Addis Ababa University

College of Natural Sciences

School of Information Science

Application of web usage mining for Extracting employee internet access pattern by URL category: The Case of Commercial Bank of Ethiopia

Yohannes Mesfin

October 2015

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

Application of web usage mining for Extracting employee internet access pattern by URL category: The Case of Commercial Bank of Ethiopia

A Thesis Submitted to the College of Natural Science of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

BY:

Yohannes Mesfin

October 2015

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

Application of web usage mining for Extracting employee internet access pattern by URL category: The Case of Commercial Bank of Ethiopia

BY:

Yohannes Mesfin

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	_____
Million Meshesha (PhD)	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

Table of Content

ACKNOWLEDGEMENT

First, I would like to thank my Advisor Dr. Million Meshesha for his unreserved guidance and support. He has shown me the right direction to complete my research successfully.

I would like to express my sincere gratitude to Security Administrators of Commercial Bank of Ethiopia for their support in many aspects whenever I required their professional assistance: Specially, Tamirat Andarge is standing on my side, and helping me on every situation.

I would also like to express my thanks to my office Document Authentication and Registration office IT Staff for acknowledging, covering my duty and facilitating the resources that I needed to carry out this research:

Finally, a very big thank you goes to my Family especially for my father whose encouragement, help, and motivates for total support staying on graduate school.

Yohannes Mesfin

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

CBE	Commercial Bank of Ethiopia
CSV	Comma-Separated Values
FP-Growth	Frequent Pattern Growth
FTP	File Transfer Protocol
HTTP	Hyper Text Transfer Protocol
IIS	Internet Information Service
KDD	Knowledge Discovery in Data
NCSA	National Center for Supercomputing Applications
SMTP	Simple Message Transfer Protocol
URL	Uniform Resource Locator
WAN	Wide Area Network
W3C	World Wide Web Consortium
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

The Internet offers e-mail, e-commerce and research tools to enhance productivity of the organization and employees. However, In Ethiopia internet bandwidth allocation is limited and costly. As a result, implementing efficient utilization of internet resources is a major concern. To this end, this study aims at extracting employee's internet access pattern and internet usage statistics using web usage mining and analytical tools to indicate efficient bandwidth utilization.

In this research, the Web Usage Mining processes model suggested by Sharma [1] is followed which consists of, data collection, data preprocessing, pattern discovery and pattern analysis. Iron port web appliance log data was used for pattern discovery. Moreover, URLProfiler and Mcafee trusted source online tools are used to classify and categorize URLs. Next the log files were preprocessed by applying tasks such as, data cleaning, data integration, feature selection, data separation into different time group, transaction identification is made. Finally a total of 225,448 transactions are used for statistical analysis and by parsing child URL paths and by counting frequency of duplicates 11,523 unique URL parent classes were used for the association rule pattern discovery experiment.

After the preprocessing was completed, experiment was conducted with the datasets using Weka Software and applying association rule to discover interesting patterns in different time groups. MS Excel were also employed to yield different useful statistical reports including frequent off time URL category, frequent work time URL category, denied service request time registered, and successful response time replied.

The finding indicates that Entertainment and Social URL categories are the most frequently accessed on off duty hours, whereas Internet Services and Business URL categories most frequently accessed on duty hours. In addition to this, employees request denying service during duty hours, Lack of caching algorithm efficiency, employees' violation for internet access policy are some of the findings. The major challenge that needs further investigations are categorizing URLs based on their context, which is recommended as future research direction.

Keywords: Web Usage Mining, Pattern Discovery, Association rule, Log file, URL Profiler, Mcafee trusted source

CHAPTER ONE

INTRODUCTION

1.1 Background

The Internet has revolutionized the way individuals, organizations and the whole society communicate. During its existence, the characteristics of the Internet have changed and it has become a more interactive platform than it was previously. People are increasingly discovering the new communicative possibilities that the Internet can offer. They are no longer connecting to the Internet only in order to find information on different web pages but also contributing to discussions with their opinions, experiences or other types of content [2].

The fast growth of web site over the 'World Wide Web' (WWW) has not only raised many concerns but also opened a window of opportunity for organizations to analyze the lifetime value of their customers, and also improve their cross marketing strategies. As more and more organizations rely on the WWW to conduct business, the traditional strategies and techniques for market analysis needs to be revisited. The new strategies involve analyzing a large collection of unstructured data. The data over the web is collected in the form of server access logs that are generated by the interaction of clients with the web site and stored automatically by web sequencing the form of transaction logs. Different genres of organizations can make use of this data by analyzing for respective purposes, such as, incorporate load balancing, efficient bandwidth utilization and implementing quality of service [3].

In today's digital age, Firewalls, clients, cookies and servers including Internet security appliance etc, collect large amounts of log data about their users and their users' online activities [4].Large-scale mining and sharing of this data has been a key driver of innovation and improvement in the quality of internet services.

According to Cisco Iron Port Web User Guide [5] the web security appliance is a robust, secure, efficient device that protects corporate networks against web-based malware and spy ware programs that can compromise corporate security and expose intellectual property. The Web Security appliance extends Iron Port's SMTP security applications to include protection for standard communication protocols, such as HTTP, HTTPS, and FTP. Malware ("malicious software") is software designed to infiltrate or damage a computer system without the owner's consent [5].

To this end, assessing the company's internet usage behaviour is the major task using technologies, like web mining. Web Mining is the area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web [3][4]. It has three taxonomies, such as web content mining, web structure mining and web usage mining. Web usage mining is the automatic discovery of user access patterns from Web logs. "The identified browsing and visit patterns can help in understanding the overall access patterns and trends of employees" "and allow for Web site design and reconstruction that is responsive to business goals and customer needs, such as user-level customization" [6] [7].

1.1.1. Overview of CBE

The Commercial Bank of Ethiopia (CBE) has established itself as the leading bank in the country and won the trust and support of its esteemed customers [8]. Even if a number of private commercial banks have come into the picture with the thriving of free market economy during the past years, the CBE maintains the lead in all parameters to customer base, capital, deposit, branch network and asset. As of June 30, 2013, its total asset reached Birr 158.11 billion. The number of its branches also reached 830 on June 30, 2014. And all these are reinforced by the close to 12,800 competent staff.

Moreover, the CBE is heavily investing in banking technology in an effort to bring about transformation in its service delivery and overall performance. Consequently, a number of IT projects including card banking and core banking solutions have been successfully implemented [8].

Up on preliminary investigation currently CBE have more than eight hundred computers that are networked to a server running Windows sever 2012. These computers are arranged in seven different administrative branches average 150 Pc's per branch.

Accordingly, CBE has computers ranging from five year old Pentiums to new Dell 790 computers, though most of the computers are 1-2 years old. The network is 2-3 years old, the wiring is all CAT5 and CAT6 and it is organized in designed full base, and is connect to the switches. The switches are (24 ports) and having 100Mbps speed. Moreover, CBE has a network infrastructure that allows it to administer its data centrally across branches and head office. Each branch has their own Network Address, subnet mask and default gateways connected to each other through data line leased from ISP (ETC). In addition to this, the up link is connected in fiber for getting an internet connection from the head office and also they are sharing a 30MB internet connection.

In doing so, CBE rent out a 30 Mega bits/sec internet access bandwidth from Ethio-Telecom in order to get an internet connection for privileged employees to access different resources. In addition to this CBE formulate and implement Internet usage policy for employees with rules and guidelines about the appropriate use of company equipment, network and Internet access. However, access to the Internet through company is a privilege and all employees must adhere to the policies concerning Computer, Email and Internet usage. Violation of these policies could result in disciplinary and/or legal action leading up to and including termination of employment. Employees may also be held personally liable for damages caused by any violations of this policy. Furthermore, every privileged employee located either head office or branch office accessing internet through one external link that enables a centralized management. In addition, CBE uses Cisco Iron Port web security appliance in order to filter all privileged employees internet access as well as applying usage policy.

1.2. Statement of the Problem

The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view, users, web service providers, business analysts. The users want to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the website suited for the different group of users [9].

The Internet has historically offered a single level of service, that of "best effort," where all data packets are treated with equity in the network [10]. Employee Internet access has become commonplace. Providing instant access to information and a vehicle for employee communication, as we know Internet connectivity presents the company with new risks that must be addressed to safeguard the facility's vital information assets. These risks include: Access to the Internet by personnel that is inconsistent with business needs results in the misuse of resources. These activities may adversely affect productivity due to time spent using or "surfing" the Internet. Additionally, the company may face loss of reputation and possible legal action through other types of misuse. All information found on the Internet should be considered suspect until confirmed by another reliable source.

However, while the Internet offers e-mail, e-commerce and research tools, it also opens more avenues for employee distraction, ranging from the innocent entertainment and shopping to

the offensive, such as gambling and pornography. As a result, the web is affecting productivity and creating more serious problems, like hostile workplace lawsuits. This means the management of employee Internet access is shifting from a strictly information technology (IT) matter. Achieving a balance between a company and its employee interests is advantageous, especially when the Internet is considered an employee benefit these days.

Internet services play an increasingly important role both in global and national economy. To be able to survive in the increasing competition, CBE should organize their operation according to the needs expressed (or in several cases even not expressed) by their employees.

Based on the preliminary investigation employee's of CBE assumed that there is 30 Mega bit internet connection speed and expecting to access the internet at a high-speed or needs to open a page on a single click without any delay. But most employee's are not satisfy with the speed of the internet access. There are several factors that contribute for this problem. One of the factors that contribute to this is that managing and allocating the bandwidth are poorly designed because employee's requirements and employees trends of accessing internet resources are not often incorporated into the bandwidth management. Besides that the security administrators a doubt that Ethio-Telecom does not give the rented internet connection speed as of expected.

Different Web service for improvement and to quantitatively measure the efficiency of the bandwidth, several related aspects of the network service are often that can be divided into user-independent and user-dependent [31]. Values of the user-independent bandwidth efficiency measurement properties (e.g., price, popularity, etc.) are usually advertised by service providers and identical for different users. On the other hand, values of the user-dependent bandwidth efficiency measurement properties (e.g., failure probability, response time, throughput, etc.) can vary widely for different users influenced by the unpredictable Internet connections and the heterogeneous user environments [31].

Studies have been conducted using web log data in order to discover Web usage pattern to explore users' access pattern, describe usage statistics using data mining technology and analytical tools so as to suggest a better and targeted way of redesigning and restructuring the website [11][12]. Those studies have used only web server log data for discovering user patterns and recommend future studies incorporate another log files like Firewalls, clients cookies and Internet security appliance etc, because these technology devices collect large amounts of log data about their users and their users' online activities [4]. On this specific study Firewall web security appliance log data is used for discovering new knowledge.

Recently Web usage mining has been gaining a lot of attention because of its potential commercial benefits. The ability to observe potential users as they browse through a virtual store promises to raise business intelligence to a new level. While demographic data can always be added to the web usage mining process.

However, as per the knowledge of the researcher, there is no research undertaken for studying on user's internet access usage behaviour by URL category to enhance efficiency of accessing internet resources by using Firewall web security appliance log data in the Commercial Bank of Ethiopia (CBE). Because CBE having connected different branches through a single residential gateway to access internet resources. So this connected branches employee's internet access navigation is registered on web security appliance log data. Therefore, in this research, an attempt is made to investigate the current internet access usage behaviour by categorizing URLs of employee's access usage on which internet accessing time for determining the most popular internet access service for describing employees' navigation and internet access trends so as to indicate a better bandwidth utilization for those internet services.

To this end, this research explores and answers the following research questions:

- How to prepare appropriate dataset with relevant attributes for the usage mining task?
- How to classify URLs for defining employees' internet access by time?
- How to group URLs based on its category?
- Which tools and algorithms are used for data preprocessing and extract Employee's internet usage behavior?
- What interesting patterns are discovered to describe employees' navigation and internet access behavior?

1.3. Objective of the Study

1.3.1. General Objective

The general objective of the study is to extract employee's internet access pattern and internet usage statistics based on URL categorization using web usage mining and analytical tools so as to indicate efficient internet bandwidth usage for improving quality of Internet service.

1.3.2. Specific Objectives

To achieve the general objective of the current study, the following specific objectives are identified:

- To review relevant literature to understand the problem area and web usage mining techniques and algorithms.

- To classify URLs for defining employees' internet access by time.
- To prepare training and test dataset required for web usage mining.
- To select appropriate algorithms to discover interesting patterns.
- To generate employees internet access behavior and usage statistics using different measurement metrics.

1.4. Scope and limitation of the study

This study focuses on extracting employee's internet access pattern and internet usage statistics based on URL categorization using web usage mining and analytical tools to determine employees of CBE internet access pattern so as to indicate efficient internet bandwidth usage. To deal with the Cisco Iron port web security appliance log data of different types such as access log, error log and SSL log. In this research access log is used for the web usage mining task to pre-process and to identify access patterns.

The main challenge of the research is dealing with huge amount of log files which posed many challenges while pre-processing the log files in terms of computational resources (CPU and Memory). The average log file exceeds 100GB and also which is not clearly defined on how much period of time the log file is incorporated and which makes it difficult to process the data of longer period. As a result, the researcher is arranged to configure to start a new log file to collect daily logs for limiting the amount of the dataset to a manageable size. Hence, the recent log files of 10 days is picked the year of 2015 august 15 to august 30 2015 (2 weeks data) from the web security appliance log data.

This study describes statistically and classifies employees and type of internet service in order to explore internet access usage of employees towards efficient bandwidth utilization. Association rule discovery techniques are used to extract hidden interesting patterns by URL category.

1.5. Significance of the study

As improving employees' service is one of the major concerns of CBE, the result of this study will help to know the various demands of employees with regard to the internet usage and address those demands. The study will give an opportunity to restructure the network infrastructure in a more effective way which will help the company to carry out effective internet bandwidth utilization and to render a better quality of service to access internet resources. In addition, to an increase in the satisfaction of employees' requirement; network infrastructure performance will also be improved.

1.6. Research Methodology

For better understanding of the research area, a number of related works reviewed from books, journal articles, and conference proceedings on the area. In order to highlight the key concepts of classification of URLS and users internet access trend related to log mining previously conducted researches and related publications reviewed. The main objective of this section to explain how the research was designed, to discuss the theoretical framework within which this research was conducted, to state the problem statement and the procedures for recording and manipulating the data, and to comment on the reliability of the research findings.

1.6.1. Research Design

There are different types of data mining models is designed like CRISP, SEMMA and HYBRID model in order to discover interesting patterns from raw log data [35].

Table 1-1 Summary of the correspondences between KDD, SEMMA and CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

From these models many researchers like Gethahun Nigatu and Awet Fesseha [11][12] are used one of KDD process suggested by Sharma web mining process model to discover remarkable patterns on their MSc thesis. In this specific research the study employs a three step Web Usage Mining process suggested by Sharma [1]. As depicted in figure 1-1 is applied, because, it covered all stages and easier to understand and implement. The steps includes: Pre-processing, Pattern discovery and Pattern analysis stages.

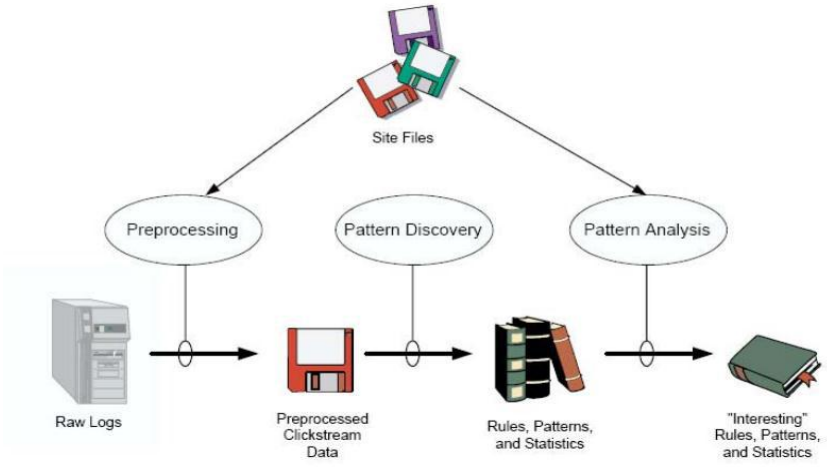


Figure 1-1 the overall architecture for the Web mining process

1.6.2. Data collection and Sampling Techniques

The researcher used recent iron port web security appliance log data over longer period of time provided all the required data are available. The pre-processed data labelled with Transaction ID to signify each employee's session. Then optimum sample size is determined through experiment. To select the sample data from the appliance daily log many experiments are applied, the average log file exceeds 100GB and also it is not clearly known on how much time the log file has incorporated which makes it difficult to process the data of longer period. As a result, the researcher has arranged to configure to start a new log file to collect daily logs to limit the amount of the dataset to a manageable size. Due to the bulk nature of log data and heavy computational resource processing requirements sample data is taken from two weeks log data by Proportionate Stratified Sampling picking only 10 hours from on duty hours and 10 hours from off duty hours. The data in each daily stratum had been identified by its time stamp. Then, by cutting only the necessary time stamp of daily log record and combined together with one file to make ready for pre-processing task.

1.6.3. Pre-processing

Web Usage Mining performs mining on web usage data, or web logs. Web logs can be regarded as a sequence of web pages accessed by users. Sometimes it is referred to as click stream data because each entry corresponds to a mouse click in a client-side browser. Web usage data includes data from web server logs, proxy server logs, browser logs, user profiles, registration data, cookies, user queries, etc. [15]

The data present in the log file cannot be used as it is for the mining process [15]. Therefore the contents of the log file should be cleaned in this pre-processing step. The unwanted data are removed and a cleaned log file is obtained. As depicted in figure 1-2 data pre-processing involves three major tasks [15].

