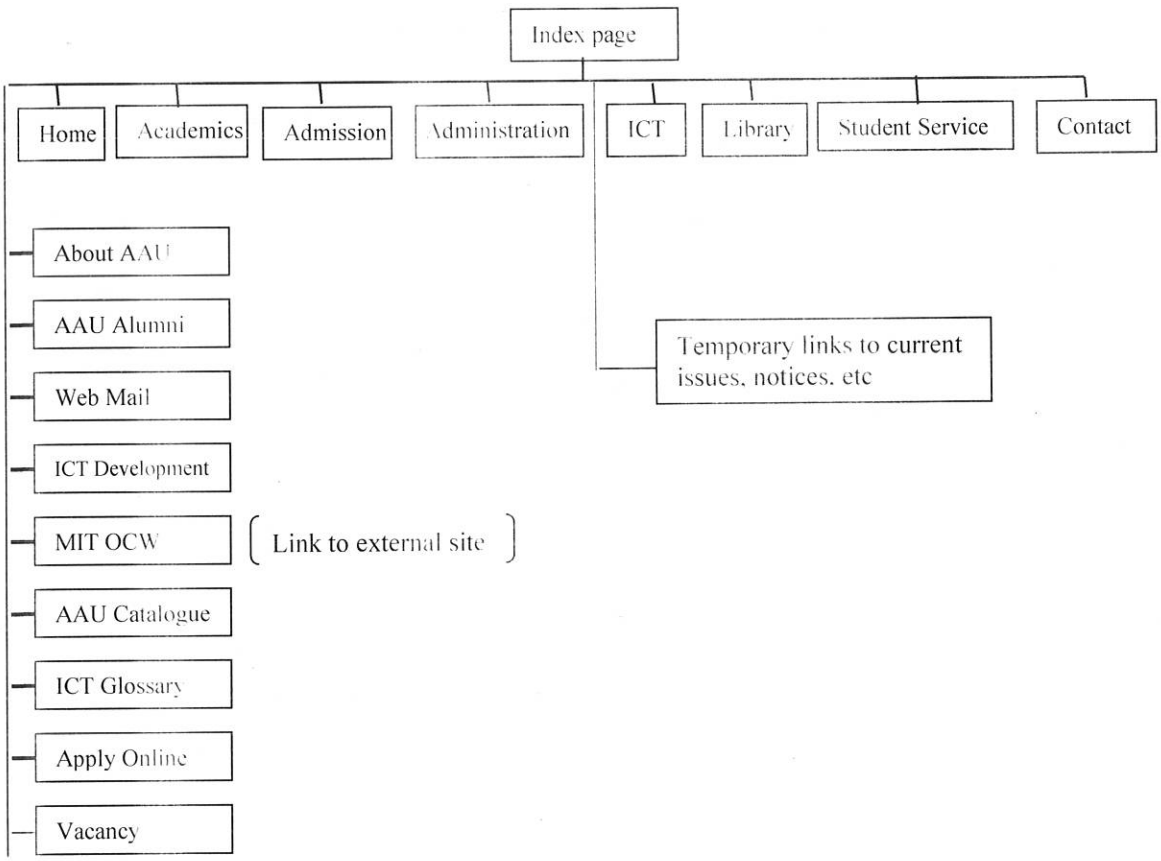


Figure 2.2-1: AAU official Web site map (level one view)

2.3. Data Mining Overview

Data mining, according to Han and Kamber [2006], refers to as the extraction and mining of knowledge from a huge amount of data. They used the term as analogous to the gold mining from rock or sand. However, several alternative terminologies have been given for the data mining, such as knowledge mining in database, data/pattern analysis, data archeology, data dredging, and knowledge extraction. Hand et al. [2001] tried to give a comprehensive definition for the data mining as the analysis of (often large) observational datasets to find unsuspected relationship and to summarize the data in novel ways that both understandable and useful to the data owner.

Knowledge discovery in database, or KDD, is assumed by many people as the same as data mining [Hand & Kamber, 2006]. But, some other scholars argue that data mining is an essential step to knowledge discovery in database, meaning that KDD is at the higher level and the final goal of data mining. The knowledge is discovered based the resulting patterns after data mining is performed on given dataset [ibid]. Data mining is often set in the broader context of KDD [Hand et al., 2001]. Most scholars argue that data mining is the core of KDD because one can not think of KDD without data mining.

2.3.1. Motivation for Data Mining

Sine the 1960s, database and information technology have been evolving systematically from the primitive file processing system to sophisticated and powerful database systems [Hand & Kamber, 2006]. The advancement of digital data acquisition and storage technology has resulted in the collection of huge databases [Hand et al., 2001]. Among other things, the dramatic and accelerated growth of computer hardware technology has led to a large number

of manufacturers and suppliers of higher performance and affordable computing machines, data collection equipments, and data storage media. This trend, in turn, facilitates the data collection task and enables storage of huge databases [Hand & Kamber, 2006]. This has happened in the human activities from daily business transaction data, such as sale records, to extensive and complex data communication records, such as telephone call details, in huge organizations, etc [Hand et al., 2001].

Apparently, statistical analysis can be performed on the collected database to find out some useful information about the database for decision making purpose. As the amount of the collected data increases, the database owners have needed to go beyond the classical statistical analysis. They rather they tend to extract some hidden patterns from their database. For instance, in addition to calculating the average daily sale of a supermarket, the owner may like to know which items are sold together most frequently in order to make some marketing related decisions. This aspiration makes the introduction of techniques and algorithms to extract some hidden patterns that may interest the database owner and useful for decision making. This gradually caused the birth of the concept of data mining.

2.3.2. Data Mining Techniques and Algorithms

In order to extract hidden patterns and discover new knowledge from a large database, there are different kinds of techniques and algorithms proposed and developed by various scholars. They are discussed as follows [Han & Kamber, 2006; Hand, et al., 2001]:

Association Rule Generation: This involves discovery of association rules showing attribute values that occur frequently together in a given set of data. This is frequently used for market-basket or transaction analysis in commercial sectors.

All the generated rules may not be sound enough for the concerned decision maker. The rules that seem interesting are considered for further actions. There are two commonly used types of interestingness measures in association rule mining: *Support* and *Confidence*.

For association rule: {A, B} → {C} ; A and B are premises, and C is the consequence.

Here, it is noted that the items proceeding the "→"symbol shows the number of items covered by the premises of the rule, and, the item(s) following the "→" symbol shows the number of items covered by the consequences (consequent items) of the rule. Thus, support and confidence can be calculated as follows:

$$Support = \frac{\#of\ consequent\ items}{Total\ \#\ of\ occurrence} \qquad (a)$$

$$Confidence = \frac{\#of\ consequent\ items}{\#of\ precedents\ (premises)} \qquad (b)$$

Equation 2.3-1: Formula for calculating Support (a) and Confidence (b)

For example, the following is a factious association rule output of a data mining process:

$$Age(X, \text{"20-29"}) \ \&\ income\ (X, \text{">2000Birr"}) \rightarrow buys(X, \text{"DVD Player"})$$

$$[Support=2\%,\ Confidence=\ 60\%]$$

The rule implies that if a customer X is in the age group 20 – 29 and has income greater than 2000Birr/ month then he or she is likely to by DVD player (i.e. he/she has bought also DVD player).

A support of 2% is that 2% of the transactions under analysis show that this is true. A confidence of 60% means that 60% of the customers in the given age group and income greater than 2000Birr bought DVD player.

Thus, support indicates the 'usefulness' whereas confidence indicates the 'certainty' of the rule. However, it is up to the data owner to decide the threshold in both cases.

There are also several algorithms to do association rule mining. The popular algorithm for association mining is *Apriori* method. It uses an iterative approach, which is known as level-wise search, where k-item sets are used to explore (k+1) itemsets. Hence, *Apriori* algorithm is based on candidate generation for identifying the most frequent items. At the first scan of the dataset, the most frequent items that satisfy the minimum support are identified. In the second run, they are joined to get their joint frequency and they checked for fulfilling the threshold (minimum support and confidence). Those itemsets that are not fulfilling the minimum threshold are rejected (pruned). Each step the *joining* and *pruning* task continues in the same fashion until all possible combinations of the items are finished. It can perform several cycles until the desired number of rules is obtained and/or until the minimum support and threshold are used up.

Despite the fact that A*priori* algorithms is the fundamental and the dominant one for association rule generation, there is also other methods by which frequent itemsets are extracted with out generating any candidate itemsets. *Frequent Pattern Growth (FP-growth)* is one of such algorithms. It is similar with *Apriori* in the first scan the database. Then, in the next steps it first compresses the database representing frequent items into a *frequent-pattern tree* (*FP-tree*), which retains the itemsets association information. It then divides the compressed database into a set of conditional databases. Each associated with one frequent item and mines each such database separately.

Classification and Prediction: These are two forms of data analysis that can be used to extract models describing important data classes or predict future data trends. A classification model can be used to categorize objects having similar attributes. For example, a classification model can be built to categorize users or pages.

A prediction model can be built to predict future trends. For example, knowing some group of customers' income and age, the prediction model can be used to forecast their potential to spend on particular equipment or simply to predict the potential buyers of a particular item.

In both techniques, two different data groups are prepared; one is for training purpose so as to develop a model. Then, the model is used for classifying or predicting class attribute of the main dataset to test its accuracy. The correctness of the model (the classier or the predictor) is its ability to classify or predict the values of the class attribute for the test dataset, which is previously unseen i.e. independent of the training dataset.

Classification can be done by Decision Tree Induction, Bayesian theorem based algorithm, Rule-based algorithm, or Neural Network Learning algorithm, etc.

Clustering: Clustering involves grouping objects so that objects within a cluster have high similarity and but are very dissimilar to object in other clusters. It is based on a principle of maximizing the intra-class similarity and minimizing inter-class similarity. Unlike, classification and predication, no class label is known or assigned to cluster the objects.

Clustering can be applied in many areas. In business, it is used to cluster customers according to their buying habit; in Web services, it is used to cluster users who access similar URLs based on the analysis of their usage pattern, etc.

Various types of methods are used to cluster a group of objects, such as Partitioning, Hierarchical, Density based, Grid based method, etc. For instance, Partitioning methods starts by grouping n objects into k groups, where $k \leq n$. It then uses an iterative relocation technique to improve the partitioning. The Hierarchical method uses a hierarchical decomposition of a given set of data objects. It may use top-down or bottom-up approach.

Outlier Analysis: Sometimes a database may contain a data object that may not comply with the general behavior of a data. Such data objects, for example, may represent an exaggerated or unexpected value. These data objects are called outliers. They can be revealed by applying some data mining algorithms. Even if they have unfavorable impact on other type of data mining tasks, they are useful for applications such as fraud detection and network intrusion detection. The analysis of outlier data is referred to as outlier mining. Outlier may be detected using statistical test that assume a distribution or probability model for data, or it may be detected using distance measures where objects may have considerable distance from the cluster of the same dataset.

2.3.3. Pattern Interestingness in Data Mining

A data mining has the potential to generate hundreds of or even thousands of patterns. Even though, these patterns have some level of usefulness regardless of a particular user or user group. Some patterns may provoke an interest of a user based on various criteria. Some patterns are also interesting if they are validated by the data owner. An interesting pattern represents knowledge. So, interestingness of a pattern refers to the extent to which the pattern assumed to represent some knowledge from both the user's perspective and the pattern itself.

To determine that a given pattern is interesting, there are several measures based on objective measures: statistics and structure of the pattern, etc, and subjective measures: users' belief in the data, unexpectedness, novelty, etc. Therefore, a given pattern is interesting if it is *a)* easily understood by the human, *b)* valid on new or test data with some degree of certainty, *c)* potentially useful, and *d)* novel. In addition, interestingness measure associated with a threshold. A user must put some sort of benchmark or ad-hoc rules to establish a limit in which the pattern is said to be interesting. Patterns that are below the threshold are regarded as uninteresting. They reflect noise, exceptions, or minority cases and are of less value.

The following are the measures of interestingness patterns [Han & Kamber, 2001, 2006]:

- **Simplicity**: pattern represents knowledge, but this can be true if it is easily understood by the human. Patterns that can not be understood by the user catch little of the user's interest. In other words, patterns that can not be comprehended by the user can not represent knowledge and hence it is uninteresting. This diminishes the significance of the pattern in order to convey some knowledge. Therefore, for the interestingness of the pattern, it must be easily grasped by the user.

- **Certainty**: the measure of certainty, or also called trustworthiness, of a pattern is based on the association rule. It is usually expressed in terms of the probability of the transaction of two or more items among the given datasets under consideration. For example, let that two items, A and B, happen to be occurring together (AuB) in 10% of the database and the chance of the occurrence of A is 55% provided that B occurs. These percentile figures show the *support* and the *confidence* level of the

co-occurrence of the two items, respectively. They are used to confirm the usefulness and certainty level of the pattern in order to determine that the pattern is interesting. However, the level of certainty is determined based on a threshold set by an expert or a user. If support and confidence level do not meet the minimum threshold, they may be rejected.

- **Utility**: as stated above, patterns may be generated fulfilling the minimum threshold; however, all the patterns may not have equal value to all users. Some patterns may have higher level of usefulness for a specific user at any given time and situation. This also may be impacted by the objective nature of the data mining task. For example, a market analyst may perform data mining on the transaction database to get which items are purchased more frequently in which season of the year. In this case, getting a pattern that shows items sold together may not be his interest, hence such pattern may has little or no use for his purpose. In addition, the level of support also reflects whether that pattern is potentially useful or not.

- **Novelty**: users usually do data mining with presumed expectations or beliefs. However, in the course of the investigation some unexpected patterns may be discovered. They are supposed to be new and, therefore, may attract the user's attention, means they are not discovered before. This concept refers to the interestingness measure that we call novelty. Such patterns may be contradicting the user's belief and, however, they offer strategic information on which the user can act.

2.3.4. Application of Data Mining

As compared to Online Transaction Analysis (OLAP) and statistical analysis, data mining have broader applications. Here are some of data mining applications [Han & Kamber, 2006, Signhal, 2001]:

- **Financial Data Analysis**: Financial institutions offer a variety of finance related services, such as credit and investment services. In this case, data warehouse may be used to generate monthly report. But, data mining can be used for more advanced purpose. Data mining techniques here are used for predicting risks related to loan payment and credit policy analysis.

- **Data Analysis in Retail Industry**: This industry is a suitable environment for data mining application because there is huge amount of data collected in day to day transactions. So, different types of data mining techniques are applied on these data thereby uncover some interesting patterns for decision making.

- **Intrusion detection and Network Security**: Intrusion detection is the process of identifying and responding to malicious activity targeted at computing and networking resources. Thus, intrusion detection system takes raw inputs from sensors and log records regarding the intrusion then analyses using a data mining techniques in order to take actions.

- **Data Analysis in Telecom Industry**: The telecom industry is one of the promising sectors for application of data mining techniques. So, data mining techniques are applied to improve the telecom services. For example, by analyzing calling patterns, it

is possible to determine what kind of calling plans to offer to improve profitability. Fraud detection is one of the activity on which data mining can be applied. Outlier analysis is a typical technique that can be applied in fraud detection.

- **DNA Data Analysis:** A great deal of biomedical research is focused on DNA data analysis. DNA data analysis has enabled the discovery of genetic causes of many diseases as well discovery of new medicines. But, the problem in genetic analysis is similarity search and comparison among the DNA sequences. Thus, data mining techniques can be used to solve these problems.

- **Web Data Analysis:** WWW is a global information center for news, advertisement, consumer information, financial management, education, government, e-commerce, and many other information services. It contains a rich and dynamic collection of hyperlink information, Web usage and access information. This provides a rich source for Web data mining. Hence, data mining is applied on such data to analyze the link structure of a Web and the interlinking among various pages to find out the Web sites that are authoritative and to determine the relevancy of a given page on a given topic. Web content mining can be applied for extracting useful information or knowledge from web pages. Apart from these two applications, web mining is also used for analyzing Web usage patterns on order to improve the Web service by analyzing the Web usage pattern.

2.4. Data Warehouse

In today's world business, the proliferation of ICT and data processing machines makes business firms and other organization to collect and store data electronically and simplifies the analytical and summarization process. In this respect, data warehousing plays an important role. A data warehouse is a repository of information collected from multiple sources, such as branch offices, market survey, sales transaction, etc. stored under a unified schema, and that usually resides at a single site [Han & Kamber, 2006]. Data warehousing supports information processing, for instant to make statistical summary for decision making, by providing a solid platform of integrated, historical data from which to do analysis. This means that it provides an integrated facility in a world of nonintegrated application systems [Sinha, 2001] because it encompasses algorithms and tools for bringing together data from distributed information repositories into a single repository that can be suitable for data analysis [Singhal, 2007]. The data warehouse organizes and stores the data needed for information, analytical processing over a long period of time [Sinha, 2001].

Singhal [2007] states the following features of data warehouse:

- **Subject oriented**: the data in the data warehouse is organized around major subjects such as customer, supplier, and sales. It focuses on modeling data for decision making.

- **Integration**: It is constructed by integrating multiple heterogeneous data sources such as RDBMS, flat files, and OLTP records.

- **Time-Variant**: Data is stored to provide information from a historical perspective.

