Machine Learning Methods for Mining Web Access Patterns

طرق التعلم بالآلة لتنقيب أنماط الوصول لصفحات الانترنت

BY

MOHAMMED HAMED AHMED ELHEBIR

A dissertation submitted in Partial fulfilment of the requirements for the degree of Doctor of Philosophy

in

(Computer Science)

Supervisor

Prof.Dr.Ajith Abraham

College of Computer Science and Information Technology

Sudan University for Science & Technology

May 2016

ACKNOWLEDGEMENTS

Many different people helped me with various parts of this thesis. Technical support, moral support, mental support. I want to thank my supervisor, Ajith Abraham, for his advice, support, and guidance. I would also like to thank my committee members, for the time and effort they spent to read and comment on my work. The most important “acknowledgement" goes to my parents and my wife; they both supported me, and pushed me when I needed motivation.

ABSTRACT

Web Usage Mining (WUM) can be defined as an application of data mining to extract the knowledge hidden in the Web log files, such as user access patterns from Web data. However, the structure of these log files does not present accurately a picture of the users' accesses to Web sites. The WUM process goes through three phases: data pre-processing, patterns discovery and pattern analysis. Data pre-processing can be used to filter and organize only the appropriate information before using Web mining algorithms on the log files' data for pattern discovery and analysis. Pattern analysis is the process of analyzing the access patterns of the log file. There are many efforts that have been conducted to accomplish the work of clustering and classification. These efforts resulted in the development of various tools and techniques, which can generate fixed reports from web logs; they typically do not allow ad-hoc analysis queries. Moreover, such tools cannot discover hidden patterns of access embedded in the access logs. In addition, they do not take an approach such as ensemble when used as machine learning tool. Moreover, the tools that have been developed for the analysis using data warehouse, populate the fact table directly from the web logs without prepressing step, which is necessary for data cleansing and enrichment. Therefore, the proposed work focuses on closing the gaps of the developed tools especially in the aforementioned issues. It takes the SUST log file as a case study. A preprocessing step is conducted before loading the log file data in a database table to make the data of the log file ready for accomplishing the mining and analysis task. The results obtained after the pre-processing were satisfactory and contained valuable and adequate information about the log files. In the mining process, the following tasks are curried out: clustering, rule based mining, and classification. In clustering, K-means clustering algorithm and Density based clustering are used to cluster web log based on the two types of clusters: user clusters and page clusters. It was found that the Density-based clustering has a better performance compared to K-means clustering with and without features selection. A priori algorithm is used for the task of rule-based mining to discover relationship among data. In this study the accuracy of ensemble models, which take advantage of groups of base learners is compared with the accuracy of several base classifiers. Stacking and Voting are used as an aggregation method to combine the results of the multiple base learners. The results show that the ensemble machine learning models using voting can significantly improve users sessions classification. To accomplish the task of pattern analysis, the log data is extracted transformed and loaded in a data warehouse. Online Analytical Processing (OLAP) is used to analyze the data that is loaded in the data warehouse. As for future work, there is a need to solve problems related to parallel processing, especially for large-scale data that resulted from the click streams of the growing usage of the web. Also due to the complexity of the dataset and the difficulty in understanding them, a visualization tools are needed to render the information related to these complex dataset in an easy and understandable way. In addition, an efficient way to analyze such large scale and complex data is needed, and it can be carried out through the use of parallel algorithms.

