Application of Web Usage Mining on Proxy Server Access Log File to Investigate Internet Users' Access Behaviors in ASTU

Enatnesh Minale

A Thesis submitted to the Department of Computer Science and Engineering,

School of Electrical Engineering and Computing

Presented in partial fulfillment of the requirement for the Degree of Master's in

Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

ADAMA

FEB, 2019

Application of Web Usage Mining on Proxy Server Access Log File to Investigate Internet Users' Access Behaviors in ASTU

Enatnesh Minale

Dr. N. Satheesh Kumar

A Thesis submitted to the Department of Computer Science and Engineering,

School of Electrical Engineering and Computing

Presented in partial fulfillment of the requirement for the Degree of Master's in

Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

ADAMA

FEB, 2019

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by _____ have read and evaluated his/her thesis entitled ***"Application of Web Usage Mining on Proxy Server Access Log File to Investigate Internet Users' Access Behaviors in ASTU"*** andexamined the candidate. This is, therefore, to certify that the thesis has been accepted in partialfulfillment of the requirement of the Degree of Masters of Science in Computer Science and Engineering.

_____	_____	_____
Advisor	***Signature***	***Date***
_____	_____	_____
Chair Person	***Signature***	***Date***
_____	_____	_____
Internal Examiner	***Signature***	***Date***
_____	_____	_____
External Examiner	***Signature***	***Date***

Declaration

I hereby declare that this MSc Thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name: Enatnesh Minale

Signature: _____

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

Name: Dr. N. Satheesh Kumar

Signature: _____

Date of submission: Feb 21, 2019

Acknowledgments

First of all, I would like to thanks my **God** for his omnipresent support. God I thank you so much. My sincere gratitude should go to my advisor ***Dr. N. Satheesh Kumar*** for his constructive comments and guidance.

My special thanks should go to ***Mr. Gizachew Belayneh*** for his love, encouragement, comment and kindness to help me in any way.

I am very happy to thank ***Mr. Kifle Gebre*** and ***Miss Tsion Eshetu*** for their material support. I would like to thanks all ASTU ICT Staff for their cooperation concerning data especially Mr. Addis.

Last, but most important, thanks go to my family; especially to my Mother ***Dege Melaku*** for her love and encouragement without which I would have not been who I am today.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

ASTU =	Adama Science and Technology University
ARFF =	Attribute-Relation File Format
BDU=	BahirDar University
CLF =	Common Log Format
CRM =	Customer Relation Management
CSV =	Comma Separated Values
DNS=	Domain Name System
ECLS =	Extended Common Log Format
FPG =	Frequent Pattern Growth
FTP =	File Transfer Protocol
GB=	Giga byte
HTTP=	Hyper Text Transfer Protocol
ICT =	Information communication technology
IP =	Internet Protocol
ISP=	Internet Service Provider
JU=	Jimma University
KDD=	Knowledge Discovery of Databases
MOST=	Minister of Science and Technology
NCTTE =	Nazareth College of Technical Teacher Education
NTC =	Nazareth Technical College
UOG=	University Of Gondar
URL =	Uniform Resource Location
WEKA=	Waikato Environment for Knowledge Analysis
WUM =	Web Usage Mining
WWW =	World Wide Web

Abstract

Nowadays Each and every activities have been becoming unimaginable to do without internet. When an internet user request a particular page on web, an entry is logged into a special file called server log file. Some organization use a proxy server for caching services and administrative control. Proxy server is a server that sits between client computer and internet, and provide indirect network services to client. Since proxy server is in between client and web server, every request from clients will pass through the proxy server to corresponding web server. Analyzing proxy log file has numerous advantages for access behavior investigation process. The main objective of the study is investigating internet users' access behavior by analyzing proxy log file of ASTU Proxy Server. By keeping this in mind, the study was done by following the three web usage mining process phases such as *data preprocessing*, *pattern discovery,* and *pattern analysis*. In Pattern discovery phase, data mining methods such as association rule mining and statistical analysis has been used with the intension of discovering patterns. Apriori algorithm of WEKA data mining tool has been used for association rule mining. The study tried to investigate users' access behavior based on two time situations such as ***class time*** and ***exam time***. By considering this two time situations, the study answered the research questions. This study will be an input for creating a platform that changes the university internet usage trend in the way to be essential for effective teaching–learning process. In addition, the study will show how analyzing web usage mining on proxy server access log data is essential and motivate other researchers.

Chapter One: Introduction

1.1. Overview of Adama science and Technology University

Adama Science and Technology University (ASTU) were first established in 1993 as Nazareth Technical College (NTC), offering degree and diploma level education in technology fields [36]. Later, in 2003 the institution was renamed as Nazareth College of Technical Teacher Education (NCTTE), a self-explanatory label that describes what the institution used to train back then candidates who would become technical teachers for TVET colleges/Schools across the country. In 2003, a new addition to NCTTE came about introduction of business education [36].Following its inauguration in May 2006 as Adama University, the full-fledged university started opening other academic programs in other areas, an extension to its original mission [36].However, it was not until it was nominated by the Ministry of Education as Center of Excellence in Technology in 2008 that it opened various programs in applied engineering and technology. For its realization, it became a university modeled after the German paradigm. The same period saw pioneering of the university in introducing PhD by Research and MA/MSC by Research programs. Following its renaming by the Council of Ministers as Adama Science and Technology University in May 2011, the university has started working towards the attainment of becoming a center of excellence in science and technology, thereby allowing for the realization of goals setin the Growth and Transformation Plan (GTP).

To this end, a South Korean has been appointed as President of the University. Currently, ASTU is setting up a Research Park, in collaboration with stakeholders and other concerned bodies: one of a kind in the Ethiopian context. The university is also venturing out to the wider community and is currently engaged in various joint undertakings [36]. Even though the official web site of ASTU was modified depending of the technology advancement the first time established in 2007[36].Starting from 2015 ASTU’s reach was limited to its only campus in Adama town which is concerned only with Science and Technology beside Addis Ababa Science and Technology University under Minister of Technology. Those two Universities are concerned only on producing

educated manpower in science and technology discipline through selecting best performer students in preparatory all over the country and using entrance exam [36]. Other schools except School of Engineering and Natural Science already moved to a new university, called Arsi University established in Asella town. The university has two divisions: Liberal arts& social science and division of freshman program and five schools: School of Electrical engineering & Computing, School of Civil engineering & Architecture, School of Mechanical, Chemical & Materials engineering, School of Applied Natural Science and Graduate school of Technology Management.

1.1.1. Information & Communication Technology Unit in ASTU

Information & Communication Technology has been established as a unit at the beginning September 2008 with the objectives of building state-of-the-art ICT infrastructure and mainstreaming the ICT-based services in the teaching, learning, research, consultancy and administrative activities of the university. The unit also aims at ensuring sustainable use of the technology by providing hardware, software and network maintenance services. To accomplish the above stated tasks the unit has been structured into five major sub-units: Information technology infrastructure team, Business Application Development Team, Technology Teaching Learning Unit, User support and maintenance Team and training and Consultancy Team. The five sub-units are organized at the university level providing support services to all campuses. The ICT center belongs to vice president of institutional development and business. The Teams provide their own functions for the university [36].

Unit Name	Their respective services
Information technology infrastructure	Focus on the overall network and system installation, configuration and maintenance.
Business Application Development	It Mainly works on application development or software development and software maintenance.
Technology Teaching Learning	working on e learning, website administration and development, content development and video conferencing
User support and maintenance	works on supporting all admin and academic staff on maintaining all ICT related devices and facilities
Training and Consultancy	facilitating training for ICT staff and organizing training for other staff from ICT like Cisco

Table1.1 ICT units and their respective functions [62].

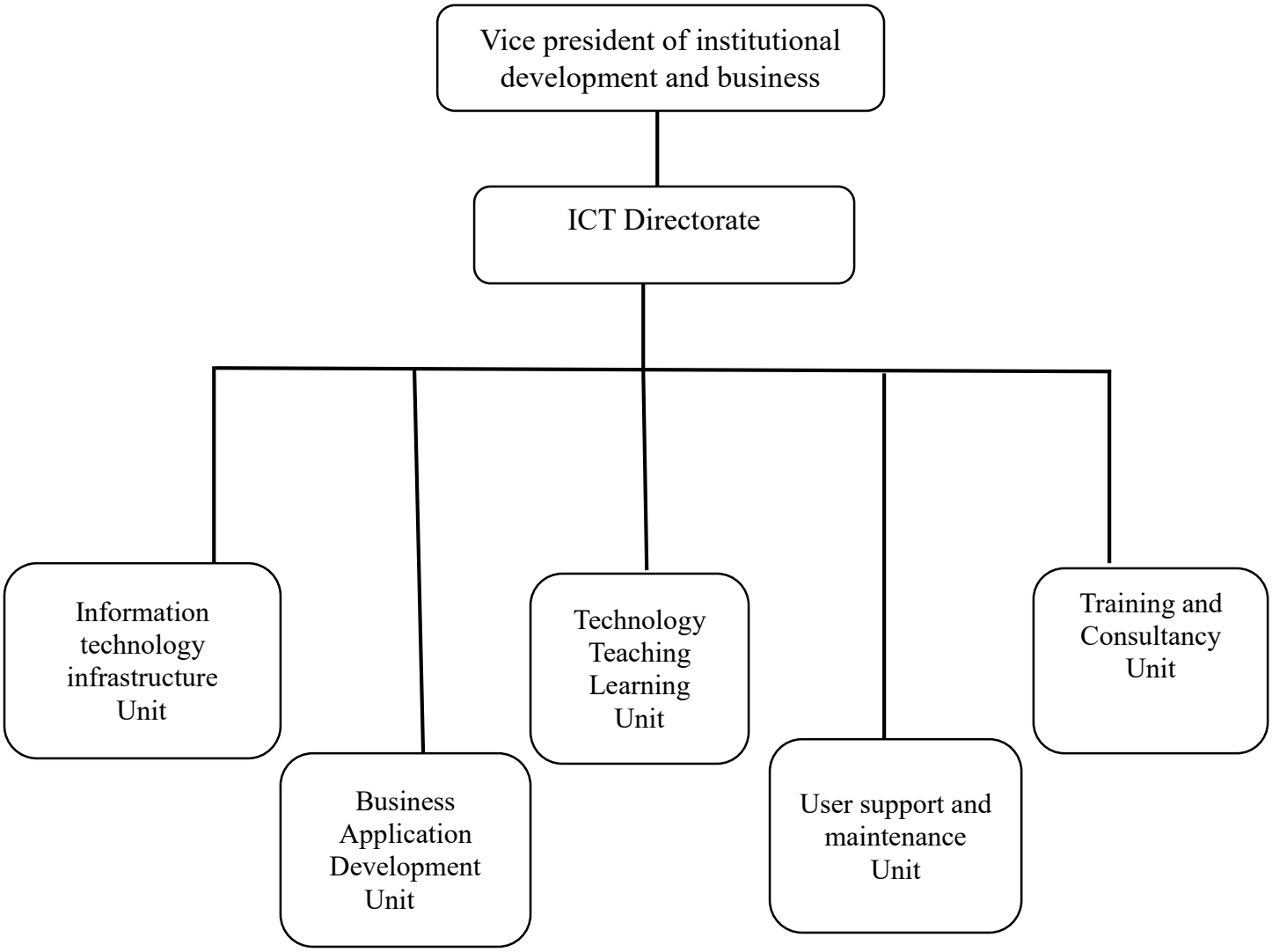

Figure 1.1. ICT center Structure

1.1.2 ASTU and its network infrastructures

The ICT unit, in accordance with the vision and Mission of the university is established mainly to support the entire University communities which include Students, Academic and Administrative staffs and External Users of the University with ICT based solutions. The ICT unit so far has done several remarkable activities inside and outside of the university, such as network expansion and upgrading in its campus, different application software development and customization using open source and commercial software, and continues user support [36].

The University Internet service is increasing progressively with the increase in number of users and the program expansions of the university. The ICT in consultation with the management of the University in this regard has done several remarkable activities to expand and upgrade Information Technology Infrastructure and services. Currently the university buildings are interconnected using Fiber Optic Cable. The center provides fast broadband internet services with capacity of 400MBps. Furthermore, campus wide huge wireless internet connection is deployed which has 32 access points (with its wireless controller) which are mounted in the university mainly focusing on student's dormitory, class rooms, Staff offices, Staff residents and library. More than 15 computer laboratories with the capacity of 22 computers are actively working. These computer laboratories are mainly intended for Internet browsing, to deliver specialized courses which need laboratories and for e-learning services. Ever since its establishment, the ICT unit has been working closely with the university community to create awareness of ICT, giving training for staff members of the university and most importantly to expand network and Internet facilities for the university community. The office is currently established as an ICT Directorate Office level with one Director, four Team Leaders, 15 professionals, and 26 technical support staffs. There are several ways to connect using both wired and wireless connections. The network also provides access to the Internet, support for secure solutions for data integrity, and backup links limit the impact of equipment failure.

The University's network is available 24 hours a day, seven days a week and provides both wired and wireless connections[30].

- Wired: staff offices, some of the class rooms, student ICT pools and Library.
- Wireless: widely available across campus and Staff halls of residence and in some student halls of residence.

Unit	**No of staff per each unit**
Information technology infrastructure	14
Business Application Development	16
Technology Teaching Learning	10
User support and maintenance	28
Training and Consultancy	8
Total Staff in ICT	**76**

Table 1.2 Summarized table of ICT staff

1.1.2. Proxy servers in ASTU

Proxy servers has been used in ASTU since 2008. They started with one proxy server with the IP address on 192.168.0.5 then after two years' time they add the second server with the IP address on 192.168.0.6 currently there are 4 proxy servers used in the university.

According to the interview held with the system administrator Mr. Amanuel, ASTU's proxy server's uses squid proxy server as an application to manage Web caching and prefetching. Squid is a caching proxy for the Web supporting HTTP, HTTPS, FTP, and more. It reduces bandwidth and improves response times by caching and reusing frequently-requested web pages. Squid has extensive access controls and makes a great server accelerator.

1.1.3. SWOT Analysis

Strength

The ICT center strictly store the log file. Therefore, the researcher didn't face any problem in order to get enough data.

Weakness

Even if it stores the log file, it doesn't conduct any analysis for further users' access behavior investigation.

Opportunities

The ICT center is near for the researcher, this provides an opportunity to get further explanations and advice easily and early concerning the center and data.

Threats

The nature of network installation was not depending on the type of users in the campus (admin staff, academic staff, and students). This leads the researcher to not identify from the three users' perspectives.

1.2. Background of the study

As we all know, nowadays each and every activities have been becoming unimaginable to do without internet. World Wide Web is increasing rapidly and huge amount of information is generated due to users' interactions with websites.

Every user spends their most of the time on the internet and their behavior is different from one another [31]. It serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services [34]. It contains a rich and dynamic collection of information about web page contents with hypertext structures and multimedia, hyperlink information, and access and usage information, providing fertile sources for data mining [35]. Web mining is an application use of data mining techniques to discover and extract information from web documents and services. It is the extraction of interesting, useful knowledge and hidden information from activities related to the WWW.

Web Mining can be classified with respect to data it uses. They are web structure mining, web content mining and web usage mining. Web content mining is one of the category of web mining that used to extract useful information from the content of the web document. Web structure mining is the process of discovering pattern from hyperlinks in a web. Web Usage Mining which is the concern of this study, is the category of web mining that is the application of the various data mining techniques for extracting or discovering interesting usage patterns from large data generated as a result of user interactions with websites. Since web usage mining is extracting of hidden pattern from usage details, it uses log data (because of this web usage mining also called web log mining). There are three log data sources that contains helpful user usage details such as [IP address, date and time, request method, status code, URL, user agent (browser), operating system]; they are web server, proxy server and client browser. Web server is a server that holds the actual web pages. When an internet user request a particular page on web, an entry is logged into a special file called server log file. The recorded log file in web server contains the access of single website by multiple users. Proxy server is a server that sits between client computer and internet, and provide indirect network services to client. Since proxy server is in between client and web server, every request from clients will pass through the proxy server to corresponding web server. The clients' usage details will be recorded in proxy servers as proxy server log; the recorded log file contains the access of multiple websites by multiple

users. Proxy server used to improve a performance of web response by caching web page that is requested by the client and respond from the proxy server when similar web page is requested rather than sending the request to the corresponding web server. It stores internet users' usage details which is the main concern of this study and also used to restrict and grant websites in an organization.[36].

In the context of our country, even if we are late to introduce with internet, the current motivation to take advantage of internet is very promising. But we are still late when we are compared to others. Most of our country internet coverage is in universities, governmental institutions and big private sectors.
Universities have high internet coverage and several skillful users as compared to outside internet users. In order to control the internet users, their behavior should be known. As stated above, proxy server used to grant and restrict websites, to cache web pages for fast response and to store internet users' usage details on multiple websites which is called access log. Analyzing that access log helps to investigate users' access behavior. That's why it is called web usage mining. However, from 43 universities ASTU, JU, BDU, and UOG are the only four universities those understood the functionalities of proxy server and using it partially. Although the four universities are using the proxy server, they don't allow the system to store the access log and they don't conduct analysis on access log for the support of internet usage management policy designing. Rather they used only to grant and restrict some websites as whole in the campus which is unfair for genuine users. Analyzing access log file has a magnificent utilization to understand the nature of the web traffic, to understand the users' reaction and to improve or establish well-structured internet usage management policy that strengthen the effectiveness of teaching-learning process in the university. Especially for commercial websites, this kind of study has unreplaceable role to be productive and profitable. Because knowing the behavior of the customer helps the commercial organizations to make it acceptable their production as per the users need. The researcher has been motivated by thinking that Universities should be the first to aware the outside community about the essentiality of web usage mining on access log by doing this kind of practical study. This study is conducted by taking ASTU proxy server access log data as a case study. [30]

The general objective of the study is to apply web usage mining on ASTU proxy server access log data to investigate internet users' access behavior.

The study was done by following the three web usage mining process phases such as *data preprocessing*, *pattern discovery,* and *pattern analysis*. Data preparation, data cleaning, user identification

and session identification has been performed under preprocessing phase. In Pattern discovery phase, data mining methods such as association rule mining and statistical analysis has beenused with the intension of discovering patterns. Apriorialgorithm of WEKA data mining tool has been used for association rule mining.

The study answered the question like top visited websites, number of request per a day, transferred byte per a day, number of unique user per a day, busy and free traffic hours, website classification, how much time does the user spent on internet etc. It clearly investigated ASTU internet users' access behavior.

This study will have its own contribution in order to tell the trend of internet usage behavior in ASTU and to aware the essentiality of analyzing web server access log file for all internet based business organizations of the country. Also, it can be an input for creating a platform that changes the university internet usage trend in the way to be essential for effective teaching–learning process. In addition, the study will show how analyzing web usage mining on proxy server access log data is essential and motivate other researchers.

1.3. Statement of the problem

The three main goal of universities are teaching, research and community service. From this we can understood that teaching learning is the primary activity in universities. This means every movement in the university should perform towards to strengthen the effectiveness of teaching- learning process. Nowadays, internet is becoming a mandatory to do anything in the world. Education center is a place where is internet used widely. As universities are one of the giant education centers, they need better internet facility. By considering this, in our country many of the universities are trying to establish well-equipped ICT centers. But, almost all ICT centers worrying about fulfilling the hardware and software infrastructure of the internet instead of designing well-structured internet usage management policy that strengthen the effectiveness of teaching-learning process. Because of the limitation on internet usage management policy.

This addiction has negative impact, especially on students to not properly use their time for learning and even it could be the reason to dropout. The other limitation of ICT centers is that, for the above problem they choose restricting usage of some selected websites by using proxy server. This has the disadvantage that a genuine user who needs access to information is also denied.

The proxy server not only used to restrict websites and cache web pages rather it stores the users usage details which is called access log. Access log file is very essential to investigate internet users' access behavior and to improve websites and internet usage management policy. ASTU uses Proxy server to reduce the LAN's shared Internet connection traffic by caching frequently requested web pages by using web

caching. Though ASTU uses a proxy server as a solution to the problems arising from growing network traffic on the limited network infrastructure. They were being able to solve the problem with the infrastructure side but still they are unable to solve the problem fully, the network traffic is still unstable and inconsistence. With this study result they can know where the exact problem resides from the user side.

Since the access log file informs who uses which website, where and when; analyzing this kind of log file has a magnificent utilization to understand the nature of the web traffic, to understand the users' reaction and to improve or establish well-structured internet usage management policy in the university. As obviously known, internet based activities are controlling the world's education, communication, and business movement; and users joining this movement are increasing rapidly. As the users increased, different access behavior will be available as well. And also several technologies that has an impact on users' access behavior will be invented within a year. Because as new customers join the websites, new access behavior might be available.

[30], tried to analyze adama science and technology university's proxy server access log file but, the sample date that the author selected to use for analysis is not appropriate to identify genuine access behavior of internet users' of the campus. He selected the access log from October 01/2015 to October 31/2015 in Gregorian calendar. That means from Meskerem 21/2008 to Tikimit 21/2008 in our calendar. According to the university trend Meskerem 21- Tikimit 5 (local time) is registration time. The class starts after registration but, registration with penalty, re-grading, add and drop courses are processed in this time. By keeping in mind this, majority of students will back to their home after registration and majority of staff will be late to start class by considering the absence of many students. So, it is difficult to get the full participation of the users from Meskerem 21 to Tikimit 21 (local time). This leads us to wrong conclusion. In campus there are two time situation that can vary users' behavior which are class and exam time. They are ***class time*** and ***exam time***. The sampled date by the author does not include these two time situations.

So, this study has conducted to analyze the proxy server performance from user's perspective, understanding the nature of Web traffic; and understanding user reaction and motivation in addition to overcome the limitation of [30] in well manner. Generally, it has been conducted to tell ASTU internet users' usage status of 2010 (local time) and it will be used as a systematic primary source for the process of establishing well-structured internet usage management policy that strengthen the effectiveness of teaching-learning process in the university.

1.4. Research Questions

The study tried to address the following research questions

- How long do the users spent on internet on average per a day?
- Which day and hour is busy?
- Which are top 20 frequently accessed Websites in both situations?
- Which websites are frequently accessed together by the users?
- What motivating rule is discovered that could be an input for ASTU ICT?

