DEBRE BERHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

WEB DATA ANALYSIS TO DISCOVER WEB USAGE NAVIGATIONAL BEHAVIOUR FOR WOLKITE UNIVERSITY INTERNET USERS'

BY

ABUNU TESFAW

ADVISOR: SOLOMON DEMISSIE (Ph. D)

DEBRE BERHAN UNIVERSITY, ETHIOPIA

JUNE 2022

DEBRE BERHAN UNIVERSITY
COLLEGE OF COMPUTING
DEPARTMENT OF INFORMATION SYSTEMS

WEB DATA ANALYSIS TO DISCOVER WEB USAGE NAVIGATIONAL BEHAVIOUR FOR WOLKITE UNIVERSITY INTERNET USERS'

A Thesis Submitted to the Department of Information Systems of Debre Berhan University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Systems.

BY

ABUNU TESFAW

Debre Berhan University, Ethiopia

June 2022

DEBRE BERHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

WEB DATA ANALYSIS TO DISCOVER WEB USAGE NAVIGATIONAL BEHAVIOUR FOR WOLKITE UNIVERSITY INTERNET USERS'

BY

ABUNU TESFAW

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Solomon Demissie (Ph.D)	Advisor	_____	_____
_____	External examiner	_____	_____
_____	Internal examiner	_____	_____
_____	Chair person	_____	_____

Declaration

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

ABUNU TESFAW

June 2022

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

Solomon Demissie (Ph. D)

June 2022

ACKNOWLEDGEMENT

First and foremost, praises and thanks to God, the Almighty, for His showers of blessings throughout my research work to complete the research successfully. I would like to express my deep gratitude to my advisor Solomon Demissie (Ph. D), without his valuable guidance, support, and continuous follow-up; this research would never have been possible.

Secondly, I would like to thank WKU ICT staff members, especially Mr. Siraji (ICT director of directorate), Mr. Bethemichel, network administrator to provide the web log data and supporting every information about the proxy data and valuable information for this study.

Thirdly, I would also like to extend my thanks to my friends who have been supporting advice throughout my work.

Finally, I would like to thank my family, especially my mother for her love, prayers, and sacrifices in strengthening and educating me for my future. I am very much thankful to my wife and my daughters for their love, understanding, prayers, and continuing support to complete this research work.

ABUNU TESFAW

_____

June 2022

Table of Contents

List of Tables

List of Figures

ABSTRACT

The objective of this research is to discover web user navigational behavior for Wolkite University web users. in this study, experimental research has been used as a research design. Sharma's web usage mining process model has been followed to discover web users' behavior. In this study, the dataset is collected from Wolkite University proxy server data with a total of three-month data starting from February 01/ 2021 to April 30/2021. For data cleaning, to extract the URL path Python programming language has been used and to split the VLANs from the IP address MS-Excel 2021 have used for VLAN identification. Since the data is a huge, in addition, Minitab and Python have been used for statistical analysis and association rule mining respectively. To discover association rules FP-growth and Apriori algorithms has been used in this study. From the statistical analysis result, most of the time Facebook, and YouTube websites are the top-level websites accessed by the student. However, in terms of website category Entertainment websites have been accessed by the student as the primary interest, Education websites as the second interest, and social media websites as the third web interest. Whereas in the staff dataset most of the time Gmail, Facebook, and YouTube websites are accessed at the top level. However, in terms of website category, educational websites have been accessed as the primary interest, entertainment websites, social media websites, and email websites as second, third, and fourth web interest by the staff users. in terms of web traffic, some of the VLANs in the student dataset have more web traffics especially VLAN (90,120) have more web traffics as compared to the other. Whereas VLANs such as (78, 81) have low web traffics as compared to the remaining VLANs. On the other hand, in staff VLANs, VLAN (2) have more web traffic as compared to the other whereas VLAN (50) has low web traffic as compared to the other. From the association rule discovery, the FP-growth algorithm shows that entertainment websites and social media websites have been browsed together by the student. Whereas in staff users, email and social media, email and entertainment, entertainment and social media, educational and educational websites have been accessed together by the staff VLAN users. The key challenges in this work include preparing log files due to their enormous, noisy, and complex nature of weblog data due to the existing network VLANs is complex, and it is challenging to identify the requests from which users are submitted and identify their behavior accordingly.

Keywords: web usage mining, association rule mining, VLAN

List of Abbreviations

CGI---------- Common Gateway Interface

CSV----------Comma Delimited Value

FTP----------File Transfer Protocol

HTML---------- Hyper Text Markup Language

HTTP----------Hyper Text Transfer Protocol

ICT----------Information Communication Technology

URL----------Uniform Resource Location

VLAN----------Virtual Local Area Network

WKU----------Wolkite University

WWW---------- World Wide Web

XML----------Extensible Markup Language

CHAPTER ONE
INTRODUCTION

1.1 Background

The World Wide Web is growing rapidly and vast quantities of information are produced due to users' experiences with websites. This growth of the World Wide Web has been the creation of various tools both on the client and on the server-side to extract knowledge from the web. Analyzing this data allows the organization to understand the site user's interest and provide web services according to the interest of their client[1]. However, the rapidly increasing mass of the data that is collected and saved in various and broad data repositories goes well beyond our human understanding capability. Most of this data is typically by the web servers automatically and collected in access or server logs during the regular operation of the organization, either as customer or transaction log data or as part of the World Wide Web process. An organization that works on the World Wide Web must collect useful information or pattern from these enormous stored data to make decisions. Analysis of this server data can also provide useful information on how to structure a website, online search engine, information retrieval, and network management in a better way. A new technique, called data mining, in particular, web mining has been introduced to solve such problems[2].

Today, the World Wide Web is one of the dominant internet traffic components. This explosive growth in web traffic occurred due to various reasons. These are ease of use on the web, the availability of graphical user interfaces to browse the web, editors and support tools available for the creation and publishing of web documents, protocols used for constructing and exchanging web documents, as well as the continuing growth of the internet hosts and users. The web server data is the user logs that are generated on the web server. These logs let the analyst to monitor the behaviors of the users visiting the website and analyze them. The server performance analysis begins with log file collection covering a period of time to make an analysis, and there is a need to collect logs over a certain period to understand traffic trends. A lot of research has been triggered by a fantastic growth in web traffic for improving the World Wide Web[3].

Users typically initiate web traffic through the use of web browsers. The flow of traffic begins with a mouse click, which sends the browser information to a server that uses default rules and

methods to obtain user browser requests. The server then decides what action is required based on these rules. Therefore, web traffic analysis tools are required, these tools manage and categorize traffic and increase the web server's workload handling ability[3].

Generally, analyzing web log files has a magnificent utilization to understand the nature of the web traffic, understand the users' interaction, and enhance or establish a well-structured internet usage management policy that strengthens the effectiveness of the teaching-learning process in the university[4]. Additionally, for commercial websites, this kind of study has an unreplaceable advantage to be productive and profitable, because understanding the behavior of the customer helps the commercial organizations to make acceptable their production based on the need of the clients.

Web mining is an interesting discipline in the domain of data mining where information mining strategies are utilized for extracting data from the web servers. It is the way toward applying knowledge mining techniques to web data to automatically and rapidly extract useful information and also discover interesting patterns. The unpredictability of web mining is based on the unstructured nature of web data. web mining entails analyzing the server logs of a specific website, essentially called web server logs, which include the complete contact list of a specific user when accessing the website. the extracted information from such analysis from server logs is very helpful in almost all web applications[5].

The web server stores and maintains client log files to get feedback about activities, the performance of the server, and problems occurring in the web server. Such log files play a very important role in pattern recognition as analyses of log files help in identifying relationships and Patterns between messages requested from the user[6]. A log data is a file that records the activities of clients on a particular server, this log data can be collected from different sources like web servers, proxy servers, and a web client. Analyzing this data is important for extracting knowledge but a pre-processing task is needed to analyze it. Therefore mining those data can provide useful information for decision-making [7].

According to Vidya [8], web mining can be categorized into three according to the data to be mined. Those are web content mining, web structure mining, and web usage mining. Web content mining is a method of extracting valuable knowledge from the content of web documents.it can provide useful and interesting user needs and contributes to behavior patterns. Also referred to as the scanning and mining of images, text, audio, video, or structured records such as lists and

tables. Web structure mining is the process of discovering and finding the structural information of the website. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Technically, web content mining focuses primarily on the inner document structure, but web structure mining attempts to discover the link structure of hyperlinks at the inter-document level. Web usage mining is the application of data mining techniques to discover and investigate interesting usage patterns or information from the web usage data to understand and provide better services based on the needs of web users through their browsing behaviors at the website.

Essentially, web usage mining allows the web-based organizations to collect interesting information about users' browsing behavior which can be later used for personalizing web content, enhancing system performance, understanding web traffic, identifying marketing strategies, and e-commerce applications, developing adaptive websites, and improving the web server performance in terms of content and bandwidth based on the needs of web users[8].

This study contributes in identifying the trend of internet usage behavior at Wolkite University and to aware of the essentiality of analyzing web server access log files for all web users of Wolkite University. Also, it can be an input for creating a platform that changes the university internet usage trend in a way to is essential for an effective teaching-learning process. Therefore, to understand web users' navigational behaviors at Wolkite University web server log data is used to conduct the current study to describe web user navigational behaviors by applying web usage mining.

Furthermore, the 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 the load of the network by analyzing the accessibility of web users' behavior. Also, it enables the system administrators to inform when there is a slow connection or busy traffic to restrict certain users from accessing some URLs. A massive amount of data is collected from web servers in the form of web-access log files. So, this rich source of information makes for understanding web user surfing behavior for the university internet users. As a result of this, exploring user navigation behavior will provide other universities to adopt, analyze, and understand the needs of their web users and enhance the internet usage for the community of the campus based on the findings of their web users' analysis other universities can establish a well-structured internet usage policy for their web users.

1.2 Background of the study

Wolkite University (WKU) is one of the third-generation higher institutions that have been founded in 2012. It is established to provide and promote higher education learning, research, and outreach programs in the country to ensure the realization of the national vision of reaching the level of middle-income countries by 2020. The University is located in the Southern Nation Nationalities Regional State, in Guraghe zone, 158 km southwest of the capital city, Addis Ababa, on the way to Jima. In November 2009 the late prime minister, his Excellency Mr. Meles Zenawi, laid the foundation stone of the University in a plain landscape that is quite ideal for academic pursuit. It is situated at Gubreye sub-city, 14 km away from Wolkite town, of the gubreye-butajira road. The major link road to the University is a direct route to Wolkite-Jimma, Wolkite-Hossana, and wolkite-Butajira[9].

The university mainly focuses on fields like natural science with an emphasis on biotechnology as indicated with the smallest cellular structural symbol at the center of the left outer ring, and engineering and agro-processing sciences as expressed with the smallest engineering symbol at the center of the right outer ring. Other business and social science academic programs are expressed with the opened knowledge book that is placed in the southern central part of the circle. The collaborative efforts in the academic, research, and community linkage programs produce knowledge that is expressed as the candlelight symbol placed in the most canters in the circle that is sourced from the knowledge book that lightens the development process in the region and nation at large[9].

Currently, the WKU ICT is working on several projects to meet client needs. Information and communication technology (ICT) plays a critical role in the implementation of business processes. Currently, all the campus has complete internet access, both wired and wireless. A well-organized data center has been constructed to securely handle all of the university's data and network components. are used to solve problems with scalability, security, and network management. Broadcast filtering, security, address summarization, and traffic flow control are all provided by routers in VLAN topologies[9]. VLANs enable traffic patterns to be controlled and relocations to be handled rapidly. It enhances the freedom to react to changing network requirements while also simplifying administration[10].

WKU networks have been divided into VLANs to properly manage web users. Specifically, the VLANs are categorized under two circumstances those are student VLANs and staff VLANs. Squid proxy servers have been built to manage all users' access web logs to control every web incoming record.

1.2.1 Information and Communication Technology Unit in WKU

Since the university was launched in 2004 E.C, Information &Communication Technology has been created, with the objectives of creating the state-of-the-art ICT infrastructure and mainstreaming ICT-based teaching, learning, research, consulting, and university administrative services. It is also intended to make the technology via hardware, software, and maintenance of the network sustainable use. To accomplish the above different stated tasks the unit defined five major structured sub-units as software development team, ICT infrastructure development team, ICT training and consultancy, technical support and maintenance, and technology teaching and learning development. The teams provide their functions for the university[11].

Table 1. 1 ICT units and their respective functions[11]

Unit Name	Their Respective Services
Software Development Team	Focus on the development & implementation of different software
ICT infrastructure development	Focus on the installation of network infrastructure
ICT training and consultancy	Delivering & proposing short-term & long-term ICT-related training for the community of the university & outsiders
Technical support and maintenance	Maintaining & troubleshooting different office machines, and computers & supporting users
Technology Teaching &learning development	Managing & delivering different technology-based teaching

1.3 Statement of the problem

The World Wide Web (WWW) is a massive, interconnected, semi-structured, widely dispersed, highly heterogeneous, and hypertext information repository. The internet is still growing. As an information gateway, it is expanding at an astonishing rate[12]. In addition to providing an enormous amount of information, the web is a massive, explosive, diversified, dynamic, and largely unstructured data store that adds to the complexity of handling the information from various perspectives from users, web service providers, and business analysts. The study of such web traffic and the usage of websites by users has become very relevant because of the alarming pace at which the World Wide Web (WWW) is increasing in both the sheer amount of traffic and the complexity of websites every day[4]. Understanding the interests of users is becoming a fundamental need for Website owners to offer better services to their visitors by making adaptive the content and usage, and the structure of the site to their preferences. The analysis of web log files permits the identification of usage patterns of the browsing behavior of users on the web and is exploited in the process of navigational behavior[7].

Web sites, as one of the e-business tools, can meet people everywhere and anywhere around the world. Therefore, it is very important to research the actions of web users to enhance web-based services. Analyzing such a large volume of data will help organizations evaluate consumer lifetime, cross-marketing strategies between goods, and promotional campaign effectiveness, among other things[2]. But, the extreme complexity of the data which is collected and stored on the web has become beyond our human ability for understanding.

Users want to have powerful search tools to quickly and precisely acquire relevant information on the web. The web service providers want to find the best way to predict the habits of the users and customize information to minimize traffic load and the website suitable for the clients. Business analysts want to provide software to learn the needs of users. They all expect tools or strategies to help them fulfill their demands and/or solve the problems regarding the web. Since the growth of information available on the web, various web service tools have been developed to support users to access information from the web. Meanwhile, the current web services were not enough to satisfy the needs of different web users on the web. Therefore, further research needs to be carried out to identify new intelligent techniques and services for web users[13].

In addition to this, web users want to have good internet service, within the least amount of time they want to download web resources. Similarly, service providers want to efficiently offer web resources to web users. The amount of web traffic and the details of all clients' history on the web has a great value to a web-based service as well as bandwidth management[14]. The exponential rate of the web and the growing number of users have led organizations to publish their information on the web and offer sophisticated web-based services such as distance education and online shopping. However, this rapid growth of web information leads difficulties to manage the information. Therefore, further research is conducted to identify new intelligent techniques and services for web users, and to achieve this, web mining becomes an active and popular research area[13].

Wolkite University is one of the higher education institutions in Ethiopia and the institution relies on WWW to enhance the teaching and learning progress. As the researcher tries to investigate the existing practice of WKU the way using WWW services, even though, the institution highly depends on using WWW resources, it does not realize the interest of web users. Therefore, studying web users' behavior is an important task to analyze the needs of users on the web. Also, the university wants to study the web users 'behavior or interest to understand and satisfy the need of web users. Furthermore, domain experts want to manage network bandwidth and develop web usage policies based on the web users 'behavior regarding internet access.

Generally, the research could be discovered users' internet usage status and analyze web traffics, and investigate the interest of web users through study. It also provides a primary source for the process of establishing a well-structured internet usage management policy that strengthens the effectiveness of the teaching-learning process in the university.

There are different researchers to identify and analyze the behaviors of web users.

Yohannes[15], conducted web usage mining for extracting employee internet access patter on the Commercial Bank of Ethiopia. He discovered some useful usage patterns about employee internet access behavior. His final result shows that entertainment and social media websites were accessed during employees' off-duty hours and business sites have been accessed during duty hours. He recommended that future researches need to consider data obtained from the proxy server to identify users 'access and browsing patterns. Another research was also done by Amare[4], on log file analysis to discover user navigational behavior on Adama Science and Technology University web users. He also analyzed the web log file and discovered some

interesting patterns. Amare recommended that further research needs to be done by considering specific web users (who are the web users) through network VLAN which is one of the potential research areas that improves the performance of the web usage mining.

Neha[16], conducted on analyzing users behavior from web access logs using automated log analyzer tool. His primary focus is on identifying top visitors, errors, and general statistics about web users. according to his final result, general statistics, access statistics, activity statistics, visitors and top entry pages have been discovered.

Arvind, and Gupta[17], have conducted a study on the Analysis of Web Server Log Files to Increase the Effectiveness of the Website Using Web Mining tool. the study was focused on analyzing the web server log files of an Educational Institution's Website that is www.davkota.orgto discover web usage behavior of the Website users. the researchers used experimental research methodology and web log expert tool to discover users' behavior. Their main findings were identifying the number of the visited website within a day or a week, identifying the top downloaded files by the users, the top browsers used by the users, and the total number of errors that occurred during the user's access behavior has been discovered.

S. Padmaja, and Ananthi[18], have conducted on Identification of Interested Web Users' Behavior by Analyzing the Web Server Access Log File. They have used a web log analyzer tool to discover user behavior. The researchers mainly focused on the in-depth analysis of Web Log Data of the Vels University server to find information about top errors, website, and potential visitors of the website.

Generally, as to the researcher's knowledge, there is no study conducted to investigate web users' navigational behavior in Wolkite university. Therefore, the aim of this study is to discover web user navigational behavior so as to extract usage pattern behaviors of web users in Wolkite university.

At the end of this research, the study attempts to answer the following research questions:

RQ 1: What are the top frequently accessed websites by both student and staff VLAN users?

RQ 2: What interesting rules and patterns discovered that could be an input for the Internet usage policy for Wolkite University ICT center?

RQ 3: Which tools and algorithms are suitable for web log data preprocessing and web usage pattern discovery?

RQ 4: Which VLAN services have more web traffic in terms of web resource usage?

RQ 5: How to represent web users' navigational behavior of Wolkite University internet users on proxy server data?

1.4 Objective of the study

1.4.1 General Objective

The general objective of the study is to discover web user navigational behavior and extract usage pattern behavior of web users in Wolkite University.

1.4.2 Specific Objectives

- Review different works of literature
- Collect weblog data from the proxy server
- Preprocess web log files to prepare a dataset for web usage mining
- Analyze web usage traffic
- Conduct experiment on user web navigational behaviors log file data
- Extract interesting rules in the weblog data
- Evaluate the interesting rules and patterns

1.5 Scope and Limitation of the study

The scope of this research is limited to explore the usage patterns of Wolkite University web users. Web mining can be divided into three sub-categories such as web content mining, web structure mining, and web usage mining. However, the scope of this research focused on web usage mining particularly discovering web users' navigational behaviors using statistical analysis and association rule mining. In web mining, data can be gathered from web servers, client sites, and proxy servers. Due to all web requests passing through the proxy server, the dataset has been taken from the proxy server for this study. The major challenges involved in this study are preprocessing of log files due to the large, noisy, and complex nature of log records. The data in the proxy

server is categorized according to the existing network infrastructure as student VLANs and staff VLANs. In this study, only wired VLANs have been incorporated. Therefore, wireless VLANs have been out of the scope of this study.

1.6 Significance of the study

This research aims to investigate the web access behavior of WKU internet users. The outcome of this study will be used as guidance and recommendations to support other organizations to understand the reaction of their web users. It is difficult to handle users' internet access without understanding the actions of web users. If the activity of web users is known, the bandwidth is dependent on individual tasks that can be prioritized and controlled. Therefore, identifying which users need access, and for what reason is imperative. Thus, it is useful for network administrators and system managers to research web usage behaviors in university surroundings to strengthen their policies and increase the efficiency of internet services. This research describes the community of web users and their web interests and analyzes web traffic accordingly. As a consequence, the output becomes a bandwidth control input. Web usage mining focuses on techniques that can predict user behavior when the user activities are predicted that the system administrator can provide the web user with the required web service based on their need. Thus, the outcome becomes a recommendation for an internet usage policy.

1.7 Organization of Thesis

The research is organized into six chapters. The **first chapter** describes the introduction to the general study as background, the background of the study, the statement of the problem, the objective of the study, the scope of the study, the limitation of the study, and the significance of the study. **Chapter Two** is about literature review, data mining, web mining techniques and tools, web data source, types of web server logs, web mining processes, web data source, application of web mining and related works have been reviewed in this study. **Chapter three** is about the research methodology. It includes the research design method and web usage mining process model for the study**. Chapter four** deals with data preparation such as data collection, data cleaning, attribute selection, and data categorization. **Chapter five** deals with the experiment on statistical analysis and association rule discovery. **Chapter six** deals with the conclusion and recommendations for future works.

CHAPTER TWO
LITERATURE REVIEW

2.1 Data Mining

The development of Information Technology has generated a large number of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to storing and manipulating this precious data for further decision-making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called a knowledge discovery process, knowledge mining from data, knowledge extraction, or data pattern analysis[19].

Data mining is the process of non-trivial discovery of useful knowledge from implied, previously unknown, and potentially useful information from data in large databases[20]. Therefore, taking this definition as fact, data mining can be called a core element in knowledge discovery, often used synonymously with knowledge discovery. The data is integrated and cleaned so that the relevant data is retrieved. Data mining presents discovered data that is not just clear to data mining analysts but also to domain experts who may use it to derive actionable recommendations. It is an interdisciplinary field, drawing from various areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. It has various application areas such as banking, education, and e-commerce, and also recently which is become popular in new data mining applications such as World Wide Web, spatial data, and multimedia data. Successful applications of data mining include the analysis of genetic patterns, graph mining in finance, expert systems to get proper advice, and consumer behavior in marketing. Traditional data mining uses structured data stored in relational tables, spreadsheets, or flat files in tabular form. With the growth of the Web and text documents, Web mining and text mining are becoming increasingly important and popular[20].

2.2 Web Mining

Web mining is one application of data mining to explore patterns from the World Wide Web. It becomes one of the paramount areas in Computer and Information Sciences because of its direct applications in e-commerce, web analytics, information retrieval, filtering, and web information

systems. In addition to this, web mining is a data mining technique to discover and extract knowledge from web data. Even though web mining applies different data mining techniques, it does not describe a pure application of traditional data mining because of the heterogeneity, semi-structured, and unstructured nature of data on the web. The World Wide Web has a variety of data, which includes text, images, videos, audio, and other forms of data. To analyze these tremendous data and extract meaningful patterns or information and knowledge there is a need to develop some new techniques and tools. Data mining is a process of extracting useful information from a huge data set, when it is applied to web data is called web mining. Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), Machine Learning, etc.[21].

Web mining is the process of extracting and analyzing patterns for knowledge discovery from different available data on the web. The explosive growth of the World Wide Web has resulted in a large amount of data that is now available for any web user. So different types of data have to be managed and organized using different web mining techniques[22]. Nowadays due to the pace of data on the internet combined with its rapid and disordered structure, the World Wide Web has evolved into a network of data without a proper organizational structure. Besides, the growth and heterogeneous nature of web data lead web browsing an intricate task not only for inexperienced users but also for expert users. As a result, when browsing the Web users often feel lost, and overwhelmed by a huge amount of data that continue to enlarge over time. Besides, e-business and web marketing are rapidly developing and providing a variety of services on the web. So analyzing the needs of customers has become an imperative activity for web applications to understand the preference of web users as well as to enhance the usability of web services[23].

With the proliferation of web-based services, a huge amount of user data is collected and stored during their interactions with web applications. Therefore, analyzing such data is important to achieve different goals related to the nature of the considered web applications. These include offering personalized content or services to users, designing marketing strategies, optimizing functionalities of web applications, etc. the analysis of such collected data from the web is essential to discover meaningful patterns from large collections of web data. Generally, the adoption of machine learning techniques reveals to be an appropriate way to analyze and extract useful knowledge from the web. The efforts carried out in this direction have led to the growth of an interesting research area named Web mining which essentially refers to the application of

Data Mining methods to automatically discover and extract knowledge from data generated by the Web[23].

2.2.1Taxonomy of web mining

According to Rajinder Singh Rao[24], Web Mining is broadly divided into three categories such as web content, web structure, and web usage mining.

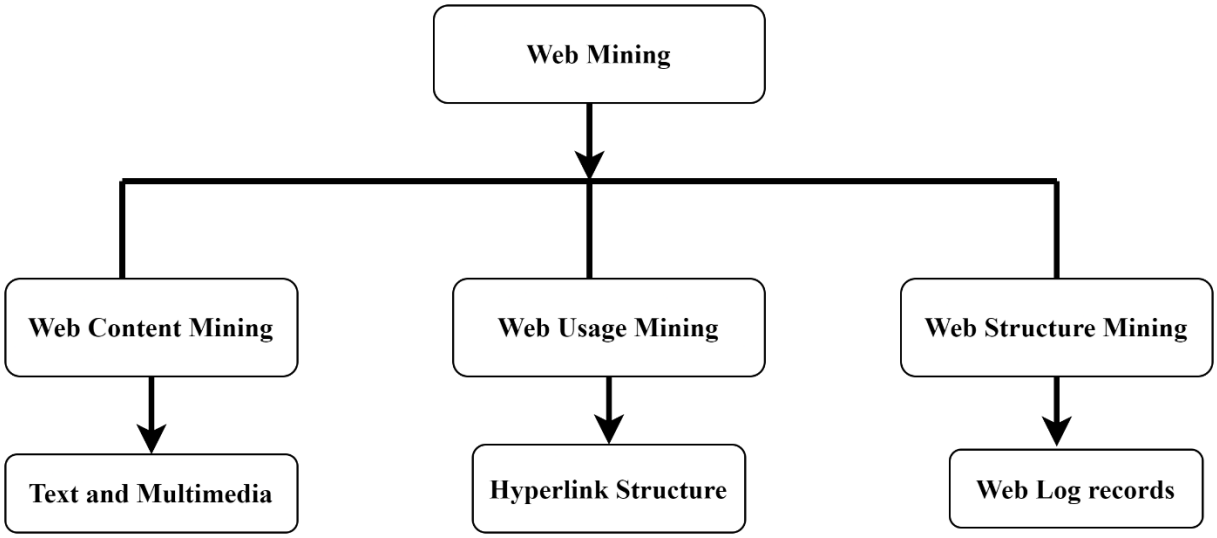


Figure 2. 1 Taxonomy of Web Usage Mining[24]

2.2.1.1 Web Content Mining

Web Content Mining is the process to discover useful knowledge from the contents of Web data. It may consist of text, images, audio, video, or structured records such as lists and tables[24]. Web Content mining refers to attaining useful information from the contents of the webpage using different text-mining techniques. It involves the extraction of structured data/information from web pages, detection, similarity, and integration of data with similar meaning, view extraction from web sources, and concept hierarchy[25]. In web content mining, patterns are extracted from online sources like HTML files, text documents, images, e-books, or e-mail messages. The concept of Web content mining is far broader than searching for any specific term or only keyword extraction or some simple statistics of words and phrases in documents[26]. The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information retrieval)[27].

2.2.1.2 Web Structure Mining

Web structure mining is the process of discovering information from the web to analyze the node and connection structure of a website. Nodes of a graph are web pages and lines are hyperlinks[28]. It deals with the arrangement of hyperlinks inside web pages themselves. Based on the hyperlinks, web structure mining classifies the web pages and generates detailed information. The structure of a typical wave graph consists of web pages as nodes and hyperlinks as edges connecting two related graphs. The main purpose of web structure mining is to generate structural information about web pages and websites. It illustrates the relationship between the user and the web and discovers the link structure of hyperlinks at the inter-document level. It also helps in discovering the structure of a document, which is used in revealing the structure of web pages, and it is possible to compare the web page schemes[21].

According to the type of web structural data, web structure mining can be divided into two kinds. The first kind of web structure mining is extracting patterns from hyperlinks on the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of web structure mining is mining the document structure for the analysis of the tree-like structure of page structures to describe HTML or XML tag usage[28]. The main focus of this research is web usage mining. Usage mining as the name implies focuses on how the users of websites interact with the website, the web pages visited, the order of visits, timestamps of visits, and duration of them.

2.2.1.3 Web Usage Mining

Web usage mining is one application of data mining to analyze and discover interesting patterns or information from various web logs (i.e., web user history). Various weblogs are server logs, client logs, and network logs. This is the process of finding out what users are looking for on the Internet and the usage of web pages.[28]. Web usage mining is the process to predict patterns of the user while the user interacts on the web. It uses a different collection of data from weblogs to discover the patterns of web users' behavior. Weblog records are unformatted text file which contains data like User name, date, time, IP address, status code, etc. whenever the user interacts with a website, the information is recorded and maintained in web servers[25].

While web content and structure mining utilize real or primary data on the web, Web usage mining works on the secondary data such as Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, and bookmark data[4].

2.3 Web Data Source

Any website visitor ‘s activity is stored in a weblog which is called a web log file. The data is stored in different format types in the web log file. A weblog file is automatically generated by the web server whenever a user visits or accesses different web services on a given website. A Weblog is a file in which the server takes the knowledge/data each time a user requests a site from a particular server[24]. There are three kinds of web data sources such as server-level data, client-level data, and proxy-level data.

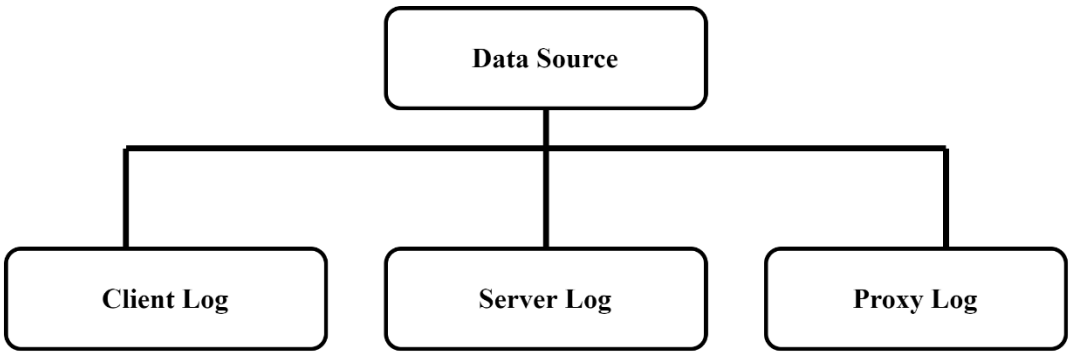


Figure 2. 2 Data Sources for web usage mining[29]

2.3.1 Server Level Data

The first kind of web data source is web server data which is one of the richest and most common sources of web data because they explicitly record a huge amount of data for the browsing behavior of site visitors**.** The data recorded on the web server reflects the access to the website by, multiple users, in chronological order[23].

At the server level, log files can be stored in different formats such as common log or extended log formats. However, the usage data recorded in the web server may not be entirely reliable due to the effect of various levels of catching within the web environment and misinterpretation of the

IP user addresses. The requests for cached web pages were not logged into log files. In reality, if a user accesses the same web page rather than making a new request to the web server, the cached copy is returned to the user. During this, the user request does not reach the web server holding the page and, as a result, the server is unaware of the behavior and access to the page made by the web users. Cache-busting is one solution to this first problem. This requires the use of unique headers, specified either on Web servers or on Web pages, which provide instructions to specify the objects to be cached and the time to cache them[23].

2.3.2 Proxy Level Data

A proxy server is an intervening level of caching that lies between the client browsers and web servers. Proxy caching describes the way to reduce the loading time of a web page as well as the network traffic load both at the server and client sides. The main advantage of proxy caching is to reveal the actual HTTP request from multiple clients to multiple web servers, this can be taken as a valuable source of data characterizing the browsing behavior of a group of anonymous users sharing a common proxy server[30].