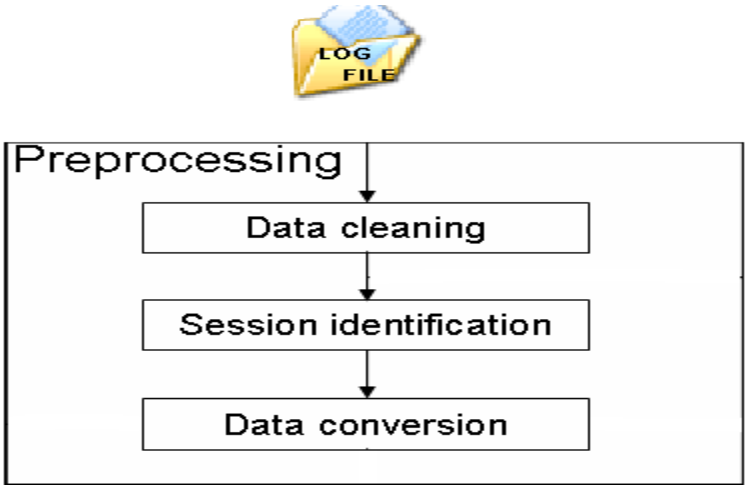


Figure 1-2 Pre-Processing steps

- *Data cleaning:* In this process the entries made in the log file for the unwanted view of images, graphics, Multi media etc., made by the users are removed. Once these data are removed the size of the file is minimized to a greater extent.
- *Session Identification:* This is done by using the time stamp since UNIX epoch details of the internet service. The total time (latency) used by each user of each visiting URL. This can also be done by noting down the user id (IP) those who have visited the internet access. Session is the time duration spent in the accessing internet service.
- *Data conversion:* This is conversion of the log file data into the format needed by the mining algorithms.

For such specific task, the researcher used em-editor from em-editor suite for text processing of the iron port log .s extension. Unlike other tools em-editor is a fast, lightweight, yet extensible, easy-to-use text editor for Windows file and it is easier and comfortable for text processing. For further pre-processing task the researcher used python 2.7 picked from python software foundation, which is a command based object oriented program editor for detailed pre-processing files. The reason behind to select and use this tool is? It is an open source and highly available and easier to familiarize. It presents the configuration options of the different python built in functions and scripts. In addition to this the researcher used MS-Excel 2007 for customizing result and some calculation.

1.6.4. Pattern discovery

After the conversion of the data in the log file into a formatted data the pattern discovery process is under gone [10]. With the existing data of the log files many useful patterns are discovered either with user ids, session details, time outs etc.

To discover usage patterns several experiments are conducted using statistical and data mining techniques. In this study, Statistical analysis is also conducted to give descriptive analysis of the web usage logs. In doing so, MS Excel is used for statistical analysis and charting, calculation and summary purpose. In addition to this the researcher is used weka 3.7 for pattern and association rule discovery because

1. A comprehensive collection of data pre-processing and modelling techniques.
2. It is freely available under GNU General Public License.
3. Weka s/w contain very graphical user interface, so the system is very easy to access.
4. As weka is platform independent & portable.
5. There is very large collection of different data mining algorithms [33].

Moreover weka implements the popular algorithms for association mining such as Apriori and FP-Growth and it is available for free and also these algorithms are suitable for examining Web Usage Mining [11]. Web Usage Mining is the process by which we identify

the browsing patterns by analysing the navigational behaviour of user. It focuses on techniques that are used to predict the user behaviour while the user interacts with the web. It uses the secondary data on the web. This activity involves the automatic discovery of user access patterns from one or more web servers. Through this mining technique can ascertain what users are looking for on Internet. It consists of three phases, namely pre- processing, pattern discovery, and pattern analysis. For the purpose of association rule mining, Apriori and FP-Growth algorithm in WEKA 3.7 data mining tool is used. Refer description of the algorithm in Chapter two (Literature Review). These algorithms construct a candidate set of large item-sets, counting the number of occurrences of each candidate, and then determine large item-sets based on the predetermined minimum support [33].

- *Association Rules*: This technique is used to predict the correlation of items where the presence of one set of items in a transaction implies the presence of other items
- *Sequential Patterns*: sequential patterns discover the user's navigation behavior. The sequence of items occurring in one transaction has a particular order between the items or the events. The same sequence of item may re-occur in the same order.

1.6.5. Pattern analysis and validation

This analysis process would eliminate the irrelevant rules or patterns that were generated. They tend to extract the interesting rules or patterns from the output of the pattern discovery process [4]. The statistical reports and patterns discovered are analyzed using different techniques such as Literature Review, business knowledge and discussion with domain experts. After analyzing the findings, recommendations are forwarded for future researches.

1.7. Organization of the thesis

This Thesis report is organized into five chapters. Chapter one discusses introduction, background of the study, problem statement, objective of the research, significance of the research, scope and limitation of the research and research methodology. Chapter two is the literature review part in which different conceptual and empirical concepts regarding Web Usage Mining are discussed. Chapter three discusses data preparation. In this chapter, data collection, data cleaning, session identification, feature selection, transaction identification and data transformation are discussed. Moreover, algorithms used for data pre-processing are described. Chapter four discusses experimentation and analysis. Here, different statistical analysis and experimentation of pattern discovery is presented and summarizes the research findings. Chapter five provides conclusion and recommendations based on the findings.

CHAPTER TWO

LITRATURE REVIEW

For better understanding of the research area, a number of related works reviewed from books, journal articles, and conference proceedings on the area. In order to highlight the key concepts related to log mining and pattern discovery previously conducted researches and related publications reviewed. Furthermore, in this review the researcher try to consult previous works in the area to underscore the factors and variables attributed to the successful pattern extraction by URL category.

2.1. Web Mining

The data over the web is collected in the form of server access logs that are generated by the interaction of clients with the web site and stored in the form of transaction logs. The web servers in the form of server or access logs generally automatically store this information. Different genres of organizations can make use of this data by analyzing for respective purposes. The analysis of the server logs can provide them with information on how to better structure the web site for effective use and benefit of the organization. For organizations working with intranet technologies, such analysis can lead to important information about managing the workgroup communication and the infrastructure of the organization. For advertising organizations, such analysis can help them in targeting a specific customer group after analyzing the user access patterns [14].

Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. Web mining is the application of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services [16][17]. Web Mining has three major sub-disciplines; namely, **Web Content Mining**, **Web Structure Mining** and **Web Usage Mining** [17].

Web Content Mining is that part of Web Mining which focuses on the raw information available in web pages. The type of data is mainly of textual in nature. Typical applications are content-based categorization and content-based ranking of web pages.

Web Structure Mining is that part of Web Mining which focuses on the structure of websites. Mainly it consists of the structural information in web pages such as links to other pages. Typical applications are link-based categorization of web pages, ranking of web pages through a combination of content and structure.

Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files, client logs and proxy server logs. The main source of data consists of the textual logs that are collected when users access web servers which are represented in standard formats.

2.2. Web Log File

Web log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text [18].

Log files are used to monitor web traffic. To configure the appliance to create log files, you create log subscriptions. A log subscription is an appliance/ configuration that associate a log file type with a name, logging level, and other parameters, such as size and destination information. You can subscribe to a variety of log file types [5].

2.2.1. Location of Web log File

Web log file is located in three different locations, **Web server logs**, **Web proxy server** and **Client browser** [18]. **Web server logs** or **Web log files** provide the most accurate and complete usage of data to web server. The log file do not record cached pages visited. Data of log files are sensitive, personal information so web server keeps them closed [18].

Web proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Client send request to web server via proxy server.

Client browser log file can reside in client's browser window itself. HTTP cookies used for client browser. These HTTP cookies are pieces of information generated by a web server and stored in user's computer that are ready for future access.

2.2.2. Type of Web log file

There are four types of server logs [18]: Access log file, Error log file, Agent log file and Referrer log file. Access log file is data of all incoming request and information about client of server. Access log records all requests that are processed by server. On the other hand, Error log file is the list of internal errors. Whenever an error occurs while the page is being requested by the client to web server the entry is made in error log. Agent log file contains information about user's browser version. Finally, referrer log file provides information about link and redirects visitor to site.

The log file type indicates what information is recorded in the generated log, such as web traffic or system data. By default, the Web Security appliance has log subscriptions for most

log file types already created. However, there are some log file types that specific to troubleshooting the Web Proxy [5].

2.2.2.1 Access Log File

The access log file provides a descriptive record of all Web Proxy filtering and scanning activity. Access log file entries display a record of how the appliance handled each transaction [5].

The W3C access log also records all Web Proxy filtering and scanning activity, but in a format that is W3C compliant. The following text is an example access log file entry for a single transaction [5]:

1278096903.150 97 172.xx.xx.xx TCP_MISS/200 8187 GET http://my.site.com/ - DIRECT/my.site.com text/plain DEFAULT_CASE_11-AccessOrDecryptionPolicy-Identity-OutboundMalwareScanningPolicy-DataSecurityPolicy-ExternalDLPPolicy-RoutingPolicy <IW_comp,6.9,"Skipped","-",-,-,-,"-","-",-,-,-,"-","-","-","-","-","-",-,-,IW_comp,-,"-","-","Unknown","Unknown","-","-",198.34,0,-,"-","-"> -

Table 2-1 Access Log File Entry

Field Value	Field Description
1278096903.150	Timestamp since UNIX epoch.
97	Elapsed time (latency) in milliseconds.
172.xx.xx.xx	Client IP address. **Note that:** It is possible to choose to mask the IP address in the access logs using the advancedproxyconfig > authentication CLI command.
TCP_MISS/	Transaction result code.
200	HTTP response code.
8187	Response size (headers + body).
GET http://my.site.com/	First line of the request. **Note:** When the first line of the request is for a native FTP transaction, some special characters in the file name are URL encoded in the access logs. For example, the "@" symbol is written as "%40" in the access logs. The following characters are URL encoded: and # % + , : ; = @ ^ { } [ ]
-	Authenticated username. **Note:** You can choose to mask the username in the access logs using the advancedproxyconfig > authentication CLI command.
DIRECT/my.site.com	Code that describes which server was contacted for the retrieving the request content.Most common values include: • **NONE.** The Web Proxy had the content, so it did not contact any other server to retrieve the content. • **DIRECT.** The Web Proxy went to the server named in the request to get the content.

	• **DEFAULT_PARENT.** The Web Proxy went to its primary parent proxy or an external DLP server to get the content.
text/plain	Response body MIME type.
DEFAULT_CASE_11	ACL decision tag. **Note:** The end of the ACL decision tag includes a dynamically generated number that the Web Proxy uses internally. You can ignore this number.
AccessOrDecryptionPolicy	Access Policy or Decryption Policy group name. When the transaction matches the global Access Policy or global Decryption Policy, this value is "DefaultGroup." Any space in the policy group name is replaced with an underscore (_).
Identity	Identity policy group name. Any space in the policy group name is replaced with an underscore (_).
OutboundMalwareScanning Policy	Outbound Malware Scanning Policy group name. Any space in the policy group name is replaced with an underscore (_).
DataSecurityPolicy	IronPort Data Security Policy group name. When the transaction matches the global IronPort Data Security Policy, this value is DefaultGroup." This policy group name only appears when IronPort Data Security Filters is enabled. "NONE" appears when no Data Security Policy was applied. Any space in the policy group name is replaced with an underscore (_).
ExternalDLPPolicy	External DLP Policy group name. When thetransaction matches the global External DLPPolicy, this value is "DefaultGroup." "NONE" appears when no External DLP Policy was applied. Any space in the policy group name is replaced with an underscore (_).
RoutingPolicy	Routing Policy group name as: *ProxyGroupName/ProxyServerName*. When the transaction matches the global Routing Policy, this value is "DefaultRouting." When no upstream proxy server is used, this value is "DIRECT." Any space in the policy group name is replaced with an underscore (_).
<IW_comp,6.9,"Skipped","-",-,-,-,"-","-",-,-,-,"-","-","-","-","-","-",-,-,IW_comp,-,"-","-","Unknown","Unknown","-","-",198.34,0,-,"-","-">	Scanning verdict information. Inside the angled brackets, the access logs include verdict information from various scanning engines.
-	Suspect user agent.

2.2.2.2 Traffic Monitor Log

The L4 Traffic Monitor log file provides a detailed record of monitoring activity. You can view L4 Traffic Monitor log file entries and track updates to firewall block lists and firewall allow lists [5].

The L4 Traffic Monitor log file provides a detailed record of monitoring activity. You can view L4 Traffic Monitor log file entries and track updates to firewall block lists and firewall allow lists. Consider the following example log entries:

172.xx.xx.xx discovered for blocksite.net (blocksite.net) added to firewall block list

In this particular example, a match becomes a block list firewall entry. The L4 Traffic Monitor matched an IP address to a domain name in the block list based on a DNS request which passed through the appliance. The IP address is then entered into the block list for the firewall.

2.2.2.3. Web Proxy Logging

By default, the Web Security appliance has one log subscription created for Web Proxy logging messages, the "Default Proxy Logs." The Web Proxy information stored in this log covers all aspects, or modules, of the Web Proxy. The appliance also includes log file types for each Web Proxy module. The following list includes all Web Proxy module log types [5]:

• Access Control Engine Logs
• AVC Engine Framework Logs
• Configuration Logs
• Connection Management Logs
• Data Security Module Logs
• DCA Engine Framework Logs
• Disk Manager Logs
• FTP Proxy Logs
• HTTPS Logs
• License Module Logs
• Logging Framework Logs
• McAfee Integration Framework Logs
• Memory Manager Logs
• Miscellaneous Proxy Modules Logs
• Request Debug Logs
• SNMP Module Logs
• Sophos Integration Framework Logs
• WBRS Framework Logs

• WCCP Module Logs

• Webcat Integration Framework Logs

• Webroot Integration Framework Logs

The appliance also creates other log file types, such as the system log file.

Nowadays, due to high rate of data generation from different transactions and complex nature of data, extracting useful information and knowledge manually have become almost impossible. Hence, for the purpose of problem solving, and fast and accurate decision making organizations are forced to look for automated processes which are capable of extracting useful information and knowledge from a large accumulation of data [19].

Data mining is a field of study that is used to discover new and useful knowledge from large data sources such as data stored in large databases, warehouses and other massive information repositories. Data mining involves different techniques to uncover hidden patterns form a large dataset. Data mining tasks are generally categorized as classification, clustering, and association rule discovery [20].

2.2.3. Transaction Result Codes

Transaction result codes in the access log file describe how the appliance resolves client requests. For example, if a request for an object can be resolved from the cache, the result code is TCP_HIT. However, if the object is not in the cache and the appliance pulls the object from an origin server, the result code is TCP_MISS.The following table describes transaction result codes.

Table 2-2 Transaction Result Codes

Result Code	Description
TCP_HIT	The object requested was fetched from the disk cache.
TCP_IMS_HIT	The client sent an IMS (If-Modified-Since) request for an object and the object was found in the cache. The proxy responds with a 304 response.
TCP_MEM_HIT	The object requested was fetched from the memory cache.
TCP_MISS	The object was not found in the cache so it was fetched from the origin server.
TCP_REFRESH_HIT	The object was in the cache but had expired. The proxy sent an IMS (If-Modified-Since) request to the origin server and the server confirmed that the object has not been modified. Therefore the appliance fetched the object from either the disk or memory

	cache.
TCP_CLIENT_REFRESH_MISS	The client sent a “don’t fetch response from cache” request by issuing the ‘Pragma: no-cache’ header. Due to this header from the client the appliance fetched the object from the origin server.
TCP_DENIED	The client request was denied due to Access Policies.
NONE	There was an error in the transaction. For example, a DNS failure or gateway timeout.

2.2.4. Web log file Format

Web log file is a simple plain text file which records information about each user. Log files data have three different formats [18]: **W3C Extended log file format**, **NCSA common log file format**, and **IIS log file format.**

NCSA and IIS log file format the data logged for each request is fixed. W3C format allows user to choose properties, user want to log for each request [18].

W3C extended log file format is default log file format on IIS server. Fields are separated by space, time is recorded as GMT. It can be customized that is administrators can add or remove fields depending on what information is needed to record [18].

NCSA common log file format is a fixed ASCII text-based format, so it cannot be customized. The NCSA Common log file format is available for Websites and for SMTP and NNTP services, but it is not available for FTP sites [18].

IIS log file format is a fixed ASCII text-based format, so it cannot be customized [18].

2.3. Overview of Knowledge discovery in log data

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate 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 as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis [21].

.

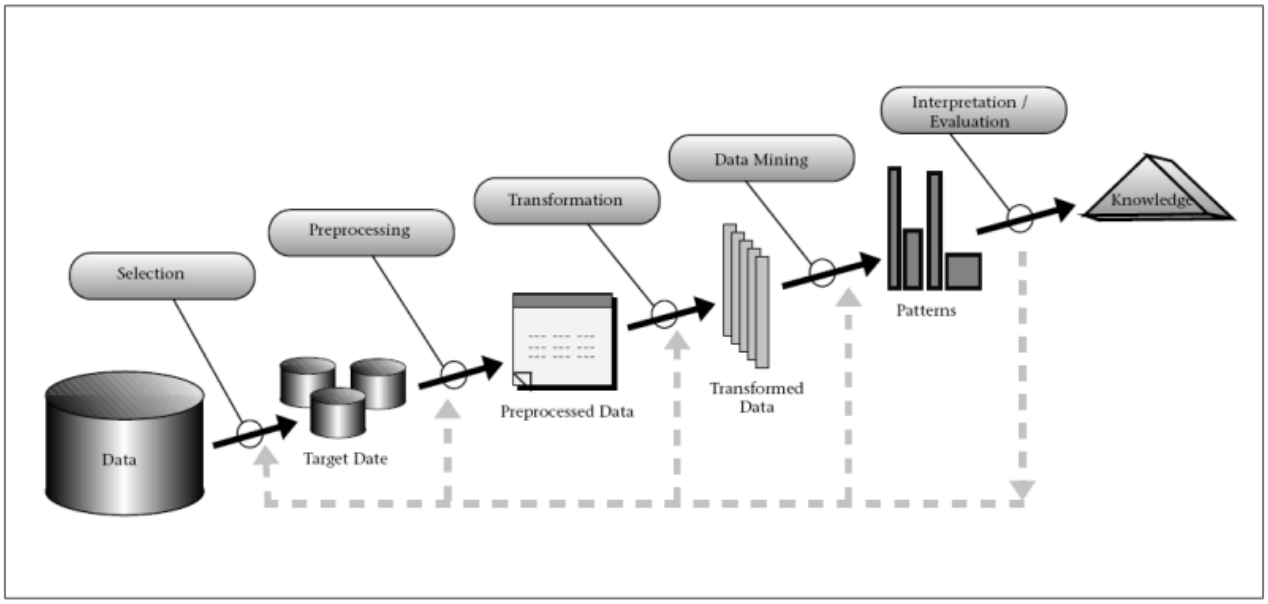


Figure 2-1. Knowledge discovery Process

2.4. Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Rule induction, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases [21].