1.5. Significant of the study

This study is focused on investigating access behavior of ASTU internet users'. Then, the result will be used to generate a guidance and suggestion to benefit all other universities in order to know their users' reaction. This kind of study can find out the information related to the user's access behavior which is valuable to improve internet usage management and enable to know load of the network by analyzing the accessing time. And also, enable to predict the most preferable day and time when to shut down the server for maintenance and when low traffic is available for online analysis. Enables the proxy server administrators to inform when there is slow connection or busy traffic to restrict certain users from accessing some URLs. For the proxy server administrators in their plans to improve network efficiency by determining which Internet resources negatively affect the network. Generally, the study will be used as a systematic primary source for:

- The process of establishing well-structured internet usage management policy that strengthen the effectiveness of teaching-learning process.
- Researcher who are interested to conduct this kind of study for Ethiopian e-commerce websites.

1.6. Objective

1.6.1. General Objective

The general objective of this study is to analyses web log files of ASTU by applying web usage mining for extracting usage patterns.

1.6.2. Specific Objective

- To collect proxy server access log file from ASTU ICT proxy servers.
- To preprocess the collected proxy servers access log file in the way it will be suitable for mining process.
- To discover patterns of preprocessed log file by applying data mining techniques.
- To analyze the discovered pattern.
- To interpret and visualize the result.
- To recommending directions for the future.

1.7. Methods

1.7.1. Literature Review

Comprehensive literature review was done to investigate the fundamental theories of the various approaches, algorithms, techniques and tools that were employed in the research. Several books, journals, magazines, articles and papers relating to the domain knowledge have been referred to understand the possible applicability of web usage mining in user access behavior investigation.

1.7.2. Time situations Selection

As tried to mention above, in order to include users' access behavior on different situations that can vary the access behavior of the users in the campus, the researcher decided to focus on two main time situations such as **class time** and **exam time.** These two time situations has different impact on all activities of users in the campus. Class time enables us to know the access behavior of the campus users at the time of class delivery. Exam time enables us to know the access behavior of the campus users at the time of preparing themselves for exam and at the time of exam delivery.

1.7.3. Data Selection and Collection

Even though, there are three log file data sources such as web server, proxy server and client/browser to do web usage mining, the data source of this study was ASTU proxy servers. Proxy servers log file contains different types of files (cache log, store log, and access log), from those access log file is selected because of its appropriateness for investigating internet users' access behavior on multiple websites. Even if ASTU ICT doesn't use the proxy server access log files for further advanced analysis, it stores proxy log files. This study aimed to conduct ASTU internet users' access behavior investigation based on two time situations such as **class time** and **exam time**. It is obvious regular academic year has two semesters. Even if analyzing one year log file will leads to better conclusions, the researcher has chosen second semester because of time constraints to analyze one year log file and second semester log file was the latest than first semester log file. So, the researcher took second semester access log file ***(38 GB)*** from the system administrator.

1.7.4. Data Analysis

The study was done by following the three web usage mining process phases such as *data preprocessing*, *pattern discovery,* and *pattern analysis*.

A. Data Preprocessing

Since the access log file collected via proxy server is raw data and it is not suitable for data mining process, all unnecessary data has been removed to make the data appropriate for data mining process.

Although data preprocessing is time consuming, there is no suitable tool to perform the task. Specially, a researcher who needs to preprocess proxy server access log file will be forced to perform the preprocessing task manually. Due to this, the researcher performed partial of the preprocessing task manually and used Microsoft excel 2016 for the rest of preprocessing task.

Several techniques are available under data preprocessing but, by considering the objective of the study; data preparation, data cleaning, user identification and session identification was done in this phase. In data preparation, since the only free log file viewer is notepad, it cannot able to view the attribute separately and clearly and even the attributes has no header title. In order to view each attributes separately, the researcher imported the log file to Microsoft excel 2016 and give descriptive header title for the attributes. In data cleaning, removal of irrelevant and incomplete records is done. In user and session identification, the process of identifying the number of users and sessions is done.

B. Pattern Discovery

Pattern discovery is the process of applying data mining methods and techniques on the preprocessed data. From these data mining techniques association rule mining and statistical analysis have been done in this study. Microsoft Excel 2016 was used for statistical analysis and Apriori algorithm of ***WEKA*** tool was used for association rule mining. The researcher preferred Apriori algorithm is because gives output level by level and WEKA is chosen because of its popularity in data mining and easy to use for beginners.

C. Pattern Analysis

In this phase, separates the interesting and uninteresting pattern from the overall pattern discovered during pattern discovery phase. The discovered access behaviors are assessed based on the research objective expectation and through validating the association rule mining and statistical analysis techniques experiment results.

1.8. Scope and Limitation

- Even though proxy server stores different types of logs such as cache log, store log and access log, in this study only access log was used because of its suitability for users' access behavior investigation.
- Since ICT hadn't a mechanism to differentiate the users of the university such as academic staff, administrator staff and students, the study didn't present the access behavior separately.
- Although the objective of the study is analyzing ASTU proxy servers' access log file to investigate all users' access behavior, this study doesn't include summer program users.

1.9. Organization of the Thesis

This thesis is organized into five chapters. The first chapter focused on the problems that form the basis of this study. The general and specific objectives of the study, statement of the problem, research methods as well as research questions were discussed. In the second chapter, different literatures are discussed from domain knowledge and related works point of view. Moreover, some researches related to web usage analysis were briefly presented. The third chapter provide detailed process methods of algorithms of the research at hand. Each of the steps that were contained in the web usage analysis methodology proposed was discussed. In the fourth chapter, results and discussions are reported. The web usage patterns were discovered by integrating statistical analysis and data mining applications. Interesting rules and patterns of the experiments results identified. In the fifth chapter, concluding remarks and recommendations were made. Finally, lists of references and appendices were presented at the end of this paper.

Chapter Two: Literature Review

2.1. Overview of Web Mining

One of the applications areas of data mining is WWW. It serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services [7]. It contains a rich and dynamic collection of information about web page contents with hypertext structures and multimedia, hyperlink information, and access and usage information, providing fertile sources for data mining [8]. According to the survey in December 2016, the number of website was ***1.1 billion***, in 2017 it was ***1.7 billion*** and currently, it is above ***1.9 billion*** [6]. These statistics tells us, how the number of website is growing in unimaginable speed. As the number of website grow, a massive amount of information is generated due to users' interactions with websites. Websites are launching in different category such as e-commerce, health, educational, communication, file sharing, promotional, entertainments, news, game, and so on.

While interaction with web various problems like finding useful information, personalization of information, to learn about consumers or individual users, creating new knowledge from the information available on web arises. Web mining has emerged as most popular and effective technique to overcome these problems. [18].

For instance, e-commerce is one of the website category which is currently growing in high speed. E-commerce website is a place where big market exchange is occurred from anywhere at any time. As it is commercial website, it should provide its service as per the customer needs in order to be productive and profitable. To know the needs of online customer, transaction analysis is needed. Web mining plays a vital role here in order to analyze the transactions.

Web mining is an application use of data mining techniques to discover and extract information from web documents and services. It is the extraction of interesting and useful knowledge and hidden information from activities related to the WWW. Web mining enables us to discover unknown but potentially useful knowledge from a massive amount of web data. Although Web mining uses many data

mining techniques, it is not purely an application of traditional data mining due to the heterogeneity and semi structured or unstructured nature of the Web data.

Two different approaches were proposed for defining Web mining. The first approach is a ***'process-centric view'***, which defines Web mining as a sequence of ordered tasks. Second one is a ***'data-centric view'***, which defines web mining with respect to the types of web data that is used in the mining process. The data-centric definition has become more acceptable. Web Mining can be classified with respect to data it uses. Web involves three types of data; the actual data on the WWW, the web log data obtained from the users who browsed the web pages and the web structure data. Thus, the web mining should focus on three important dimensions; web structure mining, web content mining and web usage mining. The brief overview of web mining categories is given as follows [30].

2.1.1. Types of Web Mining

A. **Web Content Mining.**

Web content mining is one of the category of web mining that used to extract useful information from the content of the web document. Content data is the collection of facts a web page is designed to contain. It may consist of texts, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. [3].

B. **Web Structure Mining.**

Web structure mining is the process of discovering pattern from hyperlinks in a web. The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages [3]. Web structure mining uses graph theory to analyze the node and connection structure of a web site.

C. **Web Usage Mining.**

Web usage mining is the application of web mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications.

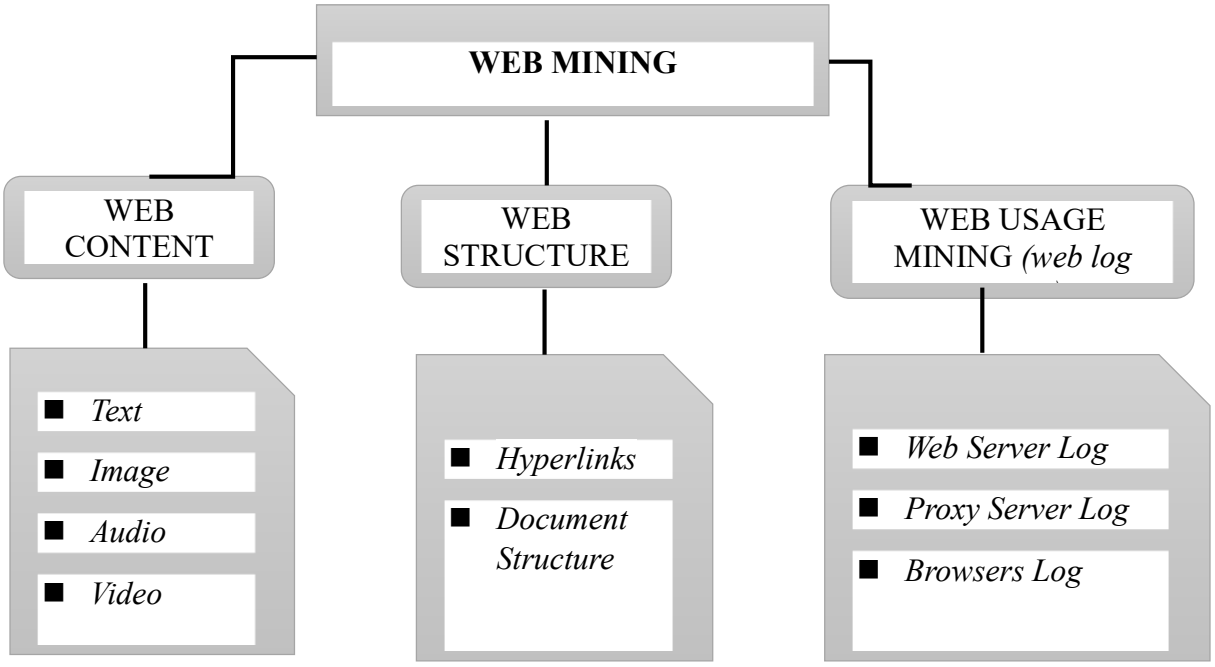

Figure2.1: Taxonomy of Web Mining.

2.1.2. Application of Web Mining

Nowadays, due to everything in the world is becoming performing via internet, web mining become rewarding research area and it has a lot of application area. Common application area of web mining are listed below.

- E- Commerce
- Search Engine and Web Search (information Retrieval)
- Website Design
- Recommendation Engines
- Web Communities and Web market places
- Network Management.
- Etc.

2.2. Web Usage Mining.

As mentioned above, web usage mining which is the main concern of this thesis, is the application of web mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site [3]. Web search companies routinely conduct web usage mining to improve their quality of service [8]. The usage data collected at the different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, and single-site browsing behavior to multi-user, and multi-site access patterns.

2.2.1. Web Log

Web log is a log automatically created and maintained by a web server. Every hit to the website, including each view of HTML document, image or other object is logged. It contains information about who was visiting the site, where they came from, and exactly what they are doing on the website. It is an essential part for the web mining to extract usage patterns and study the visiting characteristics of user [10].

2.2.1.1. Data Source

The usage data collected at the different sources will represent the navigation patterns of different segments of the overall web traffic, ranging from single-user, and single-site browsing behavior to multi-user, multi-site access patterns [10]. The data will be collected at different level as follows.

a. ***Server Level Collection***

A Web server log is an important source for performing web usage mining because it explicitly records the browsing behavior of website visitors. When an internet user request a particular page on web, an entry is logged into a special file called server log file. This file is not accessible by general internet user, only administrative person or server owners can access these file [18]. The data recorded in server logs reflects the access of a website by multiple users. These log files can be stored in various formats such as common log or extended log formats [10].

b. ***Proxy Level Collection***

A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy servers are used by ISPs and organizations like ASTU. Proxy caching can be used to reduce the loading time of a web page experienced by users as well as the network traffic load at the server and client sides [10]. Since proxy server is in between client and web server, every request from clients will pass through the proxy server to corresponding web server. And the data recorded in proxy server logs contains the access of multiple website by multiple users.

c. ***Client Level Collection.***

This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window, the entries to the log file is done only by the Web server.

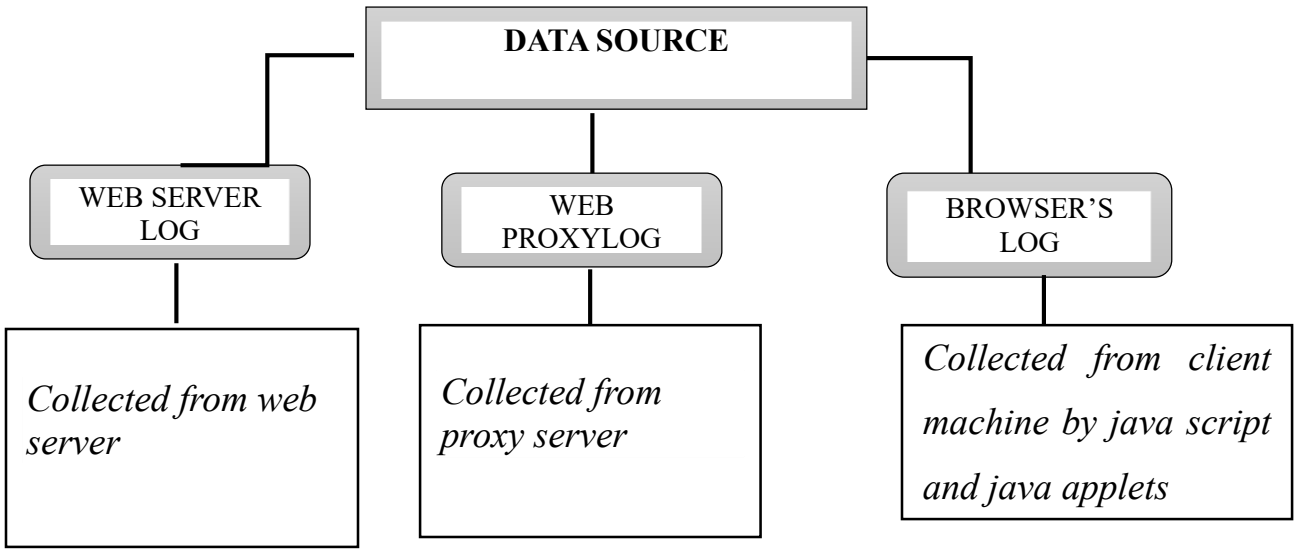

Figure 2.2: Types of log data and its source

2.2.1.2. Log File Parameters

Log files contain various parameters which are very useful in recognizing user browsing patterns. Below is the list of some of the parameters.

- ***Remote-host:*** is the remote hostname or its IP address;
- ***Log-name:*** is the remote log name of the user;
- ***Username:*** is the username with which the user has authenticated himself,
- ***Date and Time***: is the date and time of the request,
- ***URL (Request):*** is the exact request line as it came from the client,
- ***Status:*** is the HTTP status code returned to the client, and
- ***Bytes:*** is the content-length of the document transferred

2.2.1.3. Types of Log File Format

There are mainly two types of log file formats that are used by majority of the servers.

- ***Common log format:*** is a standard non-customized format (fixed no of attributes) suitable for http web sites. This type of log includes user's IP address/hostname, rfcname, log name, date and time, page access method, URL, http version, status code and byte received. [18]
- ***Extended log format:*** is a customizable log file format which can add some additional attributes like referrer_url, http_user_agent and cookies.[18].

Attribute	**Description**
IP Address/ Host Address	It identifies the visitor's machine address
Rfcname	Identifies user's authentication. ("-" character shows that this field is empty).
Log_name	Provides the login name of the user. ("-" character shows that this field is empty).
Date and time	Returns date and time of the user's request.
Page access method	Describes the mode of the request (GET, PUT, HEAD, POST)
URL	Give the path of requested page on web server.
HTTP Version	Return the version of HTTP protocol.
Status Code	Provides the status of response given by the server. (for example 404, represents 'file not found')
Byte received	Describes file size of file sent from the server.
Referrer_url	Provides the url from where the requested page is coming. ("-" character shows that this field is empty).
User_agent	Identifies user's OS and browser's version
Cookies	It is a piece of information sent to visitor's by the web server to identify the details of a particular user

Table 2.1: Attribute of common log and extended log format

2.2.2. Web Usage Mining Process

There are three phases for web usage mining: data preprocessing, pattern discovery, and pattern analysis. The three phases have been briefly explained below.

A. Data Preprocessing

Data preprocessing is an important one because of log data is unstructured, redundant and noise full. Data is preprocessed in order to help improve the quality of the data and to improve the efficiency and ease of the mining process. It takes almost 80% time of whole web usage mining process [18]. Here, some of the data preprocessing techniques have been explained.

- ***Data cleaning:*** The main purpose of data cleaning method is to eliminate useless data that may affect the whole mining process. [9].
  - ***Multimedia (images, videos and audio) Files:***
    - Data redundancy will occur and that lead the result for inaccuracy.
    - Categorized as useless files in web log preprocessing.
    - Web log files size can be reduced to less than 50% of its original sizes by eliminating the image request.
  - ***Web Robots Request:***
    - Request which is not from human (real user).
    - Dramatically affect the websites traffic statistics.
    - These are not important from the mining perspective and hence must be removed.
  - ***HTTP Status Codes:***
    - Log files with unsuccessful HTTP status code are usually eliminated during the web log cleaning process.
    - The widely acceptable definition for unsuccessful HTTP status codes is a code under 200 and over 299.
  - ***Other Files:***
    - Remove web log with file extension(.css, .pdf, .txt and .doc)
- ***User Identification:*** After cleaning HTTP log file, next step is the identification of the user; a user is a person or a client who interactively retrieves and extracts resources. User identification is a complex process just because of the presence of proxy servers. Web Mining methods are used to resolve these problems. However, it's not easy because of security and privacy. Following heuristics identify the user. [9].

  1. Each user has its unique IP address.
  2. If the IP address is same for more than one log entry, but if there is a change in browser software and operating system, then the combination of IP address, browser and OS represents a different user.

3. Using the URL and Referrer fields values of web access log the site topology can be constructed that shows the browsing path sequence for each user. If a page is in the browsing sequence that is not directly referred from the page visited by the user, it shows that there is another user who has same IP address.

- ***Session Identification:***Group of activities performed by a user from the moment he/she entered the website to the moment he/she left it. A set of user clicks usually referred to as a click stream, across web servers is defined as a user session. A sequence of web pages user browse in a single access. Session Identification isgrouping the different activities of a single user. It is process of segmenting the access log of each user into individual access sessions.

The goal of session identificationis group the page access of each user into individual access sessions. Identifying which user has spent how much time on the website. Each heuristic scans the user activity logs to which the web server log is partitioned.

The session identification heuristics: [12]

- ***Timeout:*** if the time between pages requests exceeds a certain limit (30 minute is default), it is assumed that the user is starting a new session.
- ***IP/Agent:*** Each different agent type for an IP address represents a different sessions.
- ***Referring page:*** If the referring page file for a request is not part of an open session, it is assumed that the request is coming from a different session.
- ***Same IP-Agent/different sessions (Closest):*** Assigns the request to the session that is closest to the referring page at the time of the request.
- ***Same IP-Agent/different sessions (Recent):*** In the case where multiple sessions are same distance from a page request, assigns the request to the session with the most recent referrer access in terms of time.

- ***Path Completion:*** After identify unique user session there is need to determine important page accesses that are not logged into the log file due to presence of client side or proxy side caching. If a user accesses a page by using back button in browser then it return copy of that page which is stored in cache. This kind of accessing does not record any entry in log file

that causes problem of missing references hence path completion techniques are required to fill these entries in log file.

B. Pattern Discovery

Pattern discovery is the next phase of web usage mining after data preprocessing is completed. Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. [5] In this section some of the techniques to discover pattern have been explained.

- ***Statistical Analysis***

Statistical techniques are the most common method to extract knowledge about visitors to a website. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. [5].

- ***Association Rule***

Association rule is popular research method for discovering interesting relations between variables in large data repositories. Association rule generation utilized to relate pages that are most often accessed together in a single server session. With regards to web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some predefined threshold. Aside from being applicable for business and marketing applications, the existence or nonexistence of such rules can help Web designers to rebuild their website. The most popular algorithm used to discover association rules are Apriori and FP Growth algorithm.

A. *Apriori Algorithm*

Apriori algorithm is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. [4]. It is an influential algorithm for mining frequent itemsets for Boolean association rules. It uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data).

The two basic concepts in association rule are ***support*** and ***confidence.***

Support: known as (frequency) is simply a probability that a randomly chosen transaction t contains both items A and B.

$$\text{Support } (A \Rightarrow B)t = P(A \subset t \wedge B \subset t) = \frac{\#\ of\ Transactions\ Containing\ both\ A\ and\ B}{Total\ \#\ of\ Transactions}$$

Confidence: known as (accuracy) is simply a probability that an itemset B is purchased in a randomly chosen transaction t given that the item set A is purchased.

$$\text{Confidence } (A \Rightarrow B)t = P(B \subset t \mid A \subset t) = \frac{\#\ of\ Transactions\ Containing\ both\ A\ and\ B}{Total\ \#\ of\ Transactions\ Containing\ A}$$

Support [62]

The rule X=>Y holds with support s if s% of transaction in D contains X**U**Y

Rules that have a s great than a user-specified support is said to have minimum support.