2.4.1. Data Warehouse vs. Data Mart

For many people data warehouse and data mart seem to be similar. But, Griffin [1998] put the distinction as a data warehouse incorporates information about many subject areas, often the entire enterprise, while the data mart focuses on one or more subject areas. The data mart represents only a portion of an enterprise's data, perhaps data related to a business unit or work group. Typically, a data mart's data is targeted to a smaller audience of end users or used to present information on a smaller scope. In short, data warehouse is enterprise-wide whereas data mart is departmental-wide [Han & Kamber, 2006].

2.4.2. Role of Data Warehouse for Data Mining

Data warehousing has emerged as an increasing popular and powerful concept for applying information technology to turn the huge island of data into meaningful information for better business decision [Sinha, 2001].

Data warehousing is an enabling technology for data analysis in the area of retail, finance, telecommunication/ Web services and bio-informatics. For traders, data warehouse is used for various purposes such as for determining how the sales are differ from across regions/countries; to identify which region or branch store require the products in what interval to keep the stock full; and to assess the impact of promotion in a given region, etc. In other areas, data warehousing is used in Web services application for collecting the usage information and then identify usage patterns, catch fraudulent activities, make better use of resources and improve the quality of service [Singhal, 2007]. In all these cases, data warehouse is served as a source of information for data mining application. This is because, unlike statistical sampling, data mining needs much more datasets for increasing its accuracy in generating rules or models.

Chapter Three

Web Mining

3.1. An Introduction to Web Mining

The World Wide Web (WWW) shows a continuous growth at an amazing speed in the quantity and complexity of Web sites. It has greatly impacted every aspect of our societies and our life. The impact ranges from information dissemination to communication and from e-commerce to process management. By browsing through a Web site, users complete different tasks, such as buying products, registering for classes, attending classes online [Khasawneh & Chan, 2006], watch news stuffs, get map and weather information, etc. Along with the growth of WWW, the complexity of tasks such as Web site design, Web server design, and/or simply navigating through a Web site has increased. People at different times tried to address such issues by analyzing how a Web site is being used [Cooley, et al., 1999], how the site structure is made, and what and how the Web contents are organized. By their endeavor to address such issues, a range of possible techniques were proposed. All these lead to the introduction of various new ways to discover knowledge from a Web site data. Hence the term Web mining is coined to such endeavor. According to Kosala and Blockeel [2000], Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data. So, it can be defined as the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining merely uses the techniques that are used in the conventional data mining process except the source of data i.e. is the Web site. In this regard, Web mining can be defined

roughly as data mining using data generated by the Web and includes the following sub areas: web content mining, web usage mining, and web structure mining [Araya, et al., 2004] because the purpose of Web mining is just to try to acquire useful information and knowledge from the huge amount of information in WWW [Like, et al., 2004].

Kosala and Blockeel [2000] suggested that Web mining can be decomposed into the following subtasks, namely:

a) Resource Finding: Selection of tasks of retrieving intended Web documents

b) Information Selection and Pre-processing: Automatically selecting and pre-processing specific information from retrieved Webs as well as across multiple sites.

c) Generalization: Automatically discovers general patterns at individual Web sites.

d) Analysis: Validate or interpretation of the mined patterns.

In general, several scholars categorize Web mining into three areas of interest based on what part of the Web to mine: Web content mining, Web structure mining, and Web usage mining [ibid]. They are discussed in the following sections.

3.2. *Web Structure Mining*

Unlike the Web usage mining, Web structure mining concerns on the hyperlink structure of Web pages. A Web page contains not only contents but also hyperlinks to other pages. These hyperlinks contain an enormous amount of latent human annotation that can help to automatically infer the notion of authority. When an author of a Web page creates a hyperlink to another page, it can be considered as the author's endorsement of the other page's content.

Similarly, as several pages are pointing to a particular page, its importance and relevance for the content of pointing pages is approved, hence such page becomes authoritative. On the other side, the page that is pointing to the authoritative page is called a hub [Han & Kamber, 2006]. Web Structure Mining aims at finding the underlying topology of the interconnections between Web objects. As a result, the model built can be used to categorize as well as to rank Web sites. In addition, it can be used to find out similarity between Web sites [Baglioni, et al., 2003]. The motivation for Web structure mining comes from the method that was used in the 1970s. By that time researchers were using citation analysis to evaluate the strength of research articles. But, according to Han and Kamber [2006], Web documents have their own unique features. Web hyperlinks may not always represent an endorsement, rather the links perhaps be created just for paid advertisement or for navigational purpose.

3.3. Web Content Mining

Web content mining is one of the three Web mining categories. The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek by mining the Web contents [Iváncsy & Vajk, 2006]. It focuses on techniques for searching the web for documents whose content meets web users queries [Batisa & Silva, 2002].

The Web content is the data that the Web page was designed to convey to the users. This usually consists of text and graphics, but not limited to these two types [Srivastava, et al., 2000]. Web content consists of several types of data such as text, image, video, metadata as well as hyperlinks [Kosala & Blockeel, 2000]. Thus, content mining covers data mining techniques to extract models from web object contents including unstructured contents, such as plain text; semi-structured documents, such as HTML or XML; structured documents, such

as data in a table or in a database; dynamic documents, multimedia documents. The extracted models are used to classify web objects, to extract keywords for use in information retrieval, to infer structure of semi-structured or unstructured objects [Baglioni, et al., 2003].

Web Content Mining deals with problems of automatic information filtering and categorization, intelligent search agents, and personalization of web agents [Punin, et al., 2001]. So, it is the task of discovering useful information available on-line in order for organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents, etc [Iváncsy & Vajk, 2006].

In some recent researches, mining multi-types of data is termed as multimedia data mining. Moreover, much of the Web contents are unstructured documents (text data). Researches in this area employ data mining techniques to extract knowledge from the Web text data. Thus, both multimedia data mining and text mining are considered as instances of content mining [Kosala & Blockeel, 2000].

3.4. Web Usage Mining

An important area in Web mining is usage mining, which is the discovery of patterns in the browsing and navigation data of Web users. It has been an important technology for understanding users' behaviors on the Web [Fu & Shih, 2002]. Web usage mining can also be viewed as the extraction of usage patterns from Web access log data containing the browsing behavior of users [Batista & Silva, 2002]. The Web usage analysis includes straightforward statistics, such as frequency of page access, number of visits within a specified period of time, etc, as well as more sophisticated forms of analysis, such as finding the common traversal paths through a Web site , group of pages accessed together, etc [Cooley, et al., 1999].

From the server side there are three types of logs, namely access log, error log, and proxy server log. The client-side log is collected by implementing a remote agent or by modifying the source code of an existing browser to capture every activity of the client while interacting with Web sites [ibid].

The server error log is a file in which diagnostic information, or simply errors, which are encountered in processing requests, are recorded. It is the first place to look when a problem occurs with starting the server or with the operation of the server, since it will often contain details of what went wrong and how to fix it [11].

Web proxy servers act as an intermediate level of caching between client browsers and Web servers [Batista & Silva, 2002]. Proxy log, usually collected from and stored in the proxy server, is used to record requests and the cached pages.

Web access log, also known as Web server log, contains every users request to a specific Web site. Each request is recorded in the log by using the clients' IP Address or domain name as a record entry.

3.4.2. The Purpose of Web Usage Mining

Web usage information can be used to restructure a Web site in order to better serve the needs of users of a site. Complex traversal paths or low usage of a page with important site information could suggest that the site links and information are not laid out in an intuitive manner. The design of a physical data layout can be enhanced by knowledge of how users typically navigate through the site. Usage information can also be used to directly aid site navigation by providing a list of "popular" destinations from a particular Web page [Cooley, et al., 1999]. Thus, the purpose of usage mining here is used to identify the common paths users

frequently follow while navigating through a given Web site. Re-designing the Web page and re-arrangement of the Web content can be done based on the users' preference that is gained by mining their visiting behavior, for example by identifying which pages they most frequently access together. The algorithms that are used for association rule generation help to identify which pages are most frequently requested together by a user. Such kind of information can not be found by the Web log analyzer tools that generate the common statistical reports. But, mining for association rules among different Web pages together with the statistical summary helps a lot to get an insight of a given Web site's usage pattern.

3.4.3. Web Usage Mining Techniques

As Web mining is one form of data mining, it is possible to use any of the data mining techniques. But, few of these techniques are most commonly used in the Web mining endeavor. For example, some of the data mining algorithms that are commonly used in Web usage mining are association rule generation, sequential pattern generation, and clustering. Association Rule mining techniques discover unordered correlations between items found in a database of transactions. In the context of Web Usage Mining, a transaction is a group of Web page accesses, with an item being a single page access [Cooley, et al., 1999].

In usage mining the association rule is the correlation among access to various files on a server by a given client. For example, in association rule mining one can find the following correlation: 40% of clients who visited a given company Web site accessed the Web page with URL */Company/products/product1.html* also accessed the page with URL */Company/products/product2.html* [Mobasher et al, 1997].

In sequential patterns generation, it is possible to determine sequential patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. By analyzing this information, a Web usage mining system can identify relationships among data items (usually visited URLs). For example, 40% of the users who visited a URL */company/products/product1.html* visited the other URL */company/products/product2.html* after fifteen days [ibid]. The duration may not be necessary in days, but also in the same user session, the order of the visiting can be mined. For example, the following sequential pattern rule is generated based on order of visit in a single user session of and Official Web site of the 1996 Atlanta Olympic: 9.81% of the site visitors accessed the Atlanta home page followed by the Sneakpeek main page.

Clustering techniques usually applied to group Web site users according to their surfing behavior and to customize the Web page layout and content accordingly. In fact, this is possible whenever the users' identities are known.

3.4.4. Limitations and Challenges in Web Usage Mining

To do Web usage mining, it is possible to use server side data, client side data, or user registration data. However, there are some issues raised while trying to use these sources. To use a client side data, either cookies or a remote Java agent are required to collect her/his browsing trend and session, but there is privacy issue [Cooley, et al., 1999]. For such reasons the server side log data is preferred for mining usage pattern. The server side log data refers to the three types[3] of files: Web access log file, Error log file, and Proxy log file.

[3] They have been discussed in section 3.3.1

The Web access log holds records of each hit by a visitor. When the user requests for a single page, the browser parses it and generate requests for all embedded files in it (such as inline images) [Arlitt, 2000]. For example, the following scenario shows how the log registers the request when the user requests a Web page with URL `/index.html` [Barrett, 2008].

- The web browser asks for the URL *index.html*.
- The server sees the request and sends back the HTML page.
- The web browser notices that there are two inline graphic links in the HTML page, so it asks for the first one, *welcome.jpg*.
- The server sees the request and sends back the graphic image.
- The web browser then asks for the second image, *logo.jpg*.
- The server sees the request and sends back the graphic image.
- The browser displays the web page and graphics for the user.

In the Web server access log, the following lines would be added:

```
192.168.45.13 - - [24/May/2007:11:20:39 -0300] "GET /index.html HTTP/1.1" 200 117
192.168.45.13 - - [24/May/2007:11:20:40 -0300] "GET /welcome.jpg HTTP/1.1" 200 231
192.168.45.13 - - [24/May/2007:11:20:41 -0300] "GET /logo.jpg HTTP/1.1" 200 432
```

Thus, this situation causes a creation of a huge log file within a shorter period of time, which may not be handled by many Windows application to open. It also makes the preprocessing task difficult.

The other difficulty in processing access log is identification of user session. A user session is considered to be all of the page accesses that occur during a single visit to a Web site. The information contained in a raw Web server log does not reliably represent a user session file

for a number of reasons [Cooley, et al., 1999]. It is difficult to identify a user because a search engine usually does not have much information about the user unless he/she has registered, and the IP address is not a reliable resource due to the use of proxy servers and dynamic IP allocation [He & Goker, 2000].

In general, there are a number of difficulties involving in cleaning the raw server logs to eliminate outliers and irrelevant items, reliably identifying unique users and user sessions within a server log, and identifying semantically meaningful transactions within a user session [Cooley, et al., 1999]

However, if the Web access log is divided into sessions with the optimal session interval, at least the data collected within a session can be sure to be related to a particular user in a topic most of the time. So with the optimal session interval, much more data become available for the mining [He & Goker, 2000]. There are different methods to identify user session, such as time based heuristic method. In this method some scholars claim that users may have average time interval to stay in one Web site [Catledge & Pitkow, 1995; Berendt, et al., 2002]. For instance, Shen, et al [1999] used 1:00 hour duration for a session; Catledge & Pitkow [1995] used 25.5 minutes where as He and Goker [2000] found ten to fifteen minutes session duration from two Web log data. Usually, thirty minutes are taken for such a time-based sessionization [Araya et al., 2004; Srivastava, et al., 2000].

3.5. Applications of Web Mining

Information users (both the information provider and seekers) could encounter, among others, the following problems when dealing with the Web [Kosala & Blockeel, 2000]:

- *Finding Relevant Information*: Users may take more that the usual time to get some information relevant to their need and they may end up with no relevant information after surfing the web for several hours.

- *Creating New Knowledge out of the Information Available on the Web*: Some users particularly Web based information providers lack the skill and the tools to extract useful information from WWW data and services.

- *Learning about Individual Users*: This is a problem that specifically deals with personalization of the information, which simply knowing what the users do and want. The tasks needed to do are either mass customizing to the intended users or personalize the Web service to individual users.

Web mining techniques are aiming to deal with those problems. As a result, it is possible to design a Web site that is suitable for the targeted user community because one of the applications of Web usage mining is to personalize the Web contents and services. Web usage mining can help in addressing some of the shortcomings of the standard approaches for web personalization. However, it should be noted that the discovery of patterns from usage data is not by itself sufficient for performing the personalization tasks [Batista & Silva, 2002].

In addition, the extracted knowledge can also be used for other applications, such as improving site usability, business intelligence, and usage characterization [ibid]. By improving site usability, it is to mean that the discovered usage pattern can help the Web designer and the owner in identifying the users visiting habit or interest thereby they are able to improve the Web site's content and layout design. This, in turn, facilitates the users' interaction with the Web and stamp positive experience in their mind.

Applying Web mining technology for business intelligence is becoming common particularly among e-commerce companies. Understanding user access patterns in a Web site using mining techniques not only helps to improve Web system design, but also leads to wise marketing decisions, such as putting advertisement in proper places, classifying users, etc [Zaine, et al., 1998].

According to Araya et al. [2004], analyzing Web data can also be used for system improvements by providing the key to understand the Web traffic behavior. Advanced load balancing, data distribution or policies for Web caching as well as higher security standards are potential benefits of such improvements.

In addition to what is stated above, Web mining techniques could be used to solve the information overloaded problems directly or indirectly. However, it is not claimed that Web mining techniques are the only way to solve those problems [Kosala & Blockeel, 2000].

3.6. Statistical Approach in Web Usage Analysis

The discovery of Web usage patterns in this research is accomplished by using both the data mining and the statistical techniques. Statistical reports can be useful for identifying usage patterns from the Web log records. To do this, a number of Web usage statistical report generators are available both for free and commercially.

Statistical techniques are the most common method to extract knowledge about visitors to a Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, etc.) on variables such as page views, viewing time and length of a navigational path, etc. Many Web traffic analysis tools produce a periodic report

containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This report may include error analysis such as detecting unauthorized entry points or finding the most common invalid URLs. Despite lacking in-depth analysis, this type of knowledge can be potentially useful for improving the system performance; enhancing the security of the system; facilitating the site modification task; providing support for marketing decisions; etc [Srivastava, et al., 2000]. Statistical summary reports, comprising of information such as client sites; types of browsers; and the usage time statistics, are used to give understanding about the Web sites usage and it is also very important for web masters to know the efficiency of the web server [Punin, et al., 2001].

On other hand, data mining techniques are used for getting hidden patterns from the log records. Therefore, using statistical techniques together with data mining techniques for Web usage pattern discovery gives a better insight how the Web site is being used.

3.7. Related Works

Different aspects of Web 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 most efficient approach for Web access association mining. Their approach consists of three steps: *a)* transform raw web logs to a relational table; *b)* convert the relational table to a collection of access transactions; and then *c)* 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 1:00 hour time duration for each session and finally presented the experiment results.

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. They discouraged making sessions in Web log based on the 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. According to these scholars, this tendency lacks a more user oriented view. Their argument is that a session on the Web can be defined as a group of user activities that are related to each other not only through an evolving information need but also through close proximity in time. They did the experiment on two Web sites log data. Finally, their experiment revealed a 10 to 15 minutes threshold between user activities for an appropriate session interval.

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 URLs for personalizing the Web service based on the reading pattern of the users. Using commercial data mining software systems, they have identified and characterized several reading patterns within the news site. These patterns would define user profiles which integrate a news recommendation system based on web user preferences.

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. They used user registration data in order to use it together with the Web access log data. They have conducted experiments that built classification model for inferring an association between sex and interest of users based on their navigational behavior. Another experiment they have conducted is to predict whether a user might be interested in visiting a section of the web site based on the sections the user has already visited.

In general, the majority of the articles in which the writer of this thesis came across show efforts to provide a general framework / architecture for Web mining and in few of them attempts were made to perform mining activities using techniques for association rule generation.

Chapter Four

Modeling the Research Process

4.1. Overview

Web usage mining is relatively a young research area that is being still researched; hence there is no universal and general approach to be adopted while undertaking a Web usage analysis. As a matter of fact, the writer, experienced that the Google™ search engine retrieved the usual statistical analysis tools for the search phrase '*Web usage mining tool*', which probably shows that there is no full-fledged tool for Web usage pattern discovery that incorporates both the statistical and the data mining features.