2.3.3 Client Level Data

Client-side usage data can be collected through the use of a remote agent (which is implemented using Javascript or Java applets) embedded in web pages. In addition, client data can be collected by modifying the source code of an existing browser to ameliorate its data collection capabilities[30].

The implementation of client-side data collection methods involves user cooperation, either in enabling the functionality of the JavaScript and Java applets or interested to use the modified browser. Those data collection has an advantage over the server-side collection because it improves both the caching and session identification problems. However, Java applets do not perform better than server logs in terms of determining the actual view time of a page. JavaScript consumes little interpretation time but it cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user, single-site browsing behavior[30].

2.4 Types of Web Server Logs

According to[22][24], there are four kinds of web server log files such as access log file, agent log file, error log file, and referrer log file. An access log file is a file that records all incoming requests and information about the web service users[31]. As cited in [34], the access log file is one of the major weblog servers, it will record each click, hit, and access of the users. Some of the captured information about the user has some attributes such as Client IP, Client name, Date and Time, Server site name, and Server IP. Agent log file describes the users' browser, browser version, and operating systems. An error log file is a list of internal errors which records the error found on the websites, particularly when the user clicks on a certain link and the browser does not display the specified page or website and the user receives the error 404 not found. The referrer log file provides the information that a user came from a particular website by using the user's page link[34].

2.4.1 Web log file formats

According to Rathod, Prajapati, and Joshi [35], there are three kinds of log file formats such as:

- W3C (World Wide Web Consortium) Extended log file format
- IIS (Internet Information Services) Log file format
- NCSA (National Center for Supercomputing Application) Log file format

W3C Log File Format is an adaptable ASCII design that has various sorts of fields. These fields can be separated by spaces. Also, time can be recorded in UTC (Coordinated Universal Time). It can be customized by the administrators which can be added or removed based on what information it wants to be recorded. IS log file format: - Microsoft IIS is a non-flexible ASCII group. This organization can record more data than the NCSA design. The IIS log file can provide some additional information such as the client's IP address, client name, Service status code, demand date-time, and the number of bytes got. Besides, it encompasses definite things like the slipped by time, the number of bytes sent to the activity, and the target document. All the fields are finished with a comma. A hyphen fills in as a placeholder for a specific field that has no substantial worth[35].

NCSA log file format is a non-adaptable ASCHI group that is accessible for web destinations. But it is not available for FTP sites. It caches information about the clients' name, remote-host name, time, date, the number of bytes sent by the server, and the status code for HTTP. In addition, time can be recorded as nearby time, and files can be separated by spaces[35].

2.5 Web Usage Mining Process

According to Jaideep et al. [36] there are three main tasks for performing web usage mining or web usage analysis. Such as preprocessing, pattern discovery, and pattern analysis as shown below in figure 2.4

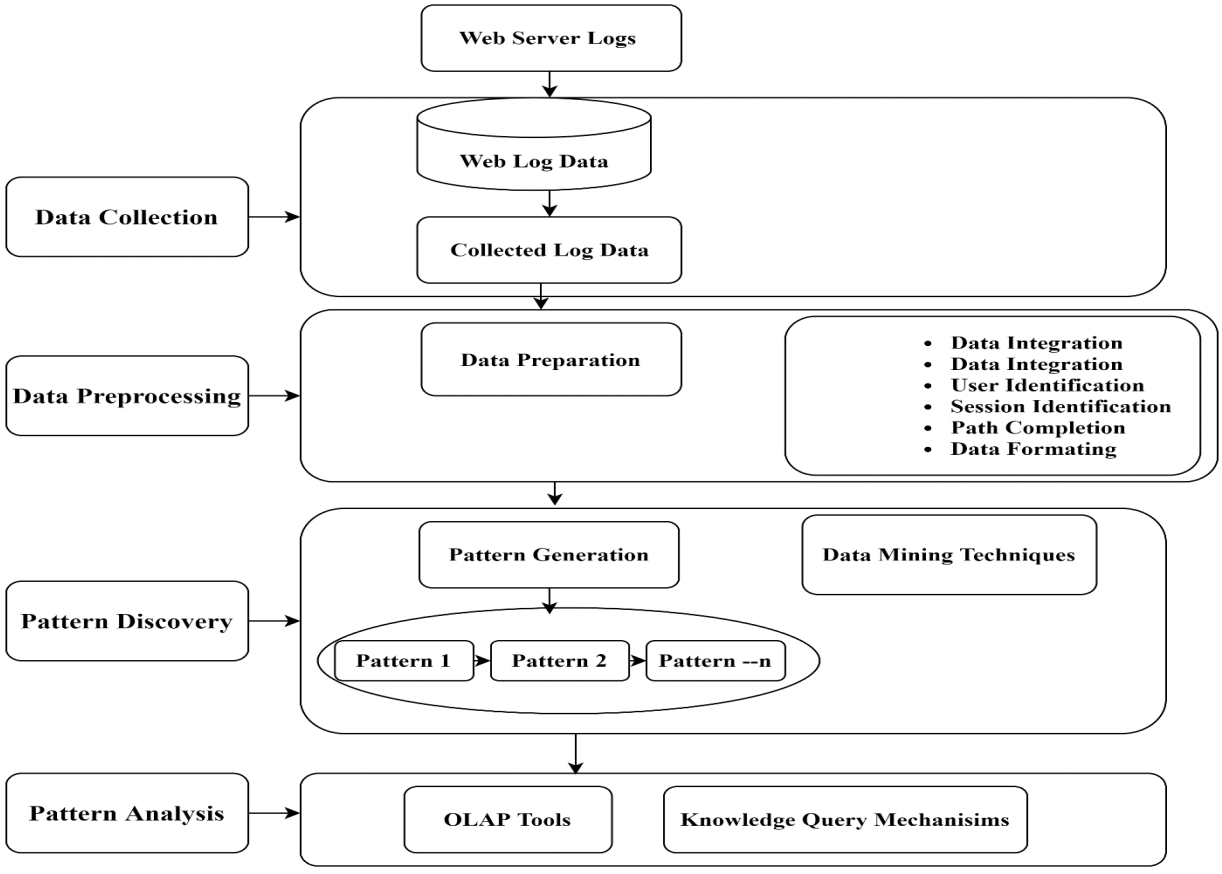


Figure 2. 3 High-Level Web Usage Mining Process[37]

2.5.1 Preprocessing

Preprocessing is the preliminary step for the web usage mining process. It includes converting the usage, content, and structure information contained in the various available data sources because data should be consistent and integrated to be used in pattern discovery. Data preparation includes data cleansing, user identification, session identification, path completion, and transaction identification[38].

2.5.1.1 Data Cleaning

Data cleaning is a fundamental activity in Web usage data preprocessing to clean raw web data from noise data. Since web log data stores all activities of web users on the website, the web server records a lot of useless information (for example the browser version, browser name, the version of the HTTP protocol used, etc.). therefore the goal of data cleaning is to remove irrelevant or redundancy data from the log files that do not support significant information for pattern discovery or analysis tasks[23].

Redundant records in the log file are due to a separate access request generated for every file, image, and multimedia object embedded in the web page. In such a way, a single user request for a particular web page might result in several log entries that correspond to files automatically downloaded without an explicit request of the same user. So the method to eliminate these items can be performed by checking the file name suffix such as gif, jpeg, GIF, JPEG, jpg, JPG, etc[23]. In addition to this, cascading style sheets and robot results can be removed from the log files and erroneous files can be removed by checking the status of the request (such as status code 404).

2.5.1.2 User Identification

One of the most complex activities in web usage data preprocessing is user identification, which is the identification of users who access a particular web page. Therefore the task of identifying a single user is crucial to characterize the browsing behavior on the website[23].

This identification can be performed with the help of the user agent and IP address. In web usage analysis identifying a user does not need, However, it is essential to differentiate among different users. Since a user may visit a site more than once, the web server records a variety of sessions for every user. The most widespread approach to distinguishing a unique user is performed through

client-side cookies. But it is impossible to use cookies for the whole website due to the privacy issue in which client-side cookies are disabled by the users. Another approach is the use of IP addresses. Users with different IP addresses can be treated as unique users [39]. The IP address or remote hostname is present in standard log files, and it is deemed to identify unique users. However, identification by IP address is largely intricated due to certain policies of internet providers, by proxy servers, by individuals using multiple computers, and by multiple individuals using the same computer[39].

2.5.1.3User Session Identification

Session identification is the method of identifying a group of activities for a particular user when he or she login and logout from a website. So, the session is the period when a particular user entered or left the website, and the user visits several pages during this time. A user session is a delimited set of pages visited by a similar user within a particular duration of a specific visit to the website. To group, the activities of a single user from the web log files are referred as a session. Those sessions can predict the page sequences and trace the activity of the web user, and the user may have multiple or single user sessions during the time of visiting the website. Once the user session has been identified, each user clickstream is divided into logical clusters. the method to build this session is known as session reconstruction or sessionization[39].

Sessionization can be defined as the process of identifying a particular user session from the web log data. Sessionization can be done using two approaches as time-oriented approach and the navigation-oriented approach. when the user is connected immediately to the website, it is called the session of that particular user. Most of the time, 30 minutes time-out was taken as a default session time-out[39].

After recognizing users' sessions, if both IP address, browser version, and operating system are the same the next consideration is the referrer URL filed. If the URL in the referrer field in the existing record is not accessed in the previous or if the referrer field is empty then it is considered a new user session[40]. According to [40] there are two heuristics methods for user session identification such as the time-oriented and navigation-oriented approaches. In the time-oriented approach, two pieces of information were considered such as total session time and a single page stay time. In the first example, consecutive access is considered to belong to the same session within a minimum fixed timeout which varies from 25.5 minutes to 30 minutes. Two consecutive

accesses exceeding a maximum set timeout are considered to belong to separate sessions. The second method is based on a single page stay time which is taken the difference between two timestamps. if it exceeds 10 minutes the user is considered as a new session. The second heuristic approach uses web topology in graph format for session identification. It considers webpage connectivity to identify user sessions.

2.5.1.4 Path Completion

Due to proxy servers, cache agent, POST technique, and the use of the back button by clients the URL paths are not completely recorded in the web log file. Therefore, many missed pages are logged in the user's session. So, to identify the missing pages, a path completion step is needed in web data preprocessing. Path completion is done using the URL and the referrer URL in the user's session. If the page request is not explicitly connected to the last requested page, the recent session history is checked, and if the page was previously accessible as a referrer URL, the related URL of the previous entry is added to the path. Full paths provide a user with full navigation during a specific visit[41].

2.5.1.5 Transaction Identification

As cited in [41] transaction identification is one important activity in web usage data preprocessing. The main goal to conduct transaction identification is to create meaningful clusters of references for each web user. The identification of transactions varies from case to case depending on the web usage mining. Transaction identification can be done through the use of the divide or merge approach. To find out the user's travel pattern and the user interests, there are two kinds of transaction approaches such as travel path and content-only transactions. In travel path transactions both auxiliary and content pages are accessed by the user, while in content-only transactions only content pages which used to discover the user's interests and cluster users visiting the same website.

2.5.2 Pattern Discovery

After web data preprocessing is completed, the next stage of the web usage mining step is pattern discovery. Pattern discovery is the process to discover interesting patterns or knowledge from the pre-processed web log data. different techniques are applied to discover the web patterns such as statistical analysis and data mining techniques. These methods are applied to web log data to discover web user behavior. Once the patterns or knowledge is discovered, it can be represented in the form of a table, graph, charts, etc.[42].

2.5.2.1 Statistical Analysis

Statistical analysis is the most powerful technique to extract knowledge from website visitors. Statistical analysis can be done on a session file through the use of variables such as page views, viewing time, and length of a navigational path. The result of statistical analysis can provide a periodic report including information such as the most frequently accessed pages, average view time of a page, or average length of a path through a website. also, the result might include unauthorized entry points or finding the most common invalid URL. Despite, lacking a depth of statistical analysis the result can help improve system performance, and site modification, enhance System security, and provide support for decision-making [37]. Different statistical analysis tools are used to discover or extract knowledge from the web log files. Some of the tools are as follows[43].

Web Log Expert: is a fast and powerful access log analyzer. It provides detailed information about website visitors, such as activity statistics, accessed files, paths through the site, referral pages, search engines, browsers, operating systems, and more. The tool provides easy-to-read reports including text information and charts. It can also offer reports in different formats like HTML, PDF, and CSV formats. Additionally, this tool can analyze both Apache and IIS web server log files, and it can read compressed files so, it doesn't need to unpack them manually. Generally, the tool has an intuitive interface and the built-in wizards help to create easily and quickly for the website profile and analyze it[44][45].

AWStats: is a functional tool that graphically generates advanced web, streaming, FTP, or mail server statistics. This log analyzer works on the CGI or command line and shows all the information from the log files. To be able to process huge log files, sometimes and rapidly, it uses

a partial information file. All major server tools such as Apache log files, WebStar, IIS(W3C log format), and many other webs, proxy, streaming servers, mail servers, and some FTP servers can be analyzed using this tool[44][46].

W3Perl: W3Perl is a free Web analytics platform focused on CGI. It provides the ability to monitor page data using a page bug without looking at log files or the ability to read and report log files[44][47].

Webalizer: Webalizer is a little free research tool for Web log analytics that is easily ported to several different systems. For reports, it comes with many distinct languages and a lot of stats to report on. The Webalizer is a free fast web server that programs for log file review. It generates very detailed, easily configurable HTML-format usage reports for viewing with a regular web browser[44][48]**.**

Data Preparator: It is a free software application designed to assist in web log data analysis and data mining with data preparation (or data preprocessing) activities. In particular, Data Preparator can help to explore and prepare data for data cleaning, discretization, counting, scaling, attribute collection, missing values, outliers, statistics, visualization, balancing, sampling, row selection, and many other tasks, and also manages large volumes of data[49].

Google Analytics: is a free tool offered by Google that focuses primarily on marketing purposes. Tracing their course helps to evaluate website statistics and provide a general report about the web visitors and their requirements. To aid the consumer efficiently, it supports mobile app platforms[44].

Deep Log Analyzer: It has the extensible ability, unlike other software, to examine various types of logs, including FTP logs. It can create a list of keywords containing keywords and hits on web pages. It is very useful for optimizing search engines[44][50].

Open Web Analytics: It can process very big logs and can also optionally fetch those from a database format directly. This open-source platform, unlike many other technical tools, can provide a click-stream report. This helps to troubleshoot the website code, to know exactly what the web user did, and to try to repeat those steps to reproduce the issue. It can also produce a report-style heatmap whereby the website statistics are divided into most-hit and least-hit sites, displayed for easy understanding in the form of color gradients[44][51].

Glogg: is a fast, smart weblog explorer and is a multi-platform GUI application that supports various log file formats to explore, browse, and search complex log files. It is designed to take

into account programmers and system administrators. Glogg can be seen as an explorer of graphical and interactive weblogs. To load large files, and search and browse log files, the glogg program is easy to use and very quick in terms of loading times[52].

Web Log Expert: is a statistical log analyzer tool that provides information about the visitors to your site such as activity statistics, accessed files, a path across the site, referring pages, search engine browsers, operating systems, and more. The program generates reports that incorporate both text information(tables) and charts in an easy-to-read manner. The log analyzer can generate reports in HTML, PDF, and CSV formats. It also comes with a web server that can handle dynamic HTML reports. Weblog Expert can examine Apache and IIS web server logs. It can even read GZ and ZIP compressed log files, eliminating the need to manually unpack them. The program has an easy-to-use interface. Built-in wizards will assist you in quickly and easily creating and analyzing a profile for your site[43].

Minitab: is a statistical analysis software that aids in data analysis. This is the most extensively used software for small, medium, and big businesses. It gives a quick and easy way to enter statistical data, change it, spot trends and patterns, and extrapolate answers to present problems[53].

2.5.2.2 Data Mining Techniques

Generally, Data mining tasks can be classified into two categories[54] such as descriptive data mining and predictive data mining .predictive data mining is used to predict unknown or future values of the attributes of interest using other attributes in the database. Also, predictive data mining is used to construct one or a set of models, perform inference on the available set of data, and attempt to predict the behavior of new data sets. Classification is one example of predictive data mining. While descriptive data mining refers to describing the data to understand it easily and interpretable to humans. Additionally, descriptive data mining is used to describe the dataset in a precise and summary manner and presents interesting patterns of the dataset. Clustering and association rule mining are examples of descriptive data mining.

2.5.2.2.1 Association Rule Mining

Association rule mining is one of the most popular ways of data mining techniques used to find regularities or patterns in data. association mining has been used in many applications, one of the best is in the field of business for the discovery of purchase patterns or associations between

products to provide effective marketing and decision making[55]. According to [56] association, rule mining is the most data mining technique, and at the same time the most used technique in web usage mining. When applied to web usage mining, association rules are used to find associations among web pages that frequently appear together in users' sessions. The typical result has the form "A.html, B.html=>C.html" which states that if a user has visited page A.html and page B.html, it is very likely that in the same session, the same user has also visited page C.html.

Additionally, this technique depends on generating frequent patterns and rules. After preprocessing is done the data in the web log file describes interesting facts such as the number of URL visits by which one can identify frequently accessed web pages by users which can help to understand user needs. The association rule focuses on the discovery of relations between pages visited by users on a website and can be used to relate the web page that is most often used by the single user session[5].

Generally, association rule discovery is applied to databases of transactions where each transaction contains a set of items. In such a way the problem is to discover both associations and correlations among data items where the presence of one set of items in a transaction implies the presence of other items. But in the context of web usage mining, this issue involves discovering the correlations among references to various files available on the server by a given client. Each transaction is made up of a set of URLs accessed by a client on a visit to the server. For example, using association rule mining we can find correlations such as the following. 40% of clients who accessed the web page with URL/company product1, also accessed/company/product2; or 30% of clients who accessed/company/special placed an online order in /company/product[57].

2.5.2.2.2 Measures of Association Rule Mining

There are various metrics to find the strength of the association rule and filters to avoid insignificant generated rules from the association rule mining. The following are some of the metrics to evaluate the association rules[58].

Support and Confidence: Support and Confidence are the two common indicators of the objective measure to evaluate the association rule; the former measures the usefulness of the rules while the latter reflects the effectiveness of the rules. **Suppor**t [58]refers to the frequency that the concurrence of data domains A and B involved by the association rule occupies in all item sets,

during the research data item sets. The accuracy will be higher only when the researching association rule frequently appears in item sets. Only when the support of the concurrence of A and B is greater than or equal to the designated minimum support threshold, A and B will be confirmed to be frequent item sets. The Support can be expressed as:

$$\boldsymbol{s}(\boldsymbol{A}\rightarrow\boldsymbol{B})=\boldsymbol{P}(\boldsymbol{AB})=\boldsymbol{N}(\boldsymbol{AB})/|\boldsymbol{D}|$$

Among these, $\boldsymbol{N}(\boldsymbol{AB})$ stands for the number of records of the concurrence of A and B, and |D| refers to all the number of records of the data sets.

Confidence [58] is the statistical probability of the occurrence of the consequent after the occurrence of the antecedent among the transactional data sets. Confidence is used to measure the reliability of the rules. The formula is as follows:

$$\boldsymbol{c}(\boldsymbol{A}\rightarrow\boldsymbol{B})=\boldsymbol{P}(\boldsymbol{B}|\boldsymbol{A})=\boldsymbol{P}(\boldsymbol{AB})/\boldsymbol{P}(\boldsymbol{A})$$

The process of mining association rules can be broken down into two steps[58]: first, identify the maximal frequent item sets that satisfy the requirement, and second use that frequent item sets to construct the association rules. Strong association rules are generated after the weak association rules that cannot achieve the support-confidence level are filtered out. Support (s) and confidence (c) measures of rule interestingness and they respectively reflect the usefulness and certainty of the discovered rule.

Lift (Correlation Analysis): Due to the deficiency of a support-confidence frame, some scholars carried out the correlation analysis for the mined association rules, namely Lift. Lift [58] is also called Correlation or Interestingness in some references. Lift refers to the ratio of the rule's confidence to the consequent's occurrence probability of the rule, reflecting the positive and negative correlation between the antecedent and consequent. The research of the correlation can partly remove some rules that have little correlation among the rules mined based on the support-confidence frame. The correlation reflects the probability ratio of B occurrence under the condition of A to the B occurrence without the condition of A, and reflects the relation between A and B. Lift does not possess the downward closure or the problem of a rare itemset.

$$\boldsymbol{Lift}(\boldsymbol{A}\rightarrow\boldsymbol{B})=\boldsymbol{c}(\boldsymbol{A}\rightarrow\boldsymbol{B})/\boldsymbol{P}(\boldsymbol{B})=\boldsymbol{P}(\boldsymbol{AB})/\boldsymbol{P}(\boldsymbol{A})\boldsymbol{P}(\boldsymbol{B})$$

The value range of the lift is, a closer value to 1 indicates that the rule has a greater value.

The mathematical framework of support and confidence

Let S be the set of all possible purchases and let n be the number of transactions. Each transaction record is a *subset* of S. We consider rules of the form ―(x1, x2… xj) implies (y1, y2… yk) where

X1, x2… y1, y2 … are elements of S. The collection (x1, x2… xj) is called an *item set*; read this as ―x1 and x2 and … and xj''. The *support* of the rule is defined as

Supp(x1,x2, ….) implies (y1,y2,….)) = (No. of transaction containing x1,x2,… and y1,y2,..) / N

More generally it defines the support of an item set as

Supp(x1,x2) = (No. of transaction containing x1,x2) / n

The confidence in the rule is

Conf ((x1,x2, …) implies (y1,y2, …)) = supp (x1,x2, …) implies (y1,y2, ..) / supp(x1,x2)

To consider a rule, we impose a minimum support, indicating a reasonable amount of data about the rule. The confidence measures how good a predictor the rule is. If we specify a minimum support s0 and a minimum confidence c0, then a *strong rule* is one which has Supp((x1, x2, …) implies (y1, y2, …)) > s0 and Conf((x1, x2, …) implies (y1, y2, …)) > c0

2.5.2.2.3 Association Rule Mining Algorithms

There are several association rule discovery algorithms, however, the widest algorithm in association rule discovery is Apriori and Frequent pattern growth(FP-growth) [59].

Apriori Algorithm

Apriori algorithm captures massive datasets during its initial database passes and the outcome will become the base for discovering other large databases during subsequent passes. Item sets that have the degree of support above the minimum are referred to as large or frequent itemset, and those below are small itemset. Apriori algorithm is the most supervised and imperative algorithm for mining frequent itemset. It uses a bread first search and a Hash tree structure to identify candidate itemset efficiently and effectively. Apriori algorithm identifies individual items within the information and extends them to larger and bigger itemset as long as those itemset seem sufficiently typically within the information. The algorithm confirms frequent itemset that may be accustomed to determining association rules that highlight general within the information[60].

According to Ganjir and Chopra [61], the procedure of the Apriori algorithm is the following:

- Generate a candidate of size N from the dataset
- Any(N-1) itemset that is not frequent cannot be a subset of frequent K-itemset
- Generate C-n: candidate itemset of size N: L-n: frequent itemset of size N
- Generate C_n+1 itemset from L-n

- Generate L_n+1 candidate with minimum support
- Returns generated candidates in L_n+1

FP-Growth Algorithm

The FP growth algorithm from FP-Tree generates frequent item sets by traversing fashion from the bottom up. The algorithm discovers frequent items without candidate generation item sets[59]. the method of the FP-growth algorithm turns the problem of finding long frequent patterns into recursively looking for shorter ones and then concatenating the suffix. As a suffix it uses the least popular objects, providing good selectivity. The technique decreases the search costs considerably when the database becomes large[61]. FP-growth algorithm covers the drawbacks of the Apriori algorithm. It is one of the fastest algorithms in association rule mining.it has a two-step process, in the first step the database is scanned twice, and find the support count of each item and infrequent items are deleted from the list, the remaining items are stored in descending order. In the second step, the FP tree is constructed using a frequent itemset[62].

FP-Growth algorithm has three unprejudiced significances: first, the database is scanned only two times and computational cost decreases dramatically. second thing is that no candidate item sets are generated. third, it uses the divide and conquers approach which consequently reduces the search space[62].

As cited in Amare [4] the following steps are involved in the FP-Growth algorithm.

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

Table 2. 1 Comparison between Apriori and FP-Growth Algorithm[62]

Parameters	Apriori Algorithm	FP-Growth Algorithm
Memory required	Large due to large candidate generation	Less memory due to no candidate generation
No.of visits to the Database	Multiple times for candidate set	Twice only
Techniques used	Use Apriori property and prune property	It constructs a conditional frequent pattern tree and conditional pattern base from the database which satisfies the minimum support
Execution time	More time needs due to candidate set generation	Small as compared to Apriori

Apriori algorithm continuously scans all web log data and checks a large set of candidates by pattern matching. Because of this, it takes a long running time. Whereas, since the FP tree algorithm scans the web log only twice, it is faster than the Apriori algorithm. But the FP tree algorithm only takes a binary value.

2.5.2.2.4 Sequential Patterns

Sequential mining is a data mining method to discover or extract subsequences from a large volume of web data[14]. For the identification of navigational patterns in web use data, sequential pattern discovery turns out to be especially useful. To discover a sequential pattern, time is added to the process of finding trends in this method. This technique seeks to discover time-ordered sequences of pages that often occur in user sessions. In general, a typical sequential pattern is expressed in a form syntactically similar to association rules. For example, a sequential trend might be as follows: 80% of users who first visited the A.html page and then visited the B.html page also accessed the C.html page during the same session[23]**.**

Using sequential pattern discovery, useful user patterns can be discovered, visit pattern predictions can be made, website navigation can be enhanced and website content can be adapted to individual customer requirements or automatically suggested to customers better suited to customer profiles[14].

2.5.3 Pattern Analysis

The final procedure for web usage mining is pattern analysis. In this step, irrelevant rules or patterns were eliminated or removed from the generated rules or patterns. This step attempts to extract interesting rules or patterns from the output of the pattern discovery process[32]. different techniques and tools are needed for pattern analysis to make those patterns understandable for the analysts and to maximize the benefits from those generated patterns. Techniques such as database querying, graphics, visualization, statistics, and usability analysis are included pattern analysis step. but the most common methods are visualization and online analytical processing(OLAP) tools[38].

2.6 Application of Web Usage Mining

Web mining application is related to the exponential growth of the world wide web, web mining is a very popular and hot subject in the field of web science. Web mining also plays an important role in e-commerce and e-service websites to understand the use of their websites and services and to provide better service to both consumers and users. some of the web mining applications are as follows[23][63][8].

E-Learning: one of the applications of web mining is e-learning which is used for enhancing the teaching and learning process in the e-learning environment. Machine learning techniques and web usage mining improve web-based learning environments by providing online education courses and materials for users[8].

Digital Libraries: digital services or repositories provide useful information to spread around the globe. So, through the use of web mining, it can investigate the needs of users from different aspects of digital library books. Additionally, it removes the physical presence in various sections of different libraries[8].

E-Government: The key aspect of e-government systems is related to the use of technology to provide services online, focusing on public needs, and delivering better information and enhanced service in support of the government. E-government systems will offer personalized services to people resulting in better information and enhanced service. Organizations that communicate with the citizen of the country contribute to better social services[8]

Website Personalization: is the process of customizing the content and/or the layout of a website based on the needs or interests of users by taking the input from the analysis of the user's navigational behavior. personalization of the website gathers user information, and certain privacy is the major issue. collaborative filtering includes personalization technology, where filtering is applied to various sites for the selection of relevant information that might appeal to a particular user. User profiling data can be obtained from numerous websites as a result of a customized web page. Using web mining data analysis, users can forecast future forecasts and locate a session with other websites with interesting links. So web mining is an imperative research area to solve the problems which arise from the web users regarding the contents of the website[63].

Website Reorganization: the attractiveness of the website is an essential factor that gives the website a strong structure. Website reorganization can be done based on how users communicate with the websites, how they link, and what are the underlying behavior habits of users with the website. The layout of the website can be reorganized by website navigation. It dynamically updates the relationships between web pages. Reorganization can be carried out with the extraction of frequent patterns of mining for web use. Information on web use offers information on the user behavior of any website. The content and layout also relate to an adaptive website[63].

Web Prefetching: to keep user satisfaction, the system's output is a very critical concern. Web usage mining is an important field of research for web traffic detection. Web traffic consists of between the PC and the server. For a web-based organization, the amount of traffic and specifies each visit is incredibly imperative information. The server computer tracks each user's request for a web page and decides which sites get the most attention. Web traffic analysis provides companies with concrete, reliable information about their customer's interests. The more traffic a website receives, the more sessions it receives, and the more server processes it hits. Every time a web server processes a file request, the computer creates a server log entry, a dedicated file on the hard drive of the server. So, to develop the server performance new policies should be used**.** The user experience can be slowed down by downloading files. Therefore this prefetch approach is

useful for both client and server-level web caching, load balancing, and the transmission of data distributed information[63].

Web Site Improvements: according to [23] the analysis result of web usage patterns which is discovered by the web usage mining method might provide valuable recommendations for improving the design of websites by reorganizing their web pages and improving their usability and accessibility. Many automated methods have been built in this application context to create sites that can redesign themselves based on the past actions of individual users. this leads to the development of adaptive websites, which are capable of enhancing their organization and presentation automatically by learning from user access trends. The integration of this study into web systems not only enables information about the most common links on the website to be collected but also recommends an optimal page design based on the frequency of visitors and the time spent on those links.

Usage Characterization: the user uses a website in several ways. Based on the user information which is previously stored on the web server logs, user activities can be classified and clustered according to the navigational behavior of users' interests [63].

Business Intelligence: web usage mining offers data to enhance the needs of the customer in the marketing field. It also improves the competitive advantage of an organization through internet marketing. Through this technology, organizations understand the interest of users by analyzing the previous history of data available on the server log file and increasing the demand for product sales based on the habits of users[63].

Bandwidth management: The largest amount of data that can be transferred in a particular amount of time, usually measured in seconds, is known as bandwidth. The amount of data to be sent is known as data transfer, while the rate of data transfer is known as bandwidth. When the quantity of traffic on the network is modest in comparison to its capacity, data flows fast and smoothly. The speed at which data travels begins to slow with the amount of traffic approaching the network's capacity**.** When a user visits a website, the communications between the user's computer and the site's server are referred to as web traffic. To a web-based service, the amount of online traffic and the details of each visit are extremely valuable data**.** Web traffic analysis helps in monitoring and controlling internet traffic, particularly in determining when additional bandwidth should be added. HTML, photos, video, or sound may be downloaded by web users. The larger the files (video, image, and audio) and the more people that use them, the more traffic

there will be, necessitating a large amount of bandwidth. If consumers don't have enough bandwidth, the system will fail. Traffic becomes congested, causing delays for guests. Setting bandwidth limitations helps ensure that network access is only used for purposes that are in line with the university's mission. Web traffic is one approach that system and network administrators use to manage the network. Ensuring that bandwidth is accessible for academic, research, and administrative purposes in line with the mission of the university. User download content type, number of web users available, requesting item size, and number of connections established are some of the criteria used to analyze online traffic, according to various scholars. These are increasingly becoming a determining factor in web response time.in response to a web user's request.[64].

2.7 Related works

In the last few years, various national and international scholars have worked in the field of web usage mining to investigate various aspects of the web mining endeavor, ranging from web development to the application of statistical and data mining techniques for web mining.

Yohaness Mesfin[15] has researched the application of web usage mining for extracting employee internet access patterns by URL category: the case of Commercial Bank of Ethiopia. His study is mainly focused on employees' internet access behavior on duty and off duty hours. The researcher collected data from the Commercial Bank of Ethiopia Cisco Iron port web appliance device for his study and he utilized a Hybrid knowledge discovery model for his study. He also used the URL profile tool to categorize URLs and MacAfee for further categorization of URLs. Besides, he used MS-excel for statistical analysis and the WEKA tool for association rule discovery. The main finding from statistical analysis indicates that the Entertainment and Social URL categories are the most frequently accessed internet services during off-duty hours, while the Internet Services and Business URL categories are the most frequently accessed internet services during duty hours.