TID	ITEMS	Support = Occurrence / Total support
1	ABC	Total Support = 5 Support{AB} = 2/5= 40% Support {BC} =3 /5= 60% Support {ABC} = 1 /5= 20%
2	ABD	
3	BC	
4	AC	
5	BCD	

Confidence [62]

The rule X=>Y holds with confidence c if c% of transaction in D that contain X also contain Y. rules that have a c great than a user-specified confidence is said to have minimum confidence.

TID	ITEMS	Given X➔Y Confidence = Occurrence of {Y} / Occurrence of {X}
1	ABC	
2	ABD	Confidence {A=>B} = 2/3 = 66%
3	BC	Confidence {B=>C} = 3/4 =75%
4	AC	Confidence {AB=>C} =1 / 2=50 %
5	BCD	

B. *FP-Growth Algorithm*

The FP-Growth Algorithm is an efficient and scalable method for mining the complete set of frequent patterns. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset association information. [4]. FP-Growth is an improvement of apriori designed to eliminate some of the heavy bottlenecks in apriori. It allows frequent item set discovery without candidate item set generation. In an FP-Tree each node represents an item and its current count, and each branch represents a different association.

- ***Sequential Patterns***

The technique of sequential pattern discovery attempts to find inter- session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can foresee future visit patterns which will be helpful in setting promotions aimed at certain user groups. [5].

- ***Classification***

Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users having a place with a specific class or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc. [5].

- ***Clustering***

Clustering is a method to group together a set of items having comparable attributes. In the web usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. Clustering of users tends to build groups of users exhibiting similar browsing patterns. Clustering analysis in web usage mining intends to find the cluster of user, page, or sessions from web log file, where each cluster represents a group of objects with common interesting or characteristic. User clustering is designed to find user groups that have common interests based on their behaviors, and it is critical for user community construction. Page clustering is the process of clustering pages according to the users' access over them. Clustering of pages will discover groups of pages having related content. This information is useful for the Internet search engines and Web assistance providers. [11].

C. Pattern Analysis

Pattern Analysis is the final stage of WUM, which involves the validation and interpretation of the mined pattern. Validation to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process and interpretation is need because of the output of mining algorithms is mainly in mathematic form, not suitable for direct human interpretations [5].

2.2.3. Web Usage Mining Applications

Each of the applications can benefit from patterns that are ranked by subjective interesting. Web usage mining is used in the following areas: Web usage mining offers users the ability to analyze massive volumes of click stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs. Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages. [11]

By determining frequent access behavior for users, needed can be identified to improve the overall performance of future accesses. In addition to modifications to the linkage structure, identifying common access behaviors can be used to improve the actual design of web pages and to make other modifications to the site. Web usage patterns can be used to gather business intelligence to improve customer attraction, customer retention, sales, marketing and advertisement, cross sales. Mining of web usage patterns can help in the study of how browsers are used and the user’s interaction with a browser interface. [11].

2.2.4. Web Usage Mining Tools

In the area of World Wide Web, there are many websites are active over the internet. To perform the log analysis on huge number of available websites, numerous featured log analysis tools are existing. But the big difficulty is in selection of suitable tools. This table help to review the set of tools those are currently available tools.

S. No	Name of analysis tool	Current version	Log file format	Language	Plat form	Report format	Real time analysis	Source URL
1	Google analytics	Single	CLX,XLF,ELF	JavaScript	Window,Mac,linux	HTML,PDF,CSV	Yes	http://www.google.com/analytics
2	Deep log analyzer	7.0	Apahce,IIS,CLF,XLF,ELF	In-built	Window	HTML /Ms-Excel	Yes	http://www.deepoftware.com
3	Web logexpert	9.3	Apache,IIS	In-built	Window	HTML,PDF,CSV	No	http://weblogexpert.com
4	Visitors	0.7	CLF,Apache,IIS,W3C	C	Unix,linux	HTML	yes	http://www.hping.org/visitors/
5	Webalizer	2.23-08	CLF,XLF,ELF,FTP	C	Unix,linux	HTML	No	http://www.webanalizer.org
6	Analog	6.0	CLF,IIS,W3C	C	Window	HTML	No	http://analog.gsp.com/
7	Piwik	2.1 6.5	Apache,IIS,Ngnix	Phyton/ruby	Window/mac	HTML/PDF	Yes	http://www.piwik.org
8	Open web analytics	1.5.7	CLF,XLF,ELF	PHP	window	HTML	No	http://www.openwebanalytics.com
9	AWStats	7.3	CLF,XLF,ELF,W3C	Perl	window	HTML/PDF	Yes	http://www.awstats.org
10	Web log storming	3.2	IIS,W3C,etnedelog,apache	C	window	HTML/PDF	Yes	http://www.weblogstorming.com

11	Web trax	23	NCSA,Combined Format	Perl	Window/unix/mac	HTML	No	http://multicians.org/thvv/webtraxhelp.html
12	Dailystast	3.0	CLF,XLF,ELF,W3C	Perl	Window/mac	HTML	No	http://www.perlfect.com/freescripts/dailysta
13	Relax	2.80	Apache combined,NCSA,exteneed/CTL,webstar	Perl	linux	HTML/CSV/Text	No	http://ktmatu.com/software/relax/
14	StatCounter	3	Embedded in webpage	___	Windows	CSV	Yes	https://statcounter.com
15	SAWMILL	8.7.8	IIS,W3C,ELF,apache	C	Windows	HTML	Yes	http://sawmill.net/
16	GoAccess	1.0	Apache,ginx,CLF,ELF	C	Unix/linux	HTML, CSV,JSON	Yes	https://goaccess.io/
17	Nihuno log analyzer	4.19	Apache,ZEW,CLF,ELF, cloud front	Javascript, HTML	Window/linux	HTML	Yes	http://loganaler.net
18	HTTP analyze	2.4	NCSA,CLF,W3C,ELF	C	Window/linux	HTML	Yes	http://httpanalyze.org/index.php
19	Log analytics sense	2.3	More than 40 format	___	Window	HTML	No	http://www.statspire.com/
20	AlterWind log analyzer	4.0	Apachecommon combined IIS	___	Window	HTML	No	http://www.alterwind.com

Table 2.2: Web usage mining tools [22].

2.3. Related Works

Here, different related works have been reviewed in different perspectives.

2.3.1. Basics of web usage mining.

In this perspective, papers have been reviewed regarding the basics of web usage mining

Monika Dhandi and Rajesh Kumar Chakrawarti [9], in this paper, web usage mining and its process has been explained in detail. It stated that *"Web usage mining is the process of extracting useful information from web server logs based on the browsing and access patterns of the users".* It presents a comprehensive explanation by focusing on web usage mining process phases such as data preprocessing, pattern discovery, and pattern analysis. The paper has been prepared in the way to give general overview about web usage mining and its' process.
Nisha Yadav [17], presents a research area in web mining and discussed that web usage mining include enormous degree techniques for variety applications like web search, classification, personalization point of view. In the paper, web usage mining process phases explained in well manner. The author also discussed source of web usage mining and application of web usage mining.

The researcher has reviewed the above two papers to gather concepts regarding the basics of web usage mining. The papers had a great contribution for the study in order to provide domain knowledge concerning web usage mining.

2.3.2. Preprocessing

In this perspective, papers have been reviewed regarding data preprocessing in web usage mining process.
MitaliSrivastavaet. al, [18], the paper is a survey and discussed about the nature of log data and its' source in order to emphasize the essentiality and complexity of preprocessing. The paper stated that *"Preprocessing is considered as more complex and time consuming process due to diverse nature of log data".* It presents the four common preprocessing techniques such as data cleaning, user identification, session identification and path completion. The authors conducted a critical literature review on six papers which are written related to preprocessing techniques.

A. Deepa and P. Raajan [19], discussed some of the preprocessing techniques such as data cleaning, user identification, session identification, data filtering, and data formatting. It presents the algorithm of data cleaning and data filtering. The authors reviewed ten papers which are written related to preprocessing techniques. The paper stated that *"Preprocessing aims to reduce the size of file and enhance the quality of data".* Then the authors used IIS log file format web log dataset to apply preprocessing techniques and they have got reduced file and quality data for the mining process.

J. SoonuAravindan and Dr. K. Vivekanandan [20], provides an overview of various data pre-processing techniques used in web usage mining. It clearly explains the essentiality of data preprocessing. It stated that *"User accessed websites are recorded as a web log file, which may contain noisy & ambiguous data which may affect the results of the mining process. So there is a necessity to pre-process the web data before extracting knowledge from web log files".* It discussed seven preprocessing techniques such as data fusion, data cleaning, page view identification, user identification, session identification, path completion, and data integration using an example. The authors used IP address and User agent for user identification and time heuristics for session identification. The researcher used the above three papers to have a clear understanding on data preprocessing techniques in web usage mining. In order to create a better understanding, the papers show the preprocessing techniques by the help of algorithm and examples. All of the papers emphasized on four common preprocessing techniques such as data cleaning, user identification, session identification and path completion.

2.3.3. Tools

In this perspective, papers have been reviewed regarding web log analyzer tools

Vinod Kumar and Ramjeevan Singh Thakur [21], the paper will help to review the set of tools currently available for web log analysis. It discussed around twenty web analyzer tools. The authors also list the general features of web log analyzers. Today, large collections of web log analyzer software tools are available but, it is a boring and time consuming task to find the appropriate tool to perform the analysis economically, efficiently and effectively. This paper presents a summarized and comparative demonstration of 20 web log analyzer tools using table. And it concludes by stating that *"If someone wish to use proprietary software, the deep web log analyzer, and web log Expert analyzer are very powerful and useful giving almost similar result. In case of open source, Piwik is very strongly helpful for web log analysis. For online and real time web log analysis, google analytics and statcounter are very useful and powerful web log analyzer".*

ChhaviRana [22], discussed many web log analyzer tools and presents feature and function of twenty web log analyzer tools. The author informs that web usage mining research promises lot of space for advancements in the techniques and tools.

On the above two papers, many web log analyzer tools have been discussed and presented in the form of summarization and comparative demonstration. Even if numerous tools are available, most of them are very limited in the results they can provide. Additionally, most of the tools are made to analyze web server log data. There is lack of analyzer that conduct analysis of proxy server log data easily. After all the researcher gained a brief understanding concerning web log analyzer tools from the papers.

2.3.4. User Access Behavior from Web Server Log

In this perspective, papers have been reviewed regarding user access behavior identification using web server log file.

K. R. Suneetha and Dr. R. Krishnamoorthi [23], discussed the three location of log files with their drawbacks. It concentrated on preprocessing stage of web usage mining and then analyzed the preprocessed data for performance improvement of website design. The author used seven days 195 MB web server log file of NASA website. It applies two preprocessing techniques such as data cleaning and user identification. After data cleaning the log files reduced to less than 50 % of the original size.

NeetuAnand and Prof. SabaHilal [24], briefly describe the application of web usage mining. The authors used ten days of 400,000 records common log file format of NASA website web server log. In this paper three even though the authors focused on data cleaning, user identification and session identification had been used as preprocessing techniques. K-means clustering is used in the pattern discovery phase and WEKA is used as a tool.

Mustafa M.H. Ibrahim et al [25], this paper is different from others because it used data that has been captured by capturing device directly from the router in the form of port mirroring, did not used server log files. The authors conducted the research in University Utara Malaysia main campus as case study. They used Tcpdump tool to capture mirrored packets. Total capture files size was 197.2 GB within 5 days from 10:00-11:00 AM. Wireshark is used for packet analyzing process and Microsoft Excel is used as visualization tool. AmitDipchandjiKasliwal and Dr. Girish S. Katkar [26], divides the web usage mining

process phases in to four in contrast with other papers, preprocessing, pattern discovery and pattern analysis are common in many papers but, in this paper, before preprocessing phase input phase is added to retrieved and separate the server log files in to access, referrer and agent logs. The authors used MATLAB (matlab rules) to clean and filter the web server log file. In this paper data cleaning, data filtering, user and session identification and path completion had been used as preprocessing techniques. After preprocessing task, the file has been changed in to arfffile format in the way it will be suitable for FP-Growth association rule mining on RapidMiner tool.

2.3.5. User Access Behavior from Proxy server Log

In this perspective, papers have been reviewed regarding user access behavior identification using proxy server log file.

Jan Kerkhofs et al [27], focused to aware the essentiality of analyzing proxy server log data in web usage mining. To conduct the study, the authors used Belgian ISP proxy server log data. They have used data cleaning, user and session identification as preprocessing techniques, and association rule, sequence analysis and clustering method as pattern discovery techniques. After preprocessing task, the data reduced from 600 MB to 185 MB. In the paper, 50 websites which produced a minimum of 120 hits has been selected and categorized in to five sections such as newspaper, banks, finance, TV and Radio {7, 9, 17, 13, 4} respectively. The authors used generalized rule induction and apriori algorithm to perform association rule mining, Capri algorithm to perform sequence analysis and Kohenen algorithm to perform clustering of sessions. An input for the clustering algorithm was the same file that was used for association rule algorithm and identified 9 clusters of sessions. SPSS clementine used as a visualization tool.

BekaluShiferaw [30], in this paper web usage mining, its process and its application area has been broadly explained. The author conducts the research as case study of Adama Science and Technology University to identify internet users' general access pattern. He used one month (October 2015) proxy server log access file which original size is 10 GB (600 MB after preprocessing). He used data cleaning technique in preprocessing phase and association rule mining in pattern discovery phase. The data cleaning is performed manually. The author used proxy log explorer as a tool to analyze and illustrate user navigational behaviors and Microsoft Excel for visualization. The author categorized the data in to two categories such as *session time interval cluster* (access log time interval period) and *URL cluster* (grouping of similar URLs). The first

category has 8 clusters (morning, lunch time, afternoon, night, late night-early morning, Saturday, Sunday, and holiday). The second category has 13 clusters (educational, entertainment, search engines and job searches, social media, sport and general news, advertisement and shopping, computer and internet, downloads, government and organizations, religion and health, pornography, malicious sites, and others). The paper tried to visualize the results as per the above two categories even if the way of visualization is poor in order to perceive the results effortlessly.

Nattapol Kiatkumjoun wong et al. [28], proposed a model to classify web proxy traffic into normal, medium, high and burst traffic. The proposed system consists of three main processes; data preprocessing, outlier detection and log classification. The authors conducted their experiment on five datasets having the total number of log records {6447, 4581, 2325, 8058, and 8434}. In data preprocessing phase, data cleansing is done by removing all records containing incomplete and invalid data. The outlier detection phase consists of three steps; numeric attribute discretization, minimum support selection, and execution of Apriori algorithm. To separate frequent itemsets, the well-known Apriori algorithm is applied. WEKA software is used for the execution. The authors consider only five file category as they constitute the majority of web access. They are text, image, video, application and others. Each file category has its own several file types. For example, image file category have gif, jpeg, png, bmp etc. the normal logs are separated during this phase. Final phase log classification, to do the log classification under the same criteria would not be effective because each file type has its own characteristics. The medium, high and burst traffic are classified based on the computed rates (***computed rate=bandwidth / duration***) and the setting of thresholdvalues (80%). The authors consider under 80% as medium rated logs and all logs above 80% as high and burst rated logs. Finally to classify high or burst rated logs, the average and the standard deviation of the traffic rates are computed, and the doubled standard deviation is used as the threshold. If the logs have the rate less than the average plus the double standard deviation, they are considered as high rated logs and the greater one are considered as burst rated logs. In this paper, phases and steps are accomplished in deep detail and clear manner.

Kartik Bommepally et al. [29], proposed web based tool to analyze internet activity. The authors briefly discussed the design and implementation of a proxy analyzer which is implemented by them. The tools have three components; log parser, database loader and data analyzer. The data is collected from proxy servers at a campus wide network, 5 months' log data. The paper focuses on investigating average access time and analysis of web traffic originated from different users. The average access time analysis report focuses on;

individual statistics *(users' access time for a particular website)****, relative statistics*** *(average access time across users)* ***and general statistics*** *(average access time of users across different websites)****.***Even though the developed analyzer has some limitations, it has contributed a lot for the authors to meet their objectives.

The limitation to [30] is that, the author didn't perform user identification and session identification. As the behavior of campus internet users is varied during **class time** and **exam time**, the sample date that the author selected to use for analysis is not appropriate to identify genuine access behavior of internet users'. The author selected the access log from October 01/2015 to October 31/2015 in Gregorian calendar. That means from Meskerem 21/2008 to Tikimt 21/2008 in our calendar. In this time, it is difficult to get the full participation of the users. This will lead us to wrong conclusion. And also, the author's website categorization method is not reasonable and persuasive. For instance, he categorized sport & news together as one category and religion & health together as one category.

This study overcome the above limitations by performing user and session identification. Since teaching-learning is the main objective of universities, the two time situations **(class time** and **exam time)** have direct impact on universities overall activities. By keeping in mind this, the researcher used access log data that helps to investigate the internet users' access behavior on the two time situations and obtained fair, better and comprehensive conclusions. On this study, websites are categorized based on their highly cohered similarities. To indicate the key ideas in the papers noticeably, summary of selected papers are presented in table below.

No.	Author(s)	Publication Year	Objectives	Used log file	Preprocessing Techniques	Pattern Discovery (Outlier Detection) Techniques	Tools	Remarks
1	K. R. Suneetha and Dr. R. Krishnamoorthi [23]	2009	Identifying user behavior	Web server log file	Data cleaning and user identification	----	---	Used only preprocessing phase to identify user behavior. Poor conclusion.
2	NeetuAnand and Prof(Dr.)SabaHilal [24]	2012	Identifying user access pattern	Web server log file	Data cleaning, user & session identification	Clustering (K-means Clustering)	WEKA	Didn't use a complete pattern discovery techniques, and didn't meet its' objective.
3	Mustafa M.H. Ibrahim et al [25]	2014	Analysis of Internet traffic	Live captured packets	---	---	Wireshark & Ms. Excel	It presents the users behavior clearly and effectively
4	AmitDipchandjiKaslikwal and Dr. Girish S. Katkar [26]	2015	Prediction of user access behavior	Web server log file	Data cleaning, data filtering, user & session identification and path completion	Association Rule (FP-growth algorithm)	MATLAB &RapidMiner	It's good in analyzer tool selection but poor in result presentation and visualization.

5	B.Shiferaw [30]	2016, Unpublished	Investigating Internet Users' general access pattern	Proxy server log file	Data Cleaning	Association Rule (Apriorialgo.) & statistical analysis	Proxy Log Explorer & Ms. Excel	User and Session identification are ignored in preprocessing phase. Sampled access log is not inclusive. Web site categorization is not convincing.
6	Jan Kerkhofs et al [27]	2001	to give an indication of what kind of informationcan be extracted from the proxy log files	Proxy server log file	Data cleaning, user & session identification	Association Rule, Sequence Analysis, Clustering	SPSS Clementine	Used E-metrics in addition to common preprocessing & pattern discovery techniques, this helped the authors to obtain a good conclusion.
7	NattapolKiatkumjoun wong et al. [28]	2014	Classification of web proxy logs based on patterns of traffic rates	Proxy server log file	Data cleaning	Frequent Itemset, Statistical analysis (Apriori algorithm)	WEKA & Ms. Excel	Phases and steps are done in deep detail and clear manner. The authors met their

								objective.
8	KartikBommepally et al. [29]	2010	analyze a large proxy log to explore user access patterns	Proxy server log file	---	---	Proxy analyzer (developed by the authors)	Develop a web based analyzer tool (Even though the tool has limitations), this helped the authors to meet their objective.

Table 2.3: Summary of literature

Chapter Three: Research Methodology

3.1. Overview

Research methodology used to understand what research methods and approaches are going to be used for the purpose of the study strategy of data collection, preparation, organization, analysis, visualization and interpretation. The nature of this study enforced the researcher to follow mixed approach (i.e. both qualitative and quantitative). The data is qualitative; the approach used to analyze the data was quantitative and finally interpreted in qualitative manner. Mostly, the study followed qualitative case study approach; because by conducting this kind of study in the case of ASTU, the researcher implicitly needs to awake/recommend other universities and internet based business organizations on two basic concepts; (1) use proxy server that enables them to store users' usage details and manage users/customers, (2) applying web usage mining on the proxy server's access log data in order to understand users'/customers' need early. Generally, the study has been done to obtain qualitative results by using qualitative data and following quantitative analysis approaches. To meet the objective of the study, the following methodological steps are followed.

3.2. Time Situations Selection

As tried to mention above, in order to include users' access behavior on different situations that can vary the access behavior of the users in the campus, the researcher decided to focus on two main time situations such as **class time** and **exam time.** These two time situations have different impact on all activities of users in the campus. Class time enables us to know the access behavior of the campus users at the time of class delivery. Exam time enables us to know the access behavior of the campus users at the time of preparing themselves for exam and at the time of exam delivery.

3.3. Data Selection and Collection

Even though, there are three log file data sources such as web server, proxy server and client/browser to do web usage mining, as tried to mention on the title, the data source of this study was ASTU proxy servers. Proxy servers log file contains different types of files (cache log, store log, and access log), from those access log file is selected because of its appropriateness for investigating internet users' access behavior on multiple websites. Even if ASTU ICT doesn't use the proxy server access log files for further advanced analysis, it

stores proxy log files. This study aimed to conduct ASTU internet users' access behavior investigation based on two time situations such as **class time** and **exam time**. As we know regular academic year has two semesters. Even if analyzing one-year log file will lead to better conclusions, the researcher has chosen second semester because of time constraints to analyze one-year log file and second semester log file was the latest than first semester log file. So, the researcher took second semester access log file ***(38 GB)*** from the system administrator.

3.4. Sampling Technique

As we can see above the size of the semester log file is 38 GB because each and every request of multiple users was stored day to day. Due to this, still it is difficult to analyze one semester log file within the given time. So, the researcher planned to analyze by taking sample log files from the whole semester log files. It's known that the best sampling leads to the best conclusion. Since this study focused on two time situations (class time and exam time), the date we consider for each situation matters on the conclusion. For each situation, strict presence of users is a mandatory. By keeping in mind this, the researcher followed cluster sampling technique. First of all, the semester divided in to 16 weeks by considering the 8^{th} week is mid-exam week and the 16^{th} week is final-exam week based on course outline preparation format. From those 16 weeks; 4^{th}& 5^{th} week before mid-exam and 12^{th}& 13^{th} week after mid-exam selected for class time situations; 7^{th}& 8^{th} week during mid-exam and 15^{th}& 16^{th} week during final exam selected for exam time situations. The reason that the week selects is in order to get full participation of users. And the size of the data of 8 weeks is 17GB.