2.4.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier [21].

Types of classification models include Decision tree induction, Bayesian Classification, Neural Networks, Support Vector Machines (SVM) and Classification Based on Associations (Refer [SOURCE] for detail description).

2.4.2 Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality [21].

Types of clustering methods include Partitioning Methods, Hierarchical Agglomerative (divisive) methods, Density based methods, Grid-based methods and Model-based methods (Refer [SOURCE] for detail description).

2.4.3 Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models [21].

Types of regression methods include Linear Regression, Multivariate Linear Regression, Nonlinear Regression and Multivariate Nonlinear Regression (Refer [SOURCE] for detail description).

2.4.4 Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behaviour analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value [21].

Types of association rules include Multi-level association rule, Multidimensional association rule and Quantitative association rule [21].

One of the fundamental data mining tasks is mining of association rules. It is perhaps the most important model invented and extensively studied by the database and data mining community. Its objective is to find all co-occurrence relationships, called associations, among data items. It has attracted a great deal of attention [22].

Association Rule mining deals with discovering patterns of the form (A->B) from frequent item sets where A is an antecedent and B is a consequent. This means that the occurrence of A implies the occurrence of B with a given support and confidence level.

The following basic steps are involved in association rule discovery [25].

- Generate item frequent k-item sets that satisfy the minimum support threshold value where k is one or multiple items.
- Generate association rules on those frequent k-item sets with the minimum confidence constraint.

In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. For example, association rule discovery may reveal a correlation between users who visited a page containing electronic products to those who access a page about supporting equipment. Aside from being applicable for business and marketing applications, the presence or absence of such rules can help Web designers to restructure their Website. The association rules may also serve as a heuristic for pre-fetching documents in order to reduce user-perceived latency when loading a page from a remote site [23].

Let I = {I1, I2…Im} be an item set. Let D, the task-relevant data, be a set of database transactions where each transaction T is a non-empty item set such that $T \subseteq I$. Each transaction is associated with an identifier, called a TID. Let A be a set of items. A transaction T is said to contain A if $A \subset T$. An association rule is an implication of the form A =>B, where $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ and $A \cap B = \emptyset$. The rule A =>B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$. This is taken to be the probability P ($A \cup B$).

The rule A =>B has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A) [23].

Support: The support of a rule, X ->Y, is the percentage of transactions in T that contains X ->Y, and can be seen as an estimate of the probability, Pr(X->Y). The rule support thus

determines how frequent the rule is applicable in the transaction set T. Let n be the number of transactions in T.[10]The support of the rule X ->Y is computed as follows:

Support (X=> Y) = P (X∪ Y)

Support is a useful measure because if it is too low, the rule may just occur due to chance. Furthermore, in a business environment, a rule covering too few cases (or transactions) may not be useful because it does not make business sense to act on such a rule (not profitable) [22].

Confidence: The confidence of a rule, A ->B, is the percentage of transactions in T that contain A also contain B. It can be seen as an estimate of the conditional probability, Pr (A | B) [22].

It is computed as follows:

- Confidence thus determines the predictability of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict B from A.A rule with low predictability is of limited use.
- Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong [22].

2.4.4.1 Association Rules Discovery Algorithms

Even though there are a number of techniques for association rules discovery, as noted by Han and Kamber [25], the most widely used algorithms for pattern discovery are Apriori and Frequent Pattern Growth (FP-Growth).

i. Apriori

Apriori employs an iterative approach known as a level-wise search, where k-item sets are used to explore (k+1)-item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scans of the database [25].

To improve the efficiency of the level-wise generation of frequent item sets, an important property called the Apriori property is used to reduce the search space. Apriori property states that, all nonempty subsets of a frequent item set must also be frequent [25]

The Apriori algorithm works in two steps: [22]

1. **Generate all frequent item sets:** A frequent item set is an item set that has transaction support above min support. The Apriori algorithm relies on the Apriori downward closure property to efficiently generate all frequent item sets.

Downward Closure Property: If an item set has minimum support, then every non-empty subset of this item set also has minimum support i.e. any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set. Hence, if any (k-1)-subset of a candidate k-item set is not in Lk-1, then the candidate cannot be frequent either and so can be removed from Ck. This subset testing can be done quickly by maintaining a hash tree of all frequent items sets [25].

2. **Generate all confident association rules from the frequent item sets**: A confident association rule is a rule with confidence above min conf.

Limitations with Apriori approach [22]

Apriori algorithm may suffer from the following limitations:

- The main problem with association rule mining is that it often produces a huge number of item sets (and rules), tens of thousands, or more, which makes it hard for the user to analyze them to find those useful ones. This is called the **interestingness** problem [22].

- While implementing Apriori property, it requires two nontrivial computationally expensive processes which make the algorithm costly in both space-wise and time-wise [25]. An efficient implementation of the Apriori algorithm involves sophisticated data structures and programming techniques.

- The Apriori algorithm reduces the size of candidate frequent item sets by using "Apriori property by which potential sources to associative rules may be lost.

ii. FP-Growth

To overcome the inefficiency of Apriori mentioned above, an algorithm named frequent pattern growth (FP-Growth) was introduced. FP-growth was first proposed by Han et al. [25]. FP-Growth adopts a divide-and-conquer strategy. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree, which retains the item set association information. It then divides the compressed database into a set of *conditional databases* (a special kind of projected database), each associated with one frequent item or pattern fragment and mines each such database separately. With FP-Growth algorithm, no candidate generation is required [25].

As presented in Algorithm 2.1, in FP-Growth algorithm, the first scan of the database is the same as Apriori, which derives the set of frequent items (1-itemsets) and their support counts (frequencies). The set of frequent items is sorted in the order of descending support count. This resulting set or *list* is denoted *L*. An FP-tree is then constructed as follows. First, create

the root of the tree, labelled with "null". Scan database *D* a second time. The items in each transaction are processed in *L* order (i.e., sorted according to descending support count), and a branch is created for each transaction [25]. In general, when considering the branch to be added for a transaction, the count of each node along a common prefix is incremented by 1, and nodes for the items following the prefix are created and linked accordingly [25].

A study on the performance of the FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm [25]. The main advantage of FP-Growth algorithm is that it uses compact data structure and avoids repeated database scan.

2.5. Interestingness of Patterns

A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful and novel [25]. A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents new knowledge.

The support and confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, a correlation measure can be used to augment the support–confidence framework for association rules. That is, a correlation rule is measured not only by its support and confidence but also by the correlation between item sets A and B. There are many different correlation measures from which to choose. One of these is given below [22].

$$\text{Lift (A, B)} = \frac{P(AUB)}{P(A)P(B)}$$

The above equation is equivalent to which is also referred to as the lift of the association rule (or correlation) rule A B. In other words, it assesses the degree to which the occurrence of one ―lifts the occurrence of the other [25].

If the value of lift is less than 1, then the occurrence of A is negatively correlated with the occurrence of B, meaning that the occurrence of one likely leads to the absence of the other one. If the value of lift is greater than 1, then A and B are positively correlated; which means that the occurrence of one implies the occurrence of the other. If the value of lift is equal to 1, then A and B are independent and there is no correlation between them [25].

2.6. Related work

Hofgesang[30] has conducted a study with a goal of analyzing user behaviour by mining enriched web access log data on association rules mining algorithm is used for mining

frequent page sets and for generating interesting rules. The study also applied the mixture model to build a predictive model of navigational behaviours of users. The major results of the study are content and origin based data enrichment and a tree like visualization of frequent navigational sequences [30]

According to Iváncsy [29], User identification is one of the most challenging tasks in Web Usage Mining. This is because multiple users can use the same computer or a single user can use multiple computers to browse the website. Furthermore, proxy servers can hide relevant information about unique users as multiple computers appear on the internet using the same IP address through the proxy server.

Frequent pattern mining is a heavily researched area in the field of data mining with wide range of applications. One of them is to use frequent pattern discovery methods in Web log data. Discovering hidden information from Web log data is called Web usage mining. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behaviour of the users [27].

The characteristics of service quality which is intangible, heterogeneity, inseparability and perishability cannot be measured objectively (Patterson and Johnson. 1993). However, many researchers stated that service quality can be measured by making the comparisons between customers' expectations and perceptions [26]. The authors have distinguished the service quality into four types namely expected service; desired service; adequate service; and predicted service. Expected services referred to the services customers intend to obtain from the service provider. Desired services are the level of service which the customer wishes to obtain. Adequate service refers to the minimum level of services expected from the service provider and finally, predicted services are what the customers believe the company will perform.

There are a number of local studies conducted in the area of web mining especially in the area of assessing user behaviour of accessing the website and provide a modification for adjusting the website. In this study three different but related works reviewed as shown in table 2-3.

Table 2-3 summary of related works

Author	Title	Method and Algorithm	Tools	Data Source	Finding
Awet Fesseha (2011)	Web Usage: Exploring Navigational Behaviour of Users, the Case of the Official website of AAU	WUM Process, Apriori	WUMprep, Web Utilization Miner	Web Server Log	Most frequently accessed are The Home Page, Registrar and Graduate Admission pages
Getahun Nigatu (2014)	Web Usage Pattern Discovery and Analysis By Region: The Case Of Ethiopian Airlines Official Website	WUM prep, Apriori and FP-Growth algorithms	Google Analytics, MS Excel, Java, Mach5, Weka	Web server log data.	ShebaMiles, Contacts, NetworkBaggage related pages are the most frequently accessed pages by the customers in most of the regions. There are similarities and variations from one region to another with respect to the accessibility and usage pattern.
Hailay Weldegebriel (2011)	Developing dynamic bandwidth allocation prototype model for campus network based on network traffic analysis	Ntop, MRTG	Ntop, MRTG	Web Server log data.	TCP and UDP is the highest bandwidth consuming protocols

Awet [28] has conducted his study on exploring the navigational behaviour of users of Addis Ababa University (AAU) official website. The study mainly intended for utilizing Web server log files by using different analytical and knowledge discovery tool Moreover, it depicts about a suggested approach redesigning the website.

The study shows application of web mining will have a positive impact on the development or restructuring of the website. In order to develop and implement web mining the log data would follow the steps; Data pre-processing, pattern discovery and pattern analysis in his study. He used Apriori algorithm for pattern discovery.

Moreover, he utilized WUM-prep tool for data pre-processing, Web Utilization Miner for statistical analysis and pattern discovery. In his research, he has described users' access behaviour of the website with different parameters and mapping the current knowledge and collaboration patterns within the organization is also important since it will help to determine the starting point for redesigning the website. His finding concludes to show the home page of AAU's website is the most frequently accessed and used page. He has also concluded that this same page is the top exit page which indicates that AAU website needs improvement in terms of its content. He also indicated which pages are accessed frequently after the first page. According to Awet's study AAU website is reached mostly either by directly typing the URL or via search engine which entails that there is less referral to AAU website from other websites.

Getahun's [12] in his study undertakes general website access statistics and pattern discovery. The statistical analysis was done based on one year's website usage and the pattern discovery was based on 30 days server log data which spans over 3 months, September, October and November 2013. The log files were thoroughly pre-processed in order to make them suitable for the mining task. Pre-processing of the log files was a challenging and time taking task mainly because of the log file size. After the processing and data transformation was completed, several experiments were conducted in order to discover interesting patterns in the website. Accordingly some useful results were found.
On his study an attempt is made to compare the accessibility of Ethiopian Airlines website by different regions. As can be seen a ratio of Number of Transaction against Number of Passengers is used for the comparison and accordingly the regions are listed descending order as follows: North America, Europe, Asia, Middle East and Africa. The data for Latin America is not available. This data can be used to show in which regions the website is more

accessed and in which regions it is less accessed relative to the number of passengers in the region.

He has used both data mining and statistical analysis and the result of his study shows that customers in North America region are mostly interested in Sheba Miles, Contacts, Network, FAQ and Call enter related pages. On the other hand, for Latin America, no interesting rule was discovered.

In general, Sheba Miles, Contacts, Network and Baggage related pages are the most frequently accessed pages by the customers in most of the regions. Hence, further attention should be given for these categories in order to make the pages in each category more informative and easily accessible. Moreover, he proposed for minor restructuring of the website should be made in order to rearrange them according to their access pattern and frequency. Moreover he also indicated that there are similarities and variations from one region to another with respect to the accessibility and usage pattern. Finally, he has forwarded a recommendation for further researches and improvement areas in order to improve the website.

Hailay [29] attempts to develop dynamic bandwidth management conceptual framework for intercampus network based on network traffic analysis. In his work, it is used Mekelle University web server log data a period of 6 days. Protocol distribution and network traffic load analysis are conducted on different links to assess the potential application of ntop for measuring the global protocol distribution based on the real data and as a result it shows TCP and UDP is the highest bandwidth consuming protocols currently present in the Mekelle University network. Moreover, the ingoing and outgoing traffic load is analyzed using MRTG every five minutes, it reads the inbound and outbound octet counter of the gateway router, and then logs the data to generate daily, weekly, monthly and yearly graphs for web pages. Based on the traffic analysis at different links and web server of MU he proposed, a dynamic bandwidth allocator algorithm which consists of modules such as administration tool, which provides a graphical interface for configuring bandwidth allocation based on the different bandwidth demand in the intercampus network; policy agent, which implements the configuration and handles communication with the kernel module; kernel module, which implements the packet classifier, packet selector, bandwidth estimator, unique IP address counter etc. During development of dynamic bandwidth management he suggest to consider the factor like the geographic locations in the campus, the number of VLANS in a specific building, number of users and traffic types passing through the link.

Lastly, He believed that full investigation is not done on the campus. Hence future work is needed to investigate further analysis on the network traffic and generate usage patterns even Ethiopian universities intercampus network too.

CHAPTER THREE

DATA PREPARATION

The web protocol requires a separate connection for every file that is requested from Cisco iron port. Therefore a user's request to view a particular page often results in several log entries. Since the main intent of web usage mining is to get a picture of user's behaviour, it is unnecessary to include file requests that the user didn't explicitly request [24]; and hence, before using the log files for web mining, it is necessary to thoroughly prepare it in order to remove noisy data and make it ready for Web Usage Mining task. Pre-processing on web usage data is required to eliminate noisy data and make data effective for further analysis task. The main pre-processing tasks include data cleaning, user identification, session identification, transaction identification and data transformation.

3.1. Data Collection

In this research, datasets are collected from CBE Cisco Iron port web appliance device. On this device the user access logs for the full year of 2007 is captured and stored in a file. Availing mining process to all record is un-imaginable. Due to this reason an attempt is made for configuring the web security appliance to capture new log file with new storage and then, getting and picking a sample of only two weeks log file is captured separately for this research. The original size of the data was huge before pre-processing. The Iron port log data is in common Log Format containing information about the access logs, error logs, transaction log ...etc of the usage access pattern. Figure 3-1 shows the sample log file captured by web security appliance.

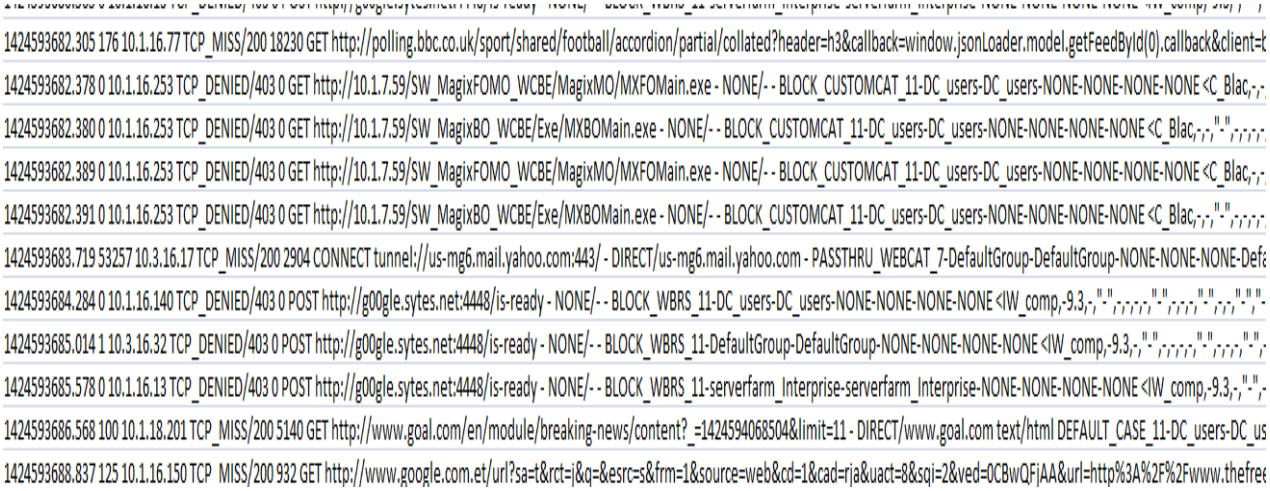


Figure 3-1 sample web log file

3.2. Data pre-processing

Basically, several pre-processing tasks need to be done before implementing web mining algorithm on web server logs. There are five preprocessing tasks such as, data cleaning, user identification, session identification, path completion and transaction identification [24].

3.2.1. Pre-Processing steps

To prepare the web server log for mining process, the data needs to be cleaned and preprocessed. Data cleaning is an important stage in data preprocessing. In data cleaning, certain techniques are used to remove irrelevant and non-significant items from the web appliance logs. The inputs for data preprocessing phase are the access logs transferred from the web security appliance and the outputs are user session file, transaction file and transformed to csv file which will be used for pattern discovery task. In this paper, the researcher arranged the tasks of cleaning the log data for determining the stages of data pre-processing is depicted in figure 3-2 in order to clean data, identify session, select feature, identify transaction and transform data.