Tadele Asitatikie[65] has researched web usage pattern discovery: the case of Addis Ababa University's official website. The researcher has utilized web server log files and applied three main steps such as data preparation, pattern discovery, and pattern analysis for his study. He used Match5 statistical analyzer to explore the general statistics of the website and WEKA for association rule discovery and he utilized sequence mining with an Apriori algorithm to discover

common navigational sequences in the website. He also used Python code for data preparation and the WUMprep tool for data processing. According to the statistical analysis, his finding shows that most users of the website begin navigation at the home page of the website and the page of the Institute of Ethiopian Studies program has been visited more frequently than other program pages. In terms of the daily user visit trend, Monday through Thursday have relatively higher hits. Meanwhile, the weekends have the fewest hits, while Friday has the most.

Furthermore, several requests for pages about AAU return the error of some pages being unavailable. This includes the president's office, the president's office/communication team, the president's office/reform team, human resources, and policies and procedures pages all returning a response indicating that the pages are unavailable. Moreover, his final pattern discovery demonstrates that the majority of users of the official website who visited the academics page would more likely visit the library page. Similarly, the academics page and admissions page was accessed together, followed by online applications for programs. Finally, he recommended that the website be restructured in a user-friendly manner and that the pages be accessed as error-free as possible.

Senait Mezgebu[66] has researched web usage pattern discovery and analysis for website optimization: the case of Ethio Telecom's official website. She collected the data from the server of Ethio-telecom's official website. The researcher used different tools like Google Analytics for statistical analysis, WUMprep for data processing, and MS-excel for further data processing. Additionally, she used python code to prepare appropriate data format for pattern discovery and the WEKA tool for association rule discovery followed by the FP-growth algorithm. The researcher has adopted an experimental research design followed by the Sharma model. Under the statistical analysis, the researcher categorizes the report in three different formats audience, acquisition, and behavior. From the location report, the traffic sources show that most of the visitors originated locally and most of the users access the home page by directly typing the URL address. Moreover, 93% of visitors who visited business Internet following visits to the Fixed Broad Band Unlimited ADSL fiber page, and more than 88% of the visitors who visited the home page after the bid and vacancy pages are the most frequent access to the results of the Associative Rules, and the hose visitors who had visited the internet page and business page also visited troubleshooting page.

Gashaw Bekele[67] researched exploring users' navigational behavior using web usage mining: the case of Ethiopia commodity exchange official website. He took a one-month duration web server log file collected from the Ethiopia commodity official website. He used a variety of tools for his research, including a weblog storming tool to explore general website statistics, a log file viewer for data preprocessing, MS-excel for additional data cleaning, and WEKA for association rule discovery, which was followed by the FP-growth algorithm. His findings show that the most widely accessed pages, the top-level input and output pages, the user's top navigation paths, and the majority of visitor countries and cities were analyzed through statistical analysis and frequent errors. Also, half of the users of the Ethiopian commodity exchange official website begin their navigation at the website's root page, while the others use a search engine to find the page they want to visit. Regarding the association rule, the mining technique reveals that the majority of visitors who visited the root page and home page of the Ethiopian commodity exchange official website did not visit the maize product daily price document, and the majority of visitors who visit the root page also visit the website's home page

Anteneh Legesse[68] researched discovering frequent navigational patterns for constructing user profiles: the case of e-business online solutions for private limited company official website. In his study, he has used WUMprep for preprocessing the data, a weblog expert tool also used for statistical analysis, WEKA for association rule discovery, and clustering to find out the interesting rules for his study. The data source for his study is server web log data from the e-business official website. He adopted experimental research as a research design for his study. according to his finding the statistical analysis reports the most frequent access pages, top entry and exit pages, top accessed countries, and top website referrals for the users' behavior. Generally, his research found the most frequently accessed pages were shown by statistical analysis of the home page, tenders, and user page. Also, the association rule discovery using the FP-growth algorithm revealed that 90% of the users who accessed the page found themselves accessing the tender's page. Also, 90% of users who accessed the tender page have accessed user and index pages. Therefore, a tender page is the most widely accessed resource successfully identified. As a result, tender is chosen for the user profile.

Amare Mulatie[4] has researched log data analysis to discover web user navigational behavior. He used the Adama Science and Technology web server log file for his study. he followed the hybrid model data mining process model for his study. he utilized a data preparatory tool for statistical

analysis, a log file viewer and MS-excel for data cleaning, and WEKA for association rule discovery. He had discovered some useful interesting patterns and statistical analysis reports for his finding. In terms of the statistical analysis most frequently accessed pages (URL) and frequently accessed categories of the site were analyzed. He has conducted six different experiments with six different categories of the partitioned dataset for both statistical analysis and association rule discovery. Finally, according to the statistical analysis, social media and entertainment sites are the most frequently accessed websites by users. Also, in association rule discovery social media and entertainment site has a great probability to be browsed together relative to educational and organization sites.

Naha Goel [16] researched analyzing user's behavior from web access logs using an automated log analyzer tool. The researcher was taken web access log data from the astrology website. His study is mainly focused on analyzing web users using statistical analysis study. He also used the Google Analytics tool to discover statistical reports for the website. According to the statistical analysis report, the result shows general statistics, access statistics, visitors, errors, and activities of the web uses were discovered.

Awet Fesseha[69] researched exploring the navigational behavior of Addis Ababa University's official website users. He used web server log files for his study, and he utilized the WUMPrep tool for data processing and the WUM tool for statistical analysis. In his study, he discovered the most requested pages, top entry pages, top exit pages, and referrer pages of the website. His final finding shows that most of the users of AAU's official website start navigation at the home page of the website. He also stated which pages are commonly visited after the home page. According to Awet's research, AAU's website is mainly accessed by directly typing the URL or using a search engine, implying that AAU's website receives fewer referrals from other websites.

Arvind, and Gupta[17], have conducted a study on the Analysis of Web Server Log Files to Increase the Effectiveness of the Website Using Web Mining tool. the study was focused on analyzing the web server log files of an Educational Institution's Website that is www.davkota.orgto discover web usage behavior of the Website users. the researchers used experimental research methodology and web log expert tool to discover users' behavior. Their main findings were identifying the number of the visited website within a day or a week, identifying the top downloaded files by the users, the top browsers used by the users, and the total number of errors that occurred during the user's access behavior has been discovered

2.8 Summary of Related Work

Table 2. 2 Summary of related works

Author & Year	Title	Methods and approaches	Tool	Finding
Tadele Asitatkie (2011)	Web usage pattern discovery: the case of AAU official website	WUM process, statistical analysis, and Apriori algorithm applied	WUMprep, Match5, WEKA, Python	Top frequent AAU website pages have been identified
Senait Mezgebu (2015)	Web usage pattern discovery: the case of Ethio telecom's official website	Experimental research design, statistical analysis, FP-growth algorithm, MS-excel	Google analytics, WEKA, WUMprep	Top frequent Ethio - telecom website pages have been identified
Gashaw Bekele (2015)	Exploring users' navigational behavior: the case of Ethiopian commodity exchange website	Statistical analysis, WUM process, FP-growth algorithm	Weblog storming, WEKA, MS-excel, log file viewer	Top ECX frequent navigated webpages were identified
Awet Fesseha (2011)	Exploring the navigational behavior of the AAU official website	Statistical analysis, sequence mining	WUMprep, python, web utilization miner, Perl	Top frequent AAU website pages were identified
Amare Mulatie (2015)	discover web user navigational behavior: the case of Adama Science & Technology	Hybrid knowledge discovery model, association rule using FP-growth algorithm,	Datapreparator-1.7, log file viewer, WEKA	Social media and entertainment websites are the most frequently accessed sites

	University	statistical analysis		
Yohaness Mesfin (2015)	Web usage mining extracting employee internet access patterns by URL category in the case of CBE	He utilized a hybrid knowledge model, Apriori algorithm, and statistical analysis	WEKA, MS-excel, URL profile	His finding shows that entertainment and social media URLs are the most frequently accessed services during off-duty hours and business URLs services are the most accessed services during duty hours

As summarized in Table 2.3 different researchers have studied web usage mining. All the researcher's data sources were the web server data. In addition, most of the researchers except Amare were focused on discovering users' behavior on a single official website, which page is most frequently accessed and their research objective is almost the same which is mainly focused on website modification based on their final result. But the area of web usage mining had a different application in addition to site modification, optimization, and redesigning tasks. In addition, identifying users' behavior using network VLANs and analyzing web usage traffic is an imperative study to apply efficient network bandwidth allocation for an organization's web users Though the last two researchers 'recommendation is taken as motivation for this study. this research is different from the above-listed research work. First, in this study, the web users' behavior is discovered and analyzed based on VLANs. Using VLANs. since VLANs are used to identify every user's accessed behavior. Secondly, in this study, the web traffic is analyzed based on the resources accessed by the users and it can identify in which VLANs the more web traffic has occurred through the statistical analysis. Thirdly, this study is mainly for investigating the result that is an input for internet usage policy, efficient bandwidth allocation, and identifying web usage traffic for Wolkite University internet users. generally, this research is different from the above researchers by objective, scope, data source, attributes, and methodology.

CHAPTER THREE
Research Methodology

3.1 Overview

Research methodology is a road map that shows how the researcher will study from the beginning to the end and is used to understand which research methodologies and procedures will be used for the goal of the data collection, preparation, organization, analysis and visualization, and interpretation for the study. Generally, in this chapter, the researcher describes the general research approach, research process models, and algorithms.

3.2 Research Design

In this research, experimental research methodology has been used. Experimental research is conducted by performing experimental analysis and identifying the cause and effect of variables for the experiment. The reason to select experimental research for this study, firstly its finding is depends on conducting experimental results. Second, it helps to identify the case and the effect of showing the final result. Thirdly, this research is conducted in different experiments based on the collected dataset and the final results have been discovered and analyzed according to the experimental results. Fourthly, this research design is best suited to the research objective.
based on the collected dataset and then performing different experimental results

In this research, the Sharma web usage mining process model has been used. The reason to choose the Sharma model[71] are the following, First, this model is the most popular one for web usage mining and used by scholars and previous researchers[66]. Second, this model is best suited for the research objective as compared to the other web mining process model. The experimental research approach is used by following four detailed web usage mining process models suggested by Sharma[70]. The process model consists of steps, such as data collection, data preprocessing, pattern discovery, and pattern analysis.

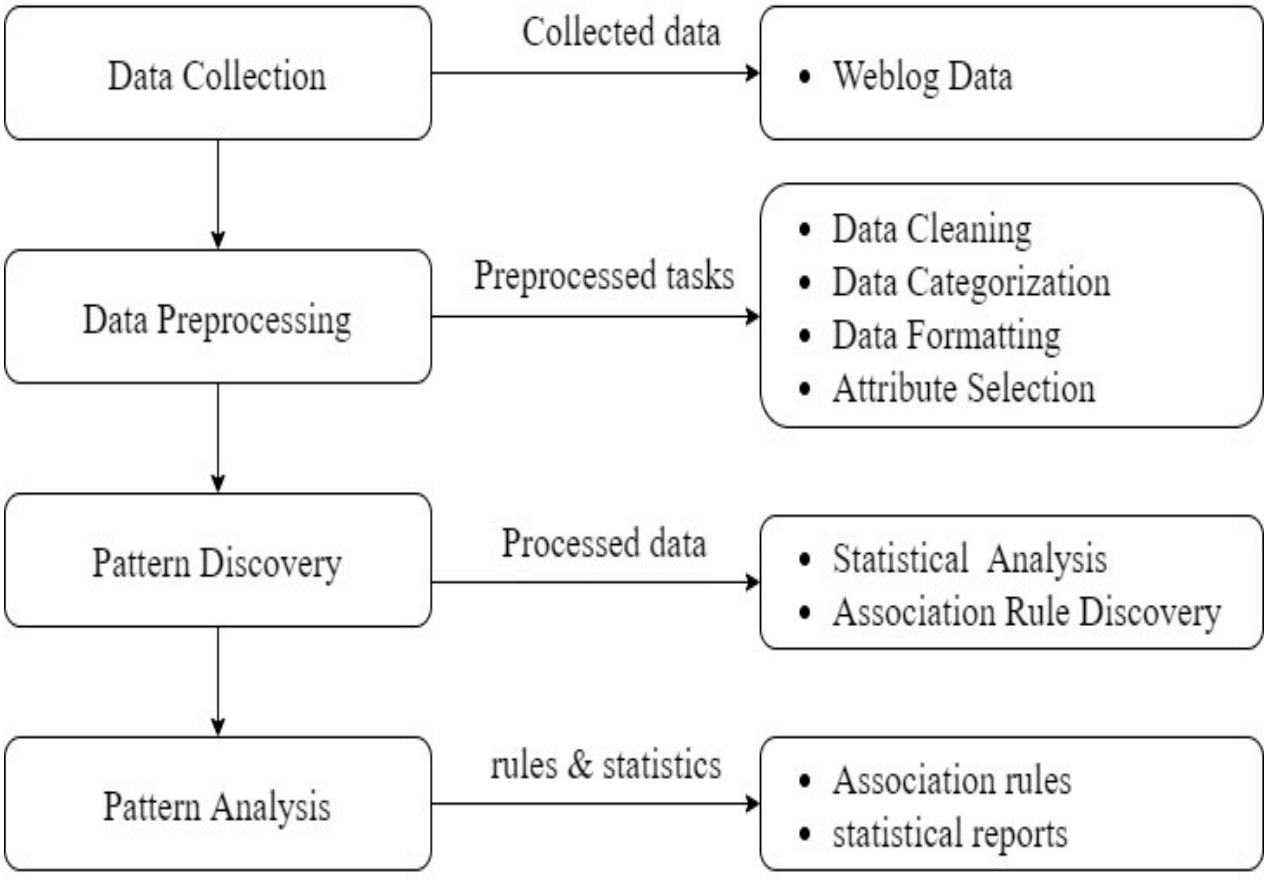


Figure 3. 1 Web Usage Mining Process Model adapted from [70]

3.2.1 Data collection

In this research, the source of data is web proxy server data. The data is collected from Wolkite University internet web users. The data is collected from February 1/2021 to April 30 /2021 with a total of three months of data taken for the study. the total amount of web log data was 6,875,218 records. The researcher has discussed the nature of the data with the domain experts especially the network administrator for WKU. The reason that the researcher took a three-month dataset is to represent the behaviors of internet web users, because of the massiveness of the data size, it is difficult to use and prepare additional datasets. A proxy server is a server that sits in the middle of the client and the web server. Web traffic is routed through this proxy server and handles all the requests and responses of web users. Some of the proxy server attributes are date/time, elapsed time, client IP address, requested URL, requested item size, transactional result code, port number, method, type of web protocol, cache retrieval, and content type[71].

```
1612196308.0 5006 10.194.10.40 TCP_REFRESH_HIT/304 214 GET http://www.goonernews.com/styles.css DIRECT/207.58.145.61 -
1612541908.0 2864 10.194.8.19 TCP_MISS/200 3614 GET http://www.dubmagazine.com/ DIRECT/65.61.150.253 text/html
1613665108.0 3856 10.194.8.109 TCP_MISS/302 519 GET http://mail.yahoo.com/ DIRECT/209.73.177.115 text/html
1614097108.0 1464 10.194.10.142 TCP_HIT/200 802 GET http://www.hornymatches.com/images/line_3.jpg NONE/- image/jpeg
1616515920.0 1452 10.194.1.22 TCP_IMS_HIT/304 218 GET http://www.thisdayonline.com/images/back.gif NONE/- image/gif
1615479120.0 1686 10.194.30.4 TCP_MISS/302 -1 GET http://eunis.eea.europa.eu/sites/343779 - NONE/- text/plain
1615417920.0 3657 10.194.48.1 TCP_MISS/200 -1 GET http://bio2rdf.org/pubmed:2748346 - NONE/- application/rdf+xml
1616109120.0 3667 10.194.13.29 TCP_MISS/498 -1 GET http://gutenberg.org/image/pdb:1H0A - NONE/- null
1617059520.0 116 10.194.48.64 TCP_MISS/200 559 GET http://ninemsn.com.au/images/hd/4.gif? - DIRECT/202.58.56.1 image/gif
1617145920.0 1982 10.194.11.10 TCP_MISS/200 124 GET http://gutenberg.org/image/ NONE/- null
1618528320.0 2420 10.194.2.246 TCP_MISS/302 683 GET http://cgi.sexlist.com/counter.cgi? - DIRECT/207.246.138.127 text/plain
```

Figure 3. 2 WKU Proxy Log File Sample

There are different sources of web log data such as client log files, web server log files, and proxy log files. In this research, the source of data is taken from the proxy server log file. The reason to select proxy server log file is that the web server contains only WKU official website web user statistics which means if the user access other than the WKU website the user log file could not be stored on the web server rather than a proxy server. The second reason is, according to the objective of the research full source of information is found on the proxy server. In addition, according to the network administrator any web user request whether from within or outside the university is filtered by the proxy server. If the queries or requests are superfluous or inappropriate the proxy server will deny them. This log data consists of several attributes such as timestamp in Unix time, elapsed time(duration), client IP address, transactional result code, HTTP response code, total bytes transferred, the request method, user name, URL, result code, suspect user agent and others. Because all requests go through a proxy server, the researchers used the proxy server's web log data. Furthermore, a proxy server can be used to identify or categorize web surfers based on their network VLAN. In Addition, the researcher had discussed the proxy server and web server properties in-depth with domain experts.

3.2.2 Data Preprocessing

After the data is collected the next phase is data preprocessing, in this phase different activities are carried out such as data cleaning, feature selection, and data formatting. To perform the data preprocessing.

- ✓ First, the log file data is converted into a text file format using the Notepad++ tool.
- ✓ Second, the text file has been split into different partitions using the Text File Splitter tool to make the data preprocessing easy since the data size is huge.

3.2.2.1 Data Cleaning

In the data cleaning session, the irrelevant file extensions such as gif, jpeg, GIF, JPEG, jpg, JPG, Cascading Style Sheet files (CSS), and scripts, have been removed and only the relevant URL path is extracted from the text file using Python. To perform this task Python library Regex package has been used.

3.2.2.2 Data Categorization

After data cleaning has been performed, the data is categorized based on the nature of network VLANs. To do this, MS-Excel 2021 has been used to split the VLANs from the IP address. Using Python similar VLANs have been selected and merged at once.

3.2.2.3 Data Formatting

The purpose of this data formatting is to set up the data format that is suitable for statistical analysis. Therefore, the data is converted into a comma-delimited file (CSV) using MS-Excel-2021.

3.2.3 Pattern Discovery

Following preprocessed activities are completed, the next task is pattern discovery and the web data can be ready for pattern discovery using statistical and data mining approaches. In pattern discovery statistical analysis, association rules, are among the methods that are being used in this

study. In this study statistical analysis and association rule mining algorithms specifically, Apriori and FP-growth algorithms have been used to discover patterns.

3.2.3.1 Statistical Analysis

The most prevalent analysis used to obtain information about web users' behavior is statistical analysis. This analysis technique can provide useful statistical information. This statistical data is utilized to generate an actionable report from the site such as the most frequently visited or accessed websites, and the least frequently visited or accessed websites have been identified.
To discover the web users' behavior in statistical analysis two basic experiments have been conducted. First, using statistical analysis with two separate experiments been conducted. Those are:

- ✓ Statistical analysis using students' VLAN
- ✓ Statistical analysis using staffs' VLAN

Through the above two statistical analysis reports web users' navigational behavior has been identified and analyzed. Therefore, to perform this statistical analysis selecting an appropriate statistical tool is an imperative task. As stated in the literature review different statistical tools have been described. In this, study Minitab 21 tool is selected to extract statistical information about WKU internet web users. Because the Minitab 21 tool has a better user-friendly user interface for statistical analysis.

3.2.3.2 Association Rule Mining

For a big amount of data objects, association rule mining discovers interesting patterns and rules. An implication form of X->Y, where X and Y are sets of items known as association rules. Association rule mining discovers all such implications in a given data collection that meet particular criteria such as minimal support and confidence. It usually assesses the degree to which the item sets X and Y are linked. To forecast web user behavior a variety of prediction approaches are available. In web usage mining, the mining of association rule is a critical study. In association rule mining a variety of algorithms are used such as Apriori and FP-growth algorithms[72][4].

Therefore, to discover such interesting patterns and rules we have used association rule mining algorithms. To discover usage patterns, we have used Python 3.10 programming language to conduct association rule mining experiments for this study.

3.2.3.2.1 Apriori Algorithm

Apriori algorithm is one of the most classical algorithms for mining frequent item sets which is proposed by Agrawal and Shrikant. it is used to find all frequent item sets in a given database DB. The key issue in this algorithm is to make multiple passes over the database. The following is the general algorithm for the Apriori algorithm[73].

Algorithm: Apriori Algorithm

Input: S- Log Data, rare_visitor - Minimum visitor page

Threshold

Output: F- Large Frequent item set

Begin

Step 1: k =1

Step 2: Find frequent visitor set Fk from Ck

Scan database S and count each visitor set in Ck

If the count > rare_visitor then Add that visitor set to Fk

Step 3: Form Ck+1 from Fk

For k = 1, C1 = all visitor sets of length-1

For k >1, generate Ck from Fk-1 as

Ck = k-2 way joins of Fk-1 with itself;

If both {I1 ,,Ik-2 , Ik-1} and { I1 ,…,Ik-1 , Ik } are in Fk-1 then

Add {I1… Ik-2, Ik-1, Ik} to Ck

Remove {I1… Ik-2, Ik-1, Ik} if do if it does not contain a large (k-1) subset;

Step 4: k = k + 1

Step 5: Repeat steps 2, 3, and 4 until Ck is empty

End

In general, the Apriori algorithm may have the following limitations:

- ❖ It produces a large number of item sets, due to this more search space is required and the cost of I/O will increase.
- ❖ As the number of database scan is increased it requires high computational cost in candidate generation

3.2.3.2.1 FP-Growth Algorithm

FP-Growth algorithm is one the frequent pattern mining which is used in the development of association rule mining. FP-growth algorithm solves the issue observed in the Apriori algorithm. The FP-growth algorithm was discovered to be faster than the Apriori algorithm by eliminating the candidate generation phase and making fewer passes over the databases. It employs a divide and conquers approach. First, it creates a frequent pattern tree or FP-tree from the database representing the frequent items. It keeps track of item set associations, and compressed databases are divided into a series of conditional databases, each one is associated with a frequent item. It also stores the associated information in a set of conditional databases, the process saves a lot of memory and helps in minimizing the processing time[39].

The FP-growth technique may directly extract frequent Itemset from FP-tree by using FP-tree. The FP-growth technique will be used to extract frequent itemset, which will generate a data tree structure called FP-tree. After a collection of transaction data has gone through the FP-tree building stage, the FP-growth algorithm will be used to hunt for important frequent itemset. The three main steps of the FP-growth algorithm are as follows: 1) Frequent itemset; 2) Conditional pattern basis; 3) Conditional FP-tree The FP-growth algorithm takes the following form[74]:

Algorithm: FP-growth

```
Input: D, a transaction database; and min sup, the minimum support count threshold.
Output: The complete set of frequent patterns.
// Construct FP-tree
        Scan the transaction database D once. Collect F, the set of frequent items, and their
        support counts.
        Sort F in support count descending order as L, the list of frequent items.
        Create the root of an FP-tree, and label it as null.
        For each transaction Trans in D do the following.
                Select and sort the frequent items in Trans according to the order of L.
                Let the sorted frequent itemlist in Trans be [p|P], where p is the first element
                and P is the remaining list.
                Call insert tree([p|P], T),which is performed as follows.
                        If T has a child N such that N.item-name=p.item-name, then
                        IncrementN's count by 1;
                        else
                        create a new node N
                        let its count be 1
                        its parent link be linked to T
                        itsnode-link to the nodes with the same item-name via the node-link
                        structure.
                                If P is nonempty
                                callinsert_tree(P, N) recursively.
// The FP-tree is mined by calling FP growth (FP tree, null), which is implemented as follows.
Procedure FP growth(Tree, a)
        if Tree contains a single path P then
for each combination (denoted asβ ) of the nodes in the path P
                generate pattern βUα with support count = minimum support count of nodes in
                β;
        else
                for each ai in the header of Tree
        generate patternβ = aiU αwith support count = ai:support count;
        constructβ's conditional pattern base and thenβ 's conditional FP tree Tree ;
if Tree β≠Ø then
call FP-growth(Treeβ ,β )
```

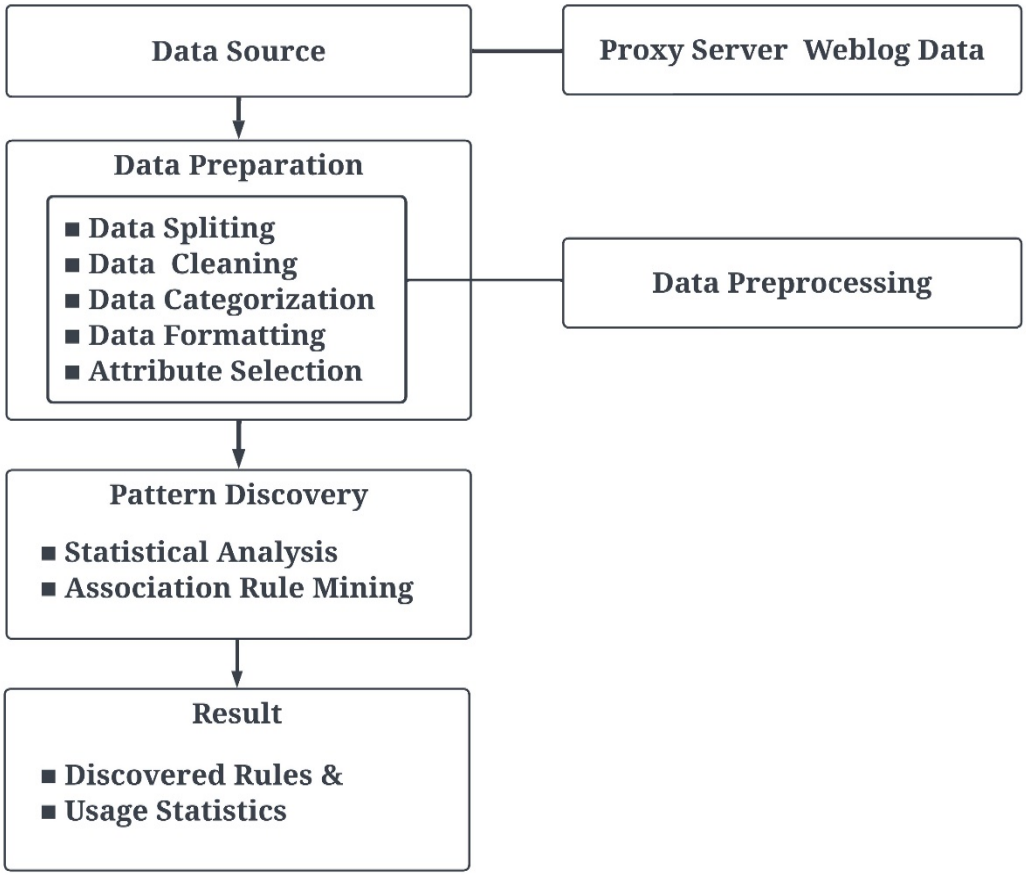
Generally, we have described Apriori and FP-Growth association rule mining algorithms, in this study the researcher selected FP-Growth algorithms for this study due to the following criteria. First, FP-growth algorithms have better memory utilization due to no candidate generation being required, second, it scans the database only twice, third the execution time is better than the Apriori algorithm[75].

3.2.4 Pattern Analysis

Once the patterns are discovered, the final stage of web usage analysis is pattern analysis. The goal of this stage is to extract the interesting patterns from the output of the pattern discovery patterns by eliminating the irrelative patterns[76]. Therefore, there needs to analyze those discovered patterns and determine how that information can be used. In this phase, interesting rules that can be generated during the pattern discovery phase have been described and discussed based on the objective of the study. in addition, the results are analyzed with discussion with the domain experts, after finding the results recommendations are forwarded to WKU ICT concerned staff that has an input for internet usage policy and traffic management for the WKU internet web users.

CHAPTER FOUR
DATA PREPARATION

Data preparation or preprocessing is a crucial task to discover internet web usage behaviors in the web data. Through this step, only appropriate data can be filtered and prepared before applying further web mining procedures. During this technique, the logfile sizes can be reduced since some of the data can be filtered and removed from the original log file data. Therefore, data preparation aims to improve the quality of data and increase the data mining accuracy results.



Data Source: The first step is identifying the type and source of the dataset for the study. As it has been described earlier the data is collected from proxy server weblog data in Wolkite University with a total of three months of weblog data. The reason that we have chosen proxy server data is due to that, only the official website of the university is stored on the server data. Therefore, the most useful information is found on the proxy server data.

Data Preparation: The second step is data preparation, in this stage, we have to prepare the dataset to make the data that is suitable for experimental analysis. During this, we have done different data preprocessing tasks such as data splitting, data cleaning, data categorization, attribute selection, and data formatting.

- **Data Splitting**: the first task in data preprocessing is data splitting. Since the dataset is huge it is difficult to browse in MS-excel in order to clean unwanted data. To perform this task, we have used the text splitter tool. Therefore, to manage easily the dataset is split into a different manageable size.
- **Data Cleaning**: in this stage, the unwanted data is removed from the original dataset. Moreover, file extensions like gif, jpeg, GIF, JPEG, jpg, Cascading Style Sheets (CSS), and JPG files can be removed from the weblog dataset. Additionally, the data which has server status request code 404 has been removed. To perform data cleaning Python programming language has been used.
- **Data Categorization**: in this stage, the weblog dataset has been categorized into two according to the existing virtual local area networks. So, the weblog dataset is categorized as student and staff VLANs. To perform this task, we have used MS-excel to identify VLANs from the IP address.
- **Attribute Selection**: after the data has been categorized, the next task is attribute selection. Based on the objective of the study some of the attributes have been selected. The researcher selected those attributes based on their importance to the research objective.
- **Data Formatting**: this phase is the last preprocessing task, which is converting the log file format to a CSV file format to make it suitable for statistical analysis and association rule discovery.

Pattern Discovery: after the data preparation has been completed. The next step is pattern discovery. In this section, web usage patterns have been discovered using statistical analysis and association rule discovery.

- **Statistical Analysis**: in statistical analysis, web users' behavior has been extracted using the statistical analysis method. To perform the statistical analysis, we have used Minitab

statistical software tool. Moreover, the statistical analysis has been conducted on staff VLANs, student VLANs, and both staff and student VLANs.

- **Association rule discovery**: in association rule mining, different rules have been discovered using Apriori and FP-growth algorithms. To perform association rule discovery, we have used Python programming language.

Pattern Analysis: in this phase, the result has been analyzed from the statistical analysis report and association rule discovery. Moreover, the analysis focused on the statistics and rules.

4.1 Data Collection

In this research, the source of data is web proxy server data. The data is collected from Wolkite University internet web users. The data is collected from the duration of February 1/2021 to April 30/2021 with a total of three months of data taken for the study. the total amount of web log data was 6,875,218 records. During the data collection the researcher, have discussed the nature of the data with the domain experts especially the network administrator for WKU.