According to the university (ASTU) 2017/2018 regular calendar, registration date was on Mar 03, 2018 to Mar 04, 2018, class begin date was on Mar 05, 2018, mid-exam date was from April 23, 2018 to April 30, 2018, and final-exam date was from Jun 11, 2018 to Jun 20, 2018. Based on this schedule, the researcher used access log files of appropriate sample date for the two time situations. The date sampled for the two time situations are clearly shown in the below table.

Situations		***Date Interval***	***Week***	***Remarks***
Situation 1	**Class time**	MARCH 26, 2018 -APRIL 01, 2018	4th	*Two weeks before mid-exam*
		APRIL 02, 2018 – APRIL 08, 2018	5th	
		APRIL 30, 2018 -MAY 06, 2018	12th	*Two weeks after mid-exam*
		MAY 07, 2018 – MAY 13, 2018	13th	
Situation 2	**Exam time**	APRIL 16, 2018 – APRIL 22, 2018	7th	*One week before- mid-exam*
		APRIL 23, 2018 – APRIL 29, 2018	8th	*Mid-Exam*
		JUN 04, 2018 – JUN 10, 2018	15th	*One week before- final-exam*
		JUN 11, 2018 – JUN 20, 2018	16th	*Final-Exam*

Table 3.1: Date based time situations' partition

3.5. The web log file

As tried to mention above, access log data has different format such as common log format, extended log format and native log format which is default log format of squid proxy server. Since ASTU ICT uses squid proxy server, the access log was stored in native format. Native format is recommended due to its greater amount of information made available for later analysis.

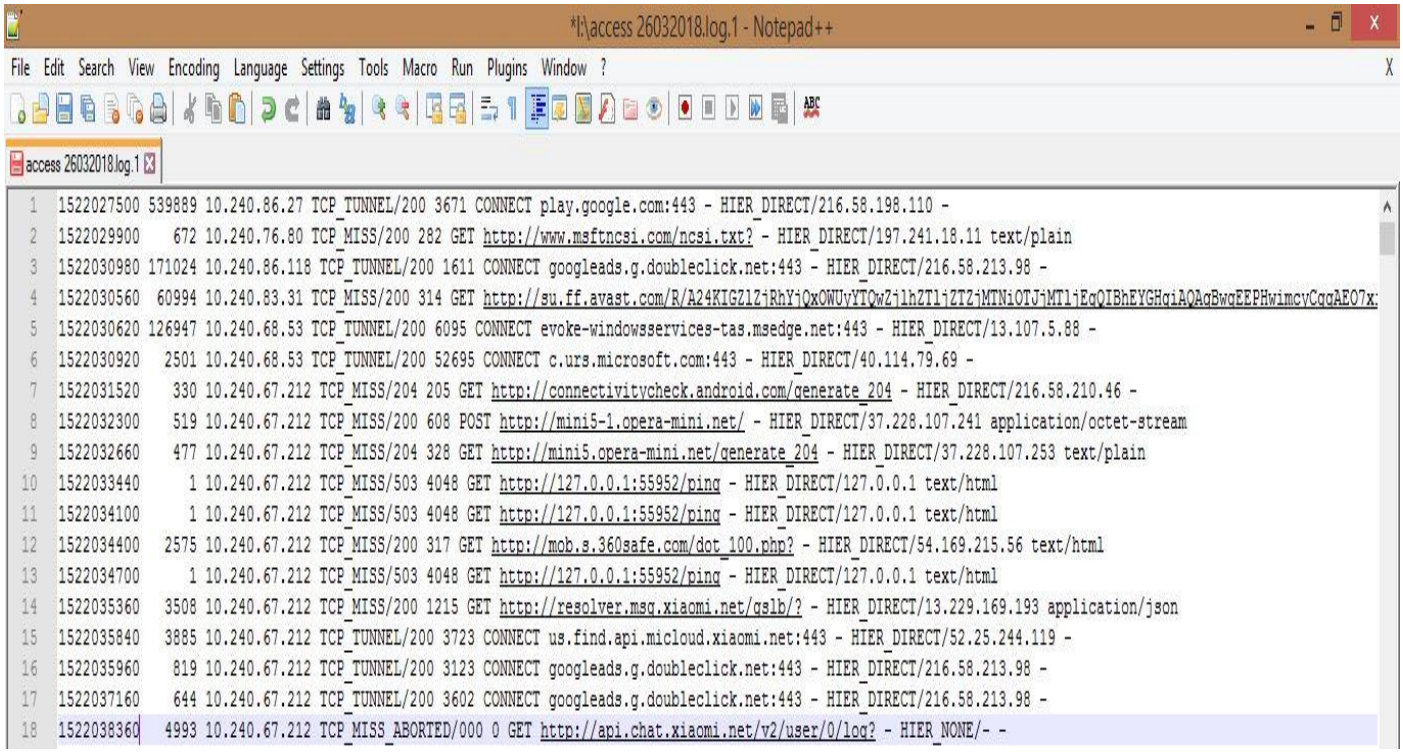

Figure 3.1: Sample Access Log file

In general, an access.log native format entry usually consists of (at least) 10 columns separated by one or more spaces [30].

Sample record of access log file

152202750049134 10.240.210.62 TCP_MISS/502 4292 POST http://nob1.tinvupdates.ru/index.asp - DIRECT/109.236.84.46 text/html

No	*Attributes*	*Description of the attribute*	*Example*
1	Date and Time	The client request time stamp is in Squid format; it is the time of request has been sent. We can convert this Unix squid time stamp into human readable by using excel formula ***(=Unix time reference cell/86400+25569)***	**1522027500** (26-03-2018 1:25 AM)
2	Duration(Response Time) in millisecond	The time the proxy spent processing the client request. The timer starts when the proxy server receives the HTTP request andstops when the response has been fully delivered.	**49134**
3	IP Address	Client's IP address	**10.240.210.62**
4	Result/Response status Code	Consists of two tokens separated by a slash. The first token, result code, classifies the protocol and the result of a transaction (e.g., TCP_HIT or UDP_DENIED). The codes that begin with TCP_ refer to HTTP requests, while UDP_ refers to ICP queries.The second token is the HTTP response status code (e.g, 200, 304, 404, etc.).	**TCP_MISS/502** ***TCP_MISS***- *means squid didn't have a cached copy of the requested resource.* ***502***- *means Bad Gateway*
5	Transfer Size in Byte	This field indicates the number of bytes transferred to the client. Strictly speaking, it is the number of bytes that Squid told the TCP/IP stack to send to the client.	**4292**
6	Request Method	The action the client was tryingto perform.(GET, POST, HEAD). The most common HTTP request method is GET.	**POST**

7	Accessed URI	Squid uses a special format for certain failures. These are cases when Squid can't parse the HTTP request or otherwise determine the URI. Instead of a URI/URL, you'll see a string such as "error:invalid-request."	**http://nob1.tinvupdates.ru/index.asp**
8	Client Identity	The authenticated client's user name. if you use proxy authentication, Squid places the given username in this field	**-**
9	Peering Code/Peer Host/	The peering information consists of two tokens, separated by a slash. It is relevant only for requests that are cache misses. The first token indicates to where the request has been sent (destination). The second token is IP address of the destination.	**DIRECT/109.236.84.46** ***DIRECT*** *– means squid forwarded the request directly to the origin server.* **109.236.84.46 -** *Origin server's IP address*
10	File Type	The final field of the default, native access.log is the content type of the HTTP response.	**text/html**

Table 3.2: Native log attributes and their description

Response Status Code	Description		Response Status Code	Description
1xx	**Informational**		403	Forbidden
100	Continue		404	Not Found
101	Switching Protocols		405	Method Not Allowed
2xx	**Successful**		406	Not Acceptable
200	OK		407	Proxy Authentication Required
201	Created		408	Request Timeout
202	Accepted		409	Conflict
203	Non-Authoritative Information		410	Gone
204	No Content		411	Length Required
205	Reset Content		412	Precondition Failed
206	Partial Content		413	Request Entity Too Large
3xx	**Redirection**		414	Request-URI Too Long
300	Multiple Choices		415	Unsupported Media Type
301	Moved Permanently		416	Requested Range Not Satisfiable
302	Found		417	Expectation Failed
303	See Other		**5xx**	**Server Error**
304	Not Modified		500	Internal Server Error
305	Use Proxy		501	Not Implemented
306	(Unused)		502	Bad Gateway
307	Temporary Redirect		503	Service Unavailable
4xx	**Client Error**		504	Gateway Timeout
400	Bad Request		505	HTTP Version Not Supported
401	Unauthorized		**6xx**	**Proxy Error**
402	Payment Required		600	Unparseable Response Headers

Table 3.3: HTTP Response Status Codes

3.6. Data Analysis

The study was done by following the three common web usage mining process phases such as *data preprocessing*, *pattern discovery,* and *pattern analysis*.

3.6.1. Data Preprocessing

Since the access log file collected via proxy server is raw data and it is not suitable for data mining process, all unnecessary data will be removed and corrected in this phase to make the data appropriate for data mining process. Even though several techniques are available under data preprocessing, by considering the objective of the study; data preparation, data cleaning, user identification and session identification was done in this study. As data preprocessing is time consuming, there is no suitable tool to perform the task. Specially, a researcher who needs to preprocess proxy server access log file will be forced to perform the preprocessing task manually. Due to this, the researcher performed partial of the preprocessing task manually and used Microsoft excel 2016 for the rest of preprocessing task. Tasks those are done under data preprocessing phase is clearly explained below.

3.6.1.1. Data Preparation

The raw access log file is text file that needs advanced tool to view rather than notepad. The researcher used Microsoft Excel 2016 to view the log files in the way all attributes are visible separately. Since the raw file had no header title, it was confusing to know the attributes easily. So, descriptive header title giving for each column has been done. Since Proxy Squid uses UNIX epoch format ((Thu Jan 1 00:00:00 UTC 1970), instead of something more human-friendly, to simplify the work of various log file processing programs [31].Converting the UNIX time epoch format into human readable date-time string format has been done here. The researcher used the below formula in excel for timestamp conversion.

Human readable date-time = (UNIX Timestamp/86400) + 25569

= (CELL VALUE/ 86400) + 25569 (if the UNIX timestamp is in excel cell)

The formula assumes that the UNIX timestamp is on cell A1:

= (A1 / 86400) + 25569

Note:

86400 is the number of seconds in the day = (60*60*24)

25569 is the number of days between 1970-01-01 and 1900-01-01 and which are the basesets for Excel and UNIX time stamps (min date in windows excel and UNIX).

	A	B	C	D	E	F	G	H	I	J
1	Date and Time	Converted DT	Response Time	IP Address	Status Code	Transfer Size	Request Method	URI	Client Identity	Peering Code
2	1522027500	3-26-18 1:25 AM	539889	10.240.86.27	TCP_TUNNEL/200	3671	CONNECT	play.google.com:443	-	HIER_DIRECT/216.58.198
3	1522029900	3-26-18 2:05 AM	672	10.240.76.80	TCP_MISS/200	282	GET	http://www.msftncsi.com/ncsi.txt?	-	HIER_DIRECT/197.241.18
4	1522030980	3-26-18 2:23 AM	171024	10.240.86.118	TCP_TUNNEL/200	1611	CONNECT	googleads.g.doubleclick.net:443	-	HIER_DIRECT/216.58.213
5	1522030560	3-26-18 2:16 AM	60994	10.240.83.31	TCP_MISS/200	314	GET	http://su.ff.avast.com/R/A24KIGZlZjRhYjQxOW	-	HIER_DIRECT/77.234.43.2
6	1522030620	3-26-18 2:17 AM	126947	10.240.68.53	TCP_TUNNEL/200	6095	CONNECT	evoke-windowsservices-tas.msedge.net:443	-	HIER_DIRECT/13.107.5.8
7	1522030920	3-26-18 2:22 AM	2501	10.240.68.53	TCP_TUNNEL/200	52695	CONNECT	c.urs.microsoft.com:443	-	HIER_DIRECT/40.114.79.6
8	1522031520	3-26-18 2:32 AM	330	10.240.67.212	TCP_MISS/204	205	GET	http://connectivitycheck.android.com/generat	-	HIER_DIRECT/216.58.210
9	1522032300	3-26-18 2:45 AM	519	10.240.67.212	TCP_MISS/200	608	POST	http://mini5-1.opera-mini.net/	-	HIER_DIRECT/37.228.107
10	1522032660	3-26-18 2:51 AM	477	10.240.67.212	TCP_MISS/204	328	GET	http://mini5.opera-mini.net/generate_204	-	HIER_DIRECT/37.228.107
11	1522033440	3-26-18 3:04 AM	1	10.240.67.212	TCP_MISS/503	4048	GET	http://127.0.0.1:55952/ping	-	HIER_DIRECT/127.0.0.1
12	1522034100	3-26-18 3:15 AM	1	10.240.67.212	TCP_MISS/503	4048	GET	http://127.0.0.1:55952/ping	-	HIER_DIRECT/127.0.0.1
13	1522034400	3-26-18 3:20 AM	2575	10.240.67.212	TCP_MISS/200	317	GET	http://mob.s.360safe.com/dot_100.php?	-	HIER_DIRECT/54.169.215
14	1522034700	3-26-18 3:25 AM	1	10.240.67.212	TCP_MISS/503	4048	GET	http://127.0.0.1:55952/ping	-	HIER_DIRECT/127.0.0.1
15	1522035360	3-26-18 3:36 AM	3508	10.240.67.212	TCP_MISS/200	1215	GET	http://resolver.msg.xiaomi.net/gslb/?	-	HIER_DIRECT/13.229.169
16	1522035840	3-26-18 3:44 AM	3885	10.240.67.212	TCP_TUNNEL/200	3723	CONNECT	us.find.api.micloud.xiaomi.net:443	-	HIER_DIRECT/52.25.244.1
17	1522035960	3-26-18 3:46 AM	819	10.240.67.212	TCP_TUNNEL/200	3123	CONNECT	googleads.g.doubleclick.net:443	-	HIER_DIRECT/216.58.213
18	1522037160	3-26-18 4:06 AM	644	10.240.67.212	TCP_TUNNEL/200	3602	CONNECT	googleads.g.doubleclick.net:443	-	HIER_DIRECT/216.58.213
19	1522038360	3-26-18 4:26 AM	4993	10.240.67.212	TCP_MISS_ABORTED	0	GET	http://api.chat.xiaomi.net/v2/user/0/log?	-	HIER_NONE/-

Figure 3.2: Sample Access Log file after preparation

3.6.1.2. Data Cleaning.

Data cleaning was done to remove irrelevant and missing data which is not useful for future work in mining process. Data Cleaning enables to filter out useless data which reduce the log file size to use less storage space and to facilitate upcoming tasks. The quality of results strongly depends on the cleaning process. Proper cleaning of data has high effects on the performance of web usage mining. The researcher applied the following steps to perform the data cleaning:

A. Removal of records with request methods other than 'GET'

Since this study is interested on users information request from server and the only cacheable request method is 'GET', request method other than 'GET' should be removed.

B. Removal of records in file extension other than (.html, .htm, .php, .asp)

Since the study is dealing with user behavior, the records should be intentional user request. There are records that are automatically downloaded without the intention of the users; such as video, audio, graphics, and robot.txt files. Records with these file type should be removed.

C. Removal of records with the failed HTTP status code

Status code shows the success or failure of a request. In this study records with successful status code (in between 200 and 299) were considered. Records with other status code are not useful for session identification because of time spent with the user is unavailable.

D. Removal of incomplete records

Incomplete records means records which have no value for all access log file attributes. As tried to list above, among the attributes of access log files are date and time, Duration, IP Address, Transferred Size, URI, File Type etc. If two or more of these attributes are empty, the record should be removed to do not draw mistaken conclusion.

E. Removal of records with bytes transferred field zero

Byte transferred size zero means, the server doesn't reply for the users' request. So, if the server doesn't reply to the user, there is no communication between user and server and that means there is no time spent.

F. Only DNS name is considered.

The last but not the least task of data cleaning is editing the URI column in to its DNS name. Since the study is dealing about websites, suffixes after DNS name are not required. For instance if the URI is www.websitename.com/gallary/image09878.htm, only www.websitename.com is enough. Otherwise it will be difficult to investigate frequently accessed websites.

As mentioned above, the researcher took 38 GB one semester data. After random week selection, it become 17 GB. But, after all of the above data cleaning steps, the data has been reduced to 1.2 GB. Table 3.4 shows the number of records before and after data cleaning per the selected week.

Class Time			**Exam Time**		
	#Records Before and After Cleaning			**#Records Before and After Cleaning**	
Week	**Before**	**After**	**Week**	**Before**	**After**
4th	37,681,712	422,933	7th	34,155,199	332,095
5th	35,711,993	340,696	8th	34,018,243	234,869
12th	46,584,777	324,310	15th	33,827,310	226,548
13th	36,840,090	327,888	16th	30,019,151	283,637
Total	**156,818,572**	**1,415,827**	**Total**	**132,019,903**	**1,077,149**

Table 3.4: Number of Records Before and After Data Cleaning

3.6.1.3. User Identification

After the log file has been prepared and cleaned, the next step is identifying users. User identification is a process of identifying the unique users who are accessed different websites. In this case, the researcher identified unique users per a day based up on "***Different IP addresses refer to different users***" rule.

3.6.1.4. Session Identification

After user identification, the researcher identified each session per a single user and total sessions per a day. A rule of thumb that is commonly used is that when there is an interval of 30 minutes between two page views, the click stream should be divided in two sessions. This rule has been applied since Pitkow [37] established a timeout of 25.5 minutes, based on empirical data.

3.6.2. Pattern Discovery

Pattern discovery is the process of applying data mining methods and techniques on the preprocessed data. From these data mining techniques association rule mining and statistical analysis have been done in this study.

3.6.2.1. Statistical Analysis

Statistical analysis has been done in order to explain things those should be told in number for clear description. The following tasks has been done under this phase and explained using graph.

- Number of records has been counted before and after preprocessing.
- Average request per hour.
- Top 20 Websites of both situations (class time and exam time).
- Daily average number of users, sessions and websites in both situations.
- Average request of days of a week (Monday to Sunday)
- Number of error code of both situations.

3.6.2.2. Association Rule Mining.

After a completion of statistical analysis, association rules mining has been done. Association rule mining is a process of discovering websites frequently accessed together.

I. Algorithm and Tool Selection

Apriori algorithm of WEKA tool was used for association rule mining. The researcher preferred Apriori algorithm is because it gives output level by level (i.e. most frequently accessed support and confidence of 2 itemsets, 3 itemsets, 4 itemsetsetc. based on the support threshold). WEKA is chosen because of its popularity in data mining and easy to use for beginners.

II. Data preparation

Since here the aim is to discover patterns of websites which are frequently accessed together, excel data (in csv file format) that contains top 20 websites as column and sessions of unique users as row is prepared. The researcher used Boolean representation to prepare the data which is suitable for the Apriori algorithm. **"T"** (if the website is accessed in the session) and **"NULL"** (if the website is not accessed in the session). In order to prepare excel data (in csv file format) input for apriori algorithm, the researcher used Wednesday log file for class time and Sunday log file for exam time because while statistical analysis Wednesday and Sunday were busy day during class time and exam time respectively.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Session/Website	playstore	google	facebook	youtube	en.softon	getintopc	diretube	filehippo	bbc	sodere	tewnet
2	S-1	T		T	T			T			T	T
3	S-2	T		T	T			T			T	
4	S-3	T		T	T			T			T	T
5	S-4	T		T	T			T			T	
6	S-5		T	T	T					T		
7	S-6	T		T	T			T			T	T
8	S-7	T		T	T			T			T	
9	S-8		T	T	T			T		T		
10	S-9	T		T	T	T			T		T	
11	S-10	T		T	T	T		T	T		T	T
12	S-11	T		T	T	T		T	T		T	
13	S-12		T	T		T		T	T	T		
14	S-13	T		T	T	T			T	T	T	
15	S-14	T		T	T	T		T	T		T	T
16	S-15	T		T	T			T	T		T	
17	S-16		T	T	T	T		T	T	T		
18	S-17	T		T	T	T			T		T	

Figure 3.3: Class Time: Sample of prepared .csv file for Apriori algorithm

III. Mining Process

As mentioned above, the excel data is prepared in the way it will be suitable for the Apriori algorithm and imported to WEKA. Steps followed by the researcher to execute the data is explained below

1. Import CSV File
2. Check the data type of the attributes (It should be nominal).
3. Go to **"Associate"** tab set minimum support and confidence.
4. Start the Algorithm.