In this study, the writer attempts to model the approach for Web usage mining in which both the data mining and statistical techniques are employed to find out the Web usage pattern based on the premises discussed in the following sections.

4.2. Format of Web Log Data

The format of a log record may vary depending on the type of server and its configuration. Most of the logs hold almost similar information; however, the format and order of the attributes may vary. The *common log format*, as set by *World Wide Web Consortium (W3C)* [16], contains the following fields:

For example, the following fictitious log entry shows these fields populated with values in a *Common Log Format* file record:

```
125.125.12.25 - - [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043
```

The following are the fields in the *Common Log Format*:

- ***Host*** (125.125.12.25 in the example): The IP address or host/ sub-domain name of the HTTP client that made the HTTP resource request.

- ***rfc931*** ("-" in the example : This is an identifier used to identify the client making the HTTP request. If no value is present, a "-" is substituted.

- ***Username*** ("-" in the example): The username (or user identification) used by the client for authentication. If no value is present, a "-" is substituted.

- ***date:time time_zone*** ([10/Oct/1999:21:15:05 +0500] in the example): The date and time stamp of the HTTP request.

- ***request*** ("GET /index.html HTTP/1.0" in the example): The HTTP request. The request field contains three pieces of information. The main piece is the requested resource (`index.html`). The request field also contains the HTTP methods either `GET` or `POST` and the HTTP protocol version (`1.0` or `1.1`).

  The `GET` method is used when a user types a URL into the address bar of the browser; or she/he clicks on link in a document displayed in the browser; or when the browser download images for displaying within the `HTML` document. The `POST` method is

typically is used for sending information collected from a form displayed within a browser [Jackson, 2007].

- ***statuscode*** (200 in the example) : The status is the numeric code indicating the success or failure of the HTTP request.

  The numbers appeared in the status column indicate various forms of the request status. Here are some of the common status codes: 200 indicates successful status (OK); 301 tells that the requested URL has been changed (Moved Permanently); 307 indicates that the RUI for requested resources has been redirected or at least changed temporarily (Temporary Redirect); the status code 401 could be printed if the requested resource is password protected (Unauthorized); 403 indicates the requested resource available but read-protected and it could be intentional or error from the server administer (Forbidden); 404 appears when there is no the requested resource at the server(Not Found); and 500 returned when an internal server software failure detected (Internal Server Error) [Jackson, 2007].

- ***bytes*** (1043 in the example) : The bytes field is a numeric field containing the number of bytes of data transferred as part of the HTTP request, not including the HTTP header.

In the *Extended Log File Format*, there are additional fields incorporated in the log, such as the Referrer (the other Web sites that leads to the current Web site); the client's browser type and version; etc.

The input for the Web usage mining process is a Web log file, which is referred to as a user session file that contains the access log records for each user. It gives an exact accounting of who accessed the Web site; what pages were requested and in what order; and how long each page was viewed [Cooley, et al., 1999]. The file format can be a simple ASCII text file, a spreadsheet, database file, or any other format that is specified by the data mining tool like that of WEKA's (.arff) file format. Thus, it requires some preprocessing tasks to make the data suitable for the selected tool.

4.3. WEKA File Format

WEKA, a data mining tool, has its own format for the data file it handles. The WEKA file needs to have an ".arff" file name extension. ARFF files have two main sections. The first section is the *Header* information, which is followed by the *Data* information. The *Header* of the ARFF file contains the name of the relation; list of attributes (the columns or field names in the data); and their data types. The *Header* section is identified by the two tags: *RELATION* and *ATTRIBUTE*. The data section has a *DATA* tag. All tags are preceded by @. Comments, which form the optional section, are identified by '%' mark (See Figure 5.4-5).

4.4. Association Mining by Apriori Algorithm

The *Apriori* algorithm is used for association mining by selecting frequent items for building candidate itemsets in order to generate association rules. The items are supposed to be in separate columns i.e. one item per column. This means that the association rule is made between/ among items in different columns. But, when we see a Web log record, we can observe that the log records need further data preparation task in addition to the usual data cleaning task so as to make associations between/among URLs.

The different algorithms implemented in WEKA also require different data types. For instance, the *Apriori* algorithm requires the values to be nominal and missing values should be replaced by a '?' (question mark). Then, the algorithm excludes the attributes with a '?' from the frequent itemsets.

4.5. Cleaning Web Log Record

Data preprocessing for data mining task is one of the indispensable activities in KDD. It usually starts from data cleaning process. In discovering knowledge from Web log data, data cleaning can be used to remove irrelevant data, such as log records for images, scripts, help files, and cascading style sheets. Only data that are relevant to the user identification process are kept [Khasawneh & Chan, 2006]. On the other hand, the data cleaning features available in data mining tools, such as filtering features in WEKA, do not fit to handle the cleansing task in Web log data due to the very nature of log records. For example, the data cleaning in relational database record may be replacing the missing value with any one of the available values; or it may be replacing outliers with some common values, etc; however, this may not work for Web log data.

Consequently, there is a need to adopt some other mechanism to carry out such task. In this case, the only option is to write some codes using programming language.

4.6. User Sessions and Transactions

As it has been stated in the previous chapter, session identification must be done before submitting a Web log data to the usage mining tool. Each identified session based on the chosen criteria represents duration of a user's stay at the Web site. A user is defined as a unique client to the server during a specific period of time [Khasawneh & Chan, 2006]. Thus, a user session means a time from the user enters into the site up to she/he leaves the site.

The task of user identification is to identify records that are belonging to the same user from log records which are recorded in a sequential manner. Then, transaction identification should be done for mining the association among the URLs the users may request. In transaction identification, the URLs accessed by the same user are put together alongside the session identification that serves as an entry to the transaction.

4.7. Tools Selection

As stated earlier in the current chapter, data mining in general, Web usage mining in particular, is relatively young fields of study; hence it is hard to get a full-fledged system for discovering knowledge from a database of Web access records. Therefore, two different tools have been selected for this research: one for statistical report generation and the other for data mining.

4.7.1. Tool for Data Mining

The primary tool used for the Web usage mining is WEKA (*Waikato Environment for Knowledge Analysis*). The main reasons why WEKA has been chosen are:

- WEKA is now among the newly emerging and popular data mining tools;
- It is also easily obtained as a freeware;
- An *Apriori* algorithm, which is the popular algorithm for association mining, has been implemented in WEKA;
- It has user friendly graphic user interface; and
- The writer has a better understanding of the tool.

However, WEKA had been originally developed for conventional data mining purpose hence it lacks features for preprocessing and analyzing Web log data as it is. To overcome this limitation, a WEKA plug-in, which have been downloaded from the Internet for free, for

preprocessing (filtering and session identification) of Web log data for mining in WEKA has been used. *WUMPrep4WEKA* is a plug-in to be integrated with WEKA for preprocessing Web log data; however, it has some limitation, too. For example, it has no a feature for Transaction identification for performing association rule mining since WEKA requires the dataset to be put in distinct columns to generate association rule among them.

WEKA uses a file format of its own. An *ARFF* (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

4.7.2. Tool for Statistical Analysis

For the statistical analysis, *Mach5 Analyzer* has been selected. The main reasons to choose this tool for the statistical analysis are:

i) It is capable of producing the statistical reports discussed in this work and many more;

ii) It can present the report in both chart and graph;

iii) It can show the common visitors entry and exit pages graphically;

iv) Is can help to generate a hit report specific to some directories; and

v) It can be downloaded from the Internet for free.

4.7.3. Programming Language

Python programming language is used for writing codes for some preprocessing tasks, such as for attribute selection; transaction identification; session-URL matrix generation; unique URLs identification; etc. The language has been chosen because it is more appropriate for text processing as the log file stored as a text file. Moreover, the language has simple syntax and built-in data structure to easily manipulate text data.

4.8. Model for the Web Usage Pattern Discovery

As it has been indicated in the previous discussions in the current chapter, the discovery of Web usage pattern is a young research area. In this research, the usage pattern discovery is carried out by using both the data mining and the statistical techniques. As a result, additional and/or different data preprocessing tasks are required.

Considering all what have been discussed above, the general approach used in this research has been modeled and the following illustration depicts it.

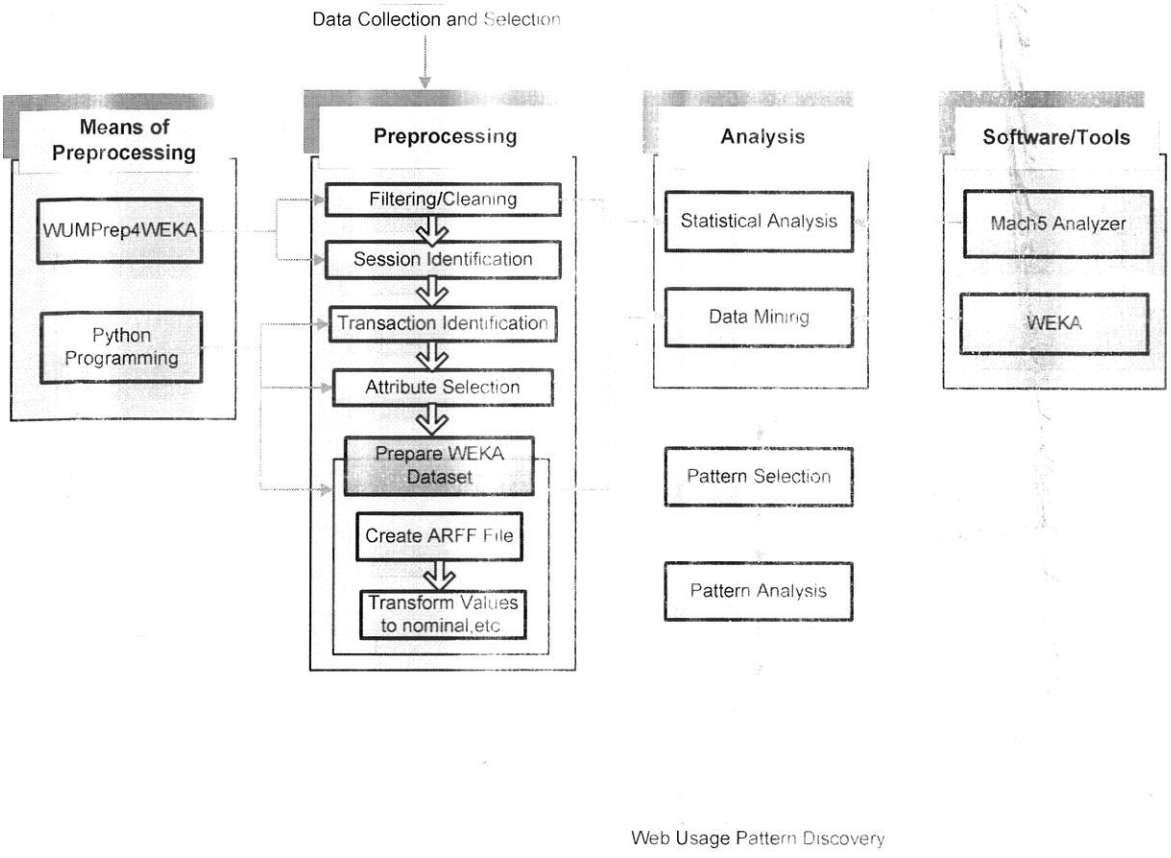


Figure 4.8-1: Process model of a Web Usage Mining Approach used in the study

Chapter Five

Experiment

5.1. Overview

In this chapter, the experiment has been conducted based on the model endorsed in the previous chapter. The data have been analyzed statistically and mined for finding any association rule between/among URLs.

5.2. Experiment Setup

The experiment has been conducted on the following Setup:

- Computer Type: *Personal Computer (X86-based PC)*
- Operating System: *OS Name Microsoft® Windows Vista™ Home Premium*
- Processor: *AMD Turion™ 64 X2 Mobile Technology TL-56, 1800 Mhz, 2 Core(s)*
- Primary Memory: *1024 MB*
- Data Mining Tool: *WEKA 3.4.11*
- Statistical Analysis Tool: *Mach5 Analyzer 4.1.7*

5.3. Data Collection and Selection

The data for this study is a Web access log data of AAU official Web site. As mentioned in the previous chapters, a Web log data is favored by many for Web usage analysis. The data warehouse that is considered in this study is the Web log records of the Web site that have been recorded and accumulated on the AAU Web site server since the launching of the Web site.

A three months Web access log data have been selected by judgment sampling method and they a they comprise 25% of the whole year. The data latest data available from the data owners. Thus, double advantages are gained: Firstly, the Januar when the regular students are in campus (the represent the time when regular students are on v (the 'Kiremt' season, which is the shorter one representative for one academic year data (1 separately to mitigate the impact of any possibl similar result, it is possible to conclude that the otherwise the Web usage trend varies on differen

5.4. Data Preprocessing

5.4.1. Dividing the Log File

The log data that was obtained from the Web se to be processed as it is. The *Python* program c into manageable size files.

For this study, the data preparation includes later re-merging after filtering individually. As user's request for a single page resulted in mu by excluding such entries reduces the log size

Accordingly, the log file has been divided into smaller size files. Then, the files have been fed into the *WEKA* plug-in, *WUMPrep4WEKA*. Using the *WEKA's KnowledgeFlow* graphic user interface application and the Python codes, further preprocessing have been performed. (See Appendix A and B for knowledge flow diagram of request filtering and session identification, respectively).

5.4.2. Filtering/ Cleaning

The Web access log records file may have entries that contain irrelevant records for the discovery of usage pattern. Thus, such records should be filtered and excluded from the dataset. Most of the statistical analysis tools have their own facility to let the user fill in the criteria to filter out the irrelevant request records. However, due to the size of the log file, it has been found advantageous to use the same filtered dataset for both the statistical analysis and data mining tools.

Primarily, the entries of the log that contain inline or embedded image files are removed. The common entries identified in the AAU Web Site are GIF and JPG/JPEG format images. Entries with these file name extensions have been removed. Table 5.4-1 shows the items served as criteria during filtering.

S.No	Entries Excluded from the Dataset Include:	Reason/ Explanation
1	*.gif *.jpg or *.jpeg *.mov *.png, etc	By scanning the log records, it has been found that the files with these extensions are embedded images; such as logos, welcome images, etc, which have nothing to do with the content of the page.
2	*.css	Cascading Style Sheets (CSS), a control over the presentation of the HTML documents, i.e. they

		used to determine the look of the Web page through browsers. They do not have contribution for the usage pattern discovery [Jackson, 2007].
3	*.cgi *.js	Items with those suffixes are common scripts that have no contribution for the usage mining. A URL with path beginning with /cgi-bin/ is interpreted by many servers as request for CGI generated content.
4	POST, HEAD	These methods are used for conveying a request to the server that are emanated from forms, such as Web user registration forms, etc.
5	*.doc, *.xls, *.pdf, *.txt	Entries with files with these extensions are rare, hence they are removed.

Table 5.4-1: Log Filtering criteria and their respective explanations

Entries with the request method POST have been also excluded from the dataset because the AAU Web Site has no a feature to entertain this method, such as forms to be filed out by users, etc.

These are the main criteria to exclude some log entries; however, some other filtering criteria also applied for individual files specific to the mining techniques, such as removing attribute values that are not important for the analysis.

In addition to filtering the entries, cleaning the log data have been done. To avoid inconsistencies, all the pages having ".html" extension have been converted to ".htm" and the texts have been converted to lower case (small letter) form.

5.4.3. Selecting Attributes

Attribute selection has been applied for some dataset depending on the type of the Web usage mining techniques used in the analysis. For example, for the URLs association mining, all the attributes except the URL have been removed because later the URLs are concatenated horizontally where each URL is located in a column.

5.4.4. Session Identification

For user session identification, the time heuristic method is used because the log format does not allow using other methods and thus 30 minutes, which is the default value for the preprocessing tool, has been used for this experiment.