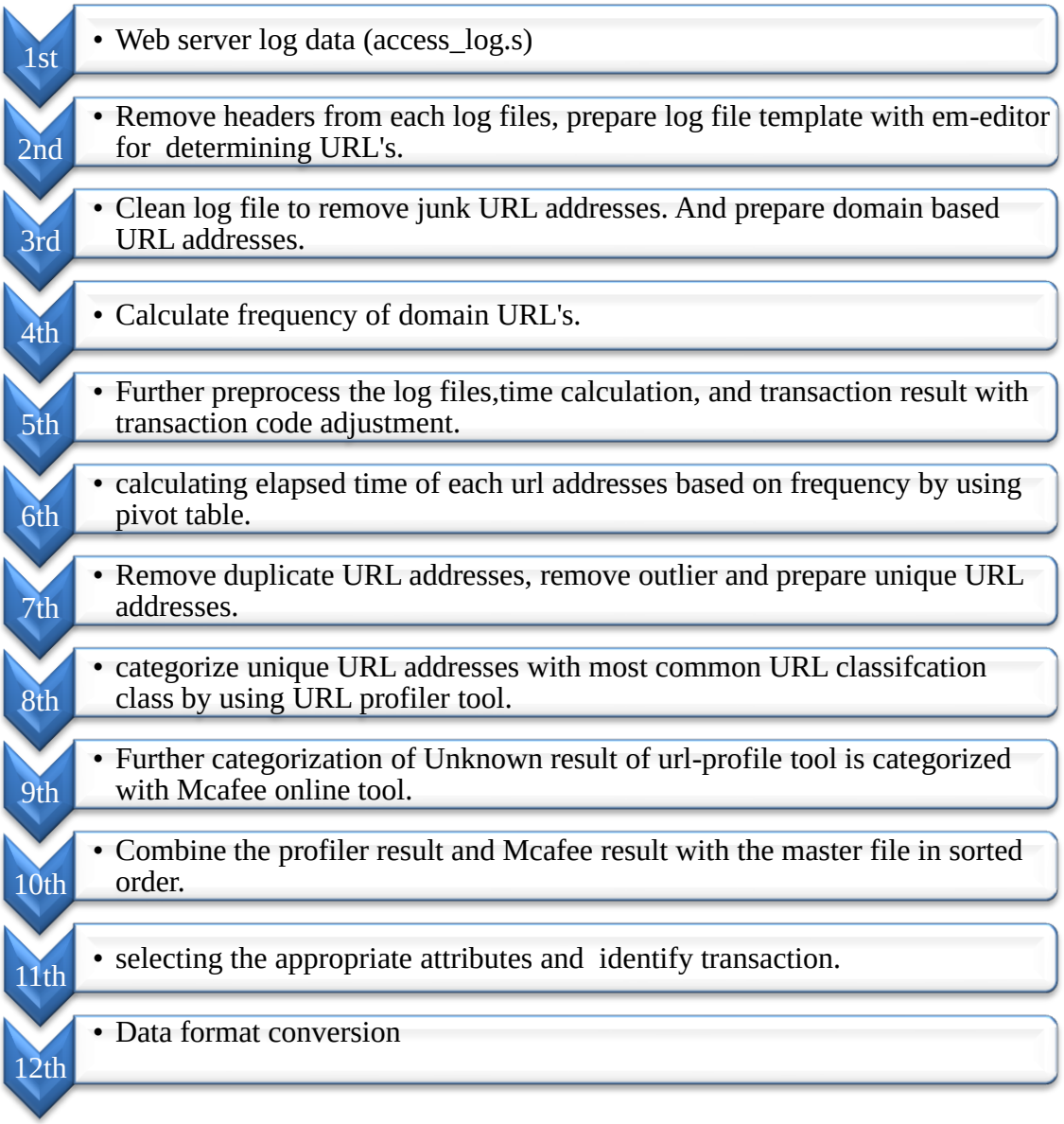


Figure 3-2 Preprocessing Steps followed in preparing server log file for mining

Next, a brief description of the steps in figure 3-2 is given.

3.2.1.1 Data cleaning

The appliance server logs are given in the .s format. The first step taken was to compile the web server logs based on the format. Most log files have their own unique characteristics format. As for this web appliance logs, configured the appliance format according to w3c standard. Then, the first stage of preprocessing of data cleaning; remove unnecessary files.

For such specific task, the researcher used em-editor from em-editor suite for text processing of the iron port log .s extension file and using python 2.7 from python software foundation which is a command based object oriented program editor for detailed pre-processing files.

Remove headers from each log files, prepare log file template with em-editor for determining URL's

In this step, software information that logged the data, version information, log server date and field names are removed from each day appliance log file header. And converting .s extension to .txt extension in order to prepare for further lists. Finally, all files are collected in one folder for better data management and easiest utilization of pre-processing activity.

Clean log file to remove junk URL addresses. And prepare domain based URL addresses

In this step, the researcher write a python code that reads lists of bulk URLs addresses and then trimming only the base class of URL address and remove the rest details of URL paths. See Appendix C.

Calculate frequency of domain URL's

In this step, the researcher calculate frequency of each occurrences of the URL base class with Ms-Excel and copying and pasting with paste special command only values to avoid the calculation point of reference. Then the calculated frequency result pointed in column. In Excel, by using COUNTIF function to count the duplicates of URL addresses.

Select a blank cell adjacent to the first data of your list, and type this formula *=COUNTIF($A$1:$A$8, A1)* (the range $A$1:$A$8 indicates the list of data, and A1 stands the cell you want to count the frequency, you can change them as you need), then press **Enter**, and drag the fill handle to fill the column you need.

Further preprocess the log files, time calculation, and transaction result with transaction code adjustment.

In this step, further pre-processing activities made. The log file time is presented in UNIX epoch format in order to convert this epoch to or based class “Off time” and “work time” the following activities are done.

First, calculate each UNIX epoch to a readable date and time (See appendix A).

Second, split dates and times that captured in one cell in to two separate cells remembering that in the internal Excel system the date value is stored as a whole part and the time value as a fractional part of a decimal number, you can extract the date using the INT function, which rounds the cell value down to the nearest integer (See appendix B).

Third, to extract the time portion, subtract the date returned by the above step from the original date and time value (See appendix B).

Finally, now we have only time with" if and time function "it can made easily for categorizing the time attribute categorized in to "Off time" and "work time".

Remove duplicate URL addresses, remove outlier and prepare unique URL addresses

In this step, by using pivot table sum up elapsed time associated with duplicate URL addresses frequency and then execute python code which is used to remove those duplicate records which are not required for the mining task (See appendix C).

Categorize unique URL addresses with most common URL classification class by using URLProfiler tool.

In this step, downloading the URLProfiler trial tool with registered account and it works on internet connection. There are different online tools for categorizing URLs such as check point software, trusted source, and similar web-Api. These tools are categorizing single URLs at a time. Unlike these tool URLProfiler tool which has significantly improved several of my data pre-processes for many different things can accept multiple URLs at a time to categorizing URLs based on its content. The researcher picked this URLProfiler tool that carries out numerous checks on a list of URL's which really help to speed up link classification.

On this regard URL classification or URL categorization stands for merging similar contents together by identifying URL and IP address.

Steps and consideration to use URLProfiler tool,

First, it is a licensed tool downloading the tool for trial purpose with registered account for only twenty days.

Second, installing on the machine and opens it after connecting the machine to the internet. If there is no internet connection the application didn't run.

Third after running the application choosing and selecting the appropriate domain class to meets your interest. After different experiment on this study only selecting "Site Type" and "IP Address" are necessary for URL classification.

Finally, The URLProfiler tool accepts a lot of common file formats. Such as a CSV or TXT file with the data you want to analyze.

See the Screen Shoots,

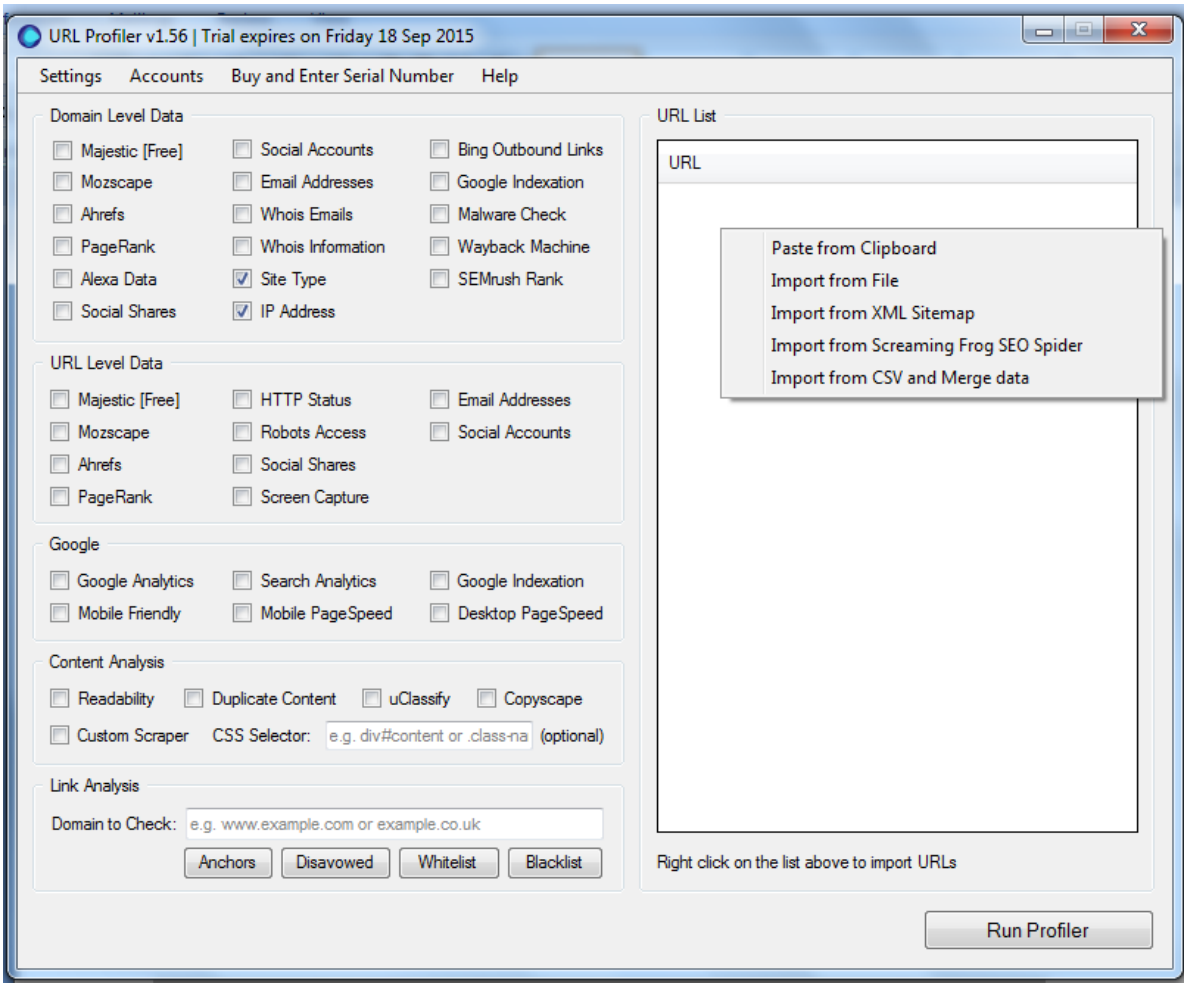


Figure 3-3 URLProfiler main window

URLProfiler facilitates URL classification task greatly for further analysis. However, As of getting URL profiler result two URL categories i.e. “unknown” and “not checked” will almost covered 65 percent of the recorded result; This means only 35 percent meets its purpose. To such extent, further experiments are made to avoid those things. Upon experiment the record “Not checked” result amount produced from URLProfiler is presenting local addresses. Therefore, filtering “Not checked” records and replaces its category to local addresses is maximizes URL profiler result to 65 percent.

The sample result is shown in figure 3-5,

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	URL	Path	Domain	Root Dom	TLD	Scheme	Name Ser	Name Ser	Server Co	IP Addres	Domains c	IP Addres	Domains c	Site Type	CMS
2	http://0.0.0.0/	/	0.0.0.0	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
3	http://0.gravatar.com/	/	0.gravatar	gravatar.c	com	http	cs91.wac.c	cs91.wac.c	United Sta	68.232.35.	5	68.232.35	29	Unknown	Unknown
4	http://0.s3.envato.com/	/	0.s3.envat	envato.co	com	http	d2mdw06	ns-1347.a	United Sta	54.192.184	1	54.192.184	6	Not Check	Not Checke
5	http://08d3c60f.seriousdeals.net/	/	08d3c60f.s	seriousde	net	http	lex-rewrit	lex-rewrit	United Sta	50.31.6.13	2	50.31.6	2	Unknown	Unknown
6	http://0haxor.blogspot.com/	/	0haxor.bl	0haxor.bl	blogspot.c	http	blogspot.l	blogspot.l	United Sta	173.194.11	1	173.194.11	54	Blog	Blogspot
7	http://1.bp.blogspot.com/	/	1.bp.blogs	bp.blogsp	blogspot.c	http	photos-ug	photos-ug	United Sta	173.194.11	3	173.194.11	54	Not Check	Not Checke
8	http://1.gravatar.com/	/	1.gravatar	gravatar.c	com	http	cs91.wac.c	cs91.wac.c	United Sta	68.232.35.	5	68.232.35	29	Unknown	Unknown
9	http://1.resources.www.etonline.	/	1.resource	edgesuite	net	http	a1824.b.al	a1824.b.al	South Afri	41.193.163	6	41.193.163	84	Not Check	Not Checke
10	http://1.www.s81c.com/	/	1.www.s8	s81c.com	com	http	global.ibn	global.ibm.com.edg	104.86.224	2	104.86.224	2	Unknown	Unknown	
11	http://10.1.11.8/	/	10.1.11.8	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
12	http://10.1.6.8:9095/	/	10.1.6.8	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
13	http://10.212.22.26/	/	10.212.22.	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
14	http://10.212.22.28/	/	10.212.22.	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
15	http://10.3.0.80:8014/	/	10.3.0.80	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
16	http://10.8.16.14:8014/	/	10.8.16.14	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
17	http://10.8.16.148:58064/	/	10.8.16.14	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	
18	http://10.8.16.220:8014/	/	10.8.16.22	IP Addres	IP Addres	http	a.root-servers.net.	nstld.verisi	Not Found	231	Not Found	231	Not Check	Not Checke	

Figure 3-4 single file sample result of URLProfiler.

The result of URL profiler tool meets the research perspective 65 percent that's also not sufficient in order to get correct Employee's behavior pattern so further investigation is made. In order to maximizing better inputs to research and tackle those challenges further study accomplished to get another option and gets online tool from McAfee.

Further categorization of Unknown result of url-profile tool is categorized with Mcafee online tool.

In this step, The McAfee TrustedSource.org website helps to maintain the highest possible level of security. By logged into the TrustedSource.org portal; getting access to additional functionalities.

This online tool enables to check if a site is categorized within various versions of the Smart-Filter Internet Database or the Web-washer URL Filter Database.

The Trusted Source Web Database uses categories to organize similar types of URLs into groups based on the content of the web page. For example, www.mcafee.com, www.trustedsource.org, and www.webwasher.com are grouped into the Business category. The categorization of a particular URL is a defined process using objective standards and definitions. Together and rate potential websites, McAfee uses various technologies, artificial intelligence techniques, such as link crawlers, security forensics, honey pot networks, sophisticated auto-rating tools, and customer logs. In some cases, one URL will be in more than one category. For example, www.mcafee.com and www.trustedsource.org are in the Business category. However, the two websites also contain software and hardware

information. Therefore, they are also in the Software/Hardware category. Due to overlapping content, taking the first category and remove the rest for focusing single URL categorization.

The rest 3649 record or 35 percent of the record is “unknown” URL classification those records is categorized with Mcafee online tool. This tool also having some limitation it takes only 100 rows with only 5KB file size. In order to arrange those records with those requirements is a tedious task. In doing so the researcher writes python code and execute to separate 3649 records in to 37 different files which are sequentially input to McAfee tool for classification (See appendix A).

The following picture shows Mcafee online tool.

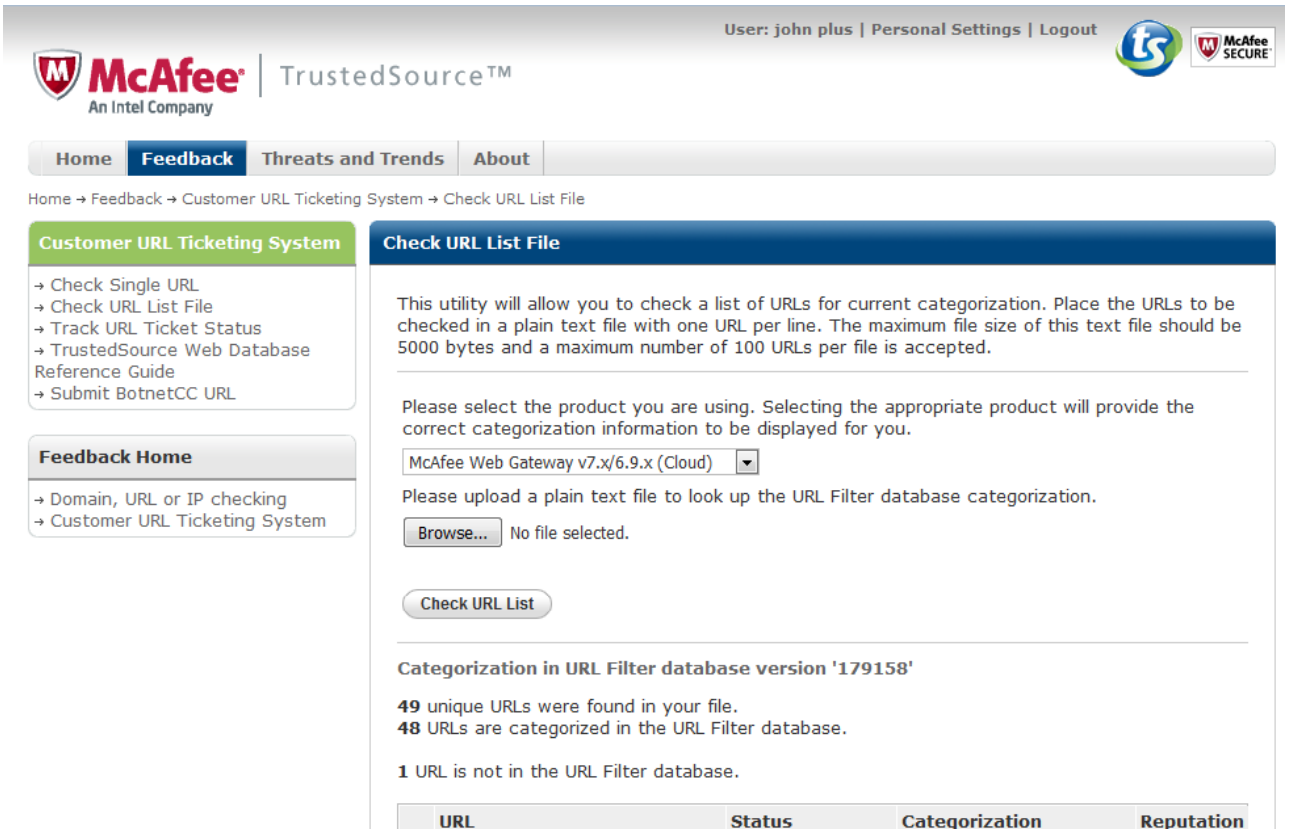


Figure 3-5 Mcafee trusted source lists of URL category checker online tool

Sample result in one file

Categorization in URL Filter database version '179158'

100 unique URLs were found in your file.