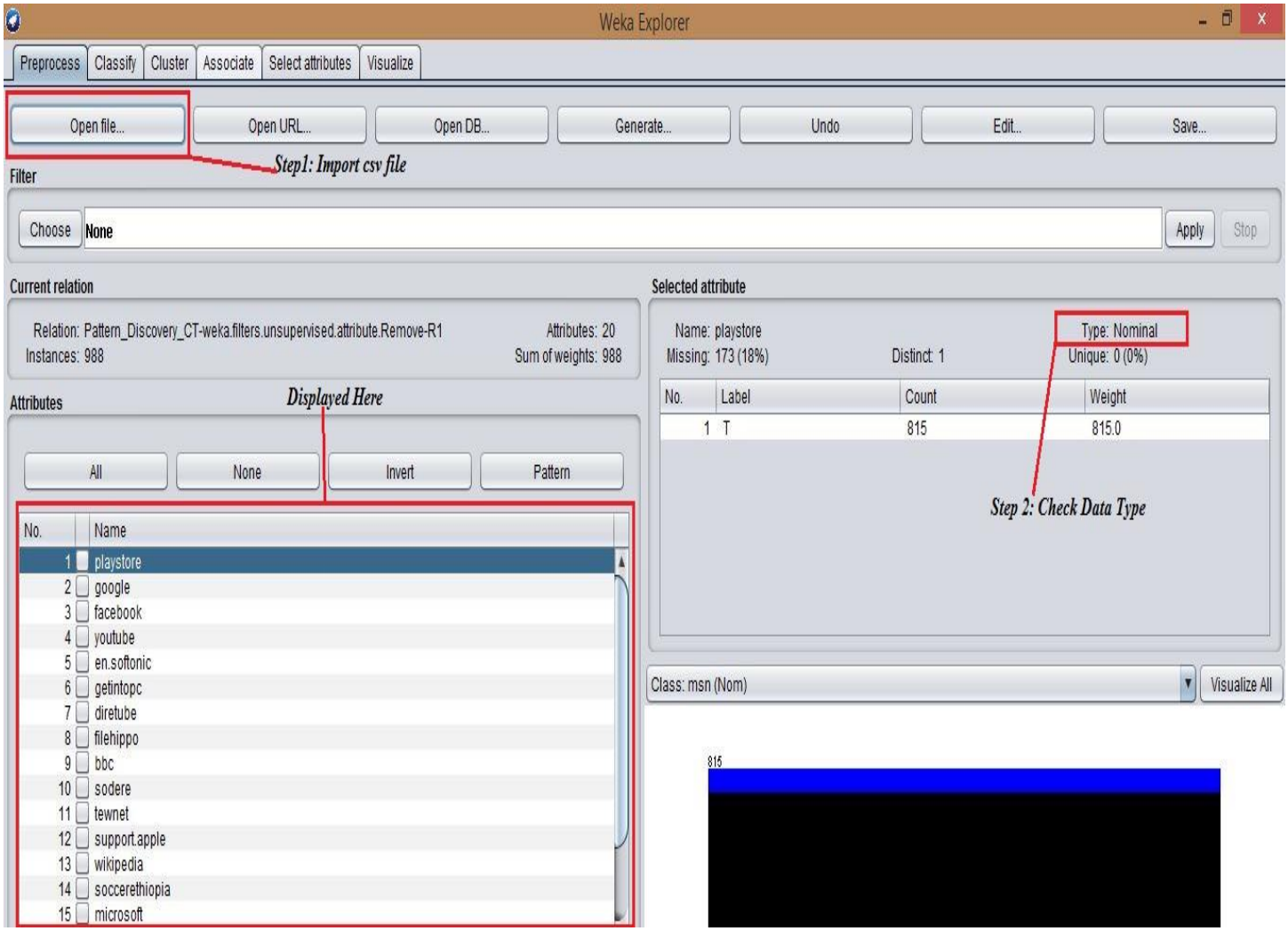

Figure 3.4: Import CSV file to WEKA

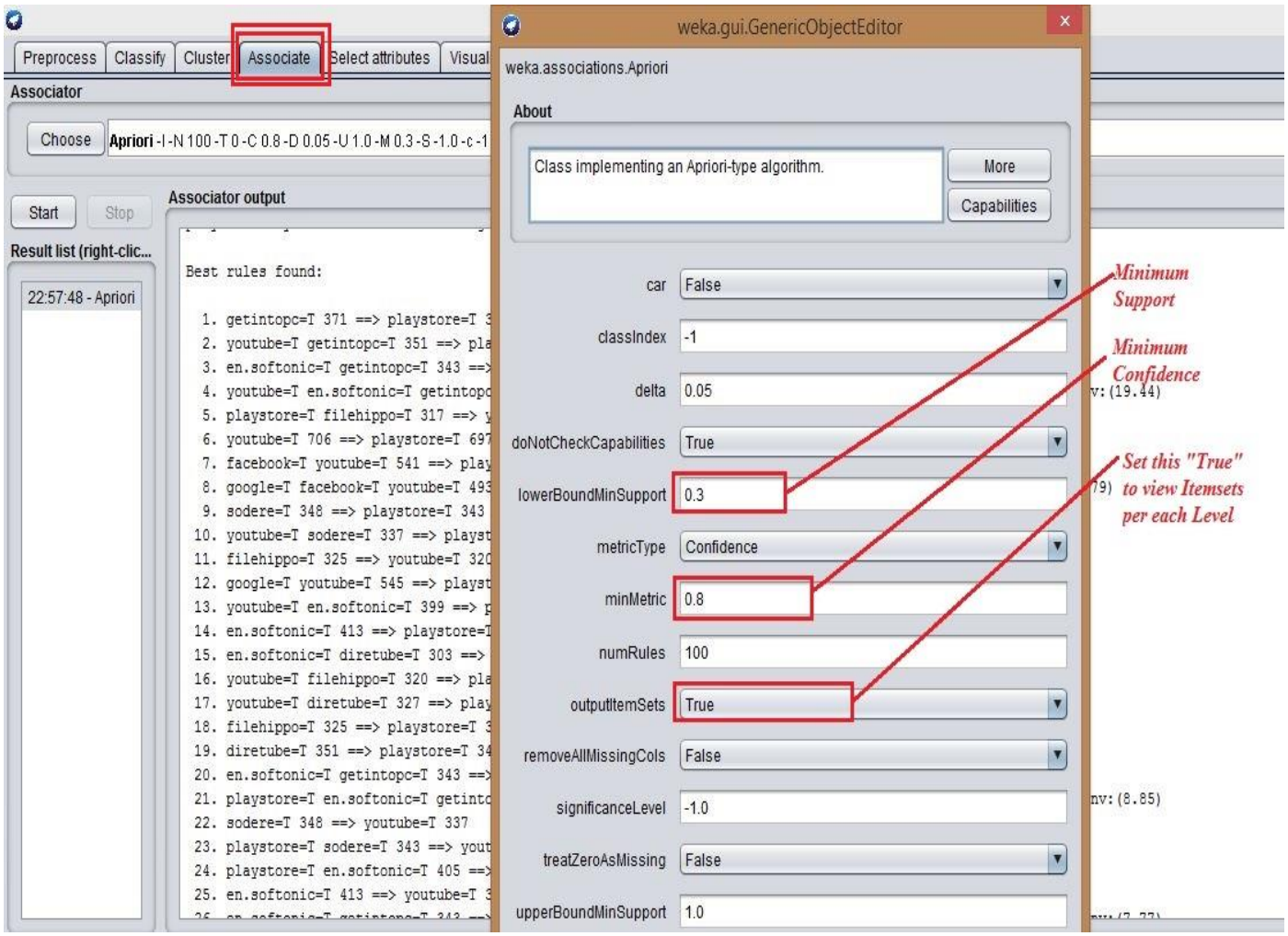

Figure 3.5: Set minimum support and confidence

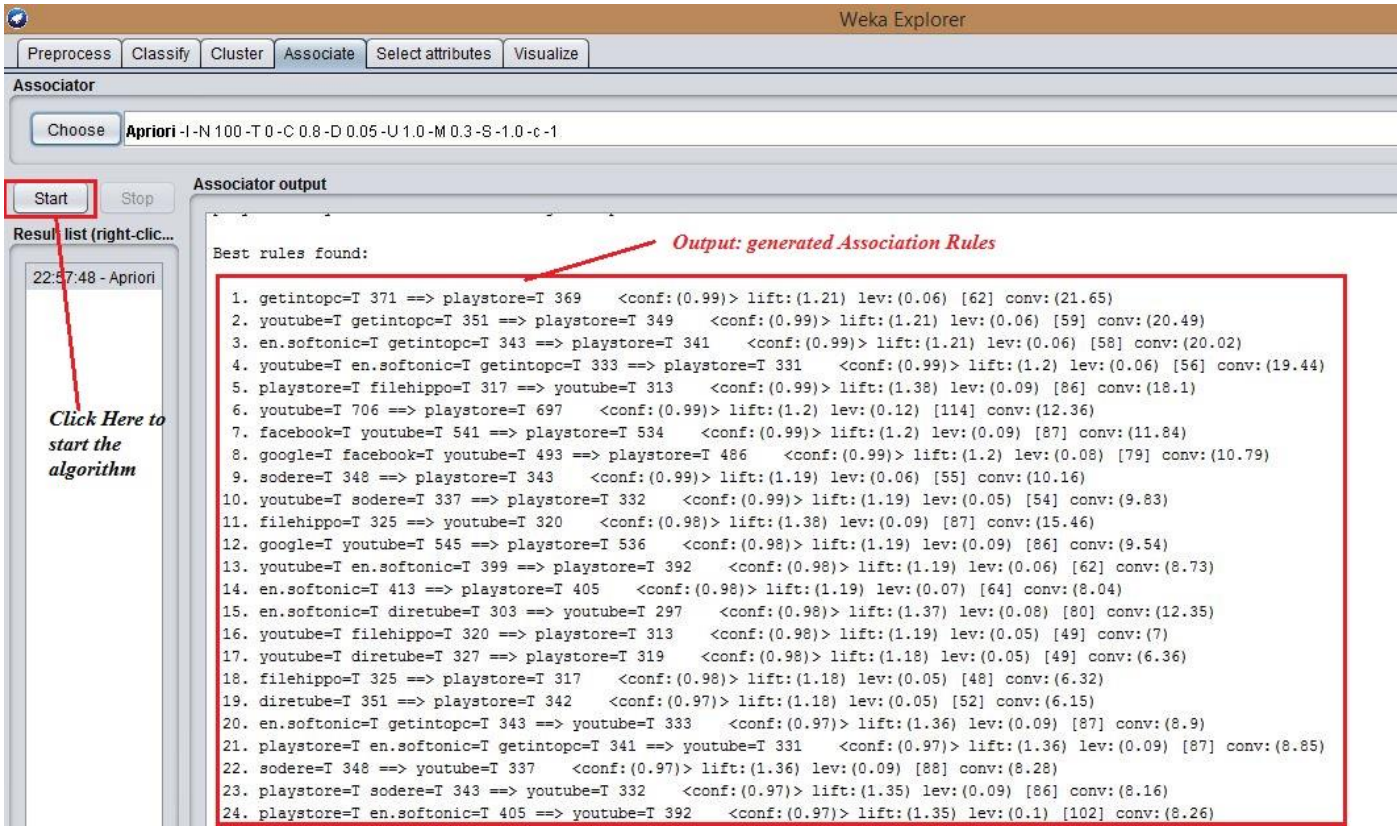

Figure 3.6: Execute the Algorithm

3.6.3. Pattern Analysis

In this phase, separates the interesting and uninteresting pattern from the overall pattern discovered during pattern discovery phase. The discovered access behaviors are evaluated based on the research objective expectation and through validating the association rule mining technique and statistical analysis techniques experiment results.

Chapter Four: Results and Discussions

In this chapter results and interpretation are reported so as to draw the implications of the study. The analyses of data are presented in tabular form and in figures or pictorial presentation. The results are interpreted in detail. The researcher depicts the result and the discussion in parallel.

4.1. Results

4.1.1. Statistical Analysis

4.1.1.1. Users' Trend

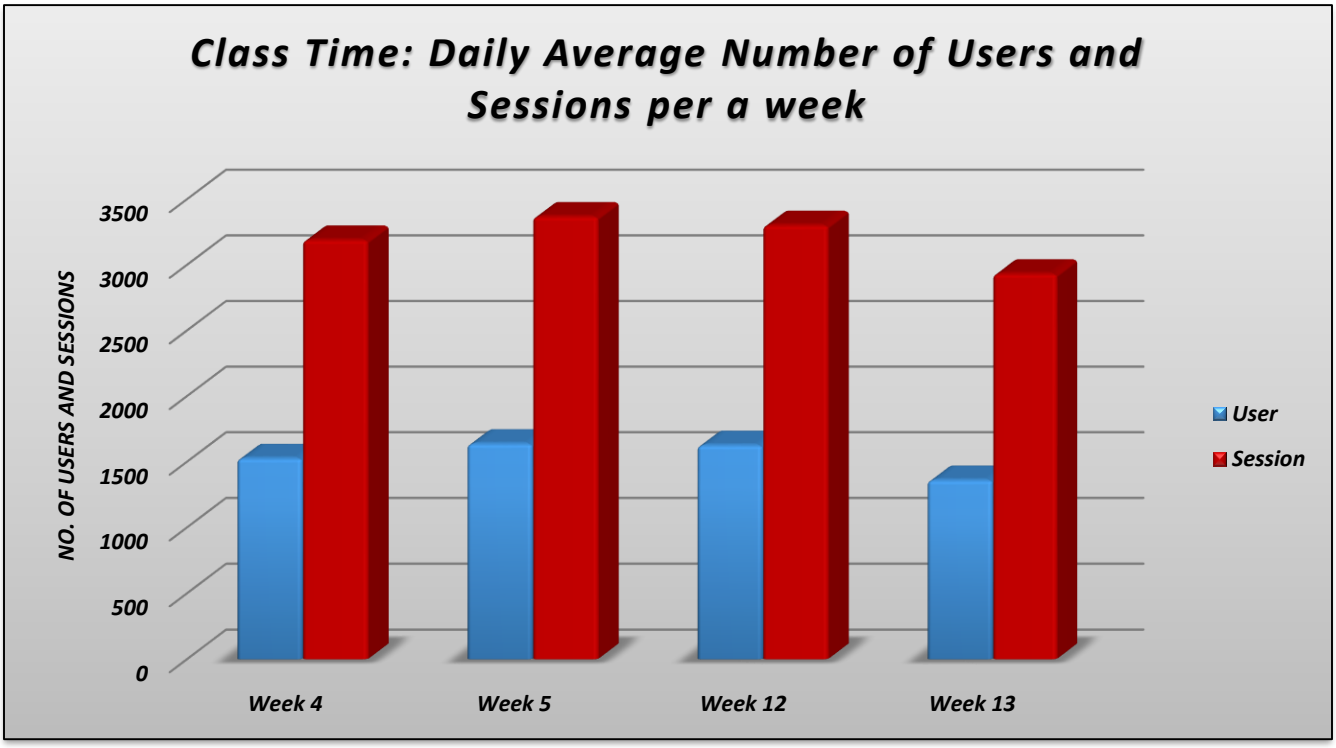

Figure 4.1: Class Time: Daily Average Number of Users and Sessions per a Week

Figure 4.1 shows daily average number of users and sessions per each week during class time. Maximum of 1637 users and 3368 sessions were available. Most of the users have two and above sessions per a day.

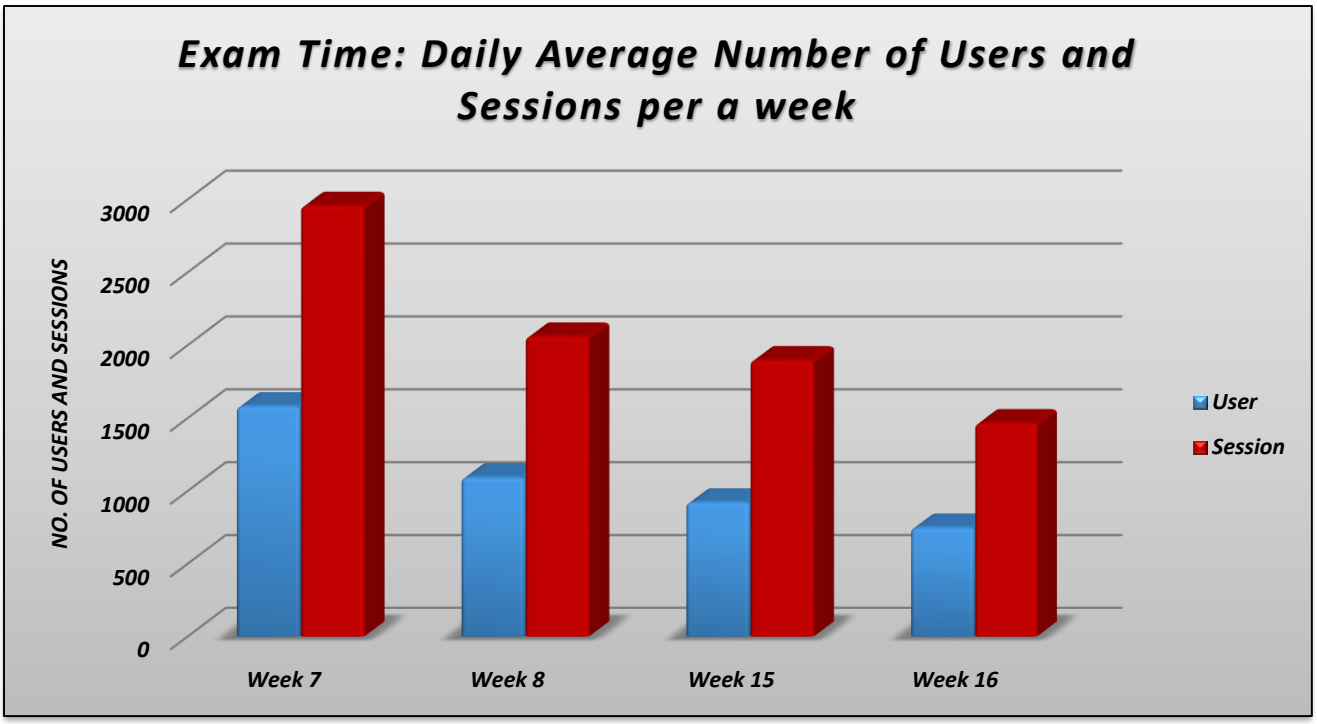

Figure 4.2: Exam Time: Daily Average Number of Users and Sessions per a Week

Figure 4.2 shows daily average number of users and sessions per each week during exam time. Maximum of 1581 users and 2957 sessions were available. Most of the users have two sessions per a day.

As we can have understood from the above two figures, number of users and sessions are decreased during exam time. This is because most of the students spend their time by reading books and handouts in order to prepare for exam and most of the instructors spend their time by invigilating and checking exams.

4.1.1.2. Request's Trend

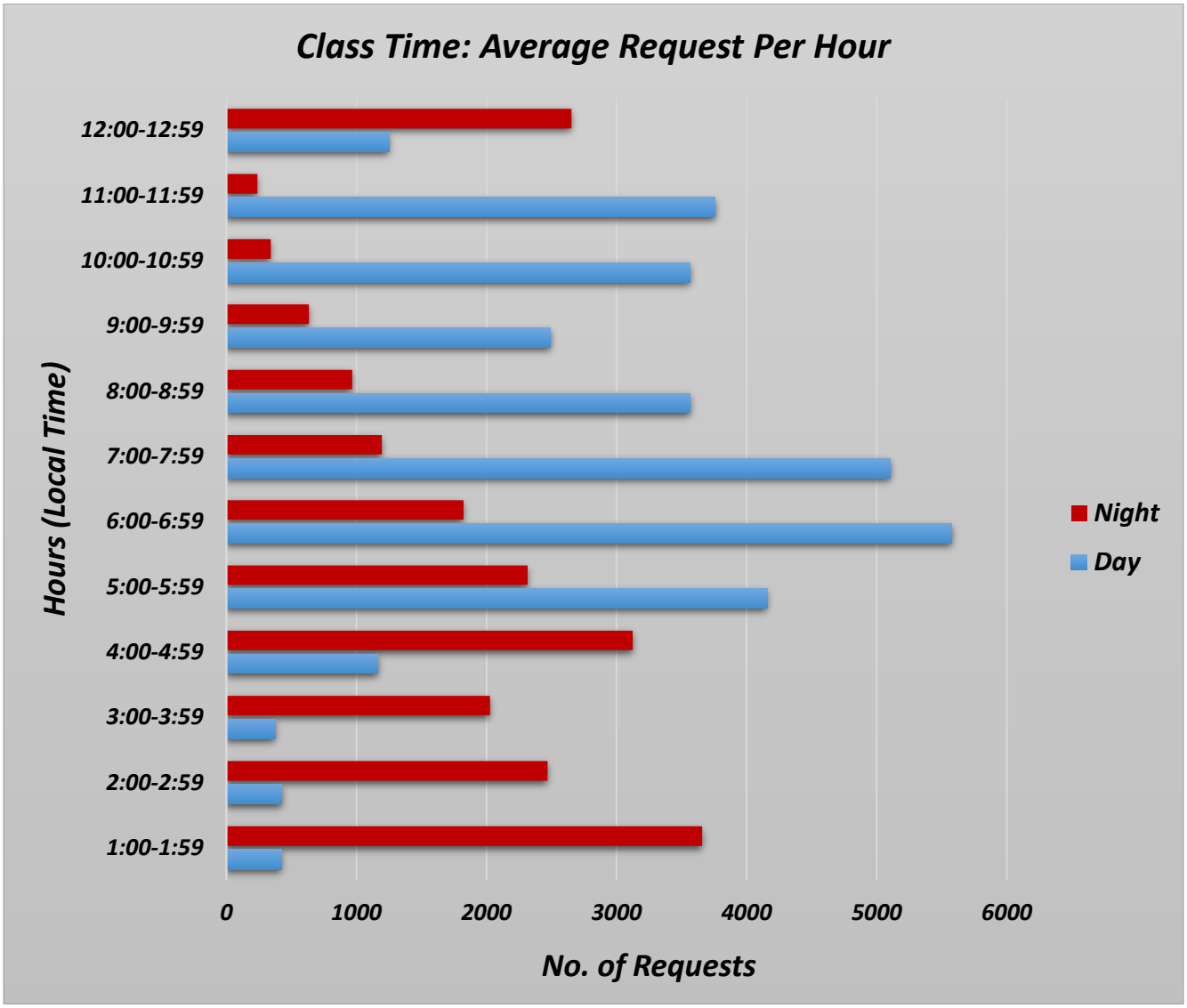

Figure 4.3: Class Time: Average Request per Hour

Figure 4.3 shows average request per hour during class time. Maximum requests has been sent from 12:00 to 2:00 at night and from 6:00 to 8:00 at day time. This means at night 11:00 to 1:00 and at day time 1:00 to 3:00 less request has been sent which is preferred time for the ICT in order to give network maintenance service.

