ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

EXPLORING USERS NAVIGATIONAL BEHAVIOR USING WEB USAGE MINING: THE CASE OF ETHIOPIA COMMODITY EXCHANGE OFFICIAL WEBSITE

GASHAW BEKELE KABTIMER

JUNE, 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

EXPLORING USERS NAVIGATIONAL BEHAVIOR USING WEB USAGE MINING: THE CASE OF ETHIOPIA COMMODITY EXCHANGE OFFICIAL WEBSITE

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

GASHAW BEKELE KABTIMER

JUNE, 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

EXPLORING USERS NAVIGATIONAL BEHAVIORUSING WEB USAGE MINING: THE CASE OF ETHIOPIA COMMODITY EXCHANGE OFFICIAL WEBSITE

BY

GASHAW BEKELE KABTIMER

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chair person,	_____	_____
_____	Advisor,	_____	_____
_____	Examiner,	_____	_____
_____	Examiner,	_____	_____

DEDICATION

This research work is dedicated to my nephew Hunyalew Shume.

TABLE OF CONTENTS

ACKNOWLEDGEMENT…… …IV
LIST OF TABLES……..V
LIST OF FIGURES……...VI
LIST OF APPENDICES……..VII
LIST OF AKRONYMS…….VII
ABSTRACT……..IX
CHAPTER ONE 1
INTRODUCTION 1
1.1 Background 1
1.2 Overview of Ethiopia Commodity Exchange 3
1.3 Statement of the Problem 4
1.4 Objective of the Study 7
1.4.1 General Objective 7
1.4.2 Specific Objectives 7
1.5 Scope and Limitation of the Study 8
1.6 Significance of the Study 8
1.7 Research Methodology 9
1.7.1 Web usage Mining Process 9
1.7.2 Data Collection 10
1.7.2 Data Preprocessing 10
1.7.3.1 Data Cleaning 10
1.7.3.2 User and Session Identification 11
1.7.3.3 Transaction Identification 12
1.7.4 Pattern Discovery 12
1.7.3 Pattern Analysis 13
1.8 Organization of the Study 14
CHAPTER TWO 15
LITERATURE REVIEW 15

2.1 Data Mining 15
2.2 Web Mining 15
2.2 Taxonomy of Web Mining 16
2.3 Source of Web Log Data 17
2.3.1 Types of Web log File 18
2.3.1 Types of Web log file formats 18
2.4 Preprocessing 19
2.4.1 Data Cleaning 19
2.4.2 User Identification 20
2.4.3 Session Identification 20
2.4.4 Path Completion 20
2.4.5 Transaction Identification 20
2.5 Pattern Discovery 21
2.5.1 Statistical Analysis 21
2.5.2 Data Mining techniques 23
2.5.3 Association rule mining algorithm 24
2.6 Pattern Analysis 25
2.7 Challenges of Web Usage mining 25
2.8 Application of Web Usage mining 26
2.9 Related Works 28
CHAPTER THREE 31
METHODS AND ALGORITHM 31
3.1 Architecture of the system 31
3.2 Statistical Analysis 32
3.3 Association Rule Mining 32
3.3.1 Apriori Algorithm 34
3.3.2 FPGrowth Algorithm 35
CHAPTER FOUR 36
DATA PREPARATION 39

4.1 Data Collection 39
4.2 Data Preprocessing 41
4.2.1 Data Cleaning 41
4.2.2 Session Identification 44
4.2.4 Transaction Identification 45
CHAPTER FIVE 49
EXPERIMENTATION 49
5.1 Statistical Analysis 49
5.2 Association Rule Discovery and Analysis 50
5.2.1 Experimental setup 60
5.2.2 Pattern Discover and Analysis with all country dataset 61
5.2.2.1 Apriori Algorithm Experiment 61
5.2.2.2 FP-Growth Algorithm Experiment 62
5.2.3 Pattern Discover and Analysis with Africa dataset 63
5.2.4 Pattern Discover and Analysis with Asia dataset 64
5.2.5 Pattern Discover and Analysis with Europe dataset 65
5.2.6 Pattern Discover and Analysis with North America dataset 66
5.7 Discussion and Interpretation 67
5.8 Strength of the study 69
CHAPTER SIX 70
CONCLUSION AND RECOMMENDATION 70
6.1 Conclusion 70
6.2 Recommendation 73
Appendices 79

ACKNOWLEDGEMENT

First of all, I would like to thanks God almighty who has been giving me everything to accomplish this thesis: patience, health, wisdom.

Next, I would like to express my deepest gratitude to my advisor, Dr. Million Masresha, for his excellent guidance, thoughtful, patience, and providing me with an excellent atmosphere for doing research. I have been strongly impressed by his constructive comments and guidance.

Then, I would like to express my sincere gratitude to staffs of Ethiopia Commodity Exchange Iyob Tilahun, Manager of Data Warehousing and Business Application, Yonas Tizazu, Network Security Expert, Anteneh Degef, Corporate Communication Specialist for their cooperation in data collection and data analysis of the research work.

Finally, I would like to extend my thanks to all of my family members. They were always supporting me and encouraging me with their best wishes.

LIST OF TABLES

Table 4.1 Description of IIS log file description----------40

Table 4.2 Summary of preprocessed daily log data----------43

Table 4.3 Summary of user and session identification report----------44

Table 4.4 Sample session record attribute----------45

Table 4.5 Sample page/file frequently accessed by the visitor----------45

Table 4.6 Sample session created on root page of the website----------46

Table 4.7 Sample session created on historicalprice page of the website----------46

Table 4.8 Sample identified transaction data----------47

Table 4.9 sample transformed dataset----------48

LIST OF FIGURES

Figure 1.1 High level Web usage mining process model----------9
Figure 2.1 Taxonomy of web mining----------16
Figure 3.1 Architecture of web usage mining process----------31
Figure 3.2 Algorithm for Apriori----------36
Figure 3.3 Algorithm for FP-Growth----------38
Figure 4.1 Sample web log file----------39
Figure 4.1 Data cleaning----------42
Figure 5.1 most frequently accessed page/file report----------50
Figure 5.2 Summary report of top entry page----------51
Figure 5.3 Summary report of top exit page----------51
Figure 5.4 Summary of top navigational path of users----------52
Figure 5.5 Summary of top visitor country report----------53
Figure 5.6 Summary top visitor city report----------53
Figure 5.7 Summary top coffee product visitor countries report----------54
Figure 5.8 Summary top Seasame product visitor countries report----------55
Figure 5.8 Summary top White pea product visitor countries report----------55
Figure 5.10 summary top Wheat product visitor countries report----------56
Figure 5.11 Summary top Maize product visitor countries report----------56
Figure 5.12 Summery of frequent error type----------57
Figure 5.13 pages that case for frequent partial content error type----------57
Figure 5.14 pages that case for internal server error type----------58
Figure 5.15 Summary of frequently used operating system----------58
Figure 5.17 Summary of visitors hit by hours of day report----------59

LIST OF APPENDICES

Appendix A: list of selected attribute for association rule discovery----------79

Appendix B: Sample transaction identification MSQL statement----------80

Appendix C: Weka Association rule discovery outputs----------81

LIST OF ACRONYMS

CRM	Customer Relation Management
CLF	Common Log Format
CSV	Comma Separated Values
ECLS	Extended Common Log Format
ECX	Ethiopia Commodity Exchange
FPG	Frequent Pattern Growth
FTP	File Transfer Protocol
HTTP	Hyper Text Transfer Protocol
ID	Identification
IIS	Internet Information Services
IP	Internet Protocol
KDD	Knowledge Discovery of Databases
URL	Uniform Resource Location
Weka	Waikato Environment for Knowledge Analysis
WUM	Web Usage Mining
WWW	World Wide Web

ABSTRACT

The use of websites for distributing information is one of the most common media in today's global market. It changes the business environment and has dramatic impact to build and maintain customer relationships through online activities to facilitate the exchange of ideas, products, and services that satisfy the goals of the organization. ECX is a business organization that works to revolutionize Ethiopia's tradition bound agriculture to the global market. It has an extensive web site which provides users with access to services, products, and to advertise their organization. Despite the website's potential of providing the information to the visitor, however, the website of ECX does not guarantee to being delivered and understood correctly. As a result, exploring user navigation behavior is expected to redesign the website based on user requirement and experience.

Web Usage Mining is the process of applying statistical analysis and data mining techniques to discover interesting usage navigation patterns of website. To explore usage patterns of the ECX official website the researcher followed Web usage mining process such as data collection, data preprocessing, pattern discovery and pattern analysis. The web server access log prepared by using log file viewer tool to clean irrelevant record from the log data, user and session identified by using Web log storming tool, preprocessed log record converted into the form appropriate for pattern discovery tool by using MYSQL statements.

After preprocessing of log file experiments conducted using statistical analysis with web log storming and association rule mining with Apriori and FPGrowth algorithm. The result of statistical analysis shows that half of the user of ECX website starts navigation at the root page (www.ecx.com.et/), others are directly access the page they want to visit with the help of search engine. More than two third (2/3) of visitors of ECX website exit from entry page without visiting the other page, others navigate other page before exiting from the entry page. Some visitors visit the same page or link frequently within the same session. Country wise visitor's analysis shows that Iceland, Ethiopia, United States, China, and United Kingdom are the most frequent visitor countries. Most frequent visitor countries looks for coffee, white pea, maize and seasame. The most frequent page that case to error response is root page, which display partial content of the page (error type 206). The path analysis from statistical analysis and association rule mining shows that most of the pages of ECX website are not accessed together accept root and home page of the website. The major challenges that involved in this study are preprocessing of log file due to its large, noisy, and complex nature of log record, and identifying rules and patterns that are potentially interesting. Finally recommendation were done for decision makers, web designers, and further researchers to improve the website.

CHAPTER ONE

INTRODUCTION

1.1 Background

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge [1]. It has been adopted by the mass market more quickly than any other technology over the past century. Then, it was becoming very popular by the community. In the area of business websites are used to bring together consumers and merchants all over the world to a virtual marketplace where customization, direct marketing, market segmentation and customer relationship management can take place [2].

As a result organizations create their own websites to offer and advertise their services to the customers and customers visit these websites to perform any kind of business transactions. As more businesses moved online, the competition between businesses to keep their old customers and attract new customers has increased, since a competitor's website is just one click away. However, as only an effective website can fulfill what it intends to achieve, marketers and web designers find it necessary to understand the effectiveness of their sites and to take appropriate action [3].

On the other hand organizations want to know who their customers are, and how they react to their websites. In this new marketplace, most marketers find their customer behaviors and need. Because understanding the customers will help to develop and distribute the product, service as well as getting the right price point and developing successful promotional activities [4]. Utilizing the opportunities provided by online customer analytics also helps to promote right products in real time to the right customer. Generally, knowledge about the user profile and usage pattern is fundamental for the improvement of effective Web services [5]. All these reasons have motivated organizations to analyze the usage activity data that results from online transactions in order to better understand and satisfy their website users on an individual or group basis, for instance to support Customer Relationship Management (CRM) [4].

However, millions of clicks and online transactions take place every day, and the data that results from these transactions is becoming so huge that trying to use conventional analysis methods is neither possible nor cost-effective. As a result, it has become imperative to use effective methods to turn this raw data into knowledge that can help organizations to better understand their users. Web Usage Mining is the process of applying data mining techniques on web log data (transactions) to extract the most interesting users' online activity patterns that can be used for site improvement, recommendation or personalization services, and hence help increase the online users' satisfaction and maintain their loyalty [5]. The analysis of the discovered usage patterns can also help online organizations gain a variety of business benefits, such as developing cross-product marketing strategies, uncovers real-time e-business opportunities across geography, enhancing promotional operations, and decision making [3].

A common source of this usage pattern is web server logs. Web logs are textual data collected by servers over a period of time, which are enriched with information on users' browsing behavior and website usage statistics. It is a massive repository of users' activities in a website tracked by the web server. It includes entries of document traversal, file retrieval and unsuccessful web events among many others that are organized according to the date and time it was made to the web server. It also contains IP addresses indicating the area, region they belong to and frequency of using the website.

There are different data mining techniques and algorithms specifically tailored to discover knowledge from these usage data such as association rules discovery, sequential pattern discovery, clustering and classification [6]. Association rules discovery is a basic mechanism for extracting frequent patterns, associations and correlations among sets of items. In the web domain, association rules are used to find correlations between pages that are visited by the user during a certain browsing session. Sequential pattern discovery is an extension of association rules finding in such a way that it indicates patterns of co-occurrence with the incorporation of time sequence. Clustering is a method of categorizing/groups items that have similar characteristics. Classification is a process of mapping a data item into one of the pre-organized classes.

In the web domain classes usually represent different user profiles and classification is then applied using selected features which describe each user's category. In this study Web usage mining has been applied to explore usage access patterns of Ethiopia Commodity official Website using statistical analysis and association rule discovery techniques.

1.2 Overview of Ethiopia Commodity Exchange

The Ethiopia Commodity Exchange (ECX) is a new initiative for Ethiopia and the first of its kind in Africa. The vision of ECX is to revolutionize Ethiopia's tradition bound agriculture through creating a new marketplace that serves all market actors such as farmers, traders, processors, exporters and consumers [7]. Its' objective is to connect all buyers and sellers in an efficient, reliable, and transparent market by harnessing innovation and technology and based on continuous learning, fairness, and commitment to excellence. To achieve its objective ECX have different stakeholders such as rural farmers, cooperative unions, traders, exporters, foreign buyers, policy makers and media.

Rural farmers are the producer of the Ethiopian agricultural product. Cooperative unions are small scale farmers to form cooperative unions of around ten to twenty farmers. Traders can be either supplier companies that buy up production from farmers in their region of the country, or individual brokers who act as representatives for cooperative unions; supplier companies. The exporters connect the local market to the international market to meet the requirements of the foreign buyer. Foreign buyers are international buyers of Ethiopian agriculture product. To attract the attention of buyers on the global market the Ethiopian Commodity Exchange needs to be highly transparent and competitive. Policy Makers are political decision makers. It require an understanding of historical, current and future market economics to best be able to promote action that benefits the long term development of the country. ECX can aid in this process by providing a transparent, efficient and stable market with high international reputation. Medias are used to advertise and make transparent market information to the national and international (global) market. ECX Website is one of the best media to advertise Ethiopian agricultural product and to disseminate market information through internet.

The main purposes of the website are to make the exchange accessible to the global market and to provide a comprehensive source of information for people who need to know how to trade on the exchange, what rules and regulations apply to different commodities and operative information. The website also used to advertise Ethiopian agricultural product such as coffee, sesame, wheat, maize, haricot beans to transforms the economy through a dynamic, efficient and transparent marketing system [8].

1.3 Statement of the Problem

Website is one of the most revolutionary technologies that changes the business environment and has dramatic impact on the electronic commerce [9]. It helps to build and maintain customer relationships through online activities to facilitate the exchange of ideas, products, and services that satisfy the goals of the organization. But many ecommerce websites are still too hard to use and fall short on consumer/user expectations [10]. Scholars suggested that website designers should include essential features into any commercial website design such as determining how to design responsive Website infrastructure that provides a sustainable competitive advantage through a better understanding of target customers based on user experience [11]. Because, the Website depends on interrelated factors such as page content, content design and Website design, services it provide, and the unpredictability of e-customer behavior [11].

Web users are rarely on a website to enjoy its design, therefore page design (surface appearance of website) should provide users easy access to the content. Users are also goal-driven and sometimes impatient when reading on the web; they rather scan the content than actually read. So, the page content should be brief, make the text shorter (without losing depth of content), use easily visible and meaningful hyperlinks for more information, use meaningful headlines, and emphasize relevant words. The site design should reflect the users' view of the site and its information and services; providing the user with a logical and intuitive navigation system that reflects the site's structure and hierarchy; the information architecture. Because, the user will at all times want an indication of where they currently are, where they can go next, and where they have been. Many users are more search-dominant than link-dominant, especially in situations when they are looking for something specific.

On the other hand, in contrast to the dynamic growth of the web, most websites are not useful for most of the users [12]. As the researchers claimed, 99% of the websites are not useful for 99% of the users [13]. This is because large number of websites are poorly designed, since user requirements are not often incorporated into the web design process. Websites focus more on quantity rather than content quality that could suit the user's requirement. Thus, the effectiveness of a certain website should be evaluated regularly so as to assure that it fits the need of users [12].

Because everyone having the Web experience knows that how the connection to a popular website may be very slow during rush hours and it is well known that web users tend to leave a site if the wait time for a page to be served exceeds a given value [3]. Therefore, performance and service quality attributes have gained enormous relevance in service design and deployment. As a result, organizations are promoting many initiatives concerning user's profile, in order to implement better sites, more functional and close to customers' needs. Due to this, collecting and mining Web usage data from e-commerce Websites has become increasingly important for marketing, advertising, and decision analysis activities. Therefore, to have effective and efficient Web service, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized web pages to the user by learning user navigational pattern knowledge [9]. The business analyst of e-business organization also wants to know the user (customer) profile for decision making and competitive business strategy of the organization [12].

ECX has an extensive website which provides users with access to services, products, and advertises of their organization. However, according to preliminary investigation, the website is not efficiently organized. The website pages are not well designed. Most of the interface of the website covered by animated image and tables. The page content of the website not easily understandable, for example the root page of the website have more than 15 links of coffee, 8 links of seasame product daily price document, which are not easily by new users of the website. These different links of the website display the same pdf document for each product. For example, important links of the website that required for content personalization such as the types of coffee product that exists with different link have no content of their own.

Users may want to access information about each product price information, however they did not rather than accessing pdf document that contain all coffee product information. This implies that hub pages (contain links to highly important pages) and authoritative pages (highly important pages) not match. It leads user for confusion, because user have expected to access each product with their specific type of product type based on the link. The site structure also not logically organized, for example coffee, seasame and white pea product daily price exist on the home page and wheat and maize product price exist on the other page of the website. The website have not search box that helps user to find preferred navigation links from the website.

As a global competitive organization, ECX seeking to investigate about their Website information source such as the frequent product accessed by the visitor, the demography of customers who access their product and service frequently, the effectiveness and efficiency of the Website to improve their website and decision making process. However, requirements, preferences, behaviors, and demographic traits of a customer were not studied. It has become a concern as to whether the site is being used, how it is being used, does the website achieve the expected objective of designer and whether the site structure is optimal or if it needs to be reorganized. The web developer of ECX try to collect feedback from website users of ECX staff members using questioner to know about the effectiveness and efficiency of the website. However, the expected users of ECX website are dispersed throughout the world. Staff members of ECX are not enough to measure the effectiveness and efficiency of the website and user access patterns. This research therefore, aims to explore user's access pattern of the official website of ECX using web usage mining technique.

To this end, this study attempts to experiment and answer the following research questions.

- ✓ How to prepare appropriate quality dataset with relevant attributes?
- ✓ What are the patterns and association rules that helps to optimize ECX website?
- ✓ What interesting rule are identified that reflecting frequent user access patterns?

1.4 Objective of the Study

1.4.1 General Objective

The general objective of this study is to explore user's access pattern of ECX official website that helps to identify potential user and redesigning of website based on user experience using web usage mining and statistical analysis tools.

1.4.2 Specific Objectives

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

- ✓ To prepare log file for the official website of ECX for web usage mining.
- ✓ To select appropriate Web usage mining tools, techniques and algorithms for association rule discovery.
- ✓ To investigate website usage statistics such as frequently accessed page, visitors and user agents.
- ✓ To discover web usage pattern of ECX website by continent and country.
- ✓ To evaluate interesting patterns of the discovered association rules.

1.5 Scope and Limitation of the Study

The scope of this research is limited to explore the usage patterns of ECX official website using web usage mining techniques. The source of data for this study is Web server log file of ECX official website. Appropriate techniques and algorithms of web usage mining are investigated and selected to discover patterns that describes browsing behavior of users of ECX official website. The log file is prepared to make suitable data for Web usage mining statistical analysis and association rule discovery techniques are employed for pattern discovery. Despite the useful information available in log files, the log data suffer from limitations, creating challenges for use. In this study the limitations of Web log files are raised, because certain types of visitor data are not logged, such as information about the specific product type (i.e. the type of coffee product that accessed by the visitor) visitor record is not logged. Some of the data that are logged are incomplete, such as visit duration, referrer site, user ID, accessed URL. Due to this, the preprocessing task of this research is challenging, because the web log contain noisy, huge and complex records that needed to clean.

User and session identification is also a challenging task, because a single user can use multiple computer and multiple user can use single computer to access a website.

1.6 Significance of the Study

The result of this study provide preliminary data for reconstructing existing websites and for building new user-orientated ones for Ethiopia Commodity Exchange. Furthermore, the result provide fundamental knowledge for developers of websites to identify interesting patterns or rules that helps for decision making. The output of this study also help to provide efficient and effective promotional and marketing strategy by identifying the right place of advertising and interest of customers. Because, as the main objective of ECX to revolutionize Ethiopia's tradition bound agriculture to the global market. And to make accessible the services, products, and advertises of the organization throughout the world. It also provide the user preference and navigational behavior, accessibility and effectiveness of ECX official websites. This will help the ECX official website developer to understand about the performance and user navigational behavior of the website. Because, knowing what the users' need and preference and navigational behavior was becomes a significant step towards improving the website with respect to customer satisfaction and addressing the goal of the organization.

1.7 Research Methodology

Research methodology used to understand what research methods are going to be used, the choice of the study design and a strategy of data collection, organization and analysis. Different literature reviewed from books, journal articles and the internet to identify Web usage mining research techniques, processes and tools.

1.7.1 Web usage Mining Process

In this study high level web usage mining process model suggested by Sharma [14], is used. This Web usage mining process is a four step process such as data collection, data preprocessing, and pattern discovery and pattern analysis. Figure 1.1 shows a generalized approach followed for the web usage mining process.