After the preprocessing, user sessions have been identified, which means that the Web site is visited virtually by the number of users that is equal to the number of sessions within the given period of time

5.4.5. Transaction Identification

Since the format of a Web log data is not suited for direct import into the mining algorithm, further data preparation tasks are required. Transaction identification is the task of grouping of page references based on the user sessions like that of market-basket analysis where the transaction definition is the items purchased by a customer at one time [Cooley et al. 1999]. This is because the existing data mining techniques for association rule mining require the attribute to be in distinct columns. For this experiment, the URLs requested by the same user have been grouped and put in columns, like *{URL1, URL2,...URLn}* where *n* is the number of unique URL requested in the log. Then, the URLs requested by a user should be fall in any of the columns. To do this, a Python code has been written. Figure 5.4-1 shows the algorithm

developed for transaction identification. Figure 5.4-2 also shows the transactions created for each user session using the algorithm. The IP numbers are hidden for privacy reason.

```
Algorithm: Concatenating URLs alongside their session ID

Input: F, sessionized log data file
Output: T, a file containing user request transactions

L=( )          // each request in F, such that L1 ..Ln, where n is the number of requests
SID=[ ]        // session id
newSID= [ ]    // current session id
prevSID=[ ]    // previous session id
List= [ ]      // list of items in a line
Count=0        // initialized to count number of requests
WHILE i<=Count:
        Read L from an input file
        List={array of items in the line L after separated by space}
        tempList={array of null items list}
        newSID= List[0]          // List[0] is the session id for current line's List
        IF newSID==prevSID THEN
                Append the List[6] to tempList  // List[6] is the URL for current line's List
        ELSE
                Append prevSID and tempList to myList
                Write Mylist to T
                Assign the current session ID to prevSID
        END IF
        Count=Count + 1
END WHILE
```

Figure 5.4-1: Algorithm for transaction identification

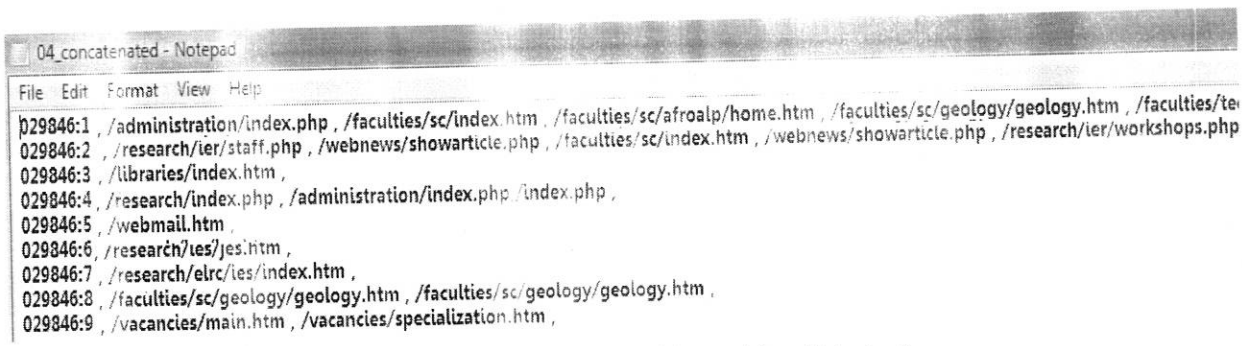


Figure 5.4-2: Comma separated transaction for log data of August (partial view).

5.4.6. Preparing the Dataset for Mining Tool

After the transaction identification has been performed, some tasks should be done to prepare the data for the mining tool. The identified transactions have long text rows that reduce the readability, which makes it difficult for human to easily assimilate the dataset contents. Hence

each URL needs to be represented by a shorter name and then the column headings have been named as URL1, URL2,...URLn, as it is stated above. The values under each column have been represented either by '1's or '0's, where '1' means that the URL was requested with that user and '0' otherwise. But, '0' is used for attribute definition only and for actual sessions it needs to be replaced by '?' to tell the mining algorithm that the value is missed and then it is excluded from the mining process while generating large itemsets. Then, the log data need to put in WEKA understandable format i.e. *Attribute-Relation File Format* (.ARFF).

```
Algorithm: To create Session-URLs matrix

Input: T, a file containing user request transactions
Input: Q, a file containing unique URLs list that fulfill min support
Output: V, a file containing Session-URLs matrix

L=( )          // each request in F, such that L1 ..Ln, where n is the number of requests
SID=[ ]        // session id
newSID= [ ]    // current session id
prevSID=[ ]    // previous session id
List= [ ]      // list of items in a line
Count=0        // initialized to count number of requests
Read uniqueList from Q
Read inputList from T
Count the lines in T
WHILE i<= Count:
        IF i<=Count THEN
           FOR (k=0; k<=lengthOfUniqueList; k++) DO
                Assign 'null' to UniqueList[k]        'null' can be '0' in the code
           END FOR
        END IF
        Read inputList from an input file T
        Calculate the length of inputList
        FOR (k=0; k<=lengthOfInputList; k++) DO
            FOR (j=0; j<=lengthOfUniueList; j++) DO
                IF inputList[k] == UniueList[j] and VectorList[j] < > inputList[k] THEN
                  Assign inputList[k] to VectorList[j]
              END IF
            END FOR
           AssignCounter++
        END FOR
        IF AssignCounter >=2 THEN    //to skip singleton transactions
          Write VectorList to file V
        END IF
        AssignCounter=0
        i++
        VectorList=[ ]     // initialize for the next iteration
END WHILE
```

Figure 5.4-3: Algorithm for creating Session-URLs matrix

The following figure shows what has been (can be) done by the above algorithm.

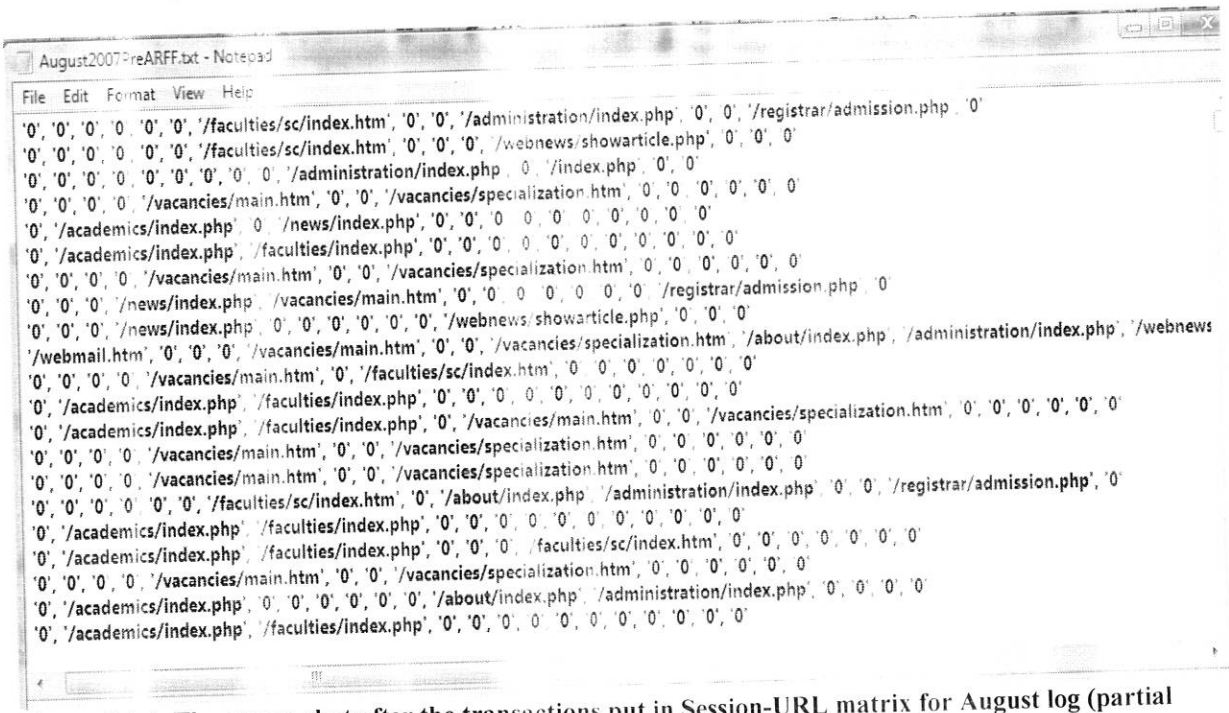


Figure 5.4-4: The screen-shot after the transactions put in Session-URL matrix for August log (partial view).

Figure 5.4-4 shows the identified transactions, after they are put in Session-URL matrix, where each row is a single transaction (user session) and the column represent a unique URL. The first column is URL1; the second is URL2; and so on. If the URL was not requested by a user, the program automatically puts '0' under that URL column. For example, the first user requested only URL7, URL10 and URL13. It is also noted that it is not necessary to include the user session identification for association rule generation since a row represents a single session.

The following figure shows after the transactions are transformed to '1's and '0's and put in WEKA file format. But, each '0' has been replaced by '?' later for generation of rules.

In fact, the general assumption is to create Session-URL vector for each user session. But, in reality this could not be possible due to two reasons: for one thing, creating a column for each

The following figure shows what has been (can be) done by the above algorithm.

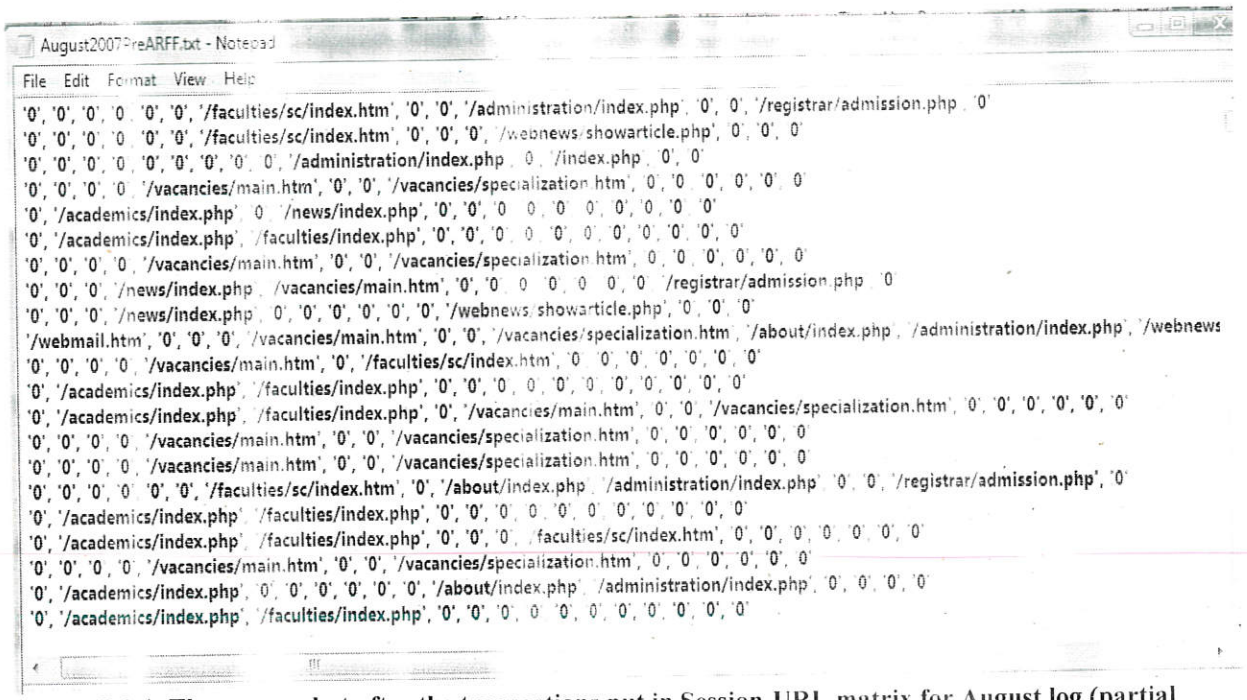


Figure 5.4-4: The screen-shot after the transactions put in Session-URL matrix for August log (partial view).

Figure 5.4-4 shows the identified transactions, after they are put in Session-URL matrix, where each row is a single transaction (user session) and the column represent a unique URL. The first column is URL1; the second is URL2; and so on. If the URL was not requested by a user, the program automatically puts '0' under that URL column. For example, the first user requested only URL7, URL10 and URL13. It is also noted that it is not necessary to include the user session identification for association rule generation since a row represents a single session.

The following figure shows after the transactions are transformed to '1's and '0's and put in WEKA file format. But, each '0' has been replaced by '?' later for generation of rules.

In fact, the general assumption is to create Session-URL vector for each user session. But, in reality this could not be possible due to two reasons: for one thing, creating a column for each

unique URL make the dataset nonsense where large numbers of missing values were recorded as some URLs were requested very rarely, for instance, once.

05_Transformed - Notepad

File Edit Format View Help

```
%     Title:  ARFF File for AAU Web Access Log Data
%     Created By:  Mekonnen Tsegaye
%     Date Created: November, 2008

@relation  AAUWebLogAugust

@attribute  URL1  {1,0}
@attribute  URL2  {1,0}
@attribute  URL3  {1,0}
@attribute  URL4  {1,0}
@attribute  URL5  {1,0}
@attribute  URL6  {1,0}
@attribute  URL7  {1,0}
@attribute  URL8  {1,0}
@attribute  URL9  {1,0}
@attribute  URL10  {1,0}
@attribute  URL11  {1,0}
@attribute  URL12  {1,0}
@attribute  URL13  {1,0}
@attribute  URL14  {1,0}

@data
0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0
0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0
0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0
0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0
1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0
0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0
0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0
```

Figure 5.4-5: The transformed dataset in WEKA standard file format for August log (partial view).

For this experiment, the general rule has been used is that the uniquely identified URL must fulfill a minimum 3% support of the total number of transactions in the best-case scenario. The best-case scenario here is that if the whole transaction, t, is processed in the association rule generation and if the URL, n, found only once per transaction.

Then, any given URL, *n*, could be identified as a candidate large itemset if it could gain at least 3% support threshold. Here, 3% chosen as a minimum because if it goes beyond that, the number of columns may increase and complicates the mining process. Moreover, those items that are below this threshold have less probability to be among the best association rules due to their low support value: Thus,

$$\frac{n}{t} \geq 0.03$$, n is any URL in the transaction and t is number of transactions.

Equation 5.4-1: A formula to determine the minimum support threshold for selecting URLs for association rule generation

Accordingly, fulfilling the minimum threshold, some numbers of URLs have been selected for preparing the dataset (See Table 5.4-2).

S.No	Months (2007)	Total Sessions	Unique Pages	Sessions for Experiment	Selected Pages
1	January	22314	1429	2423	18
2	May	23037	2226	2161	18
3	August	18640	2089	3348	14

Table 5.4-2: Month-wise statistics after transactions identification has been done.

5.5. Data Analysis

5.5.1. Statistical Analysis

The dataset which have been filtered have been fed into the statically analysis tools to find out a usage pattern. The Web access logs for some sample months have been analyzed individually. Thus, using the statistical analyzing tool, the following outputs have been generated.

5.5.1.1. Hits Statistics for the Sample Months

S.No	Items	Months (of year 2007)		
		January (31 days)	May(31 days)	August(31 days)
1	Hits (Request Records)	170033	118702	106792
2	Average Hits per Day	5484.94	3829.10	3444.90
3	Total Failed Requests	3477 (2.04%)	5656 (4.76%)	4376 (4.10%)

Table 5.5-1: Hits statistics for the three months.

The total hit is relatively higher in January than the remaining two months. However the numbers of hits recorded on pages showed decrement proportional to the total hits. This proposition also exhibited on the average number of hits per day. Higher percentage of failed requested were recorded during May and August. Despite the fact that higher number of visitors was recorded in January, less failed requested were shown in the same month (Table 5.4-2 and 5.5-1).

5.5.1.2. Most Requested Pages

The following table shows the top ten most accessed pages during the month of May, 2007, Web access log data. See appendices for January and August months' graph.

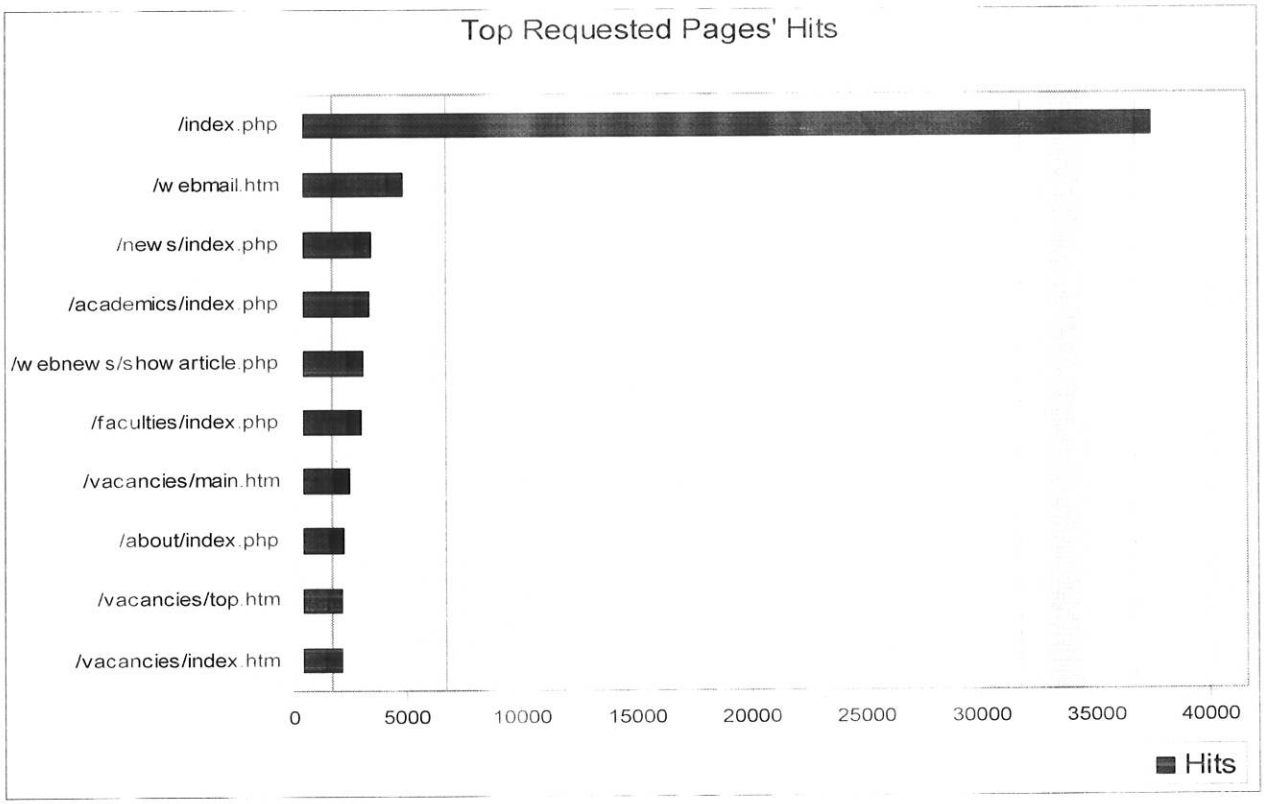


Figure 5.5-1: The top-ten frequently requested pages during May.