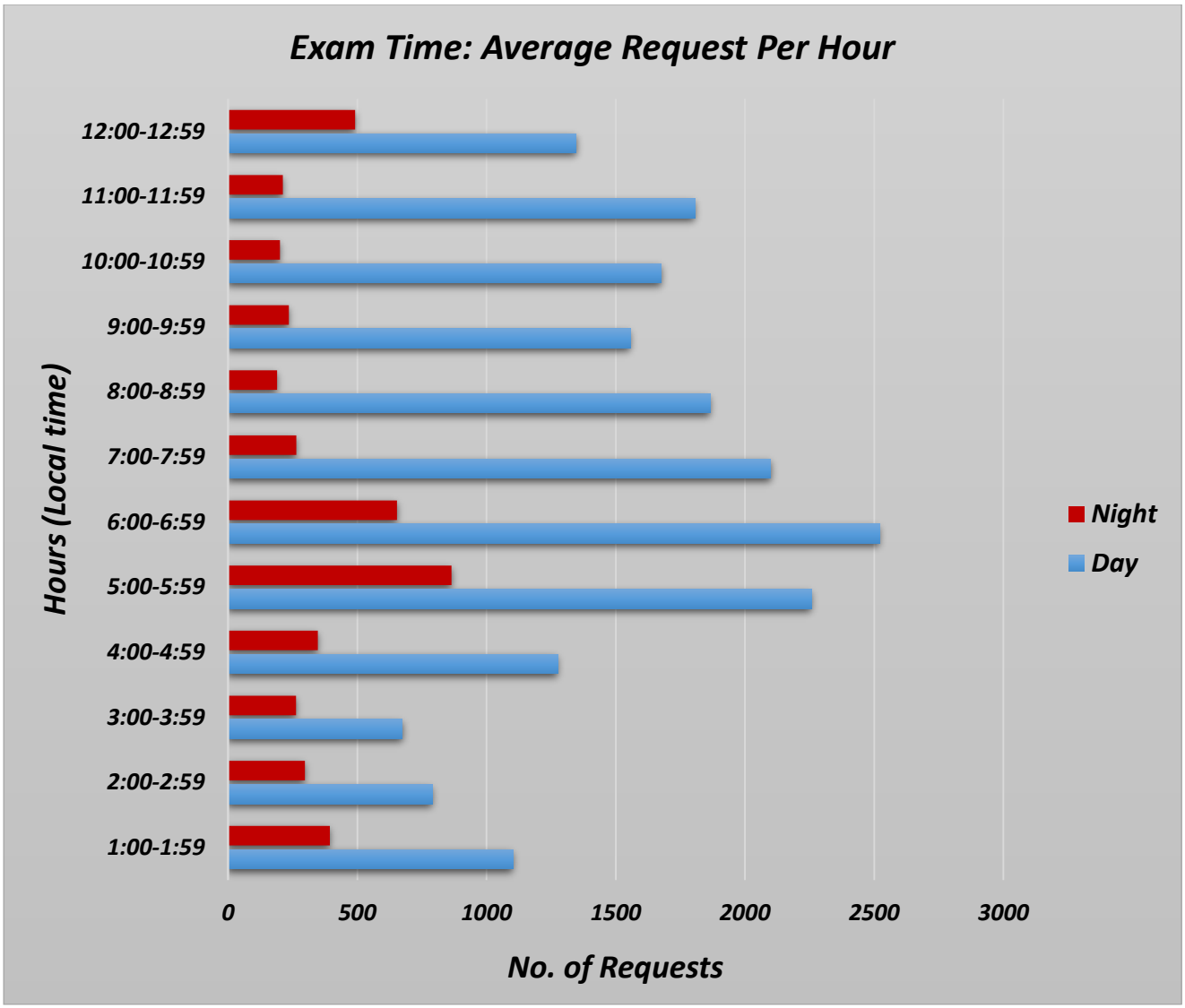

Figure 4.4: Exam Time: Average Request per Hour

Figure 4.4 shows average request per hour during exam time. Maximum requests has been sent from 5:00 to 6:00 at night and from 5:00 to 7:00 at day time. This means at night 8:00 to 11:00 and at day time 3:00 to 5:00 less request has been sent which is preferred time for the ICT in order to give network maintenance service.

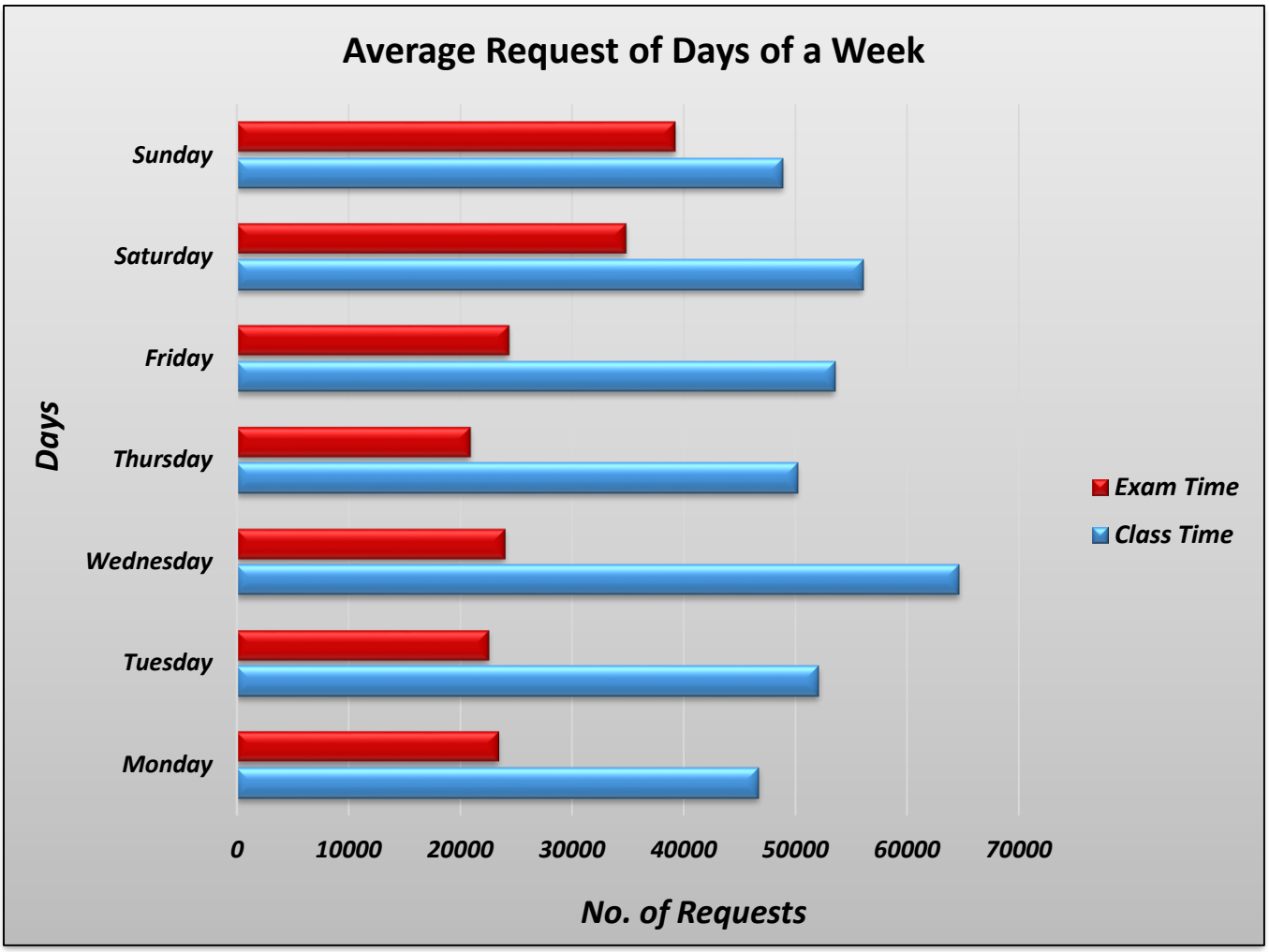

Figure 4.5: Average Request of Days of a Week

Figure 4.5 shows average request of days of a week during both class time and exam time. Maximum request has been sent on Wednesday during class time and on Sunday during exam time. Based on number of request, Wednesday is busy day during class time and Sunday is busy day during exam time.

4.1.1.3. Websites' Trend

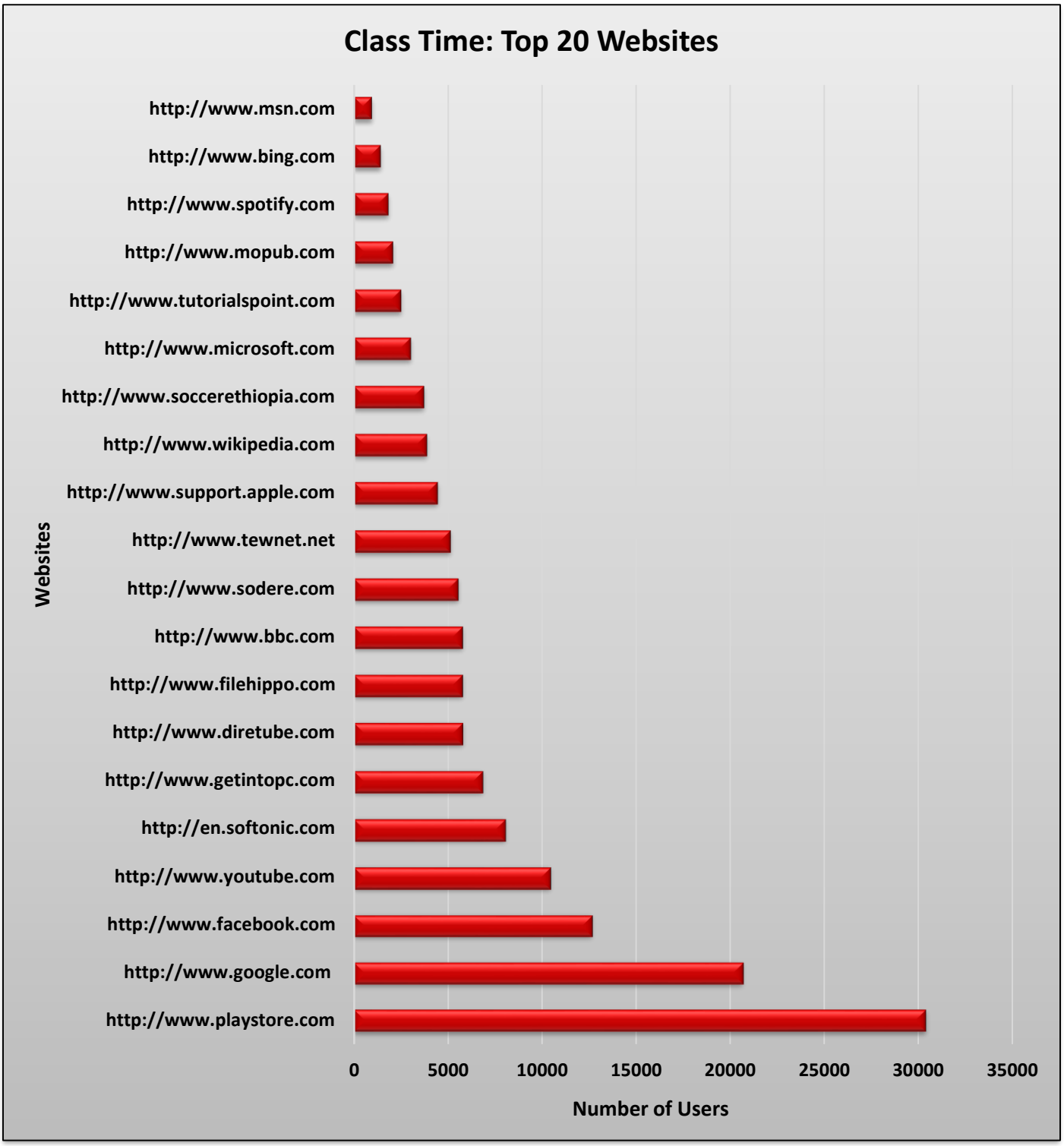

Figure 4.6: Class Time: Top 20 Websites

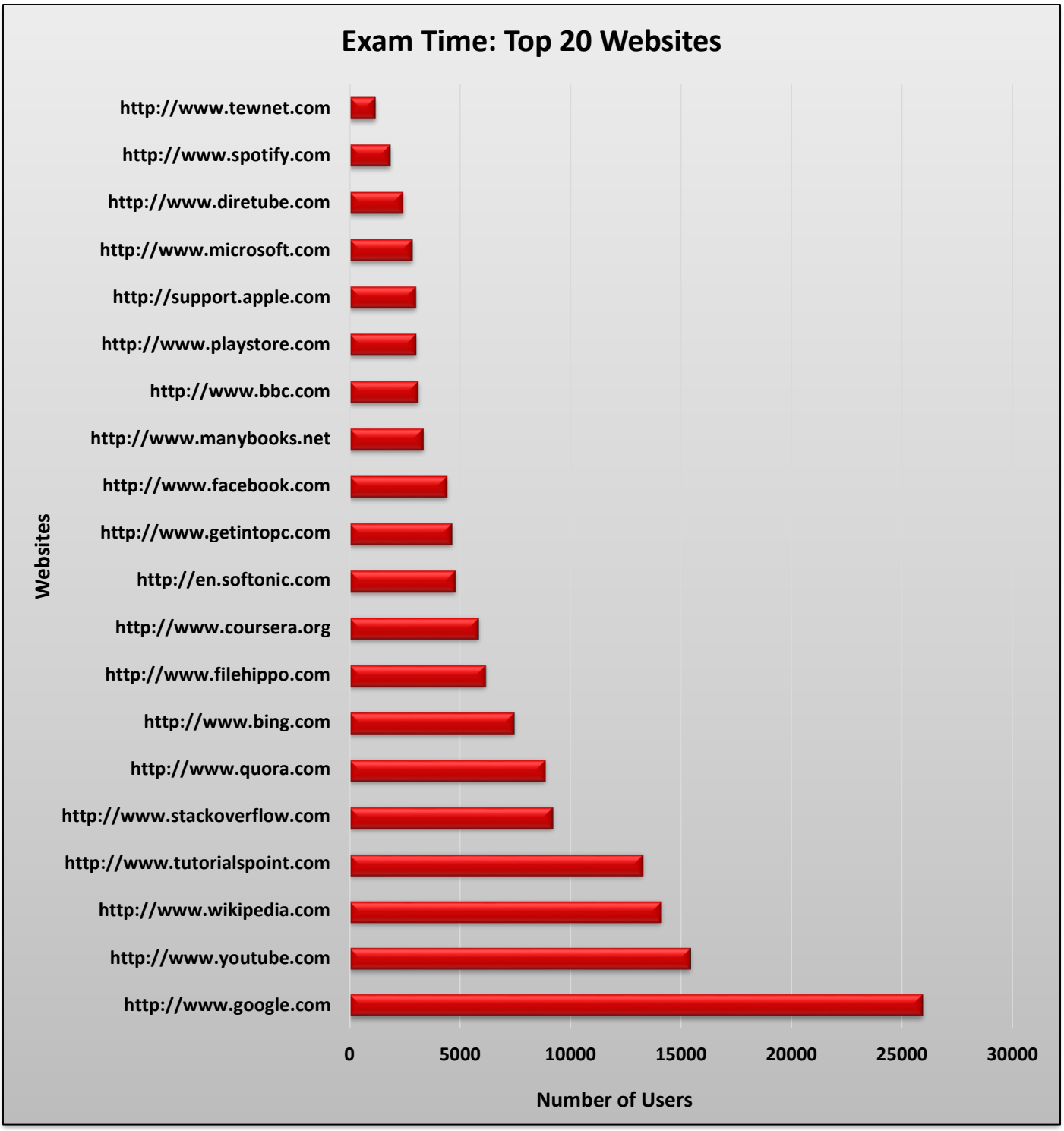

Figure 4.7: Exam Time: Top 20 Websites

The reason why the researcher filtered top 20 websites is to keep the inclusiveness of the selection (i.e. to obtain variety of websites from different category).

Figure 4.6 shows Top 20 most frequently accessed websites during class time based on number of unique users. As we can have understood from the figure, most of the users spend by browsing social network and multimedia websites and downloads various applications from different websites.

Figure 4.7 shows Top 20 most frequently accessed websites during exam time based on number of unique users. Here, since number of students and instructors dominate the campus communities, the user access behavior tends to educational websites.

No.	**Website**	**Category**	**Description**
1.	www.google.com	Search engine	Website used to search anything
2.	www.youtube.com	Multimedia	Website used to search any videos
3.	www.facebook.com	Social Network	Website used communicate with other peoples
4	www.msn.com	Search engine	Website used to search anything
5	www.microsoft.com	Company	Microsoft company website that used to get solutions concerning Microsoft product.
6	www.wikipedia.com	-	Website used get meaning of anything
7	www.tutorialspoint.com	educational	Website used access different course materials
8	www.stackoverflow.com	-	Website used get answers for different random questions
9	www.quora.com	-	Website used get answers for different random questions
10	www.bing.com	Search engine	Website used to search anything
11	www.filehippo.com	Download	Website used to download applications
12	www.coursera.org	Educational	Website used learn courses online and offline
13	www.en.softonic.com	Download	Website used to download applications
14	www.getintopc.com	Download	Website used to download applications
15	www.manybooks.net	Educational	Website used to get e-books

16	www.bbc.com	News	Website used to see world news
17	www.playstore.com	Game	Website used to download mobile games
18	www.support.apple.com	Company	Apple company website used to get apple applications and solutions
19	www.diretube.com	Multimedia	Ethiopian Website used to search videos
20	www.spotify.com	Multimedia	Website used to search music
21	www.tewnet.com	Multimedia	Ethiopian Website used to search video
22	www.sodere.com	Multimedia	Ethiopian Website used to search video
23	www.soccerethiopia.net	Sport	Ethiopian website used to search sport news
24	www.mopub.com	Application	Website enables developer to publish their applications.

Table 4.1: Websites with their category & description

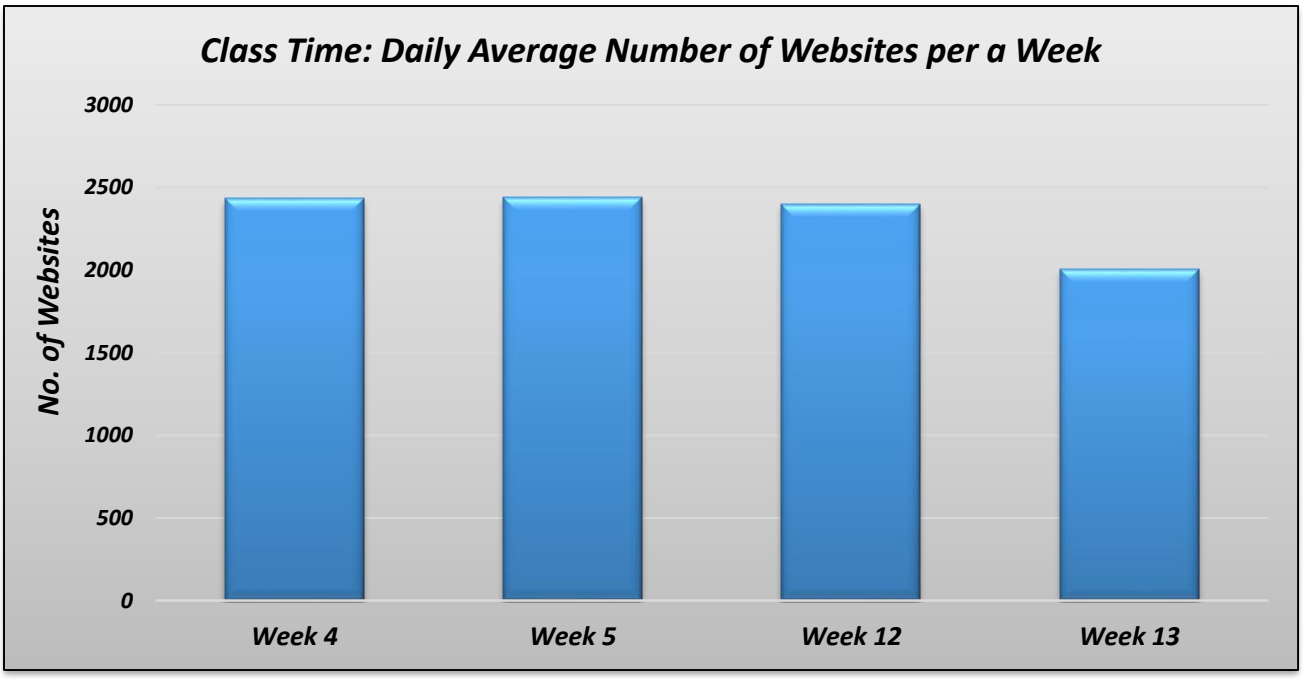

Figure 4.8: Class Time: Daily Average Number of Websites per a Week

Figure 4.8 shows daily average number of websites per each four weeks during class time. Maximum of 2439 websites were available.

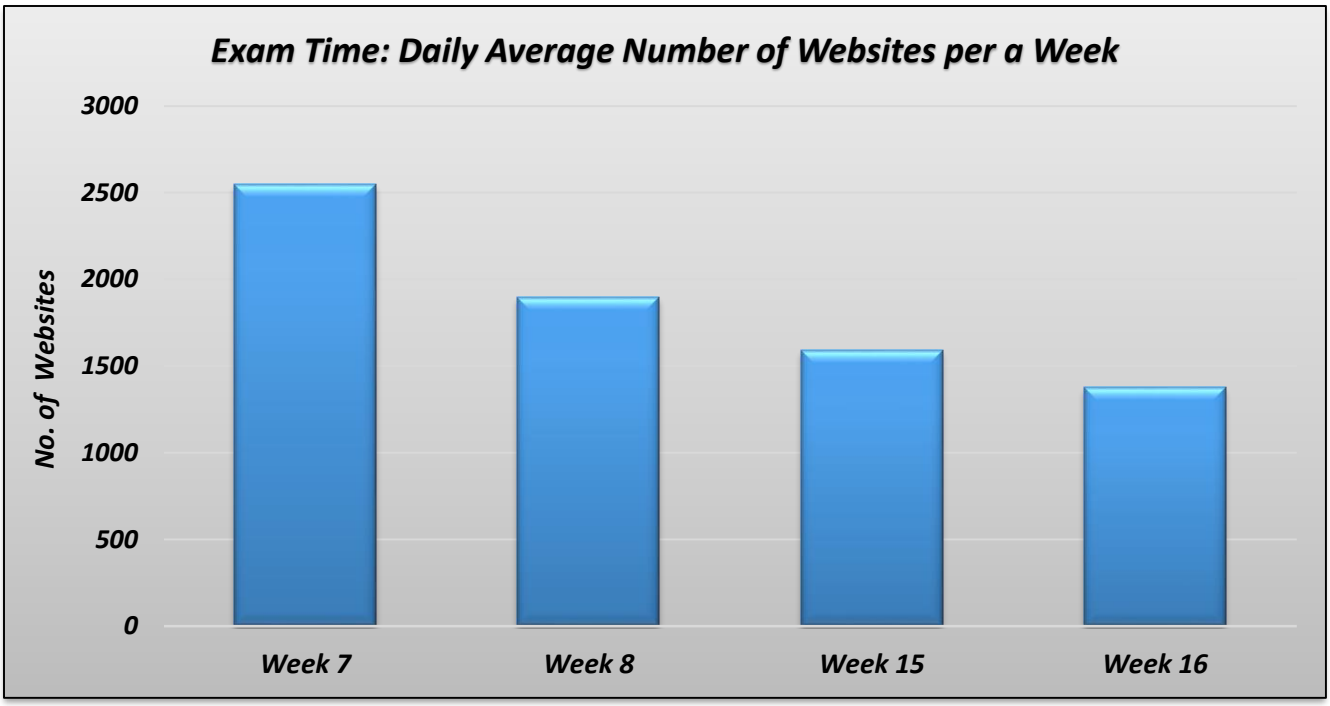

Figure 4.9: Exam Time: Daily Average Number of Websites per a Week

Figure 4.9 shows daily average number of websites per each four weeks during exam time. Maximum of 2548 websites were available.