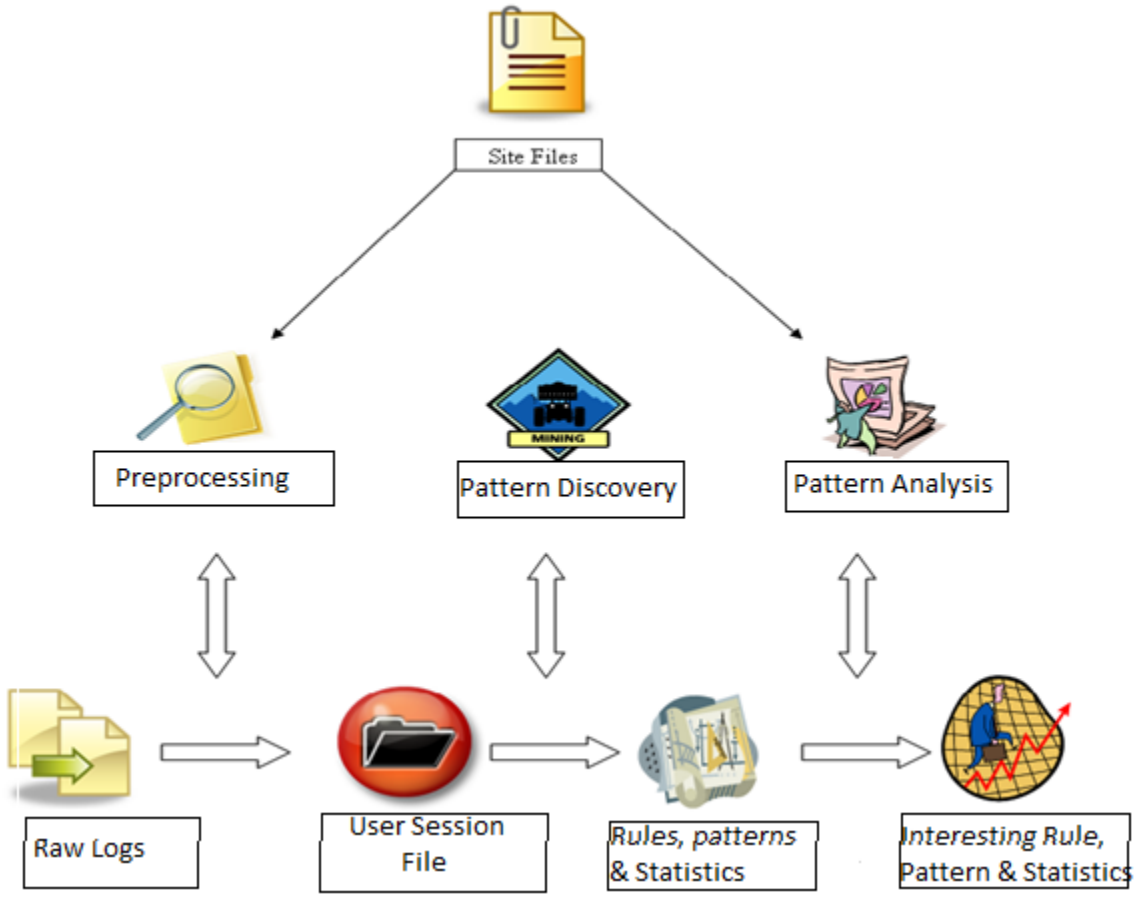

Figure 1.1 High level Web usage mining process model [14]

In the web usage mining process, the techniques of statistical analysis and data mining are applied so as to discover the **t**rends and the patterns in the browsing nature of the visitors of the ECX official website. There is extraction of the navigation patterns as the browsing patterns could be traced. When it is talked about the browsing nature of the user it deals with frequent access of the website. This information can be extracted from the log file. Only these log files record the session information about the web pages.

1.7.2 Data Collection

Data gathered from web servers is placed into special files called logs and can be used for web usage mining. Usually this data is called web log data as all visitors activities are logged into this file [15]. The source of data was 30 days (from January 11, 2015 to February 10, 2015) one month Web server access log file of ECX official website.

The access log file contains records of each user requests that contains remote host IP address (identify who has visited the website), date (access time of the website), visiting path (path taken by the user while visiting the website), path traversed (the path taken by the user within the website), success rate (status code returned by the server), URL (the page that accessed by the user), request type (GET or POST) and user agent (browser type that user use to send request).

1.7.2 Data Preprocessing

The Web server usually registers all users', spiders and bots access activities of the website as Web server logs. The data present in the log file cannot be used as it is for the mining process. Therefore, data preprocessing is mainly one phase in Web usage mining. Generally, several preprocessing tasks need to be done before performing web mining algorithms on the Web server logs [16]. Data processing includes task such as data cleaning, user and session identification, Transaction Identification, and data conversion.

1.7.3.1 Data Cleaning

Data cleaning consists of removing all the data tracked in Web logs that are useless for mining purposes [5]. Since all the log entries are not valid, we need to eliminate the irrelevant entries. Usually, this process removes requests concerning non-analyzed resources such as multimedia files, and page style files. For example, requests for graphical page content (*.jpg & *.gif images) and requests for any other file which might be included into a web page or even navigation sessions performed by robots and web spiders.

By filtering out useless data, we can reduce the log file size to use less storage space and to facilitate upcoming tasks. Thus, data cleaning includes the elimination of irrelevant entries like: requests executed by automated programs, such as web robots, spiders and crawlers; these programs generate the traffic to websites, can dramatically bias the site statistics.

Log file viewer –standard edition tool used to filter relevant records from the log file record by cleaning irrelevant record. A filter is composed of one or more matching conditions with a particular logic. Each condition has a matching pattern (is, is not, contains, excludes, wildcard or regular expression), applied in a selected column or the whole loaded file. MS Excel also use for further data cleaning to remove record that are not cleaned by log viewer tool. It used to remove access of web log record by administrator of the site, staff of the website owner organization and incomplete records. This is usually done by referring to the remote hostname, by referring to the user agent or by checking the access to the robots.txt file.

1.7.3.2 User and Session Identification

It is the most important step in data pre-processing as without identifying the users who access the website, web usage mining cannot be achieved [17]. The task of user and session identification is to find the different user sessions from the available information in the log file. The purpose of user's identification is to identify users those who access website and pages. The need of session identification is to divide the page accesses by each user at a time based on individual sessions.

In this study, various heuristics are employed using the data available in the server log files such as the IP address, user agent and referrer for user identification. The different IP addresses represent different users [18]. If the IP addresses are same, user agents such as, the different browsers type and operating systems show different users. User sessions are used to reconstruct visitor's sessions from Web log records. A common timeout method have been selected to sessionise the filtered log record. According to Spiliopoulou et. al. [19] more than 90% of genuine visit sessions can be identified by the timeout method and a time period of between 20 and 30 minutes is considered optimal.

Web Log Storming 3.0 used for user and session identification process (www. weblogstorming.com). It's easy to track sessions based on the given time period. In this study 20 minutes time period is taken.

1.7.3.3 Transaction Identification

Transaction identification used to prepare data in the format appropriate for the specific data mining algorithm to be used [16]. In this study, transaction are identified based on merge approach, which identify transaction consisting of all of the page reference for a given user session (merging small transactions into larger transaction). Then, by using MYSQL database management tool, total session and session table of each frequent page/file is stored. Finally, by using MYSQL statement the all session data are merged into one, which shows ach page preference of user session.

1.7.4 Pattern Discovery

After data preparations have been completed, the next step is pattern discovery. The pattern discovery phase consists of different techniques derived from various fields such as statistics analysis and association rule discovery are used [5].

Statistical techniques are the most powerful tools in extracting knowledge about visitors to a Website [16]. Web log storming tool used as a tool for statistical analysis technique. It displays detailed website statistics with interactive graphs and reports. It used generate detailed log analysis of activity from every visitor to website. It's easy to track sessions, hits, page views, downloads, or whatever metric is most important to each user. Website behavior, from the top entry and exit pages, to the paths that users follow, can be analyzed. And to identify from which countries and cities visitors came from, and which operating systems and browsers they use. In addition to statistical analysis web log storming used to identify web accessed by spiders, robots and crawlers that are not removed in the data cleaning step.

Association rule mining techniques FPGrowth algorithm used to find association among webpages that frequently appear together in user session and discover correlation between items found in preprocessed web log transaction [16]. It used to find associations among Web pages that frequently appear together in user's session and associations between groups of users with specific interests [20]. In this study Weka 3.7.9 used as a tool for association rule discovery.

1.7.3 Pattern Analysis

Pattern analysis is a final stage of the whole Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. Patterns discovered by statistical and data mining are analyzed in consultation with literature review and domain experts. The output of data mining algorithms also is often not in the form suitable for direct human consumption, and thus need to be transform to a format that can be assimilated easily.

1.8 Organization of the Study

This paper is organized into six chapters. The first chapter provides general introduction about the study such as background, statement of the problem, objective, scope, and methodology of the study. In the second chapter, literature on data mining and web mining particularly process, tools, and techniques of Web usage mining and its application that are necessary to understand the methods and terms that are introduced in the study are reviewed. Finally, a few researches which address problems similar to this study are reviewed. In the third chapter methods and algorithms with the overall architecture of the study presented. Chapter four is about data preprocessing activities such as data cleaning, user and session identification and data conversion process of the study. Chapter five presents experimentation part of this research with different steps that are incorporated in the adopted methodology were described. The web usage patterns were discovered by integrating statistical analysis and association rule mining. Interesting rules and patterns of the experiments 'results identified. In the six chapters, concluding remarks and recommendations were made. Finally, lists of references and appendices were presented at the end of this paper.

CHAPTER TWO
LITERATURE REVIEW

2.1 Data Mining

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories [13]. It is an interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Data Mining has various application areas such as banking, education, and e-commerce. On the other hand, the new data mining applications such as World Wide Web, spatial data, multimedia data.

World Wide Web is one of the largest and most widely known data source. Today, www contains billions of documents edited by millions of people. The total size of the whole documents can be interpreted in many terabytes. World Wide Web is growing at a very large rate in size of the traffic, the amount of the documents and the complexity of websites. Because, different organizations, individuals or societies provide their public information such as news, markets, company advertisements. The World Wide Web serves to a broad diversity of user communities through web. Due to this, the demand for extracting valuable information from this huge amount of data source is increasing every day. This leads to new area called Web Mining [21], which is the application of data mining techniques to World Wide Web.

2.2 Web Mining

According to Etzioni [21], Web mining is the application of data mining techniques to automatically discover and extract useful information or patterns from Web data. It discover pattern or knowledge from semi-structured and unstructured, readily available data, whereas traditional data mining from structured and relational data defined organized in columns, rows, keys, and constraints. Web mining can be used for studying varied aspects of a site can recognize the patterns and relationships in the user behavior so as to get the insight in crucial information [22].

2.2 Taxonomy of Web Mining

The web mining can be generally classified into three different categories such as web content mining, web structure mining and web usage mining [23] as shown in figure 2.2.1.

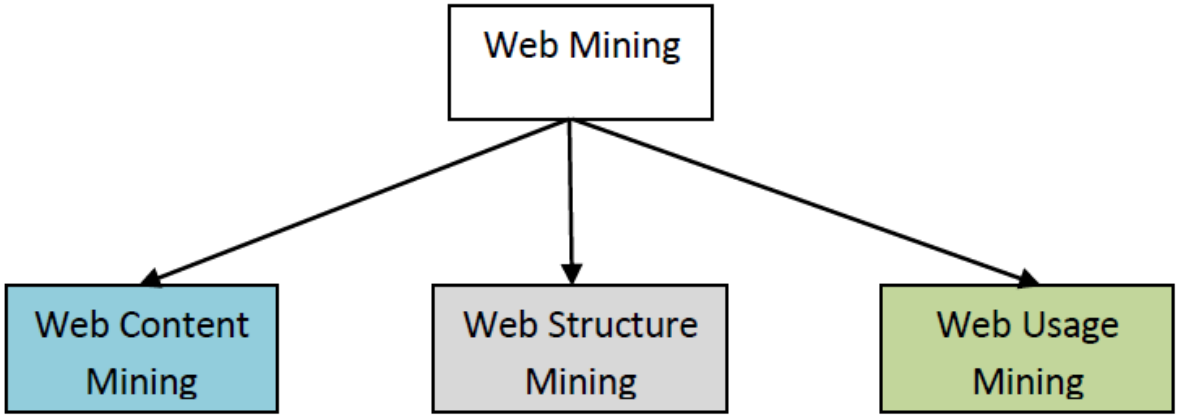

Figure 2.1 Taxonomy of web mining [23]

Web content mining deals with discovery of useful information from unstructured, semi structured or structured contents of web documents. Text, images, audio, video comprised by unstructured document, semi structured data includes HTML documents and lists and tables represent structured documents. The main aim of web content mining is to act as tool to retrieve information easily and quickly. Web Content Mining works by organizing a group of documents into related categories which helps web search engine to extract information more quickly and efficiently.

Web structure mining mines the information by utilizing the link structure of the web documents. It works on inter document level and discovers hyperlink structure. It helps in describing the similarities and relationships between sites. Web structure mining categorizes the pages into authorities and hubs. Authority pages are considered as high quality pages which are related to particular query and hub pages provide pointers to authority pages.

Web usage mining is the process of applying statistical and data mining techniques to extract useful information/usage patterns from the web log data [5]. It provides better understanding for serving the needs of Web-based applications. Web usage mining includes discovering knowledge from Web log files.

When a user accesses a website, its interactions are recorded in the log file. The mining of this kind of data has emerged as a necessity in today's world as users expect to find the relevant information on the web as quickly as possible. Web usage mining defines several procedures leading to the discovery of the desired knowledge. According to Sirivastava [2000], web usage have four main processes. They are preprocessing, pattern discovery, and pattern analysis.

2.3 Source of Web Log Data

The source of the Web data can be either implicit or explicit [24]. Explicit data comes from registration forms which the user fill's while signing up with the website. Additional data such as demographic (i.e. study of the characteristics of visitors) and application data (example: e commerce transactions) can also be used. Implicit data includes past activities/clickstreams as recorded in Web server logs or via cookies or session tracking modules. This helps to study use's behavior at the website. Implicit data can be collected at the server-side, client-side, proxy servers, or obtain from an organization's database, which contains business data or consolidated Web data [5].

Server Level Collection: The most important source is the server log files since they explicitly record the clickstreams for all the users on a particular website and they are much easier to obtain since they are owned by the organization that owns the web servers. Web server collects client requests and stored in the server as web logs. Because, server log explicitly records the browsing behavior of website visitors. These logs contain user name, IP address, date, and time of the request, the request line exactly came from the client. These data can be bound together as a single text file, or divided into different logs, like access log, referrer log, or error log [5].

Proxy Level Collection: A web proxy is basically a server midway between the client and the server, that caches the most recently requested pages in order to reduce the loading time of web pages and reduce the network traffic load. It shares the same concept of caching by web browsers on local machines, however a web proxy is typically shared by multiple users and caches pages from multiple web servers. Hence, a web proxy server can provide accurate and rich information about the common interests of multiple users on multiple websites, which makes it a valuable source for the web usage mining process. On the other hand, local (client-side) caching provides information about only a single user. However, obtaining the data from web proxies is hard and raises more privacy concerns, because web proxies contain the activity of multiple web servers.

Client Level Collection: It is the client itself which sends information to a repository regarding the user's behavior. This is done either with an ad-hoc browsing application or through client side application running standard browsers. Client level data collection can be implemented by using a remote agent (such as Java applets or Java Scripts).

2.3.1 Types of Web log File

Web server log files comprise access logs, referrer logs, agent logs and error logs [25].

Access Logs: provide the bulk of the Web server data, including the date, time, users IP address, and user action. The following is some of the information that can be obtained from an access log: The IP (Internet Protocol) address of the computer making the request for a document the time stamp (user access date and time) the user's request (e.g., html document or image requested, or data posted).

Referrer Logs: provide information on what Web pages, from both the site itself and other sites, contain links to documents stored on the server. The log provides information such as the URLs of sites and pages on sites that referred visitors to a particular page. For example, users may often arrive at a particular Website through a search engine, and the referring search engine along with the keywords used in the originating query, can be obtained from the Referrer log.

Agent Logs: supply data on the browser, browser version, and operating system of the user who accessing the web.

Error Logs: Error Logs contain information on specific events such as file not found, document contains no data; the time, user domain name, and the page on which a user received the error is recorded, providing a server administrator with information on problematic and erroneous links on the server.

2.3.1 Types of Web log file formats

According to Lakshmi et al. [25], there are three kinds of log file formats to record log files. They are: Common Log Format (CLF), Extended Common Log Format (ECLF), and Microsoft IIS (Internet Information Services).

Common Log Format: This log format is supported by a variety of web server applications and includes the following seven fields: Remote host field, Identification field, Authuser field, Date/time field HTTP request, Status code field and Transfer volume field.

Extended Common Log Format, ECLF is a variation of the common log format, formed by appending two additional fields onto the end of the record, the referrer field, and the user agent field. ECLF Log File Format have two additional fields they are referrer and user agent.

Microsoft IIS Log files Format: The IIS format records more fields than the other formats. The Microsoft IIS log format includes the following fields: Client IP address, User name, Date, Time, Service and instance, Server name, Server IP, Elapsed time, Client bytes sent, Server bytes sent, Service status code, Windows status code, Request type, Target of operation, and Parameters

2.4 Preprocessing

Preprocessing is preliminary and essential step in web usage mining. Because, a web usage data (Web log file) is generally diverse, incomplete, inconsistent, noisy and difficult to be used directly for pattern discovery [25]. A Web log file also have different format from the database or data warehouse data which has a good data structured. All this has made the work of pretreatment face many technical problems. Due to this, data preprocessing has become the most difficult task in the Web usage mining. This preprocessing method is used to process the actual web logs before the real usage mining process.

Web Log Data Preprocessing Steps

Web log data pre-processing is a complex process. It can take up to 80% of the time in knowledge discovery process [26]. The aim of data preprocessing is to select essential features, to clean irrelevant records and finally transform raw data into sessions. To achieve its goal Web log data preprocessing have the following steps: data cleaning, user identification, and session identification, path completion and transaction identification [16].

2.4.1 Data Cleaning

The purpose of data cleaning is to remove irrelevant items stored in the log files that may not be useful for analysis purposes. When a user accesses a HTML document, the embedded images, if any, are also automatically downloaded and stored in the server log. For example, log entries with file name suffixes such as gif, jpeg, GIF, JPEG, jpg and JPG can be removed. Since the main objective of data preprocessing is to obtain only the usage data, file requests that the user did not explicitly request needs to be eliminated. This can be done by checking the suffix of the URL name.

In addition to this, erroneous files can be removed by checking the status of the request (such as status code 404). The cleaned log represents the user's accesses to the Website.

2.4.2 User Identification

Identification of users who access a website is an important step in web usage mining. The simplest method is to assign different user ID to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. If the IP address of a user is same as previous entry and user agent is different than the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address.

2.4.3 Session Identification

User session is a delimited set of web pages visited by a particular user in single visit to the website. Identification of user sessions has also received significant attention as it reveals the navigational behavior of users, which forms the foundation of personalization system. A user may have a single or multiple sessions during a particular time period. Various heuristic methods have been used for identifying user sessions. Spiliopoulou [18], divides these methods into time-based and context-based. In time-based approach, a page viewing time is defined, and a single user session consists of all those web pages, which are requested by a particular user within the page viewing time. Context-based approach is not very strict, and depends on the users' perspective. A web page, which is a navigational page (that contain primarily hyperlinks to other web pages and are used just for browsing purpose) for one user might be a content page (that contain the actual information of user's interest) for the other.

2.4.4 Path Completion

Determining the missing important web page access due to the proxy server and the browser is essential for mining information. This is accomplished by path identification process. If the requested page is not linked to the previous accessed page by the unique user, then from which page request came is identified using the referrer log file. If the page is available in the user's history, then it is assumed that the user pressed back button. Hence each and every session reflects the complete path, including the web pages that have been backtracked.

At the end of path completion the user session file gives the paths consisting of a group of page references including repeated page accesses made by a user. At the end of this stage the user session file is ready for transaction identification process.

2.4.5 Transaction Identification

In order to group individual Web page references into meaningful transactions for the discovery of patterns such as association rules, an underlying model of the user's browsing behavior is needed [16]. As a result, further preprocessing step is required, which is transaction identification. Transaction identification used to prepare data in the format appropriate for the specific data mining algorithm to be used. According to Cooley [16], for this process two kinds of transaction are identified, travel path transactions and content only transactions. The travel path is a combination of auxiliary and content pages accessed by a user. Auxiliary pages are pages that used to facilitate the browsing of a user while searching for information. The content only transactions are only content pages which are user interest.

User session in a user session file transaction identified thought either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. Therefore, the task of identifying transactions is one of either merging small transactions into larger transaction or dividing a large transaction into multiple smaller ones or in order to create transactions [16]. Transaction identification approach defined as ether a merge or a divide approach. Merge approach is based on identify transaction consisting of all of the page reference for a given user session. The divide approach such as transaction identification by reference length is depending upon the time taken a user spends on a page correlates to whether the page should be classified as auxiliary or content pages for that user.

2.5 Pattern Discovery

When data preprocessing is completed, the next phase of web usage mining is pattern discovery. There are several methods that can be used to discover patterns, and these methods are rooted from fields such as data mining, statistics, and machine learning [5].

2.5.1 Statistical Analysis

Statistical techniques are the most common method to extract useful information about visitors to a website [5]. Statistical analysis can be performed on the session file variables such as page views, viewing time and length of navigational paths.

The output of applying statistical methods could be determining the most frequently accessed pages, average viewing time of a page, average length of navigation paths to a specific page, or the most common invalid URI. Despite its lack of depth, the output of statistical analysis can occasionally help in reorganizing web content, making better marketing decisions and enhancing system performance, facilitating the site modification task and provide support for marketing decision. There different statistical analysis tools that helps to discover descriptive pattern from the web log file, such as Web log storming, Web log expert, weblog expert, google analytics, awstats and analog [27].

Web Log Storming: is an interactive, desktop-based Web Log Analyzer for Windows (www.weblogstorminig.com). It discover detailed website statistics with interactive graphs and reports. Very complete detailed log analysis of activity from every visitor to the website is only a mouse-click away. It's easy to track sessions, hits, page views, downloads, or whatever metric is most important to each user. Website behavior, from the top entry and exit pages, to the paths that users follow, can be analyzed. We can learn which countries and cities visitors came from, and which operating systems and browsers they use. Due to these strength, Web log storming selected for statistical pattern discovery of web usage mining of this study.

Web Log Expert: is a powerful access log analyzer. It can help to reveal important statistics about website usage: activity of visitors, access statistics, and paths through the site, visitors' browsers, and much more (www.weblogexpert.com). The HTML reports of web log expert include multiple graphical charts to show the number of visitors, and what pages they viewed after arriving.

Google Analytics: It is a free utility provided by Google which helps in keeping a track of unique visitors (www.googleanalytics.com). Google Analytics allows to dig down deep into stats to see breakdowns of individual regions, states/provinces, cities and numerous other items to better identify your site visitors.

AWStats: is a free powerful tool that generate advanced web usage statistics (www.awstats.org). This tool works as a CGI Script or from command line. It displays all sorts of information that the log contains. It uses a partial file information to be able to process large log files.

Analog: An easy to use and install freely available log analysis tool. It is extremely fast, highly scalable, works on any operating system and an easy to install tool.

2.5.2 Data Mining techniques

Data mining tasks can be classified into two categories: descriptive data mining and predictive data mining [28]. Predictive data mining like classification used to constructs one, or a set of, models, performs inference on the available set of data, and attempts to predict the behavior of new data sets. Descriptive data mining such as clustering, association rule mining, used to describes the data set in a concise and summary manner and presents interesting general properties of the data.

Classification

Classification is supervised way of learning which maps the data items to one of many predefined classes [5]. In the Web domain, this technique used to establish a profile of users belonging to a particular class or category. It helps to develop a profile for items belonging to a particular group according to their common attributes. This requires extraction and selection of features that best describe the properties of a given class or category. According to Srivastava et al. [5] classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifier, and support vector machines.