Figure 5.5-1 shows the top ten most requested pages during the month of May 2007. As it is shown the most requested page is the `/index.php` page followed by `/webmail.htm` page and the `/news/index.php` is the third top page.

This is a reflection of that the `/index.php` page is most poplar by most users in all the three months. In fact, this shows that most visitors enter into the site directly by typing the Web site address as it is shown in the following sections (5.5.1.3 & 5.5.1.4).

5.5.1.3. Most Visited Directories

As Table 5.5-2 shows, the root directory "/" is the most accessed directory where the `/index.php` is located. Most users also show interest on the contents under the /vacancies/ directory. It is also possible to say that `/faculties/tech/` is the third popular directory. The other directories are also having almost similar popularity in the three months.

Rank	January		May		August	
	Directory Name	Hits	Directory Name	Hits	Directory Name	Hits
1	/	75%	/	61%	/	60%
2	/vacancies/	6%	/vacancies/	10%	/vacancies/	9%
3	/faculties/tech/	3%	/webnews/	6%	/strategicplanning/	5%
4	/faculties/linguistics/	3%	/libraries/	4%	/faculties/tech/	4%
5	/webnews/	3%	/faculties/tech/	3%	/academics/	4%
6	/academics/	2%	/academics/	3%	/news/	4%
7	/libraries/	2%	/news/	3%	/libraries/	4%
8	/faculties/	2%	/faculties/coe/	3%	/webnews/	4%
9	/research/ies/	2%	/faculties/	3%	/faculties/	3%
10	/administration/	2%	/faculties/linguistics/	3%	/administration/	3%

Table 5.5-2: Percentage of the most frequently access directories in the three months.

In the three months. the most accessed directory is the root directory. In fact, this may not be surprising as many of the users visited the `/index.php` page many folds times than other pages. Even if it is to far from the `/index.php` page access frequency. the `/vacancies/` directory have got many visitors.

5.5.1.4. Most Frequent Entry and Exit Page

Entry pages are pages that the Web site users visited first as he/ she enters to the Web site where as the exit page is the last page the user visited during his/her session. Figure 5.5-2 illustrates the main entry and exit page for the January log data. (See the Appendices C and D for the other two months output).

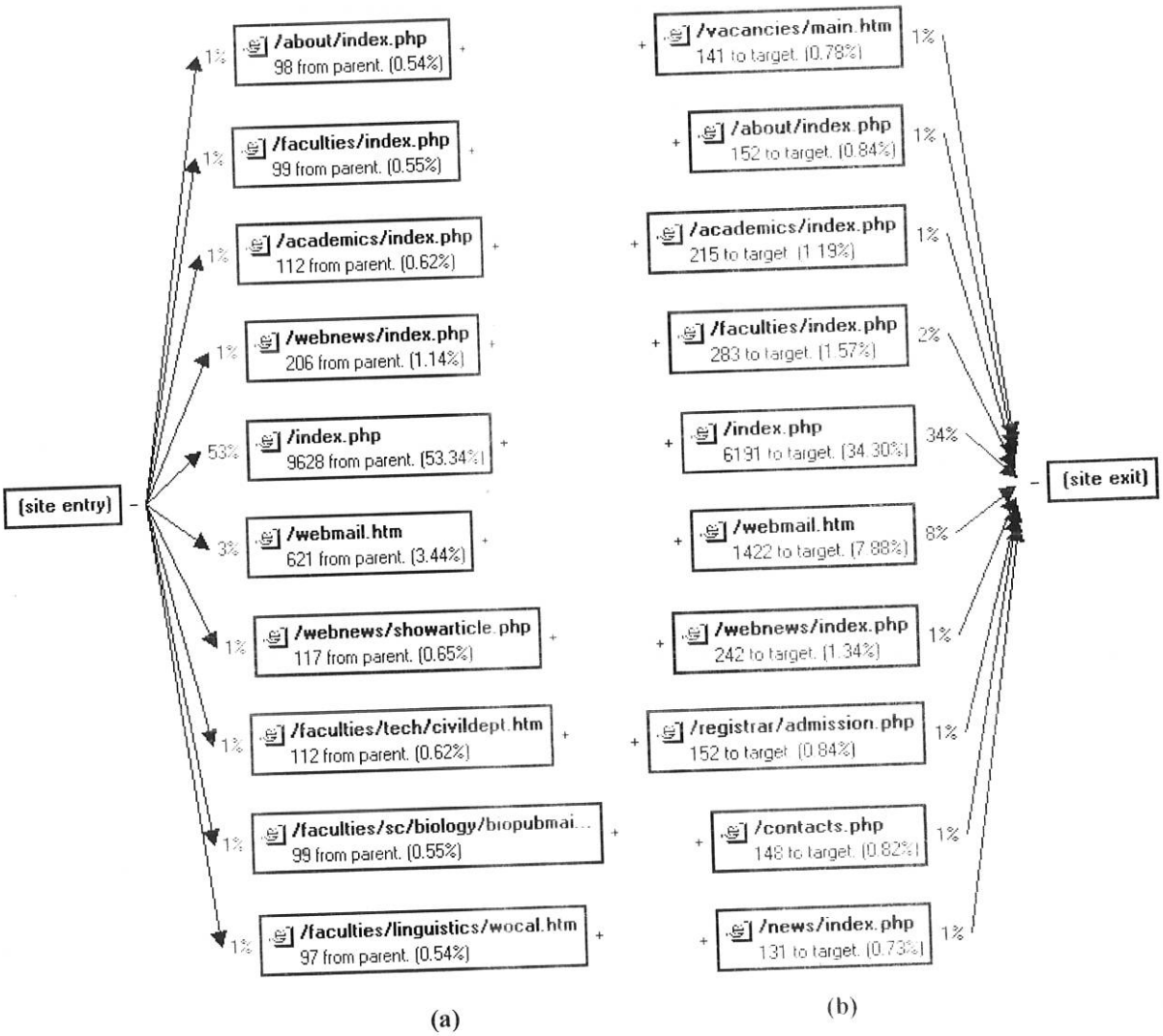


Figure 5.5-2: The common entry and exit page illustration from the log data of January.

For the month of January, 2007, as it is shown in the Figure 5.5-2a, 53% of the visitors have entered into the Web site directly through the `/index.php` page. This is also same as the May visitors whereas 52% of August's visitors also entered to the site via the home page.

Figure 5.5-2b also shows that 34% of the users have left the Web site from the home page for the month of January. Similarly, 35% of the May and 29% of the August visitors left the Web site from the home page.

5.5.1.5. Users' Visiting Time

Visiting by Days of the Week

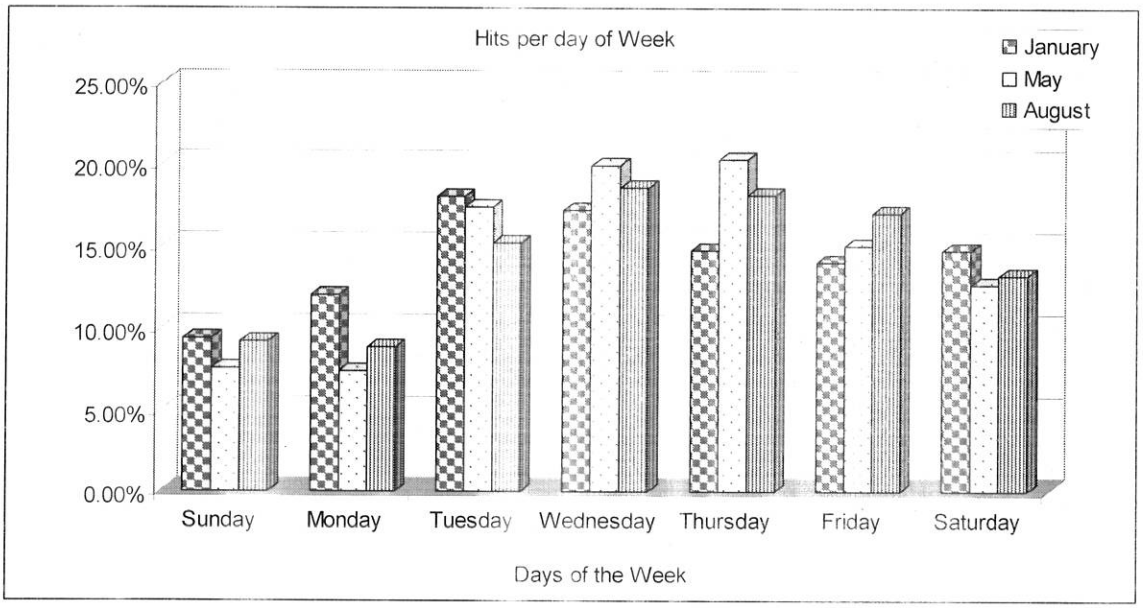


Figure 5.5-3: Number of visitors per days of a week for January, May, and August.

Visiting by Time of the Day

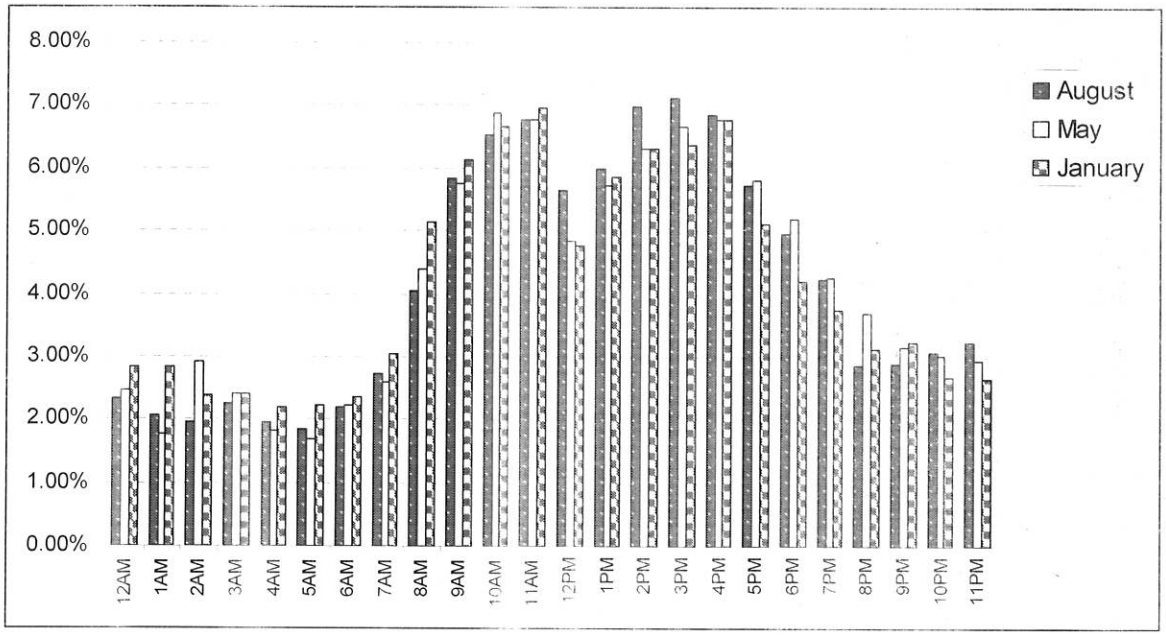


Figure 5.5-4: Number of visitors per time of a day for the three months.

As it is well depicted in the above figures, highest number of visitors have been recorded on Tuesday, Wednesday, and Thursday; and higher number of user also recorded on Friday and

Saturday relative to Sunday and Monday. From week days (working days), least number of users has been recorded on Monday.

The time is shown in full time label with 12-hours format. The log entries were recorded based on a -3:00 time zone; hence it has been modified into +3:00 time zone for this graph as Ethiopia is located in +3:00. In addition, when we mean 9:00AM, it represents the 60 minutes from 9:00 to 9:59AM.

Regarding the users' distribution per hours of a day, highest number of users has been recorded from 10:00AM to 11:00AM in the morning and from 2:00PM to 4:00PM in the afternoon. The second higher number of visitors has been recorded 9:00AM in the morning and 1:00PM and 5:00PM in the afternoon.

5.5.1.6. Common Errors Encountered

While people try to explore the Web resources they may come across error pages. The top errors have been registered for the `/search/` directory in all the three months log data. The `/search/` directory error shares 22.38%, 10.77%, and 14.33% of the total failed requested recorded in the month of January, May, and August, respectively.

This could be an indication that the stated directory has no any page to access or there may be some kind of naming error.

Month	Directory	Error Occurrences	Percentage
January	/search/	778	22.38%
	/applyonline/reginfo/textstyle	547	15.73%
	/textstyle	186	5.35%
May	/search/	609	10.77%
	/applyonline/reginfo/textstyle	448	7.92%
	/faculties/sc/index.htm	432	7.64%
August	/search/	627	14.33%
	/applyonline/reginfo/textstyle	504	11.52%
	/textstyle	189	4.32%

Table 5.5-3: Directories with common errors encountered.

5.5.2. Mining for Association Rules

The association mining experiment has been conducted using WEKA version 3-4-11, with the *Apriori* algorithm for association rule generation.

The ARFF file format that has been prepared during the preprocessing phase has been used for the experiment. Then, the experiment has been conducted for each of the three months data separately.

Apriori works with categorical values only and that is the reason, among others, why 1 and 0 are used in this experiment.

In the default parameter setting, the experiment result may show the following line together with other run information:[4]

"weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 "

And some of the parameters are the following:

- *numRules* (N): Number of rules to find.

- *metricType (T 0)*: Set the type of metric by which to rank rules. *Confidence* is the proportion of the examples covered by the premise that are also covered by the consequence.

  Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support.

[4] Source: WEKA's Context help for Apriori algorithm

- *minMetric* (C): Minimum metric score. Consider only rules with scores higher than this value.

- *delta* (D) : Iteratively decrease support by this factor. Reduces support until min support is reached or required number of rules has been generated.

- *upperBoundMinSupport* (U): Upper bound for minimum support. Start iteratively decreasing minimum support from this value.

- *lowerBoundMinSupport* (M): Lower bound for minimum support.

From all unique URLs, herein called items, only few of them have been selected as they fulfill the minimum support threshold of 3% out of the total transactions. The rest have been excluded during preparing the ARFF format dataset for either of the following reasons:

- if transaction might have contained only one URL, then such transactions are rejected as a singleton transaction are considered uninteresting for association rule mining;

  OR

- the transaction might have contained URLs that could not fulfill the minimum support threshold, i.e. they could not being among the large itemsets.

Moreover, it is difficult for readability to use the URLs for column name; hence they have been labeled as URL1, URL2....URLn, where *n* is the number of selected URLs.

5.5.2.1. Association Rules for *January* Dataset

For the January, 2007, AAU Web access log data, 22314 user sessions (i.e. distinct Transactions) have been identified and of which 2423 sessions have been used in the association mining experiment.