As we can have understood from the above two figures, number of websites are decreased during exam time. This is because most of the students spend their time by reading books and handouts in order to prepare for exam and most of the instructors spend their time by invigilating and checking exams.

4.1.1.4. Error's Trend

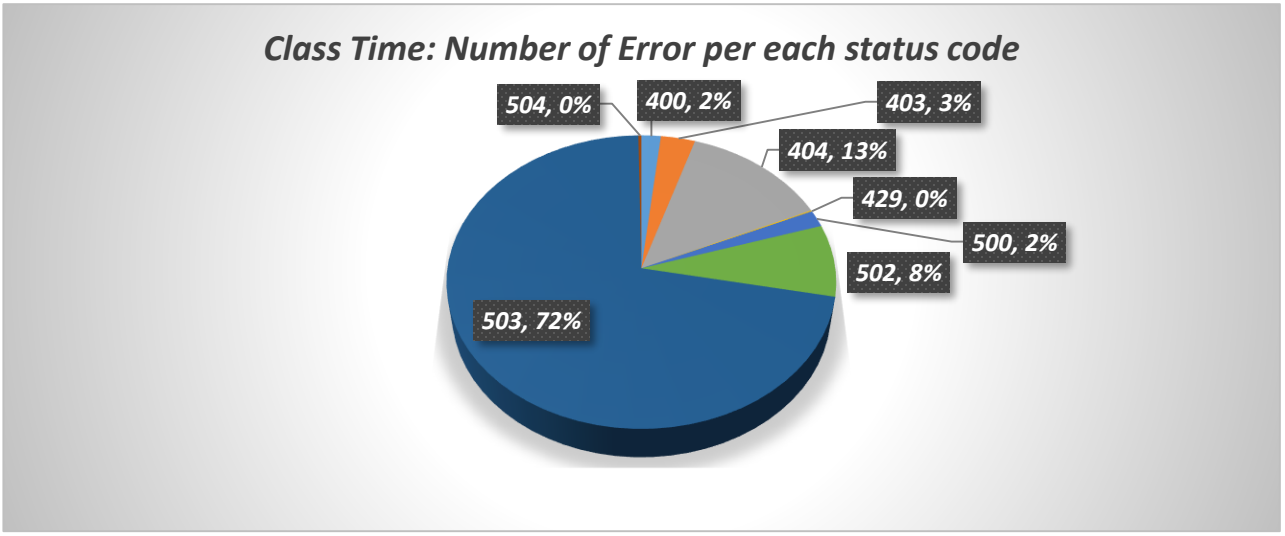

Figure 4.10: Class Time: Number of Errors in terms of Code

Figure 4.10 shows number of errors per each status code during class time. Status code 503 (Service Unavailable *see Table 3.3*) was the most frequently occurred error.

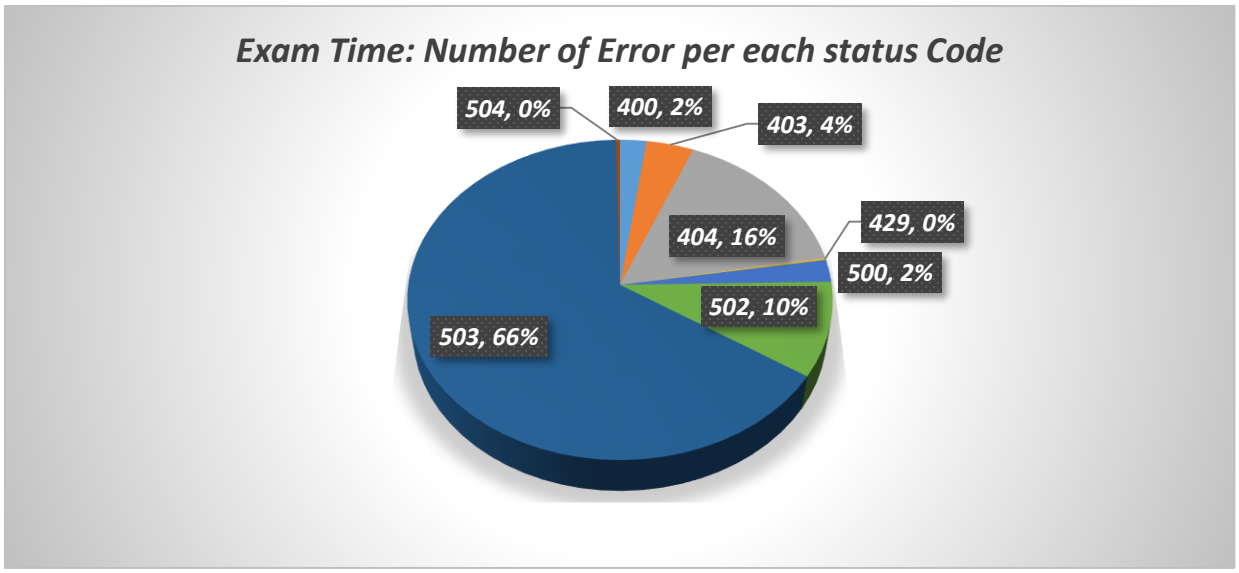

Figure 4.11: Exam Time: Number of Errors in terms of Code

Figure 4.11 shows number of errors per each status code during exam time. Status code 503 (Service Unavailable *see Table 3.3*) was the most frequently occurred error.

4.1.1.5. Summary of statistical analysis

	@ Class Time	**@ Exam Time**
Number of Unique Websites	6901	5255
Number of request before Preprocessing	156,818,572	132,019,903
Number of request after Preprocessing	1,415,827	1,077,149
Number of errors	2,731,381	2,051,979
Average Unique Users per a day	1553	829
Average Session per a day	3564	2532
Average Session per a day	3222	1470
Average Session per a User	2	3
Average Total Time Spent of Users per a day	8	5
Busy Hour	6:00-8:00 [Day] 12:00-2:00 [Night]	5:00-7:00 [Day] 5:00-6:00 [Night]
Busy Day	Wednesday	Sunday
Most Accessed file type	Image	
Most occurs error code	503	
Top accessed Search Engine	www.google.com	
Top accessed Educational Website	www.tutorialspoint.com	
Top accessed Social Network Website	www.facebook.com	
Top accessed Multimedia Website	www.youtube.com	
Top accessed download Website	www.filehippo.com	
Top accessed Ethiopian Website	www.diretube.com	

Table 4.2: Summary of statistical analysis

4.1.2.Association Rules Mining

After statistical analysis is completed, pattern discovery through association rule mining has been done. The fundamental purpose of web usage mining in this study is to discover websites those are accessed frequently together. As showed above in statistical analysis sub-phase top 20 websites during both situations were filtered based on number of unique users. So, the association rule mining is performed on those top 20 websites and the session of unique users who accessed them. Here, the prepared CSV file that contain top 20 websites as column and unique users' sessions as row is imported to WEKA. The researcher set the minimum support=0.35 and confidence=0.8. Finally, the CSV file is executed using Apriori algorithm and gives the belowmined association rules.**24** Rules during class time and **21** Rules during exam time.

4.1.2.1. Class Time

1. facebook ==> google <conf:(0.93)>
2. google ==>facebook<conf:(0.92)>
3. facebook, youtube ==>google<conf:(0.91)>
4. google, youtube ==>facebook<conf:(0.9)>
5. facebook, youtube ==>playstore, google<conf:(0.90)>
6. google, youtube ==>playstore, facebook<conf:(0.89)>
7. getintopc ==>playstore<conf:(0.99)>
8. youtube ==>playstore<conf:(0.99)>
9. facebook, youtube ==>playstore<conf:(0.99)>
10. google, youtube==>playstore<conf:(0.98)>
11. youtube, en.softonic ==>playstore<conf:(0.98)>
12. en.softonic ==>youtube<conf:(0.97)>
13. en.softonic==>playstore, youtube<conf:(0.95)>
14. getintopc ==>youtube<conf:(0.95)>
15. playstore,getintopc ==>youtube<conf:(0.95)>
16. playstore, google ==>facebook<conf:(0.92)>
17. playstore, facebook ==>google<conf:(0.91)>

18. playstore, facebook, youtube ==> google<conf:(0.91)>
19. playstore, google, youtube ==> facebook<conf:(0.91)>
20. playstore ==>youtube<conf:(0.86)>
21. playstore, google ==>youtube<conf:(0.83)>
22. playstore, facebook ==>youtube<conf:(0.82)>
23. playstore, google, facebook ==>youtube<conf:(0.82)>
24. facebook ==>playstore<conf:(0.81)>

4.1.2.2. Exam Time

1. google, tutorialspoint ==>wikipedia<conf:(0.98)>
2. tutorialspoint ==>wikipedia<conf:(0.98)>
3. youtube, tutorialspoint ==>wikipedia<conf:(0.98)>
4. google, quora ==>stackoverflow<conf:(0.94)>
5. quora ==>stackoverflow<conf:(0.94)>
6. youtube, wikipedia ==>tutorialspoint<conf:(0.92)>
7. wikipedia ==>tutorialspoint<conf:(0.92)>
8. google, wikipedia ==>tutorialspoint<conf:(0.91)>
9. stackoverflow ==>google<conf:(0.88)>
10. stackoverflow, quora ==>google<conf:(0.88)>
11. quora ==>google<conf:(0.88)>
12. stackoverflow ==>quora<conf:(0.87)>
13. google, stackoverflow ==>quora<conf:(0.87)>
14. wikipedia ==>google<conf:(0.87)>
15. wikipedia, tutorialspoint ==>googleconf:(0.87)
16. tutorialspoint ==>google<conf:(0.87)>
17. tutorialspoint ==>google, wikipedia<conf:(0.85)>
18. youtube, wikipedia ==>google<conf:(0.84)>

19. quora ==>google, stackoverflow<conf:(0.83)>
20. youtube ==>google<conf:(0.82)>
21. tutorialspoint ==>youtube<conf:(0.8)>

Here, all association rules mined in the case of the two time situations are mentioned with their confidence of being accessed together frequently. For example let's take the last exam time association rule. **tutorialspoint ==>youtubeconf:(0.8)**this means the probability in percentage of a user who access www.tutorialspoint.com will access www.youtube.comis 80%.

4.2. Discussions

The findings from this study suggests the overall internet usage of the campus during class time higher than during exam time. And educational websites and education supportive multimedia websites are mostly accessed during exam time than class time. This fact is obviously expected by the researcher because most of users of the campus are students and instructors and their access behavior will become exam focused. The reason why internet users are decreased is because of students mostly preferred hardcopy in order to prepare for exam and instructors spent most their time by invigilating checking of exams.

This study tried to answer all of the research questions which are raised by the researcher. In statistical analysis experiment questions like how long do the users spent on internet on average per a day? Which day and hour is busy? Which are top 20 frequently accessed Websites in both situations? Are answered. According to the results obtained during experiment, users spent 8 hours, 5 hours in average during class time and exam time respectively. Wednesday and Sunday are busy day during class time and exam time respectively. **6:00- 8:00** Day Time, **12:00-2:00** at Night and **5:00-7:00** Day Time, **5:00-6:00** at Night are busy hours during class time and exam time respectively.

www.google.com, www.youtube.com, www.wikipedia.com, and www.playstore.com, www.google.com, www.facebook.com are among top 3 websites during class time and exam time respectively. In association rule mining experiment questions like which websites are frequently accessed together by the users?, What motivating rule is discovered that could be an input for ASTU ICT? Are answered. According to the results obtained during experiment www.google.com, www.youtube.com&www.facebook.com are frequently accessed together during class time and (www.google.com, www.tutorialspoint.com&www.wikipedia.com), (www.tutorialspoint.com, www.wikipeida.com), (www.qoura.com&www.stackoverflow.com),

(www.youtube.com, www.tutorialspoint.com, &www.wikipedia.com) and (www.google.com&www.stackoverflow.com) are frequently accessed together during exam time.

As ASTU ICT trend of controlling internet usage, sometimes www.youtube.com, www.facebook.com and pornographic websites are denied by the proxy server. But the researcher needs to suggest the ICT to do not deny www.youtube.com during exam time because especially students may use youtube for educational purpose.

The researcher obtained unexpected resulthttp://ak.staticimgfarm.com which is recorded as most frequently accessed website during both time situations. While the researcher checked this website, the website is accessed without the intention of users and it is a malware website that change browsers setting. The users were disturbed by this website. There is a way to stop the website to do not disturb the users, so ICT can resolve this problem for all users by studying about the problem.

[30], tried to investigate users' access behaviors but the researcher tried to generalize one year users' access behavior using only 30 consecutive days (October 01-31, 2015), so it is difficult to compare the results with this study. Because this study based on two basic time situations such as class time and exam time and sampled the users' usage data from four months.

Generally, the study tried to cover all users' access behavior by applying web usage mining on ASTU proxy server access log file and answered the research questions. It will better if the coming researcher gather users' behavior by preparing well questionnaire in addition to proxy server access log file in order to compare and contrast the real users' interest and the actual recorded usage file.

Chapter Five: Conclusions and Recommendations

This chapter presents the conclusions drawn from the findings and the researcher's recommendations on further actions that can be taken and future research areas.

5.1. Conclusions

This research showed users access behavior from different point of view such as from users' trend, request's trend and websites' trend point of view. The proxy server access log data was taken from ASTU ICT Office. The researcher taken 38 GB second semester data of 2018 (2010 local time). The researcher decided to investigate the users' access behavior based on two time situations such as class time and exam time in order to obtain inclusive conclusions. Then strictly followed the three common web usage mining phases such as data preprocessing, pattern discovery and pattern analysis. The access log data has been properly preprocessed in the way it could be appropriate for data mining process. In preprocessing phase data preparation, data cleaning, user identification and session identification has been done using Microsoft Excel 2016. It is because of unavailability of proxy access log data preprocessor. Most of the tools are developed for web server log data. In pattern discovery phase before preparing the data for data mining process, statistical analysis was done using Microsoft Excel 2016. Since the main target is discovering which websites are frequently accessed together, association rule mining was used as data mining techniques. The researcher preferred Apriori algorithm is because it gives output level by level (i.e. most frequently accessed support and confidence of 2 itemsets, 3 itemsets, 4 itemsets etc. based on the support threshold). WEKA is chosen because of its popularity in data mining and easy to use for beginners.

The statistical analysis reported that number of requests, users, sessions and websites are decreased from class time to exam time. The average session per user is two. Based on the average number of requests of days in a week, Wednesday is busy day during class time and Sunday is busy day during Exam time. Based on the average number of requests per hour, during class time night (12:00 to 2:00), day (6:00 to 8:00) and during exam time night (5:00 to 6:00), day (5:00 to 7:00) are busy hours. Based most frequently accessed websites, social network, multimedia and download websites are mostly accessed during class time and educational & related to educational websites (supportive websites) are mostly accessed during exam time.11:00-1:00 at night & 1:00-3:00 at day time during class time and 8:00-11:00 at night & 3:00-5:00 at day time during exam time are hours which have less traffic. So, ICT can use these hours when it needs to provide short network

maintenance. From the association rule mining experiment the researcher obtained www.google.com, www.youtube.com&www.facebook.com are frequently accessed together during class time and (www.google.com, www.tutorialspoint.com&www.wikipedia.com), (www.tutorialspoint.com, www.wikipeida.com), (www.qoura.com&www.stackoverflow.com), (www.youtube.com, www.tutorialspoint.com, &www.wikipedia.com) and (www.google.com&www.stackoverflow.com).

This study answered the research question as for the first research question since the question

Asks how long do the users spent on internet on average per a day? Which answered According to the results obtained during experiment, users spent 8 hours, 5 hours in average during class time and exam time respectively. For the second question which day and hour is busy? Is answered by Wednesday and Sunday are busy day during class time and exam time respectively. Which are top 20 frequently accessed Websites in both situations? Is the third question which is addressed by the researcher is to discover user navigational behavior of websites in ASTU which are more accessed by the users is answered by the statistical analysis by listing in order of access ranking in both situation class and exam time respectively.

The other question is which websites are frequently accessed together by the users? is answered by association rule mining.

What motivating rule is discovered that could be an input for ASTU ICT? is answered. According to the results obtained during association rule mining experiment www.google.com, www.youtube.com&www.facebook.com are frequently accessed together during class time and (www.google.com, www.tutorialspoint.com&www.wikipedia.com), (www.tutorialspoint.com, www.wikipeida.com), (www.qoura.com&www.stackoverflow.com), (www.youtube.com, www.tutorialspoint.com, &www.wikipedia.com) and (www.google.com&www.stackoverflow.com) are frequently accessed together during exam time.

5.2. Recommendations

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

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

- ASTU ICT should have to design well-structured internet usage management policy that strengthen the effective ness of teaching-learning process.
- It will be better to obtain further strong and accurate results and conclusions if ASTU ICT made the network infrastructure installation in the way it could be easy to differentiate the exact visitor of the specific website (who is accessing? Is that student? Instructor? Or administrator staff?).

The following points are recommended for interested future researchers on this kind of thesis.

- The future researcher can conduct web usage mining on one year or more proxy server access log data and implementing Big Data analytics for prediction.
- The future researcher can take the motivation to conduct this kind of study on Ethiopian e-commerce websites by using this thesis as a guideline.

References

1. Kartik B et al.(2010).“Internet Activity Analysis through Proxy Log.”, https://www.researchgate.net .
2. Raymond Kosala and HendrikBlockeel (2000). "Web Mining Research: A Survey." *ACM SIGKDD* Volume 2, Issue 1, https://arxiv.org
3. JaideepSrivastava, PrasannaDesikan and VipinKumar. “Chapter 21 *Web Mining — Concepts, Applications, and Research Directions.”, dmr.cs.umn.edu.*
4. www.wikipedia.org
5. Qiaoyuan Jiang (2003).*Web Usage Mining: Processes and Applications*, PPT, https://www.slideshare.net
6. http://www.internetlivestats.com
7. Jiawei Han and MichelineKamber (2006). *Data mining-Concepts and Techniques*, 2nd ed., USAElsevier Inc, https://www.sciencedirect.com
8. Jiawei Han and MichelineKamber (2012).*Data mining-Concepts and Techniques*, 3rd ed., USAElsevier Inc, https://www.elsevier.com
9. Monika Dhandi and Rajesh Kumar Chakrawarti (2016),“A Comprehensive Study of Web Usage Mining.”,*2016 Symposium on Colossal Data Analysis and Networking, ieee.xplore.ieee.org.*
10. Jaideep, Robert, Mukund, Pang (2000), “Web Usage Mining: Discovery and Applications of UsagePatterns from Web Data”, *ACM SIGKDD,* Volume 1, Issue 2, www.powershow.com
11. RachitAdhvaryu (2008). “A Review Paper on Web Usage Mining and Pattern Discovery.”, *Journal Of Information, Knowledge And Research In Computer Engineering,*Volume – 02, Issue – 02, https://doaj.org/toc/2248-9622
12. Juan D. Velásquez Silva (2006). *“Mining web data: Techniques for understanding the user behavior in the Web”,* PPT, University of Tokyo, Japan, *mate.dm.uba.ar* .
13. www.slideshare.net
14. J. Han and et al.(2004), “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach∗.”, *Data Mining and Knowledge Discovery, hanj.cs.illinois.edu.*
15. Nanhay Singh, Achin Jain, and Ram Shringar Raw (2013). “Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques.”, *International Journal of Data Mining & Knowledge Management Process* , Vol.3, No.4 , https://www.academia.edu

16. KamikaChaudhary and Santosh Kumar Gupta (2013). "Web Usage Mining Tools & Techniques: A Survey.", *International Journal of Scientific & Engineering Research,* Volume 4, Issue 6, https://www.ijser.org
17. Nisha Yadav (2014)."Web usage mining: A Research Area in Web Mining.", *International Journal for Scientific Research & Development,* Vol. 2, Issue 02, www.ijsrd.com
18. MitaliSrivastava and et. al (2014)."Preprocessing Techniques in Web Usage Mining: A Survey", *International Journal of Computer Applications*, Volume 97– No.18, https://research.ijcaonline.org
19. A. Deepa and P. Raajan (2015)."An Efficient Preprocessing Methodology of Log File for Web Usage Mining.", *International Journal of Computer Applications, National Conference on Research Issues in Image Analysis and Mining Intelligence,* https://research.ijcaonline.org
20. J. SoonuAravindan and Dr. K. Vivekanandan (2017)."An Overview of Pre-processing Techniques in Web usage Mining.",*International Journal of Computer Trends and Technology,* Volume 48 Number 1, https://www.researchgate.net .
21. Vinod Kumar et al (2017)."A Brief Investigation on Web Usage Mining Tools", *Saudi Journal of Engineering and Technology,* Vol-2, Iss-1, *scholarsmepub.com.*
22. ChhaviRana (2012)."A Study of Web Usage Mining Research Tools.", *International journal advanced Networking and Applications,* Volume:03 Issue:06, www.ijana.in
23. K. R. Suneetha and Dr. R. Krishnamoorthi (2009)."Identifying User Behavior by Analyzing Web Server Access Log File", *International Journal of Computer Science and Network Security*, Vol.9 No.4, https://www.researchgate.net .
24. NeetuAnand and Prof(Dr.)SabaHilal (2012)."Identifying the User Access Pattern in Web LogData",*International Journal of Computer Science and Information Technologies*, Vol. 3 (2), www.ijcsit.com
25. Mustafa M.H. Ibrahim et al.(2014). "Analysis of Internet Traffic in Educational Network Based on Users' Preferences.", *Journal of Computer Science,* 10 (1), citeseerx.ist.psu.edu.
26. AmitDipchandjiKasliwal and Dr. Girish S. Katkar (2015)."Web Usage mining for Predicting User Access Behavior.", *International Journal of Computer Science and Information Technologies,* Vol. 6 (1), https://pdfs.semanticscholar.org
27. Jan Kerkhofs et al (2001)."Web Usage Mining on Proxy Servers:A Case Study.", *Limburg University Centre,* https://www.researchgate.net.