Clustering

Clustering is a technique to group users who exhibit similar browsing patterns, or web pages which exhibit similar contents [28]. According to Pierrakos et. al, [28], Web usage mining allows the overlapping of clusters. Clustering is a method of gathering items that have similar characteristics. In the context of web mining, we can have two distinct cases, user clusters and page clusters. User clustering identifies groups of users that seem to have similar behaviors when browsing through a website. Page clustering results in groups of pages that are apparently related to each other in terms of user's perception. Such clustering information is then used for personalizing a website.

Association Rule Mining

Association is the discovery of association relationships or correlations among a set of items [13]. They are often expressed in the rule form showing attribute-value conditions that occur frequently together in a given set of data. In the context of Web usage mining, association rule used to discover associations between Web pages based on their co-occurrence in user sessions [29]. It refer to set of pages that are accessed together with a support value more than some specified threshold.

Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests.

2.5.3 Association rule mining algorithm

There are number of association rule discovery algorithm, however the most widely algorithm for pattern discovery are Apriori and Frequent Pattern Growth (FP-Growth) [13].

Apriori Algorithm

Apriori algorithm is an efficient algorithm, which is one of the best available methods for discovering frequent patterns [30]. Apriori runs breadth-first search algorithm and uses a hash tree structure to count candidate items at each step. Apriori algorithm searches for large itemsets during its initial database pass and use its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent itemsets and those below are called small itemsets. The algorithm is based on the large itemset property which states. Any subset of a large itemset is large and any subset of frequent item set must be frequent.

FP-Growth Algorithm

The FP-growth algorithm is frequent items set mining uses the FP-tree structure to attain a divide-and conquer to break down the mining problem into a set of smaller problems [13]. It retains the item set association information and compressed databases are divided into a set of conditional databases, each one associated with a frequent item. It takes the help of prefix tree representation of the given database of transactions (called FP tree), which saves considerable amount of memory for storing the transactions. Every node in the tree represents one item and each path represents the set of transactions that involve with the particular item. All nodes referring to the same item are linked together in a list, so that all the transactions that containing the same item can be easily found and counted. Performance of FP-growth compared with Apriori in large databases [31]. Frequent patterns from web log data using FP for finding the most frequently access pattern generated. Apriori for association rule mining. Finding frequent sets using candidate set generation. Apriori sets many candidates so the cost is high compared to FP and its getting slow. FP requires less memory, low cost, very fast than Apriori.

2.6 Pattern Analysis

Pattern analysis is about undergoing further interpretation of the discovered patterns before applying the discovered rules to useful application [5]. It is used to extract the interesting rules, patterns or statistics from the output of the pattern discovery process by filtering the irrelative rules or statistics. Another aim of this analysis is to obtain some information can offer valuable insights about users‟ navigational behavior. For example we can understand the number of users that started from a page and proceeded through some certain pages and finally visited their goal page. Also, we can obtain some information about page popularity or some pages that contain the most information for a visitor.

2.7 Challenges of Web Usage mining

There are different challenges in web usage mining, which need to be addressed in order to discover useful knowledge from them [32]. The first challenge is the volume of document requests that are amassed in a single log file. The log file can be regarded as a huge collection of records of events that took place in a website. It includes entries of document traversal, file retrieval and unsuccessful web events among many others that are organized according to the date and time it was made to the web server. Distinguishing which document requests is needed and which is not, may vary significantly based on the purpose and specification of the study undertaken. Thus, it is important to carefully consider which type of document requests will maximize the chances of discovering knowledge about users and the website. Otherwise it could result in the removal of document requests that carry vital information, thus leading to data loss and undesirable effects of the study undertaken.

The second challenge involves finding and grouping all document requests that was made by the same user during the user's visit to the website. This is commonly referred to as session modeling. This task can be difficult and is highly dependent on the type of information available in the web logs. The term 'session' is generic and could mean different things to different people based on the scope and purpose of one's study. According to Cooley, et al., [16] a session is all of the document accesses that occur during a single visit by a user to a website. Apart from the volume of the data and its low quality, the data is not completely structured. It is in a semi-structured

format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information.

In addition, due to the cache present on client browser, most of the requests, if they are present in the cache are not sent to Web server. Most of the time, the users do not visit the home page of a Website [33]

2.8 Application of Web Usage mining

The discovered knowledge from the usage data of web has emerged as a necessity in today's world as users expect to find the relevant information on the web as quickly as possible and web administrators want to track user's behavior for many reasons [5]. There are a range of application provided by web usage mining techniques like personalization, system improvement, site modification, business intelligence.

Web personalization: is the process of customizing a Website to the needs of each specific user or set of users, taking advantage of the knowledge acquired through the analysis of the user's navigational behavior [34]. Web personalization simply means to understand the needs and interest of visitors of the site and respond accordingly. It is the process of gathering and storing information about site visitors, analyzing the information, and, based on the analysis, delivering the right information to each visitor at the right time. The personalization of Web services is a leap in the direction of alleviating the information overload problem and making the Web a friendlier environment for its users.

System Improvement: performance and other service quality attributes are crucial to user satisfaction and high quality performance of a web application [5]. Web usage mining of patterns provides a key to understanding Web traffic behavior, which can be used to deal with policies on web caching, network transmission, load balancing, or data distribution. Web usage is also useful for detecting intrusion, fraud, and attempted break-ins to the system.

Site Modification: The attractiveness of a website, in terms of content and structure, is crucial for its usability [19]. Web usage mining could provide knowledge on user navigation behavior. Then, this knowledge could be utilized by website designers to design the website for user's cease of navigation. Besides, structure of a website may be set to change automatically on the basis of usage patterns discovered from server logs as in the case of adaptive website [35]. It can assist in making

a website adaptive by dynamically changing the content and structure of the site according to the patterns mined from user behavior.

Business Intelligence: Information on how customers are using a Website is critical information for marketers of e-commerce businesses. It can be used to provide information on products bought and advertisement click-through rates. According to Buchner et al. [36], a knowledge discovery process in order to discover marketing intelligence for web data.

User Characterization: Mining of web usage patterns can help in the study of how browsers are used and the user's interaction with a browser interface. Usage characterization can also look into navigational strategy when browsing a particular site. It focuses on techniques that could predict user behavior while the user interacts with the Web.

2.9 Related Works

Different national and international scholar's works in the area of web usage mining in the past few years to explore various aspects of the web mining endeavor that ranges from developing a web mining architecture to application of statistical and data mining techniques for web mining.

Anand [37] conducted a research on data mining of Web access log on computer science website of Royal Melbourne Institute of Technology University. He used data mining techniques to find access patterns hidden inside huge volumes of web access data. His objective is to investigate the access patterns between visitors from within Australia and visitors from outside Australia, visitors from within Royal Melbourne Institute of Technology University and visitors from outside Royal Melbourne Institute of Technology University. Anand used the data mining techniques such as classification, association rules, with three pattern discovery process such as transaction identification and feature extraction, discovery of the access patterns, analysis of the discovered patterns. He identified long transaction of access log manually. His finding shows, visitors from Australia generally visit the root page while visitors from outside Australia do not. However, some Visitors from outside Australia visit the root page and pages about post graduate programs (such as Master of Technology). Finally, he recommend for further researcher, that data mining techniques with better preprocessing like cleaning Web robots through the web browser information , to improve the discovered interesting patterns.

Wei [38], conduct research on exploring health website users on Clarian Health website in India by web mining. His objective is to examine the navigation behavior of different user groups and to make some suggestions to reconstruct a website for more customized Web service. Wei, used access weblog files from one local health provider's website. He used WUM-prep, a Perl-based tool supporting data preparation for mining Web server log files. Web Utilization Miner (WUM), to discover navigation patterns over the aggregated view of the web log. Rapid Miner, to generate and compare classifiers of naive Bayesian and Support Vector Machine (SVM). His findings show that users are not searching health information as much as was thought. The top two health topics which patients are concerned about are children's health and occupational health. Patients and doctors have different search strategies when looking for information on the website.

Wei recommended, to redesign and improve the website by adding more intuitive portal and more customized links for both user groups.

Mekonnen [39], conducted a research on web usage pattern discovery on Addis Ababa University official website. He used combination of statistical analysis and data mining approaches for pattern discovery. Mekonnen used a Weka plug in called WUMPrep4 to clean the raw log and develop session file and transform the session file to a format that is compatible to Apriori algorithm of Weka. Regarding usage patterns pattern of the research, association mining, i.e. discovery of common pages of the site that are accessed together, was accomplished for the prepared transaction file. Besides, he used Mach5 statistical analyzer tool to discover possible statistical reports of the web log record. Finally, he recommended to use combination of statistical analysis and data mining based pattern discovery for discovery of effective usage patterns of websites.

Tadele [40], attempts web usage pattern discovery on Addis Ababa University official website. He follow the web usage mining process that is suggested by Sirivastava [5]. He also used python code for data preparation, Mach5 statistical analyzer and Weka tool for association rule and sequences mining with Apriori algorithm. His finding shows the daily access trend, top entry and exit page and page that display error response of the website.
He also discover the correlation between pages that accessed together. His recommendation is mainly the need for reconstruction of Addis Ababa University official website in user friendly manner.

Awet [41], also conducted a research on exploring the navigational behavior of users of Addis Ababa University official website. He used the same web usage mining process as Tadele [42] used. He used WUMprep tool for data preparation, Web Utilization Miner for statistical analysis and pattern discovery. . His finding shows most requested pages, top entry and exit pages, referrer page and pages that accessed frequently after the home page of the website. He recommended to go for combining web usage mining such as content mining with web usage mining for efficiency of exploring user's behavior on the website.

Getahun [42] conducted a research on Web usage pattern discovery and analysis by region taking the case of Ethiopian Airline official website. He follows the Web usage mining process model suggested by Sharma [13]. He used IANA IP assignment dataset to divide the server log dataset into different region based on IP address. He used WUMprep tool, Java programming, and MS Excel for data preparation, Google Analytics for statistical analysis and Weka tool for association rule mining with FPGrowth algorithm. His finding shows the other page of the website except home page needs optimization and more promotion is required to make the website more accessible with respect to referrers. He also show the navigational behavior of user across region have similarity and difference. Finally, he recommended for feature researcher to include proxy server and client server logs data and proxies and client cookies method of user identification method for better usage pattern discovery.

CHAPTER THREE

METHODS AND ALGORITHM

In this study statistical and association rule techniques are applied to explore user access patterns of ECX website. An attempt was made to construct a model that describes web usage access patterns using Web log storming and Apriori and FPGrowth algorithm. In this section the methods and algorithms used to build the models are discussed in detail.

3.1 Architecture of the system

Figure 3.1 shows the over process of web usage mining pattern discovery of this study.

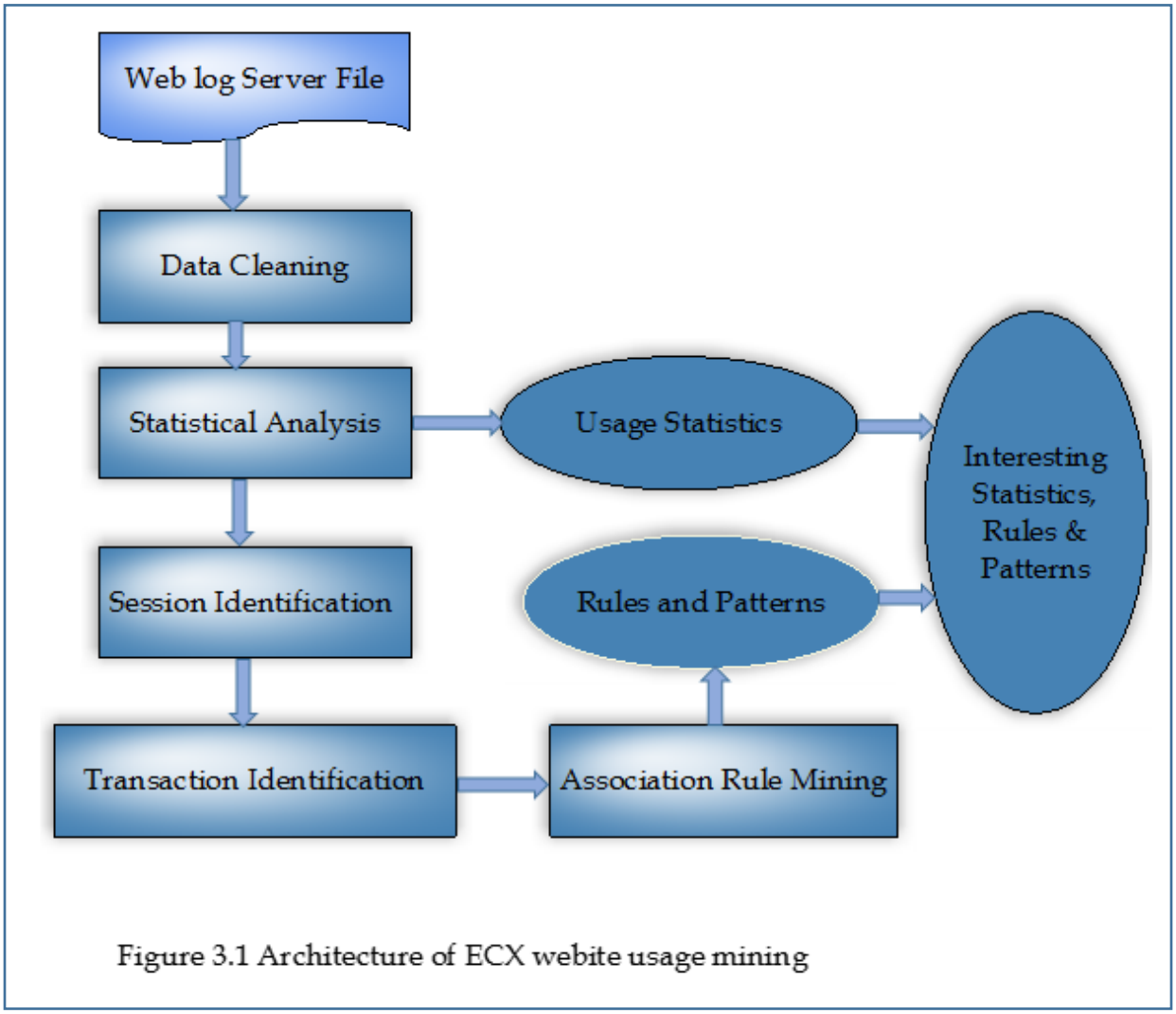

Figure 3.1 Architecture of ECX webite usage mining

As shown the above figure, Web server log is the input for the pattern discovery process. The web server log cleaned by using web log preparation tools to remove irrelevant attribute from the log record, user session identified to determine hits from a single visitor in a predefined timeframe, then transactions are identified from precisely identified session in to the form appropriate for association rule mining algorithm user transactions. Finally, usage statistics from statistical analysis and rules and patterns from association rule mining are discovered.

3.2 Statistical Analysis

Statistical analysis techniques are used to extract descriptive knowledge about visitors of a Website. In this study Web log storming used as a tool to analyze the Web usage data of ECX official Website. This tool does far more than just generate common reports - it displays detailed website statistics with interactive graphs and reports. Web log storming splits all the pages accessed by the user into session using a certain time limit by using IP address and user agent. The referrer based method is used for identifying the new session in the web log files, when IP address, user agents are the same. The next session is calculated if the time duration exceeds more than the given time limit. Then the web log storming tool generate frequent visited page of the website by counting the frequency of pages/files that accessed by the visitor from the Web server log record. The top visitor countries of the website generated by converting the visitors IP address code into county and by counting frequency of country that visit the ECX website.

3.3 Association Rule Mining

Association rule mining is a data mining method originally invented to extract patterns from transactional databases. It has been widely used from traditional business applications such as cross-marketing attached mailing, catalog design, store layout, and customer segmentation to e-business application such as Web personalization [43]. Currently, Web server log file e- business application is the basic data source for web usage pattern discovery.

Web usage log files generated on web servers contain huge amount of information appropriate for applying data mining methods to discover potentially useful knowledge [15]. According to B. Mobasher et al. [15], discovering web usage association rules is one of the popular data mining methods that can be applied on the web usage log data. The association rule mining is a technique for finding interesting frequent patterns, associations, and correlations among sets of items [34].

Association rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. The information contained in association rules can be used to learn about website visitor behavior patterns, and improve web marketing movements [16].

Given a set of transactions where each transaction is a set of items (itemset), an association rule implies the form $X \Rightarrow Y$, where X and Y are itemsets; X and Y are called the body and the head, respectively. A rule can be evaluated by two measures, called confidence and support. Support for the association rule $X \Rightarrow Y$ is the percentage of transactions that contain both itemset X and Y among all transactions. The confidence for the rule $X \Rightarrow Y$ is the percentage of transactions that contain an itemset Y among the transactions that contain an itemset X. Support (usefulness) can be measured with a minimum support threshold (minsup).

Confidence (certainty) can be measured with a minimum threshold for confidence (minconf). Association rule mining is the task of finding all rules with support *S* and confidence *C* such that $S \geq minsup$ and $C \geq minconf$, where *minsup* is support threshold and *minconf* is the confidence threshold [44].

The rule X ➔ Y holds in the transaction set D with support S, where S is the percentage of transactions in D that contain $X \cup Y$ (i.e., the union of itemsets X and Y, or say, both X and Y). This is taken to be the probability, $P(X \cup Y) = \frac{\#\ of\ transaction\ with\ itemset\ X \cup Y}{\#\ of\ total\ transaction}$ Support shows the probability that all the predicates in X and Y fulfill together.

$$\text{Support}(X \rightarrow Y) = \frac{\#\ of\ tuble\ bothe\ contain\ X \text{ and } Y}{total\ number\ of\ tuble}$$

The rule X ➔ Y has confidence C in the transaction set D, where C is the percentage of transactions in D containing X that also contain Y. This is taken to be the conditional probability.

$$P(Y|X) = \frac{\#\ of\ transaction\ with\ itemset\ X \cup Y}{\#\ of\ transaction\ with\ itemset\ X}$$

Confidence measure how often predicates Y fulfilled if predicate X get fulfilled.

$$\text{Confidence (X/Y)} = \frac{\#\ of\ tuples\ containing\ both\ \ X \text{ and Y}}{\#\ of\ tuples\ containing\ X}$$

According to Agrawal et al. [44), association rule mining can be viewed as a two-step process. **First, find all frequent itemsets,** by definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, *min sup*.

Then generate strong association rules from the frequent itemsets, by definition, these rules must satisfy minimum support and minimum confidence.

In the context of web usage mining, item sets are sets of web pages accessed and association rule mining is used to discover the set of web pages accessed together in a user session. Given a set of web pages accessed by the user, other frequently co-occurred pages may be recommended to the user. The most popular algorithm used to discover association rules are Apriori and FP Growth algorithm [30].

3.3.1 Apriori Algorithm

Apriori algorithm is an efficient algorithm, which is one of the best available methods for discovering frequent patterns [30]. Apriori is most widely used when working on databases containing transactions. The algorithm serves as a basis for many other pattern discovery methods. Apriori runs breadth-first search algorithm and uses a hash tree structure to count candidate items at each step. Its search is complete and bottom up with a horizontal layout and discovers all frequent *itemsets*.

Apriori is an iterative algorithm that counts *itemsets* of a specific length at each step while going over the database. The initial tasks of the method are scanning all records in the database and finding the first frequent item sets. After this step, using these frequent items, it forms potential frequent candidate 2-itemsets. An additional pass for scanning all transactions in the database is performed for determining supports of these patterns. By observing supports, the infrequent ones are eliminated from candidate 2-itemsets, while the remaining ones will form frequent 2-itemsets. This process is repeated until all frequent itemsets have been discovered. The Apriori algorithm will have the following steps [30].

Initialization: Generate all the frequent itemset with cardinality of 1 (i.e. L_1) in which each elements are sorted lexicographically.

Join Step: Generate the candidate k-itemsets by joining L_{k-1} with itself (i.e. $C_k = L_{k-1} \infty L_{k-1}$) using the following procedure:

- ✓ Take any two element from L_{k-1} where each of them are similar in all their elements except the last
- ✓ Form k-itemset set by union operation of the two (k-1)-itemset
- ✓ Repeat the procedure for all possible such elements

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. For example L_3= *{abc, abd, acd, ace, bcd}.* Generating C_4 from L_3

- ✓ *abcd* from *abc* and *abd*
- ✓ *acde* from *acd* and *ace*

Pruning: *acde* is removed because *ade* is not in L_3**,** C_4={*abcd*}.

Apriori Algorithm Pseudo code

Figure 3.2 gives an overview of the Apriori algorithm for finding all frequent itemsets. The first pass of the algorithm simply counts item occurrences to determine the large 1-itemsets. A subsequent pass, say pass *k*, consists of two phases. First, the large itemsets L_{k-1} found in the (*k*-1) the pass are used to generate the candidate itemsets C_k *(*Set of candidate item set of size *k)*, using the Apriori candidate generation function (apriori-gen) described below. Next, the database is scanned and the support of candidates in *Ck* is counted. For fast counting, an efficient determination if the candidates in *Ck* that are contained in a given transaction *t* is needed.

The Apriori algorithm is:

```
L1 = {large 1-itemsets};
for (k = 2; Lk-1 ? Ø; k++ ) do
    Ck = apriori-gen(Lk-1);
    forall transactions t ? D do
      Ct = subset(Ck, t);
        forall candidates c ? Ct do
        c.count++;
        end
    Lk = { c ? Ck | c.count = minsup}
  end
return ∪ k Lk;
```

Figure 3.2 Algorithm for Apriori [30].

3.3.2 FPGrowth Algorithm

The FP-growth algorithm: mining frequent patterns without candidate generation [45]. It compress a large database into a compact Frequent-Pattern tree (FP-tree) structure.

FP-Tree frequent pattern mining is used in the development of association rule mining. FP-Tree algorithm overcomes the problem found in Apriori algorithm. By avoiding the candidate generation process and less passes over the database, FP-Tree was found to be faster than the Apriori algorithm [45]. It adopts a divide and conquer strategy. Firstly it compresses the database representing frequent items into a frequent pattern tree or FP-tree. It retains the item set association information and compressed databases are divided into a set of conditional databases, each one associated with a frequent item. It takes the help of prefix tree representation of the given database of transactions (called FP tree), which saves considerable amount of memory for storing the transactions.

An FP-Tree is a prefix tree for transactions. Every node in the tree represents one item and each path represents the set of transactions that involve with the particular item. All nodes referring to the same item are linked together in a list, so that all the transactions that containing the same item can be easily found and counted. Large databases are compressed into compact FP tree structure. FP tree structure stores necessary information about frequent item sets in a database [45].

A frequent-pattern tree (or FP-tree in short) is a tree structure defined below [46].

- ✓ It consists of one root labeled as "null", a set of item-prefix subtrees as the children of the root, and a frequent-item-header table.
- ✓ Each node in the item-prefix sub-tree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none.
- ✓ Each entry in the frequent-item-header table consists of two fields, (1) item-name and (2) head of node-link (a pointer pointing to the first node in the FP-tree carrying the item-name).