99 URLs are categorized in the URL Filter database.

1 URL is not in the URL Filter database.

Table 3-1 sample Mcafee trusted tool result.

	URL	Status	Categorization	Reputation
☐	http://www.way2sms.com/	Categorized URL	- Messaging	Minimal Risk
☐	http://www.webestools.com ...	Categorized URL	- Internet Services	Minimal Risk
☐	http://www.webmd.com/	Categorized URL	- Health	Minimal Risk
☐	http://www.webpopulation. ...	Categorized URL	- Internet Services	Minimal Risk
☐	http://www.westbromnews.c ...	Categorized URL	- Sports - Blogs/Wiki	Minimal Risk
☐	http://www.whitesmoke.com ...	Categorized URL	- Software/Hardware	Minimal Risk
☐	http://www.woolwicharsena ...	Categorized URL	- Sports	Minimal Risk
☐	http://www.writinglocalbu ...	Uncategorized URL		Unverified

As of the result obtained from McAfee tool More than 90% successfully categorize URLs with more than 50 different URL categories.

Combine the profiler result and Mcafee result with the master file in sorted order.

In this step, merge the records together in one file for easier management. After merging the log file records are sorted using the keys (Domain, Frequency and time) in order to arrange same sessions together.

3.2.1.2 Feature Selection for data transformation

Time-Group, latency, Client-IP, Transaction-result, Transaction-code, Response-size, method, contact-server, address-groups, frequency, and URL-Categorization attributes are relevant for association rules discovery. Hence, the other attributes are removed from the dataset. But as of records remove FTP starting URLs because of it doesn't have any URL categorization calculation on the tools and also remove outliers (URLs having one or two repetition on records is removed because it doesn't represent its URL category) for better association rule discovery result. As a result, the pre-processed and consolidated log file was edited with MS-Excel and grouped by path with frequency so that unique URL in the data with their counts is obtained. Then, the frequency of each URL category is calculated. The result was sorted in descending order of URL count so that the most frequently accessed are displayed on top using excel sheet. As of the result gaining from the online tools categories of URL addresses is more detail and it is still difficult to mine and get interesting pattern. For

standard guidance of website classification approach like, Search: e.g. Google, YouTube, Flickr… – the key activity is search for information, Social: e.g. Face book, Match.com, Stack Overflow… – the key activity is socialize with other members;

Up on these concepts an attempt is made for the following table to classify detail categories of child URL classes into a more parent grouped class.

Table 3.2 major category of URL classification

Categorization	**Group**
Blog	Blog
Business, Finance, Marketing/Merchandising, Mobile Phone, Online Shopping, Shopping, Software/Hardware, Stock Trading	Business
Cms	Cms
Entertainment, Gambling, Sports, Games, Public Information, Streaming Media	Entertainment
Education/Reference, Internet Services, Malicious Sites, Potential Illegal Software, Spam URLs, Technical/Business, Forums, Travel, Web Ads, Government/Military, Health, Information Security, Interactive Web Applications, Major Global Religions, Motor Vehicles, Non, Phishing, Real Estate, Remote Access, Technical Information, Text Translators, Web Mail	Internet Services
Not Checked	Local
Content Server, Job Search, Parked Domain, Search Engines	Search Engines
Instant Messaging, Media Downloads, Media Sharing, Messaging,P2P/File Sharing, Personal Network Storage, Portal Sites, Shareware/Freeware, Social, Social Networking, Wiki	Social

3.2.1.3. Transaction Identification

In this phase of pre-processing, transactions were identified for the consolidated data. The log data is further passed through some cleaning steps to remove irrelevant logs that escaped from previous processing activities. The pre-processed log file so far is containing all the access logs with all attributes and records.

Association rule mining can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes. Discretization is the process of putting values into buckets so that there are a limited number of possible states. The buckets themselves are treated as ordered and discrete values. Discretize can made both numeric and String columns [32].

Table 3.3 Discretization method

Discretization method	Description
AUTOMATIC	Analysis Services determines which discretization method to use.
CLUSTERS	The algorithm divides the data into groups by sampling the training data, initializing to a number of random points, and then running several iterations of the Microsoft Clustering algorithm using the Expectation Maximization (EM) clustering method. The CLUSTERS method is useful because it works on any distribution curve. However, it requires more processing time than the other discretization methods. This method can only be used with numeric columns.
EQUAL_AREAS	The algorithm divides the data into groups that contain an equal number of values. This method is best used for normal distribution curves, but does not work well if the distribution includes a large number of values that occur in a narrow group in the continuous data. For example, if one-half of the items have a cost of 0, one-half the data will occur under a single point in the curve. In such a distribution, this method breaks the data up in an effort to establish equal discretization into multiple areas. This produces an inaccurate representation of the data.

There are 3 such attributes in this data set: "Elapsed time ", "Response size ", and "Response Code" having continuous data so an attempt is made for changing these continuous data to discrete data. Firstly, sorting the data and dividing into three equal sections then, applying simple if else algorithm for changing those sections into discrete data.

The following calculation changes continuous elapsed time value into minimum, medium and maximum category.

```
If Elapsed time is less than 3000 Then
    Elapsed time is Minimum
Else if  Elapsed time is less than 6000 Then
    Elapsed time is Medium
Else
    Elapsed time is Maximum
```

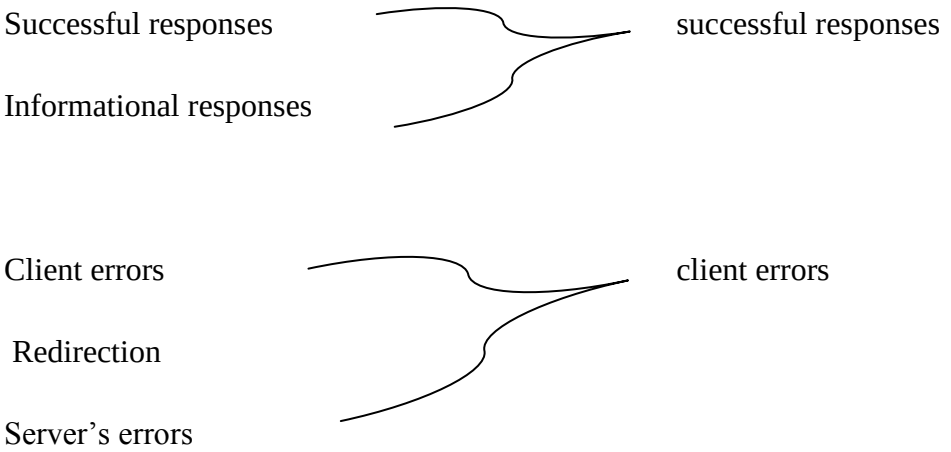
Same logic applies for changing response size to discrete one. The following calculation changes continuous response size value into small, medium and large category.

```
If  response size is less than 25000 Then
     response size is  Small
Else if   response size is less than 50000 Then
     response size is  Medium
Else
    response size is  Large
```

The next changes are applied on Hypertext Transfer Protocol (HTTP) response status codes. This includes codes from IETF internet standards as well as other IETF RFCs, other specifications and some additional commonly used codes. The first digit of the status code specifies one of five classes of response; the bare minimum for an HTTP client is that it recognizes these five classes. The phrases used are the standard examples, but any human-readable alternative can be provided. Unless otherwise stated, the status code is part of the HTTP/1.1 standard (RFC 7231).The Internet Assigned Numbers Authority (IANA) maintains the official registry of HTTP status codes [30].

HTTP Response Codes indicate whether a specific HTTP requests has been successfully completed. Responses are grouped in five classes: informational responses, successful responses, redirections, client errors, and server's errors.

In this specific study the researcher merges and transforms 5 categories into two successful response and client error responses.



However the paths are filtered in feature selection stage; moreover remove uncategorized URLs of logs and left 11, 523 unique records. This is the final log data which is used for transaction identification.

Once having the cleaned log data, next step was to sort each data in order to arrange the same transaction codes together. See table 3.2

Table 3-4 sample dataset used for experimentation

Time-Group	Elapsed-Time(Milliseconds)	IP-Address	Transaction-Result	Transaction-code	Response-Size	Method	Contact-Server	Frequency	URL	Categorization
OFF TIME	230	10.3.16.129	TCP_MISS	200	4652	GET	DIRECT	6	http://1.bp.blogspot.com/	Blog
OFF TIME	539	10.1.16.172	TCP_MISS	200	10807	GET	DIRECT	38	http://1.www.s81c.com/	Entertainment
WORK TIME	1894	10.8.17.188	TCP_DENIED	403	0	POST	NONE	63	http://10.2.0.80:8014/	Local
OFF TIME	1489	10.8.16.169	TCP_DENIED	403	0	POST	NONE	1	http://1.www.s81c.com/	Business
Work TIME	786	10.1.16.229	TCP_DENIED	403	0	POST	NONE	1	http://widgets.pinterest.com/	Internet service

After dataset preparation, association rule discovery algorithms are used for extracting patterns during experimentation using WEKA knowledge discovery tool.

3.2.1.4. Data Format Conversion

The last step in Data Pre-processing was to transform the transactions to a Weka understandable format. As a result, the files in above section were edited with MS Excel in order to remove un-necessary column, add the column labels to exact identifiable name. Then the resulting file, which is saved in a separate folder with ―.csv format.

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS

In this study an experiment is conducted using both data mining and statistical analysis so as to mine web usage patterns in CBE internet users.

To discover the pattern of internet access and make analysis on the result, an experiment is conducted using the pre-processed data. Some modifications have been done on the dataset in the process of getting interesting patterns whenever required. The experimentation and findings focus on two main aspects: Statistical analysis and pattern discovery.

The computer used for conducting this experiment is Dell Laptop computer having Intel Core(TM) 2 duo processor@2.93 GHz speed and 2 GB RAM size. The Operating System is Windows 7 Ultimate (32-bit). The analysis tools used for discovering the pattern are Weka 3.7.9 knowledge discovery software for pattern discovery and MS Excel 2010 for Statistical Analysis.

4.1 Statistical analysis

4.1.1. Most Frequently Accessed Services

Out of the sampled log file, it was found that the following URL categories are the top frequently accessed services (See Table 4.1). The entertainment, the Internet service and Business pages are the most three frequently accessed services while, CMS search engine and locals are not frequently accessed services. On this regard special attention should be given to those most frequently accessed URL services in terms of efficient bandwidth management for maximizing employee's satisfaction and also adding quality of internet access service.

Table 4.1 frequency of Accessed Services

URL Category	Frequency	Percentage
CMS	4371	1.94%
Search Engine	7116	3.16%
Local	8093	3.59%
Blog	25136	11.15%
Social	26222	11.63%
Business	46736	20.73%
Internet Service	53785	23.86%
Entertainment	53989	23.95%

4.1.2. Transaction access result

The appliance provides caching services for user access patterns in order to make fast response to clients whether by its own caching algorithm mechanism. When the log file generates depends up on the transaction result stated in chapter two. The following lists of transaction result shows most user requests is not cashed in the storage (See Table 4.2). As of the result shows, transaction access result TCP_MISS is more than 70% of the record of the sampled log file. This indicated most services are not replied from caching storage thus, employees request is trying to contact resources from origin server as a result the caching service needs to restructure to replied requests from its cash so as to do fast response time for employees and also employees are requesting to access denied services is the second frequently transaction result. This indicates that employees are lack of knowhow or violated the rule stated on internet access policy.

Table 4.2 frequency of transaction result

Transaction Result	**Frequency**	**Percentage**
TCP_DENIED_SSL	138	0.06%
TCP_IMS_HIT	221	0.10%
TCP_MEM_HIT	332	0.15%
TCP_CLIENT_REFRESH_MISS_SSL	442	0.20%
NONE	2765	1.23%
TCP_MISS_SSL	5945	2.64%
TCP_CLIENT_REFRESH_MISS	8434	3.74%
TCP_REFRESH_HIT	12084	5.36%
TCP_DENIED	36307	16.10%
TCP_MISS	158779	70.43%

4.1.3. Top response size

Every paged accessed by the user have been calculated for its page size. It describes how does the services consumed associated bandwidth, shorter page size having shorter response size consumed less bandwidth while larger page size having larger response size consumed high bandwidth. From the above result one hand, shorter response size of URL category is CMS then local and then searches engines. These URL category services are having quick response. On another hand, social, Internet Services, and Business URL category, having larger response size and they are lately respond (See Table 4.3).

Table 4.3 response size of URL category

Categories of URL	Sum of Response-Size	Percentage
Cms	6,642,917	0.41%
Local	14,608,392	0.91%
Search Engines	51,941,462	3.24%
Entertainment	141,471,869	8.84%
Blog	179,197,769	11.19%
Business	261,219,337	16.31%
Internet Services	352,511,689	22.02%
Social	593,562,010	37.07%
Grand Total	1,601,155,445	100.00%

4.1.4. Top URL Category by Time-Group

The knowledge of how the URL is being accessed by different time parameters could be useful in order to improve the URL category by different time group for making and apply appropriate Internet usage policy. The following table figure 4-1 shows that most frequently accessed URL category on different time group. It is required to analyze which URL categories are frequently on which time group. Up on statistical result during off time, Entertainment, social and internet service is frequently accessed, whereas Internet services Business and Blog is most frequently accessed on work time. From aggregate result frequent accessed services are Entertainment, Internet services and Business.

Table 4.4 URL category with both off time and work time accessing frequency

URL Category	OFF TIME	Percentage	WORK TIME	Percentage	Grand Total	Percentage
Blog	5,446	5.07%	19,690	16.69%	25,136	11.15%
Business	13,609	12.66%	33,127	28.08%	46,736	20.73%
Cms	3,505	3.26%	866	0.73%	4,371	1.94%
Entertainment	37,614	35.00%	16,375	13.88%	53,989	23.95%
Internet Services	15,886	14.78%	37,899	32.13%	53,785	23.86%
Local	600	0.56%	7,493	6.35%	8,093	3.59%
Search Engines	5,291	4.92%	1,825	1.55%	7,116	3.16%
Social	25,527	23.75%	695	0.59%	26,222	11.63%
Grand Total	**107,478**	**100.00%**	**117,970**	**100.00%**	**225,448**	**100.00%**

4.1.5. Top IP Divisions by Time-Group

On this regard tries to identify the knowledge of which department or divisions accessing internet based on time divisions could be higher. It is useful in order to improve which

department on which time group needs more time to access internet .The following table 4.5 shows that debrework building (IT staff) internet divisions is most frequently accessing internet on both "OFF TIME" and "WOKK TIME" time group. In addition to this, CATS insurance building internet divisions are needs more special treatment on "WORK TIME" rather on "OFF TIME". Moreover, Data center IT internet divisions is needed more special treatment on "OFF TIME" as well as "WORK TIME". Finally, those categorized knowledge is useful in order to incur a better bandwidth management so as to implement quality of service.

Table 4.5 IP divisions with both off time and work time accessing frequency

Divisions	**OFF TIME**	**Percentage**	**WORK TIME**	**Percentage**	**Grand Total**	**Percentage**
CATS insurance building internet	9795	9.52%	17128	13.97%	26923	11.94%
Data center IT internet	13889	13.50%	28511	23.26%	42400	18.81%
debrework building users Internet	2387	2.32%	2031	1.66%	4418	1.96%
debrework building (IT staff) internet	51411	49.97%	54731	44.65%	106142	47.08%
Head office internet	16578	16.11%	11969	9.77%	28547	12.66%
kera building internet	7013	6.82%	6844	5.58%	13857	6.15%
saris building internet	1805	1.75%	1356	1.11%	3161	1.40%
Grand Total	**102878**	**100.00%**	**122570**	**100.00%**	**225448**	**100.00%**

4.2 pattern discovery

In this study, an experiment is conducted using weka software to discover patterns about CBE Iron port web usage. The data set used for this experiment is an access log dataset of CBE web appliance which passed through various pre-processing stages and finally converted to .csv format required by the Association rule algorithms in weka. In this specific research both algorithms having similar result and picking FP-Growth for work time dataset and Apriori Algorithm for off time data set is used for pattern discovery.

The end result of weka after applying Apriori algorithm and Fpgrowth is a list of association rules. These rules were analyzed to interpret the interesting rules. As all strong patterns are not necessarily interesting, the researcher first identified which of the strong rules are interesting and which are not. Then the interpretation of the interesting rules is made.