4.1.1 Description of Collected Data

The aforementioned data was taken from the WKU proxy server, so the proxy data is implemented in Linux operating systems due to this the result of the collected data log data is formatted in Linux format can be described in the following figure.

```
1612196308.0 5006 10.194.10.40 TCP_REFRESH_HIT/304 214 GET http://www.goonernews.com/styles.css DIRECT/207.58.145.61 -
1612541908.0 2864 10.194.8.19   TCP_MISS/200 3614 GET http://www.dubmagazine.com/ DIRECT/65.61.150.253 text/html
1613665108.0 3856 10.194.8.109 TCP_MISS/302 519 GET http://mail.yahoo.com/ DIRECT/209.73.177.115 text/html
1614097108.0 1464 10.194.10.142 TCP_HIT/200 802 GET http://www.hornymatches.com/images/line_3.jpg NONE/- image/jpeg
1616515920.0 1452 10.194.1.22   TCP_IMS_HIT/304 218 GET http://www.thisdayonline.com/images/back.gif NONE/- image/gif
1615479120.0 1686 10.194.30.4   TCP_MISS/302 -1 GET http://eunis.eea.europa.eu/sites/343779 - NONE/- text/plain
1615417920.0 3657 10.194.48.1   TCP_MISS/200 -1 GET http://bio2rdf.org/pubmed:2748346 - NONE/- application/rdf+xml
1616109120.0 3667 10.194.13.29 TCP_MISS/498 -1 GET http://gutenberg.org/image/pdb:1H0A - NONE/- null
1617059520.0 116   10.194.48.64   TCP_MISS/200 559 GET http://ninemsn.com.au/images/hd/4.gif? - DIRECT/202.58.56.1 image/gif
1617145920.0 1982 10.194.11.10 TCP_MISS/200 124 GET http://gutenberg.org/image/ NONE/- null
1618528320.0 2420 10.194.2.246 TCP_MISS/302 683 GET http://cgi.sexlist.com/counter.cgi? - DIRECT/207.246.138.127 text/plain
```

Figure 4. 2 Sample proxy server log data

Table 4. 1 Web proxy server attribute description

1612282708.0 1686 10.194.13.29 TCP_MISS/200 10182 GET http://www.goonernews.com/ - DIRECT/207.58.145.61 text/Html			
No	Attribute	Description	Example
1	Date and time	This attribute shows that the time and date of the web request were made and the client request is in Linux format since the proxy server is deployed on the Linux platform. So, to make it human-understandable the timestamp is converted using Python	**161122282708.0** (2/02/2021:19:18:28)
2	Duration (Response time)	The time required to process the client request. The timer starts when the proxy server receives the HTTP request and stops when the response has been fully delivered. This helps to know how much time takes the machine to process the request	1686
3	IP-Address	A client IP address is very important to identify client web users.	10.194.13.29
4	Transactional code	This describes the requested item transaction status. It consists of two tokens separated by a slash. The first token classifies the protocol and the second is the response status code of the transaction It might be the requested object found in the cache or not. This attribute helps to know whether the requesting resource is available at the cache or not and also it tells us whether the request is denied or allowed.	TCP_MISS/200 TCP_MISS- the requested object was not in the cache. /200 status of the user request-response
5	Transfer size in byte	This indicates the number of bytes transferred to the client. It is the number of bytes that the Squid told the	10182

		TCP/IP stack to send the client.	
6	Request Method	The action that the client was trying to perform (GET, POST, HEAD)	GET
7	Accessed URL	The resource is accessed by the user. It may be an HTML page or a script	http://www.goonernews.com/
8	Client Identity	The authenticated client's user name. if you use proxy authentication, proxy-server 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 where the request has been sent (destination). The second token is the IP address of the destination	DIRECT/207.58.145.61 **DIRECT** – means squid forwarded the request directly to the origin server. 207.58.145.61-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

4.2 Data Preprocessing

The primary aim of data preprocessing is to convert the raw access log file data into a set of user profiles, as the access log files obtained via a proxy server are raw data and it is not suitable for data extraction process or statistical analysis. Even so, several data preprocessing techniques are available, depending on the objective of this study; Data categorization, data cleaning, data consolidation, attribute selection, URL selection, data transformation, and data formatting were performed in this study. Also, Text File Splitter, MS Excel 2021, and Python 3.10 were used for

data preprocessing. The figure below shows the general procedures and steps performed during data preprocessing.

4.2.1 Data cleaning

The goal of data cleaning is to remove irrelevant objects stored in the log file that may not be useful for analysis purposes. Data cleaning permits clearing out vain records which reduces the log record size to apply to a much smaller log file and facilitate upcoming tasks. The pleasant consequences are strongly relied upon in the cleansing process. Proper cleansing of records has excessive effects on the overall performance of internet utilization mining. As has been mentioned in the methodology section, Python programming has been used to clean irrelevant data and extract useful data using Python Regex package libraries. When a web user accesses an HTML document, the embedded images, if any, are also automatically downloaded and stored in the server log. For example, registry entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, Cascading Style Sheets (CSS), and JPG files can be deleted. Since the main purpose of data pre-processing is to get only the usage data, requests from the file that are not explicitly requested by the user can be eliminated. In addition, erroneous files can be removed by checking the status of the request (such as status code 404)[77] using MS-Excel. In this study, before cleaning the data, each network VLAN was identified from logs and the data was organized based on network VLAN as summarized in the following table.

Table 4. 2 Summary of unprocessed and preprocessed weblog data

No	Network VLANs	Network VLANs Name	No. of record before cleaned	No. Of record after cleaned
1	1,2,4,6,8,10,11,12,13,20,21, 30,40,42,44,45,48,50,55,60, 61,62,65,68,70,71,74,78,80, 81,86,90,91,96,100,101,110, 114,118,120	Staff VLANs && Student VLANs	6,875,218	1,440,414

Microsoft Excel has been used to split the VLAN from the IP address to categorize the data into two parts according to the VLAN. To categorize the VLAN whether the VLAN is staff or student

we have used Python to select those VLANs which has been categorized under the student as well as the staff. Therefore, similar VLANs have been merged as students or staff.

	A	B	C	D	E	F	G	H
1	Unix-time	Response-time	IP-address	VLAN	Transactional-code	Byte-transferred	Methods	URL
2	1612196308	5006	10.194.10.40	10	TCP_REFRESH_HIT/304	214	GET	http://www.goonernews.com/ styles.css/DIRECT/207.58.145.61 text/html
3	1612541908	2864	10.194.8.19	8	TCP_MISS/200	512	GET	http://www.dubmagazine.com/ DIRECT/65.61.150.253 text/html
4	1613665108	3856	10.194.8.109	8	TCP_MISS/304	210	GET	http://www.yahoo.com/ DIRECT/209.73.186.238 text/html
5	1614097108	1454	10.194.10.42	10	TCP_MISS/200	1024	GET	http://www.hornymatches.com/images/hot_on_h.gif / NONE/- image/gif
6	1616515920	1452	10.194.1.22	1	TCP_MISS/200	854	GET	http://www.clitical.com/images/down2.gif - NONE/- image/gif
7	1615547920	1686	10.194.30.4	30	TCP_MISS/304	145	GET	http://mail.yahoo.com/ DIRECT/209.73.177.115 text/html
8	1615417920	3657	10.194.48.1	48	TCP_MISS/302	256	GET	http://www.google-analytics.com/urchin.js NONE/- text/javascript
9	1616109120	3667	10.194.13.29	13	TCP_MISS/304	233	GET	http://www.goonernews.com/graphics/newslogo.gif / DIRECT/207.58.145.61 -
10	1617145920	116	10.194.11.10	11	TCP_MISS/408	477	GET	http://www.gutenberg.org/dirs - NONE/- application/rdf+xml
11	1617069520	1982	10.194.2.245	2	TCP_MISS/200	452	GET	http://bio2rdf.org/uniprot:P00807_16 - NONE/- application/rdf+xml
12	1618528320	2420	10.194.13.29	13	TCP_MISS/200	530	GET	http://eunis.eea.europa.eu/species/175881 - NONE/- text/plain

Figure 4. 3 VLAN Identification

To identify the VLAN number from the IP address we have used MS-excel, and first, the file is opened in excel during file opening the column value is separated by using the dot operator. Then after all the cell values should be split based on the (.) dot criteria. As a result, the final VLAN identification is shown in figure 4.3.

Furthermore, because the proxy server runs on the Linux platform, the original weblog date-time format should be changed to a human-readable format to identify the monthly data from the dataset. As displayed in the aforementioned figure the Linux time format is like 161122282708.0 so, to change the human-understandable format the researchers have used the Python date-time conversion method to perform such operations we have used the Python built-in module. To convert Linux time format using Python, first, we have to import data time library packages. The

following is the general formula to convert Linux time to a human-readable date format. For example, a timestamp for ***161122282708.0***

D = datetime.fromtimestamp(161122282708.0) ==> 2/02/2021:19:18:28

4.2.2 Attribute Selection

After the log file has been cleaned, attribute selection is very essential to make statistical evaluation and association rule discovery. As it's been defined above, weblog data has a variety of attributes. But, taking into account the objective of the research, the researcher selects only imperative attributes such as date and time, client IP address, and URLs. In addition to the current attribute-derived attributes such as network VLANs and site category have been selected based on [4][78]. The reason to select these attributes is that these attributes must be used to conduct statistical analysis and association rule discovery. For instance, ***date-time*** is used to identify the Unix-time date to human readable date format. As a result, we have selected this attribute. The ***IP address*** was selected as an important attribute because it helps to identify the web users and is used to identify the VLANs that is why we have selected it as an important attribute. ***URLs*** this the most important attribute because it represents the websites that have been accessed by the users. ***VLAN*** attribute is also important to identify and categorize the web users as student and staff users due to this we have used and selected it as an important attribute.

But, due to the web users accessing a variety of URLs for their interest it was difficult to take into account select all available accessed URLs, therefore it was necessary to select only the most significant URLs from those listed URLs. to select the most frequent URLs in Microsoft Excel, an excel pivot query has been used, and the result of each URL frequency has been calculated and based on the minimum and maximum frequency distribution and the result has been displayed using the pivot table in descending order by looking forward to each URL's frequency. As a result, the researcher decided to select only 41 URLs found at the top of frequent URLs. due to this, from the total of 1,440,414 cleaned weblog data, we have used 185,627 since some of the URLs other than 41 URLs were removed from the cleaned weblog data.

4.2.3 Data Categorization

Requests from web users are routed through the proxy server. As a result, it is possible to identify the web request from which the network VLAN can be categorized. Furthermore, the web request can be identified whether the request is originating from the student or staff member based on the existing network VLAN. Data categorization has been summarized as directed by domain experts as summarized in the following table.

Table 4. 3 Categorized VLANs

No	Web user type	Network VLANs	Total Data
1	Staff VLANs	1,2,4,6,8,10,11,12,13,20,21,30,40,42,44,45,48,50,55,60	104,768
2	Student VLANs	61,62,65,68,70,71,74,78,80,81,86,90,91,96,100,101,110,114,118,120	80,859
Total			185,627

4.2.4 Data Formatting

The weblog data has been modified numerous times to accommodate data preparation, data cleaning, identifying VLANs, and selecting the top frequent URLs for statistical analysis and association rule mining. Further, the final data is converted to CSV and similar URLs have been converted into a common standard URL. For example, if a user accesses the Twitter webpage through any means, the request has been translated into a common URL www.twitter.com. The following table describes the most frequented sites, their common converted URLs, site category, and descriptions are summarized as follows.

Table 4. 4 frequently accessed websites and their representation

URLs	Domain Name	Description of Domain Name	Site Category
URL1	www.facebook.com	The user who accessed Facebook contents	Social media
URL2	www.youtube.com	The user who accessed YouTube website	Entertainment
URL3	www.gmail.com	The user who accessed Gmail website	Email
URL4	www.sodere.com	The user who accessed sodere website	Entertainment
URL5	www.tutorialspoint.com	The user who accessed tutorials point website	Educational
URL6	www.pogo.com	The user who accessed pogo website	Entertainment
URL7	www.slideshare.net	The user who accessed SlideShare website	Educational
URL8	www.freepornforu.com	The user who accessed porn website	Entertainment
URL9	wikipedia.org	The user who accessed Wikipedia website	Educational
URL10	www.newadvent.org	The user who accessed new advent website	Educational
URL11	www.microsoft.com	The user who accessed Microsoft contents	Educational
URL12	www.ethiojobs.net	Any user who accessed job-seeker website	Job Search
URL13	myflixer.to	The user who accessed movie contents	Entertainment
URL14	www.ethiopianorthodox.org	Any user who accesses orthodox website	Educational
URL15	www.twitter.com	Any user who accesses twitter website	Social Media
URL16	www.medscape.com	The user who accessed Medscape website	Educational

URL17	www.abebooks.com	The user who accessed book website	Educational
URL18	et1lib.org	The who accessed book repository website	Educational
URL19	www.yahoo.com	The user who accessed yahoo website	Email
URL20	mereja.com	The user who accessed mereja website	News
URL21	www.linkedin.com	The user who accessed LinkedIn website	Social Media
URL22	www.researchgate.net	Any user visited research website	Educational
URL23	yts.mx	Any user who accesses movie website	Entertainment
URL24	getintopc.com	The user who accessed software website	Software Site
URL25	www.freeprojectz.com	The user who accessed source code website	Educational
URL26	www.cisco.com	The user who accessed cisco contents	Educational
URL27	www.wheresthematch.com	The user who accessed football website	Entertainment
URL28	www.arifzefen.com	The user who accessed music website	Entertainment
URL29	www.adobe.com	The user who accessed adobe website	Educational
URL30	www.wku.edu.et	The user who accessed WKU website	Governmental
URL31	www.sciencedirect.com	The user who accessed research contents website	Educational
URL32	www.talkenglish.com	Any user who accessed talk English website	Educational
URL33	www.coursera.org	Any user visited coursera website contents	Educational

URL34	www.w3schools.com	The user who accessed W3school content	Educational
URL35	sourceforge.net	Any user visited source forge website contents	Software Site
URL36	www.listscholarship.com	The user who accessed scholarship website	Educational
URL37	etd.aau.edu.et	The user who accessed AAU repository	Educational
URL38	www.gutenberg.org	The user who accessed Gutenberg website	Educational
URL39	www.good-amharic-books.com	The user who accessed amharic-book website	Educational
URL40	www.tribalfootball.com	The user who accessed football website	Entertainment
URL41	bio2rdf.org	The who accessed bio2rdf website contents	Educational

CHAPTER FIVE
EXPERIMENTATION AND ANALYSIS

After the completion of data preparation, the next phase presents the detailed procedures followed during the experiments. In this study, the experiment is conducted using statistical analysis and data mining techniques such as association rule mining to discover web usage navigational patterns in WKU web usage data. To experiment, we have used Minitab 21.1 tool as a statistical analysis tool and Python 3.10 for association rule mining.

5.1 Experiment Setup

The experiment has a total of nine different experiments using a total of 185,627 preprocessed weblog datasets. During the experiment, the dataset is categorized as staff, and student datasets, and a combination of both staff and student datasets has been prepared based on the objective of the study. All experiments were conducted for two different purposes, first for statistical analysis and second for association rule discovery. In each dataset category, we have conducted three different experiments. Generally, the experimental setup has been described in the following way:

- Experiment I: a statistical analysis using Minitab with students' weblog dataset
- Experiment II: a statistical analysis using Minitab with staffs' weblog dataset
- Experiment III: a statistical analysis using Minitab with all weblog dataset
- Experiment IV: association rule mining using the Apriori algorithm with students' weblog dataset
- Experiment V: association rule mining using FP-growth algorithm using students' weblog dataset
- Experiment VI: association rule mining using Apriori algorithm using staffs' weblog dataset
- Experiment VII: association rule mining using FP-growth algorithm using staffs' weblog dataset
- Experiment VIII: association rule mining using Apriori algorithm using all weblog dataset
- Experiment IX: association rule mining using FP-growth algorithm using all weblog dataset

5.2 Statistical Analysis

The most frequent strategy for extracting descriptive facts about user navigational behaviors of web resources is to employ statistical analysis techniques. Minitab 21 statistical analysis tool has been used to analyze and investigate user navigation behaviors of WKU internet users. in this study, a summarized statistical web analytics report is generated mainly on frequently accessed sites, frequently accessed categories of sites, and each VLANs web traffic the detailed statistical analysis is described below for each student and staff categorical dataset

Experiment I: Statistical Analysis using students' weblog dataset

As it was described in previous chapters four and three, the weblog dataset is categorized according to the nature of the existing VLANs system in WKU. For the sack of this experiment, 80,859 weblog datasets have used to extract statistical reports about WKU student web users. The figure below illustrates the top frequently accessed sites and their categorical sites which are accessed by WKU student internet users.

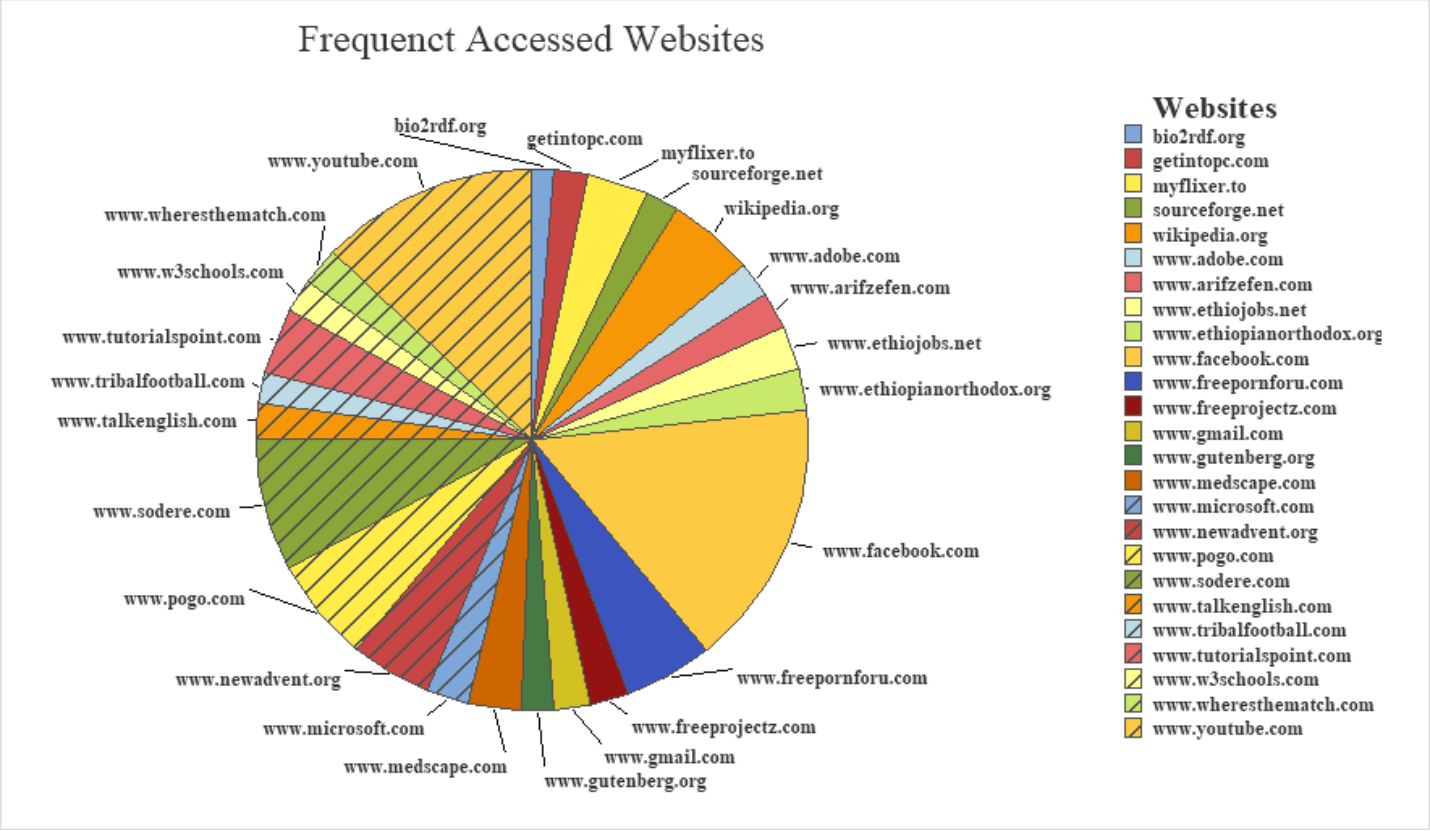


Figure 5. 1 Top Frequent Accessed Websites by Students'

From the above figure, we have been observing that Facebook, YouTube, sodere.com, and freepornforu.com were the top frequent accessed sites by the student. Whereas Wikipedia, new advent, and tutorials point, were the second most frequently accessed sites as compared to the remaining sites. Therefore, students have an interest in accessing Facebook, YouTube, and freepornforu as the priority whereas pogo, Wikipedia, new advent, and tutorial point websites are the second priority interest as compared to the remaining accessed websites.

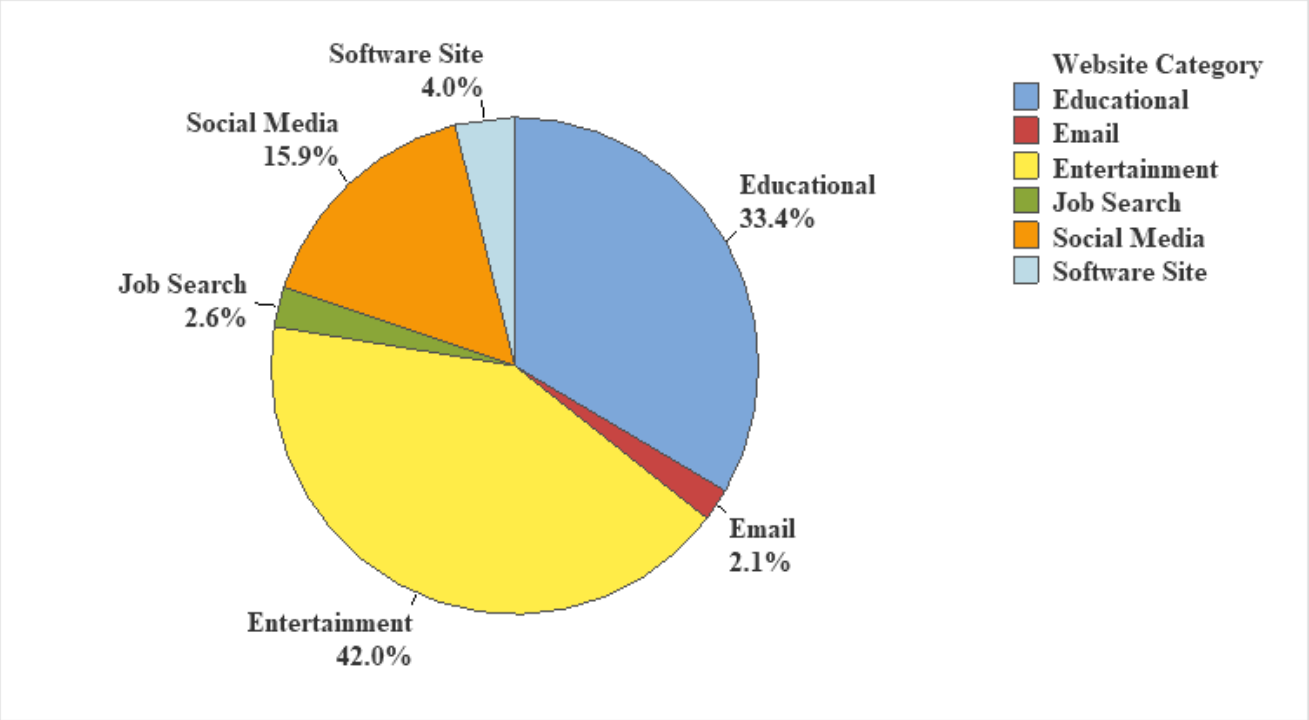


Figure 5. 2 Statistical reports for categorized accessed sites by student

Figure 5.2, depicts that Entertainment, educational, and social media categorized sites were the top three categorized sites accessed in the student's dataset. According to the aforementioned graph, most of the time the students' priority interest is accessing Entertainment websites first, educational websites second, and social media websites as the third priority than the other categorized websites. Therefore, accessing Entertainment sites is the top priority in students' weblog dataset as compared to the other sites and this shows that most of the time the students' priority interest is accessing entertainment websites rather than accessing the remaining categorized websites. However, in figure 5.1 Facebook is the first top accessed website among the listed accessed websites and YouTube is the second most accessed website next to Facebook, but,

in the case of categorized websites as shown in figure 5.2 Facebook is accessed in the second priority since Facebook is categorized under the social media, whereas YouTube is accessed in the priority due to YouTube is categorized under entertainment websites.

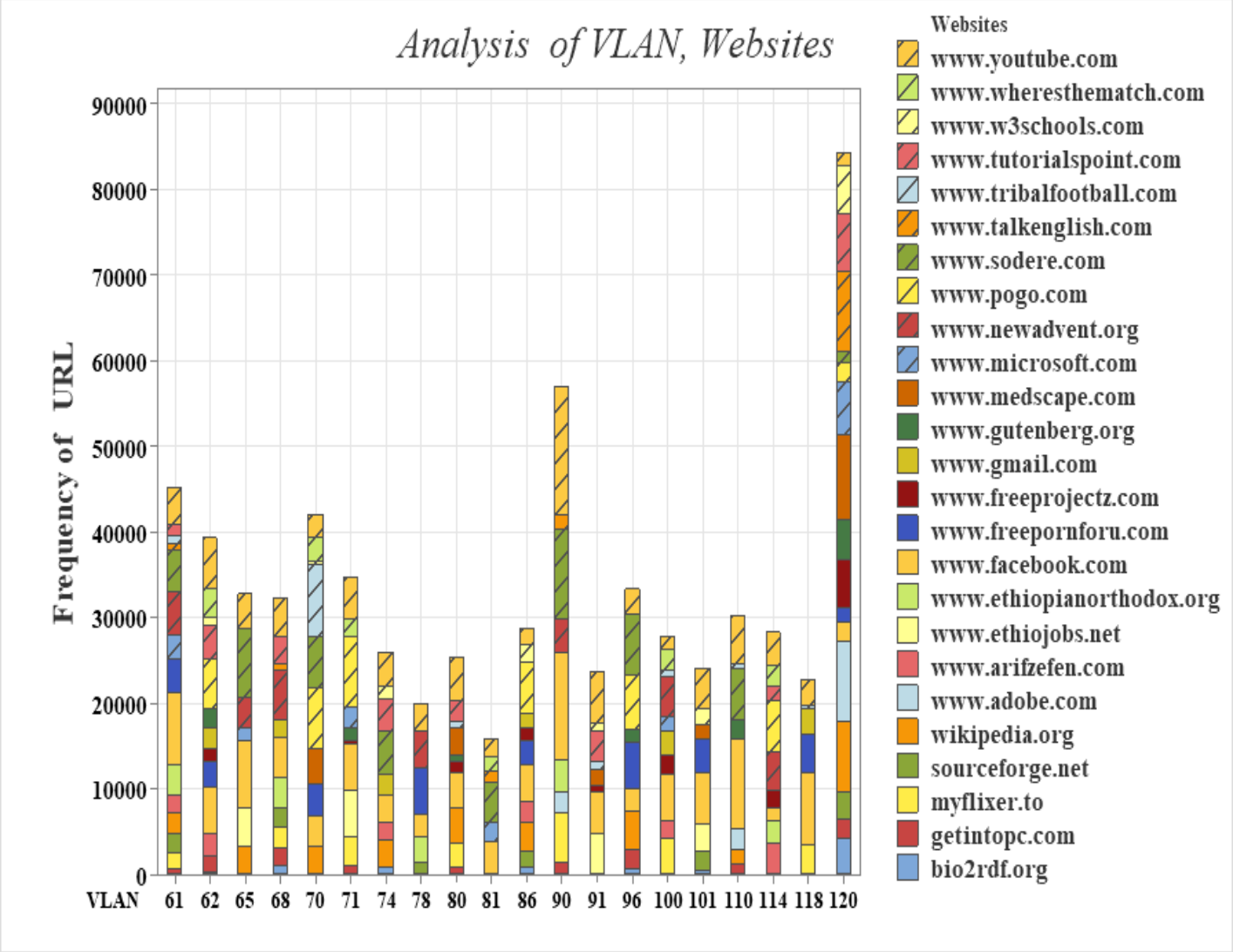


Figure 5. 3 Top frequently accessed sites in students' VLAN

As has been demonstrated in above figure 5.3, some of the listed websites are accessed in certain VLANs and some of the websites are not accessed in certain VLANs. For example, websites, pogo.com, medscape.com, ethiojobes.net, freeprojectz.com, wherethematch.com, adobe.com, w3schools.com, gutenberg.org, and bio2rd.org are not accessed in VLAN (61), on the contrary, some these websites are accessed in some of the VLANs. For example, Medscape.com and wherethematch.com websites are accessed in VLAN (70). Therefore, each VLAN has its URL property.

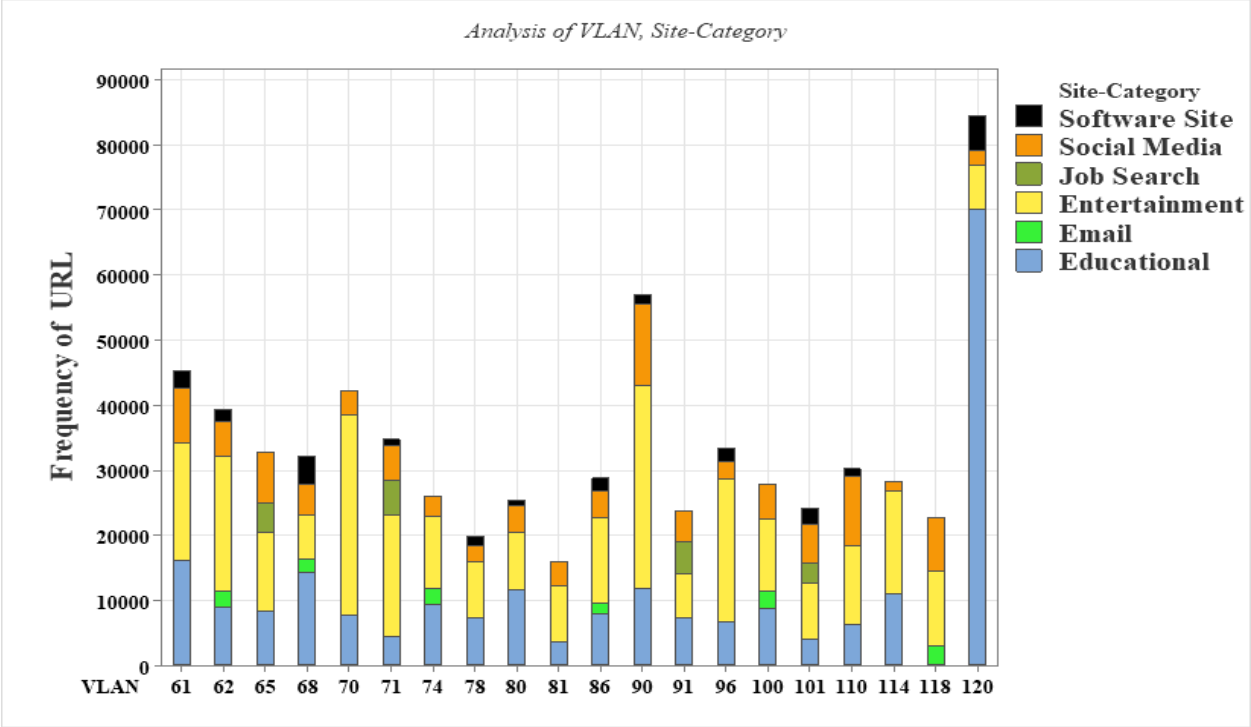


Figure 5. 4 Statistical analysis for site category in students' VLAN

In the second figure 5.4, we have observed that, in some, of the VLANs, educational websites have been accessed, and in some of the VLANs entertainment and social media websites have been accessed. Firstly, in VLANs (61, 62, 65, 70, 71, 74, 78, 81, 86, 90, 96, 100, 101, 110, 114, 118) Entertainment website have been accessed more frequently. Because of this, most of these VLAN users were in students' class room computer laboratory. Secondly, in VLANs (68, 80, and 120) Educational websites have been accessed more frequently next to entertainment sites. Because such VLANs are primarily designed for library computer laboratories for students. As a result, students in this room were not allowed to browse social media or entertainment websites other than educational websites. According to this statistical report we have observed the fact that most of the time Entertainment websites have been accessed in the students' VLANs specifically in students' laboratory VLANs. However, in some of the VLANs educational websites has been accessed specifically in library laboratory since social media or entertainment website has been restricted in those VLANs. In terms of website traffic, some of the VLANs have the highest web traffic occurred, on the other hand, some of the VLANs have low web traffic. For example, VLANs such as (120, and 90) have more web traffics since the number of accessed websites is

high as compared to the other, and VLANs (61, 70) are also second VLANs which have high web traffic next to VLAN (120, 90). On the other side, some of the VLANs have low web traffic has occurred, for example, VLANs (78, 81) have low web traffic since the number of accessed URLs in a minimum number of requests as compared to the other VLANs.