مستخلص

التنقيب عن استخدام الشبكة العنكبوتية يمكن تعريفه بأنه تطبيق لتنقيب البيانات لإستخلاص المعرفة المخفية في ملفات سجل الشبكة العنكبوتية مثل أنماط وصول المستخدمين من بيانات الشبكة العنكبوتية. ولكن تركيبة ملفات السجل هذة لا تقدم بدقة صورة وصول المستخدم إلى مواقع الشبكة العنكبوتية. وتسير عملية التنقيب عن استخدام الشبكة العنكبوتية عبر ثلاث مراحل هي: ما قبل معالجة البيانات، إكتشاف الأنماط وتحليل الأنماط. وتلعب مرحلة ماقبل معالجة البيانات دوراً مهماً في تقنية التنقيب عن استخدام الشبكة العنكبوتية. مرحلة ماقبل معالجة البيانات يمكن استخدامها في تنقية وتنظيم المعلومات الملائمة فقط قبل إستخدام خوارزميات تنقيب بيانات الشبكة العنكبوتية فى ملفات السجل من أجل اكتشاف وتحليل الأنماط. وتحليل الأنماط هي عملية تحليل أنماط الوصول في سجل الملفات. هنالك جهود عديدة بذلت لانجاز أعمال التصنيف والتجميع العنقودي. ونتج عن هذة الجهود تطوير أدوات وتقنيات مختلفة. ويمكن لهذة الأدوات أن تنتج تقارير ثابتة من سجل الشبكة العنكوبتية وهي تحديداً لا تتيح تحليلاً غير مجدول للاستعلامات. علاوة على ذلك، مثل هذة الأداوات لاتستطيع اكتشاف الأنماط المخفية في سجلات الوصول. أيضاً، هي لاتأخذ بالنظريات مثل التجميع عندما تستخدم كاداة تعليم آلية. علاوة على ذلك‘ فإن الأدوات التي تم تطويرها للتحليل باستخدام مستودعات البيانات تزود جدول الحقائق مباشرة من سجلات الشبكة العنكبوتية. ملء جدول الحقائق مباشرة يتجاوز خطوة ما قبل المعالجة والتي هي ضرورية لتصفية وإثراء المعلومات. لذلك‘ فإن العمل المقترح يركز على ردم هوة الأدوات المطورة خاصة في الموضوعات السابق ذكرها .وقد أخذّ ملف السجل لجامعة السودان للعلوم والتكنولوجيا كدراسة حالة. وتجرى خطوة ما قبل المعالجة لجعل بيانات ملف السجل جاهزة لانجاز مهمة التنقيب والتحليل. النتائج المتحصل عليها بعد خطوة ماقبل المعالجة كانت مرضية وأحتوت على معلومات قيمةعن ملفات السجل. في عملية التنقيب تجرى المهام التالية: التجميع العنقودي، التنقيب المبني على القواعد والتصنيف. في التجميع، تستخدم خوارزمية مركز ك والتجميع المبني على الكثافة، لتجميع سجل الشبكة استناداً على نوعين من التجميع: تجميع المستخدمين وتجميع الصفحات. وأضحت الدراسة أن التجميع المبني على الكثافة له أداء أحسن مقارنة بتجميع مركز ك باختيار وبدون إختيار الصفات. يُستخدم اللوغاريثم الاستدلالي في مهمة التنقيب المبني على القواعد لكشف العلاقة بين البيانات. في هذة الدراسة فإن دقة نماذج التجمبع والتي تستفيد من تجميع من قواعد المصنفات تمت مقارنتها مع دقة العديد من قواعد المصنفات. ويستخدم الحشد والتصويت كطريقة تجميع لدمج النتائج الخاصة بقواعد المصنفات المتعددة. وأظهرت النتائج أن نماذج آلة التجميع التعليمية باستخدام التصويت يمكنها أن تحسن تصنيف جلسات المستخدمين بصورة كبيرة. ولانجاز مهمة تحليل الأنماط، بيانات السجل المستخلصة يتم تحويلها وتحميلها فى مستودعات البيانات. وتُستخدم المعالجة التحليلية المباشرة فى تحليل البيانات فى مستودعات البيانات. وبالنسبة لأعمال المستقبل، هنالك حاجة لحل المشكلات المتعلقة بالمعالجة المتوازية خاصة للبيانات ذات القياسات الكبيرة والناتجة من تدفقات الاستخدام المتنامي للشبكة العنكبوتية. وأيضاً بسبب تعقيدات مجموعة البيانات والصعوبة في فهمها، يُحتاج إلى أدوات تصورية لتحليل المعلومات المتعلقة بمجموعات البيانات المعقدة بطريقة سهلة ومفهومة. أيضاً، هنالك حاجة إلى طريقة فعالة لتحليل مثل هذة البيانات ذات القياسات الكبيرة والمعقدة والتي يمكن تنفيذها عبر إستخدام اللوغاريثمات المتوازية.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure No | Page No.

LIST OF ABBREVIATIONS

BN: Bayes Net.

CMC: Combination of Multiple Classifiers.

CPU: Central processing unit.

GB: Gigabyte.

HTTP: Hyper Text Transfer Protocol.

KNN: K-Nearest Neighbor.

MAE: Mean absolute error.