The dataset used for the pattern discovery task has initially 30 attributes and 225,448 records. After conducting several experiments with this dataset, the researcher in consultation of expert noticed that most of the generated rules were not interesting. Due to this, an attempt is made for applying various rearrangements on the dataset like merging child attributes to their respective parent category, transform continuous data to discrete data and remove numeric results; the weka result has been analyzed if interesting rules have come.

Finally an experiment has been conducted on separated on “OFF Time” and “WORK Time” dataset with 11,523 unique records, so as to check whether interesting patterns on both time. The result is discussed in table below.

Table 4.6 Sample, OFF TIME data set used for the experiment

Time	Elapsed-Time	Division	Transaction-Result	Transaction-code	Response-Size	Method	Contact-Server	Frequency	Categorization
WORK TIME	Medium	CATS insurance building internet	TCP_MISS	Successful response	Large	GET	DIRECT	64	Blog
WORK TIME	Medium	kera building internet	TCP_MISS	Client error	Medium	GET	DIRECT	1	Cms
WORK TIME	Medium	saris building internet	TCP_MISS	Successful response	Small	GET	DIRECT	1	Entertainment
WORK TIME	Medium	CATS insurance building internet	TCP_MISS	Successful response	Small	GET	DIRECT	2	Blog
WORK TIME	Medium	Data center IT internet	TCP_MISS	Client error	Small	GET	DIRECT	1	Blog
WORK TIME	Minimum	kera building internet	TCP_DENIED	Client error	Small	GET	NONE	1	Entertainment
WORK TIME	Minimum	saris building internet	TCP_DENIED	Client error	Small	GET	NONE	1	Blog
WORK TIME	Medium	Data center IT internet	TCP_DENIED	Client error	Small	GET	NONE	1	Internet Service
WORK TIME	Medium	debrework building users Internet	TCP_DENIED	Client error	Small	GET	NONE	1	Local
WORK TIME	Maximum	CATS insurance building internet	NONE	Client error	Small	GET	NONE	1	Search Engines
WORK TIME	Medium	kera building internet	TCP_MISS	Client error	Small	GET	DIRECT	4	Social
WORK TIME	Medium	saris building internet	TCP _HIT	Successful response	Large	GET	DIRECT	1	Local
WORK TIME	Medium	Data center IT internet	TCP_MISS	Successful response	Small	GET	DIRECT	69	Search Engines
WORK TIME	Maximum	debrework building users Internet	TCP_MISS	Successful response	Small	CONNECT	DIRECT	16	Social

Table 4.7 Sample, WORK TIME data set used for the experiment

Time	Elapsed-Time	Division	Transaction-Result	Transaction-code	Response-Size	Method	Contact-Server	Frequency	Categorization
OFF TIME	Medium	CATS insurance building internet	TCP_MISS	Successful response	Small	GET	DIRECT	45	Cms
OFF TIME	Medium	kera building internet	TCP_MISS	Client error	Medium	GET	DIRECT	3	Entertainment
OFF TIME	Medium	saris building internet	TCP_MISS	Successful response	Medium	GET	DIRECT	48	Blog
OFF TIME	Medium	CATS insurance building internet	TCP_MISS	Successful response	Medium	GET	NONE	1	Local
OFF TIME	Medium	Data center IT internet	TCP_MISS	Successful response	Medium	GET	DIRECT	30	Entertainment
OFF TIME	Medium	kera building internet	TCP_MISS	Successful response	Medium	GET	DIRECT	1	Blog
OFF TIME	Medium	saris building internet	TCP_DENIED	Client error	Small	CONNECT	NONE	1	Internet Service
OFF TIME	Maximum	Data center IT internet	TCP_ MISS	Client error	Medium	CONNECT	DIRECT	1	Local
OFF TIME	Maximum	debrework building users Internet	TCP_MISS	Successful response	Medium	CONNECT	DIRECT	26	Search Engines
OFF TIME	Medium	CATS insurance building internet	TCP_MISS	Successful response	Small	GET	DIRECT	268	Internet Service
OFF TIME	Maximum	kera building internet	TCP_MISS	Successful response	Medium	GET	DIRECT	3	Local
OFF TIME	Medium	saris building internet	TCP_DENIED	Client error	Small	GET	NONE	361	Search Engines

In each experiment the threshold values for minimum support and confidence are selected by heuristics after conducting several experiments with different values.

Summary of experimental result on work time, off time and aggregate dataset including all attributes is discussed below. For detail Weka Output (see Appendix F).

4.2.1. Pattern Discovery and Analysis using "WORK TIME" Dataset

In order to find the rule from the particular database: If Item a,b is in transaction 100,200,300 and has a support 3. The remaining item-set is having support less than 3 therefore the "most frequent rule" for this database is a→b. Now to calculate the support of rule a→b. Support count for a,b is 3 and there are total 5 transactions, the rule support is 3/5= 0.6. Therefore the support for the rule a→b is 0.6. Let us come to how to calculate the confidence of rule a→b. We can get the rule confidence by dividing the support count of 'ab' and by dividing the support count of 'a', because 'a' appears in transaction 100,200,300,400 and the support count is 4, and the support count for ab is 3. Therefore rule confidence for a→b is 3/4 = 0.75 [34]. In this specific study an attempt is made to extract association rules in WORK TIME dataset, Apriori algorithm is applied with a minimum Support = 0.05 and Minimum Confidence = 0.85 is used. These threshold values are selected because more interesting association rules are generated at these support and confidence level after trying several experiments with plus or minus of these threshold values.

RULE 1: Transaction-Result=TCP_DENIED 1865 → Transaction-code=Client error 1865 <conf:(1)> lift:(2.85) lev:(0.13) [561] conv:(561.91)

This rule states that, while most employees accessing internet services on duty hours and the result obtained from this rule is replied TCP_DENIED, 100% of its Transaction-code receives Client error. To conclude, the appliance responds client error for denying services request.

RULE 2: Division=debrework building(IT staff) internet Response-Size=Medium 1800 → Transaction-code=Successful response 1766 <conf:(0.96)> lift:(1.47) lev:(0.06)

This rule states that, if most debrework building (IT staff) internet users accessing internet resources during work time and their associated response size is medium, the result of this rule indicates 96% successful transaction response.

RULE 3: Elapsed-Time=Medium Response-Size=Medium 1115 → Transaction-code=Successful response 1045 <conf:(0.94)> lift:(1.44) lev:(0.08) [320] conv:(5.5)

This rule states that, while most employees accessing internet services on "WORK TIME" and those employees elapsed medium time on their access and corresponding response page size is medium, their transaction code responds successfully.

4.2.2. Pattern Discovery and Analysis using "OFF TIME" Dataset

To extract association rules in OFF TIME dataset, a minimum Support = 0.05 (5%) and Minimum Confidence = 0.85 (85%) is used. These threshold values are selected because more interesting association rules are generated at these support and confidence level after trying several experiments with plus or minus of these threshold values.

RULE 1: Time=OFF TIME ResponseSize=Medium 1879 → ContactServer=DIRECT 1875 conf:(1) lift:(1.14) lev:(0.06) [225] < conv:(45.92)>

This rule states that, if most employees on off duty hours access some URL and receives medium response size; this implies that Employee's contact server communication is Direct. This result shows most off duty hour Employee's service is not similar that is the reason to access their services directly from the root server it is not cached in the appliance. To conclude, "OFF TIME" employees access pattern trend is different from each other.

RULE 2: ElapsedTime=Maximum 1242 → ContactServer=DIRECT 1221 conf:(0.98) lift:(1.12) lev:(0.03) [130] < conv:(6.9)>

This rule states that, while most employees are accessing internet services on "OFF TIME" and staying maximum time on their access, 98% of this transaction contacting the source server directly. To conclude, the appliance caching service is not usable for those internet employees.

RULE 3: [Response-Size=Large_binarized=1]: 1517 → [Transaction-Result=TCP_MISS_binarized=1]: 1486 <conf:(0.94)> lift:(1.21) lev:(0.02) conv:(3.59)

This rule indicates that, if most employees accessing internet resource on off duty hours, their accessing page response size is large and their corresponding transaction result TCP_MISS, This implies the employees get sources directly from the origin server.

4.2.3. Pattern Discovery and Analysis using "AGGREGATE" Dataset

To extract association rules in AGGREGATE dataset, a minimum Support = 0.05 and Minimum Confidence = 0.85 is used. These threshold values are selected because more interesting association rules are generated at these support and confidence level after trying several experiments with plus or minus of these threshold values.

RULE 1: [Categorization=Social_binarized=1]: 1568 → [Time=OFF TIME_binarized=1]: 1568 <conf:(1)> lift:(2.2) lev:(0.07) conv:(583.58)

This rule indicates that, social URL Category is accessed during off time. The reason behind this rule is Social URLs are protected to access during work time.

RULE 2: TransactionResult=TCP_MISS 5742 → ContactServer=DIRECT 5742 <conf:(1)> lift:(1.2) lev:(0.12) [961] conv:(961.34)

This rule states that while all employees' access some URL and its Transaction Result is replies TCP_MISS, 100% the appliance directly contacts the origin server in order to find the information.

RULE 3: Transaction-Result=TCP_DENIED 1318 → Transaction-code=Client error 1318 <conf:(1)> lift:(3.35) lev:(0.11) [924] conv:(924.36)

This rule indicates that, if most employees accessing internet resource and their corresponding transaction result replied TCP_DENIED, its associated Transaction-code replied client error.

RULE 4: [ResponseSize=Small_binarized=1, TransactionResult=TCP_DENIED_binarized=1]: 1159 → [TransactionCode=Client error _binarized=1]: 1159 <conf:(1)> lift:(2.85) lev:(0.13) conv:(558.02)

This rule states that, if most employees access some URL and receive small response size and its corresponding transaction result is "TCP_DENIED", if and only if the employee's transaction result code is client error. This is because trying to access denying services are protected and can't pass the web security appliance.

RULE 5: [Time=WORK TIME_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: 1311 → [Transaction-code=Client error_binarized=1]: 1281 <conf:(0.99)> lift:(3.35) lev:(0.11) conv:(919.45)

This rule indicates that transaction result is TCP_DENIED if and only if during work time. So it needs special attention.

RULE 6:[Time=OFF TIME_binarized=1, Response-Size=Medium_binarized=1, Categorization=Entertainment_binarized=1]: 1152 → [Transaction-code=Successful response_binarized=1]: 1140 <conf:(0.99)> lift:(1.42) lev:(0.03) conv:(36.35)

This rule states that, if most employees on off duty hours access some URL and receive medium response size and their URL categorization is entertainment, its corresponding transaction code replied successful response.

RULE 7: [Time=WORK TIME_binarized=1, Response-Size=Medium_binarized=1]: 1591 → [Transaction-code=Successful response_binarized=1]: 1520 <conf:(0.96)> lift:(1.36) lev:(0.05) conv:(6.6)

This rule states that most employees on duty hours access some URL and receive medium response size without any consideration of URL category; the result provides that Employee's transaction code reply successful response. This result shows most on duty hour employees accessing internet service medium response size having successful response.

RULE 8: [Time=OFF TIME_binarized=1, Response-Size=Large_binarized=1]: 1527 → [Transaction-code=Successful response_binarized=1]: 1487 <conf:(0.88)> lift:(1.43) lev:(0.02) conv:(157.4)

This rule states that, if most employees on off duty hours access some URL and receive large response size, their corresponding transaction code having successful response.

RULE 9: [Time=WORK TIME_binarized=1, Response-Size=Medium_binarized=1, Categorization=Business_binarized=1]: 1065 → [Transaction-code=Successful response_binarized=1]: 936 <conf:(0.87)> lift:(1.36) lev:(0.02) conv:(6.62)

This rule states that, if most employees on duty hours access some URL and receive medium response size and their URL categorization is business, their associate transaction code result responds successfully.

RULE 10: ResponseSize=Medium 3574 → TransactionCode=Successful response 3450 <conf:(0.97)> lift:(1.38) lev:(0.12) [943] conv:(8.54)

This rule states that, while most employees access some URL and its Response Size is Medium, this kind of request having a successful transaction code response.

RULE 11: Division=debrework building(IT staff) internet Time=OFF TIME 2022 → Transaction_Result=TCP_MISS 1985 <conf:(0.96)> lift:(1.11) lev:(0.02) [160] conv:(1.37)

This rule states that, while most debrework building (IT staff) internet employees access some URL during OFF TIME their transaction result is "TCP_MISS".

RULE 12: Time=OFF TIME ElapsedTime=Maximum 1242 → ContactServer=DIRECT 1191 conf:(0.94) lift:(1.12) lev:(0.03) [130] < conv:(6.9)>

This rule states that, while most employees accessing internet services on "OFF TIME" and staying maximum time on their access is contacting the source server directly. To conclude, the appliance caching service is not usable for those internet access trends.

RULE 13: Time=OFF TIME ResponseSize=Medium 1879 → ContactServer=DIRECT 1915 conf:(0.9) lift:(1.14) lev:(0.06) [225] < conv:(45.92)>

This rule states that, if most employees on off duty hours access some URL and receive medium response size, this kind of Employee's request having direct contact server communication. This result shows most off duty hour employees service is different from each other that is the reason to access their services directly from the root server it is not cached in the appliance. To conclude, "OFF TIME" employees access pattern is different from each other.

RULE 14: Categorization=Entertainment 1428 → ContactServer=DIRECT 1334 conf:(0.88) lift:(1.09) lev:(0.03) [110] < conv:(2.68)>

This rule states that, if most employees access some URL and its URL categorization is entertainment, this kind of service having direct server contact in order to get its entertainment service. To conclude, any time Employee's entertainment access pattern having direct server contact.

RULE 15: Transaction_Result=TCP_MISS Method=CONNECT Categorization=Entertainment 1211 → Time=OFF TIME 1038 <conf:(0.87)> lift:(1.89) lev:(0.05) [390] conv:(5.64)

This rule states that, if most employees access some URL and its Transaction result is "TCP_MISS" and its method is "CONNECT" and its categorization is "Entertainment", this kind of result is received mostly during off duty hours.

RULE 16: Categorization=Business Time=WORK TIME 1256 → Transaction_Result=TCP_MISS 1156 <conf:(0.86)> lift:(1.19) lev:(0.02) [171] conv:(1.85)

This rule states that, if most employees' access some URL categorized to "Business" and accessed on duty hours, its corresponding transaction result is TCP_MISS. This result shows most on duty hour employees browse Business URL and their access trend is volatile that is the reason for not cached on to the firewall appliance.

RULE 17: Categorization=Social 1068 → Transaction_Result =TCP_MISS 914 <conf :(0.86)> lift:(1.22) lev:(0.02) [161] conv:(2.04)

This rule states that, if most employees access some URL and its URL categorization is "Social", its corresponding Transaction_Result is "TCP_MISS". This result shows all "Social" URL categorization is having TCP_MISS transaction result; this implies social URL category is not cached into the firewall appliance because of its nature of uniqueness.

4.3 Summary of the Findings

Experimentation result with both off time and work time aggregate dataset result in strong and interesting Association rules presented below

- *[Time=WORK TIME_binarized=1, Response-Size=Medium_binarized=1]: → [Transaction-code=Successful response_binarized=1]:*
    - This rule states that, if most employees on duty hours access some URL and receive medium response size without any consideration of URL category, the result provides that Employee's transaction code reply successful response.
- *[Time=OFF TIME_binarized=1, Response-Size=Large_binarized=1]: → [Transaction-code=Successful response_binarized=1]:*
    - This rule states that, if most employees on off duty hours access some URL and receive large response size, their corresponding transaction code having successful response.
- *[Time=WORK TIME_binarized=1, Response-Size=Medium_binarized=1, Categorization=Business_binarized=1]: → [Transaction-code=Successful response_binarized=1]:*
    - This rule states that, if most employees on duty hours access some URL and receive medium response size and their URL categorization is business, their associate transaction code result responds successfully.
- *Time=OFF TIME_binarized=1, Response-Size=Medium_binarized=1, Categorization=Entertainment_binarized=1]: → [Transaction-code=Successful response_binarized=1]:*
    - This rule states that, if most employees on off duty hours access some URL and receive medium response size and their URL categorization is entertainment, its corresponding transaction code replied successful response.
- *[Time=WORK TIME_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: → [Transaction-code=Client error_binarized=1]:*
    - This rule states that, if most employees on duty hours access some URL and receive TCP_DENIED transaction result, its corresponding transaction code replied Client error.
- *Division=debrework building(IT staff) internet Time=OFF TIME → Transaction_Result=TCP_MISS*
    - This rule states that while most debrework building (IT staff) internet employees access some URL during OFF TIME their transaction result is

"TCP_MISS". To conclude, Debrework building IT staff internet access trend behaviour is not similar among staff members during off duty hour.

In this study the researcher getting many interesting patterns with statistical measurement and association rule discovery. To select interesting patterns the researcher discussed the generated rule with domain experts and reviewing CBE company policy for identifying indicators and provide its dimensions for restructuring the internet bandwidth allocation for better utilization and redesigning IT policy for keeping employees internet access for a manageable way that leads to design and implement quality of service for the internet access usage of CBE.

The following table shows the customized research results,

Table 4.8 Interesting rules with corresponding indicator and dimensions

Rules from association discovery and statistical result	**Indicator**	**Dimensions**
TCP_DENIED during duty hour	Lack of knowhow or Employees have violated the rule stated on internet access policy has occurred.	CBE needs to create awareness for employees on how to access internet resources during duty or off duty hours.
Transaction Result=TCP_MISS	Cisco Iron port web appliance not cached on its storage.	This appliance configured with default configuration, studying on it or attaching another caching service algorithm may increase its efficiency. Because Employee's services are cached on the appliance and then, employees getting fast response time.
Categorization=Social → Transaction Result =TCP_MISS	All social URL categories are working on off duty hours. This service did not cache to the appliance because of its volatile nature.	CBE may add more bandwidth conception for this service during off time. Because most employees are interested on accessing this category.
Entertainment and social maximum time taking services during off duty hours.	This statistical result shows employees are refreshing on off duty hours.	Employees more times spent on these URLs category, CBE may give special attention to those services for maximizing employee satisfaction.
Internet Services and business maximum time taking services during on duty hours.	Business and Internet service URL category is most employees needed services during working hours.	Adding more bandwidth to those services during work time may add employee's interest.
Debrework building (IT staff) internet users staying more time to internet resources on both off and on duty hours.	While other divisions debrework building IT staff internet users stay more time on internet resource	May those divisions needs special attention. If they are not larger in number from other divisions.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

In this study, employee's internet access trend is discovered by categorizing URLs during on duty hours and off duty hours for the purpose of getting interesting patterns so as to improve efficient bandwidth utilization by applying different techniques and algorithms in different environments. The statistical analysis and the pattern discovery was done based on log usage data which spans over 10 days of august 2015; from those days separated and picking only 10 hours from on duty hour and 10 hours from off duty hours.