28. NattapolKiatkumjounwong et al. (2014).“Analysis and Classification of Web Proxy LogsBased on Patterns of Traffic Rates.”, https://*www.ieee.org*
29. Kartik B et al.(2010).“Internet Activity Analysis Through Proxy Log”, https://*www.ieee.org*
30. B.Shiferaw (2016)“*Web usage mining on Proxy Servers to investigate the general access pattern of internet users: The case study of Adama Science and Technology University,* Unpublished” MSc. Thesis Submitted to Adama Science and Technology University
31. Anuragkumar et al. (2017).“A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining.”, *International Journal Of Engineering And Computer Science*, Volume 6 Issue 2, https://www.researchgate.net
32. Anuragkumar, VaishaliAhirwar, and Ravi Kumar Singh, “A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining.”, *International Journal of Engineering And Computer Science,*Volume 6 Issue 2, https://pdfs.semanticscholar.org.
33. MitaliSrivastava and et. al (2014). “Preprocessing Techniques in Web Usage Mining: A Survey”, *International Journal of Computer Applications*, Volume 97– No.18, https://research.ijcaonline.org .
34. J. Kerkhofs, K. Vanhoof (2001). “Web Usage Mining on Proxy Servers: A Case Study”, *LimburgUniversity Centre*, *citeseerx.ist.psu.edu*
35. C. kumar and P. Bhargavik (2013), “Analysis of Web Server Log by Web Usage Mining for Extracting Users Patterns,” *International Journal of Computer Science Engineering and Information Technology Research*, Vol. III, No. 2 , https://www.researchgate.net
36. Official Web Site of Adama Science and Technology University," Adama Science and Technology University, 18 may 2019. [Online]. Available: http://www.astu.edu.et .
37. [Accessed 24 December 2014].AkshayShenoy(2011). "Improving the Performance of a Proxy Server using Web log mining". *Master’s Thesis and Graduate Research*, San Jose State University, *scholarworks.sjsu.edu*
38. http://www.youngzsoft.net/ccproxy/use-proxyserver.htm.
39. Shiva Asadianfam and MasoudMohammadi (2014). “Identify Navigational Patterns of Web Users”, *International Journal of Computer-Aided Technologies*, Vol.1, No.1, https://issuu.com .
40. Manoj Kumar and Mrs. Meenu(2017). “A Survey on Pattern Discovery of Web Usage Mining”, *International Journal of Advance Research, Ideas and Innovations in Technology,* Volume3, Issue1, https://www.ijariit.com

41. Priyanka Patel, MitixaParmar (2014). "A Review on User Session Identification through Web Server Log", *International Journal of Computer Science and Information Technologies*, Vol. 5 (1), *citeseerx.ist.psu.edu*
42. ThanakornPamutha, et.al (2012). "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns", *International Journal of Research and Reviews in Wireless Communications,* Vol. 2, No. 2, https://pdfs.semanticscholar.org
43. Naga Lakshmi, et. al (2013). "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", *International Journal of Innovative Technology and Exploring Engineering*, Volume-2, Issue-4, www.jatit.org
44. Om Kumar C. U. & P. Bhargavi (2013). "Analysis Of Web Server Log By Web Usage Mining For Extracting Users Patterns", *International Journal of Computer Science Engineering and Information Technology Research*, Vol. 3, Issue 2, https://www.researchgate.net
45. SuhasiniParvatikar and Bharti Joshi.(2014). "Analysis of User Behavior through Web Usage Mining", *International Journal of Computer Applications, International Conference on Advances in Science and Technology,* https://research.ijcaonline.org .
46. Ashwiniladekarand PoojaPawar (2015). "Web Log Based Analysis of User's Browsing Behavior", *International Journal of Computer Science and Information Technologies*, Vol. 6 (2), *ijarcet.org*
47. Mohammed Hamed Ahmed Elhiber and Ajith Abraham (2013). "Access Patterns in Web Log Data: A Review", *Journal of Network and Innovative Computing*, Volume 1, *his02.softcomputing.net.*
48. AbhishekChauhan and Dr.SandhyaTarar (2016). "Prediction of User Browsing Behavior Using Web Log Data", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, Issue 5, https://www.academia.edu
49. Yi-Hung Wu and Arbee L. P. Chen (2002). "Prediction of Web Page Accesses by Proxy Server Log", *World Wide Web: Internet and Web Information Systems,* https://link.springer.com
50. Nikhil Kumar Singh, Deepak Singh Tomar and BholaNath Roy (2010). "An Approach to Understand the End User Behavior through Log Analysis", *International Journal of Computer Applications*, Volume 5– No.11, *citeseerx.ist.psu.edu*
51. Xiu-yuZhong (2011). "The research and application of web log mining based on the platform weka", *Advanced in Control Engineeringand Information Science,* https://www.sciencedirect.com

52. Mrs. M.Kavithaand Ms.S.T.TamilSelvi (2016). "Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons", *International Journal of Computer Science Trends and Technology*, Volume 4 Issue 4, www.ijcstjournal.org
53. KamikaChaudhary, Santosh Kumar Gupta (2013). "Web Usage Mining Tools & Techniques: A Survey", *International Journal of Scientific & Engineering Research*, Volume 4, Issue 6, https://www.ijser.org
54. C.P. Sumathi, R. PadmajaValli , T. Santhanam (2011), "An Overview of Preprocessing of Web Log Files for Web Usage Mining", *Journal of Theoretical and Applied Information Technology*, Vol. 34 No.2, www.jatit.org
55. M RekhaSundari et al (2016). "A Review on Pattern Discovery Techniques of Web Usage Mining", *Int. Journal of Engineering Research and Applications,* Vol. 4, Issue 9, *core.ac.uk*
56. Wenwu Lou and Hongjun Lu (2002). "Efficient Prediction of Web Accesses on a Proxy Server", *Hong Kong University of Science andTechnology,* https://doi.acm.org .
57. ThanakornPamuthaet al. (2012). "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns", *International Journal of Research and Reviews in Wireless Communications*, Vol. 2, No. 2, https://pdfs.semanticscholar.org
58. Chaitra L Mugali et al (2015). "Pre-Processing and Analysis of Web Server Logs", *International Journal of Innovative Research in Advanced Engineering*, Issue 8, and Volume 2, www.ijetcse.com
59. L.K. Joshila Grace et al. (2011). "Analysis Of Web Logs And Web User In Web Mining", *International Journal of Network Security & Its Applications*, Vol.3, https://arxiv.org
60. VedpriyaDongre and JagdishRaikwal (2015). "An Improved User Browsing Behavior PredictionUsing Web Log Analysis", *International Journal of Advanced Research in Computer Engineering & Technology*, Volume 4 Issue 5, *ijarcet.org*
61. Mr. Tesema, Interviewee, E-learning expert, video conferencing technician at Adama Science and Technology University. [Interview]. 20 March 2019
62. Kenneth Lai and Narciso Cerpa (2014), "Support vs Confidence in Association Rule Algorithms", The Journal of Theoretical and Applied Electronic Commerce Research, https://www.researchgate.net

Appendixes

Appendix (A)

Sample of Proxy Access Log Result codes

STATUS CODES SENT BY THE SERVER

1xx Info

HTTP_INFO – Request received, continuing process

_ 100 Continue – HTTP_CONTINUE

_ 101 Switching Protocols -HTTP_SWITCHING_PROTOCOLS

_ 102 Processing – HTTP_PROCESSING

2xx Success

HTTP_SUCCESS **–** action successfully received, understood, accepted

_ 200 OK – HTTP_OK

_ 201 Created – HTTP_CREATED

_ 202 Accepted – HTTP_ACCEPTED

3xx Redirect

HTTP_REDIRECT – The client must take additional action to complete the request xx Info

_ 301 Moved Permanently – HTTP_MOVED_PERMANENTLY

_ 302 Found – HTTP_MOVED_TEMPORARILY

_ 304 Not Modified – HTTP_NOT_MODIFIED

4xx Client Error

HTTP_CLIENT_ERROR – The request contains bad syntax or cannot be fulfilled

_ 400 Bad Request – HTTP_BAD_REQUEST

_ 401 Authorization Required –HTTP_UNAUTHORIZED

_ 402 Payment Required – HTTP_PAYMENT_REQUIRED

_ 404 Not Found – HTTP_NOT_FOUND

_ 405 Method Not Allowed – HTTP_METHOD_NOT_ALLOWED

5xx Server Error

HTTP_SERVER_ERROR – The server failed to fulfill an apparently valid request.

_ 500 Internal Server Error – HTTP_INTERNAL_SERVER_ERROR

_ 501 Method Not Implemented – HTTP_NOT_IMPLEMENTED

_ 503 ServiceTemporarily Unavailable – HTTP_SERVICE_UNAVAILABLE

505 HTTP Version Not Supported – HTTP_VERSION_NOT_SUPPORTED

Due to space constraints only few status codes are discussed above. These status codes that are sent alongwith the response data is also entered in the log file.

Squid result codes

TCP_HIT

A valid copy of the requested object was in the cache.

TCP_MISS

The requested object was not in the cache.

TCP_REFRESH_HIT

The requested object was cached but *STALE*. The IMS query for the object resulted in "304 not modified".

TCP_REF_FAIL_HIT

The requested object was cached but *STALE*. The IMS query failed and the stale object was delivered.

TCP_REFRESH_MISS

The requested object was cached but *STALE*. The IMS query returned the new content.

TCP_CLIENT_REFRESH_MISS

The client issued a "no-cache" pragma, or some analogous cache control command along with the request. Thus, the cache has to refetch the object.

TCP_IMS_HIT

The client issued an IMS request for an object which was in the cache and fresh.

TCP_SWAPFAIL_MISS

The object was believed to be in the cache, but could not be accessed.

TCP_NEGATIVE_HIT

Request for a negatively cached object, e.g. "404 not found", for which the cache believes to know that it is inaccessible. Also refer to the explanations for *negative_ttl*in your *squid.conf*file.

TCP_MEM_HIT

A valid copy of the requested object was in the cache *and* it was in memory, thus avoiding disk accesses.

TCP_DENIED

Access was denied for this request.

NONE

Seen with errors and cachemgr requests.

The following codes are no longer available in Squid-2:

ERR_*

Errors are now contained in the status code.

Hierarchy Codes

The following hierarchy codes are used with Squid-2:

NONE

For TCP HIT, TCP failures, cachemgr requests and all UDP requests, there is no hierarchy information.

DIRECT

The object was fetched from the origin server.

SIBLING_HIT

The object was fetched from a sibling cache which replied with UDP_HIT.

PARENT_HIT

The object was requested from a parent cache which replied with UDP_HIT.

Sources

1. https://www.tenon.com/support/webten/papers/squidlog.shtml

2. http://wiki.squid-cache.org/SquidFaq/SquidLogs

3. http://www.comfsm.fm/computing/squid/FAQ-6.html

4. http://etutorials.org/Server+Administration/Squid.+The+definitive+guide/Chapter+13.+Log+Files/13.2+access.log/

Appendix (B)

Weka Association rule discovery outputs using Apriori Algorithm

1. Experiment one: Class time dataset

=== Run information ===

Scheme: weka.associations.Apriori -I -N 30 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.35 -S -1.0 -c -1

Relation: Pattern_Discovery_CT-weka.filters.unsupervised.attribute.Remove-R1

Instances: 988

Attributes: 20

playstore

google

facebook

youtube

en.softonic

getintopc

diretube

filehippo

bbc

sodere

tewnet

support.apple

wikipedia

soccerethiopia

microsoft

tutorialspoint

mopub

spotify

bing
msn

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

Apriori
=======
Minimum support: 0.35 (346 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 8

Large ItemsetsL(1):
playstore=T 815
google=T 813
facebook=T 802
youtube=T 706
en.softonic=T 413
getintopc=T 371
diretube=T 351
sodere=T 348

Size of set of large itemsetsL(2): 10

Large ItemsetsL(2):

playstore=T google=T 644

playstore=T facebook=T 651

playstore=T youtube=T 697

playstore=T en.softonic=T 405

playstore=T getintopc=T 369

google=T facebook=T 744

google=T youtube=T 545

facebook=T youtube=T 541

youtube=T en.softonic=T 399

youtube=T getintopc=T 351

Size of set of large itemsetsL(3): 6

Large ItemsetsL(3):

playstore=T google=T facebook=T 594

playstore=T google=T youtube=T 536

playstore=T facebook=T youtube=T 534

playstore=T youtube=T en.softonic=T 392

playstore=T youtube=T getintopc=T 349

google=T facebook=T youtube=T 493

Size of set of large itemsetsL(4): 1

Large ItemsetsL(4):

playstore=T google=T facebook=T youtube=T 486

Best rules found:

1. getintopc=T 371 ==>playstore=T 369 <conf:(0.99)> lift:(1.21) lev:(0.06) [62] conv:(21.65)
2. youtube=T getintopc=T 351 ==>playstore=T 349 <conf:(0.99)> lift:(1.21) lev:(0.06) [59] conv:(20.49)
3. youtube=T 706 ==>playstore=T 697 <conf:(0.99)> lift:(1.2) lev:(0.12) [114] conv:(12.36)
4. facebook=T youtube=T 541 ==>playstore=T 534 <conf:(0.99)> lift:(1.2) lev:(0.09) [87] conv:(11.84)
5. google=T facebook=T youtube=T 493 ==>playstore=T 486 <conf:(0.99)> lift:(1.2) lev:(0.08) [79] conv:(10.79)
6. google=T youtube=T 545 ==>playstore=T 536 <conf:(0.98)> lift:(1.19) lev:(0.09) [86] conv:(9.54)
7. youtube=T en.softonic=T 399 ==>playstore=T 392 <conf:(0.98)> lift:(1.19) lev:(0.06) [62] conv:(8.73)
8. en.softonic=T 413 ==>playstore=T 405 <conf:(0.98)> lift:(1.19) lev:(0.07) [64] conv:(8.04)
9. playstore=T en.softonic=T 405 ==>youtube=T 392 <conf:(0.97)> lift:(1.35) lev:(0.1) [102] conv:(8.26)
10. en.softonic=T 413 ==>youtube=T 399 <conf:(0.97)> lift:(1.35) lev:(0.11) [103] conv:(7.86)
11. en.softonic=T 413 ==>playstore=T youtube=T 392 <conf:(0.95)> lift:(1.35) lev:(0.1) [100] conv:(5.53)
12. getintopc=T 371 ==>youtube=T 351 <conf:(0.95)> lift:(1.32) lev:(0.09) [85] conv:(5.04)
13. playstore=T getintopc=T 369 ==>youtube=T 349 <conf:(0.95)> lift:(1.32) lev:(0.09) [85] conv:(5.02)
14. getintopc=T 371 ==>playstore=T youtube=T 349 <conf:(0.94)> lift:(1.33) lev:(0.09) [87] conv:(4.75)
15. facebook=T 802 ==>google=T 744 <conf:(0.93)> lift:(1.13) lev:(0.09) [84] conv:(2.41)
16. playstore=T google=T 644 ==>facebook=T 594 <conf:(0.92)> lift:(1.14) lev:(0.07) [71] conv:(2.38)
17. google=T 813 ==>facebook=T 744 <conf:(0.92)> lift:(1.13) lev:(0.09) [84] conv:(2.19)
18. playstore=T facebook=T 651 ==>google=T 594 <conf:(0.91)> lift:(1.11) lev:(0.06) [58] conv:(1.99)
19. facebook=T youtube=T 541 ==>google=T 493 <conf:(0.91)> lift:(1.11) lev:(0.05) [47] conv:(1.96)
20. playstore=T facebook=T youtube=T 534 ==>google=T 486 <conf:(0.91)> lift:(1.11) lev:(0.05) [46] conv:(1.93)
21. playstore=T google=T youtube=T 536 ==>facebook=T 486 <conf:(0.91)> lift:(1.12) lev:(0.05) [50] conv:(1.98)
22. google=T youtube=T 545 ==>facebook=T 493 <conf:(0.9)> lift:(1.11) lev:(0.05) [50] conv:(1.94)
23. facebook=T youtube=T 541 ==>playstore=T google=T 486 <conf:(0.9)> lift:(1.38) lev:(0.13) [133] conv:(3.36)
24. google=T youtube=T 545 ==>playstore=T facebook=T 486 <conf:(0.89)> lift:(1.35) lev:(0.13) [126] conv:(3.1)
25. playstore=T 815 ==>youtube=T 697 <conf:(0.86)> lift:(1.2) lev:(0.12) [114] conv:(1.95)
26. playstore=T google=T 644 ==>youtube=T 536 <conf:(0.83)> lift:(1.16) lev:(0.08) [75] conv:(1.69)
27. playstore=T facebook=T 651 ==>youtube=T 534 <conf:(0.82)> lift:(1.15) lev:(0.07) [68] conv:(1.57)
28. playstore=T google=T facebook=T 594 ==>youtube=T 486 <conf:(0.82)> lift:(1.14) lev:(0.06) [61] conv:(1.56)
29. facebook=T 802 ==>playstore=T 651 <conf:(0.81)> lift:(0.98) lev:(-0.01) [-10] conv:(0.92)

2. Experiment Two: Exam time dataset

=== Run information ===

Scheme: weka.associations.Apriori -I -N 30 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.35 -S -1.0 -c -1

Relation: Pattern_Discovery_ET-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1035

Attributes: 20

google

youtube

wikipedia

tutorialspoint

stackoverflow

quora

bing

filehippo

coursera

en.softonic

getintopc

facebook

manybooks

bbc

playstore

support.apple

microsoft

diretube

spotify

tewnet

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

Apriori
=======

Minimum support: 0.35 (362 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 9

Large ItemsetsL(1):
google=T 845
youtube=T 791
wikipedia=T 553
tutorialspoint=T 516
stackoverflow=T 483
quora=T 449
bing=T 416
filehippo=T 405
coursera=T 390

Size of set of large itemsetsL(2): 10

Large ItemsetsL(2):
google=T youtube=T 652
google=T wikipedia=T 482

google=T tutorialspoint=T 447
google=T stackoverflow=T 427
google=T quora=T 395
youtube=T wikipedia=T 441
youtube=T tutorialspoint=T 414
youtube=T stackoverflow=T 376
wikipedia=T tutorialspoint=T 506
stackoverflow=T quora=T 422

Size of set of large itemsetsL(3): 4

Large ItemsetsL(3):
google=T youtube=T wikipedia=T 371
google=T wikipedia=T tutorialspoint=T 439
google=T stackoverflow=T quora=T 373
youtube=T wikipedia=T tutorialspoint=T 404

Best rules found:

1. google=T tutorialspoint=T 447 ==>wikipedia=T 439 <conf:(0.98)> lift:(1.84) lev:(0.19) [200] conv:(23.13)
2. tutorialspoint=T 516 ==>wikipedia=T 506 <conf:(0.98)> lift:(1.84) lev:(0.22) [230] conv:(21.85)
3. youtube=T tutorialspoint=T 414 ==>wikipedia=T 404 <conf:(0.98)> lift:(1.83) lev:(0.18) [182] conv:(17.53)
4. google=T quora=T 395 ==>stackoverflow=T 373 <conf:(0.94)> lift:(2.02) lev:(0.18) [188] conv:(9.16)
5. quora=T 449 ==>stackoverflow=T 422 <conf:(0.94)> lift:(2.01) lev:(0.21) [212] conv:(8.55)
6. youtube=T wikipedia=T 441 ==>tutorialspoint=T 404 <conf:(0.92)> lift:(1.84) lev:(0.18) [184] conv:(5.82)
7. wikipedia=T 553 ==>tutorialspoint=T 506 <conf:(0.92)> lift:(1.84) lev:(0.22) [230] conv:(5.78)
8. google=T wikipedia=T 482 ==>tutorialspoint=T 439 <conf:(0.91)> lift:(1.83) lev:(0.19) [198] conv:(5.49)
9. stackoverflow=T 483 ==>google=T 427 <conf:(0.88)> lift:(1.08) lev:(0.03) [32] conv:(1.56)
10. stackoverflow=T quora=T 422 ==>google=T 373 <conf:(0.88)> lift:(1.08) lev:(0.03) [28] conv:(1.55)
11. quora=T 449 ==>google=T 395 <conf:(0.88)> lift:(1.08) lev:(0.03) [28] conv:(1.5)
12. stackoverflow=T 483 ==>quora=T 422 <conf:(0.87)> lift:(2.01) lev:(0.21) [212] conv:(4.41)
13. google=T stackoverflow=T 427 ==>quora=T 373 <conf:(0.87)> lift:(2.01) lev:(0.18) [187] conv:(4.4)
14. wikipedia=T 553 ==>google=T 482 <conf:(0.87)> lift:(1.07) lev:(0.03) [30] conv:(1.41)
15. wikipedia=T tutorialspoint=T 506 ==>google=T 439 <conf:(0.87)> lift:(1.06) lev:(0.03) [25] conv:(1.37)

16. tutorialspoint=T 516 ==>google=T 447 <conf:(0.87)> lift:(1.06) lev:(0.02) [25] conv:(1.35)
17. tutorialspoint=T 516 ==>google=T wikipedia=T 439 <conf:(0.85)> lift:(1.83) lev:(0.19) [198] conv:(3.53)
18. youtube=T wikipedia=T 441 ==>google=T 371 <conf:(0.84)> lift:(1.03) lev:(0.01) [10] conv:(1.14)
19. quora=T 449 ==>google=T stackoverflow=T 373 <conf:(0.83)> lift:(2.01) lev:(0.18) [187] conv:(3.43)
20. youtube=T 791 ==>google=T 652 <conf:(0.82)> lift:(1.01) lev:(0.01) [6] conv:(1.04)
21. tutorialspoint=T 516 ==>youtube=T 414 <conf:(0.8)> lift:(1.05) lev:(0.02) [19] conv:(1.18)