Based on this definition, we have the following FP-tree construction algorithm [46].

FP-tree construction

Input: A transaction database DB and a minimum support threshold

Output: FP-tree, the frequent-pattern tree of DB.

Method: The FP-tree is constructed as follows [46].

- ✓ Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as F- List, the list of frequent items.
- ✓ Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in DB do the following. Select the frequent items in Trans and sort them according to the order of F-List. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree ([p | P], T). The function insert tree ([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N's count by 1; else create a new node N, with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree (P, N) recursively.

Figure 3.3 shows mining of frequent patterns with FP-tree by pattern fragment.

```
if Tree contains a single prefix path
    then {
    let P be the single prefix-path part of Tree;
    let Q be the multipath part with the top branching node replaced by a null root;
    for each combination (denoted as ß) of the nodes in the path P do
      generate pattern ß _ a with support = minimum support of nodes in ß;
      let  freq_pattern_set(P) be the set of patterns so generated; }
      else let Q be Tree;
        for each item ai in Q do {
          generate pattern ß= ai _ a with support = ai .support;
          (construct ß's conditional pattern-base and then ß's conditional FP-tree Treeß ;
          if Treeß = Type equation here.
           then call FP-growth(Treeß, ß);
          let freq_pattern_set(Q) be the set of patterns so generated; }
    return (freq_pattern_set(P)  freq_pattern_set (Q) _ (freq_pattern_set (P)×freq_pattern_set (Q)))
```

Figure 3.3 Algorithm for FP-Growth [46].

CHAPTER FOUR

DATA PREPARATION

Data used in data mining operations is rarely directly suitable for analysis. Before the analysis is conducted data must be first cleaned of unwanted or noisy entries and then converted into the form suitable for further analysis [47]. The aims of the preprocessing step in a Web usage mining process are roughly to convert the raw log file into a set of transactions (one transaction being the list of pages visited by one user) and to remove noisy requests (e.g. implicit requests or requests made by Web robots).

The goal of preprocessing in this study is to end up with the set of minable object for ECX Official Website. It used to convert the usage data into data abstraction necessary for pattern discovery. The main preprocessing tasks include data cleaning, user identification, session identification, transaction identification, and data transformation.

4.1 Data Collection

In this research, the source of data set is web server log file and the actual website. The server log file collected from *www.ecx..com.et*. The user access log from January 11, 2015 to February 10, 2015 are taken, which are logged by default. Sample data extracted from row log file of ECX website is shown in figure 4.1.

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2015-01-21 00:20:17
#Fields: date time s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) sc-status sc-substatus sc-win32-status time-taken

2015-01-11 00:11:05 192.168 GET /KnowledgeCenter.aspx - 80 - 192.168.95.143 Mozilla/5.0+(Windows+NT+6.1;+WOW64)+Gecko/20100101+Firefox/34.0 200 0 0 1601

2015-01-11 00:45:10 192.168 GET /home.aspx - 80 - 192.168.77.52 Mozilla/5.0+(Windows+NT+5.1;+rv:35.0)/20100101+Firefox/35.0 200 0 0 1187 2015-01-11

2015-01-11 01:28:13 192.168 GET /HistoricalPrice.aspx - 80 - 192.168.110.145 Mozilla/5.0+(Windows+NT+5.1;+rv:25.0)+Gecko/20100101+Firefox/25.0 200 0 0 14437