MCC: Matthews correlation coefficient

NB: Naive Bayes.

NCSA: National Centre for Supercomputing Application.

OLAP: Online Analytical Processing.

PRC: Parameterized Rocchio Classifier.

RMSE: Root means squared error.

ROC: Receiver operating characteristic.

SQL: Structured Query Language.

SUST: Sudan University of Science and Technology.

WEKA: Waikato Environment for Knowledge Analysis.

WUM: Web Usage Mining.

WWW: World Wide Web.

CHAPTER ONE
INTRODUCTION

1.1 BACKGROUND

Nowadays, the Web has turned to be the largest information source available on the planet. It is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies an incredible amount of information, and also raises the complexity of how to deal with the information from different perspectives of users view. Users usually want to have an effective search tool for finding relevant information easily and precisely. Web Mining refers to the use of data mining techniques to automatically retrieve, extract and analyze information from web documents and services [1]. Web data mining can be divided into three different processes: “Web Content” mining, “Web Structure” mining and “Web Usage” mining [2] [3] [4]. Web Usage mining is a heavily researched area in the field of data mining. It can be described as the discovery and analysis of user access patterns through mining of log files and associated data from a particular website [5]. Although many areas and applications can be cited where Web Usage mining is useful, it can be said that the main idea behind Web Usage mining is to let users of a website use it with ease and effectively predict and recommend parts of the website to them based on their previous actions on the web site. A server log file is a file that automatically creates and maintains the activities performed on the server. This file is used to record each and every hit to a web site [6]. It maintains a history of page requests, also it helps in understanding how and when a website pages and application are being accessed by the web browser. It contains information such as the host IP address, proprietor, username, date, time, request method, status code, byte size, and referrer and user agent [7]. Generally, Web Usage mining consists of three processes: Data Pre-processing for the web log file, Pattern discovery and Pattern analysis [8] [9]. Since the origin web logs data sources are mixed with irrelevant information, data pre-processing acts as an important step to filter and organize only suitable and relevant information before presenting it to any web mining algorithm [10]. The data source affects the quality of the pre-processed data and in turn the pre-processed data influences the results of pattern discovery and pattern analysis directly [11] [12].

The Data Pre-processing process contains three sub-steps: Data Cleaning, User Identification, and Session Identification [13] [14]. After data pre-processing, the pattern discovery process should be applied. This process consists of different techniques derived from various fields such as statistics, machine-learning methods applied to the Web domain [15]. Several methods and techniques have already been developed for this process [16]. Some of the frequently used solutions are statistical analysis, clustering, association rules and classification.

Clustering is an unsupervised classification technique widely used for web usage mining with main objective to group a given collection of unlabelled objects into meaningful clusters [17]. In the Web Usage domain, there are two kinds of interesting clusters to be discovered: user clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Here we will briefly describe some techniques to discover patterns from processed data. Commonly used clustering algorithms are K-means and Density based clustering.

Feature selection is a term commonly used in data mining to describe the techniques available for reducing inputs to a manageable size for processing and analysis. Correlation based Feature Selection (CFS) measures correlation between nominal features, so numeric features are first discretized. It is also an effective dimensionality reduction technique and an essential pre-processing method to remove noise features [18].Association rule is one of the data mining tasks which can be used to discover relationship among data. Association rule identifies specific association among data and its techniques are generally applied to a set of transactions in a database. Since, amount of data handled is extremely large, current association rule techniques are trying to prune the search space according to support count [19]. Rules discovery finds common rules in the format A→B, meaning that, when page A is visited in a transaction, page B will also be visited in the same transaction. These rules may have different values of the confidence and support. There are two measurements in association rule mining are support and confidence. The support corresponds to the frequency of the pattern while confidence indicates rule’s strength.

Web usage mining involves with the application of data mining methods to discover user access patterns from web data, to better serve the needs of web-based applications. One of the most pattern discovery techniques used to extract knowledge from pre-processed data is classification. Given a training data set, the classification model was used to categorize the given training data set into attributes and the attributes were referred to as class. Classification can be performed using different techniques. Our goal was to predict the target class based on our source data (web log data) [20]. Our model takes into consideration the category type of classification in which the target attribute has only two possible variations: forenoon and afternoon.