Experiment II: Statistical Analysis Using Staff's weblog dataset

For this experiment, we have used a total collected weblog dataset of 104,768. The detailed statistical reports are described as follows.

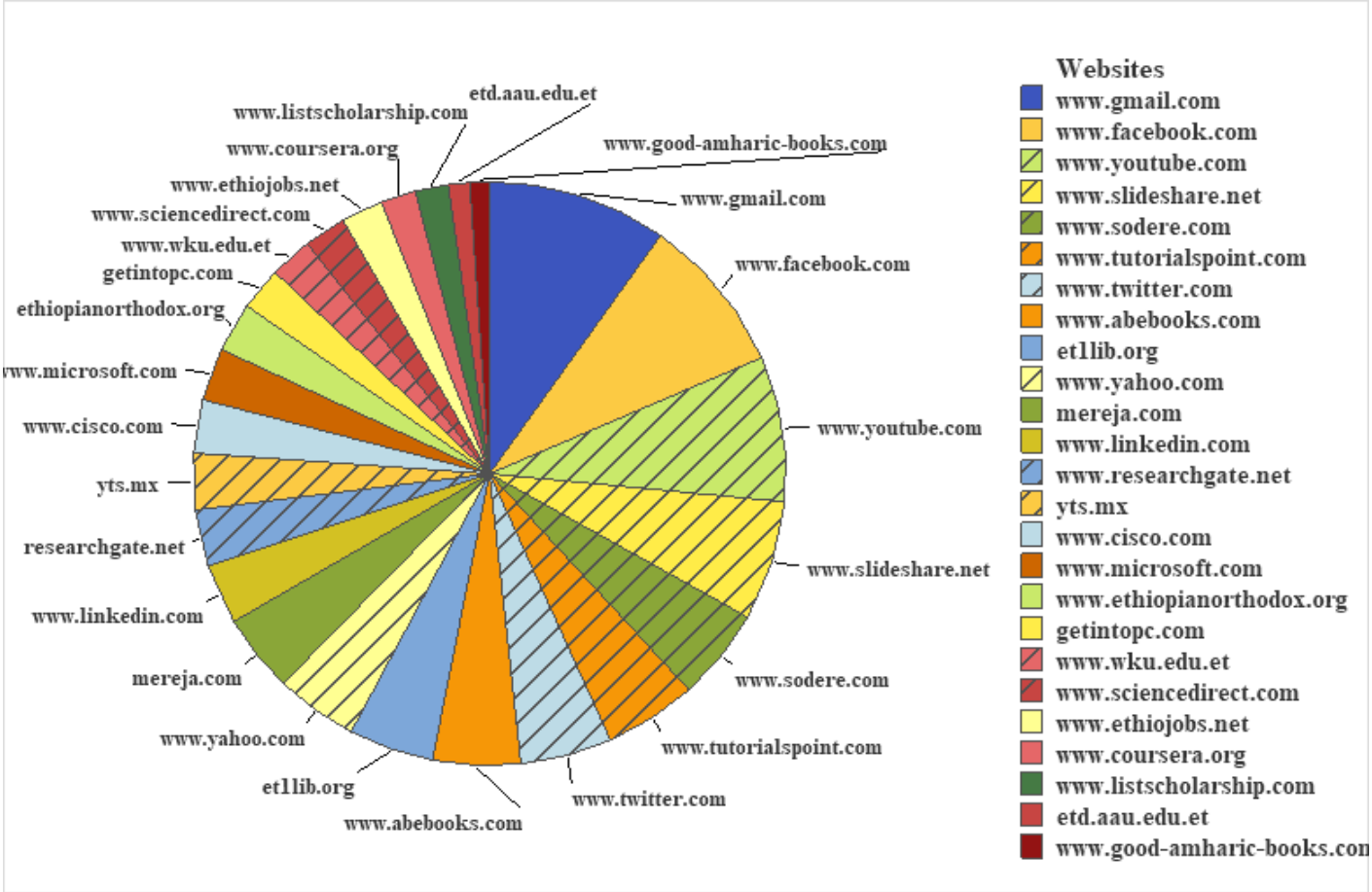


Figure 5. 5 Statistical reports on top frequent accessed sites by staff'

From the above figure 5.5, we have observed that websites such as gmail.com, facebook.com, youtube.com, and slideshare.net are the top most frequently accessed websites in the staff's dataset. Whereas websites such as sodere.com, tutorialspoint.com, tiwtter.com, abebooks.com, etlib.com, yahoo.com, and mereja.com are the second most frequently accessed websites in this scenario. This shows that most of the time the staff users are more priority interested in accessing

the Gmail website at first, Facebook second, and YouTube thirdly. On the other hand, slideshare.net, sodere.com, tuturialspoint.com, tiwtter.com, abebooks.com, etlib.com, yahoo.com, and mereja.com websites are the next interesting sites accessed by the staff dataset next to Gmail, Facebook, and YouTube.

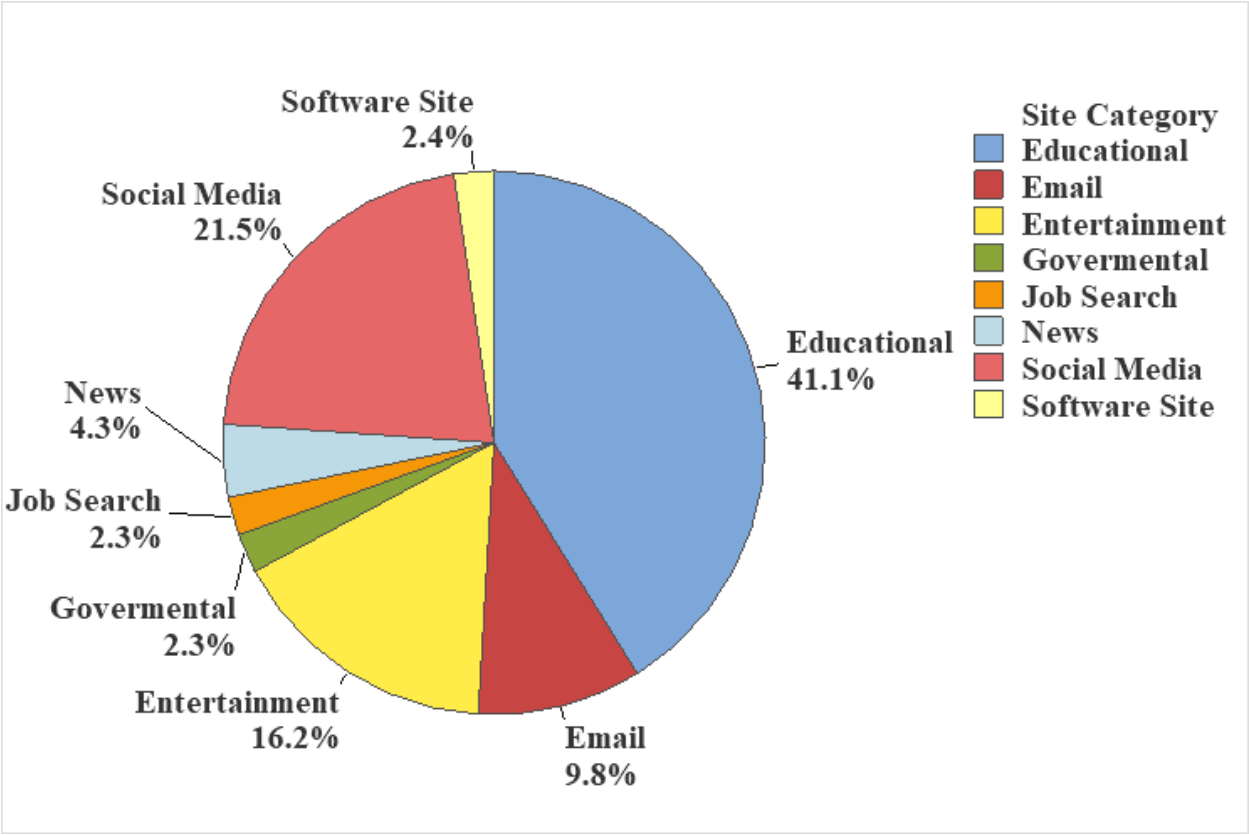


Figure 5. 6 Statistical reports on frequently accessed site category in staff dataset

In the above figure 5.6, we have observed that websites are categorized under a different category as shown in the aforementioned figure. According to the statistical report in the above figure, educational websites have been accessed as the priority in the staff dataset, whereas social media and Entertainment websites have been accessed as second and thirdly interest sites accessed in the staff dataset. However, as mentioned in previous figure 5.4, most of the time staff users are accessing Gmail sites as the priority, but when sites are categorized under some category based on the nature of the site as it has been listed in figure 5.6, Email websites are located at the fourth interesting accessed websites in the staff users. in general, from the above two figures, the statistical analysis, figure 5.5 shows that Gmail is the most priority interest website in staff web users whereas Facebook and YouTube are the second most frequent website accessed by the staff

users. but when the website is categorized under a variety of categories, educational websites are the top most frequently accessed websites. On the contrary, social media, Entertainment, and Email websites are the second most frequent websites accessed in the staff dataset.

In general, from the staff VLAN, we have discovered the fact that first, most of the time the users' priority interest is in accessing educational websites among the listed accessed site categories. Secondly, the users' interest is more focused on accessing social media websites next to educational websites. Thirdly, based on the figure report staff users' have been accessing Entertainment websites as a third interest.

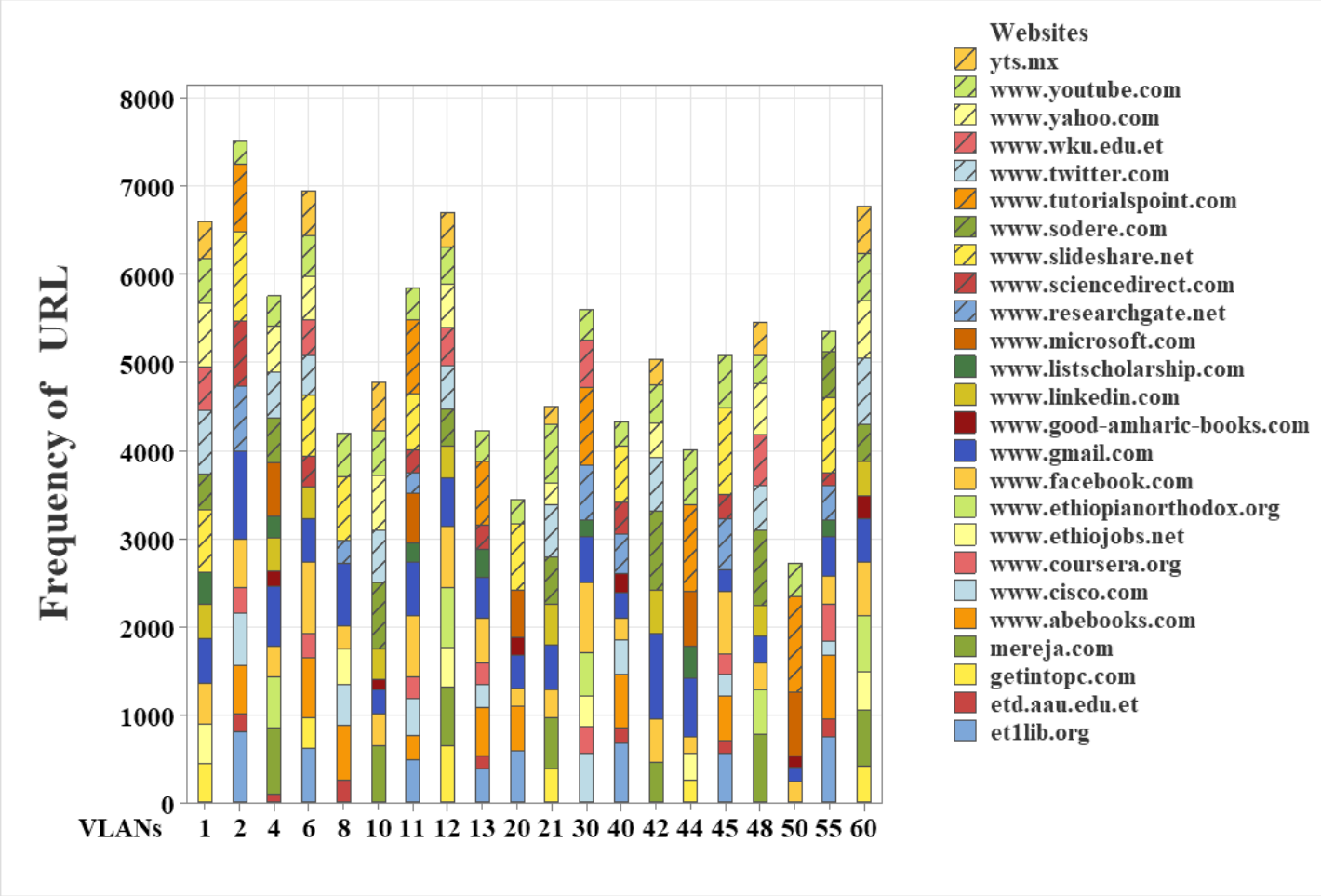


Figure 5. 7 Top frequent accessed sites in staffs' VLAN

In the above figure 5.7, we have observed that some of the websites have been accessed in certain VLANs and some websites haven't been accessed in certain VLANs. For example, in VLAN (50) websites such as Facebook, Gmail, Microsoft, Tutorials point, and YouTube websites have been

accessed in this VLAN. However, some of these websites haven't been accessed in certain VLANs. For instance, the Tutorial's point hasn't been accessed in VLAN (55, 60).

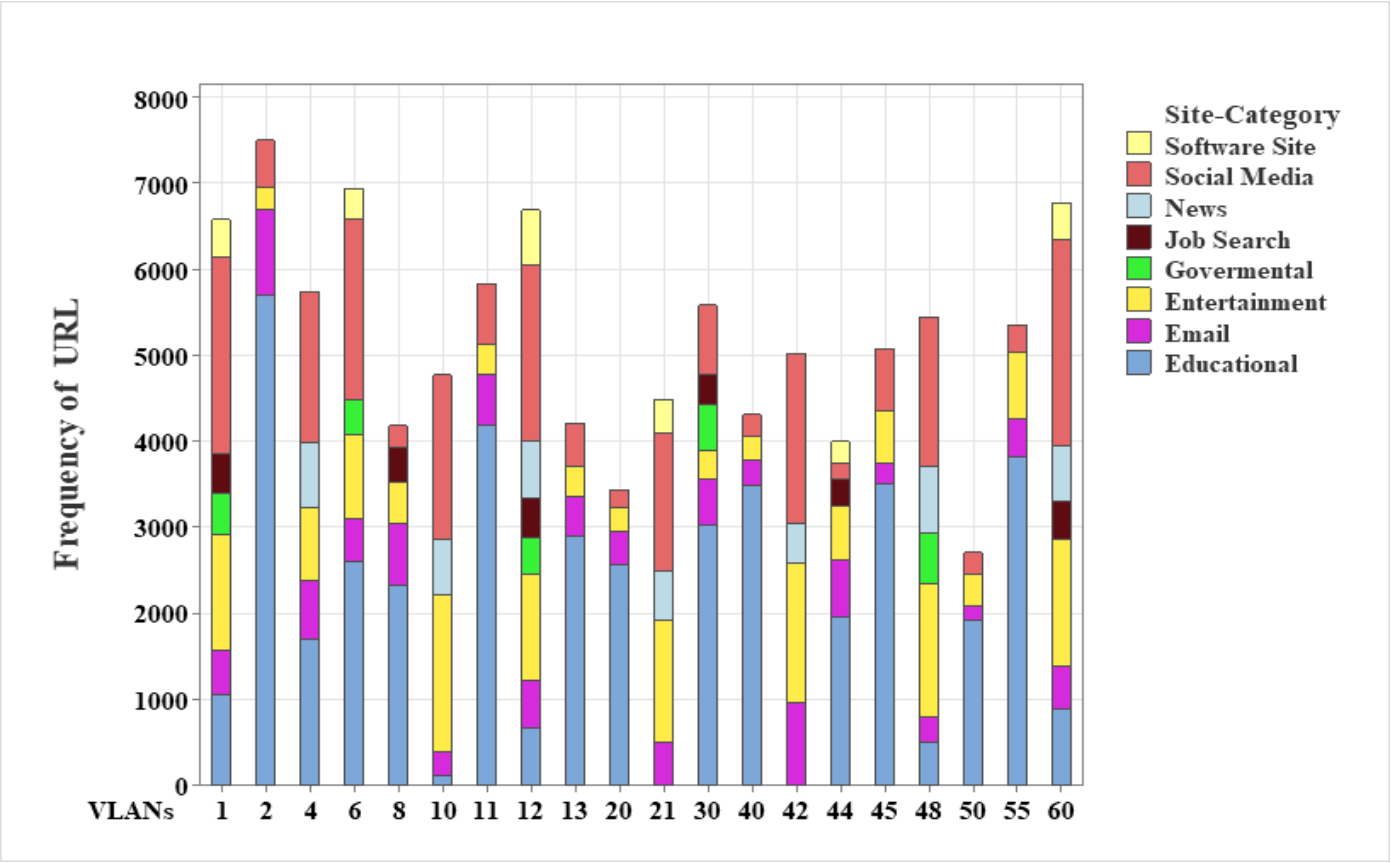


Figure 5. 8 Frequently accessed site category in staff VLANs

Figure 5.8, shows what we've discovered, in some of the VLANs Educational websites have been more frequently accessed, on the contrary, Entertainment and social media have been accessed more frequently in some of the VLANs. For instance, in VLANs such as (2, 6, 8, 11, 13, 20, 30, 40, 44, 45, 50, 55) Educational websites have been accessed more frequently. This is due to that, those VLANs are primarily designed for college VALNs. As a result, most of the time, the staff users primarily accessing educational websites. On the contrary, in VLANs (1, 10, 12, 21, 42, 48, 60) Social media websites have been accessed more frequently. Because, most the users in those VLANs were administrative and management since the VLANs were primarily designed for admin and management web users. In this scenario what we have discovered is, firstly, most of the time staff users' have more interest in accessing educational websites in certain VLANs

specifically in college VLANs. Secondly, staff users' have more interest in accessing social media websites in certain VLANs specially in administrative and management VLANs.

In terms of internet traffic supplies, VLAN (2) is the first high VLAN during which extremely web traffic source is incomed since the number of web resources in this VLAN is high as compared to the other. Whereas VLANs such as (1, 6, 12, 60) have the second listed VLANs in which high web traffic source has occurred next to VLAN (2). On the other hand, there are some VLANs in which low web traffic is incomed as compared to the other VLANs. For instance, in VLANs (50, 20) low web traffic is occurred according to the statistical report in figure 5.8.

Experiment III: a statistical analysis using Minitab with all weblog dataset

In this, the experiment is conducted using the total of 185,627 dataset with total of 41 URLs. The detail statistical report has been described in the following.

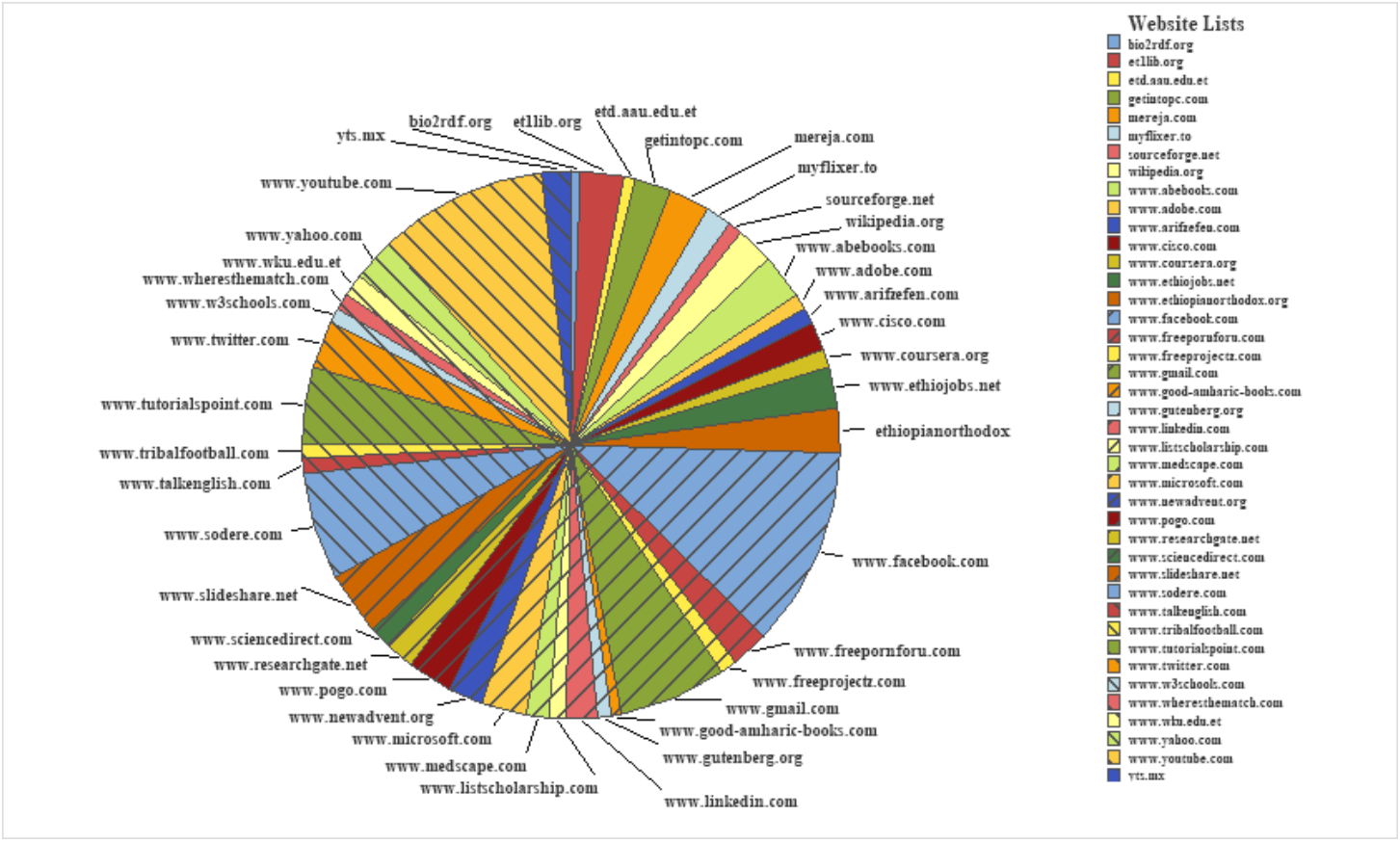


Figure 5. 9 Frequently accessed websites in all weblog dataset

From figure 5.9 we have observed that most of the time the users' primary interest is accessing Facebook, YouTube, Gmail, and Sodere websites has been accessed in the first whereas, tutorialspoint.com and slideshare.net websites have been accessed as second interest as compared to the other websites.

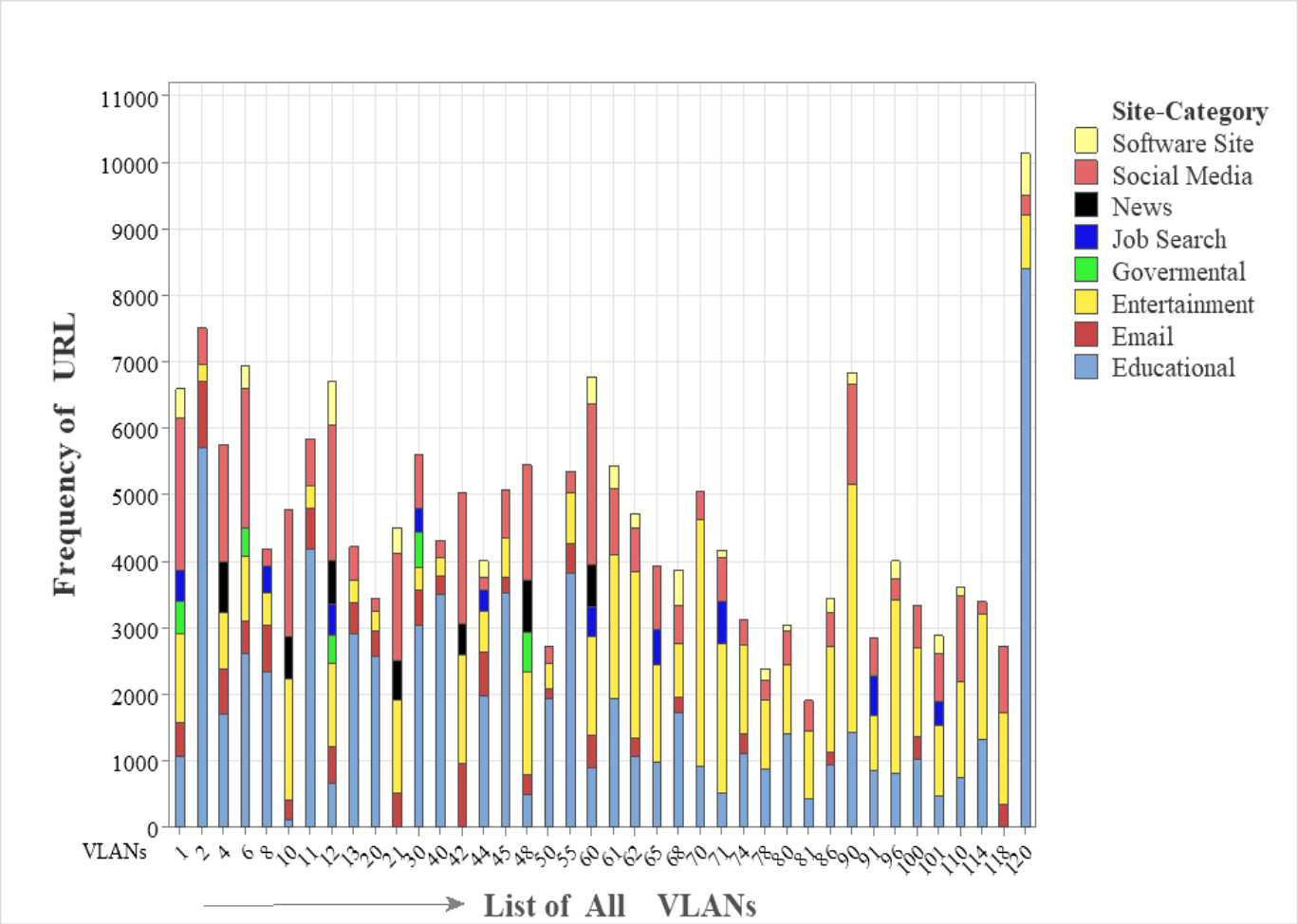


Figure 5. 10 Frequently categorical websites accessed in total weblog dataset

From figure 5.10 we observed the fact that from the total weblog dataset in some of the VLANs educational websites have been accessed most frequently. Whereas in certain VLANs entertainment and social media websites have been accessed most frequently. For instance, in VLANs (1, 4, 10, 12, 21, 42, 48, 60), in these VLANs, social media websites have been accessed more frequently as compared to the remaining VLANs. This is because most of these VLAN

users were administrative and management staff users. Whereas VLANs (61, 62, 65, 70, 71, 74, 78, 81, 86, 90, 96, 100, 101, 110, 114, 118), in these VLANs entertainment websites have been accessed more frequently. This is because these VLANs were in students' computer laboratory rooms. On the other hand, VLANs (2, 6, 8, 11, 13, 20, 30, 40, 44, 45, 50, 55, 68, 80, 91, 120), in these VLANs educational websites have been accessed more frequently as compared to the remaining VLANs. This is because some of the listed VLANs (2, 6, 8, 11, 13, 20, 30, 40, 44, 45, 50, 55) were college VLANs in the staff dataset. Whereas VLANs (68, 80, 91, 120) were student library computer laboratories, social media, and entertainment websites that have restricted access to those VLANs.

In terms of, web traffic we observed that certain VLANs have more web traffic, and on the contrary, certain VLANs have low web traffic in terms of the number of resources accessed in each VLAN. For instance, VLANs (120) have more web traffics as compared to the remaining VLANs since the number of accessed resources in this VLAN is high as compared to the other VLANs. Whereas VLANs (2, 6, 11, 60, 90, 1) were the second VLANs that have more web traffics next to VLAN (120). On the contrary, VLAN (81) is the list VLAN that low web traffic as compared to the remaining VLAN web traffics, since the number of resources accessed in VLAN (81) is small as compared to the other VLANs.

5.3 Association Rule Discovery

After the statistical analysis is conducted, the next phase is pattern discovery through association rule mining to discover websites that have been accessed frequently together by the web users. as stated in the experimental setup session, six experiments are conducted using association rules in Python environment. The web log data has been transformed into CSV format to read in the Python environment. To perform the association rules, the FP-growth algorithm and Apriori algorithm have been used on the staff, and student dataset, and using the total weblog dataset. For this experiment, all URLs that have been used in statistical analysis are incorporated under this experiment. During this phase parameters such as minimum support and minimum confidence, values have been used throughout the experiments.

Experiment IV: association rule discovery in students' dataset using the Apriori algorithm

In this experiment, we have used the Apriori algorithm to discover association rules in students' weblog datasets. As mentioned above the experiments are conducted in the Python environment

different Python package libraries such as **Pandas** package, **mlxtend** package, and **NumPy** packages have been used to conduct association rule mining. In this experiment total of 17,462 weblog datasets have been used. The reason that the difference in records of data during statistical analysis and this experiment is due to that transaction item sets are recorded in the form of a transactional dataset that is suited for the Python library.

Before conducting the experiment, we have to import important Python libraries as follows:

```
import pandas as pd
```

```
import numpy as np
```

```
from mlxtend.frequent_patterns import apriori,fpgrowth,association_rules
```

```
from mlxtend.preprocessing import TransactionEncoder
```

Figure 5. 11 Python libraries used for association rule mining

After importing these packages, the next step is importing the dataset.

```
transaction_data=[['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
```

Figure 5. 12 Transactional datasets in students' weblog data

Each row represents a transaction item sets in the students' weblog dataset. Before finding the most often occurring itemsets, we must first convert our dataset into a True - False matrix, with rows representing transactions and columns representing URLs. finally, the dataset has two values

True- represents the transaction that contains the item and ***False*** value represents the transaction that does not contain the item.

data_encode = TransactionEncoder() – initializing the transaction encoder to perform transactional encoding for transactional item sets in students' weblog dataset, and where ***data_encode*** *is a variable name.*

td = dtata_encode.fit_transform(transaction_data) – used to transform transactional encoded item sets using ***Numpy array***, and where ***td*** is the variable to hold the respective ***Python*** code**.** The final result is described in the following figure.

```
td
```

```
array([[False,  True, False, ..., False,  True,  True],
       [False,  True, False, ..., False,  True,  True],
       [False,  True, False, ..., False,  True,  True],
       ...,
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False]])
```

Figure 5. 13 Sample of transformed data in students' transactional dataset

After binarizing the transactional dataset, the next step is to transform transactions into a data frame using a Pandas Python module. For this purpose, the following line code has been used.

data_frame=pd.DataFrame(td,columns=dtata_encode.columns_)-used to transform transactions into data frame. The result is shown below.

```
data_frame
```

	adobe	arifzefen	bio2rdf	ethiojobs	ethiopianorthodox	facebook	freepornforu	freeprojectz	getintopc	gmail	...	pogo	sodere	sourceforge	talkenglish
0	False	True	False	False	True	True	True	False	True	False	...	False	True	True	True
1	False	True	False	False	True	True	True	False	True	False	...	False	True	True	True
2	False	True	False	False	True	True	True	False	True	False	...	False	True	True	True
3	False	True	False	False	True	True	True	False	True	False	...	False	True	True	True
4	False	True	False	False	True	True	True	False	True	False	...	False	True	True	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17457	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
17458	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
17459	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
17460	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False
17461	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False

17462 rows × 25 columns

Figure 5. 14 Binarized transactional dataframe in students' weblog dataset

After this, the next step is generating the frequent itemsets and association rules using specified minimum support and confidence value. After performing different experiments with different support and confidence values we have selected the minimum threshold of 0.1 as minimum support and 0.9 as the minimum confidence value. The reason to select the 0.1 support value is first, it supports the statistical analysis result. Second, based on the objective of the study we have selected the value to enhance the final rule result.