From 1429 unique URLs, 18 items have been selected as they fulfill the minimum support threshold of 3% out of the total transactions and the following are the labels for the selected URLs:

URL1 = */webmail.htm*

URL2 = */academics/index.php*

URL3 = */faculties/index.php*

URL4 = */vacancies/main.htm*

URL5 = */vacancies/index.htm*

URL6 = */about/index.php*

URL7 = */vacancies/top.htm*

URL8 = */administration/index.php*

URL9 = */vacancies/specialization.htm*

URL10 = */webnews/showarticle.php*

URL11 = */libraries/index.htm*

URL12 = */index.php*

URL13 = */news/index.php*

URL14 = */aaulib.php*

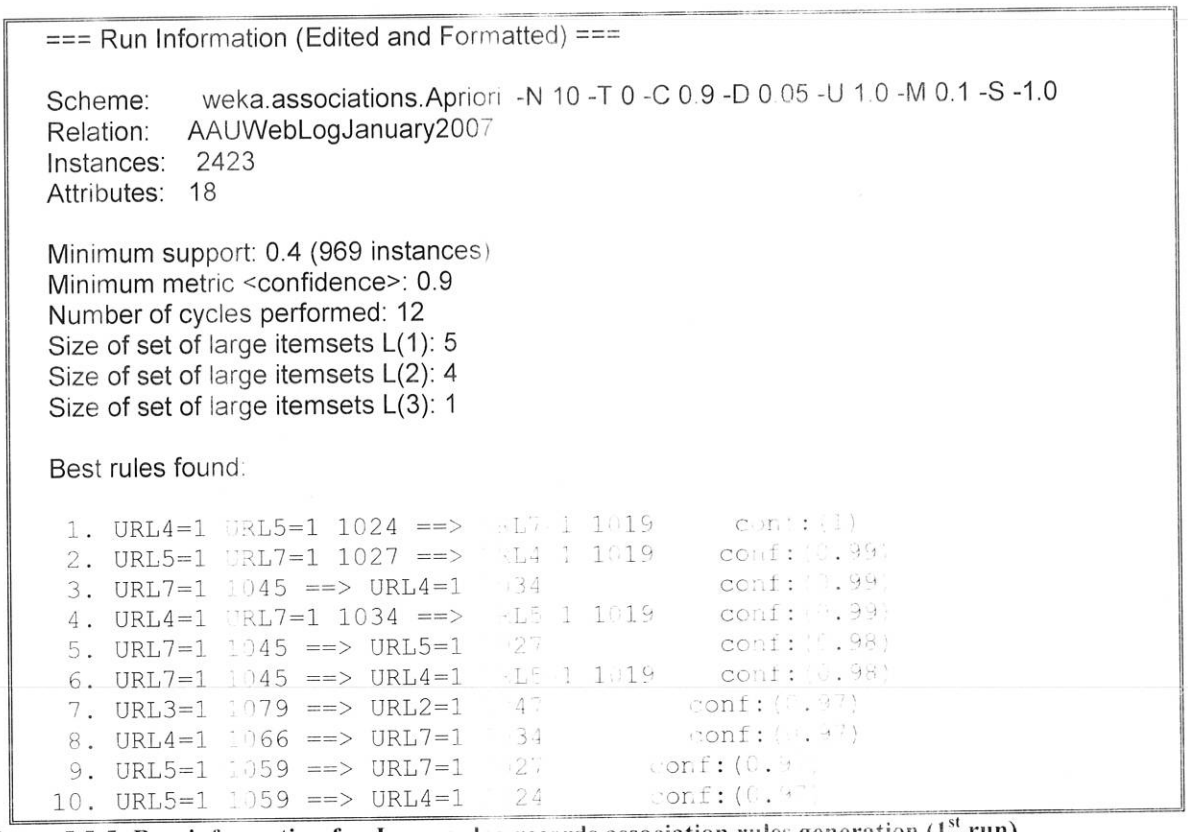
URL15 = */ict/index.php*

URL16 = */applyonline/index.php*

URL17 = */registrar/admission.php*

URL18 = */alumni/index.php*

After feeding the data and running WEKA with the default parameter setting, the following run information has been gained:



```
=== Run Information (Edited and Formatted) ===

Scheme:     weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:   AAUWebLogJanuary2007
Instances:  2423
Attributes: 18

Minimum support: 0.4 (969 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 4
Size of set of large itemsets L(3): 1

Best rules found:

 1. URL4=1 URL5=1 1024 ==>  URL7=1 1019     conf:(1)
 2. URL5=1 URL7=1 1027 ==>  URL4=1 1019     conf:(0.99)
 3. URL7=1 1045 ==> URL4=1  1034            conf:(0.99)
 4. URL4=1 URL7=1 1034 ==>  URL5=1 1019     conf:(0.99)
 5. URL7=1 1045 ==> URL5=1  1027            conf:(0.98)
 6. URL7=1 1045 ==> URL4=1  URL5=1 1019     conf:(0.98)
 7. URL3=1 1079 ==> URL2=1  47          conf:(0.97)
 8. URL4=1 1066 ==> URL7=1  34          conf:(0.94)
 9. URL5=1 1059 ==> URL7=1  27        conf:(0.9
10. URL5=1 1059 ==> URL4=1  24        conf:(0.9
```

Figure 5.5-5: Run information for January log records association rules generation (1st run)

Interpreting the rules:

- 100% of the users who requested */vacancies/main.htm* and */vacancies/index.htm* also requested */vacancies/top.htm*.
- 99% of the users who requested */vacancies/index.htm* and */vacancies/top.htm* also requested */vacancies/main.htm*.
- 99% of the users who requested */vacancies/top.htm* also requested */vacancies/main.htm*.
- 99% of the users who requested */vacancies/main.htm* and */vacancies/top.htm* also requested */vacancies/index.htm*.
- 98% of the users who requested */vacancies/top.htm* also requested */vacancies/index.htm*.

- 98% of the users who requested */vacancies/top.htm* also requested */vacancies/main.htm*.
- 97% of the users who requested */faculties/index.php* also requested */academics/index.php*.
- 97% of the users who requested */vacancies/main.htm* also requested */vacancies/top.htm*.
- 97% of the users who requested */vacancies/index.htm* also requested */vacancies/top.htm*.
- 97% of the users who requested */vacancies/index.htm* also requested */vacancies/main.htm*.

As it shown above, items under the */vacancies/* directory dominate the best rule output. But, the association among these items may not be of interest for the one who look for association between/among items in different directories. This means not that the rule is completely uninteresting because it shows, at least, that the pages in this directory have been visited together much more times than pages in other directories.

The writer assumed that it is also interesting and useful to know the association among items in different directories. To get the most interesting rules, the algorithm has been run again with a modification i.e. three of the items under the */vacancies/* directory has been excluded except */vacancies/index.htm* from the second experiment. However, only five best rules have been generated; hence the experiment has been run for the third time by setting the minimum support threshold to 0.05 to get the number of rules closer to the expected number of rules i.e. 10.

With the third run, the below output has been generated:

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

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.05 -S -1.0
Relation:     AAUWebLogJanuary2007- weka.filters.unsupervised.attribute.Remove-R4,7,9
Instances:    2423
Attributes:   15

Apriori
=======
Minimum support: 0.05 (121 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 19

Generated sets of large itemsets:
Size of set of large itemsets L(1): 14
Size of set of large itemsets L(2): 66
Size of set of large itemsets L(3): 24
Size of set of large itemsets L(4): 3

Best rules found:
1. URL3=1 URL6=1 URL8=1 191 ==> URL2=1 189    conf:(0.99)
2. URL3=1 URL14=1 235 ==> URL2=1 231    conf:(0.98)
3. URL3=1 URL17=1 232 ==> URL2=1 228    conf:(0.98)
4. URL3=1 URL5=1 276 ==> URL2=1 271    conf:(0.98)
5. URL3=1 URL6=1 URL12=1 145 ==> URL2=1 142    conf:(0.98)
6. URL3=1 URL6=1 URL14=1 129 ==> URL2=1 126    conf:(0.98)
7. URL3=1 URL6=1 407 ==> URL2=1 398    conf:(0.98)
8. URL3=1 URL8=1 287 ==> URL2=1 281    conf:(0.98)
9. URL3=1 URL17=1 282 ==> URL2=1 276    conf:(0.98)
10 URL3=1 1079 ==> URL2=1 1041    conf:(0.96)
```

Figure 5.5-6: Run information for January log records association rules generation (3rd run)

Interpreting the rules:

- 98% of the users who requested */faculties/index.php*, */about/index.php* and */administration/index.php* also requested */academics/index.php*.
- 98% of the users who requested */faculties/index.php* and */aaulib.php* also requested */academics/index.php*.
- 98% of the users who requested */faculties/index.php* and */registrar/admission.php* also requested */academics/index.php*.
- 98% of the users who requested */faculties/index.php* and */vacancies/index.htm* also requested */academics/index.php*.

- 98% of the users who requested */faculties/index.php*, */about/index.htm*, and */index.php* also requested */academics/index.php*.
- 98% of the users who requested */faculties/index.php*, */about/index.php*, and */aau/lib.php* also requested */academics/index.php*.
- 98% of the users who requested faculties/*index.php*, */about/index.php* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* and */administration/index.php* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* and */index.php* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* also requested */academics/index.php*.

From the given rules, the last rule seems less interesting. In the last rule, the algorithm associates the items found necessarily one after the other in the AAU Web site's link structure i.e., one must first go to the */academics/index.php* to get the link to */faculties/index.php*. Therefore, we may conclude that their association may not be interesting. But, for instance, if we consider the last rule alone, it prevents us from reaching that conclusion because the rule tells us that 3% of the transactions that contain "*/faculties/index.php*" do not contain "*/academics/index.php*". In other words, 32 users accessed "*/faculties/index.php*" without "*/academics/index.php*". This is an indication that either there may be some direct link to "*/faculties/index.php*" or users were able to reach this page from search engines.

5.5.2.2. Association Rules for *May* Dataset

For the May, 2007, AAU Web access log data, 23037 user sessions (i.e. distinct Transactions) have been identified and of which 2161 sessions have been used in the association mining experiment.

From 2226 unique URLs, 18 items have been selected as they fulfill the minimum support threshold of 3%, out of the total transactions and the following are the labels for the selected URLs.

URL1 = */webmail.htm*
URL2 = */news/index.php*
URL3 = */academics/index.php*
URL4 = */webnews/showarticle.php*
URL5 = */faculties/index.php*
URL6 = */vacancies/main.htm*
URL7 = */faculties/sc/index.htm*
URL8 = */about/index.php*
URL9 = */vacancies/top.htm*
URL10 = */vacancies/index.htm*
URL11 = */administration/index.php*
URL12 = */vacancies/specialization.htm*
URL13 = */applyonline/index.php*
URL14 = */index.php*
URL15 = */aaulib.php*
URL16 = */webnews/showbgstory.php*
URL17 = */ict/index.php*
URL18 = */registrar/admission.php*

After feeding the data and running WEKA with the default parameter setting, the following run information has been gained:

```
=== Run information (Edited and Formatted) ===
Scheme:  weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:   AAUWebLogMay2007
Instances:  2161
Attributes:  18

Minimum support: 0.35 (756 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 4
Size of set of large itemsets L(3): 1

Best rules found:
 1. URL6=1 URL10=1 803 ==> URL9=1 793    conf:(0.99)
 2. URL9=1 URL10=1 805 ==> URL6=1 793    conf:(0.99)
 3. URL6=1 URL9=1 810 ==> URL10=1 793    conf:(0.98)
 4. URL5=1 929 ==> URL3=1 90     conf:(0.97)
 5. URL10=1 839 ==> URL9=1 8     conf:(0.96)
 6. URL10=1 839 ==> URL6=1 8     conf:(0.96)
 7. URL9=1 850 ==> URL6=1 81     conf:(0.95)
 8. URL9=1 850 ==> URL10=1 8     conf:(0.95)
 9. URL10=1 839 ==> URL6=1 URL9=1 793    conf:(0.95)
10. URL6=1 867 ==> URL9=1 81     conf:(0.93)
```

Figure 5.5-7: Run information for May log records association rules generation (1st run)

Interpreting the rules:

- 99% of the users who requested /vacancies/main.htm and /vacancies/index.htm also requested /vacancies/top.htm.
- 99% of the users who requested /vacancies/top.htm and /vacancies/index.htm also requested /vacancies/main.htm.
- 98% of the users who requested /vacancies/main.htm and /vacancies/top.htm also requested /vacancies/index.htm.
- 97% of the users who requested faculties/index.php also requested /academics/index.php.
- 96% of the users who requested /vacancies/index.htm also requested /vacancies/top.htm.
- 96% of the users who requested /vacancies/index.htm also requested/vacancies/main.htm.
- 95% of the users who requested /vacancies/top.htm also requested /vacancies/main.htm.
- 95% of the users who requested /vacancies/top.htm also requested /vacancies/index.htm.
- 95% of the users who requested /vacancies/main.htm also requested /vacancies/index.htm and /vacancies/top.htm.
- 93% of the users who requested /vacancies/main.htm also requested /vacancies/top.htm.

The dominance of pages under the /vacancies/ directory still persists. Thus, the experiment has been done one more time by excluding those pages except the /index.htm page, and provides the following result.

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

Scheme:     weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:   AAUWebLogMay2007-weka.filters.unsupervised.attribute.Remove-R6,9,12
Instances:  2161
Attributes: 15

Minimum support: 0.1 (216 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Size of set of large itemsets L(1): 13
Size of set of large itemsets L(2): 16
Size of set of large itemsets L(3): 4

Best rules found:
 1. URL5=1 URL7=1 259 ==> URL3=1 25[illegible]     conf:(0.98)
 2. URL5=1 URL8=1 311 ==> URL3=1 30[illegible]     conf:(0.97)
 3. URL5=1 929 ==> URL3=1 900     conf:(0.97)
 4. URL5=1 URL11=1 245 ==> URL3=1 24[illegible]     conf:(0.97)
 5. URL5=1 URL14=1 233 ==> URL3=1 22[illegible]     conf:(0.96)
 6. URL3=1 URL7=1 276 ==> URL5=1 25[illegible]     conf:(0.92)
```

Figure 5.5-8: Run information for May log records association rules generation (2nd run)

Interpreting the rules:

- 98% of the users who requested */faculties/index.php* and */faculties/sc/index.htm* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* and */about/index.php* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* and */administration/index.php* also requested */academics/index.php*.
- 96% of the users who requested */faculties/index.php* and */index.php* also requested */academics/index.php*.

- 92% of the users who requested `academics/index.php` and `/faculties/sc/index.htm` also requested `/faculties/index.php`.

In the second run 10 best rules were expected; however, only 6 rules have been generated. All the rules except 1, 3 and 6 sound interesting. Rule 3 has been generated in this month data also with similar confidence like that of the previous month. This rule can be considered interesting for the argument stated in the previous section. Similarly, rule 1 and 6 could be taken as interesting rules and, for instance, rule 1 could be regarded as an interesting rule as it indicates that whoever requested the `/facilities/index.php` also requested the science faculty's page even if there are other faculty's links under the `/faculties/index.php` page.

5.5.2.3. Association Rules for *August* Dataset

For the August, 2007, AAU Web access log data, from 18640 user sessions, 3348 sessions have been used in the association mining experiment.

From 2089 unique URLs, 14 of them have been selected and the following are the labels for the selected URLs:

URL1 = /webmail.htm

URL2 = /academics/index.php

URL3 = /faculties/index.php

URL4 = /news/index.php

URL5 = /vacancies/main.htm

URL6 =/strategicplanning/viewcomments.php

URL7 = /faculties/sc/index.htm

URL8 = /vacancies/specialization.htm

URL9 = /about/index.php

URL10 = /administration/index.php

URL11 = /webnews/showarticle.php

URL12 = /index.php

URL13 = /registrar/admission.php

URL14 = /aaulib.php

After feeding the data and running WEKA with the default values of the Apriori algorithm, only the first two rules, of the following output, were generated.

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

Scheme:     weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:   AAUWebLogAugust2007
Instances:  3348
Attributes: 14

Minimum support: 0.1 (335 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 6

Best rules found:

 1. URL3=1 1535 ==> URL2=1 14[illegible]    conf:(0.96)
 2. URL8=1 873 ==> URL5=1 833    conf: 0.95
```

Figure 5.5-9: Run information for August log records association rules generation (1st run)

Interpreting the rules:

- 96% of the users who requested *faculties/index.php* also requested */academics/index.php*.
- 95% of the users who requested */vacancies/specialiazation.htm* also requested */vacancies/main.htm*.

To get more options in order to select the interesting rules, the second run has been done by setting the minimum support metric to 0.05 (reduced from 10% to 5%), and then the following run information has been gained:

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

Scheme:   weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.05 -S -1.0
Relation:   AAUWebLogAugust2007
Instances:   3348
Attributes:   14

Minimum support: 0.05 (167 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 19
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 27
Size of set of large itemsets L(3): 10
Size of set of large itemsets L(4): 1

Best rules found:

 1. URL3=1 URL14=1 200 ==> URL2=1 198          conf:(0.99)
 2. URL3=1 URL8=1 210 ==> URL2=1 207           conf:(0.99)
 3. URL3=1 URL5=1 URL8=1 201 ==> URL2=1 198    conf:(0.99)
 4. URL3=1 URL5=1 264 ==> URL2=1 258           conf:(0.98)
 5. URL3=1 URL13=1 276 ==> URL2=1 268     conf:(0.97)
 6. URL3=1 URL10=1 236 ==> URL2=1 229     conf:(0.97)
 7. URL3=1 URL12=1 246 ==> URL2=1 238     conf:(0.97)
 8. URL3=1 1535 ==> URL2=1 1486           conf:(0.96)
 9. URL3=1 URL9=1 332 ==> URL2=1 318      conf:(0.96)
10. URL3=1 URL8=1 210 ==> URL2=1 201      conf:(0.96)
```

Figure 5.5-10: Run information for August log records association rules generation (2nd run)

Interpreting the rules:

- 99% of the users who requested */faculties/index.php* and */aaulib.php* also requested */academics/index.php*.
- 99% of the users who requested */faculties/index.php* and */vacancies/specialization.htm* also requested */academics/index.php*.
- 99% of the users who requested */faculties/index.php* and */vacancies/main.htm /vacancies/specialization.htm* also requested */academics/index.php*.
- 98% of the users who requested */faculties/index.php* and */vacancies/main.htm* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* and */registrar/admission.php* also requested */academics/index.php*.

- 97% of the users who requested */faculties/index.php* and */administration/index.php* also requested */academics/index.php*.
- 97% of the users who requested */faculties/index.php* and */index.php* also requested */academics/index.php*.
- 96% of the users who requested */faculties/index.php* also requested */academics/index.php*.
- 96% of the users who requested */faculties/index.php* and */about/index.php* also requested */academics/index.php*.
- 96% of the users who requested */faculties/index.php* and */vacancies/specialization.htm* also requested */vacancies/main.htm*.

In the August data, all the rules appear interesting. The dominance of the association rules among the pages in the */vacancies/* directory has no more observed. Rather, they have appeared in the association rules with other pages. The last rule is also unexpected because the assumption was that one needs to go to the */academics/index.php* page in order to get the */faculties/index.php* page. But, the rule disproved this assumption because the */faculties/index.php* page appeared in the association rule with other pages in the absence of the */academics/index.php* page.

In general, the most interesting finding so far is that the majority of the AAU Web site visitors did not leave the Web site with out visiting the */academics/index.php* page.