This study provides comprehensive information about phases of web mining for intends to concisely describe the process of data extraction from web data and phases which are involved in each step. It is evident from literature that each phase which is involved in web usage mining has its prose and cones. With the growth in web usage and web based applications the significance of web based extraction has gained numerous importance. So this paper intends to provide the description of each phase in order to understand its approaches and methods. In this study Iron port web security appliance log file is used for experiments and also MS-Excel is used for statistical analysis and weka association rule algorithms used for pattern discovery. The data extracted from each of these phases can be used for multiple purposes specially for indicating to design a better bandwidth management so as to improve quality of service.

The major finding result on statistical analysis indicates that Entertainment and Social URL categories are the most frequently accessed internet services on off duty hours, whereas Internet Services and Business URL categories most frequently accessed internet services on duty hours. In addition to this, unlike other building internet user's debrework building (IT staff) users are accessing internet almost 50% of the bandwidth resource covered on both time divisions.

The pattern discovery was conducted for each time division separately and for the aggregate data. Several experiments were conducted by setting different minimum support values with minimum confidence of 85% until better result was obtained. Out of the experiments conducted the discovered association rules interesting patterns shows that employees are trying to access denying service during duty hours this indicates that employees' violate internet access policy or lack of knowhow of accessing internet resources, Lack of caching

algorithm efficiency (Most employee's requests are travel to origin server during both time division) this clearly indicates that the bandwidth is highly congested.

This research has attempted to answer the stated research questions and has achieved the goals. As a result, almost all of the problems stated in the statement of the problem, can be addressed by properly implementing the recommendations for indicating efficient bandwidth utilization.

The strength of this research compared to similar other researches is that, in this research employees are categorized into different time groups based on their access patterns were analyzed separately. This has really shown the differences and similarities that exist among different time groups which help to design and implement efficient bandwidth utilization on which URL category by differentiating time group. The pattern discovery in this research considered association rule discovery with access firewall appliance logs. Future researches need to consider other pattern discovery with data sources from different locations and varying status.

In this research, the more challenging task is pre-processing steps to clean the data and Categorizing URL's by using different tools; because there is no customized standard tool for Categorizing URL's. But an attempt is made for achieving research perspective as a result, with unreserved trial this study meets its goal and contributes a lot to future researches.

The final conclusion in this study is that upon result looking from statistics and association algorithms interesting patterns discovered, results of the experiments make appropriate solution for initiating to design a quality of internet access and efficient bandwidth utilization.

5.2. Recommendations

Based on the findings of this study, the following issues are recommended as a future research direction and Website improvement:

- In this research, tries to extracting employee's internet access pattern by URL category for determining employee's internet access usage pattern and browsing trend on different time group. Other researchers using different measurements studying in depth on employee's internet access usage pattern so as to design and implement quality of internet service.
- In this research, an attempt is made for categorizing URL's by using different online tools; but there is no customized standard tool for Categorizing URL's. Identifying and categorizing URL based on their context is one of another potential research area.

- In this research, the web security appliance proxy Server log is used to discover patterns. Future researches need to consider integrated data obtained from Proxy Server, cookies and router gateways so as to identify users' access and browsing patterns. Also, in this research, web security appliance logs with only URL category are used. Future researches can consider those log files with reputation status to analyze both URL category and their reputation risk patterns should experiment.
- In this research, taking less sample size from those log files due to time consumption and due to lack of high speed processing machine. In order to get more strong interesting patterns increasing sample size is one major concern.
- In this research, focused on the employee's internet access pattern by examining web security appliance log data but not considered ones of the web security appliance services; how does the web security appliance caching service works? Considering these concepts with integrating log files may improve a better implementation of efficient bandwidth utilization so as to incur quality of service.
- It is required to conduct a survey to assess the problems with CBE internet access trend during on duty hour and during off duty hour. The survey could be composed of the major areas which employees internet access behavior is adopt. Once the access trend is known, the bandwidth should be redesigned to separate with different category and giving priority up on the finding result.
- As indicated on the finding Employees are lacking knowhow for accessing internet by governing Company's policy. So training takes on several dimensions is advantageous. For starters, The Company should set up a new-employee initiation program that trains workers to focus on quality issues from their first day on the job. Furthermore, there is also a need of technology awareness for all employees when new things emerge or IT policy reformulation.
- Web security appliance Caching restructuring is crucial in the CBE internet access trend. The cache in each URL categories should arranged in order of their importance for faster access. Customizing caching service or redesign its caching service is suitable for enabling fast internet access.

EFERENCES

[1] A. Sharma. "Web Usage Mining Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data". MS Project, Rochester Institute of Technology, Rochester, NY, USA,2008.

[2] D. Glen, "Should marketers engage and how can it be done effectively", Social Media., Data and Digital Marketing Practice.No. 9, pp. 274–277, 2008.

[3] R. Cooley, B. Mobasher and J. Srivastava "information and Pattern Discovery on the World Wide Web" in International conference on Tools with Artificial Intelligence, IEEE computer society, p. 558-5 1997.

[4] F.Michele ,and P. Luca, "Recent Developments inWeb Usage Mining Research",In 5th International Conference, DaWak 2003, pp 140-150, 2003.

[5] Cisco IronPort AsyncOS 7.1 for Web User Guide, Americas Headquarters Cisco systems Inc, San Jose, 2010.

[6] O.Zaiane, Xin and Han"Multimedia Miner: a system prototype for multimedia data mining" in international conference on video data mining. Proc. ofACMSlGMOD, 1998.

[7] M. Eirinaki, M. Vazirgiannis, "Web Mining for Web Personalization", ACM Transactions on Internet Technology vol.3, no.1, pp.1-27, 2003.

[8] "Commercial Bank of Ethiopia message of the president" http://www.combanketh.et/AboutUs/CorporateMessage.aspx, Aug 28, 2014.

[9] B. Santhosh Kumar and K.V.Rukmani. ―Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms.‖ Int. J. of Advanced Networking and Applications, Volume 1, Issue 6, pp. 400-404, Department of Computer Science, C.S.I. College of Engineering, Ketti- 643 215. The Nilgiris, India, 2010

[10] P.Ferguson,and G. Huston, "*Delivering QoS on the Internet and in Corporate Networks*" John Wiley and Sons, Inc New York, NY, USA 1998.

[11] A. Fesseha. "Web usage pattern discovery: the case of Addis Ababa University official website". MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[12] G.Nigatu. “Web usage pattern discovery and analysis by region: the case of Ethiopian Airlines Official Website”: MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2014.

[14] Z. Baoyao, “ Intelligent Web Usage Mining”, MS thesis ,Division of Information Systems, School of Computer EngineeringNan yang Technological University,2004.

[15] M.Helmyand M. Djalali “Data Pre-processing on Web Server Logs for generalized Association Rules Mining Algorithm”in World Academy of Science, Engineering and Technology, 2008.

[16] Politecnico di Milano, Italy. Lanzi., Federico Michele Facca and Pier Luca,Recent Developments in Web Usage Mining Research.‖ Journal Article, Artificial Intelligence and Robotics Laboratory Dipartimento di Elettronica e Informazione.

[17] El-Yazeed., N. M. Abo. An overview of preprocessing of web log files for web usage mining.‖ Demonstrator at High Institute for Management and Computer, Port Said University, Port Said, Egypt.

[18] Gandhinagar, 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, Volume 2, Issue 12, pp. 419-423, 2012 India.

[19] R.Ahmed. Web Usage Mining‖ Internet: http://maya.cs.depaul.edu/~classes/ect584/papers/srivastava.pdf. [Online] [Cited: December 10, 2014]

[20] H. Jantan, A.Razak Hamdan and Z. Ali Othman. Data Mining Classification Techniques for Human Talent Forecasting.‖ Faculty of Computer and Mathematical Sciences UiTM, Terengganu, 23000 Dungun, Terengganu, Faculty of Information Science and Technology UKM, 43600 Bangi, Selangor, Malaysia.

[21] Ramageri, Bharati.Data mining techniques and applications. 2010.

[22] G. Kaur and S. Aggarwal. Performance Analysis of Association Rule Mining Algorithms.‖ Research Paper, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India, 2013.

[23] S.Langhnoja, M. Barot and D. Mehta. “Pre-Processing: Procedure on Web Log FileforWeb Usage Mining.” ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 2, Issue 12, December 2012(IJETAE).

[24] R. Mishra and A. Choubey. “Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining.” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9, Computer Science & CSVTU, India, 2012.

[25] J.Han, M, Kamber and J.Pei. Data Mining: Concepts and Techniques, Third Edition, Elsevier, pp. 15, 28, 208-221, 225-226, 491-492.

[26] Corporation, Two Crows. Introduction to Data Mining and Knowledge Discovery. s.l. : Two Crows Corporation. ISBN: 1-892095-02-5.

[27] R.Iváncsy, I.Vajk Frequent Pattern Mining in Web Log Data.. 2006.

[28] A.Fesseha, 2011Web Usage: Exploring Navigational Behavior of Users, the Case of the Official website of Addis Ababa University.MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia,.

[29] M. Hailay, Developing dynamic bandwidth allocation prototype model for campus network based on network traffic analysis 2011.

[30] https://developer.mozilla.org/en-US/docs/Web/HTTP/Response_codes. [Online] [Cited: august 15, 2015]

[31] Z. Zheng, Y. Zhang, and M. R. Lyu“Distributed QoS Evaluation for Real-World Web Services” IEEE International Conference on Web Services. 2010.

[32] https://msdn.microsoft.com/en-us/library/ms174512.aspx. [Online] [Cited: November 10, 2015]

[33] M. Kantardzic, J.Wiley & Sons.”Data Mining: concepts, models, methods, and algorithms.” 2003

[34] T. Steinbach, Kumar “Introduction to Data Mining” 2004

[35] Fayyad, U. M. 1996. From data mining to knowledge discovery: an overview. In fayyad, U. M. (Eds.), Advances in knowledge discovery and data mining.AAAI Press/ The MIT Press.

APPENDICES

Appendix A: UNIX epoch converter

If you import data you might encounter time values stored as Unix timestamps. Unix time is defined as the number of seconds since midnight (GMT time) on January 1, 1970 -- also known as the Unix epoch.

For example, here's the Unix timestamp for August 4, 2008 at 10:19:08 pm (GMT):

```
1217888348
```

To create an Excel formula to convert a Unix timestamp to a readable data and time, start by converting the seconds to days. This formula assumes that the Unix timestamp is in cell A1:

```
=(((A1/60)/60)/24)
```

Then, you need to add the result to the date value for January 1, 1970. The modified formula is:

```
=(((A1/60)/60)/24)+DATE(1970,1,1)
```

Finally, you need to adjust the formula for the GMT offset. For example, if you're in Ethiopia the GMT offset is +3. Therefore, the final formula is:

```
=(((A1/60)/60)/24)+DATE(1970,1,1)+(+3/24)
```

A simpler (but much less clear) formula that returns the same result is:

```
=(A1/86400)+25569+(+3/24)
```

Both of these formulas return a date/time serial number, so you need to apply a number format to make it readable as a date and time.

Appendix B:Splitting date and time

As is often the case, your Excel worksheet may contain dates and times in one cell, while you want to split them into two separate cells.

Remembering that in the internal Excel system the date value is stored as a whole part and the time value as a fractional part of a decimal number, you can extract the date using the INT function, which rounds the cell value down to the nearest integer.

Supposing your original dates and times are in column A, the following formula will accomplish the separation:

=INT(A2)
Use the Int function to extract the date values

To extract the time portion, subtract the date returned by the above formula from the original date and time value:

=A2-B2

Where column A contains the original date and time values and column B contains the dates returned by the INT function.

If you'd rather not have time values linked to the separated dates (for example, you may want to remove the date column in the future), you can use the following MOD formula that refers to the original data only:

If condition used to categorize fragmented time in to quantifiable classes.

Use =AND where both (or more) conditions have to be met to result in TRUE. Then the second IF is not required. With:

=IF(AND(F2>TIME(6,0,0),F2<TIME(16,0,0)),"A","B")

Shift A applies to just after 06:00 hrs through to just before 16:00 hrs and Shift B otherwise. IF and F so intermingled in the same formula was a little unfortunate!

Appendix C:

Python Source codes

```
## Reduce the lines by removing random number of lines

#!/usr/bin/python

        import random

        import csv

        import sys

        # create and open csv17.txt file for write purpose

        csv_file = open ('csv17.txt', 'w')

        # open original.txt file for read purpose

        txt_file = open ('original.txt', 'r')

        skipping = random.sample(range(1385927), 1373927)

        for i, line in enumerate(txt_file):

           if i in skipping:

              continue

           csv_file .write(line)

        txt_file.close()

        csv_file.close()

## Remove unwanted detail after parent URL address by checking different separator

        #!/usr/bin/python

        import csv

        import sys

        # create and open file for write purpose

        csv_file = open ('csv6.txt', 'w')

        # open file for read purpose
```

```
txt_file = open ('testdata.txt', 'r')

separator = ' - '

#iterating through reding file and split after separator

for item in txt_file:

   item=item.split(separator,1)[0]

   item=item+'\n'

   str1 = ''.join(item)

   csv_file.write(str1)

txt_file.close()

csv_file.close()
```

Remove empty line (Cariage return)

```
#!/usr/bin/python

import csv

import sys

### create and open csv15.txt file for write purpose

csv_file = open ('csv15.txt', 'w')

### open original.txt file for read purpose

txt_file = open ('original.txt', 'r')

for line in txt_file:

   if line.strip():

      csv_file.write(line.rstrip()+'\n')

txt_file.close()

csv_file.close()
```

Remove double white space

```
#!/usr/bin/python

import csv

import sys

# create and open csv10.txtfile for write purpose

csv_file = open ('csv10.txt', 'w')

# open original.txt file for read purpose

txt_file = open ('original.txt', 'r')

#iterating through reding file and remove double white space

for item in txt_file:

   if '  ' in item:

      item=item.replace('  ', ' ')

   str1 = ''.join(item)

   csv_file.write(str1)

txt_file.close()

csv_file.close()
```

##the simple python file cutter program works with the following

##First the program reading the huge files and put it f1 and then

##reading the first 100 lines from f1 and writes in to the first file and next

##reading the next 100 lines and writes in to the second file and then

##reading the next 100 lines and writes in to the third file and continue until end of file with this sequence

```
#!/usr/bin/python

import io

f1 = open('not_checked.txt','r')

#f2 = open('Output.txt', 'a')

Lines=f1.readlines()

#initializers

i=0

j=0

k=100

for i in range(43):

  with io.open("file_N_C_" + str(i) + ".txt", 'w', encoding='utf-8') as f:

       while j < k:

        f.write(Lines[j])

        j = j+1

       k=k+100

  i = i+1

f1.close()

f.close()
```

Appendix:D

Weka Association Rule Discovery Outputs

Apriori Experiment Result in WORK TIME dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 25 -T 0 -C 0.85 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: all calc 4 worktime

Instances: 6241

Attributes: 6

Elapsed-Time

Division

Transaction-Result

Transaction-code

Response-Size

Categorization

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

Apriori

=======

Minimum support: 0.15 (936 instances)

Minimum metric <confidence>: 0.85

Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 30

Size of set of large itemsets L(3): 22

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

Best rules found:

1. Transaction-Result=TCP_DENIED 1865 → Transaction-code=Client error 1865 <conf:(1)> lift:(2.85) lev:(0.13) [561] conv:(561.91)

2. Transaction-Result=TCP_DENIED Response-Size=Small 1864 → Transaction-code=Client error 1864 <conf:(1)> lift:(2.85) lev:(0.13) [561] conv:(561.26)

3. Transaction-code=Client error Categorization=Internet Services 1677 → Response-Size=Small 1660 <conf:(0.97)> lift:(1.8) lev:(0.07) [293] conv:(17.27)

4. Elapsed-Time=Maximum Transaction-Result=TCP_MISS 1670 → Transaction-code=Successful response 1653 <conf:(0.97)> lift:(1.5) lev:(0.05) [217] conv:(13.04)

5. Division=debrework building(IT staff) internet Response-Size=Medium 1800 → Transaction-code=Successful response 1766 <conf:(0.96)> lift:(1.47) lev:(0.06) [246] conv:(8.01)

6. Response-Size=Medium Categorization=Business 1665 → Transaction-code=Successful response 1636 <conf:(0.96)> lift:(1.47) lev:(0.05) [204] conv:(7.77)

7. Transaction-Result=TCP_MISS Response-Size=Medium 1430 → Transaction-code=Successful response 1353 <conf:(0.95)> lift:(1.46) lev:(0.1) [424] conv:(6.42)

8. Transaction-code=Client error 1486 → Response-Size=Small 1407 <conf:(0.95)> lift:(1.75) lev:(0.14) [603] conv:(8.53)

9. Elapsed-Time=Medium Response-Size=Medium 1115 → Transaction-code=Successful response 1045 <conf:(0.94)> lift:(1.44) lev:(0.08) [320] conv:(5.5)

10 Elapsed-Time=Minimum Response-Size=Small 1134 → Transaction-code=Client error 1051 <conf:(0.93)> lift:(2.66) lev:(0.13) [543] conv:(9.48)

11. Elapsed-Time=Medium Transaction-code=Successful response Response-Size=Small 1753 → Transaction-Result=TCP_MISS 1683 <conf:(0.91)> lift:(1.32) lev:(0.04) [165] conv:(3.32)

12. Elapsed-Time=Maximum Transaction-code=Successful response 1226 → Transaction-Result=TCP_MISS 1133 <conf:(0.9)> lift:(1.31) lev:(0.04) [153] conv:(3.07)