2015-01-11 01:28:37 192.168 GET /HistoricalPrice.aspx - 80 - 192.168.112.162 Mozilla/5.0+(Windows+NT+5.1;+rv:35.0)+Gecko/20100101+Firefox/35.0 200 0 0 1187
```

Figure 4.1 sample web log file

S.no	Field	Appears as	Description of fields
1	Date	date	The date on which the activity occurred.
2	Time	time	The time at which the activity occurred.
3	Server IP Address	s-ip	The IP address of the server on which IIS 5.0 generated the log entry.
4	Method	cs-method	The action that the client was trying to perform (for example, a GET method).
5	URL Steam	cs-uri-stem	The resource that the server accessed (eg Default.htm).
6	URL Query	cs-uri- query	The request that the client was trying to perform.
7	Server Port	s-port	The port number to which the client is connected.
8	User Name	cs-username	The name of the authenticated user who accessed the server. The hyphen represents anonymous users.
9	Client IP Address	c-ip	The IP address of the client that accessed to the server.
10		cs(User-Agent)	The browser used by the customer.
11	HTTP Status	sc-status	The status of the action in terms of HTTP protocol.
12	HTTP Substatus	sc-substatus	The substatus error code
13		sc-win32-status	The status of the action in terms Win2K.
14	Time Stamp	time-taken	The duration of the action.

Table 4.1 Description of IIS log file description

4.2 Data Preprocessing

The aim of data pre-processing is to clean data from irrelevant records, user and session identification, select essential features and finally transform raw data for mining. All these stages will be analyzed in more detail in order to understand why pre-processing plays an important role in knowledge discovery process complex web log data. The Web log data of this research preprocessed by using Web log explorer tool.

4.2.1 Data Cleaning

In this phase record contains images, videos, Cascading Style Sheet files (CSS), scripts, and flash animations that are not necessary for statistical analysis and Web usage mining are removed. The page/files with file extension such as .aspx, .asp, .html, .pdf, .docx was selected by using Log file Viewer tool. The screen shut of Log file viewer standard edition shown below in figure 4.2.1.

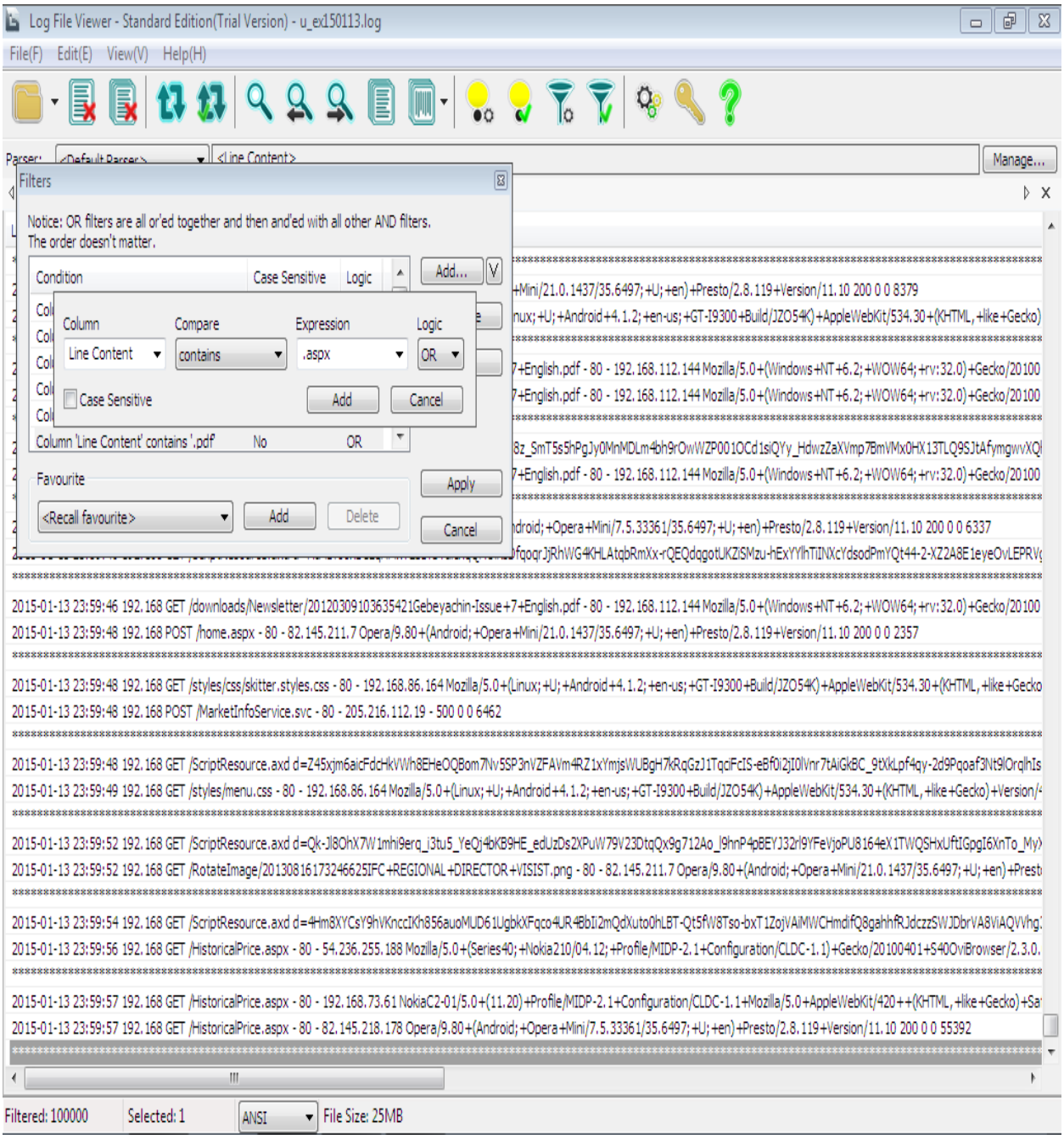

Figure 4.1 data cleaning

Further cleaning done by using MS- Excel 2013. In this cleaning phase the log record which logged by admin user, staff user removed by using IP address and user name of log record. Incomplete record such as null IP address value record. The summary preprocessed daily log data is shown below in table 4.2.

No	Log Date	# of Record Before Cleaning	# of Log Record After Cleaning by log Viewer	# of Log Record After Cleaning by Excel
1	2015-01-11	15,493	6113	1013
2	2015-01-12	59040	16216	4485
3	2015-01-13	129029	19069	6881
4	2015-01-14	92176	10817	5489
5	2015-01-15	118072	17025	6518
6	2015-01-16	92682	19857	2972
7	2015-01-17	44526	13467	2971
8	2015-01-19	1751	1267	902
9	2015-01-20	76	7	7
10	2015-01-21	1302	46	46
11	2015-01-22	28576	8128	869
12	2015-01-23	93883	22398	3464
13	2015-01-24	55253	2437	2605
14	2015-01-25	19149	6401	1198
15	2015-01-26	69425	20081	2593
16	2015-01-27	161298	10273	1133
17	2015-01-28	124595	17272	3797
18	2015-01-29	89584	22261	3044
19	2015-01-30	134373	15578	3732
20	2015-01-31	51738	12301	2142
21	2015-02-01	16719	5688	1311
22	2015-02-02	88860	37961	1996
23	2015-02-03	264868	27367	2997
24	2015-02-04	338969	35765	4008
25	2015-02-05	146374	8934	2368
26	2015-02-06	4296	3904	4294
27	2015-02-07	66227	19345	2757
28	2015-02-08	13118	7809	994
29	2015-02-09	88501	18234	2454
30	2015-02-10	97704	21035	4238
Total	30 Day	**2,506,766**	**427056**	**83278**

Table 4.2 Summary of preprocessed daily log data

4.2.2 Session Identification

User sessions are used to reconstruct visitor's sessions from Web log records. A common timeout method have been selected to sessionise the filtered log record. According to Spiliopoulou [15], more than 90% of genuine visit sessions can be identified by the timeout method. A time period of between 20 and 30 minutes is considered optimal. In this study 20 minutes time period is taken. The session identification process done by using Web log storming tool. The tool identified based on IP address and time out method. The summary of user and session identification is shown below in table 4.2.2

No	Activity	By Human	By Spider	Total
1	Total hit	52787	30491	83278
2	Average hit per day	1702.81	360.29	2607.03
3	Total session	17816	11168	28984
5	Average hits per session	2.96	2.73	2.87
6	Average different page viewed per session	1.31	1.14	1.25
7	Average different files downloaded per session	0.08	0.27	0.15
8	Total visitor IP addresses	7534	1985	9519
9	Average sessions per IP address	2.36	5.63	3.04
20	Average visitors at one moment	1.29	1.12	2.41

Table 4.3 Summary of user and session identification report

4.2.4 Transaction Identification

In this phase most important attributes that are essential for transaction identification are identified. Date and time, host, and country attribute of the session are selected from the total session file of log data. And most frequent page/file attribute are selected from accessed log data. Then the session created by each selected page/file are identified. Finally the total session of all pages/file and session of each most frequent page/file saved by MS-Excel (.CSV) format separately. Table 4.2.3.1 shows sample session of all page/file with selected attributes.

No	Date and Time	Host	Country
1	2015-01-11 00:00	82.145.208.236	Iceland
2	2015-01-11 00:06	192.168.95.144	Ethiopia
3	2015-01-11 00:07	199.16.156.126	United States
5	2015-01-11 00:46	93.186.202.239	Germany
7	2015-01-11 01:50	145.62.32.131	United Kingdom
8	2015-01-11 05:48	199.16.156.126	United States
9	2015-01-11 05:56	41.138.99.233	Burkina Faso
8	2015-01-11 10:11	178.137.212.182	Ukraine
9	2015-01-11 10:33	61.135.189.44	China

Table 4.4 Sample session record attribute

Sample frequently accessed page/file attribute that selected transaction identification shown below in table 4.4.

Page/file
/
/historicalprice.aspx
/home.aspx
/commodities.aspx
/mktdatagraph.aspx
/companyprofile.aspx
/graph.aspx
/operations.aspx
/downloads/contracts/coffee/coffeecontracts.pdf
/membership.aspx
/newsandevents.aspx
/coffeebulletinboard.aspx
/downloads/contracts/sesame/ecxsesame.pdf
/knowledgecenter.aspx
/downloads/contracts/pulse/ecxwhitepeabeans.pdf

Table 4.5 Sample page/file frequently accessed by the visitor

Table 4.6 and 4.7 shows sample session created on root and historical price page of the website respectively.

No	Date and time	Host	Country
1	2015-01-12 00:01	82.145.208.162	Iceland
2	2015-01-12 00:12	106.120.160.109	China
3	2015-01-12 00:17	93.192.215.253	Germany
4	2015-01-12 00:24	82.102.141.214	Israel
5	2015-01-12 00:26	192.181.11.10	United States
6	2015-01-12 00:29	94.198.24.91	Netherlands

Table 4.6 Sample session created on root page of the website.

No	Date and Time	Host	Country
1	2015-01-11 05:15	113.89.38.36	China
2	2015-01-12 02:34	123.63.33.89	India
3	2015-01-12 04:32	2.50.4.203	United Arab Emirates
4	2015-01-12 05:13	82.55.224.25	Italy
5	2015-01-12 06:52	83.250.35.247	Sweden

Table 4.7 Sample session created on historicalprice page of the website

After identifying a set of different user sessions that corresponding to each page/file viewed (visited). Finally, in order to provide transaction that used for pattern discovery, session data from different page/file viewed merged with total user session. The session identification process is shown in the step.

- ✓ First the session formed by all page/file saved by table name (mainsession.csv). It contain session id, date and time, host, country and URL of 40 page/file (URL1-URL40) attribute. Then mainsession.csv session data table imported into MYSQL. However the value of each URL attribute value is empty before the second step.
- ✓ Second the session formed by each page/file saved by name (URL1.csv, URL2.csv… URL40 session.csv) with session id, date and time, host attribute. Then each URL.csv session data imported into MYSQL table.
- ✓ Thirdly each URL attribute value of main session table filled by yes or no value based on existence of session record in the corresponding main session URL table by using MYSQL statement. If the first row main session table date and time and host exist in to corresponding URL table record the value of main session table URL attribute value is given yes else no.
- ✓ Finally the Main session table, which contain all URL attribute value exported from MYSQL table. Refer to Appendix A.

Sample transaction identification MYSQL statements. Sample merged session data shown below in table 4.8.

ID	Dates	Host	Country	URL1	URL2	URL3	URL4	URL5	URL6
1	2015-01-11 00:07	199.16.156.126	Iceland	yes	no	no	no	no	no
2	2015-01-11 09:46	192.168.77.40	Ukraine	no	no	no	no	yes	no
3	2015-01-11 00:07	82.145.216.198	Iceland	no	no	no	no	no	no
4	2015-01-11 16:51	82.145.211.194	Germany	no	yes	no	no	no	no
5	2015-01-11 00:16	192.168.107.24	Ethiopia	no	no	no	no	no	yes
6	2015-01-11 10:33	61.135.189.44	China	no	no	no	no	no	no
7	2015-01-11 00:35	192.168.126.4	Iceland	yes	no	no	no	yes	no
8	2015-01-11 00:45	192.168.77.52	Iceland	no	yes	no	no	no	no
9	2015-01-11 00:46	82.145.208.236	Germany	no	no	no	yes	no	no
10	2015-01-11 00:50	192.168.95.100	Iceland	no	yes	no	no	no	no
11	2015-01-11 01:40	192.168.112.190	Iceland	no	no	no	no	no	no
12	2015-01-11 01:43	82.145.209.163	Ireland	no	no	no	yes	no	no
13	2015-01-11 02:21	192.168.112.162	Iceland	no	yes	no	no	no	no
14	2015-01-11 02:43	192.168.107.48	Iceland	no	no	yes	yes	no	no
15	2015-01-11 02:46	82.145.219.131	China	no	yes	no	no	no	no
16	2015-01-11 03:08	180.153.236.200	Iceland	no	no	yes	no	no	no
17	2015-01-11 03:14	82.145.217.197	Iceland	no	no	no	yes	no	no
18	2015-01-11 03:23	82.145.208.235	Iceland	no	no	no	no	yes	no
19	2015-01-11 03:27	82.145.219.145	Iceland	no	no	no	no	no	no
20	2015-01-11 03:27	82.145.216.223	Iceland	no	no	no	yes	no	no

Table 4.8 Sample identified transaction data.

After the above preprocessing phase the transaction data transformed into Weka understandable format. Before transforming the transaction data country attribute change to continent using MS Excel. Then the data categorized into six continent (Africa, Asia, Europe, North America, Oceania, and South America) transaction data. Finally the transaction data of all continent and each continent transaction filtered. After that the other attribute except URL attribute deleted from the transaction data. Sample transformed data set are shown below in table 4.9.

URL1	URL2	URL3	URL4	URL5	URL6	URL7	URL8	URL9	URL10	URL11	URL12	URL13	URL14	URL15	URL16
yes	no	no	no	yes	no	no	no	yes	no	no	yes	no	no	no	yes
no	no	no	no	yes	no	no	no	no	no	no	no	yes	no	no	no
yes	no	no	no	no	no	no	no	yes	no	yes	no	no	no	no	no
no	yes	no	no	no	no	no	no	no	no	no	no	no	yes	no	no
no	no	no	no	no	yes	no	yes	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no	no	yes	no	no	no	no	no	no
yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no
no	yes	no	no	no	no	no	no	no	no	no	yes	no	no	no	no
no	no	no	yes	no	no	no	no	no	no	no	no	no	yes	no	no
no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	yes	no	no	no	no	no	no	no	no
no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	yes
yes	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no
no	no	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no
no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no
no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no
no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	yes	no	no	no	no	no	no	no	no	no
no	no	no	yes	no	no	no	no	no	no	yes	no	no	no	no	no
no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	yes	no	no	no	no	no	no	no	no

Table 4.9 sample transformed dataset.

The transaction data are prepared in the form of the above table based on country and continent level. Finally the outputs of transformation used for association rule mining discovery.

CHAPTER FIVE

EXPERIMENTATION

After data preparation, the experiment conducted to discover pattern or knowledge. In this study statistical analysis and data mining experiments conducted to discover pattern/knowledge from web usage data of Ethiopia Commodity Website. To accomplished experiment Web log storming 3.0 for statistical analysis and Weka 3.7 for association rule discovery are used. The experiment is done using: Satellite L755 Toshiba laptop with Intel(R) core (TM) i5 2.4 that has 32 bit window 7 ultimate.

5.1 Statistical Analysis

Statistical analysis techniques are the most common method to extract descriptive knowledge about visitors of a Website. In this study Web log storming used as a tool to analyze the Web usage data of ECX official Website. This tool does far more than just generate common reports - it displays detailed website statistics with interactive graphs and reports. It allows to dig as deep into data as we need to. Simply right-click a row in any report, whether it's a page, a host, session, country, operating system, and browser or anything else and choose from a list of sub-reports available for this item. For example, in a page views report we can select a certain page and get a list of its session, countries and cities these visitors are from.

By using this tool summary of web usage report is generated concerning most frequently accessed page, top entry and exit pages of the user, paths that users follow, countries and cities of visitors, device, operating systems and browsers they use to visit the website, and frequent visitor countries of product interest. Statistical analysis using Web log Storming conducted using 30 (thirty) days web log data with 82850 preprocessed Web log records. The detail of the report described as follows.

Most Frequently Accessed Pages

Figure 5.1 shows the summary of most frequently accessed page of ECX official website.

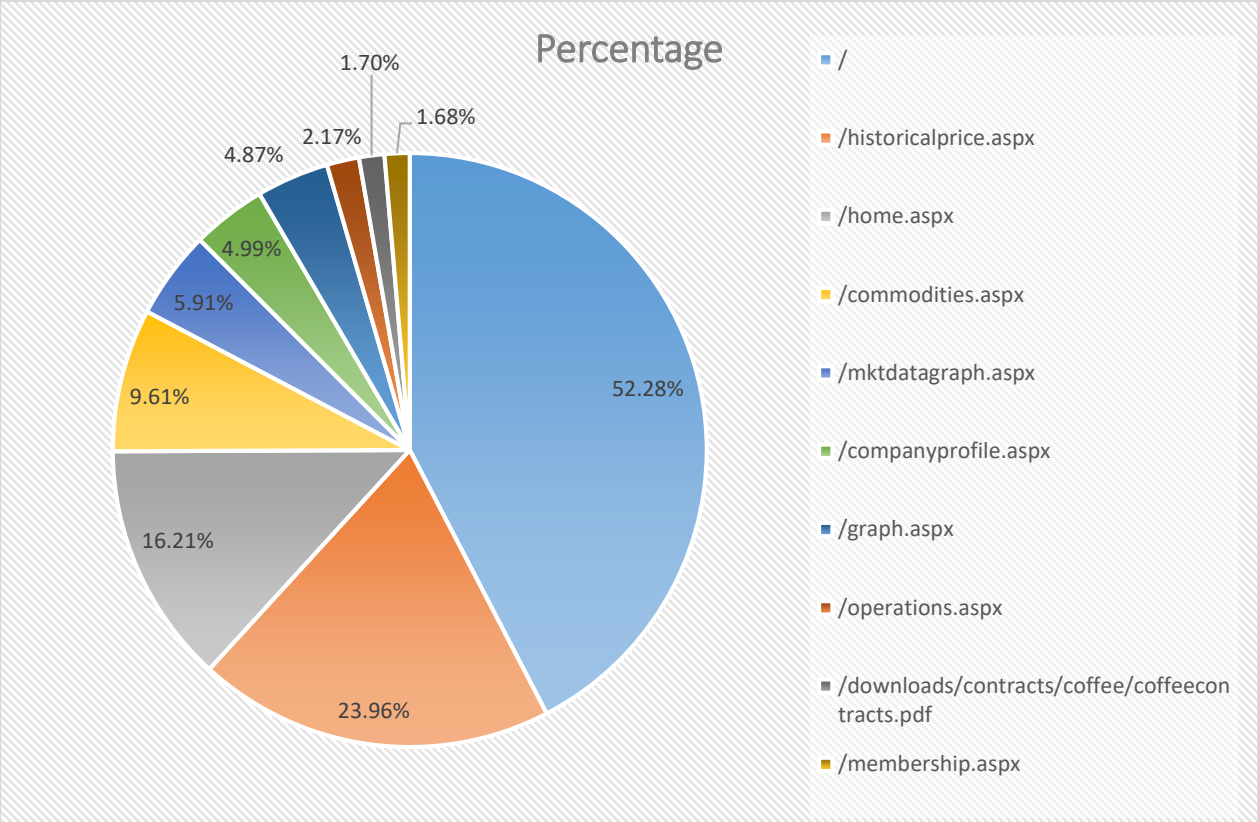

Figure 5.1 most frequently accessed page/file report

As shown in the above figure most frequently (52.28%) access page is root page. It indicates that shows that more than half of the visitor have entered into the ECX website directly or by typing the website address (wwww.ecx.com.et/). Historical price and home page also frequently access page followed by the root page.

Top Entry and Exit Page

This report shows list of entry pages that shows the users starting page when coming to visit the Website of ECX Website and exit page that user visit before exit from the website. The summary of the top entry and exit page report are shown below in figure 5.2 and figure 5.3 respectively.

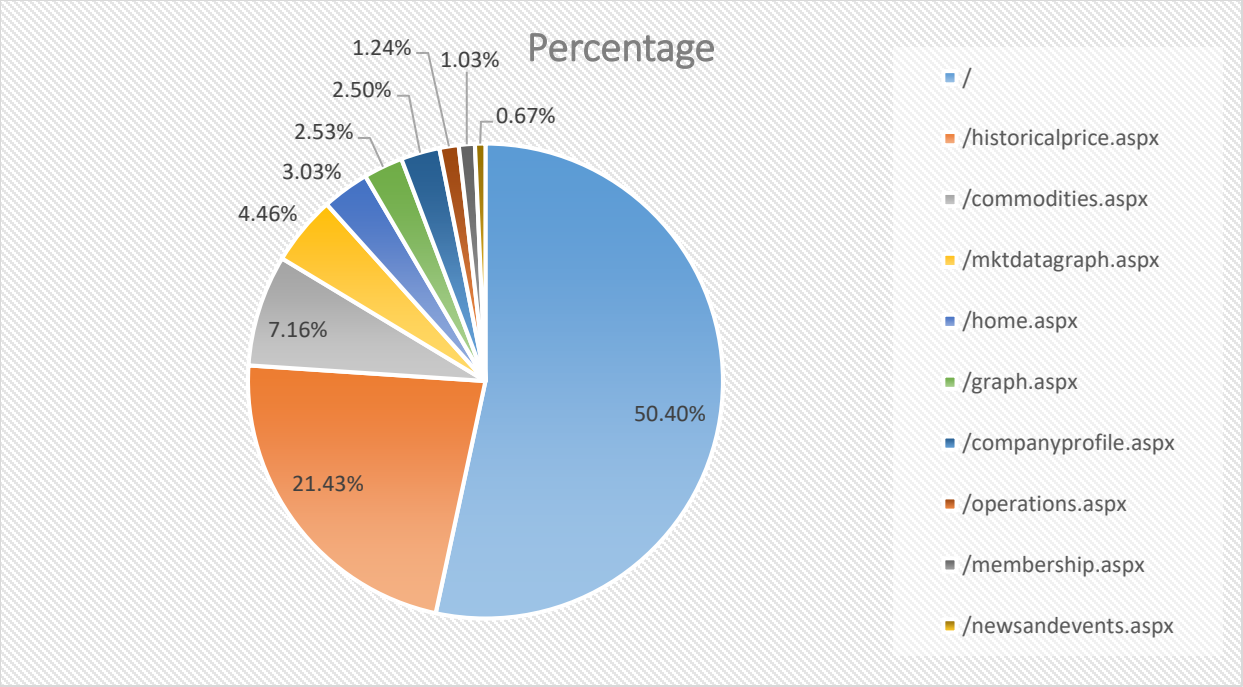

Figure 5.2 Summary report of top entry page

As shown in the above figure most of user starts navigation from root page followed by historical price and commodities page of the website. This indicates that user of ECX website can access other page of the website by the help of search engine in addition to root page. This also shows that most of the pages in addition to root page have search engine optimization.

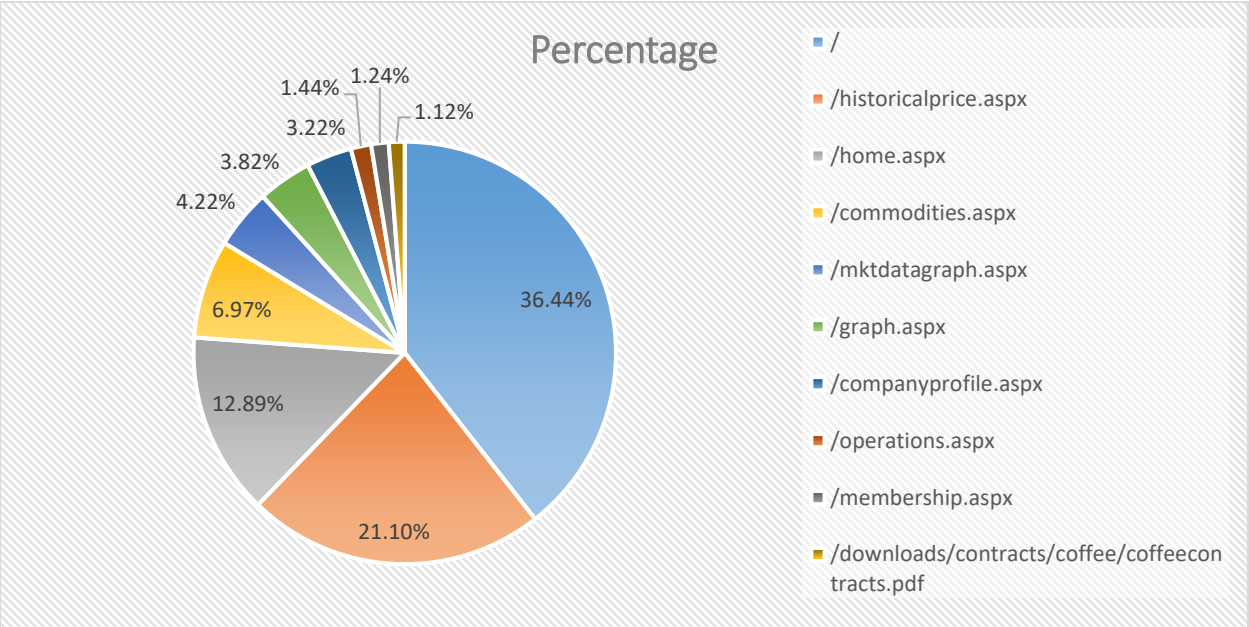

Figure 5.3 Summary report of top exit page

As shown in the above figure 5.3 root page are the top exit page of ECX website. , historical price page and home page. It indicate that after the user visit the website, and leave the website without making any click. Historical price page and home page are most frequent exit page followed by root page.

Top Navigation Path of Users

The navigational path of the user shows the path that the user follow when navigating through the website. Summary of the navigational path of the user path shown in figure 5.4.

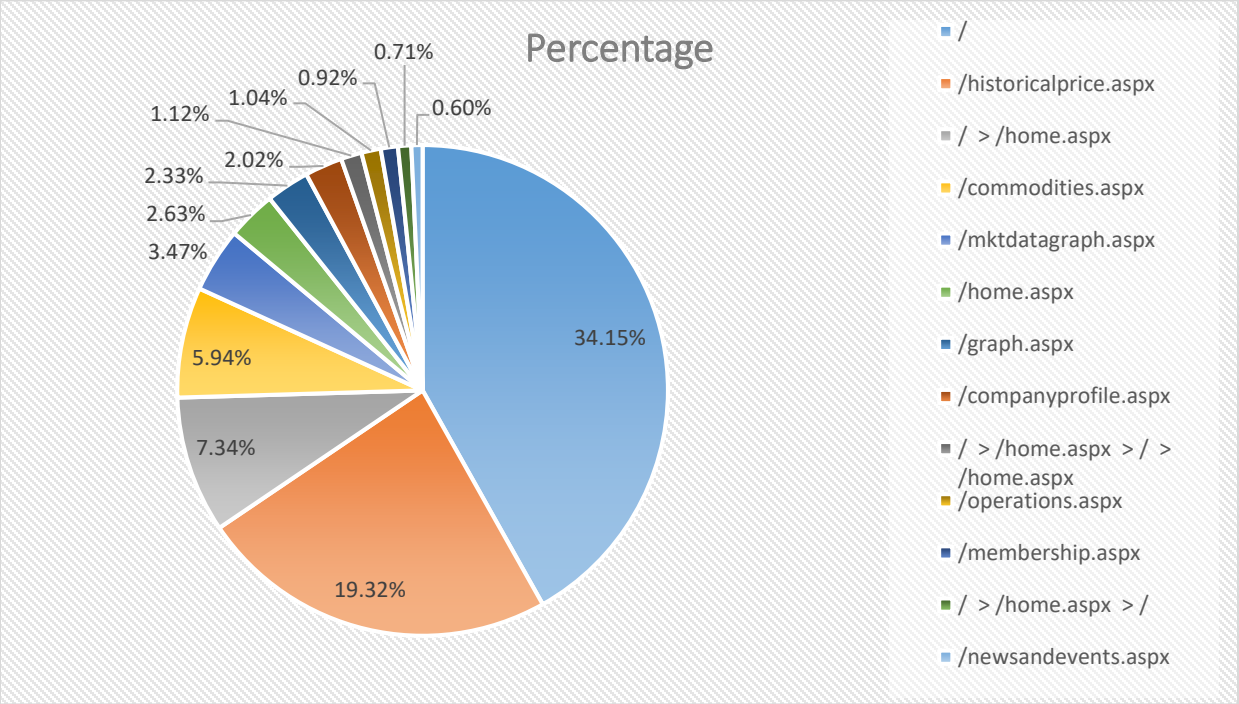

Figure 5.4 Summary of top navigational path of users

As shown in the above figure most of the user starts navigation from root page and historical price page leave the page without go through the other page. Two third of the ECX website user exit from the page they start navigation without further navigation.

Most Visitor Countries and Cities

The summary of top visitor country and city of country report shown below in figure 5.5 and figure 5.6 respectively. As shown in figure 5.5 top countries and city of ECX website visitors originate from Iceland, Ethiopia, United States, china and United Kingdom. It indicates that these countries are most active countries that search for ECX product.

This helps to identify where to advertise Ethiopian agriculture product to attract new customer. It also helps to compare countries which visit the website frequently and countries that currently the most importers (customer) of ECX product.

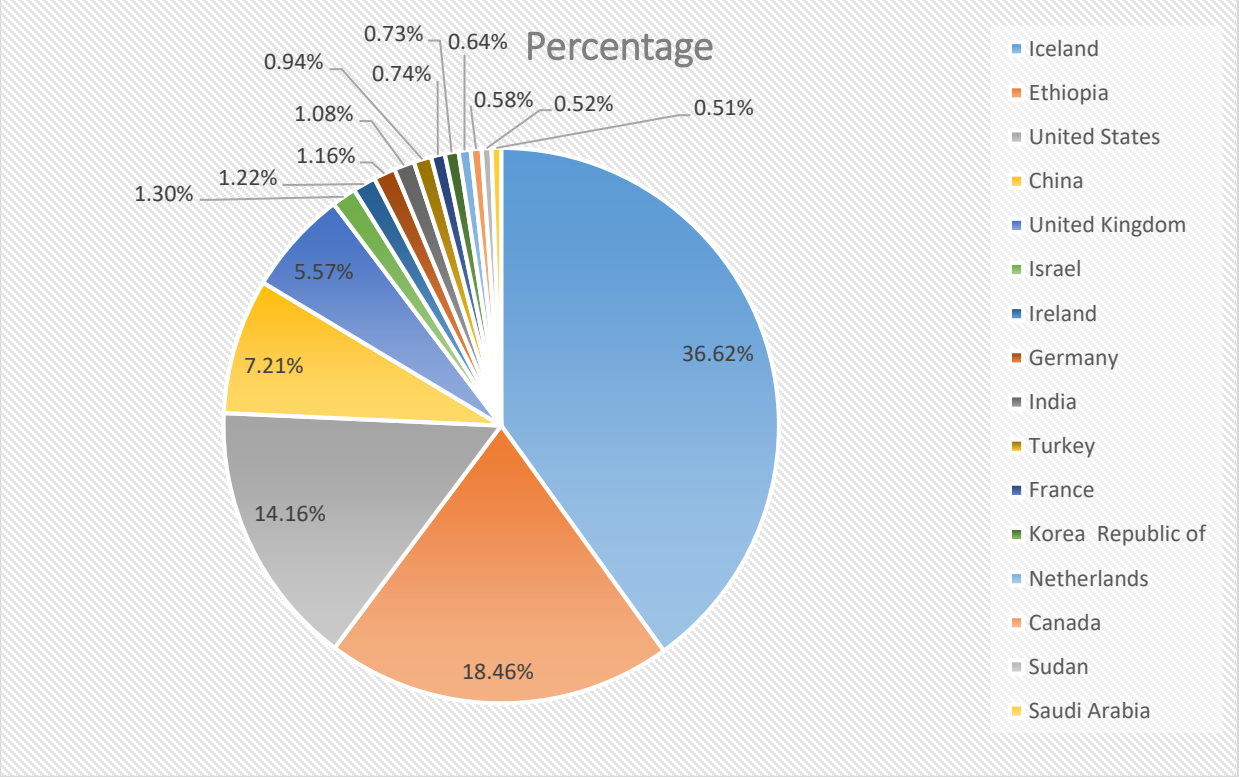

Figure 5.5 Summary of top visitor country report

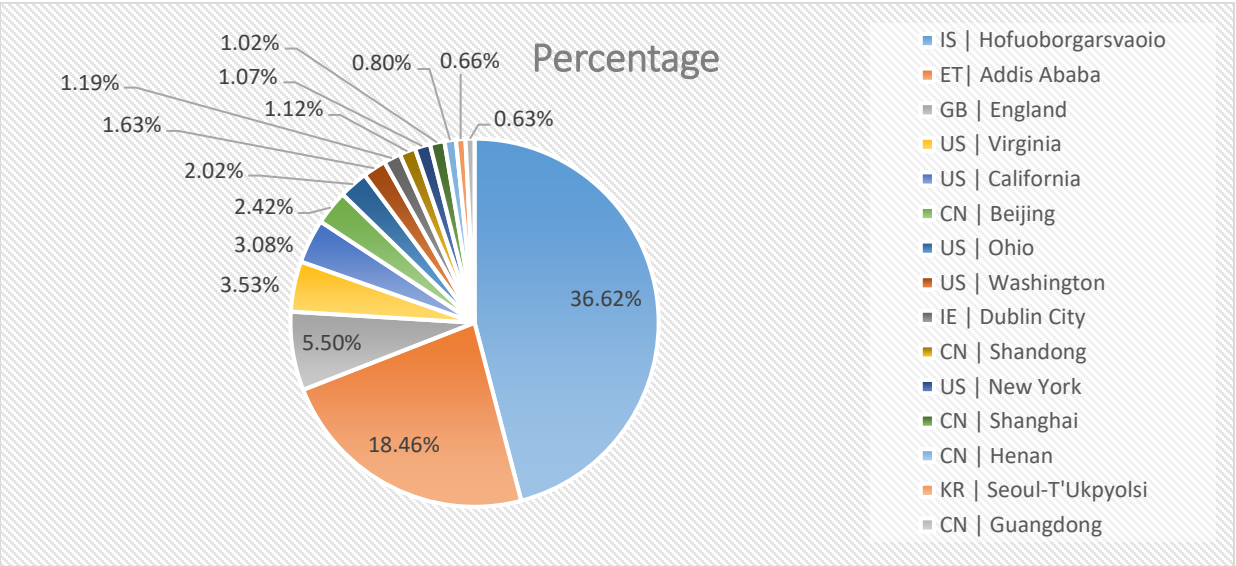

Figure 5.6 Summary top visitor city report

As shown in the above figure cities of top visitor countries are Hofuoborgarsvaoio form Iceland, Addis Ababa from Ethiopia, London United Kingdom, Virginia and California from United States, Beijing and Shandong from China.

Top Coffee, Seasame, Wheat, White pea, and Maize and Oil seed

Product Page Visitor countries

The following chart report shows the top product (coffee, Seasame, Wheat, White pea, maize and oil seed) visitor. It helps to identify the product type preference of most frequently visitor countries. Figure 5.7 shows top countries which looks for ECX coffee product.

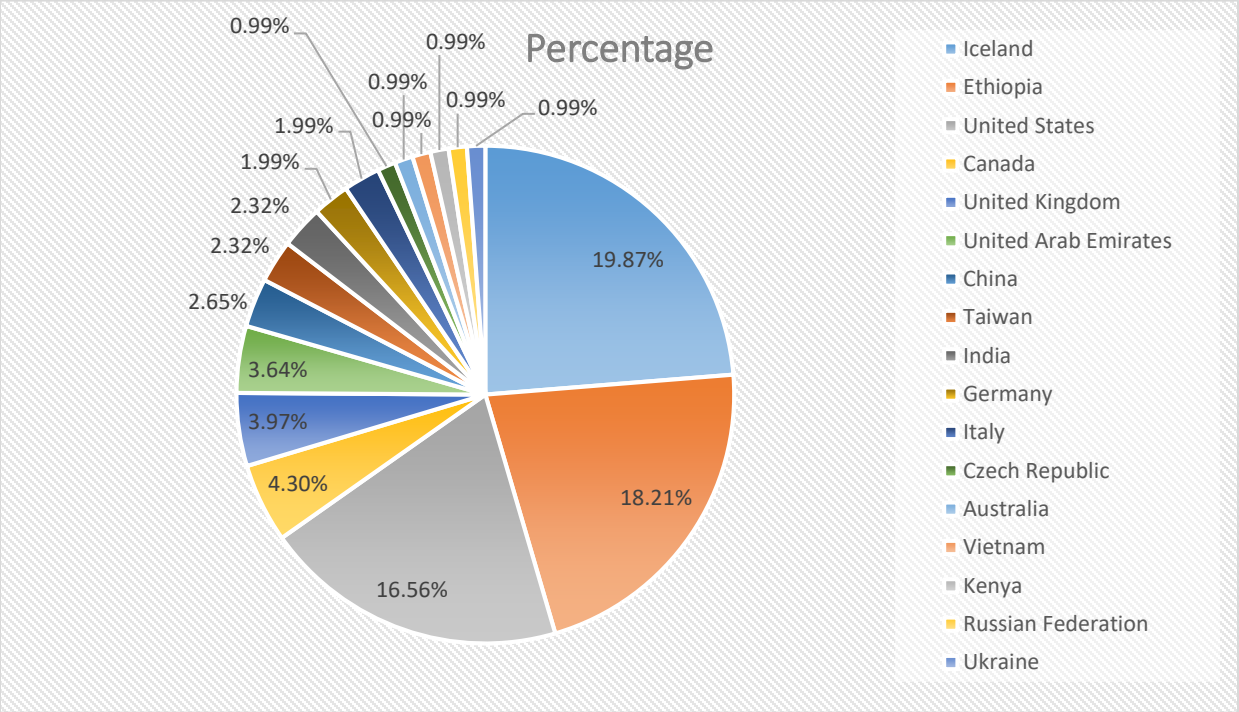

Figure 5.7 Summary top coffee product visitor countries report

As shown in the above figure Iceland Ethiopia is most frequent vistor country of Coffee product. Ethiopia, United States and Canada are most frequent frquent visitors of coffee product followed by Iceland.

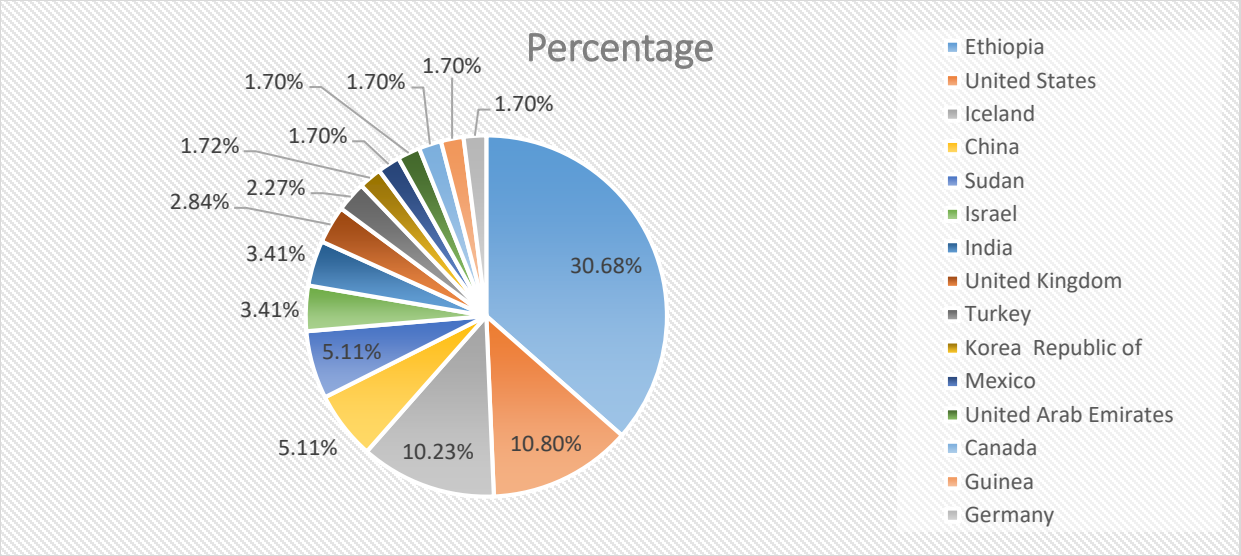

Figure 5.8 Summary top Seasame product visitor countries report

As shown in the above figure Iceland Ethiopia, United States, Iceland and China are most frequent vistor country of Seasame product.

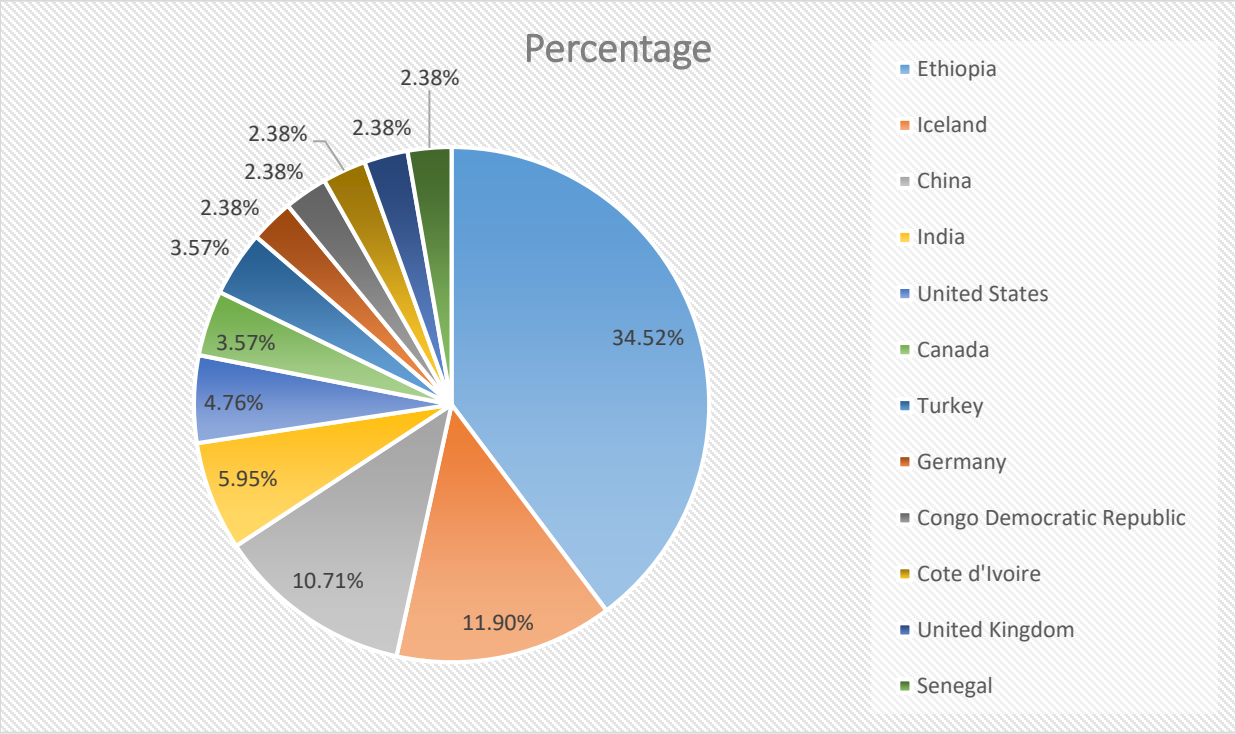

Figure 5.9 Summary top White Pea product visitor countries report

As shown in the above figure Ethiopia, Iceland, China and India are most frequent vistor country of White Pea product.

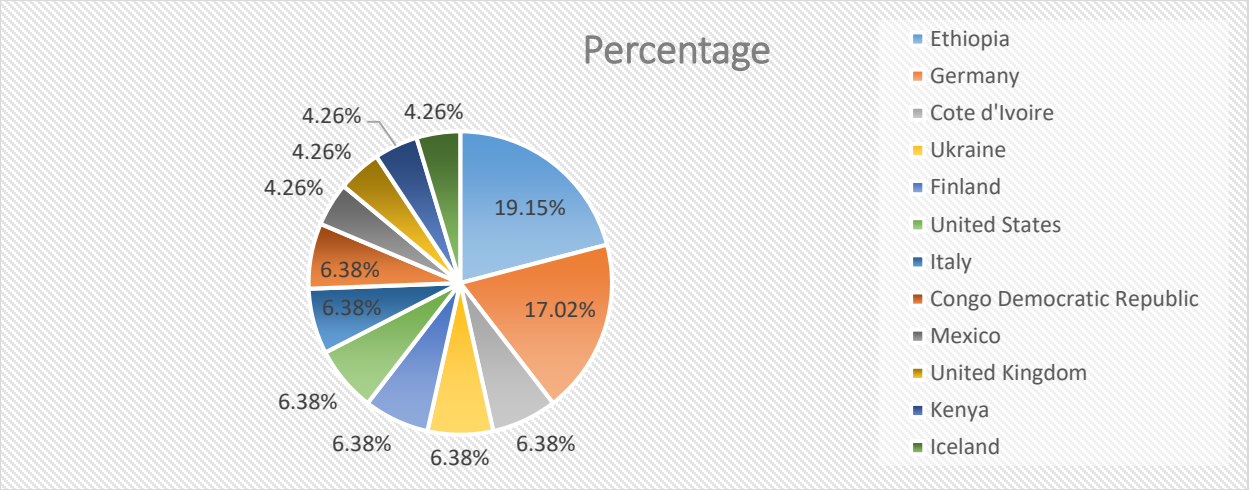

Figure 5.10 summary top Wheat product visitor countries report

As shown in the above figure Ethiopia, Germany, Cote d'voire, Ukraine and Finlad are most frequent vistor country of Wheat.

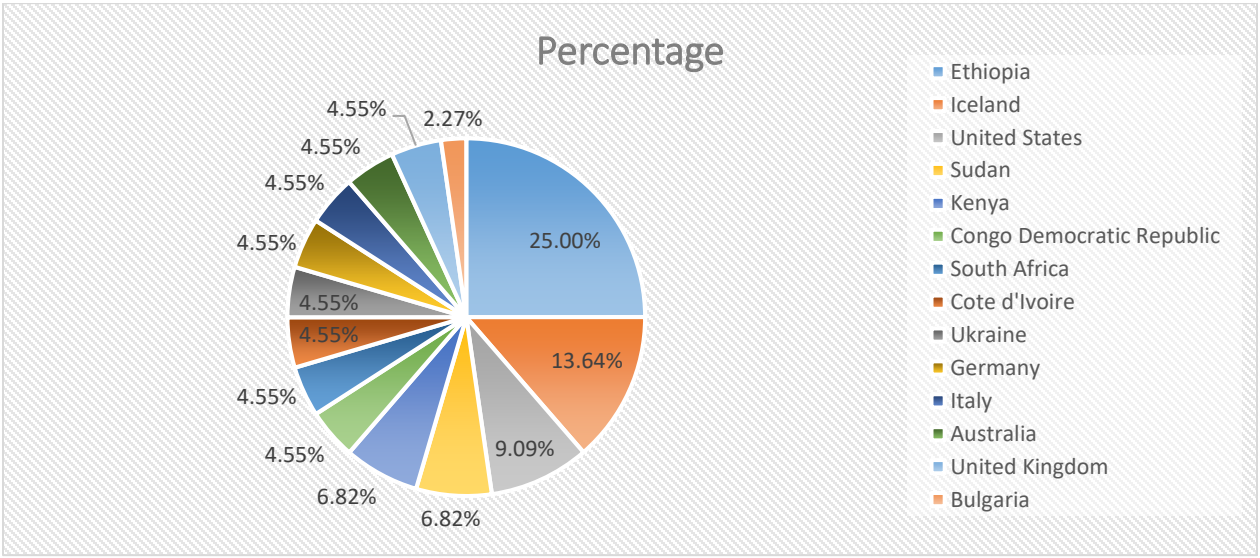

Figure 5.11 Summary top Maize product visitor countries report

As shown in the above figure Ethiopia, Iceland, United States, Sudan and Kenya are most frequent vistor country of Wheat.

Frequent error type

Figure 5.12 shows the type of error that frequently occur while user visit the ECX website.

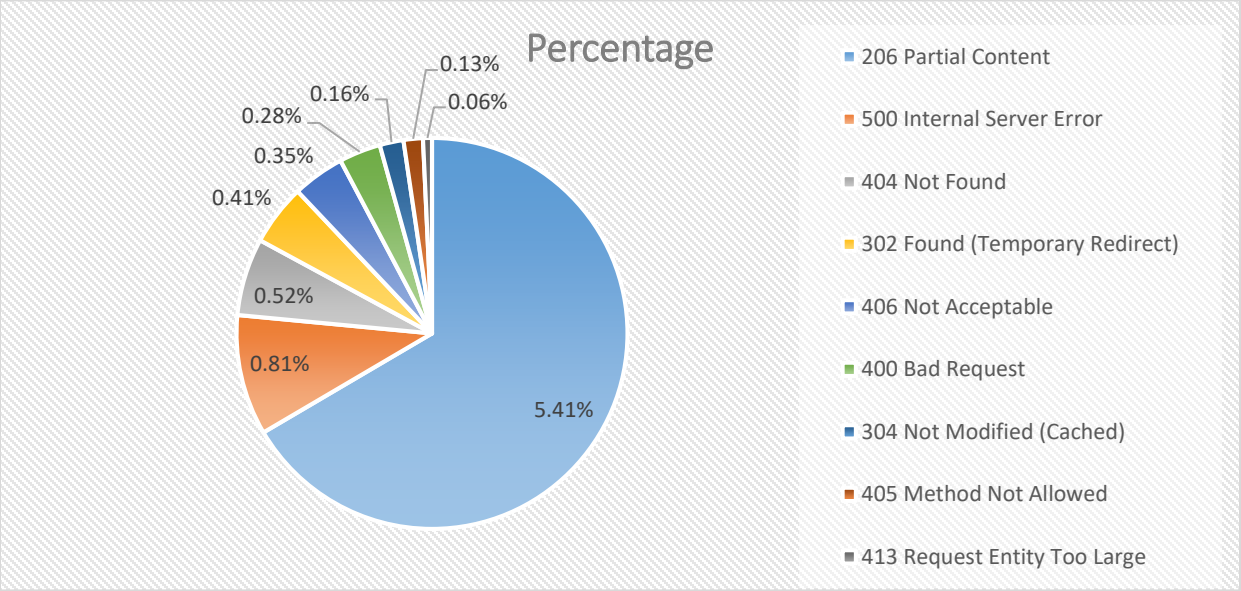

Figure 5.12 Summery of frequent error type

As shown in the above figure partial content (error 206) error type is the most frequent error type that face victors, while they visit ECX website. Internal server error type is frequent error type followed by partial content.

Partial Content and Internal Server Error type

Figure 5.13 and figure 5.14 shows frequent pages that cause for partial content and internal server error type of the ECX website respectively.

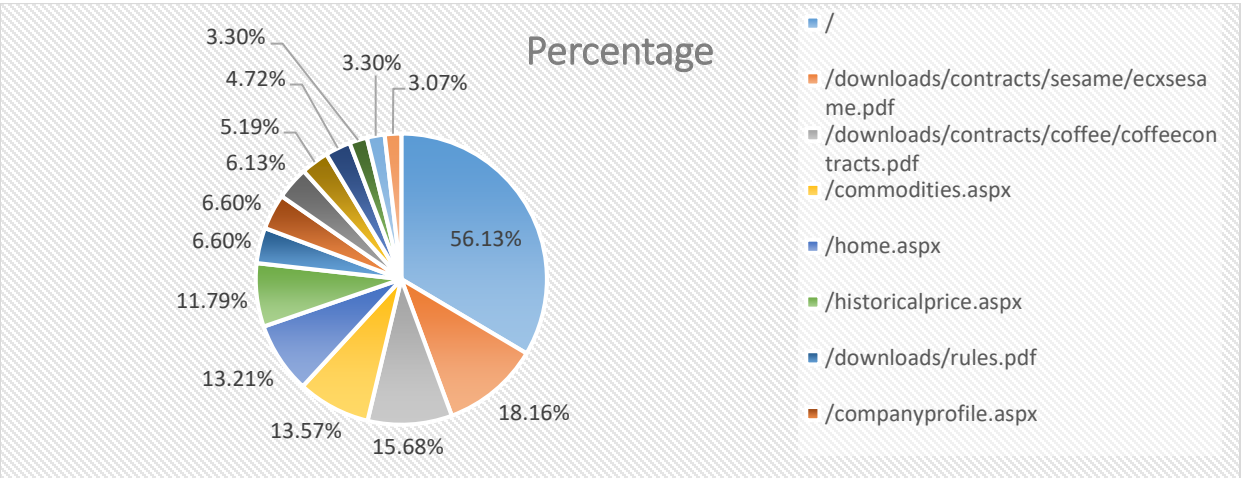

Figure 5.13 pages that case for frequent partial content error type

As shown in the above figure root page is cause for most frequent partial content error type. Coffee product file download is cause for most frequent error type followed by root page. This shows that the root page of the website have page loading problem.

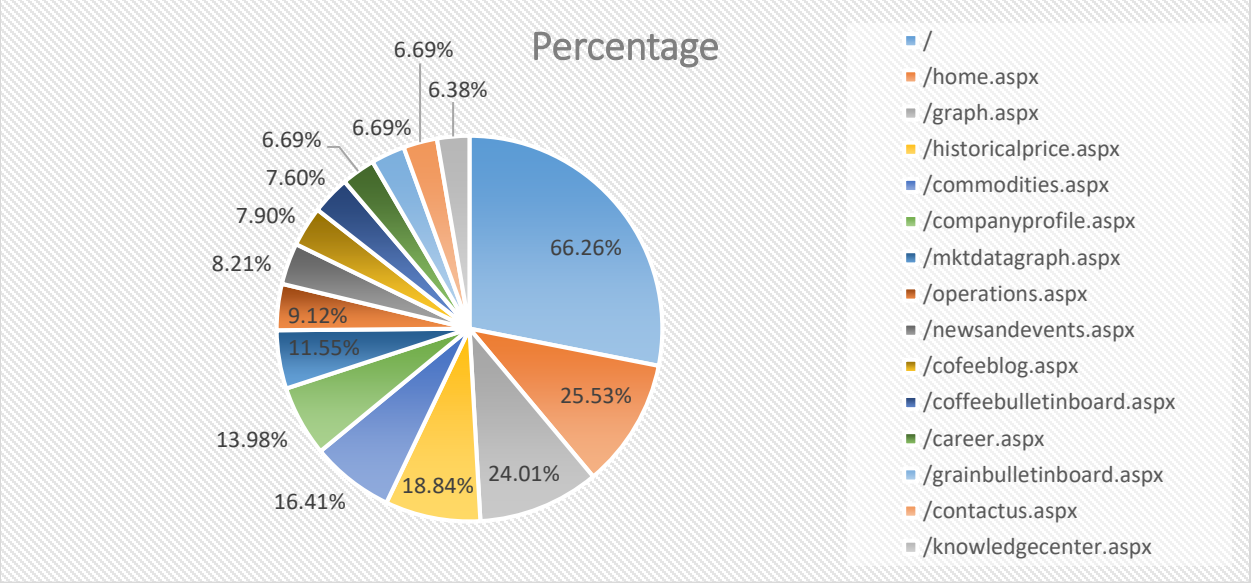

Figure 5.14 pages that case for internal server error type.

As shown in the above figure root page, home page and graph page are frequent page that cause for internal server error of the ECX website.