Chapter Six

Conclusions and Recommendations

6.1. Conclusions

Based on the course of actions done in the previous chapters, the following conclusions have been reached.

- The number of users showed a decrement in August. This may be due to various reasons, for example lack of satisfaction on the web content or any other reasons. In addition, the failed requests also increased in May and August. Specially, highest failed requests were recorded in May. This might be having a relationship with decrement of users in August.
- The statistical report shows that, from the week days and weekends, minimum number of users has been recorded on Monday and Sunday, respectively. About half of the users have been recorded on Tuesday, Wednesday, and Thursday.
- The most requested page is the */index.php*, which is the home page of the Web site; however, this page did not play a significant role on the association mining. This may be because of that more than $^1/_3$ of the users left the site from this page. This also indicates that the majority of the users finish their session after seeing only the home page. This, in turn, shows there is something wrong either with content of the page or with the Web site's structure. The */webmail.php* page has been found as the second most requested page. But, this is not because of many people requested the page as this page has been no where in the association mining. One obvious reason for this can be that the Web mail service is provided only for the university's staff and one or more of these staff members have been using the Web mail service effectively (repeatedly). The third

most requested page is also /news/index.php; however it has no role in the association mining. The /academics/index.php page has been the fourth most requested, in fact, it is nearly equal to the third one. This page has been the major actor in the association mining.

- Needless to say something about the highest hit record on the root directory, "/" because this is the reflection of what is stated above. The /vacancies/ directory got the second highest hit record. This has been also manifested during the associating mining process. So, it is possible to conclude that the /vacancies/ directory has been requested by many users in contrast to the repeated requests by few users.

- More than half of the AAU official Web site users enter into the site directly via the home page and more than $^{1}/_{3}$ of the total users also left the page from the home page. If we assume that no any user come to the home page from content pages, then we conclude that about 67% of the users who entered into the home page have left the Web site without visiting the other pages.

- The AAU Web site users repeatedly encountered error while attempting to use the search feature of the Web site. This is most probably either the link has no an object item or any other problem.

- Despite the above fact, the association rule generation provides many interesting associations. Most of the rules revealed that the /academics/index.php page is the popular one, which has been requested together with other pages. It has a high probability to be requested together with many of the other pages in other directories.

- The following group of pages are the most frequently requested pages (frequent itemsets):
    - */about/index.php*, */aaulib.php*, */faculties/index.php*, and */academics/index.php*
    - */index.php*, */about/index.htm*, */faculties/index.php*, and */academics/index.php*
    - */about/index.php*, */administration/index.php*, */faculties/ index.php*, and */academics/index.php*
    - */vacancies/main.htm*, */vacancies/specialization.htm*, */faculties/index.php*, and */academics/index.php*
    - */administration/index.php*, */faculties/index.php*, and */academics/index.php*
    - */faculties/sc/index.htm*, */faculties/index.php*, and */academics/index.php*
    - */registrar/admission.php*, */faculties/index.php*, and */academics/index.php*
    - */vacancies/main.htm*, */vacancies/top.htm*, and */vacancies/index.htm*

It is noted that in *Apriori* algorithm for association mining, if the set of items is a large itemset, then its subsets are also large itemsets provided that the subset has more than one member item.

- In Web usage mining endeavor, particularly while doing association mining, the presence of multiple pages in the same directory affects the rules generation. Similarly, having only one entry point for items in the subsections of the Web page also limits the mining algorithm from generating alternative association from which interesting rules can be generated. This particularly limits those who need to mine the association rules across multiple sections in a given Web sites.

  This research, in general, proves that relying on statistical analysis alone for Web log data analysis is not trustworthy; rather, using it with data mining technology gives more meaningful information about how the Web site is being used.

- Web mining is relatively a young discipline and have not been yet studied in detailed. Thus, modeling the Web usage mining approach is an important task for paving the way for further related researches.

6.2. Recommendations

Based on the conclusions given above, the following recommendations are forwarded:

- Most users come to the Web site directly by either typing the name or from a search engine that displays the home page. This could be a symptom that indicates the Web site has a kind of 'sickness'. The Webmaster therefore should do some kind of assessment on the departmental index pages and he/she make sure that whether those pages contain keyword for indexing by search engines.

- The site maintenance is better scheduled on Monday or Saturday or other week days in the morning from 8:00 AM to 9:00 AM or after/ around 5:00 PM in the afternoon because this time and days are when least visitors are registered.

- The office or the team in charge of maintaining the AAU Web site should periodically check the site for any accessibility problem or broken links. Such kind of problems easily solved by introducing additional features to the Website like an online form for users' feedback or broken link reporting form as many Web masters do.

- The page in the `/vacancies/` directory has shown a better usage pattern; therefore, it is advantageous for the university if it make use of its site for such kind of advertisements.

- As the `/academics/index.php` page shows higher popularity, it is a good idea to create a direct hyperlink, to this directory and to the items in it, from any sections of the Web page for the users' convenience thereby to increase their stay in the Web site and also to increase their visiting frequency. It is also a good practice to make some kind of periodic auditing on the smooth running of the Web site. Every effort should be made to make visitors enjoy their visit and find that the time spent there was worthwhile, and such users will come back some other time.

- It is also recommended that the concerned body that is in charge of the AAU Web site design should create quick links from one to the other for those pages mostly accessed together.

- It would be better if the ICT office of the university prepares standard templates for the Web pages development by individual unit; otherwise the ICT office better to develop

the Web pages centrally to avoid possible inconsistencies and errors. Moreover, using Web content management tool, such as *Joomla*®, is recommended to facilities the Web site management and mitigate request failures as well.

- The Association mining has been done using the *Apriori* algorithm, which needs the log data to be put in transaction form. But, as a future research direction, it is recommended that the development of an algorithm that can mine Web log data following the user sessions with out a need to transaction identification.
- It would be also good if others work on Web usage mining by incorporating sequential pattern generation to get any interesting rule about common navigational sequences.
- It is also a good practice employing both data mining and statistical techniques for Web usage pattern discovery to get better information how a Web site us being used.
- To those who like to do some kind of researches in this area, the following model is recommended. In addition, the model is also recommended for developers who like to produce a Web usage mining system.

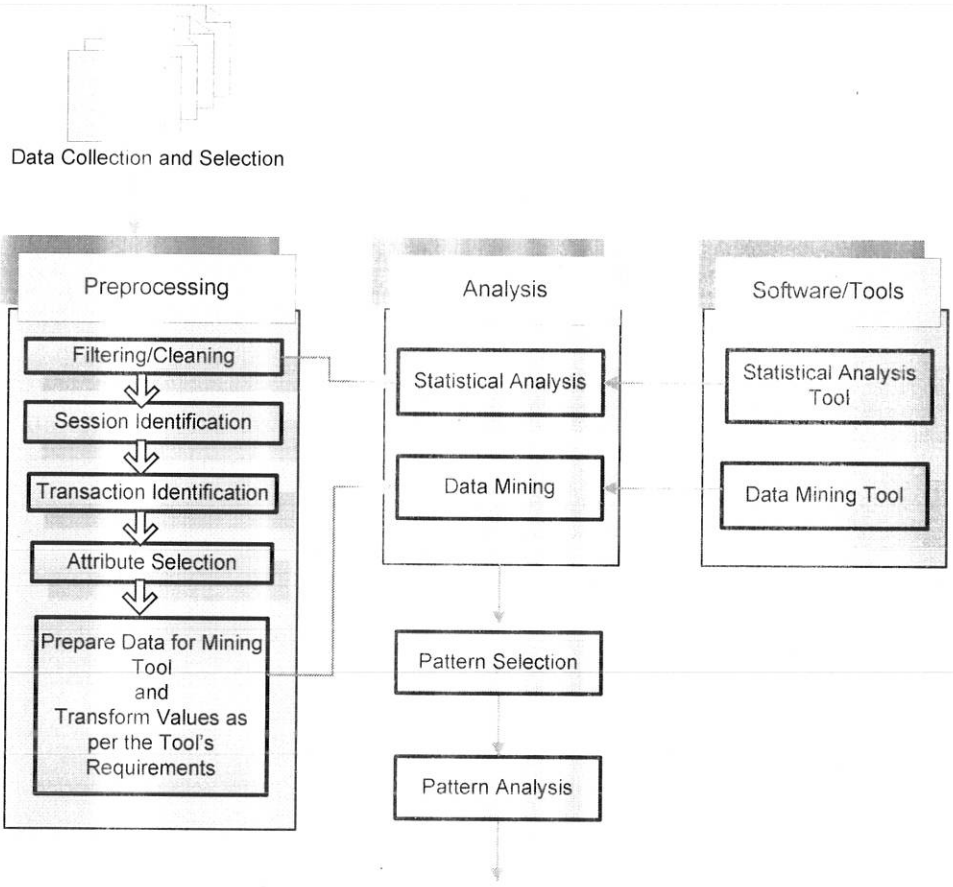
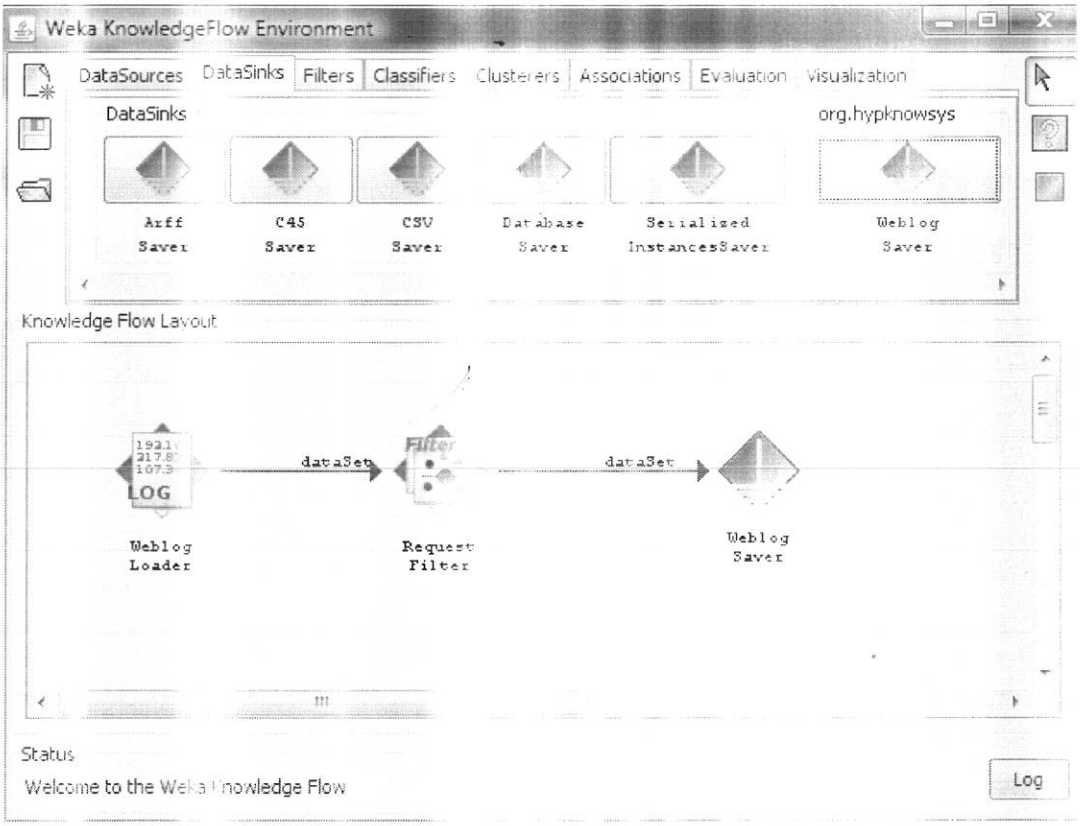


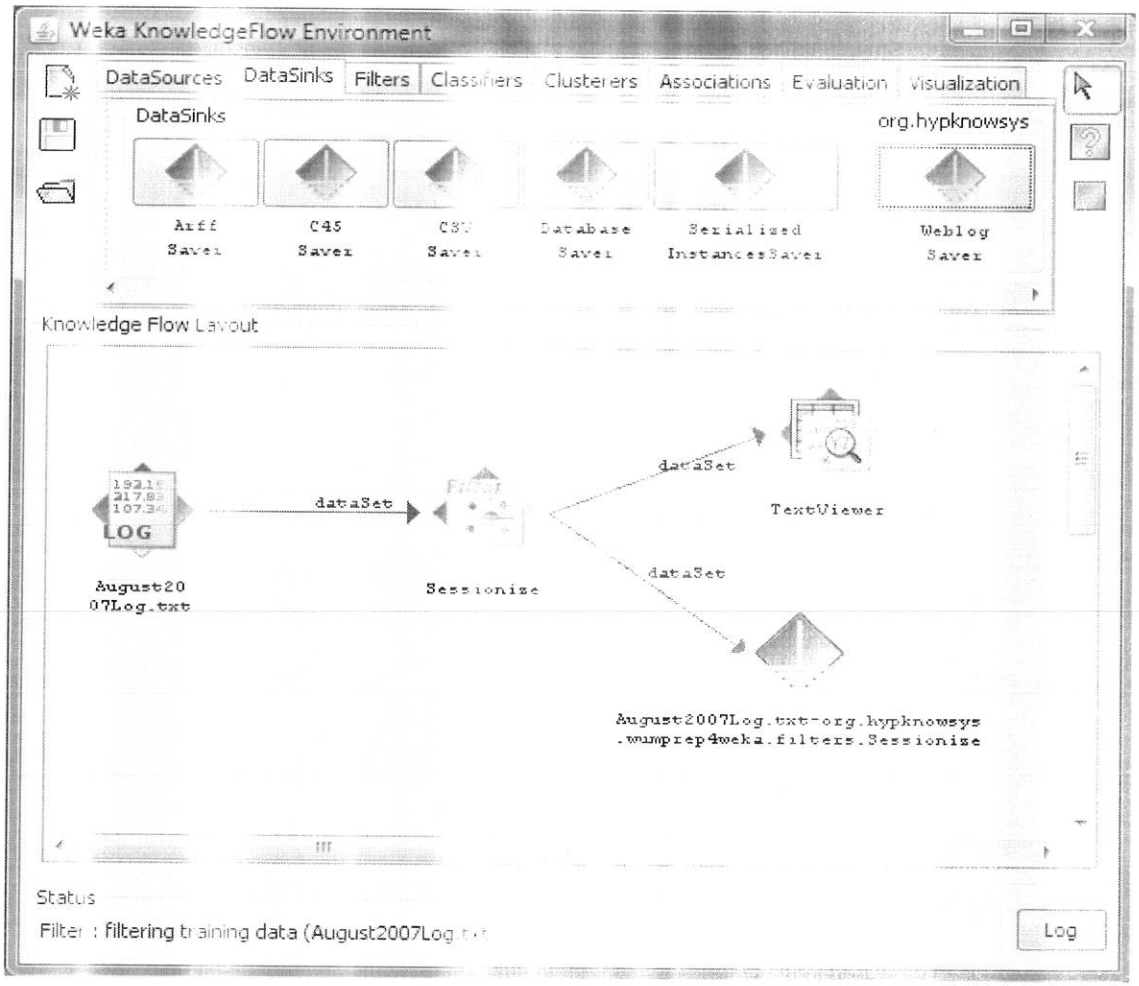
Figure 6.2-1: The recommended model for usage pattern discovery

Appendices

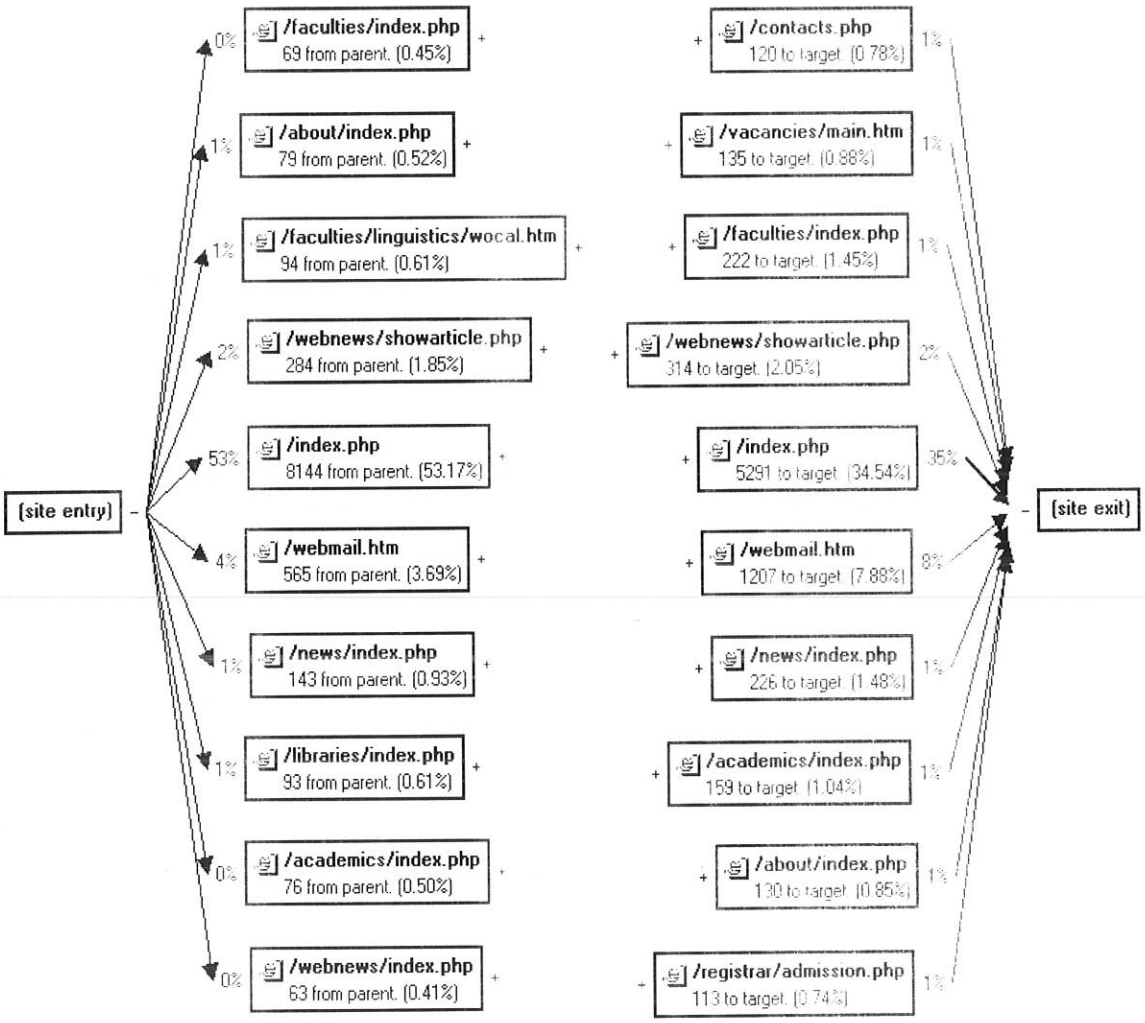
Appendix A: Screen-shot of WEKA Knowledge Flow Diagram for Filtering Requests