Conventionally an individual classifier, such as K-Nearest Neighbor (KNN), Decision Tree (J48), Naive Bayes (NB) or BayesNet (BN) is trained on web log data set. Depending on the distribution of the patterns, it is possible that not all the patterns are learned well by an individual classifier. A classifier performs poorly on the test set under such scenarios. One of the most attractive topics in supervised machine learning is learning how to combine the predictions of multiple classifiers. This approach is known as ensembles of classifies in the supervised learning area. The motivation for doing this derives from the opportunity to obtain higher prediction accuracy, while treating classifiers as black boxes, without considering the details of their functionality. Meta-learning is a process of learning from learners (classifiers); the inputs of the meta-learner are the outputs of the base-classifiers (the basic classifiers). The goal of meta-learning ensemble is to induce a meta-model that combines base-classifier predictions into a single prediction. In order to create such ensemble, both the base-classifier and the meta-learner (meta-classifier) need to be trained. Since the meta-classifier(s) training requires an already trained base-classifier, these must be trained first. After the base-classifiers are trained, they are used to produce outputs (classifications), from which the meta level dataset is made. This dataset will be used for training the meta-classifier(s). In the prediction phase, when the ensemble is already trained, the base classifiers output their predictions to the meta-classifier(s) that combines them into a final prediction (classification).

Pattern analysis is the last step in the overall Web Usage mining process. In this research one of the widely used analytical tools and techniques is used to analyze the access patterns of the University of Sudan Science and Technology ' website. In this tool, we used the data warehouse to the extracted information from web log file in terms of dimensions and facts. The dimensions were represented by time, Protocol type, Users, Agent, IP address while the number of the accesses and the document size represented facts. Then an Online Analytical Processing (OLAP) was used to analyze the data in the data warehouse.

1.2 PROBLEM STATEMENT

Now a days as the number of internet users is growing exponentially, Click stream data are collected in volumes in an easy way and the real problem is how to analyze it and how to transform it into useful information and knowledge. The quantity of the web usage data to be analyzed and its low quality are the principal problems in WUM. One of the analysis tasks is to determine patterns and associations. There are already a number of existing tools to discover pattern. However, these tools behave like a black box such that their user does not know how exactly it generates the results.

1.3 RESEARCH OBJECTIVES

The main objective of this research is to conduct an in-depth analysis of the data kept in the access log files of the server that hosts the web pages of the Sudan University of Science and Technology (SUST) to facilitate the decision-making. The specific objectives are:

1- **To improve** the quality of data by eliminates irrelevant entries from dataset to be suitable for the pattern discovery and analysis.
2- **To group** access log file data with similar patterns together.
3- **To apply** an association rule mining to extract the knowledge or pattern from the access log file.
4- **To enhance** the performance of access log file classification by using a novel ensemble approach.
5- **To develop** an analytical tool to analyze the patterns of the access log files.

1.4 RESEARCH SCOPE

The main source of the data for web usage mining was the Web server logs from Sudan University of Science and Technology. The Web server logs each visit to each web page with possibly IP address, refereed page, access time, browser type and version, and accessed page link. The period of the data source of our experiment was from 7/Nov/2008 to 10/Dec/2009. The size of the log file in this period was 567 MB containing 291642 cases.

1.5 CONTRIBUTIONS OF THE RESEARCH

In this Section we briefly describe the contributions of this Research. First, we proposed an algorithm for pre-processing to reduce the large quantity of Web usage data available and, at the same time, to increase its quality by using regular expression to separate each line in the log file into different fields and then loading these data into a database table. After that we focus on methods that can be used for the tasks of data cleaning, user identification and session identification from Web log file.

Our second contribution is discovering the minority behaviours corresponding to the association patterns and provides interesting correlations, frequent patterns from a large pre-processed log file by implementing a hybrid concrete method using algorithms of clustering and association pattern mining.

Our third contribution is producing a novel ensemble approach to enhance the performance of access log file classification.

Finally, we designed and implemented an analytical tool that enables the user to easily and selectively extract and view data from different points of view. This tool allows for quick analysis of the all-possible interesting aggregates of the log file data by employing drag and drop operation. Also this tool enables the user to analyze the complex and large quantities of data in real time and answers questions such as what is the distribution of network traffic over time (hour of the day, day of the week, month of the year).