Apriori_frequentItems=apriori(data_frame,min_support=0.1,use_colnames=True). Used to generate frequent item sets with minimum support 0.1 in Apriori algorithm. The result described as follows.

Table 5. 1 Frequent item sets using Apriori algorithm in student dataset

No	Support	Frequent Item sets
1	0.101936	(arifzefen)
2	0.121406	(ethiojobs)
3	0.113847	(ethiopianorthodox)
4	0.736399	(facebook)
5	0.240522	(freepornforu)
6	0.105143	(freeprojectz)
7	0.143168	(medscape)
8	0.114763	(microsoft)
9	0.165502	(myflixer)
10	0.220479	(newadvent)
11	0.286565	(pogo)
12	0.367655	(sodere)
13	0.183255	(tutorialspoint)
14	0.104055	(wheresthematch)
15	0.235769	(wikipedia)
16	0.593746	(youtube)
17	0.121292	(ethiojobs, facebook)
18	0.115336	(ethiojobs, youtube)
19	0.101707	(ethiopianorthodox, facebook)
20	0.113847	(newadvent, ethiopianorthodox)
21	0.113847	(ethiopianorthodox, youtube)
22	0.199061	(freepornforu, facebook)
23	0.165502	(facebook, myflixer)
24	0.179361	(newadvent, facebook)
25	0.171	(facebook, pogo)

26	0.299278	(facebook, sodere)
27	0.147635	(facebook, tutorialspoint)
28	0.180678	(wikipedia, facebook)
29	0.530581	(facebook, youtube)
30	0.115909	(freepornforu, pogo)
31	0.101878	(wikipedia, freepornforu)
32	0.179418	(freepornforu, youtube)
33	0.101077	(wikipedia,medscape)
34	0.143626	(myflixer, youtube)
35	0.167907	(newadvent, youtube)
36	0.160291	(pogo, youtube)
37	0.136468	(wikipedia, sodere)
38	0.257473	(youtube,sodere)
39	0.147406	(tutorialspoint, youtube)
40	0.162639	(wikipedia, youtube)
41	0.115222	(ethiojobs, facebook, youtube)
42	0.101707	(newadvent, ethiopianorthodox, facebook)
43	0.101707	(ethiopianorthodox, facebook, youtube)
44	0.113847	(newadvent, ethiopianorthodox, youtube)
45	0.173348	(freepornforu, facebook, youtube)
46	0.143626	(facebook, myflixer, youtube)
47	0.148379	(newadvent, facebook, youtube)
48	0.140076	(pogo,facebook,youtube)
49	0.122609	(wikipedia, facebook, sodere)
50	0.249399	(youtube, facebook, sodere)
51	0.141908	(facebook, tutorialspoint, youtube)
52	0.160062	(wikipedia, facebook, youtube)
53	0.121406	(youtube, wikipedia, sodere)
54	0.101707	(newadvent, ethiopianorthodox, facebook, youtube)
55	0.118829	(youtube, wikipedia, facebook, sodere)

After the frequent item sets have been generated, the next step is identifying the rules using the Apriori algorithm from the frequent item sets based on the minimum confidence value. To perform this we have used the following Python line of codes.

Apriori_rule = association_rules(Apriori_frequentItems,metric='confidence',min_threshold=0.9). where the ***Apriori_rule*** – is the variable, ***association_rules*** is the predefined function used to generate rules, and ***Apriori_frequentItems*** is the variable that has been used to generate the frequent item sets.

Finally, a total of 25 rules have been generated and the detailed rules have been described in Appendix. from these rules, we have selected the top rules which have a confidence value of 100% of the total rules. Therefore, the following rules have been selected. To perform this, we have used the following line of Python code.

```
result = Apriori_rule[Apriori_rule['confidence']==1]
```

```
result
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	(ethiopianorthodox)	(newadvent)	0.113847	0.220479	0.113847	1.0	4.535584	0.088746	inf
3	(ethiopianorthodox)	(youtube)	0.113847	0.593746	0.113847	1.0	1.684221	0.046251	inf
4	(myflixer)	(facebook)	0.165502	0.736399	0.165502	1.0	1.357959	0.043627	inf
8	(facebook, ethiopianorthodox)	(newadvent)	0.101707	0.220479	0.101707	1.0	4.535584	0.079282	inf
9	(facebook, ethiopianorthodox)	(youtube)	0.101707	0.593746	0.101707	1.0	1.684221	0.041319	inf
10	(newadvent, ethiopianorthodox)	(youtube)	0.113847	0.593746	0.113847	1.0	1.684221	0.046251	inf
11	(youtube, ethiopianorthodox)	(newadvent)	0.113847	0.220479	0.113847	1.0	4.535584	0.088746	inf
12	(ethiopianorthodox)	(newadvent, youtube)	0.113847	0.167907	0.113847	1.0	5.955662	0.094731	inf
14	(myflixer, youtube)	(facebook)	0.143626	0.736399	0.143626	1.0	1.357959	0.037860	inf
19	(facebook, youtube, ethiopianorthodox)	(newadvent)	0.101707	0.220479	0.101707	1.0	4.535584	0.079282	inf
20	(facebook, ethiopianorthodox, newadvent)	(youtube)	0.101707	0.593746	0.101707	1.0	1.684221	0.041319	inf
21	(facebook, ethiopianorthodox)	(newadvent, youtube)	0.101707	0.167907	0.101707	1.0	5.955662	0.084629	inf

Figure 5. 15 Top selected association rules from student dataset using Apriori algorithm

To interpret the result, we have considered the antecedent and the consequent columns with respective the total confidence value.

RULE 2: (ethiopianorthodox) ===> (newadvent) confidence = 1.0.

The rule shows that 100% of the user who visited the ethiopianorthodox.org website had also visited the newadvent.com website. Therefore, after the user visited the ethiopianorthodox.org website, the user accessed the newadvent.com website.

RULE 3: (ethiopianorthodox) ===> (youtube) confidence = 1.0

This rule tells us that 100% of the user who accessed the ethiopianorthodox.org website had also accessed the youtube.com website. Therefore, if the user accessed the educational website, it had also accessed the entertainment website.

RULE 4: (myflixer) ===> (facebook) confidence = 1.0

This rule shows that 100% of the user who accessed the myflixer website had also accessed the facebook website. Therefore, the user who visited the entertainment website had also visited social media websites.

RULE 8: (ethiopianorthodox, facebook) ===> (newadvent) confidence = 1.0

This rule shows that 100% of the user who accessed ethioipianorthodox.org and facebook had also accessed the newadvent.com website. Therefore, the user who accessed the educational website and social media websites also accessed the educational website.

RULE 9: (ethiopianorthodox, facebook) ===> (youtube) confidence = 1.0

According to this rule, 100% of the user who accessed ethiopianorthodox.org and facebook had also accessed the youtube website. Therefore, the user who accessed educational and social media websites also accessed entertainment websites.

RULE 10: (newadvent, ethiopianorthodox) ===> (youtube) confidence = 1.0

This rule shows that 100% of the user who accessed newadvent.com and ethiopianorthodox.org had also accessed youtube. Therefore, users who accessed educational websites also accessed entertainment websites.

RULE 11: (youtube, ethiopianorthodox) ===> (newadvent) confidence = 1.0

This rule shows that 100% of the user who accessed youtube and ethioipianorthodox.org websites had also accessed the newadvent website. Therefore, users who accessed entertainment and educational websites also accessed the educational website.

RULE 12: (ethiopianorthodox) ===> (newadvent, youtube) confidence = 1.0

This rule shows that 100% of the user who accessed the ethioipianorthodox.org website had also accessed newadvent.com and youtube websites. Therefore, the users who accessed educational websites also accessed entertainment and educational websites.

RULE 14: (myflixer, youtube) ===> (facebook) confidence = 1.0

According to this rule, 100% of the who accessed myflixer and youtube websites had also accessed the facebook website. Therefore, users who accessed entertainment websites also accessed social media websites.

RULE 19: (ethiopianorthodox, newadvent, facebook) ===> (youtube) confidence = 1.0

This rule shows that 100% of the user who accessed ethioipianorthodox.org, newadvent, and facebook had also accessed the youtube website. Therefore, a user who accessed educational and social media websites also accessed entertainment websites.

RULE 20: (ethiopianorthodox, youtube, facebook) ===> (newadvent) confidence = 1.0

This rule shows that 100% of the user who accessed ethioipianorthodox.org, youtube, and facebook websites had also accessed the youtube website. Therefore, a user who accessed educational, entertainment, and social media website also accessed the educational website.

RULE 21: (ethiopianorthodox, facebook) ===> (newadvent, youtube) confidence = 1.0

According to this rule, 100% of the user who accessed ethiopianorthodox.org and facebook websites had also accessed newadvent and youtube websites. Therefore, a user who accessed educational and social media websites also accessed educational and entertainment websites.

Experiment V: association rule discovery in students' dataset using FP-growth algorithm

The second association rule discovery in student dataset is conducted using FP-growth with a total of 17,462 student weblog datasets and the value of minimum support and confidence value is 0.1 and 0.9 respectively.

FP_Growth_frequentItems=fpgrowth (data_frame, min_support=0.1, use_colnames=True).

To generate the frequent item sets we have used the above line of code. The result has shown below the table.

Table 5. 2 Frequent item sets using FP-growth algorithm in student dataset

No	Support	Frequent Item sets
0	0.736399	(facebook)
1	0.593746	(youtube)
2	0.367655	(sodere)
3	0.240522	(freepornforu)
4	0.235769	(wikipedia)
5	0.220479	(newadvent)
6	0.183255	(tutorialspoint)
7	0.165502	(myflixer)
8	0.114763	(microsoft)
9	0.113847	(ethiopianorthodox)
10	0.101936	(arifzefen)
11	0.286565	(pogo)
12	0.105143	(freeprojectz)
13	0.104055	(wheresthematch)
14	0.121406	(ethiojobs)
15	0.143168	(medscape)
16	0.530581	(youtube, facebook)
17	0.299278	(sodere, facebook)
18	0.257473	(sodere, youtube)

19	0.249399	(sodere, youtube, facebook)
20	0.199061	(freepornforu, facebook)
21	0.179418	(freepornforu, youtube)
22	0.115909	(freepornforu, pogo)
23	0.173348	(freepornforu, youtube, facebook)
24	0.180678	(wikipedia, facebook)
25	0.162639	(youtube, wikipedia)
26	0.136468	(sodere, wikipedia)
27	0.101878	(freepornforu, wikipedia)
28	0.160062	(youtube, wikipedia, facebook)
29	0.122609	(sodere, wikipedia, facebook)
30	0.121406	(sodere, youtube, wikipedia)
31	0.118829	(sodere, youtube, wikipedia, facebook)
32	0.179361	(newadvent, facebook)
33	0.167907	(newadvent, youtube)
34	0.148379	(newadvent, youtube, facebook)
35	0.147635	(facebook, tutorialspoint)
36	0.147406	(youtube, tutorialspoint)
37	0.141908	(facebook, youtube, tutorialspoint)
38	0.165502	(myflixer, facebook)
39	0.143626	(myflixer, youtube)
40	0.143626	(myflixer, youtube, facebook)
41	0.113847	(newadvent, ethiopianorthodox)
42	0.113847	(youtube, ethiopianorthodox)
43	0.101707	(facebook, ethiopianorthodox)
44	0.113847	(newadvent, youtube, ethiopianorthodox)
45	0.101707	(facebook, newadvent, ethiopianorthodox)
46	0.101707	(facebook, youtube, ethiopianorthodox)
47	0.101707	(newadvent, facebook, youtube, ethiopianorthodox)
48	0.171000	(pogo, facebook)
49	0.160291	(youtube, pogo)
50	0.140076	(pogo, youtube, facebook)
51	0.121292	(ethiojobs, facebook)
52	0.115336	(ethiojobs, youtube)
53	0.115222	(ethiojobs, youtube, facebook)
54	0.101077	(medscape, wikipedia)

The next step is generating rules from the frequent item sets using the FP-growth algorithm. To perform this we have used the following Python line of codes.

FP_Growth_Rules = association_rules (FP_Growth_frequentItems,metric='confidence',min_threshold=0.9)

The algorithm generates 24 rules based on the minimum confidence value. The detailed rules have been described in Appendix. From these rules, we have selected top rules which have a confidence value of 1.0 (100%). To select those rules, we have used the following line of codes in Python.

result2 = FP_Growth_Rules[FP_Growth_Rules['confidence']==1]. Where result2 is a variable to hold the statement, FP_Growth_Rules where the variable holds all generated rules**.** The final result is shown as the following.

```
result2 = FP_Growth_Rules[FP_Growth_Rules['confidence']==1]
```

```
result2
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
7	(myflixer)	(facebook)	0.165502	0.736399	0.165502	1.0	1.357959	0.043627	inf
8	(myflixer, youtube)	(facebook)	0.143626	0.736399	0.143626	1.0	1.357959	0.037860	inf
9	(ethiopianorthodox)	(newadvent)	0.113847	0.220479	0.113847	1.0	4.535584	0.088746	inf
10	(ethiopianorthodox)	(youtube)	0.113847	0.593746	0.113847	1.0	1.684221	0.046251	inf
11	(newadvent, ethiopianorthodox)	(youtube)	0.113847	0.593746	0.113847	1.0	1.684221	0.046251	inf
12	(youtube, ethiopianorthodox)	(newadvent)	0.113847	0.220479	0.113847	1.0	4.535584	0.088746	inf
13	(ethiopianorthodox)	(newadvent, youtube)	0.113847	0.167907	0.113847	1.0	5.955662	0.094731	inf
14	(ethiopianorthodox, facebook)	(newadvent)	0.101707	0.220479	0.101707	1.0	4.535584	0.079282	inf
15	(ethiopianorthodox, facebook)	(youtube)	0.101707	0.593746	0.101707	1.0	1.684221	0.041319	inf
16	(ethiopianorthodox, newadvent, facebook)	(youtube)	0.101707	0.593746	0.101707	1.0	1.684221	0.041319	inf
17	(ethiopianorthodox, youtube, facebook)	(newadvent)	0.101707	0.220479	0.101707	1.0	4.535584	0.079282	inf
18	(ethiopianorthodox, facebook)	(newadvent, youtube)	0.101707	0.167907	0.101707	1.0	5.955662	0.084629	inf

Figure 5. 16 Top selected association rules from student dataset using FP-growth algorithm

RULE 7: (myflixer) ===> (facebook) confidence = 1.0

This rule shows that 100% of the user who accessed the myflixer website had also accessed the facebook website. Therefore, a user who accessed an entertainment website had also accessed social media website.

RULE 8: (myflixer, youtube) ===> (facebook) confidence = 1.0

According to this rule, 100% of the user who accessed myflixer and youtube websites had also accessed the facebook website. Therefore, users who accessed entertainment websites also accessed social media websites.

RULE 9: (ethiopianorthodox) ====> (newadvent) confidence = 1.0

This rule shows that 100% of the user who accessed the ethioipianorthodox.org website had also accessed the newadevent.com website. Therefore, the user who accessed the educational website had also accessed the educational website.

RULE 10: (ethiopianorthodox) ===> (youtube) confidence = 1.0

This rule shows that 100% of the user who accessed the ethioipianorthodox.org website had also accessed the youtube website. Therefore, a user who accessed educational website also accessed entertainment website.

RULE 11: (newadvent, ethiopianorthodox) ===> (youtube) confidence = 1.0

This rule shows that 100% of the user who accessed the newadvent and ethiopianorthodox.org websites had also accessed the youtube website. Therefore, a user who accessed educational website also accessed entertainment website.

RULE 12: (youtube, ethiopianorthodox) ===> (newadvent) confidence = 1.0

According to this rule, 100% of the user who accessed youtube and ethiopianorthodox websites had also accessed the newadvent.com website. Therefore, users who accessed educational and entertainment websites also accessed educational websites.

RULE 13: (ethiopianorthodox) ===> (newadvent, youtube) confidence = 1.0

According to this rule, 100% of the user who accessed the ethiopianorthodox.org website had also accessed the newadvent.com and youtube.com websites. Therefore, users who accessed educational websites also accessed educational and entertainment websites.

RULE 14: (ethiopianorthodox, facebook) ===> (newadvent) confidence = 1.0

According to this rule, 100% of the user who accessed ethiopianorthodox.org, and facebook websites had also accessed the newadvent website. Therefore, a user who accessed educational and social media websites also accessed the educational website.

RULE 15: (ethiopianorthodox, facebook) ====> (youtube) confidence = 1.0

This rule shows that 100% of the user who accessed ethiopianorthodox.org and the facebook website had also accessed the youtube website. Therefore, a user who accessed educational and social media websites also accessed entertainment websites.

RULE 16: (ethiopianorthodox, newadvent, facebook) ===> (youtube) confidence = 1.0

This rule shows that 100% of the user who accessed ethiopianorthodox.org, newadvent.com, and facebook websites had also accessed the youtube website. Therefore, a user who accessed educational and social media websites also accessed entertainment websites.

RULE 17: (ethiopianorthodox, youtube, facebook) ====> (newadvent) confidence = 1.0

According to this rule, 100% of the user who accessed ethioipianorthodox.org, youtube, and facebook websites had also accessed the newadvent website. Therefore, a user who accessed educational, entertainment, and social media websites also accessed educational websites.

RULE 18: (ethiopianorthodox, facebook) ====> (newadvent, youtube) confidence = 1.0

This rule shows that 100% of the user who accessed ethiopianorthodox.org and facebook websites had also accessed newadvent and youtube websites. Therefore, a user who accessed educational and social media websites also accessed educational and entertainment websites.

Experiment VI: association rule discovery in staff dataset using Apriori algorithm

The first association rule discovery in the staff dataset is conducted using the Apriori algorithm with a total of 16,477 records. As mentioned earlier the experiment is conducted in a Python environment the packages and the libraries that have been used in VI and V are also used in this experiment.

```
staff_data = [
    ['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listsch
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
```

Figure 5. 17 Sample of transaction data in staff dataset

Each row represents a transaction item sets in the students' weblog dataset. Before finding the most often occurring itemsets, we must first convert our dataset into a True - False matrix, with rows representing transactions and columns representing URLs. finally, the dataset has two values ***True***- represents the transaction that contains the item and ***False*** value represents the transaction that does not contain the item.

data_encode = TransactionEncoder() – initializing the transaction encoder to perform transactional encoding for transactional item sets in students' weblog dataset, and where ***data_encode*** *is a variable name.*

transform_data = dtata_encode.fit_transform(transaction_data)– used to transform transactional encoded item sets using ***Numpy array***, and where ***transform_data*** is the variable to hold the respective ***Python*** code**.** The final result is described in the following figure.

```
transform_data
```

```
array([[False, False, False, ...,  True,  True,  True],
       [False, False, False, ...,  True,  True,  True],
       [False, False, False, ...,  True,  True,  True],
       ...,
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False]])
```

Figure 5. 18 Sample of transformed data in staffs' transactional dataset

After the dataset is binarized the next task is transformed into a data frame using the Pandas Python module. To perform this task, we have used the following line of codes.

data_frame=pd.DataFrame(transform_data,columns=dtata_encode.columns_)

```
data_frame
```

	aau	abebooks	cisco	coursera	et1lib	ethiojobs	ethiopianorthodox	facebook	getintopc	gmail	...	researchgate	sciencedirect	slideshare	sodere	t
0	False	False	False	False	False	True	False	True	True	True	...	False	False	True	True	
1	False	False	False	False	False	True	False	True	True	True	...	False	False	True	True	
2	False	False	False	False	False	True	False	True	True	True	...	False	False	True	True	
3	False	False	False	False	False	True	False	True	True	True	...	False	False	True	True	
4	False	False	False	False	False	True	False	True	True	True	...	False	False	True	True	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
16472	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	
16473	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	
16474	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	
16475	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	
16476	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	

16477 rows × 25 columns

Figure 5. 19 Binarized transactional data frame in staffs' weblog dataset

After this, the next step is generating the frequent itemsets and association rules using specified minimum support and confidence value. After performing different experiments with different

support and confidence values we have selected the minimum threshold of 0.2 as minimum support and 0.9 as the minimum confidence value. The reason to select the 0.2 support value is first, it supports the statistical analysis result. Second, based on the objective of the study we have selected the value to enhance the final rule result.

staff_frequentItems = apriori(data_frame,min_support=0.2,use_colnames=True). Used to generate frequent item sets with minimum support 0.2 in Apriori algorithm. The result described as follows.

```
staff_frequentItems
```

	support	itemsets
0	0.303817	(abebooks)
1	0.295139	(et1lib)
2	0.553681	(facebook)
3	0.622383	(gmail)
4	0.211507	(linkedin)
...	...	...
67	0.205559	(yahoo, facebook, gmail, youtube)
68	0.212417	(yahoo, facebook, twitter, youtube)
69	0.200340	(linkedin, twitter, gmail, youtube)
70	0.208290	(youtube, twitter, gmail, yahoo)
71	0.204892	(twitter, gmail, yahoo, facebook, youtube)

72 rows × 2 columns

Figure 5. 20 Frequent item sets in staff dataset using Apriori algorithm

After the frequent item sets have been generated, the next step is identifying the rules using the Apriori algorithm from the frequent item sets based on the minimum confidence value. To perform this, we have used the following **Python** line of code.

Staff_Apriori_rule=association_rules(staff_frequentItems, metric='confidence',min_threshold=0.9)

where the ***Staff_Apriori_rule*** – is the variable, ***association_rules*** is the predefined function used to generate rules, and **staff_frequentItems** is the variable that has been used to generate the frequent item sets. Finally, a total of 92 rules have been generated and the detailed rules are described in Appendix.

```
Staff_Apriori_rule
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(et1lib)	(slideshare)	0.295139	0.379135	0.269102	0.911783	2.404906	0.157205	7.037922
1	(linkedin)	(gmail)	0.211507	0.622383	0.204649	0.967575	1.554631	0.073011	11.645984
2	(linkedin)	(twitter)	0.211507	0.317594	0.211507	1.000000	3.148672	0.144334	inf
3	(linkedin)	(youtube)	0.211507	0.511258	0.205317	0.970732	1.898712	0.097182	16.698681
4	(yahoo)	(twitter)	0.286521	0.317594	0.275171	0.960390	3.023952	0.184174	17.228009
...	...	...	...	...	...	...	...	...	...
87	(youtube, gmail, yahoo)	(facebook, twitter)	0.208958	0.231413	0.204892	0.980540	4.237178	0.156536	39.496170
88	(youtube, facebook, yahoo)	(twitter, gmail)	0.213085	0.253019	0.204892	0.961549	3.800300	0.150977	19.427030
89	(gmail, yahoo)	(facebook, twitter, youtube)	0.225648	0.220004	0.204892	0.908015	4.127273	0.155248	8.479610
90	(facebook, yahoo)	(twitter, gmail, youtube)	0.223949	0.228258	0.204892	0.914905	4.008214	0.153774	9.069202
91	(youtube, yahoo)	(facebook, twitter, gmail)	0.226315	0.219215	0.204892	0.905337	4.129909	0.155280	8.248013

92 rows × 9 columns

Figure 5. 21 Association rules generated from staff dataset using Apriori algorithm

From these rules, we have selected the top rules which have a confidence value of 100% of the total rules. Therefore, the following rules have been selected. To perform this we have used the following line of Python code.

```
End_result = Staff_Apriori_rule[Staff_Apriori_rule['confidence']==1]
```

```
End_result
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	(linkedin)	(twitter)	0.211507	0.317594	0.211507	1.0	3.148672	0.144334	inf
18	(gmail, linkedin)	(twitter)	0.204649	0.317594	0.204649	1.0	3.148672	0.139654	inf
30	(linkedin, youtube)	(twitter)	0.205317	0.317594	0.205317	1.0	3.148672	0.140109	inf
65	(gmail, linkedin, youtube)	(twitter)	0.200340	0.317594	0.200340	1.0	3.148672	0.136713	inf

Figure 5. 22 Top selected rules in staff dataset using Apriori algorithm

RULE 2: (linkedin) ===> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed the linkedin website had also accessed the twitter website. Therefore, a user who accessed social media website also accessed social media website.

RULE 18: (gmail, linkedin) ===> (twitter) confidence = 1.0

This rule shows that 100% of the user who accessed gmail and linkedin websites had also accessed the twitter website. Therefore, a user who accessed Gmail and social media websites also accessed social media websites.

RULE 30: (linkedin, youtube) ===> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed linkedin and youtube websites had also accessed the twitter website. Therefore, the user who accessed social media and entertainment websites also accessed social media websites.

RULE 60: (gmail, linkedin, youtube) ====> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed gmail, linkedin, and youtube websites had also accessed the twitter website. Therefore, a user who accessed gmail, social media, and entertainment websites also accessed social media websites.

Experiment VII: association rule discovery in staff dataset using FP-growth algorithm

The fourth association rule discovery in the staff dataset is conducted using FP-growth with a total of 16,477 student weblog datasets and the value of minimum support and confidence value is 0.2 and 0.9 respectively.

```
staff_frequentItems=fpgrowth(data_frame,min_support=0.2,use_colnames=True)
```

```
staff_frequentItems
```

	support	itemsets
0	0.622383	(gmail)
1	0.553681	(facebook)
2	0.511258	(youtube)
3	0.379135	(slideshare)
4	0.323117	(sodere)
...	...	...
67	0.228682	(mereja, twitter)
68	0.227893	(sodere, mereja)
69	0.212417	(mereja, yahoo)
70	0.207259	(sodere, mereja, twitter)
71	0.203435	(mereja, twitter, yahoo)

72 rows × 2 columns

Figure 5. 23 Frequent item sets in staff dataset using FP-growth algorithm

After the frequent item set is generated based on their minimum support value using the FP-growth algorithm. The next task is association rule generation using the FP-growth algorithm.

```
staff_FPgrowth_rules=association_rules(staff_frequentItems,metric='confidence',min_threshold=0.9)
```

```
staff_FPgrowth_rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(facebook, youtube)	(gmail)	0.420526	0.622383	0.386053	0.918026	1.475018	0.124325	4.606532
1	(sodere, youtube)	(gmail)	0.221217	0.622383	0.205499	0.928944	1.492560	0.067817	5.314344
2	(sodere, facebook)	(gmail)	0.205923	0.622383	0.201068	0.976422	1.568845	0.072905	16.015693
3	(twitter, gmail)	(youtube)	0.253019	0.511258	0.228258	0.902135	1.764539	0.098899	4.994032
4	(twitter, youtube)	(gmail)	0.250713	0.622383	0.228258	0.910433	1.462819	0.072218	4.216046
...	...	...	...	...	...	...	...	...	...
87	(abebooks, slideshare)	(et1lib)	0.233052	0.295139	0.228500	0.980469	3.322061	0.159718	36.088900
88	(et1lib, abebooks)	(slideshare)	0.251866	0.379135	0.228500	0.907229	2.392894	0.133009	6.692446
89	(sodere, mereja)	(twitter)	0.227893	0.317594	0.207259	0.909454	2.863572	0.134881	7.536570
90	(mereja, twitter)	(sodere)	0.228682	0.323117	0.207259	0.906316	2.804916	0.133367	7.225198
91	(mereja, yahoo)	(twitter)	0.212417	0.317594	0.203435	0.957714	3.015528	0.135973	16.137974

92 rows × 9 columns

Figure 5. 24 Generated rules in staff dataset using FP-growth algorithm

In figure 5.24 we have observed that 92 rules were generated in the staff dataset using the FP-growth algorithm based on the minimum confidence value of 0.9. from these rules, we have selected here the rule which has a 1.0 (100%) confidence value. The result rules are described as follows.

```
End_result = staff_FPgrowth_rules[staff_FPgrowth_rules['confidence']==1]
```

```
End_result
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
67	(linkedin)	(twitter)	0.211507	0.317594	0.211507	1.0	3.148672	0.144334	inf
71	(linkedin, youtube)	(twitter)	0.205317	0.317594	0.205317	1.0	3.148672	0.140109	inf
74	(gmail, linkedin)	(twitter)	0.204649	0.317594	0.204649	1.0	3.148672	0.139654	inf
81	(gmail, linkedin, youtube)	(twitter)	0.200340	0.317594	0.200340	1.0	3.148672	0.136713	inf

Figure 5. 25 Top selected association rules in staff dataset using FP-growth algorithm

RULE 67: (linkedin) ===> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed the linkedin website had also accessed the twitter website. Therefore, a user who accessed social media website also accessed social media website.

RULE 71: (linkedin, youtube) ===> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed linkedin and youtube websites had also accessed the twitter website. Therefore, the user who accessed social media and entertainment websites also accessed social media websites.

RULE 74: (gmail, linkedin) ===> (twitter) confidence = 1.0

This rule shows that 100% of the user who accessed gmail and linkedin websites had also accessed the twitter website. Therefore, a user who accessed Gmail and social media websites also accessed social media websites.

RULE 81: (gmail, linkedin, youtube) ====> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed gmail, linkedin, and youtube websites had also accessed the twitter website. Therefore, a user who accessed gmail, social media, and entertainment websites also accessed social media websites.

Experiment VIII: association rule discovery in all weblog datasets using Apriori algorithm

The first association rule discovery in all weblog dataset is conducting using Apriori algorithm with a total of 33939 records and all 41 URLs that has been used in the statistical analysis also used in this experiment. After performing different experiments with different support and confidence values we have selected the minimum threshold of 0.1 as minimum support and 0.9 as the minimum confidence value. The reason to select the 0.1 support value is first, it supports the statistical analysis result. Second, based on the objective of the study we have selected the value to enhance the final rule result.

```
Both_staff_student = [['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsof
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
```

Figure 5. 26 Transaction item sets in all weblog dataset

Each row represents a transaction item sets in the students' weblog dataset. Before finding the most often occurring itemsets, we must first convert our dataset into a True - False matrix, with rows representing transactions and columns representing URLs. finally, the dataset has two values ***True***- represents the transaction that contains the item and ***False*** value represents the transaction that does not contain the item.

data_encode = TransactionEncoder() – initializing the transaction encoder to perform transactional encoding for transactional item sets in students' weblog dataset, and where ***data_encode*** *is a variable name.*

transform_data=data_encode.fit_transform(Both_staff_student)–used to transform transactional encoded item sets using ***Numpy array***, and where ***transform_data*** is the variable to hold the respective ***Python*** code**.** The final result is described in the following figure.

```
transform_data
```

```
array([[False, False, False, ..., False,  True, False],
       [False, False, False, ..., False,  True, False],
       [False, False, False, ..., False,  True, False],
       ...,
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False]])
```

Figure 5. 27 Sample of transformed data in all weblog transactional dataset

```
data_frame=pd.DataFrame(transform_data,columns=data_encode.columns_)
```

```
data_frame
```

	aau	abebooks	adobe	arifzefen	bio2rdf	cisco	coursera	et1lib	ethiojobs	ethiopianorthodox	...	tribalfootball	tutorialspoint	twitter	w3schools	whe
0	False	False	False	True	False	False	False	False	False	True	...	True	True	False	False	
1	False	False	False	True	False	False	False	False	False	True	...	True	True	False	False	
2	False	False	False	True	False	False	False	False	False	True	...	True	True	False	False	
3	False	False	False	True	False	False	False	False	False	True	...	True	True	False	False	
4	False	False	False	True	False	False	False	False	False	True	...	True	True	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
33934	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	
33935	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	
33936	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	
33937	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	
33938	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	

33939 rows × 41 columns

Figure 5. 28 Binirized transactional data frame in all weblog dataset

The next task is frequent item generation using a specified minimum support value using Apriori algorithm.

```
Both_Dataset_frequentItems=apriori(data_frame,min_support=0.1,use_colnames=True)
```

```
Both_Dataset_frequentItems
```

	support	itemsets
0	0.147500	(abebooks)
1	0.143286	(et1lib)
2	0.133180	(ethiojobs)
3	0.143169	(ethiopianorthodox)
4	0.647691	(facebook)
...	...	...
82	0.102449	(gmail, sodere, youtube, facebook)
83	0.101211	(gmail, yahoo, twitter, facebook)
84	0.103156	(gmail, youtube, twitter, facebook)
85	0.103126	(yahoo, youtube, twitter, facebook)
86	0.101123	(yahoo, youtube, twitter, gmail)

87 rows × 2 columns

Figure 5. 29 Frequent item sets in all weblog dataset using Apriori algorithm

As we observed figure 5.29, a total of 87 frequent item sets are generated based on the minimum support value 0.1 using Apriori algorithm. After the frequent item sets have been identified the next task is generating association rules using Apriori algorithm based on the minimum confidence value 0.9.