13. Transaction-code=Successful response Response-Size=Small 1287 → Transaction-Result=TCP_MISS 1194 <conf:(0.9)> lift:(1.3) lev:(0.04) [184] conv:(2.95)

14. Elapsed-Time=Maximum 1273 → Transaction-Result=TCP_MISS 1070 <conf:(0.87)> lift:(1.26) lev:(0.03) [138] conv:(2.32)

15. Transaction-Result=TCP_DENIED Transaction-code=Client error Response-Size=Small 1064 → Elapsed-Time=Minimum 940 <conf:(0.87)> lift:(3.63) lev:(0.13) [543] conv:(5.71)

16. Elapsed-Time=Minimum Transaction-code=Client error Response-Size=Small 1271 → Transaction-Result=TCP_DENIED 1050 <conf:(0.86)> lift:(4.22) lev:(0.13) [572] conv:(5.68)

17. Transaction-code=Successful response Categorization=Business 1098 → Transaction-Result=TCP_MISS 951 <conf:(0.86)> lift:(1.25) lev:(0.04) [183] conv:(2.24)

18. Transaction-Result=TCP_MISS Transaction-code=Successful response Response-Size=Small 1094 →Elapsed-Time=Medium 983 <conf:(0.86)> lift:(1.49) lev:(0.05) [223] conv:(2.99)

19. Elapsed-Time=Minimum 1015 → Transaction-code=Client error 873 <conf:(0.86)> lift:(2.45) lev:(0.12) [517] conv:(4.61)

20. Elapsed-Time=Minimum Transaction-code=Client error 1073 → Transaction-Result=TCP_DENIED 990 <conf:(0.86)> lift:(4.21) lev:(0.13) [571] conv:(5.6)

21. Elapsed-Time=Minimum Transaction-code=Client error 1273 → Transaction-Result=TCP_DENIED Response-Size=Small 1140 <conf:(0.86)> lift:(4.22) lev:(0.13) [572] conv:(5.61)

22. Elapsed-Time=Minimum 1115 → Transaction-code=Client error Response-Size=Small 971 <conf:(0.86)> lift:(2.59) lev:(0.13) [534] conv:(4.68)

23. Transaction-code=Successful response 2755 → Transaction-Result=TCP_MISS 2362 <conf:(0.86)> lift:(1.25) lev:(0.11) [468] conv:(2.19)

24. Elapsed-Time=Medium Transaction-code=Successful response 1887 → Transaction-Result=TCP_MISS 1612 <conf:(0.85)> lift:(1.24) lev:(0.07) [314] conv:(2.14)

25. Elapsed-Time=Medium Division=debrework building(IT staff) internet Transaction-code=Successful response 970 → Transaction-Result=TCP_MISS 943 <conf:(0.85)> lift:(1.24) lev:(0.03) [145] conv:(2.13)

Apriori Experiment Result in OFF TIME dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.85 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: all calc 4 worktime

Instances: 5282

Attributes: 6

Elapsed-Time

Division

Transaction-Result

Transaction-code

Response-Size

Categorization

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

Apriori off time

=======

Minimum support: 0.15 (792 instances)

Minimum metric <conviction>: 1.1

Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 1

Best rules found:

1. ResponseSize=Midium 1879 → ContactServer=DIRECT 1875 conf:(1) lift:(1.14) lev:(0.06) [225] < conv:(45.92)>

2. Time=OFF TIME ElapsedTime=Maximum 1242 → ContactServer=DIRECT 1221 conf:(0.98) lift:(1.12) lev:(0.03) [130] < conv:(6.9)>

3. Time=OFF TIME ResponseSize=Midium 1879 → ContactServer=DIRECT 1875 conf:(1) lift:(1.14) lev:(0.06) [225] < conv:(45.92)>

4. Transaction_Result=TCP_MISS Method=CONNECT Categorization=Entertainment 1911 → Time=OFF TIME 1848 <conf:(0.91)> lift:(1.89) lev:(0.05) [390] conv:(5.64)

5. Division=debrework building(IT staff) internet 2022 → ContactServer=DIRECT 1844 conf:(0.91) lift:(1.04) lev:(0.02) [69] < conv:(1.38)>

6. Division=debrework building(IT staff) internet 2022 → Time=OFF TIME ContactServer=DIRECT 1844 conf:(0.91) lift:(1.04) lev:(0.02) [69] < conv:(1.38)>

7. Time=OFF TIME Division=debrework building(IT staff) internet 2022 → ContactServer=DIRECT 1844 conf:(0.91) lift:(1.04) lev:(0.02) [69] < conv:(1.38)>

9. Division=debrework building(IT staff) internet Time=OFF TIME 2022 → Transaction_Result=TCP_MISS 1585 <conf:(0.78)> lift:(1.11) lev:(0.02) [160] conv:(1.37)

9. Method=GET 2332 → ContactServer=DIRECT 2123 conf:(0.91) lift:(1.04) lev:(0.02) [75] < conv:(1.36)>

10.Method=GET 2332 → Time=OFF TIME ContactServer=DIRECT 2123 conf:(0.91) lift:(1.04) lev:(0.02) [75] < conv:(1.36)>

11.Time=OFF TIME Method=GET 2332 → ContactServer=DIRECT 2123 conf:(0.91) lift:(1.04) lev:(0.02) [75] < conv:(1.36)>

12.Division=debrework building(IT staff) internet ContactServer=DIRECT 1844 → Method=GET 1194 conf:(0.65) lift:(1.09) lev:(0.02) [94] < conv:(1.14)>

13.Division=debrework building(IT staff) internet ContactServer=DIRECT 1844 → Time=OFF TIME Method=GET 1194 conf:(0.65) lift:(1.09) lev:(0.02) [94] < conv:(1.14)>

Fpgrowth Experiment Result in OFF TIME dataset

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.85 -D 0.05 -U 1.0 -M 0.1

Relation: all calc 4 offtime-weka.filters.unsupervised.attribute.Reorder-R1,2,3,5,6,4-weka.filters.unsupervised.attribute.NominalToBinary-Rfirst-last-weka.filters.unsupervised.attribute.NumericToBinary-Rfirst-last

Instances: 5282

Attributes: 26

Elapsed-Time=Medium_binarized

Elapsed-Time=Maximum_binarized

Elapsed-Time=Minimum_binarized

Division=CATS insurance building internet _binarized

Division=Data center IT internet_binarized

Division=debrework building users Internet_binarized

Division=debrework building(IT staff) internet_binarized

Division=Head office internet_binarized

Division=kera building internet_binarized

Division=saris building internet_binarized

Transaction-Result=TCP_MISS_binarized

Transaction-Result=TCP_DENIED_binarized

Transaction-Result=TCP_HIT_binarized

Transaction-Result=NONE_binarized

Response-Size=Small_binarized

Response-Size=Medium_binarized

Response-Size=Large_binarized

Categorization=Blog_binarized

Categorization=Business_binarized

Categorization=Cms_binarized

Categorization=Entertainment_binarized

Categorization=Internet Services_binarized

Categorization=Local_binarized

Categorization=Search Engines_binarized

Categorization=Social_binarized

Transaction-code=Client error_binarized

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

FPGrowth found 21 rules (displaying top 10)

1. [Response-Size=Medium_binarized=1, Categorization=Entertainment_binarized=1, Elapsed-Time=Maximum_binarized=1]: 1577 → [Transaction-Result=TCP_MISS_binarized=1]: 1543 <conf:(0.94)> lift:(1.21) lev:(0.02) conv:(3.67)

2. [Response-Size=Large_binarized=1]: 1517 → [Transaction-Result=TCP_MISS_binarized=1]: 1486 <conf:(0.94)> lift:(1.21) lev:(0.02) conv:(3.59)

3. [Response-Size=Medium_binarized=1, Elapsed-Time=Maximum_binarized=1]: 1457 → [Transaction-Result=TCP_MISS_binarized=1]: 1409 <conf:(0.94)> lift:(1.2) lev:(0.03) conv:(3.44)

4. [Categorization=Entertainment_binarized=1, Elapsed-Time=Maximum_binarized=1]: 1392 → [Transaction-Result=TCP_MISS_binarized=1]: 1339 <conf:(0.93)> lift:(1.2) lev:(0.03) conv:(3.26)

5. [Division=debrework building(IT staff) internet_binarized=1, Elapsed-Time=Maximum_binarized=1]: 1260 → [Transaction-Result=TCP_MISS_binarized=1]: 1210 <conf:(0.92)> lift:(1.19) lev:(0.02) conv:(2.88)

6. [Division=debrework building(IT staff) internet_binarized=1, Categorization=Social_binarized=1]: 1244 → [Transaction-Result=TCP_MISS_binarized=1]: 1139 <conf:(0.92)> lift:(1.18) lev:(0.02) conv:(2.63)

7. [Elapsed-Time=Maximum_binarized=1]: 1242 → [Transaction-Result=TCP_MISS_binarized=1]: 1131 <conf:(0.91)> lift:(1.17) lev:(0.04) conv:(2.47)

8. [Elapsed-Time=Minimum_binarized=1]: 1101 → [Response-Size=Small_binarized=1]: 1056 <conf:(0.91)> lift:(2.35) lev:(0.07) conv:(6.67)

9. [Elapsed-Time=Medium_binarized=1, Transaction-code=Client error_binarized=1]: 1049 → [Response-Size=Small_binarized=1]: 987 <conf:(0.91)> lift:(2.34) lev:(0.06) conv:(6.4)

10.[Response-Size=Medium_binarized=1, Categorization=Entertainment_binarized=1]: 952 → [Transaction-Result=TCP_MISS_binarized=1]: 896 <conf:(0.89)> lift:(1.14) lev:(0.02) conv:(1.95)

Apriori Experiment Result in AGGREGATE dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 15 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: all calc 4 1

Instances: 11523

Attributes: 7

Time

Elapsed-Time

Division

Transaction-Result

Transaction-code

Response-Size

Categorization

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

Apriori

=======

Minimum support: 0.15 (1728 instances)

Minimum metric <confidence>: 0.85

Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 46

Size of set of large itemsets L(3): 38

Size of set of large itemsets L(4): 9

Best rules found:

1. TransactionResult=TCP_MISS 5742 → ContactServer=DIRECT 5742 <conf:(1)> lift:(1.2) lev:(0.12) [961] conv:(961.34)

2. Transaction-Result=TCP_DENIED 1318 → Transaction-code=Client error 1318 <conf:(1)> lift:(3.35) lev:(0.11) [924] conv:(924.36)

3. Transaction-Result=TCP_DENIED Response-Size=Small 1316 → Transaction-code=Client error 1316 <conf:(1)> lift:(3.35) lev:(0.11) [922] conv:(922.96)

4. ResponseSize=Medium 3574 → TransactionCode=Successful response 3450 <conf:(0.97)> lift:(1.38) lev:(0.12) [943] conv:(8.54)

5. Division=debrework building(IT staff) internet Time=OFF TIME 2022 → Transaction_Result=TCP_MISS 1985 <conf:(0.96)> lift:(1.11) lev:(0.02) [160] conv:(1.37)

6. Time=OFF TIME ElapsedTime=Maximum 1242 → ContactServer=DIRECT 1191 conf:(0.94) lift:(1.12) lev:(0.03) [130] < conv:(6.9)>

7. Time=OFF TIME ResponseSize=Midium 1879 → ContactServer=DIRECT 1915 conf:(0.9) lift:(1.14) lev:(0.06) [225] < conv:(45.92)>

8. Categorization=Entertainment 1428 → ContactServer=DIRECT 1334 conf:(0.88) lift:(1.09) lev:(0.03) [110] < conv:(2.68)>

9. Transaction_Result=TCP_MISS Method=CONNECT Categorization=Entertainment 1211 → Time=OFF TIME 1038 <conf:(0.87)> lift:(1.89) lev:(0.05) [390] conv:(5.64)

10.Categorization=Business Time=WORK TIME 1256 → Transaction_Result=TCP_MISS 1156 <conf:(0.86)> lift:(1.19) lev:(0.02) [171] conv:(1.85)

11.Categorization=Social 1068 → Transaction_Result =TCP_MISS 914 <conf :(0.86)> lift:(1.22) lev:(0.02) [161] conv:(2.04)

12.Elapsed-Time=Minimum 1516 → Time=WORK TIME 1343 <conf:(0.89)> lift:(1.62) lev:(0.06) [514] conv:(3.95)

13.Time=OFF TIME Division=debrework building(IT staff) internet 1978 → Transaction-Result=TCP_MISS 1749 <conf:(0.88)> lift:(1.21) lev:(0.04) [303] conv:(2.32)

14.Division=debrework building(IT staff) internet Transaction-code=Successful response 2860 → Transaction-Result=TCP_MISS 2516 <conf:(0.88)> lift:(1.2) lev:(0.05) [426] conv:(2.23)

15.Time=OFF TIME Transaction-code=Successful response 3145 → Transaction-Result=TCP_MISS 2762 <conf:(0.88)> lift:(1.2) lev:(0.06) [464] conv:(2.21)

Fpgrowth Experiment Result in AGGREGATE dataset

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.05

Relation: all calc 4 1-weka.filters.unsupervised.attribute.NominalToBinary-A-Rfirst-last-weka.filters.unsupervised.attribute.NumericToBinary-Rfirst-last-weka.filters.unsupervised.attribute.Reorder-R2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,1

Instances: 11523

Attributes: 28

Time=WORK TIME_binarized

Elapsed-Time=Medium_binarized

Elapsed-Time=Minimum_binarized

Elapsed-Time=Maximum_binarized

Division=CATS insurance building internet _binarized

Division=Data center IT internet_binarized

Division=debrework building users Internet_binarized

Division=debrework building(IT staff) internet_binarized

Division=Head office internet_binarized

Division=kera building internet_binarized

Division=saris building internet_binarized

Transaction-Result=TCP_MISS_binarized

Transaction-Result=TCP_DENIED_binarized

Transaction-Result=NONE_binarized

Transaction-Result=TCP_HIT_binarized

Transaction-code=Successful response_binarized

Transaction-code=Client error_binarized

Response-Size=Small_binarized

Response-Size=Large_binarized

Response-Size=Medium_binarized

Categorization=Blog_binarized

Categorization=Business_binarized

Categorization=Entertainment_binarized

Categorization=Internet Services_binarized

Categorization=Local_binarized

Categorization=Search Engines_binarized

Categorization=Social_binarized

Time=OFF TIME_binarized

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

FPGrowth found 30 rules (displaying top 15)

1: [Categorization=Social_binarized=1]: 1568 → [Time=OFF TIME_binarized=1]: 1568 <conf:(1)> lift:(2.2) lev:(0.07) conv:(583.58)

2. [Transaction-Result=TCP_DENIED_binarized=1]: 1318 → [Transaction-code=Client error_binarized=1]: 1318 <conf:(1)> lift:(3.35) lev:(0.11) conv:(924.36)

3. [Transaction-Result=TCP_MISS_binarized=1, Categorization=Social_binarized=1]: 1215 → [Time=OFF TIME_binarized=1]: 1215 <conf:(1)> lift:(2.2) lev:(0.06) conv:(499.98)

4: [ResponseSize=Small_binarized=1, TransactionResult=TCP_DENIED_binarized=1]: 1159 → [TransactionCode=Client error _binarized=1]: 1159 <conf:(1)> lift:(2.85) lev:(0.13) conv:(558.02)

5: [Time=WORK TIME_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: 1311 → [Transaction-code=Client error_binarized=1]: 1281 <conf:(0.99)> lift:(3.35) lev:(0.11) conv:(919.45)

6: [Time=OFF TIME_binarized=1, Response-Size=Medium_binarized=1, Categorization=Entertainment_binarized=1]: 1152 → [Transaction-code=Successful response_binarized=1]: 1140 <conf:(0.99)> lift:(1.42) lev:(0.03) conv:(36.35)

7: [Time=WORK TIME_binarized=1, Response-Size=Medium_binarized=1]: 1591 → [Transaction-code=Successful response_binarized=1]: 1520 <conf:(0.96)> lift:(1.36) lev:(0.05) conv:(6.6)

8. [Categorization=Internet Services_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: 1032 → [Transaction-code=Client error_binarized=1]: 1012 <conf:(0.94)> lift:(3.35) lev:(0.07) conv:(583.51)

9. [Response-Size=Small_binarized=1, Categorization=Internet Services_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: 1831 → [Transaction-code=Client error_binarized=1]: 1811 <conf:(0.93)> lift:(3.35) lev:(0.07) conv:(582.81)

10.[Time=WORK TIME_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: 1311 → [Response-Size=Small_binarized=1]: 1280 <conf:(0.9)> lift:(2.14) lev:(0.09) conv:(232.78)

11.[Elapsed-Time=Minimum_binarized=1, Transaction-Result=TCP_DENIED_binarized=1]: 1599 → [Time=WORK TIME_binarized=1]: 1565 <conf:(0.89)> lift:(1.82) lev:(0.06) conv:(99.7)

12:[Time=OFF TIME_binarized=1, Response-Size=Large_binarized=1]: 1527 → [Transaction-code=Successful response_binarized=1]: 1487 <conf:(0.88)> lift:(1.43) lev:(0.02) conv:(157.4)

13:[Time=WORK TIME_binarized=1, Response-Size=Medium_binarized=1, Categorization=Business_binarized=1]: 1065 → [Transaction-code=Successful response_binarized=1]: 936 <conf:(0.87)> lift:(1.36) lev:(0.02) conv:(6.62)

14.[Time=WORK TIME_binarized=1, Transaction-code=Client error_binarized=1, Categorization=Internet Services_binarized=1]: 1073 → [Response-Size=Small_binarized=1]: 935 <conf:(0.86)> lift:(2.1) lev:(0.06) conv:(27.28)

15.[Division=debrework building(IT staff) internet_binarized=1, Response-Size=Medium_binarized=1]: 1057 → [Transaction-code=Successful response_binarized=1]: 900 <conf:(0.85)> lift:(1.38) lev:(0.06) conv:(9.05)

DECLARATION

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

_____

Date

Yohannes Mesfin

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

_____

Million Meshesha (PhD)

Advisor