Types of operating system used

Figure 5.15 shows the type of operating system the visitor used to navigate the ECX website.

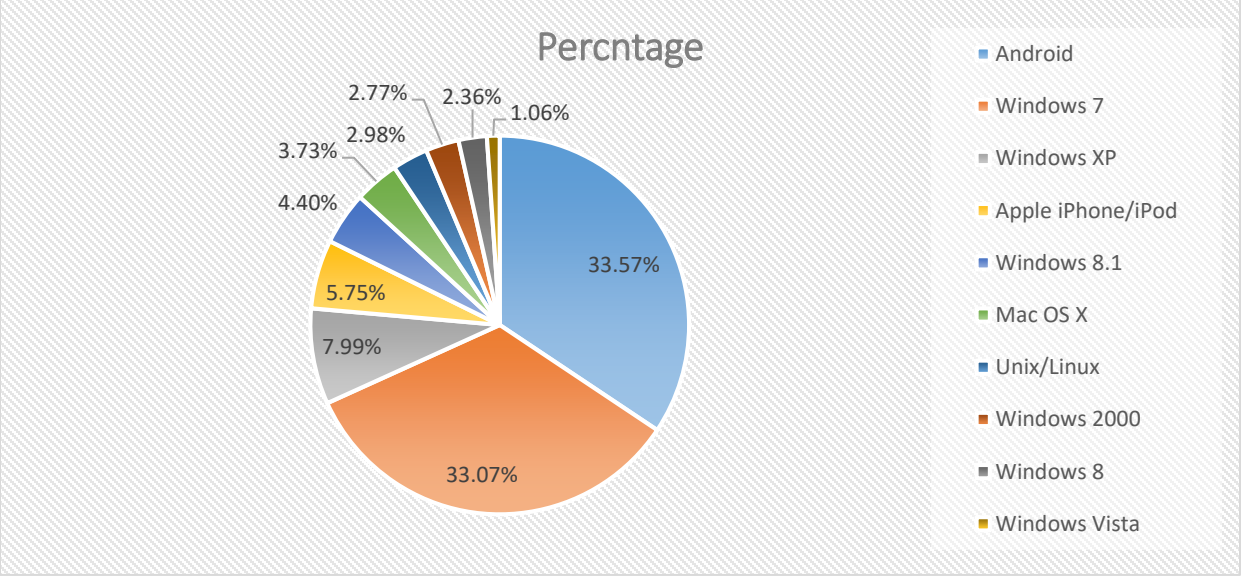

Figure 5.15 Summary of frequently used operating system

As shown in the above figure Android application is most frequently operating systems that have been used most frequently by the ECX website users to access the Website. Windows 7 operating system is most frequently used operating system followed by Android application. It indicates that most of visitor use mobile device to access the website.

Visitors Hit By Hours of Day

Figure 5.17 shows the daily hits of ECX website visitor. The highest hits of the website occur at 12AM, 12PM, 1AM, 1PM, 1PM, 2AM, 2PM and 3PM of the day.

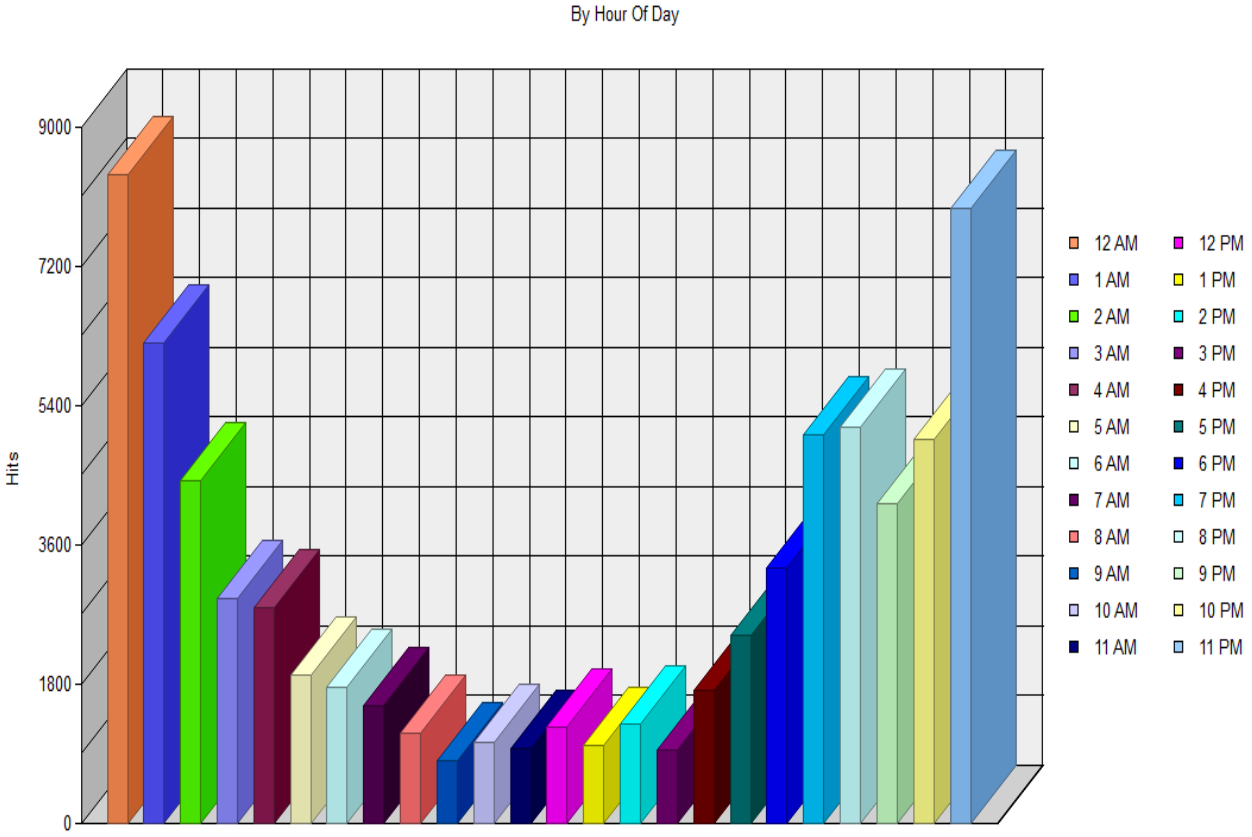

Figure 5.17 Summary of visitors hit by hours of day report

5.2 Association Rule Discovery and Analysis

In this study the main purpose of Web usage mining is to discover an interesting rule from ECX official website user access logs. Pattern discovery is performed after the process of transaction identification from session file of Web server logs. The selection of interesting patterns and analysis done with the consultation of expert Iyob [48], which helps for ECX to perform different businesses decisions over the web. The technique used for this web usage mining process is data mining algorithms. The data mining algorithms that can be performed on the preprocessed data is association rules.

5.2.1 Experimental setup

Association rule mining finds interesting associations a large set of data items. In this study association rules can be used to find correlations between web pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. In order to conduct experiments, Weka implementation is used as a tool with Apriori and FP-Growth association rule finding algorithm. The researcher used explorer environment in Weka, which allows to load experiment data set of .arff or .csv format, and to conduct experiment.

For experimental purposes, the preprocessed log file is used containing information about a user requests to the official website of ECX. Each line in the web usage log file contains information about the visitor country, date and time of request and URL requested of different user session. This preprocessed log data categorized into country and continent dataset. The original preprocessed dataset (dataset before categorized into continent) of web log contain 17810 user session of 116 countries. To conduct association rule mining experiment the data is categorized under 6 different continent (Africa, Asia, Europe, North America, Oceania, and South America). Then the dataset again categorized into 6 different group such as dataset which contain which contain African country, Europe country, Asian country, European country, North American country, Oceania country and South American Country.

Finally each six group dataset transformed to Weka understandable format with 40 URL request user session dataset. List of each selected URL request for association rule mining are shown in appendix A. While conducting the experiments, different parameter values (car, classindex, lowerbound minsupport, metrictype, minmetric, numrules, and upperbound support) are given.

5.2.2 Pattern Discover and Analysis with all country dataset

The following experiment result shows interesting patterns discovered from all country dataset of Weka 3.7.9 experiment using Apriori algorithm. To discover association rule within all county dataset by using a minimum support 0.95 and a minimum confidence 0.75. For detail Weka output, see Appendix C (I).

5.2.2.1 Apriori Algorithm Experiment

Rule 1: URL9=no URL14=no URL24=no URL28=no 17083 ==> URL25=no 17025 conf:(1)

According to rule 1, 100% visitor who is not visited about coffee daily price document (/downloads/contracts/Coffee/CoffeeContracts.pdf), seasame daily price document (/downloads/contracts/Sesame/ECXSesame.pdf), trading operation (tradingoperation.aspx) page, and oil seed daily price document (/download/oilseed.pdf) not visited about whitepea price document (/downloads/contracts/Pulse/ECXwhitepeabeans.pdf).

Rule 2: URL12=no URL13=no URL15=no URL19=no URL24=no URL25=no 17016 ==> URL28=no 16958 conf:(1)

According to rule 2, 100% visitor who is not visited about career (career.aspx) page, knowledge center (knowledgecenter.aspx) page, market reference (marketreference.aspx) page), trading operation (tradingoperation.aspx) page, and whitepea price document (/downloads/contracts/Pulse/ECXwhitepeabeans.pdf), not accessed about oil seed daily price document (/download/oilseed.pdf).

Rule 3: URL13=no URL15=no URL19=no URL24=no URL25=no 17167 ==> URL28=no 17108 conf:(1)

According to rule 3, 100% visitor who is not visited about coffee bulletin board (coffeebulleteboard.aspx) page, knowledge center (knowledgecenter.aspx) page, market coffee (marketcoffee.aspx) page, trading operation (tradingoperation.aspx) page, and whitepea price

document (/downloads/contracts/Pulse/ECXwhitepeabeans.pdf), not visited about oil seed daily price document (/download/oilseed.pdf).

Rule 4: URL12=no URL13=no URL15=no URL19=no URL25=no 17093 ==> URL28=no 17034 conf:(1)

According to rule 4, 100% of visitor who is not visited about career (career.aspx) page, coffee bulletin board (coffeebulleteboard.aspx) page, knowledge center (knowledgecenter.aspx) page, market coffee (marketcoffee.aspx) page, and whitepea price document (/downloads/contracts/Pulse/ECXwhitepeabeans.pdf) is not visited about oil seed daily price document (/download/oilseed.pdf).

The patterns discovered from the above Apriori algorithm shows that page/files of not visited by ECX website user such as coffee price document, seasame price document, knowledge center page, market coffee page, market reference page and trading operation page also not visited about whitepea and oil seed price document.

5.2.2.2 FP-Growth Algorithm Experiment

The following experiment result shows interesting patterns discovered from all country dataset of Weka 3.7.9 experiment using FP-Growth algorithm. To discover association rule within all county dataset by using a minimum support 0.05 and a minimum confidence 0.75. For detail Weka output, see Appendix C (I).

Rule1: [URL1=yes,URL3=yes]: 2244==>[URL33=no]: 2240<conf:(1)>lift:(1) lev:(0) conv:(1.11)

According to rule 1 about 100% of visitors who visited the root (/) and home (home.aspx) page of not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 2: [URL5=yes]: 1053 ==> [URL33=no]: 1050 <conf:(1)> lift:(1) lev:(0) conv:(0.65)

According to rule 2 about 100% of visitors who visited the market data graph (marketdatagraph.aspx/) page of ECX website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 3: [URL6=yes]: 889 ==> [URL33=no]: 882 <conf:(0.99)> lift:(0.99) lev:(0) conv:(0.27)
According to rule 3 about 99% of visitors who visited the graph (graph.aspx/) page of ECX official website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 4: [URL3=yes]: 2888 ==> [URL1=yes]: 2244<conf:(0.78)> lift:(1.49) lev:(0.04) conv:(2.14).
According to rule 4 about 78% of visitors who visited the home (home.aspx/) page of ECX website not accessed root page (/) of the website.

Rule 5: [URL33=no, URL3=yes]: 2883 ==> [URL1=yes]: 2240 <conf:(0.78)> lift:(1.49) lev:(0.04) conv:(2.14)
According to rule 5 about 78% visitor who visited home (home.aspx) page and visitor who not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf) of ECX official website visited the root (/) page.

5.2.3 Pattern Discover and Analysis with Africa dataset

This following experiment result shows interesting patterns discovered from all Africa dataset of Weka 3.7.9 experiment with FP-growth algorithm. For detail Weka output, see Appendix C (II). To discover association rule within all county dataset I have used a minimum support 0.1 and a minimum confidence 0.9.

Rule 1: [URL3=yes]: 500 ==> [URL20=no]: 500 <conf:(1)> lift:(1.01) lev:(0) conv:(2.69)
According to rule 1 about 100% visitor who visited home (home.aspx) page ECX website not visited about law and rules (rules.aspx) page.

Rule 2: [URL2=yes]: 495 ==> [URL20=no]: 494 <conf:(1)> lift:(1) lev:(0) conv:(1.33)
According to rule 2 about 100% visitor who visited home (home.aspx) page and historical price (historicalprice.aspx) page ECX website not visited about law and rules (rules.aspx) page.

Rule 3: [URL1=yes]: 2433 ==> [URL20=no]: 2420 <conf:(0.99)> lift:(1) lev:(0) conv:(0.93)
According to rule 3 about 99% visitor who visited root (/) page of ECX official website not visited about law and rules (rules.aspx) page.

Rule 4: [URL1=yes]: 2433 ==> [URL4=no]: 2350 <conf:(0.97)> lift:(1.04) lev:(0.02) conv:(1.94)
According to rule 4 about 97% visitor who visited root (/) page of ECX official website not visited about commodities (commodities.aspx) page.

5.2.4 Pattern Discover and Analysis with Asia dataset

This following experiment result shows interesting patterns discovered from all Asia dataset of Weka 3.7.9 experiment with FP-Growth algorithm. For detail Weka output, see Appendix C (III). To discover association rule within all county dataset I have used a minimum support 0.05 and a minimum confidence 0.76.

Rule 1: [URL3=no]: 2379 ==> [URL33=no]: 2379 <conf:(1)> lift:(1) lev:(0) conv:(0)
According to rule 1 about 100% visitor who visited home (home.aspx) page of ECX website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 2: [URL7=yes]: 207 ==> [URL33=no]: 207 <conf:(1)> lift:(1) lev:(0) conv:(0)
According to rule 2 about 100% visitor who visited graph (graph.aspx) page of ECX official website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 3: [URL7=yes]: 207 ==> [URL33=no, URL3=no]: 175 <conf:(0.85)> lift:(0.92) lev:(-0.01) conv:(0.48)
According to rule 3 about 85% visitor who visited graph (graph.aspx) page of ECX website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf) and home (home.aspx) page.

Rule 4: [URL1=yes]: 1980 ==> [URL3=no]: 1900 <conf:(0.96)> lift:(1.04) lev:(0.03) conv:(1.88)
According to rule 4 about 96% visitor who visited root (/) page of ECX website not visit the home (home.aspx) page of the website.

Rule 5: [URL33=no, URL3=no]: 2379 ==> [URL1=yes]: 1900 <conf:(0.8)> lift:(1.04) lev:(0.03) conv:(1.15)

According to rule 5 about 80% visitor who not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf) and not visited home (home.aspx) page of ECX official website not visited the root (/) page.

5.2.5 Pattern Discover and Analysis with Europe dataset

This following experiment result shows interesting patterns discovered from all Europe dataset of Weka 3.7.9 experiment with FP-Growth algorithm. For detail Weka output, see Appendix C (IV). To discover association rule within all county dataset I have used a minimum support 0.05 and a minimum confidence 0.78.

Rule 1: [URL1=yes,URL3=yes]: 1790==>[URL33=no]: 1789<conf:(1)> lift:(1) lev:(0) conv:(1.52)

According to rule 1 about 100% visitor who visited root (/) page and home (home .aspx) page of ECX website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 2: [URL2=yes]: 2988 ==> [URL33=no]: 2985 <conf:(1)> lift:(1) lev:(0) conv:(1.27)

According to rule 2 about 100% visitor who visited historical price (historical price) page of ECX website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 3: [URL4=yes]: 1058 ==> [URL33=no]: 1051 <conf:(0.99)> lift:(1) lev:(0) conv:(0.22)

According to rule 3 about 99% visitor who visited commodities (commodities.aspx) page of ECX website not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf).

Rule 4: [URL3=yes]: 1982 ==> [URL1=yes]: 1790 <conf:(0.9)> lift:(2.33) lev:(0.12) conv:(6.29).

According to rule 4 about 90% visitor who visited home (home.aspx) of ECX website not visited the root page of the website.

Rule 5: [URL33=no, URL3=yes]: 1981 ==> [URL1=yes]: 1789 <conf:(0.9)> lift:(2.33) lev:(0.12) conv:(6.29)

According to rule 5 about 90% visitor who not accessed maize daily price document (/downloads/contracts/Grain/ECXMaizeContract.pdf) and home (home.aspx) page of ECX official website not visited home (home.aspx) page.

5.2.6 Pattern Discover and Analysis with North America dataset

This following experiment result interesting patterns discovered from all Europe dataset of Weka 3.7.9 experiment with FP-Growth algorithm. For detail Weka output, see Appendix C (V). To discover association rule within all county dataset I have used a minimum support 0.05 and a minimum confidence 0.75.

Rule 1: [URL6=yes]: 161 ==> [URL9=no]: 159 <conf:(0.99)> lift:(1.01) lev:(0) conv:(1.32)
According to rule 1 about 99% visitor who visited company profile (campanyprofile.aspx) page of ECX website not accessed coffee daily price document.

Rule 2: [URL2=yes]: 680 ==> [URL9=no]: 670 <conf:(0.99)> lift:(1.01) lev:(0) conv:(1.52)
According to rule 2 about 99% visitor who visited historical price (historicalprice.aspx) page of ECX website not accessed coffee daily price document.

Rule 3: [URL4=yes]: 277 ==> [URL9=no]: 257 <conf:(0.93)> lift:(0.95) lev:(0) conv:(0.32)
According to rule 3 about 93% visitor who visited commodities (commodities.aspx) page of ECX website not accessed coffee daily price document (/downloads/contracts/Coffee/CoffeeContracts.pdf).

Rule 4: [URL3=yes]: 207 ==> [URL9=no, URL1=yes]: 165 <conf:(0.8)> lift:(1.48) lev:(0.02) conv:(2.22)
According to rule 4 about 80% visitor who visited home (home.aspx) page of ECX website not accessed coffee daily price document (/downloads/contracts/Coffee/CoffeeContracts.pdf) and visited root (/) page.

5.7 Discussion and Interpretation

This section gives a brief discussion about the dataset used for the study and the result obtained from web usage mining technique. The dataset of this study is the Web log files that taken from Web server log of ECX website. The Web log file is prepared in three main phases such as data cleaning, session identification and transaction identification. From 2,506,766 original Web log record 83,278 records selected by removing other records, which is irrelevant for web usage mining, with the consultation of experts from ECX and by reviewing related literatures. The 17816 session identified with 20 minute timeout, which is suggested by scholars. Then, transaction is identified with 40 items (page viewed) selected from the total of 71, which is selected by consultation of ECX experts. Items/pages selected, which are frequently accessed and page that are critical or page that should be included for pattern discover by the expert's points of view.

The web usage mining result shows that, most frequently visited page of ECX website is root page, and historical price page. Because, the root page is the default entry page of the website. It implies that, most of the visitor use (www.ecx.com.et) to access the website. Historical price page is the second most frequently accessed page, because it contain the previous price of each product, which helps for market reference. The visitor's path of ECX website shows that most of the visitor not travel/navigate from one page to the other page of the website, they exit immediately from the entry page without navigating through the other page of the website. This implies that the visitors of ECX website not accessed the required/complete information from the website. According to the website developers, visitors of the website should navigate through from one page of the website to the other to access expected information from the website. Seasame price document and coffee price document of the website have page loading problem, which display partial content of the website. This implies that user cannot access this page within short period of time after hitting the link of the pages. This is because of their high size of page/file content. The root page of the website also cause for internal server error. This implies that there is server side programming error. Graph and historical price page also cases for server side error of the website.

Most of the visitors of the website use Android operating system. This implies that most of the visitors of the website use mobile device to access the website. However, the current website of ECX is not compatible with mobile device. This is one of the challenge that limit users to navigate/travel through the website. Some visitors of the website used old version of Internet Explorer browser, which are not display the menu of ECX website. This is also the other challenges for the visitors of ECX website that limit users who access the website by menu.

According to, the knowledge requirement of ECX market data department, about counties who are frequently visited the ECX website as well as the association between these countries and the type of product they are frequently search/look. According, the top visitor countries of ECX website are Iceland, Ethiopia, United States, China, and United Kingdom. Visitor from Iceland search for coffee, White pea, Maize and Seasame product; United states search for seasame, coffee, maize and white pea; United Kingdom search for seasame and coffee; China search for white pea seasame and coffee. It indicates that, those countries are either interested about Ethiopia agricultural product such as coffee, seasame, white pea or they visit for their competitive advantage. So, investigation of the current customers of ECX and the top online visitors of ECX website are important to identify the objective of top visitor countries.

In this study, two association rule mining algorithm such as priori and FP-Growth algorithm tested with preprocessed web log file of ECX website. The FP-Growth algorithm is have better performance in terms of running time and small size requirement than Apriori algorithm. The Apriori algorithm takes long running time and display memory heap size error. Due to its better performance capacity of FP-Growth algorithm is selected for association rule discovery. FP-Growth algorithm conducted with all country dataset, Africa dataset, Asia dataset, Europe dataset and North America dataset. The patterns that discovered from all country dataset shows that visitor who visited about root page, historical price, commodities page or market data graph page not visited about maize price document. It indicates the most of the visitor who visit ECX website not accessed about maize price document. From Africa dataset who visited other page of the website such as root page, historical price, commodities page or market data graph page not visited company profile page.

This is because of most of Africa visitors are from Ethiopia who is familiar for ECX company profile. The pattern discovered from Asia and Europe dataset are similar to the rule discovered from all countries dataset. The pattern discovered from North American shows that visitor who visited about root page, home page and company profile page or commodities page are not accessed about daily price document of coffee product. This indicates that coffee product daily price document have weak association with the other page or user miss the link of coffee daily price page of ECX official website. This indicates that most of the visitor who access ECX website not accessed about coffee price document of the website.

Generally the patterns discovered from all dataset shows that ECX website pages are not accessed together by the visitor except home page and root page. This indicates that visitors of ECX website who access one page of the website either not interested to visit the other page of the website or can't accessed. This is because of the poor structure design of the website and by the type of device and browser they are used to visit the website.