Apriori_InBoth_Dataset=association_rules(Both_Dataset_frequentItems,metric='confidence',min_threshold=0.9) used to generate association rules using Apriori algorithm in all weblog dataset.
In this experiment a total of 60 rules have been generated based on the minimum confidence value 0.9. the detailed rules have been described in Appendix. Form these rules we have selected the rule which has 100% confidence value since it indicated a strong rule as compared to the remain rules that has been generated using Apriori algorithm.

```
Both_result = Apriori_InBoth_Dataset[Apriori_InBoth_Dataset['confidence']==1]
```

```
Both_result
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
5	(linkedin)	(twitter)	0.102684	0.154188	0.102684	1.0	6.485572	0.086852	inf

Figure 5. 30 Top selected association rules in all dataset using Apriori algorithm

As we observed from figure 5.30 out of the total of 60 association rules there is only one rule which has a confidence value 1.0 (100%) in rule 5.

RULE 5: (linkedin) =====> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed linkedin website had also accessed tiwitter website. Therefore, a user who accessed social media website also accessed social media website.

Experiment IX: association rule discovery in all weblog dataset using FP-growth algorithm

The second association rule discovery in all weblog dataset is conducting using FP-growth algorithm with a total of 33939 records and all 41 URLs that has been used in the statistical analysis also used in this experiment. The value of minimum support and confidence value is same as the previous experiment conducted using Apriori algorithm in all weblog dataset.

```
Both_Dataset_frequentItems=fpgrowth(data_frame,min_support=0.1,use_colnames=True)
```

```
Both_Dataset_frequentItems
```

	support	itemsets
0	0.647691	(facebook)
1	0.553699	(youtube)
2	0.346033	(sodere)
3	0.271281	(tutorialspoint)
4	0.149091	(microsoft)
...	...	...
82	0.110934	(et1lib, slideshare, abebooks)
83	0.111023	(twitter, mereja)
84	0.110640	(sodere, mereja)
85	0.103126	(yahoo, mereja)
86	0.100622	(sodere, twitter, mereja)

87 rows × 2 columns

Figure 5. 31 Frequent item sets generated in all weblog dataset using FP-growth algorithm

```
FPgrowth_InBoth_Dataset =association_rules(Both_Dataset_frequentItems,metric='confidence',min_threshold=0.9)
```

```
FPgrowth_InBoth_Dataset
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(et1lib)	(slideshare)	0.143286	0.184066	0.130646	0.911783	4.953577	0.104272	9.249159
1	(ethiojobs)	(facebook)	0.133180	0.647691	0.125284	0.940708	1.452401	0.039024	5.941920
2	(ethiojobs)	(youtube)	0.133180	0.553699	0.128908	0.967920	1.748098	0.055166	13.912271
3	(getintopc)	(facebook)	0.120923	0.647691	0.116975	0.967349	1.493534	0.038654	10.790107
4	(getintopc)	(youtube)	0.120923	0.553699	0.111789	0.924464	1.669614	0.044834	5.908446
5	(linkedin)	(twitter)	0.102684	0.154188	0.102684	1.000000	6.485572	0.086852	inf
6	(yahoo)	(twitter)	0.139103	0.154188	0.133593	0.960390	6.228677	0.112145	21.353351
7	(et1lib, abebooks)	(slideshare)	0.122278	0.184066	0.110934	0.907229	4.928837	0.088427	8.795138
8	(slideshare, abebooks)	(et1lib)	0.113144	0.143286	0.110934	0.980469	6.842716	0.094722	43.863732
9	(ethiojobs, youtube)	(facebook)	0.128908	0.647691	0.121011	0.938743	1.449367	0.037519	5.751306

Figure 5. 32 Sample rules in all weblog dataset using FP-growth algorithm

In this experiment a total of 60 association rules have been generated using FP-growth algorithm. Some of the rules has shown in figure 5.32. from these rules there is only one rule which has confidence value 1.0 (100%). The result has shown in the following figure.

```
Both_result = FPgrowth_InBoth_Dataset[FPgrowth_InBoth_Dataset['confidence']==1.0]
```

```
Both_result
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
5	(linkedin)	(twitter)	0.102684	0.154188	0.102684	1.0	6.485572	0.086852	inf

Figure 5. 33 Top selected association rule in all weblog dataset using FP-growth algorithm

As we observed from figure 5.33 out of the total of 60 association rules there is only one rule wihich has a confidence value 1.0 (100%) in rule 5.

RULE 5: (linkedin) =====> (twitter) confidence = 1.0

According to this rule, 100% of the user who accessed linkedin website had also accessed tiwitter website. Therefore, a user who accessed social media website also accessed social media website.

5.4 Discussion and Explanation

This part provides a succinct overview of the outcomes obtained from both the statistical analysis and association rule mining discovery results.

5.4.1 Discussion of Statistical Analysis Experiment Result

In the statistical analysis experiment, two experiments were conducted. The first experiment is a statistical analysis of the student dataset using a total of 80,859 records with 25 selected URLs. the second experiment is a statistical analysis of the staff dataset using a total of 104,768 with 25 selected URLs. The statistical analysis result demonstrates that the students and staff's web usage interests are different. The following figure shows a summarized statistical experiment analysis.

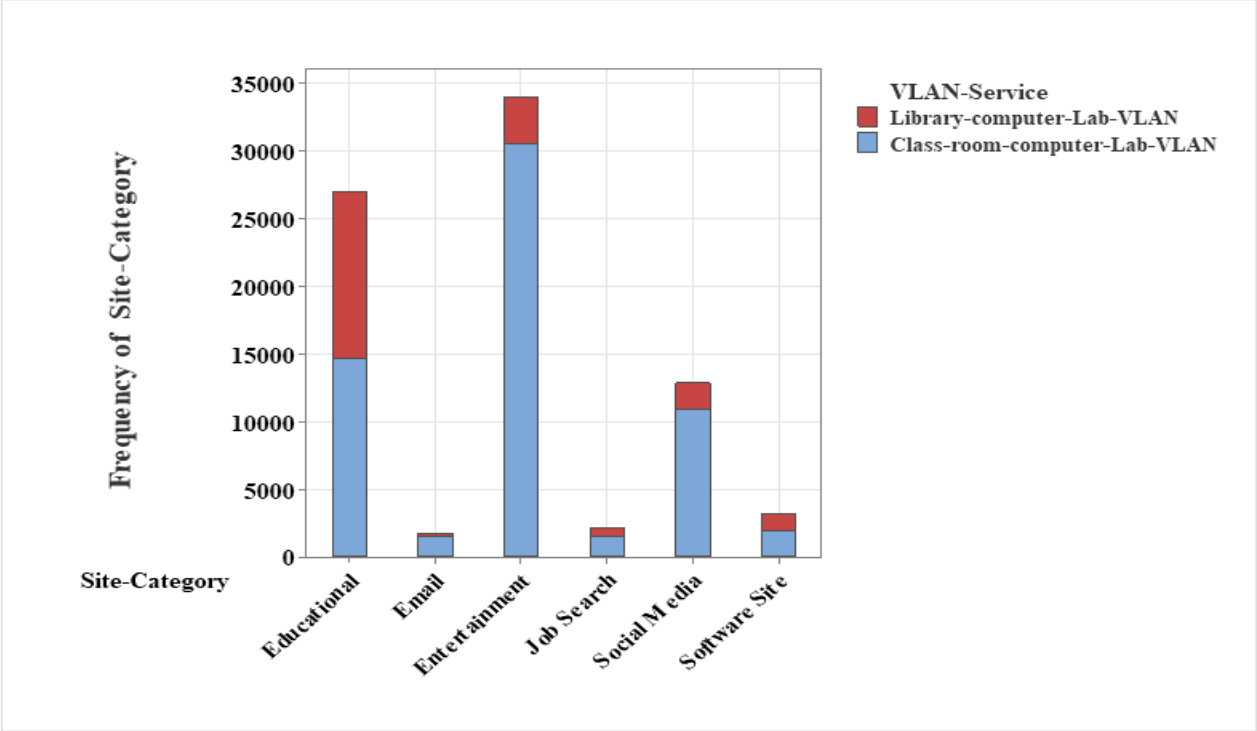


Figure 5. 34 Student web navigational behavior respect with VLAN-service

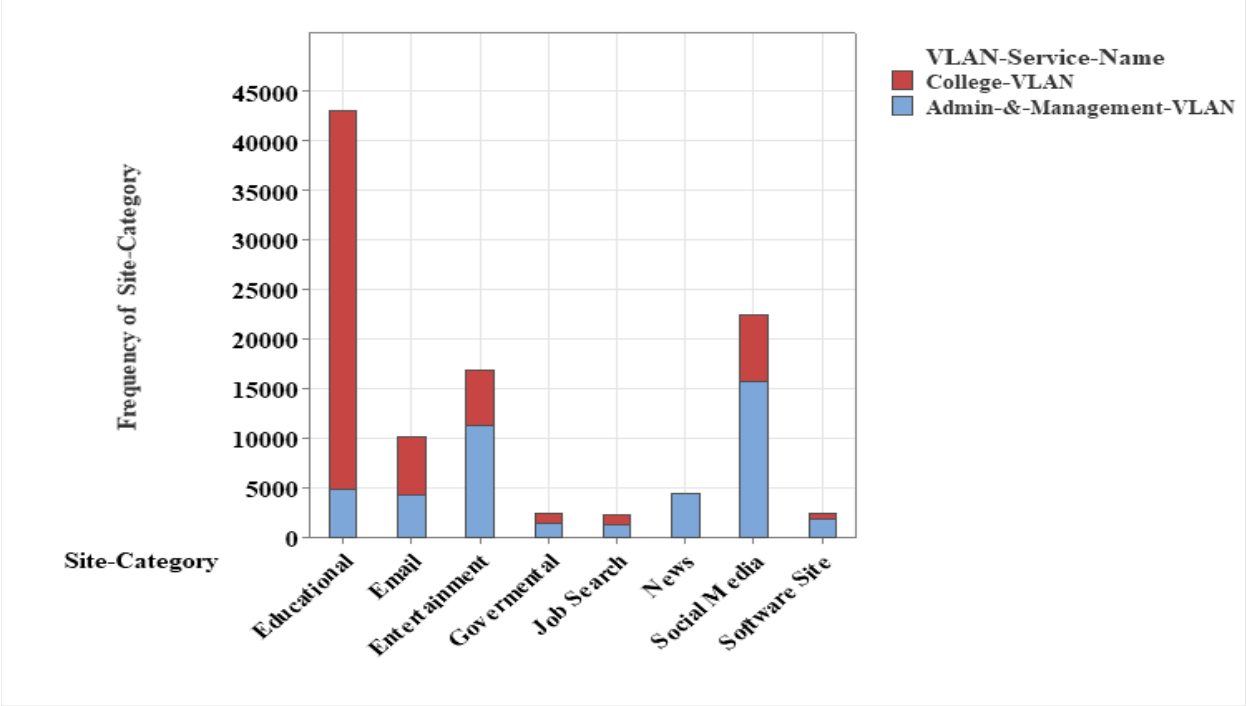


Figure 5. 35 Staff web navigational behavior respect with VLAN-service

As it has been discovered in the statistical experiment analysis, students' primary interests were accessing facebook, youtube, sodere, and freepornforu websites. Whereas wikipedia, newadvent, and tutorialspoint websites have been accessed in the second interest. However, in terms of website category as shown in figure 5.34 most of the time students' interest in accessing entertainment websites is the top priority followed by educational websites and social media websites as second and third interests respectively. Whereas the staff statistical analysis showed that most of the time the staff's web interest is accessing gmail, facebook, youtube, and slideshare.net were the top priority websites that have been accessed by the staff users. whereas websites such as sodere.com, tutorialspoint.com, twitter.com, abebooks.com, etlib.com, yahoo.com, and mereja.com were the second most frequently accessed websites by the staff users'. however, in terms of website category as shown in figure 5.35 educational websites have been accessed as the top priority. Whereas social media and entertainment websites have been accessed in the second and third frequently accessed websites by the staff users. in terms of web traffic, there is high web traffic in certain student VLANs specifically VLANs (90 & 120) have more web traffic since there is a high amount of resources accessed in these VLANs or the number of query requests to access resources in these VLANs is high as compared to the remain student VLANs. In VLAN (90) most of the time the users are accessing entertainment websites due to that this VLAN is mainly used for student's class room computer laboratory. That is why there is high web traffic since entertainment websites needs high network bandwidth because of entertainment website mostly video and audio files. Furthermore, the number of resources that has been accessed in this VLAN is high as compared to the remaining VLANs this constitutes high web traffic in terms of number or web resources in a given VLAN users Whereas in VLAN (120) most of the time the users are accessing educational websites due to that this VLAN is mainly used for library computer laboratory. That is why mostly educational websites are accessing because social media and entertainment websites have been restricted in library computer labs. Additionally, the number of resources that has been accessed in this VLAN is high as compared to the remaining VLANs this constitutes high web traffic in terms of number or web resources in a given VLAN users. on the other hand, in staff VLAN (2) have high web traffic as compared to the remaining VLANs. Due to that most of the time staff users requesting a high number of educational web resources and this VLAN mainly designed for college VLANs. Since most of the users in this VLAN is academician, they are primarily accessing educational websites as the first

priority. As the number of accessed web resource is increase it results a high number of web traffic, that is why in this VLAN there is high web traffic in terms of web resources usage. On the contrary in some of the VLANs there is low web traffic in terms of number of accessed web resources. For example, in student data VLAN (78) constitute less web traffic, this is because the number of queries accessed in these VLANs is relatively small as compared to the other VLANs and in staff data VLAN (50) has low web traffic as compared to the remaining VLANs and the cause is staff users are not mostly requesting resources in this VLAN.

5.4.2 Discussion of Association Rule Mining Experiment Result

In this section, a total of six experiments were conducted using Apriori algorithm and FP-growth algorithm in student, staff, and all weblog dataset. The first and second experiment is in student weblog dataset with a total of 17,462 record with 25 selected URLs using Apriori and FP-growth algorithm respectively. The third and fourth experiment is in staff weblog dataset with a total of 16,477 records with 25 selected URLs using Apriori and FP-growth algorithm respectively. The last two experiment five and six is conducted in both student and staff datasets with a total of 33939 with all 41 URLs using Apriori and FP-growth algorithm.
In student weblog dataset the following fundamental correlations have been identified in both algorithm:

- Educational and entertainment websites have been accessed together with 100% of confidence value
- Entertainment and social media websites have been accessed together with 100% of confidence value
- Educational, social media and entertainment websites have been accessed together with 100% of confidence value.

In the staff weblog dataset, the following correlations have been observed in both algorithm:

- Email and social media websites have been accessed together with 100% of confidence value
- Social media and entertainment websites have been accessed together with 100% of confidence value

- ❖ Email, social media, and entertainment websites have been accessed together with 100% of confidence value
- ❖ Educational, entertainment and social media websites have been accessed together with 100% of confidence value

In both student and staff weblog datasets (all weblog datasets) the following correlations have been identified:

- ❖ Social media and entertainment websites have been accessed together with 100% of confidence value.

The research questions that have been raised at the beginning of the research were the following:

RQ 1: What are the top frequently accessed websites by both student and staff VLAN users?

As it has been discussed in the statistical analysis, the dataset is categorized into two fundamental groups student and staff weblog datasets. In each dataset, unique URLs have been accessed in the student dataset as well as the staff dataset. At the same time, the are some URLs that have been accessed in both student and staff weblog datasets. In the student weblog dataset, the following 16 unique URLs have been accessed. www.pogo.com, www.freepornforu.com, www.wikipedia.org, www.newadvent.org, www.myflixer.to, www.medscape.com, www.freeprojectz.com, www.wheresthematch.com, www.arifzefen.com, www.adobe.com, www.talkenglish.com, www.w3schools.com, sourceforge.net, www.gutenberg.org, www.tribalfootball.com, and bio2rdf.org. whereas in the staff weblog dataset 16 unique URLs haven't been accessed in the students' dataset. These are www.slideshare.net, www.twitter.com, www.abebooks.com, et1lib.org, www.yahoo.com, mereja.com, www.linkedin.com, www.researchgate.net, yts.mx, www.cisco.com, www.wku.edu.et, www.sciencedirect.com, www.coursera.org, www.listscholarship.com, etd.aau.edu.et, and www.good-amharic-books.com. Furthermore, nine URLs have been accessed in both the student and staff datasets. These are the following. www.gmail.com, www.facebook.com, www.youtube.com, www.tutorialspoint.com, www.sodere.com, www.ethiopianorthodox.org, www.ethiojobs.net, www.microsoft.com, and www.getintopc.com.

RQ 2: What interesting rules and patterns discovered that could be an input for the Internet usage policy for Wolkite University ICT center?

As it has been discovered in association rule discovery a total of six different experiments have been conducted using Apriori and FP-growth algorithms. Even though, the experiment is conducted in both algorithms the rule that has been generated in each algorithm is the same but based on the time competency FP-growth has better performance to generate rules. accordingly based on the objective of the study, domain expert evaluation, and confidence value some of the rules were selected. In the students' dataset, the following rules have been identified using the FP-growth algorithm. These rules are (Rule7, Rule 8, Rule 9, Rule 10, Rule 11, Rule 12, Rule 13, Rule 14, Rule 15, Rule 16, Rule 17, and Rule 18). Whereas the staff dataset (Rule 67, Rule 71, Rule 74, and Rule 81) has been selected. In addition, all weblog dataset (Rule 5) has been selected.

RQ 3: Which tools and algorithms are suitable for web log data preprocessing and web usage pattern discovery?

Data preprocessing is conducted using Python 3.10. especially Regx Python modules have been used. In addition to Python MS-excel, and Text Splitter tools are also used for some data preprocessing tasks. To discover usage pattern through statistical analysis Minitab tool have been used. Moreover, pattern discovery is discovered using Apriori and FP-growth algorithms.

RQ 4: Which VLAN services have more web traffic in terms of web resource usage?

As it has been discovered in the statistical analysis report in the student and staff weblog dataset. To answer this question, we have considered the number of resources accessed in each VLAN as has been discussed in the statistical pattern analysis. Therefore, in the students' weblog dataset some of the VLANs have more web traffic. Accordingly, VLANs (90 &120) are the first two VLANs that constitute high web traffic as compared to the remaining VLANs. Additionally, VLANs (61 & 70) are the next VLANs that have more web traffic as compared to the remaining VLANs in the student weblog dataset. Whereas in the staff weblog dataset VLAN (2) is the first VLAN that has high web traffic since there is a high number of web resources that have been accessed in this VLAN. Additionally, VLANs (1,6,12, & 60) are the second VLANs that have high web traffic next to VLAN (2).

RQ 5: How to represent web users' navigational behavior of Wolkite University internet users on proxy server data?

To address the last question, in this study the navigation behavior of users' weblog data has been expressed in terms of the network VLANs. According to the existing network infrastructure from the university, web users have been categorized under two specified VLANs such as student VLANs and staff VLANs. As mentioned in the data preparation section, the top 41 frequent websites have been selected based on the amount of frequency as compared to the other websites. Form the top selected websites some of them have existed in the student VLAN dataset and some of them are accessed in the staff VLAN dataset. In general, among those selected websites some of the selected URLs were uniquely accessed in student VLAN and some of them were uniquely accessed in the staff VLAN and some of them have been accessed in both staff and student datasets. For ease of interpretation, the top 41 sites are categorized under different website categories such as educational, entertainment, social media, news, software site, governmental, and job search websites. In college VLANs, educational websites have been accessed more frequently since the majority of users are academic staff, whereas in admin and management VLANs social media and entertainment websites have been accessed more frequently since most of the users are administrative and management staff. So, academic staffs' behavior or interests are primarily focused on accessing educational websites whereas administrative and management users' behavior or interests are primarily focused on accessing entertainment and social media websites. Similarly, most of the time Facebook, and YouTube are accessed by the student. In addition, in terms of website category, entertainment websites, educational websites, and social media websites have been accessed as the first, second, and third primary interests by the student VLAN users.

Amare[4], discovered association rule mining and statistical analysis of Adama Science and Technology web users. he focused mainly on the list of frequently accessed websites. Because the researcher took limited attributes. However, this study looked at not just commonly visited websites but also the type of users categorized under VLANs and their web traffic has been studied. He also utilized a one-month weblog dataset this may not be as representative because the academic calendar may vary from month to month and this may modify web users' behavior. Therefore, this research is better in terms of dataset quantity since the data in this study were a three-month weblog dataset. In addition, Senait[66] studied web usage pattern discovery and

analysis for Ethio-telecom websites, Gashaw[67] studied the behaviors of Ethiopian commodity exchange website, Mekonen[79] studied the behaviors of Addis Ababa's official website, and Yohannes[15] studied the behaviors on Commercial Bank of Ethiopia. However, all these researchers except Amare focused on only a single website usage, However, it is known that web users use not only a single website but also other web resources. Therefore, this study is mostly different from these researchers since this study encompasses different web users' behavior rather than a single official website.

CHAPTER SIX
CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Web mining is one of the applications of data mining technology. It is categorized into three web usage mining, web content mining, and web structure mining. Web usage mining is used to extract data from online logs and study user web access patterns. The best way to predict user behavior, tailor information to lessen traffic, and develop web services that are appropriate for various user groups' needs is what web service providers are trying to do. Similar to this, web users require well-organized and efficient web services. Additionally, policymakers and system administrators want to understand how web users behave to create a policy for managing web usage and bandwidth.

In this study, the researcher has tried to discover web user navigational behavior using statistical analysis and association rule discovery. The research approach for this study is experimental research and the Sharma web usage mining process model have used in this study. the model has four basic web usage mining processes such as data collection, data preprocessing, pattern discovery, and pattern analysis.

To conduct this research, the researcher took three-month proxy server data starting from February 1/2021 to April 30/2021. According to the existing nature of network infrastructure, web users are categorized under different VLANs. Those VLANs are student VLANs and staff VLANs. The log files were extensively preprocessed to remove any extraneous data that could cause the mining effort to fail. Because of the log file size, preprocessing these log files was a difficult and time-consuming procedure. To describe the web user behaviors statistical analysis and association rule mining have been used in this research. Today, most of the time individuals spend a significant amount of time on the internet every day. As a result, examining internet web users' behavior is beneficial to investigate the internet usage interest of the users and the output for studying web users' interest in an imperative for network administrators to enhance the internet usage policy and internet service quality for the users.

To perform data preprocessing different tools have been reviewed in the literature part. Web log data were preprocessed, analyzed, and discovered interesting patterns and rules through statistical

analysis and association rule mining using Apriori and FP-growth algorithms. Python programming language has been used for data preparation and pattern discovery. In addition, MS-Excel 2021 has been used to split the IP address to identify the specific VLANs from each IP address and to select the most frequent URLs pivot table in MS-Excel has been used. Text File Splitter tool is also used to divide the log file into manageable sizes since the size of the web log data is huge. Moreover, Minitab 21.1 and Python 3.10 has been used for statistical analysis and association rule discovery respectively. The suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG, Cascading Style Sheet files (CSS), and scripts, have been regarded as extraneous log entries and removed from the log file. To experiment the log file has been transformed into a text file, then converted into an Excel file(.xlsx), and finally, the data is transformed into Excel(.csv) to apply statistical analysis and association rule mining.

In this study, a total of 9 experiments have been conducted, the first three experiments were conducted using statistical analysis in the student, staff, and all weblog datasets. Whereas the last six experiments were conducted in association rule mining using the Apriori and FP-growth algorithms. As the statistical analysis shows, in most cases, the users' primary access is Facebook, YouTube, and Sodere websites accessed at the top level in the student dataset. In terms of website category entertainment, educational, and social media websites are accessed in the student VLANs. Whereas Gmail, Facebook, and YouTube are the top-level websites accessed by the student. in terms of website category, educational, social media, and entertainment websites are the consecutive interest of the staff VLANs. Similarly, some of the VLANs have more web traffic in terms of web resource usage. For instance, VLANs (120 & 90) have more web traffic as compared to the remaining student VLANs since the number of resources accessed in this VLAN is high. Whereas in the staff dataset VLAN (2) has high web traffic as compared to the remaining VLANs. From the association rule experiment, interesting rules have been identified and the relationship discovered in both users are as follows. (educational and entertainment), (entertainment and social media), and (educational, social media, and entertainment) websites had a great association in the student dataset. Whereas, (Email and social media), (social media and entertainment), (email, social media, and entertainment), and (educational, entertainment, and social media) websites had more associations with the staff dataset.

6.2 Recommendation

Based on the findings of the result the following recommendations are forwarded to future researchers and important points to the system and network administrator.

6.2.1 Future Work

- ❖ Future researchers can consider sequential pattern discovery to show the time sequence of access patterns for web users.
- ❖ In this research, only web navigational behavior is considered, future researchers can work combination of navigational behavior with content web mining to discover detail web access interest for the users.

The following two critical considerations for the WKU ICT network and system administrators are advised based on the research findings and discussion:

- ♦ According to the findings of this study, the majority of WKU web users, particularly students have a priority accessing Entertainment and Social media websites. As a result, the ICT center is advised to design a social media usage policy.
- ♦ Based on the findings from the statistical analysis in some of the VLAN's social media, entertainment has been accessed, on the contrary in certain VLANs educational websites have been accessed. Therefore, it is better to provide efficient network bandwidth based on the resources accessed in each VLAN

References

[1] P. N. Srivastava, Jaideep; Cooley, R; Deshpande, M; Tan, Web Usage Mining: Discovery and Applications of Usage Patterns fromWeb Data, International Journal of Computer Science ,vol. 1, no. 2. 2000.

[2] J. Vellingiri and S. C. Pandian, "A Survey on Web Usage Mining," Global Journal of Computer Science and Technology, vol. 11, no. 4, pp. 67–72, 2011.

[3] P. V Pande, N. M. Tarbani, and P. V Ingalkar, "A Study of Web Traffic Analysis," International Journal of Computer Science and Mobile Computing, vol. 3, no. 3, pp. 900–907, 2014.

[4] A. Mulatie, "Web data analysis to discover web user navigational behavior: the case of Adama science and technology university", MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[5] B. Bhavani, V. Sucharita, and K. V. V Satyanarana, "Review on Techniques and Applications Involved in Web Usage Mining," International Journal of Applied Engineering Research, vol. 12, no. 24, pp. 15994–15998, 2017.

[6] N. Singh, A. Jain, and R. S. Raw, "Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques," International Journal of Data Mining & Knowledge Management Process, vol. 3, no. 4, pp. 137–147, 2013.

[7] O. K. C. U. & P. Bhargavi, "Analysis of Web Server Log By Web Usage Mining for Extracting Users Patterns," International Journal of Computer Science Engineering and Information Technology Research, 2013.

[8] S. Vidya and K. Banumathy, "Web Mining- Concepts and Application," International Journal of Computer Science and Information Technologies, vol. 6, no. 4, pp. 3266–3268, 2015.

[9] "Historical Background of WKU." https://www.wku.edu.et/index.php/en/about-wku/about-us/historical-background-of-wku (accessed Feb. 27, 2021).

[10] S. S. Tambe, "Understanding Virtual Local Area Networks," International Journal of Engineering Trends and Technology, vol. 25, no. 4, pp. 174–176, 2015.

[11] "'Mr. Siraj, Interviewee, Director of Directorate for ICT center in Wolkite University,'" [Interview]. 20 Feb 2021.

[12] S. M. Deshmukh and K. P. Adhiya, "A Review on Finding Users Navigation Behavior Using Web Mining Algorithm,", International Journal of Scientific Research in Science, Engineering and Technology, vol. 2, no. 6, pp. 708–712, 2016.

[13] H. Adeli and M. M. AL-Rijleh, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms," Computer-Aided Civil and Infrastructure Engineering, vol. 1, no. 06, pp. 400–404, 2010

[14] V. Soundharya, R. Ram, B. Prakash, B. Sowndarya, and B. Prathiksha, "A Survey on Pattern Discovery of Web Usage Mining," International Journal of Research in Engineering, Science and Management, vol. 1, no. 8, pp. 120–123, 2018.

[15] Y. Mesfin, "Application of web usage mining for Extracting employee internet access pattern by URL category: The Case of Commercial Bank of Ethiopia", MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia, 2017.

[16] N. Goel, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool," International Journal of Computer Applications, vol. 62, no. 2, pp. 29–33, 2013.

[17] M. Sciences, A. K. Sharma, and P. C. Gupta, "Analysis of Web Server Log Files To Increase the", International Journal of Advanced Computer and Mathematical Sciences, vol. 4, no. 1, pp. 1–8, 2013.

[18] A. Sheshasaayee, "Identification of Interested Web Users Behavior by Analyzing Web Server Access Log File", Journal of Advanced Research in Dynamical & Control System vol. 9 no. 7, November 2017.

[19] M. Ramageri, "Data Mining Techniques and Applications", Journal of Computer Science and Engineering, vol. 1, no. 4, pp. 301–305.

[20] P. B. Mohata, "Web Data Mining Techniques and Implementation for Handling Big Data", International Journal of Computer Science and Mobile Computing vol. 4, no. 4, pp. 330–334, 2015.

[21] D. A. M. M. M. Danish Ahamad, Md Mobin Akhtar, "Strategy and implementation of web

mining tools," International Journal of Innovative Research in Advanced Engineering, vol. 4, no. 12, pp. 1–7, 2017

[22] T. T. Aye, Web log cleaning for mining of web usage patterns, 3rd International Conference on Computer Research and Development, vol. 2. 2011..

[23] V. P. Juan D. Velásquez and and L. C. Jain, "Advanced Techniques in Web Intelligence-2 Web User Browsing Behaviour and Preference Analysis," and L. C. J. Juan D. Velásquez, Vasile Palade, Ed. Chile: Springer-Verlag Berlin Heidelberg, 2013, pp. 75–104.

[24] R. S. Rao and J. Arora, "A Survey on Methods used in Web Usage Mining," International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 5, pp. 2627–2631, 2017.

[25] D. A. M. M. M. Danish Ahamad, Md Mobin Akhtar, "Strategy and implementation of web mining tools," International Journal of Innovative Research in Advanced Engineering, vol. 4, no. 12, pp. 1–7, 2017

[26] R. Patel, K. Panchal, and D. Rathod, "Efficient Log Mining from Web Server Using Clustering Technique", Journal of Emerging Technologies and Innovative Research vol. 2, no. 12, pp. 94–100, 2015.

[27] I. Introduction, "Recent developments in web usage mining algorithm: theory and applications," Journal of Emerging Technologies and Innovative Research, vol. 6, no. 6, pp. 509–513, 2019.

[28] Anitha Talakokkula, "A Survey on Web usage Mining: Process, Techniques and Applications," Computer Engineering and Intelligent Systems, vol. 6, no. 2, pp. 22–30, 2015.

[29] P. Rathi and N. Singh, "A Survey of Issues and Techniques of Web Usage Mining," International Research Journal of Engineering and Technology(IRJET), vol. 4, no. 7, pp. 624–628, 2017, [Online]. Available: https://irjet.net/archives/V4/i7/IRJET-V4I7110.pdf

[30] A. H. Qureshi and R. Rajasthan, "Web traffic and log data analysis," Journal of Emerging Technologies and Innovative Research, vol. 6, no. 6, pp. 725–729, 2019.

[31] S. D. Patil, "Use Of Web Log File For Web Usage Mining," International Journal of Engineering Research & Technology (IJERT), vol. 2, no. 4, pp. 2011–2015, 2013.

[32] D. N. L.K. Joshila Grace, V.Maheswari, “Analysis of weblogs and web user in web mining,” International Journal of Network Security & Its Applications (IJNSA), vol. 3, no. 1, pp. 99–110, 2011.

[33] E. Suganya and D. S. V. Prakathambal, “Web Log Files in Web Usage Mining Research – A Review,” International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), vol. 5, no. 2, pp. 29–32, 2018.

[34] M. Abd Wahab, M. Mohd, H. Hanafi, and M. Mohamad Mohsin, “Data pre-processing on web server logs for generalized association rules mining algorithm,” World Academy of Science, Engineering and Technology, 2008.

[35] D. B. Rathod, R. T. Prajapati, and H. Joshi, “Subterranean Insect based Data Reduction in Web Usage Mining using K-implies Clustering Algorithm”, International Journal of Engineering and Advanced Technology , vol. 9, no. 4, pp. 396–400, 2020.

[36] P. N. Srivastava, Jaideep; Cooley, R; Deshpande, M; Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns fromWeb Data,” SIGKDD Explorations, vol. 1, no. 2, pp. 12–23, 2000,.

[37] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” vol. 1, no. 2, pp. 12–23, 2000.

[38] M. Aldekhail, “Application and Significance of Web Usage Mining in the 21st Century: A Literature Review,” International Journal of Computer Theory and Engineering, vol. 8, no. 1, pp. 41–47, 2016.

[39] M. E. Suganya, “Web Log Analysis Using Association Rule Mining Algorithms,” Journal of Emerging Technologies and Innovative Research, vol. 6, no. 2, pp. 333–342, 2019.

[40] V. Chitraa and D. A. S. Davamani, “An Efficient Path Completion Technique for web log mining,” 2010 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–5, 2010.

[41] V. Chitraa and D. Selvadoss Thanamani, “A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing,” International Journal of Computer Applications, vol. 34, no. 9, pp. 23–26, 2011.

[42] C. L. Mugali, A. Maniyar, and A. Padma Dandannavar, “Pre-Processing and Analysis of Web Server Logs,” International Journal of Innovative Research in Advanced Engineering

(IJIRAE), vol. 2, no. 8, pp. 46–55, 2015, [Online]. Available: www.ijirae.com

[43] S. Padmaja and A. Sheshasaayee, "Web server logs to analyzing user behavior using log analyzer tool", International Journal of Advance Research In Science And Engineering, vol. 8354, no. 3, pp. 514–525, 2014.

[44] S. Bhuvaneswari, "A Comparative Study of Different Log Analyzer Tools to Analyze User Behaviors," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, no. 5, pp. 2997–3002, 2015.

[45] "web log expert information." https://www.weblogexpert.com/features.htm (accessed Jun. 02, 2021).

[46] "AWstats information." https://awstats.sourceforge.io/ (accessed Jun. 02, 2021).

[47] "W3Perl." http://www.w3perl.com/demo/docs/uk/index.html (accessed Jun. 02, 2021).

[48] "Webalizer", Accessed: Jun. 02, 2021. [Online]. Available: http://www.webalizer.org/

[49] "Data preparator information." http://www.datapreparator.com/ (accessed Jun. 02, 2021).

[50] "Deep log analyzer information." https://www.deep-software.com/ (accessed Jun. 02, 2021).

[51] "Open Web Analytics." http://www.openwebanalytics.com/about/ (accessed Jun. 02, 2021).

[52] "Glogg information." https://glogg.bonnefon.org/description.html (accessed Jun. 02, 2021).

[53] "Introduction Minitab." https://www.greycampus.com/opencampus/minitab/introduction-on-minitab (accessed Jun. 18, 2021).

[54] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, "Web usage mining as a tool for personalization: A survey," User Modelling and User-Adapted Interaction, 2003.

[55] P. Mandave, "Data mining using Association rule based on APRIORI algorithm and improved approach with illustration," International Journal of Latest Trends in Engineering and Technology (IJLTET), vol. 3, no. 2, pp. 107–113, 2013.

[56] F. M. Facca and P. L. Lanzi, "Recent Developments in Web Usage Mining Research," Lecture Notes in Computer Science, September 2003

[57] “Available Online at www.ijarcs.info Association Rule Mining among web pages for Discovering Usage Patterns in Web Log Data,” International Journal of Advanced Research in Computer Science, vol. 4, no. 5, pp. 65–69, 2013.

[58] G. Yongmei and B. Fuguang, “The Research on Measure Method of Association Rules Mining,” International Journal of Database Theory and Application, vol. 8, no. 2, pp. 245–258, 2015.

[59] M. S. Mythili, “Performance Evaluation of Apriori and FP-Growth Algorithms,” International Journal of Computer Applications, vol. 79, no. 10, pp. 34–37, 2013.

[60] P. Prithiviraj and R. Porkodi, “A Comparative Analysis of Association Rule Mining Algorithms in Data Mining : A Study,” American Journal of Computer Science and Engineering Survey, vol. 3, no. 1, 2015.

[61] M. Ganjir and J. Chopra, “Combining Apriori and FP Growth algorithms with Simulated Annealing for Optimized Association Rule Mining,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 12, pp. 495–497, 2015,.

[62] S. Rohit, “Association Rule Mining Algorithms : Survey,” International Research Journal of Engineering and Technology, vol. 3, no. 10, pp. 500–505, 2016.

[63] D. U. Maheswari and A. Marimuthu, “A Study of Web Usage Mining Applications and its Future Trends,” International Journal of Engineering Research & Technology (IJERT), vol. 2, no. 9, pp. 1793–1797, 2013.

[64] S. Ihm and C. C. N. Miscellaneous, “Towards Understanding Modern Web Traffic,” pp. 295–312, 2011.

[65] T. Asitatikie, “Web usage pattern discovery: The case of Addis Ababa University official website”, MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[66] S. Mezgebu, “Web usage pattern discovery and analysis for website optimization: The case of Ethio Telecom official website,” MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[67] G. Bekele, “Exploring users navigational behavior using web usage mining: The case of

Ethiopia Commodity Exchange official website," MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia,2015.

[68] A. Legesse, "Discovering frequent navigational patterns for constructing user profile: The case of Ebiz online solutions Plc official website," MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[69] A. Fesha, "Web usage: exploring navigational behavior of users in the case of the official website of Addis Ababa University," MSc Thesis, School of Information Science Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[70] A. Sharma, "Data Mining of Web Access Logs," pp. 2–10.

[71] S. Parvatikar, "Analysis of User Behavior through Web Usage Mining," International Journal of Computer Applications, pp. 27–31, 2014.

[72] A. Kumar, V. Ahirwar, and R. K. 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, vol. 6, no. 2. 2017.

[73] A. Bala, "Performance Analysis of Apriori and FP-Growth Algorithms (Association Rule Mining)", International Journal of Computer Technology and Applications vol. 7, no. April, pp. 279–293, 2016.

[74] A. H. Nasyuha et al., "Frequent pattern growth algorithm for maximizing display items," Telkomnika (Telecommunication Computing Electronics and Control), vol. 19, no. 2, pp. 390–396, 2021.

[75] R. Garg and P. Gulia, "Comparative Study of Frequent Itemset Mining Algorithms Apriori and FP Growth," International Journal of Computer Applications, vol. 126, no. 4, pp. 8–12, 2015.

[76] C. P. Sumathi, R. Padmaja Valli, and T. Santhanam, "An overview of preprocessing of Web log files for Web usage mining," Journal of Theoretical and Applied Information Technology, vol. 34, no. 2, 2011.

[77] S. Langhnoja, M. Barot, and D. Mehta, "Pre-Processing : Procedure on Web Log File for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 12, pp. 419–422, 2012.

[78] “Website URL Category Check - Cyren.” https://www.cyren.com/security-center/url-category-check-gate (accessed Jun. 18, 2021).

[79] Mekonen Tsegaye, “Web usage pattern discovery using data mining and statistical analysis: The case of Addis Ababa official website,”MSc Thesis, Faculity of Informatics, Addis Ababa University, Addis Ababa, Ethiopia, 2009.

APPENDICES