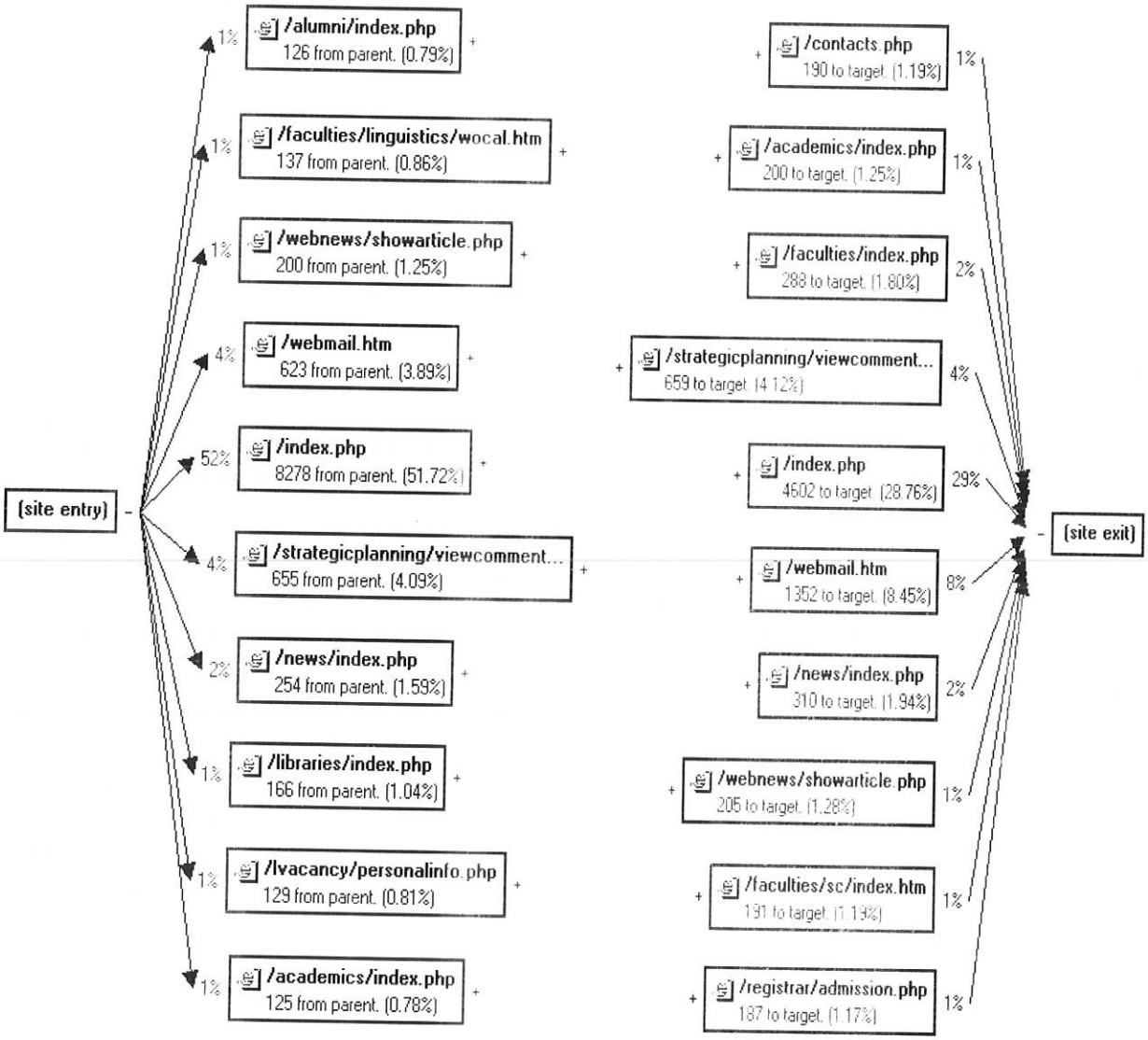
Appendix B: Screen-shot of WEKA Knowledge Flow Diagram for Sessions Identification



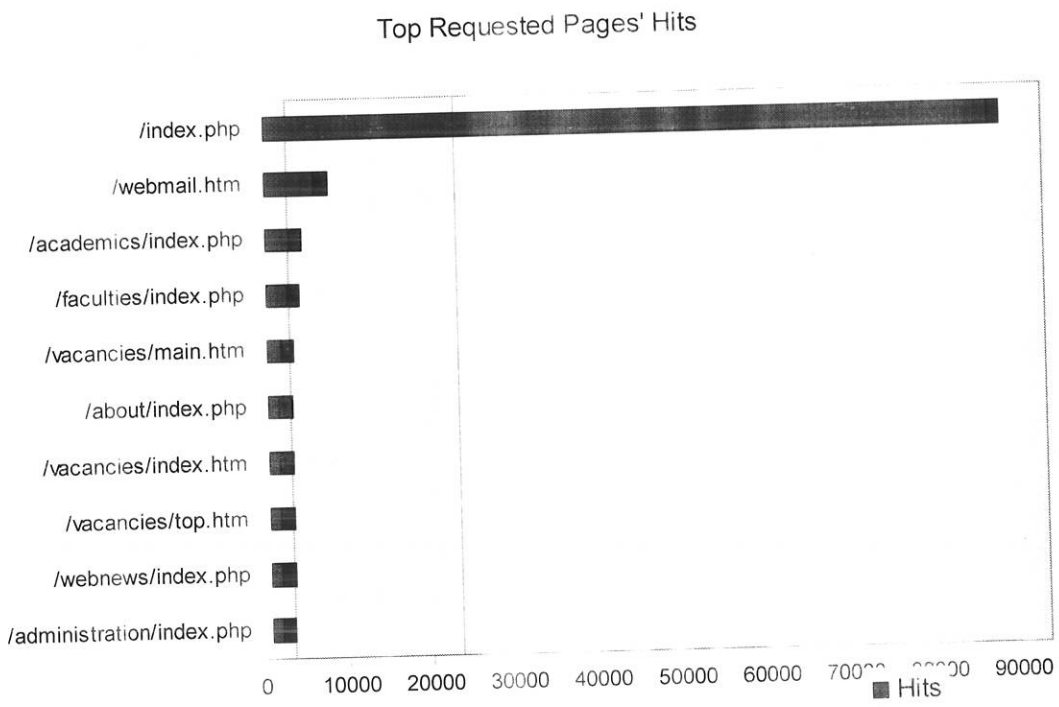
Appendix C: Common Site Entry and Exit Pages for May



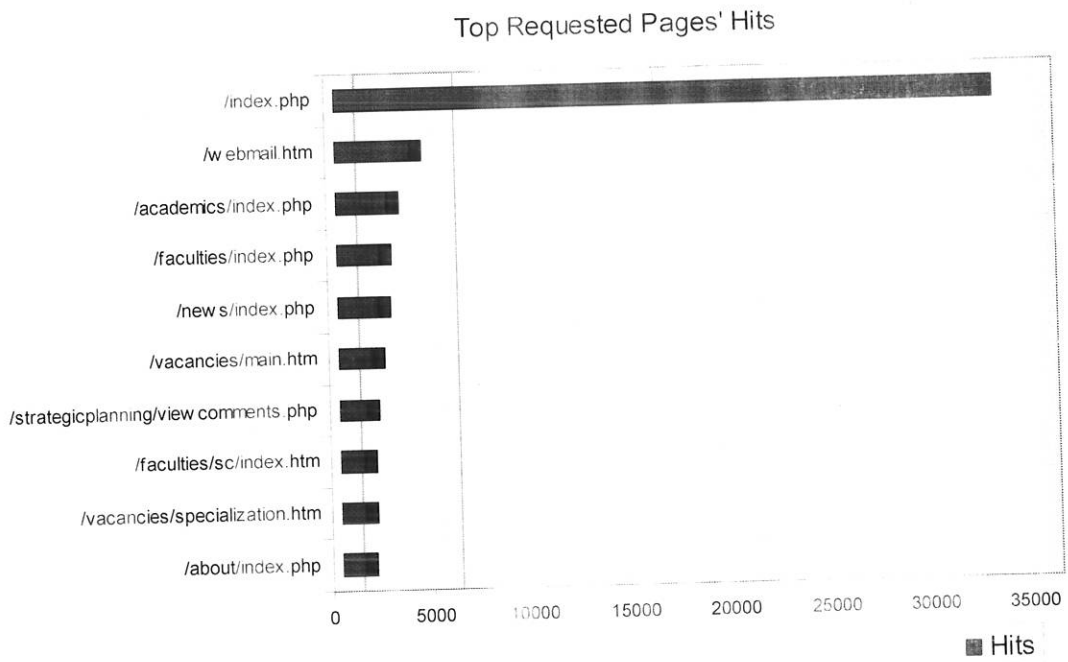
Appendix D: Common Site Entry and Exit Pages for August



Appendix E: Most Accessed Pages for January



Appendix F: Most Accessed Pages for August



References:

1. A. Aschenbrenner and A. Rauber."Minig Web Collections", in J. Masanes (ed.), *Web Archving*, Springer, 2006.

2. A. K. Sinha, "*Data Warehousing*", Thomson/Delmar Learning, India, 2001.

3. A. Singhal, "An Overview of Data Warehousing, OLAP and Data Mining Technology" in *Data Warehousing and Data Mining Techniques for Cyber Security*, Springer, 2007.

4. B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou. "The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis", *Proceedings of the WebKDD 2002 Workshop*. Edmonton, Alberta, Canada, July 2002.

5. B. L Barrett. "Simpletons Guide to Web Server Analysis", 2008. *Web article at* http://www.mrunix.net accessed on 29th September 2008.

6. B. Lydon and T. Fennell, "Web Usability: Its Impact on Human Factor and Consumer Search Behaviour", *Human-Computer Interaction: Theory and Practice (Part I)*, Vol.1. PP 793- 797, 2003.

7. C. Ratheke and V. Schreiweis. "Interaction Design Elements to Improve Information Presentation on Web Pages", *Human-Computer Interaction: Theory and Practice (Part I)*, Vol.1, PP 843-846, 2003.

8. D. He and A. Goker, "Detecting Session Boundaries from Web User Logs", *In Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, 2000.

9. H. Dai and B. Mobasher, "Integrating Semantic Knowledge with Web Usage Mining for Personalization", in A. Scime (ed.). *Web Mining: Application and Techniques*, Idea Group Publishing, 2005.

10. H. Hand, H. Mannila and P. Smyth. "Principle of Data Mining", MIT press, Massachusetts. 2001.

11. http://httpd.apache.org/docs/1.3/ accessed on 10th October 2008.

12. http://www.domaintools.com accessed on 15th June 2008

13. http://www.historyoftheinternet.com accessed on 31st March 2008.

14. *http://www.internetworldstats.com accessed on 31st March 2008.*

15. http://www.usability.gov/ - accessed on 16th June 2008

16. http://www.w3.org/pub/WWW/TR/WD-logfile-960221.html accessed on 10th October 2008

17. J. Griffin, "Data Mart vs. Data Warehouse: Information Strategy", *DM Review Magazine*, 1998, accessed from http://www.dmreview.com/issues/19980201/815-1.html on 16th June 2008

18. J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed. Morgan Kaufmann Publishers, 2006.

19. J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, 2001.

20. J. Jackson, "Web Technologies: a Computer Science Perspective", Pearson Prentice Hall, New Jersey, USA, 2007.

21. J.R Punin, M.S Krishnamoorthy and M.J Zaki, "Web Usage Mining: Languages and Algorithms", *Rensselaer Polytechnic Institute*, 2001.

22. J. Srivastava. R. Cooley, M. Deshpande, and P. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations, Vol. 1, Issue 2*, 2000.

10. H. Hand, H. Mannila and P. Smyth. "Principle of Data Mining", MIT press, Massachusetts. 2001.

11. http://httpd.apache.org/docs/1.3/ accessed on 10th October 2008.

12. http://www.domaintools.com accessed on 15th June 2008

13. http://www.historyoftheinternet.com accessed on 31st March 2008.

14. *http://www.internetworldstats.com accessed on 31st March 2008.*

15. http://www.usability.gov/ - accessed on 16th June 2008

16. http://www.w3.org/pub/WWW/TR/WD-logfile-960221.html accessed on 10th October 2008

17. J. Griffin, "Data Mart vs. Data Warehouse: Information Strategy", *DM Review Magazine*, 1998, accessed from http://www.dmreview.com/issues/19980201/815-1.html on 16th June 2008

18. J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nded. Morgan Kaufmann Publishers, 2006.

19. J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, 2001.

20. J. Jackson, "Web Technologies: a Computer Science Perspective", Pearson Prentice Hall, New Jersey, USA, 2007.

21. J.R Punin, M.S Krishnamoorthy and M.J Zaki, "Web Usage Mining: Languages and Algorithms", *Rensselaer Polytechnic Institute*, 2001.

22. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations, Vol. 1, Issue 2*, 2000.

23. L. D Catledge and J.E Pitkow, "Characterizing Browsing Strategies", *The Third International WWW Conference*, 1995.

24. L. Shen, L. Chneg, J. Ford, F. Makedon, V. Megalooikonomou and T. Steinberg, "Mining the Most Interesting Web Access Associations", *WebNet 2000-World Conference* , 1999.

25. M. Arlitt, "Characterizing Web User Sessions", International mobile Systems Laboratory, Hewlett Packard, 2000.

26. M. Baglioni, U. Ferrara, A. Romei, S. Ruggieri, F. Turini and F. Buonarroti, "Preprocessing and Mining Web Log Data for Web Personalization", *8th Italian Conf. on Artificial Intelligence* vol. 2829 of LNCS, 2003.

27. N. Khasawneh and C. Chan, "Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining", *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings (WI'06))*, 2006.

28. O.R Zaine, M. Xin, and J. Han. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", *Virtual-U Research Laboratory and Intelligent Systems Research Laboratory*, Simon Fraser University, Canada, 1998

29. P. Batista and J.M Silva, "Mining Web Access Logs of an On-line Newspaper", *12th International Meeting of the Euro Working Group on Decision Support Systems,* 2002.

30. R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", *Journal of Knowledge and Information Systems Vol* 1, no 1, 1999.

31. R. Iváncsy and I. Vajk, "Frequent Pattern Mining in Web Log Data", *Acta Polytechnica Hungarica Vol. 3, No. 1*,2006.

32. R. Kosala and H. Blockeel, "Web Mining Research: A Survey", *SIGKDD Exporations, Vol 2, Issue 1*, 2000.

33. R. Zilse and A. Moraes "An Ergonomics Analysis of the Information Architecture of Websies: Developers vs. Users: a Case Study of Brazilian websites" *Human-Computer Interaction: Theory and Practice* (Part I), Vol.1, pages 878-882, 2003.

34. Y. Fu and M. Shih, "A Framework for Personal Web Usage Mining", *International Conference on Internet Computing (IC'2002)*, Las Vegas, NV, pages 595-600, 2002.

35. Z. Chen, et al., "Efficient Web Mining for Traversal Path Patterns", in A. Scime (ed.), *Web Mining: Application and Techniques*, Idea Group Publishing, 2005.

36. Z. Like, K. Zhongbao and Z. Changshui, "Session Identification Based on Time Interval in Web Log Mining", in Z. Shi and Q. He (ed.), *Intelligent Information Processing II*, Springer-Verlag, London, UK, 2004.

Declaration

This thesis is my original work, has not been presented for a degree in any other university and that all sources of material used for the thesis have been duly acknowledged.

Mekonnen Tsegaye

12 January, 2009

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

Dr. Manoj V.N.V (Advisor)

12 January, 2009