5.8 Strength of the study

The strength of this research compared to other similar research, in this research user access patterns discovered and analyzed based on continents, countries and cities of ECX website visitor's. This helps to identify where come from the target customer, to identify similarity and difference of user navigation behavior of visitor across continent. The patterns of most frequently visitor countries and the type of product they are frequently looks/search also identified, this helps to identify each countries product type interest. The most frequent error type that occur during user navigation of ECX website and the page that cases for each frequent error type are identified. The user's access pattern of the website also discovered and analyzed based users country and continent.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

In this study, the researcher attempted to apply statistical and data mining techniques on access log files of ECX official website to explore user profile and navigation behavior. The study was conducted in four phases, such as raw log data collection, data preparation, pattern discovery and pattern analysis.

The raw log data preparation was a challenging and time taking task of the research. Since Web log files contain non-human requests such as robots, indexers, and spiders, requests such as images, icons, style sheets which are not important for analysis purpose. For the purpose of finding rules or patterns in web usage data, the researcher only interested in the pages or documents the visitors visit when traversing a website. This process deals with performing accuracy check; transforming the data into a session file; and finally structuring the data as per the input requirements. After preprocessing was completed, several experiments were conducted using statistical and data mining technique in order to extract interesting rules and patterns from the user web log record of ECX official website.

Web log storming statistical analyzer was applied on the preprocessed log to explore general statistics, about most frequently visited pages, more frequent entry and exit pages, most visitor countries and cities, most visitors country of coffee, seasame, wheat, maize, and white pea product, common errors encountered, page that frequently error occurred, types of operating system visitors used, users' visiting time by day. FPGrowth association rule algorithm of data mining technique using Weka 3.7.9 tools was used to discover association or correlations between web pages accessed together during a server session of the website. The FPGrowth algorithm conducted by categorizing the dataset of preprocessed log record into country level and continent level such as Africa, Asia, Europe, North America.

According to the statistical analysis of this study, half of the user of ECX official website start navigation at the root page (www.ecx.com.et/) of the website, others are directly access the page they want to visit with the help of search engine. This indicates that most links (content) of the website are optimized for search engines. More than two third (2/3) of visitors of ECX website exit from entry page without visiting the other page, others navigate other page before exiting from the entry page. Among the visitor the exit from the entry page most of visitors are, who starts to visit at root page and historical price of the website.

Visitor's country wise show that most of the visitor are from Iceland, Ethiopia, United States, China, and United Kingdom. Hofuoborgarsvaoio city from Iceland, Addis Ababa from Ethiopia, London from United Kingdom, Virginia, California, Ohio, Washington and New York from United States, and Beijing, Shandong, Shanghai, Henan, and Guangdong from China are city of most visitor countries. It indicates that, these countries are the most active countries who visit ECX Website frequently. Most of the users of the website also analyzed based on the type of product they visited and visitor countries. Most frequent visitors by country and product type shows that visitor from Iceland search for coffee, White pea, Maize and Seasame product; United states search for seasame, coffee, maize and white pea; United Kingdom search for seasame and coffee; China search for white pea Seasame and coffee.

The most frequent error response type that occur when the user navigating through ECX website are partial content (error type 206) and internal server error (error type 500). The most common page that cause for partial content is the root (/) page, seasame daily price information /downloads/contracts/sesame/ecxsesame.pdf) and coffee daily price document (/downloads/contracts/Coffee/CoffeeContracts.pdf) of the website. The most common page that cause for internal server error are root (/) page, home (home.aspx) page and graph (graph.aspx) of the website. The type of operating system that user used most frequently to access the website is Android application. This indicates that most of the visitor use mobile device to access the website of ECX. Windows 7 operating system is most frequently used operating system by the visitor next to Android application. From visitors hit by hours of the day, the highest hits of the website occur at 12AM, 12PM, 1AM, 1PM, 1PM, 2AM, 2PM and 3PM of the day.

The patterns discovery from all dataset of association rule mining technique shows that, most of visitor who visit root page and home page of ECX official website were not visited maize product daily price document. Most of visitors who visited the root page also visit the home page of the website. Most of visitors from Africa who visited root page, home page as well as historical price are not visit about rule and regulation of ECX website. This indicates that there is weak association between rules and regulations page with root, home and historical price page or visitor who visited root, home, and historical price page are not interested to visit rule and regulation page of the website. Generally, the statistical analysis and data mining association rule discovery shows most of the ECX website page are not accessed together except root page and home page of the website. On the other hand visitors hit the same page or link of page frequently. This indicates that there are different linked pages with the same content and link URL.

This research, attempted to answer the stated research questions, address the problem that stated in the statement of the problem and has achieved the objectives of the study. In this research, the web server log data that is used for pattern discovery prepared effectively by using different tools such as log file viewer to remove irrelevant record which have file extension such as .jpg, .png, .gif, .axd, .svc, MS- excel to remove log record accessed by ECX web admin and staff, record that is not accessed by human such as spiders, crawlers, robots cleaned by using web log storming. User patterns of the website discovered with different parameter such as by using visitor continent, country, city, the page that frequently visited by visitor, the type of product they have looked, the tool they used to visit the website, the type error of problem that visitor faced during navigation of the website, and the page that cause for error response.

In this study, transaction are identified based on travel path transaction identification approach. The travel path is a combination of auxiliary and content pages accessed by a user. Auxiliary pages are pages that used to facilitate the browsing of a user while searching for information. Content only transactions are only content pages which are user want to visit (user interest). Transaction identification by using content only approach such as transaction identification by reference length will give better result.

6.2 Recommendation

ECX aims to promote and advertise the Ethiopian agricultural product specifically coffee, seasame, white pea, wheat and maize to the global market economy. To achieve this, it is important to understand the customer need, preference, behavior and profile.

Accordingly the following recommendation are forwarded.

- ✓ As shown in the statistical analysis most of the user access the root page of the website by directly entering the website address, www.ecx.com.et or by clicking from search engine search result. The other page of the website not optimized for search engine and no referrer to the website except Google. The other page of the website are recommended to have optimized for search engine and integrated with other referrer such as Facebook, Twitter, and Whatsapp.
- ✓ Performance of website to responses for request of the visitor should be improved. According to this study the root page of the website not display the full content of the website for client side user request. According to the researcher, the main factor of this problem was the root page of the website are with more of image or graphical content.
- ✓ Coffee price document and seasame price document also not download easily, because of their large size of pdf file. To address this problem a researcher recommend for website developer to upgrade the website with database system for easily accessing of specific product cost based on users need rather than downloading the whole product price pdf document rather than collecting all product price document into one pdf document.
- ✓ Attractiveness of the site depends on its reasonable design of content and organizational structure. However, the results of web usage mining shows weak association between pages of the website, most of the visitor that visited one page may not access other page of the website. It implies that related page of website are not well organized. Therefore further investigation about web structure mining is recommended to improve the website.
- ✓ Most of user of ECX website use mobile device, therefore the web designer should consider the design of website based on mobile application compatibility in addition to Windows compatibility.
- ✓ For the better competitive advantage in the marketplace and effective marketing strategies ECX needs to work based on the customer demographic profile (country, city) that found from this study.

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

- ✓ In this study, the web server access log is used to discover usage patterns of ECX website. Future researcher need to consider proxy server and client server log file to discover a better usage pattern of the website.
- ✓ Identifying users by proxies and client cookies is one of the potential research area that improve the performance of the web usage mining study. Hence, feature researcher better to use this technique.
- ✓ Feature research work on personalization of the website for the effective and efficient handling of web customers.

REFERENCE

[1] Y.Raju, B.Prashanth Kumar, D.Suresh Babu, "An accomplishment of web personalization using web mining techniques," *International Journal of Computer Science and Information Technologies,* vol. 2, 2011.

[2] SaiMing Au, "A study of application of web mining for e-commerce: tools and methodology," *International Journal of the Computer, the Internet and Management,* vol. 10, p. 1 – 14, 2002.

[3] Pradnya Purandare, "Web mining: a key to improve business on web," *IADIS European Conference Data Mining,* 2008.

[4] Belsare Satish, Patil Sunil, "Study and Evaluation of user's behavior in e-commerce Using Data mining," *Research Journal of Recent Sciences,* vol. 1, 2012.

[5] J. Srivastava et al, "WEBKDD 2002 – Web Mining for Usage Patterns & Profiles," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* vol. 4, no. 2, p. 125, 202.

[6] Magdalini Eirinaki, Michalis Vazirgiannis, "Web Mining for Web Personalization," *ACM Transactions on Internet Technology,* vol. 3, 2003.

[7] Mitku Bade, "about.aspx," ecx, 23 2 2008. [Online]. Available: www.ecx.com.et. [Accessed 10 2014].

[8] Anteneh Degefe, Interviewee, *Corporate communication specialist of Ethiopia Commodity Exchange.* [Interview]. 14 10 2014.

[9] Irene S. Y. Kwan, Joseph Fong, and H. K. Wong, "An e-customer behavior model with online analytical mining for internet marketing planning," *Decision Support Systems,* vol. 41, pp. 189-204, 2005.

[10] Yan Wang, "Web mining and knowledge discovery of usage patterns," 2000.

[11] Jebaraj Ratnakumar, "An implementation of web personalization using web mining techniques," *Journal of Theoretical and Applied Information Technology,* vol. 19, no. 441-452, p. 2005, 2005.

[12] Bernie Lydon, Tom Fennell, "Web usability: its impact on human factor and consumer," *Human-Computer Interaction: Theory and Practice,* 2003.

[13] Jiawei Han, Micheline Kamber, Data mining: concepts and techniques, second edition ed., San Francisco: Morgan Kaufmann Publishers, 2006.

[14] Anand Sharma, "Web usage mining: Data preprocessing pattern discovery and pattern analysis on the RIT web data," MSc Project, Rochester Institute of Technology, Rochester, NY, USA, 2008.

[15] Raymond Kosla, Hendrik Blockeel, "Web Mining Research : A Survey," *SIGKDD Explorations,* vol. 2, no. 1, pp. 1-2, 2000.

[16] R. Cooley et. al, "Data preparation for mining world wide web browsing patterns," *Journal of Knowledge and Information Systems,* vol. 1, no. 1, pp. 15-32, 1999.

[17] R. Shukla, S. Silakari, K. Chande, "Existing trends and techniques for web personalization," *IJCSI International Journal of Computer Science Issues,* vol. 9, no. 4, 2012.

[18] M. Spliopoulou et. al, "An overview of preprcessing of web log files for web usage mining," *Journal of Theoretical and Applied Information Technology,* vol. 34, no. 1, 2011.

[19] M. Spiliopoulou et. al, "A framework for the evaluation of session reconstruction heuristics in web usage analysis," *Informs journal on computing,* vol. 15, no. 2, 2003.

[20] Ananthi Sheshasaayee, S. Padmaja, "Web Usage Mining and Internet User Behaviour in Web: A Survey," *International journal of engineering sciences & research,* 2014.

[21] O. Etzioni, " The World-Wide Web: quagmire or gold mine," *Communications of the ACM,* 1996.

[22] A. Rastogi et. al, "Web mining: a comparative study," International Journal of *Computational,* vol. 1, no. 2, pp. 325-331, 2012.

[23] H. Elhiber and A. Abraham, "Access Patterns in Web Log Data: A Review," *Journal of Network and Innovative Computing,* vol. 1, pp. 348-355, 2013.

[24] N. Cyriac et. al, "Web Personalization," *Computer Technology & Applications,* vol. 5, pp. 242-247, 2014.

[25] N. Lakshmi et.al, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis," *International Journal of Innovative Technology and Exploring Engineering,* vol. 2, no. 4, pp. 2278-3075, 2013.

[26] Zidrina Pabarskaite, "Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining," in *24th Int. Conf. information Technology Interfaces /TI 2002*, Cavtat, Croatia, 2002.

[27] N. Goel and C.K. Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Too," *International Journal of Computer Applications,* vol. Volume 62, 2013.

[28] D. Pierrakos et. al, "Web Usage Mining as a Tool for Personalization: A Survey," *ACM, User Modeling and User-Adapted Interaction,* 2003.

[29] B. Mobasher et. al, "Discovery and evaluation of aggregate usage profiles for web personalizatio," *Data Mining and Knowledge Discovery*, 2002.

[30] R. Agrawal, R.Srikant, "Fast algorithms for mining association rules," in *In: Proc the International Conference on Very Large Data Bases*, 1994.

[31] M. Girotra, K. Nagpal, "Comparative Survey on Association Rule Mining Algorithms," *International Journal of Computer Applications,* 2003.

[32] B. Mobasher, "Data Mining for Web Personalization," *The Adaptive Web,* p. 90–135, 2007.

[33] M. Jafari et. al, "Extracting Users'Navigational Behavior from Web Log Data: a Survey," *Journal of Computer Sciences and Applications,* vol. 1, no. 3, pp. 39-45, 2013.

[34] Magdalini Eirinaki, Michalis Vazirgiannis, "A Web Mining approach for Personalized E-learning system," International journal of avanced computer science and application, vol. 5, no. 3, 2014.

[35] M. Perkowitz, O. Etzioni,, "Adaptive Web Sites: Conceptual Cluster Mining," in *AAAI '98/IAAI '98 Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence* , Menlo Park, CA, USA, 1997.

[36] G. Buchner et al., "Discovering internet markrting intellegence throgh web log mining," *MIMIC- Mining the internet for marketing intelligence,* 1998.

[37] Anand S. Lalani, "Data Mining of Web Access Logs," MSc Thesis, *Royal Melbourne Institute of Technology Melbourne,* Victoria, Australia, 2003.

[38] Wei Kong, "Exploring health website users by web mining," MSc Thesis, *Indiana University*, India 2012.

[39] Mekonnen , "Web usage pattern discovery using data mining and statistical analysis," *Msc Thesis, Addis Ababa University*, Addis Ababa, Ethiopia, 2009.

[40] Tadele Asitatikie, "Web usage pattern discovery: the case of Addis Ababa university official web site," *M.Sc. thesis, ,* Addis Ababa University, Addis Ababa, Ethiopia,, 2011.

[41] Awet Fiseha, "Web usage: Exploring navigational behavior of users, the case of official website of Addis Ababa University," *M.Sc thesis*, Addis Ababa Unvniversity, Adis Ababa Ethiopia, 2011.

[42] Getahun Negatu, "Web usage pattern discovery and analysis by region: the case of Ethiopian Airline official website," *M.Sc. Thesis*, Addis Ababa University, Addis Ababa, Ethiopia, 2014.

[43] B.Mobasher et al., "Automatic personalization based on web usage mining," *Communications of the ACM,* 2000.

[44] R. Agrawal et al., "Mining association between sets of items in massive database," *In: Proc the ACM-SIGMOD International Conference on Management of Data,* 1993.

[45] Jaiwel Han, Jian Pei, and Yiwen Yin, "Mining Frequent Patterns without Candidate Generation," *ACM SIGMOD international conference on Management of data*, ACM New York, USA, 2000.

[46] G. Srinivasa et.al, "Implementation of association rules in web," *International Journal of Research In Science & Engineering,* vol. 1, no. 2, 2011.

[47] Pang-Ning Tan, Vipin Kumar , "Discovery of Web Robot Sessions Based on Their Navigational Patterns," *Intelligent Technologies for Information Analysis,* pp. 193-222, 2004.

[48] Iyob Tilahun, Interviewee, *Pattern analysis of web usage mining.* [Interview]. 1 6 2015.

Appendices

Appendix A: list of selected attribute for association rule discovery

URL Code	URL Description
URL1	/
URL2	/Home.aspx
URL3	/HistoricalPrice.aspx
URL4	/commodities.aspx
URL5	/MktDataGraph.aspx
URL6	/CompanyProfile.aspx
URL7	/Graph.aspx
URL8	/Operations.aspx
URL9	/downloads/contracts/Coffee/CoffeeContracts.pdf
URL10	/membership.aspx
URL11	/NewsAndEvents.aspx
URL12	/career.aspx
URL13	/CoffeeBulletinBoard.aspx
URL14	/downloads/contracts/Sesame/ECXSesame.pdf
URL15	/KnowledgeCenter.aspx
URL16	/contactUs.aspx
URL17	/downloads/Coffee.xls
URL18	/downloads/whatisnew.pdf
URL19	/downloads/Market/Coffee.pdf
URL20	/rules.aspx
URL21	/DSTIntroduction.aspx
URL22	/downloads/Newsletter.pdf
URL23	/GrainBulletinBoard.aspx
URL24	/tradingoperations.aspx
URL25	/downloads/contracts/Pulse/ECXWHITEPEABEANS.pdf
URL26	/downloads/Briefs.pdf
URL27	/dstregisteredbuyers.aspx
URL28	/downloads/Oil+ Seed.xls
URL29	/dstcatalogresults.aspx
URL30	/ReferenceMarket.aspx
URL31	/downloads/Contracts/Grain/wheat.pdf
URL32	/downloads/Articles.pdf
URL33	/downloads/contracts/Grain/ECXMaizeContract.pdf
URL34	/downloads/Brochuer.pdf
URL35	/DSTSampleRequestForm.aspx
URL36	/blog.aspx
URL37	/DSTCatalogResultsHistory.aspx
URL38	/DSTCatalogResults.aspx
URL39	/downloads/PressRelease.pdf

URL40	/CofeeBlog.aspx

Appendix B: Sample transaction identification MSQL statement