```
import pandas as pd
```

```
import numpy as np
```

```
from mlxtend.frequent_patterns import apriori,fpgrowth,association_rules
```

```
from mlxtend.preprocessing import TransactionEncoder
```

Appendix A-1: Python librariy packages used for association rule discovery

```
transaction_data=[['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
['facebook','youtube','sodere','freepornforu','wikipedia','newadvent','tutorialspoint','myflixer','microsoft','ethiopianorthodox'
```

Appendix A-2: Sample for student transactional weblog dataset

```
Apriori_rule
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ethiojobs)	(facebook)	0.121406	0.736399	0.121292	0.999057	1.356678	0.031888	279.417020
1	(ethiojobs)	(youtube)	0.121406	0.593746	0.115336	0.950000	1.600010	0.043251	8.125072
2	(ethiopianorthodox)	(newadvent)	0.113847	0.220479	0.113847	1.000000	4.535584	0.088746	inf
3	(ethiopianorthodox)	(youtube)	0.113847	0.593746	0.113847	1.000000	1.684221	0.046251	inf
4	(myflixer)	(facebook)	0.165502	0.736399	0.165502	1.000000	1.357959	0.043627	inf
5	(ethiojobs, youtube)	(facebook)	0.115336	0.736399	0.115222	0.999007	1.356611	0.030288	265.446169
6	(ethiojobs, facebook)	(youtube)	0.121292	0.593746	0.115222	0.949953	1.599930	0.043205	8.117406
7	(ethiojobs)	(youtube, facebook)	0.121406	0.530581	0.115222	0.949057	1.788713	0.050806	9.214527
8	(ethiopianorthodox, facebook)	(newadvent)	0.101707	0.220479	0.101707	1.000000	4.535584	0.079282	inf
9	(ethiopianorthodox, facebook)	(youtube)	0.101707	0.593746	0.101707	1.000000	1.684221	0.041319	inf
10	(ethiopianorthodox, newadvent)	(youtube)	0.113847	0.593746	0.113847	1.000000	1.684221	0.046251	inf
11	(ethiopianorthodox, youtube)	(newadvent)	0.113847	0.220479	0.113847	1.000000	4.535584	0.088746	inf
12	(ethiopianorthodox)	(newadvent, youtube)	0.113847	0.167907	0.113847	1.000000	5.955662	0.094731	inf

13	(freepornforu, youtube)	(facebook)	0.179418	0.736399	0.173348	0.966167	1.312015	0.041224	7.791149
14	(myflixer, youtube)	(facebook)	0.143626	0.736399	0.143626	1.000000	1.357959	0.037860	inf
15	(sodere, youtube)	(facebook)	0.257473	0.736399	0.249399	0.968639	1.315372	0.059796	8.405319
16	(tutorialspoint, youtube)	(facebook)	0.147406	0.736399	0.141908	0.962704	1.307313	0.033359	7.067801
17	(tutorialspoint, facebook)	(youtube)	0.147635	0.593746	0.141908	0.961210	1.618890	0.054250	10.473217
18	(youtube, wikipedia)	(facebook)	0.162639	0.736399	0.160062	0.984155	1.336442	0.040295	16.636150
19	(ethiopianorthodox, newadvent, facebook)	(youtube)	0.101707	0.593746	0.101707	1.000000	1.684221	0.041319	inf
20	(ethiopianorthodox, youtube, facebook)	(newadvent)	0.101707	0.220479	0.101707	1.000000	4.535584	0.079282	inf
21	(ethiopianorthodox, facebook)	(newadvent, youtube)	0.101707	0.167907	0.101707	1.000000	5.955662	0.084629	inf
22	(sodere, youtube, wikipedia)	(facebook)	0.121406	0.736399	0.118829	0.978774	1.329135	0.029426	12.418534
23	(sodere, facebook, wikipedia)	(youtube)	0.122609	0.593746	0.118829	0.969173	1.632302	0.046031	13.178620

```
FP_Growth_Rules = association_rules(FP_Growth_frequentItems,metric='confidence',min_threshold=0.9)
```

```
FP_Growth_Rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(sodere, youtube)	(facebook)	0.257473	0.736399	0.249399	0.968639	1.315372	0.059796	8.405319
1	(freepornforu, youtube)	(facebook)	0.179418	0.736399	0.173348	0.966167	1.312015	0.041224	7.791149
2	(youtube, wikipedia)	(facebook)	0.162639	0.736399	0.160062	0.984155	1.336442	0.040295	16.636150
3	(sodere, youtube, wikipedia)	(facebook)	0.121406	0.736399	0.118829	0.978774	1.329135	0.029426	12.418534
4	(sodere, facebook, wikipedia)	(youtube)	0.122609	0.593746	0.118829	0.969173	1.632302	0.046031	13.178620
5	(tutorialspoint, youtube)	(facebook)	0.147406	0.736399	0.141908	0.962704	1.307313	0.033359	7.067801
6	(tutorialspoint, facebook)	(youtube)	0.147635	0.593746	0.141908	0.961210	1.618890	0.054250	10.473217
7	(myflixer)	(facebook)	0.165502	0.736399	0.165502	1.000000	1.357959	0.043627	inf
8	(myflixer, youtube)	(facebook)	0.143626	0.736399	0.143626	1.000000	1.357959	0.037860	inf
9	(ethiopianorthodox)	(newadvent)	0.113847	0.220479	0.113847	1.000000	4.535584	0.088746	inf
10	(ethiopianorthodox)	(youtube)	0.113847	0.593746	0.113847	1.000000	1.684221	0.046251	inf
11	(ethiopianorthodox, newadvent)	(youtube)	0.113847	0.593746	0.113847	1.000000	1.684221	0.046251	inf
12	(ethiopianorthodox, youtube)	(newadvent)	0.113847	0.220479	0.113847	1.000000	4.535584	0.088746	inf
13	(ethiopianorthodox)	(newadvent, youtube)	0.113847	0.167907	0.113847	1.000000	5.955662	0.094731	inf
14	(ethiopianorthodox, facebook)	(newadvent)	0.101707	0.220479	0.101707	1.000000	4.535584	0.079282	inf
15	(ethiopianorthodox, facebook)	(youtube)	0.101707	0.593746	0.101707	1.000000	1.684221	0.041319	inf
16	(ethiopianorthodox, newadvent, facebook)	(youtube)	0.101707	0.593746	0.101707	1.000000	1.684221	0.041319	inf
17	(ethiopianorthodox, youtube, facebook)	(newadvent)	0.101707	0.220479	0.101707	1.000000	4.535584	0.079282	inf
18	(ethiopianorthodox, facebook)	(newadvent, youtube)	0.101707	0.167907	0.101707	1.000000	5.955662	0.084629	inf
19	(ethiojobs)	(facebook)	0.121406	0.736399	0.121292	0.999057	1.356678	0.031888	279.417020
20	(ethiojobs)	(youtube)	0.121406	0.593746	0.115336	0.950000	1.600010	0.043251	8.125072
21	(ethiojobs, youtube)	(facebook)	0.115336	0.736399	0.115222	0.999007	1.356611	0.030288	265.446169
22	(ethiojobs, facebook)	(youtube)	0.121292	0.593746	0.115222	0.949953	1.599930	0.043205	8.117406
23	(ethiojobs)	(youtube, facebook)	0.121406	0.530581	0.115222	0.949057	1.788713	0.050806	9.214527

```
staff_data = [
    ['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listsch
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
['gmail','facebook','youtube','slideshare','sodere','twitter','yahoo','linkedin','yts','getintopc','wku','ethiojobs','listscholar
```

```
Staff_Apriori_rule[0:21]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(et1lib)	(slideshare)	0.295139	0.379135	0.269102	0.911783	2.404906	0.157205	7.037922
1	(linkedin)	(gmail)	0.211507	0.622383	0.204649	0.967575	1.554631	0.073011	11.645984
2	(linkedin)	(twitter)	0.211507	0.317594	0.211507	1.000000	3.148672	0.144334	inf
3	(linkedin)	(youtube)	0.211507	0.511258	0.205317	0.970732	1.898712	0.097182	16.698681
4	(yahoo)	(twitter)	0.286521	0.317594	0.275171	0.960390	3.023952	0.184174	17.228009
5	(abebooks, et1lib)	(slideshare)	0.251866	0.379135	0.228500	0.907229	2.392894	0.133009	6.692446
6	(abebooks, slideshare)	(et1lib)	0.233052	0.295139	0.228500	0.980469	3.322061	0.159718	36.088900
7	(sodere, facebook)	(gmail)	0.205923	0.622383	0.201068	0.976422	1.568845	0.072905	16.015693
8	(facebook, twitter)	(gmail)	0.231413	0.622383	0.219215	0.947286	1.522031	0.075187	7.163456
9	(yahoo, facebook)	(gmail)	0.223949	0.622383	0.211750	0.945528	1.519207	0.072368	6.932377
10	(yahoo, gmail)	(facebook)	0.225648	0.553681	0.211750	0.938408	1.694853	0.086813	7.246351
11	(youtube, facebook)	(gmail)	0.420526	0.622383	0.386053	0.918026	1.475016	0.124325	4.606532
12	(yahoo, facebook)	(twitter)	0.223949	0.317594	0.220671	0.985366	3.102594	0.149546	46.631061
13	(facebook, twitter)	(yahoo)	0.231413	0.286521	0.220671	0.953580	3.328137	0.154367	15.370039
14	(facebook, twitter)	(youtube)	0.231413	0.511258	0.220004	0.950695	1.859521	0.101692	9.912621
15	(yahoo, youtube)	(facebook)	0.226315	0.553681	0.213085	0.941539	1.700509	0.087778	7.634514
16	(yahoo, facebook)	(youtube)	0.223949	0.511258	0.213085	0.951491	1.861077	0.098589	10.075182
17	(gmail, linkedin)	(twitter)	0.204649	0.317594	0.204649	1.000000	3.148672	0.139654	inf
18	(linkedin, twitter)	(gmail)	0.211507	0.622383	0.204649	0.967575	1.554631	0.073011	11.645984
19	(linkedin)	(gmail, twitter)	0.211507	0.253019	0.204649	0.967575	3.824116	0.151134	23.037412
20	(youtube, linkedin)	(gmail)	0.205317	0.622383	0.200340	0.975761	1.567783	0.072554	15.579016

21	(gmail, linkedin)	(youtube)	0.204649	0.511258	0.200340	0.978944	1.914775	0.095711	23.211798
22	(linkedin)	(youtube, gmail)	0.211507	0.449111	0.200340	0.947202	2.109061	0.105350	10.433960
23	(sodere, youtube)	(gmail)	0.221217	0.622383	0.205499	0.928944	1.492560	0.067817	5.314344
24	(yahoo, gmail)	(twitter)	0.225648	0.317594	0.222371	0.985476	3.102941	0.150706	46.984902
25	(youtube, twitter)	(gmail)	0.250713	0.622383	0.226258	0.910433	1.462819	0.072218	4.216046
26	(gmail, twitter)	(youtube)	0.253019	0.511258	0.226258	0.902135	1.764539	0.098899	4.994032
27	(yahoo, youtube)	(gmail)	0.226315	0.622383	0.208958	0.923304	1.483499	0.068103	4.923548
28	(yahoo, gmail)	(youtube)	0.225648	0.511258	0.208958	0.926036	1.811288	0.093594	6.607790
29	(youtube, linkedin)	(twitter)	0.205317	0.317594	0.205317	1.000000	3.148672	0.140109	inf
30	(linkedin, twitter)	(youtube)	0.211507	0.511258	0.205317	0.970732	1.898712	0.097182	16.698681
31	(linkedin)	(youtube, twitter)	0.211507	0.250713	0.205317	0.970732	3.871882	0.152289	25.600635
32	(mereja, sodere)	(twitter)	0.227893	0.317594	0.207259	0.909454	2.863572	0.134881	7.536570
33	(mereja, twitter)	(sodere)	0.228682	0.323117	0.207259	0.906316	2.804916	0.133367	7.225198
34	(mereja, yahoo)	(twitter)	0.212417	0.317594	0.203435	0.957714	3.015528	0.135973	16.137974
35	(sodere, yahoo)	(twitter)	0.218244	0.317594	0.210354	0.963849	3.034843	0.141041	18.876394
36	(sodere, youtube)	(twitter)	0.221217	0.317594	0.206409	0.933059	2.937897	0.136152	10.194135
37	(yahoo, youtube)	(twitter)	0.226315	0.317594	0.225648	0.997050	3.139384	0.153771	231.335559
38	(youtube, twitter)	(yahoo)	0.250713	0.286521	0.225648	0.900024	3.141220	0.153813	7.136522
39	(gmail, yahoo, facebook)	(twitter)	0.211750	0.317594	0.208472	0.984523	3.099939	0.141222	44.090996
40	(yahoo, facebook, twitter)	(gmail)	0.220671	0.622383	0.208472	0.944719	1.517908	0.071130	6.830928
41	(gmail, facebook, twitter)	(yahoo)	0.219215	0.286521	0.208472	0.950997	3.319121	0.145663	14.559817
42	(yahoo, gmail, twitter)	(facebook)	0.222371	0.553681	0.208472	0.937500	1.693214	0.085350	7.141106
43	(yahoo, facebook)	(gmail, twitter)	0.223949	0.253019	0.208472	0.930894	3.679143	0.151809	10.809249
44	(facebook, twitter)	(yahoo, gmail)	0.231413	0.225648	0.208472	0.900865	3.992351	0.156254	7.811123

```
staff_FPgrowth_rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(youtube, facebook)	(gmail)	0.420526	0.622383	0.386053	0.918026	1.475018	0.124325	4.606532
1	(sodere, youtube)	(gmail)	0.221217	0.622383	0.205499	0.928944	1.492560	0.067817	5.314344
2	(sodere, facebook)	(gmail)	0.205923	0.622383	0.201068	0.976422	1.568845	0.072905	16.015693
3	(youtube, twitter)	(gmail)	0.250713	0.622383	0.228258	0.910433	1.462819	0.072218	4.216046
4	(gmail, twitter)	(youtube)	0.253019	0.511258	0.228258	0.902135	1.764539	0.098899	4.994032
5	(sodere, youtube)	(twitter)	0.221217	0.317594	0.206409	0.933059	2.937897	0.136152	10.194135
6	(facebook, twitter)	(youtube)	0.231413	0.511258	0.220004	0.950695	1.859521	0.101692	9.912621
7	(facebook, twitter)	(gmail)	0.231413	0.622383	0.219215	0.947286	1.522031	0.075187	7.163456
8	(youtube, facebook, twitter)	(gmail)	0.220004	0.622383	0.212478	0.965793	1.551767	0.075552	11.039215
9	(gmail, facebook, twitter)	(youtube)	0.219215	0.511258	0.212478	0.969269	1.895851	0.100403	15.903925
10	(youtube, gmail, twitter)	(facebook)	0.228258	0.553681	0.212478	0.930869	1.681238	0.086096	6.456178
11	(facebook, twitter)	(youtube, gmail)	0.231413	0.449111	0.212478	0.918175	2.044428	0.108548	6.732501
12	(yahoo)	(twitter)	0.286521	0.317594	0.275171	0.960390	3.023952	0.184174	17.228009
13	(yahoo, youtube)	(twitter)	0.226315	0.317594	0.225648	0.997050	3.139384	0.153771	231.335559
14	(youtube, twitter)	(yahoo)	0.250713	0.286521	0.225648	0.900024	3.141220	0.153813	7.136522
15	(yahoo, gmail)	(twitter)	0.225648	0.317594	0.222371	0.985476	3.102941	0.150706	46.984902
16	(yahoo, youtube)	(gmail)	0.226315	0.622383	0.208958	0.923304	1.483499	0.068103	4.923548
17	(yahoo, gmail)	(youtube)	0.225648	0.511258	0.208958	0.926036	1.811288	0.093594	6.607790
18	(yahoo, youtube, gmail)	(twitter)	0.208958	0.317594	0.208290	0.996805	3.138612	0.141927	213.593008
19	(yahoo, youtube, twitter)	(gmail)	0.225648	0.622383	0.208290	0.923077	1.483134	0.067851	4.909025
20	(yahoo, gmail, twitter)	(youtube)	0.222371	0.511258	0.208290	0.936681	1.832110	0.094602	7.718751
20	(yahoo, gmail, twitter)	(youtube)	0.222371	0.511258	0.208290	0.936681	1.832110	0.094602	7.718751
21	(youtube, gmail, twitter)	(yahoo)	0.228258	0.286521	0.208290	0.912523	3.184843	0.142890	8.156219
22	(yahoo, youtube)	(gmail, twitter)	0.226315	0.253019	0.208290	0.920354	3.637484	0.151028	9.378757
23	(yahoo, gmail)	(youtube, twitter)	0.225648	0.250713	0.208290	0.923077	3.681805	0.151717	9.740730
24	(yahoo, facebook)	(twitter)	0.223949	0.317594	0.220671	0.985366	3.102594	0.149546	46.631061
25	(facebook, twitter)	(yahoo)	0.231413	0.286521	0.220671	0.953580	3.328137	0.154367	15.370039
26	(yahoo, youtube)	(facebook)	0.226315	0.553681	0.213085	0.941539	1.700509	0.087778	7.634514
27	(yahoo, facebook)	(youtube)	0.223949	0.511258	0.213085	0.951491	1.861077	0.098589	10.075182
28	(yahoo, facebook)	(gmail)	0.223949	0.622383	0.211750	0.945528	1.519207	0.072368	6.932377
29	(yahoo, gmail)	(facebook)	0.225648	0.553681	0.211750	0.938408	1.694853	0.086813	7.246351
30	(yahoo, youtube, facebook)	(twitter)	0.213085	0.317594	0.212417	0.996867	3.138807	0.144743	217.811517
31	(yahoo, youtube, twitter)	(facebook)	0.225648	0.553681	0.212417	0.941366	1.700197	0.087480	7.611993
32	(yahoo, facebook, twitter)	(youtube)	0.220671	0.511258	0.212417	0.962596	1.882799	0.099597	13.066658
33	(youtube, facebook, twitter)	(yahoo)	0.220004	0.286521	0.212417	0.965517	3.369800	0.149382	20.690902
34	(yahoo, youtube)	(facebook, twitter)	0.226315	0.231413	0.212417	0.938589	4.055898	0.160045	12.515542
35	(yahoo, facebook)	(youtube, twitter)	0.223949	0.250713	0.212417	0.948509	3.783246	0.156270	14.551940
36	(facebook, twitter)	(yahoo, youtube)	0.231413	0.226315	0.212417	0.917912	4.055898	0.160045	9.425109
37	(gmail, yahoo, facebook)	(twitter)	0.211750	0.317594	0.208472	0.984523	3.099939	0.141222	44.090996
38	(yahoo, facebook, twitter)	(gmail)	0.220671	0.622383	0.208472	0.944719	1.517908	0.071130	6.830928
39	(gmail, facebook, twitter)	(yahoo)	0.219215	0.286521	0.208472	0.950997	3.319121	0.145663	14.559817
40	(yahoo, gmail, twitter)	(facebook)	0.222371	0.553681	0.208472	0.937500	1.693214	0.085350	7.141106
41	(yahoo, facebook)	(gmail, twitter)	0.223949	0.253019	0.208472	0.930894	3.679143	0.151809	10.809249
42	(facebook, twitter)	(yahoo, gmail)	0.231413	0.225648	0.208472	0.900865	3.992351	0.156254	7.811123
43	(yahoo, gmail)	(facebook, twitter)	0.225648	0.231413	0.208472	0.923884	3.992351	0.156254	10.097543

```
Apriori_frequentItems=apriori(data_frame,min_support=0.2,use_colnames=True)
```

```
Apriori_frequentItems
```

	support	itemsets
0	0.736399	(facebook)
1	0.240522	(freepornforu)
2	0.220479	(newadvent)
3	0.286565	(pogo)
4	0.367655	(sodere)
5	0.235769	(wikipedia)
6	0.593746	(youtube)
7	0.299278	(sodere, facebook)
8	0.530581	(youtube, facebook)
9	0.257473	(sodere, youtube)
10	0.249399	(sodere, youtube, facebook)

```
Apriori_rule = association_rules(Apriori_frequentItems,metric='confidence',min_threshold=0.8)
```

```
Apriori_rule
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(sodere)	(facebook)	0.367655	0.736399	0.299278	0.814019	1.105404	0.028537	1.417352
1	(youtube)	(facebook)	0.593746	0.736399	0.530581	0.893615	1.213493	0.093346	2.477801
2	(sodere, youtube)	(facebook)	0.257473	0.736399	0.249399	0.968639	1.315372	0.059796	8.405319
3	(sodere, facebook)	(youtube)	0.299278	0.593746	0.249399	0.833333	1.403517	0.071703	2.437521

```
FP_Growth_frequentItems=fpgrowth(data_frame,min_support=0.2,use_colnames=True)
```

```
FP_Growth_frequentItems
```

	support	itemsets
0	0.736399	(facebook)
1	0.593746	(youtube)
2	0.367655	(sodere)
3	0.240522	(freepornforu)
4	0.235769	(wikipedia)
5	0.220479	(newadvent)
6	0.286565	(pogo)
7	0.530581	(youtube, facebook)
8	0.299278	(sodere, facebook)
9	0.257473	(sodere, youtube)
10	0.249399	(sodere, youtube, facebook)

```
FP_Growth_Rules = association_rules(FP_Growth_frequentItems,metric='confidence',min_threshold=0.8)
```

```
FP_Growth_Rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(youtube)	(facebook)	0.593746	0.736399	0.530581	0.893615	1.213493	0.093346	2.477801
1	(sodere)	(facebook)	0.367655	0.736399	0.299278	0.814019	1.105404	0.028537	1.417352
2	(sodere, youtube)	(facebook)	0.257473	0.736399	0.249399	0.968639	1.315372	0.059796	8.405319
3	(sodere, facebook)	(youtube)	0.299278	0.593746	0.249399	0.833333	1.403517	0.071703	2.437521

```
staff_frequentItems=apriori(data_frame,min_support=0.3,use_colnames=True)
```

```
staff_frequentItems
```

	support	itemsets
0	0.303817	(abebooks)
1	0.553681	(facebook)
2	0.622383	(gmail)
3	0.379135	(slideshare)
4	0.323117	(sodere)
5	0.364569	(tutorialspoint)
6	0.317594	(twitter)
7	0.511258	(youtube)
8	0.454027	(facebook, gmail)
9	0.420526	(youtube, facebook)
10	0.449111	(youtube, gmail)
11	0.386053	(facebook, youtube, gmail)

```
Staff_Apriori_rule = association_rules(staff_frequentItems,metric='confidence',min_threshold=0.8)
```

```
Staff_Apriori_rule
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(facebook)	(gmail)	0.553681	0.622383	0.454027	0.820015	1.317542	0.109425	2.098053
1	(youtube)	(facebook)	0.511258	0.553681	0.420526	0.822531	1.485568	0.137452	2.514911
2	(youtube)	(gmail)	0.511258	0.622383	0.449111	0.878443	1.411419	0.130913	3.106492
3	(youtube, facebook)	(gmail)	0.420526	0.622383	0.386053	0.918026	1.475018	0.124325	4.606532
4	(gmail, facebook)	(youtube)	0.454027	0.511258	0.386053	0.850287	1.663127	0.153928	3.264534
5	(youtube, gmail)	(facebook)	0.449111	0.553681	0.386053	0.859595	1.552509	0.137389	3.178789