```
UPDATE main_table_session m, root_table_session r
SET m.URL2= 'yes'
where m.dates = r.dates
and m.host = r.host

UPDATE main_table_session m, historicalprice_table_session h
SET m.URL3 = 'yes'
where m.dates = h.dates
and m.host = h.host

UPDATE main_table_session m, home_table_session h
SET m.URL4 = 'yes'
where m.dates = h.dates
and m.host = h.host


UPDATE main_table_session m, commodities_session h
SET m.URL5 = 'yes'
where m.dates = h.dates
and m.host = h.host

UPDATE main_table_session m, marketdata_session k
SET m.URL6 = 'yes'
where m.dates = h.dates
and m.host = h.host
```

Appendix C: Weka Association rule discovery outputs

Appendix C (I)

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.05 -S -1.0 -c -1

Relation: main report total continent all URL

Instances: 17815

Attributes: 22

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

Apriori

=======

Minimum support: 0.95 (16924 instances)

Minimum metric <confidence>: 0.75

Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L (1): 17

Size of set of large itemsets L (2): 105

Size of set of large itemsets L (3): 452

Size of set of large itemsets L (4): 1238

Size of set of large itemsets L (5): 1900

Size of set of large itemsets L (6): 1422

Size of set of large itemsets L (7): 449

Size of set of large itemsets L (8): 41

Best rules found:

1. URL9=no URL14=no URL24=no URL28=no 17194 ==> URL25=no 17136 conf:(1)
2. URL9=no URL14=no URL19=no URL24=no URL28=no 17089 ==> URL25=no 17031 conf:(1)
3. URL9=no URL14=no URL24=no URL28=no 17083 ==> URL25=no 17025 conf:(1)
4. URL9=no URL14=no URL15=no URL24=no URL28=no 17070 ==> URL25=no 17012 conf:(1)
5. URL9=no URL14=no URL16=no URL24=no URL28=no 17062 ==> URL25=no 17004 conf:(1)
6. URL9=no URL14=no URL17=no URL24=no URL28=no 17056 ==> URL25=no 16998 conf:(1)

7. URL9=no URL13=no URL14=no URL24=no URL28=no 17026 ==> URL25=no 16968 conf:(1)
8. URL12=no URL13=no URL15=no URL19=no URL24=no URL25=no 17016 ==> URL28=no 16958 conf:(1)
9. URL9=no URL12=no URL14=no URL24=no URL28=no 17011 ==> URL25=no 16953 conf:(1)
10. URL13=no URL15=no URL19=no URL24=no URL25=no 17167 ==> URL28=no 17108 conf:(1)
11. URL12=no URL15=no URL19=no URL24=no URL25=no 17129 ==> URL28=no 17070 conf:(1)
12. URL12=no URL13=no URL15=no URL19=no URL25=no 17093 ==> URL28=no 17034 conf:(1)
13. URL11=no URL15=no URL19=no URL24=no URL25=no 17080 ==> URL28=no 17021 conf:(1)
14. URL13=no URL15=no URL16=no URL19=no URL24=no URL25=no 17061 ==> URL28=no 17002 conf:(1)
15. URL11=no URL13=no URL15=no URL19=no URL25=no 17049 ==> URL28=no 16990 conf:(1)
16. URL12=no URL15=no URL16=no URL19=no URL24=no URL25=no 17042 ==> URL28=no 16983 conf:(1)
17. URL13=no URL15=no URL17=no URL19=no URL24=no URL25=no 17035 ==> URL28=no 16976 conf:(1)
18. URL13=no URL14=no URL15=no URL19=no URL24=no URL25=no 17018 ==> URL28=no 16959 conf:(1)
19. URL11=no URL12=no URL15=no URL19=no URL25=no 17010 ==> URL28=no 16951 conf:(1)
20. URL12=no URL13=no URL15=no URL16=no URL19=no URL25=no 17005 ==> URL28=no 16946 conf:(1)

Appendix C (II)

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 11 -T 0 -C 0.77 -D 0.05 -U 1.0 -M 0.05

Relation: country dataset

Instances: 17815

Attributes: 40

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

FPGrowth found 11 rules (displaying top 11)

1. [URL1=yes]: 9314 ==> [URL33=no]: 9304 <conf:(1)> lift:(1) lev:(0) conv:(2.09)
2. [URL2=yes]: 4269 ==> [URL33=no]: 4263 <conf:(1)> lift:(1) lev:(0) conv:(1.51)
3. [URL3=yes]: 2888 ==> [URL33=no]: 2883 <conf:(1)> lift:(1) lev:(0) conv:(1.19)
4. [URL1=yes, URL3=yes]: 2244 ==> [URL33=no]: 2240 <conf:(1)> lift:(1) lev:(0) conv:(1.11)
5. [URL5=yes]: 1053 ==> [URL33=no]: 1050 <conf:(1)> lift:(1) lev:(0) conv:(0.65)
6. [URL7=yes]: 868 ==> [URL33=no]: 863 <conf:(0.99)> lift:(1) lev:(0) conv:(0.36)
7. [URL6=yes]: 889 ==> [URL33=no]: 882 <conf:(0.99)> lift:(0.99) lev:(0) conv:(0.27)
8. [URL4=yes]: 1713 ==> [URL33=no]: 1698 <conf:(0.99)> lift:(0.99) lev:(0) conv:(0.26)
9. [URL3=yes]: 2888 ==> [URL1=yes]: 2244 <conf:(0.78)> lift:(1.49) lev:(0.04) conv:(2.14)
10. [URL33=no, URL3=yes]: 2883 ==> [URL1=yes]: 2240 <conf:(0.78)> lift:(1.49) lev:(0.04) conv:(2.14)
11. [URL3=yes]: 2888 ==> [URL33=no, URL1=yes]: 2240 <conf:(0.78)> lift:(1.49) lev:(0.04) conv:(2.13)

Appendix C (III)

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 17 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1

Relation: Africa dataset

Instances: 3723

Attributes: 40

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

FPGrowth found 17 rules (displaying top 17)

1. [URL3=yes]: 500 ==> [URL20=no]: 500 <conf:(1)> lift:(1.01) lev:(0) conv:(2.69)

2. [URL4=no, URL3=yes]: 465 ==> [URL20=no]: 465 <conf:(1)> lift:(1.01) lev:(0) conv:(2.5)

3. [URL2=yes]: 495 ==> [URL20=no]: 494 <conf:(1)> lift:(1) lev:(0) conv:(1.33)

4. [URL4=no, URL2=yes]: 454 ==> [URL20=no]: 453 <conf:(1)> lift:(1) lev:(0) conv:(1.22)

5. [URL4=no]: 3473 ==> [URL20=no]: 3456 <conf:(1)> lift:(1) lev:(0) conv:(1.04)

6. [URL4=no, URL1=yes]: 2350 ==> [URL20=no]: 2338 <conf:(0.99)> lift:(1) lev:(0) conv:(0.97)

7. [URL1=yes]: 2433 ==> [URL20=no]: 2420 <conf:(0.99)> lift:(1) lev:(0) conv:(0.93)
8. [URL20=no, URL1=yes]: 2420 ==> [URL4=no]: 2338 <conf:(0.97)> lift:(1.04) lev:(0.02) conv:(1.96)
9. [URL1=yes]: 2433 ==> [URL4=no]: 2350 <conf:(0.97)> lift:(1.04) lev:(0.02) conv:(1.94)
10. [URL1=yes]: 2433 ==> [URL20=no, URL4=no]: 2338 <conf:(0.96)> lift:(1.04) lev:(0.02) conv:(1.82)
11. [URL20=no]: 3703 ==> [URL4=no]: 3456 <conf:(0.93)> lift:(1) lev:(0) conv:(1)
12. [URL3=yes]: 500 ==> [URL4=no]: 465 <conf:(0.93)> lift:(1) lev:(0) conv:(0.93)

13. [URL3=yes]: 500 ==> [URL20=no, URL4=no]: 465 <conf:(0.93)> lift:(1) lev:(0) conv:(1)

14. [URL20=no, URL3=yes]: 500 ==> [URL4=no]: 465 <conf:(0.93)> lift:(1) lev:(0) conv:(0.93)

15. [URL2=yes]: 495 ==> [URL4=no]: 454 <conf:(0.92)> lift:(0.98) lev:(0) conv:(0.79)

16. [URL20=no, URL2=yes]: 494 ==> [URL4=no]: 453 <conf:(0.92)> lift:(0.98) lev:(0) conv:(0.79)

17. [URL2=yes]: 495 ==> [URL20=no, URL4=no]: 453 <conf:(0.92)> lift:(0.99) lev:(0) conv:(0.83)

Appendix C (IV)

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 17 -T 0 -C 0.76 -D 0.05 -U 1.0 -M 0.05

Relation: Asia dataset

Instances: 2577

Attributes: 40

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

FPGrowth found 16 rules (displaying top 16)

1. [URL3=no]: 2379 ==> [URL33=no]: 2379 <conf:(1)> lift:(1) lev:(0) conv:(0)

2. [URL1=yes]: 1980 ==> [URL33=no]: 1980 <conf:(1)> lift:(1) lev:(0) conv:(0)

3. [URL7=yes]: 207 ==> [URL33=no]: 207 <conf:(1)> lift:(1) lev:(0) conv:(0)

4. [URL3=no, URL1=yes]: 1900 ==> [URL33=no]: 1900 <conf:(1)> lift:(1) lev:(0) conv:(0)

5. [URL3=no, URL7=yes]: 175 ==> [URL33=no]: 175 <conf:(1)> lift:(1) lev:(0) conv:(0)

6. [URL1=yes]: 1980 ==> [URL3=no]: 1900 <conf:(0.96)> lift:(1.04) lev:(0.03) conv:(1.88)

7. [URL1=yes]: 1980 ==> [URL33=no, URL3=no]: 1900 <conf:(0.96)> lift:(1.04) lev:(0.03) conv:(1.88)

8. [URL33=no, URL1=yes]: 1980 ==> [URL3=no]: 1900 <conf:(0.96)> lift:(1.04) lev:(0.03) conv:(1.88)

9. [URL33=no]: 2577 ==> [URL3=no]: 2379 <conf:(0.92)> lift:(1) lev:(0) conv:(0.99)

10. [URL7=yes]: 207 ==> [URL3=no]: 175 <conf:(0.85)> lift:(0.92) lev:(-0.01) conv:(0.48)

11. [URL7=yes]: 207 ==> [URL33=no, URL3=no]: 175 <conf:(0.85)> lift:(0.92) lev:(-0.01) conv:(0.48)

12. [URL33=no, URL7=yes]: 207 ==> [URL3=no]: 175 <conf:(0.85)> lift:(0.92) lev:(-0.01) conv:(0.48)

13. [URL3=no]: 2379 ==> [URL1=yes]: 1900 <conf:(0.8)> lift:(1.04) lev:(0.03) conv:(1.15)

14. [URL3=no]: 2379 ==> [URL33=no, URL1=yes]: 1900 <conf:(0.8)> lift:(1.04) lev:(0.03) conv:(1.15)

15. [URL33=no, URL3=no]: 2379 ==> [URL1=yes]: 1900 <conf:(0.8)> lift:(1.04) lev:(0.03) conv:(1.15)

16. [URL33=no]: 2577 ==> [URL1=yes]: 1980 <conf:(0.77)> lift:(1) lev:(0) conv:(1)

Appendix C (V)

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 17 -T 0 -C 0.78 -D 0.05 -U 1.0 -M 0.05

Relation: Europe dataset

Instances: 8824

Attributes: 40

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

FPGrowth found 10 rules (displaying top 10)

1. [URL3=yes]: 1982 ==> [URL33=no]: 1981 <conf:(1)> lift:(1) lev:(0) conv:(1.68)

2. [URL1=yes, URL3=yes]: 1790 ==> [URL33=no]: 1789 <conf:(1)> lift:(1) lev:(0) conv:(1.52)

3. [URL1=yes]: 3416 ==> [URL33=no]: 3413 <conf:(1)> lift:(1) lev:(0) conv:(1.45)

4. [URL2=yes]: 2988 ==> [URL33=no]: 2985 <conf:(1)> lift:(1) lev:(0) conv:(1.27)

5. [URL5=yes]: 721 ==> [URL33=no]: 720 <conf:(1)> lift:(1) lev:(0) conv:(0.61)

6. [URL7=yes]: 451 ==> [URL33=no]: 450 <conf:(1)> lift:(1) lev:(0) conv:(0.38)

7. [URL4=yes]: 1058 ==> [URL33=no]: 1051 <conf:(0.99)> lift:(1) lev:(0) conv:(0.22)

8. [URL3=yes]: 1982 ==> [URL1=yes]: 1790 <conf:(0.9)> lift:(2.33) lev:(0.12) conv:(6.29)

9. [URL33=no, URL3=yes]: 1981 ==> [URL1=yes]: 1789 <conf:(0.9)> lift:(2.33) lev:(0.12) conv:(6.29)

Appendix C (VI)

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.05

Relation: North America dataset

Instances: 2648

Attributes: 40

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

FPGrowth found 9 rules (displaying top 9)

1. [URL6=yes]: 161 ==> [URL9=no]: 159 <conf:(0.99)> lift:(1.01) lev:(0) conv:(1.32)

2. [URL2=yes]: 680 ==> [URL9=no]: 670 <conf:(0.99)> lift:(1.01) lev:(0) conv:(1.52)

3. [URL1=yes]: 1454 ==> [URL9=no]: 1429 <conf:(0.98)> lift:(1.01) lev:(0) conv:(1.37)

4. [URL3=yes]: 207 ==> [URL9=no]: 201 <conf:(0.97)> lift:(1) lev:(0) conv:(0.73)

5. [URL1=yes, URL3=yes]: 171 ==> [URL9=no]: 165 <conf:(0.96)> lift:(0.99) lev:(0) conv:(0.6)

6. [URL4=yes]: 277 ==> [URL9=no]: 257 <conf:(0.93)> lift:(0.95) lev:(0) conv:(0.32)

7. [URL3=yes]: 207 ==> [URL1=yes]: 171 <conf:(0.83)> lift:(1.5) lev:(0.02) conv:(2.52)

8. [URL9=no, URL3=yes]: 201 ==> [URL1=yes]: 165 <conf:(0.82)> lift:(1.5) lev:(0.02) conv:(2.45)

9. [URL3=yes]: 207 ==> [URL9=no, URL1=yes]: 165 <conf:(0.8)> lift:(1.48) lev:(0.02) conv:(2.22)