The research work done during the Ph.D. studies has been presented in the following papers:

- ✓ Access Patterns in Web Log Data: A Review, Journal of Network and Innovative Computing (**JNIC**), Volume 1 (2013) pp. 348-355.
- ✓ Data Pre-Processing of Web Server Logs for Mining users Access Patterns, International Journal of Engineering Sciences Paradigms and Researches (**IJESPR**)(Vol. 23, Issue 01) and (Publishing Month: August 2015) pp. 23-31.
- ✓ Web Log Data Analysis Using a Data Warehouse and OLAP, Journal of Network and Innovative Computing (**JNIC**), Volume 2 (2014) pp. 359-365.
- ✓ Discovering Web Server Logs Patterns Using Clustering and Association Rules Mining, Journal of Network and Innovative Computing (**JNIC**), Volume 3 (2015) pp. 159-167.
- ✓ A novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification, International Journal of Computer Information Systems and Industrial Management Applications(**IJCISIM)**, Volume 7 (2015) pp. 180-195.

1.6 RESEARCH ORGANIZATION

Following the introductory chapter, rest of the Chapters are organized as follows:

Chapter 2: Literature Review and Related Work

In this chapter we briefly describe the main concepts and definitions of the Web Mining followed by log file formats. Also we present review Data Pre-processing on Web Server Logs and different techniques in Pattern Discovery and Pattern Analysis. At the end of this chapter, we present the main related works in clustering and extracting sequential patterns from Web usage data.

Chapter 3: Research Methodology

This chapter presents the methodology of the research. This methodology is divided into four main steps: data collection, data-Pre-processing, data mining technique such as clustering, association rule mining and classification for pattern discovery. Data warehouse and OLAP technique will be used for patterns analysis.

Chapter 4: The Results of the Research

In Chapter 4 of this thesis, we presented experimental results using the log files of SUST Web sites. The results showed the reduced the Web access log files down to 25% of the initial size. In this process, only the unnecessary requests are removed, all the other information is kept and can be recreated from the database that we propose. In addition, this chapter explain and discuss the results of the pattern discovery and analysis techniques that are used.

Chapter 5: Conclusions and Future Work

In the last chapter of our research, we summarize the contributions of our research and also we present the future of research.

CHAPTER TWO
LITERATURE REVIEW

2.1 INTRODUCTION

The objective of this chapter is to define web usage mining (WUM) and discuss each of the phases. WUM is the application of data mining algorithms to web click stream data in order to extract web usage patterns. WUM of large web sites may require a data warehouse to store the log file data and OLAP to extract data patterns.

2.2 WEB MINING

The term Web-mining (web data-mining), was first mentioned by Kaur [21], who suggested that traditional data mining techniques for finding hidden patterns in huge databases, can be applied to web-based information. Web mining is an emerging methodology in education research, assisting instructors and developers in improving learning environments and supporting decision-making of policymakers [22]. Web Mining is use of Data Mining techniques to automatically discover and extract information from web data [23]. Web mining is the general name of the data mining technique used in an attempt to make content analysis from the online web sites. Web mining has the facility of utilization in two different areas, the first is the analysis related to the content of the pages presented and the second is the analysis based on the user interaction. According to the differences of the mining objects, there are roughly three knowledge discovery domains that pertain to web mining [24] [25] [26]: Web Content Mining, Web Structure Mining, and Web Usage Mining, as shown in Figure 2.1.

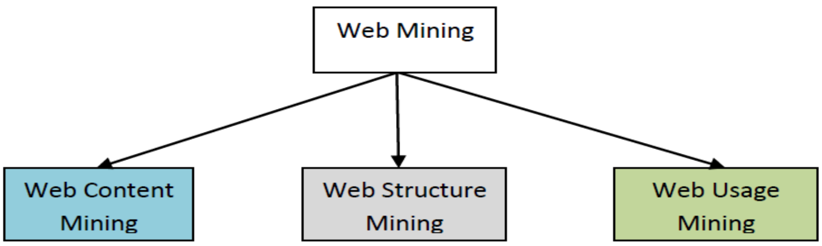


Figure 2.1: Web Mining Taxonomy

2.3 THE USAGE MINING ON THE WEB

Web Usage Mining (WUM) is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web log files. It mines secondary data (web logs) derived from the users' interaction with web pages during certain period of Web sessions [27]. Web usage mining consists of three phases, namely: pre-processing, pattern discovery, and pattern analysis [28] [29] [30], as shown in Figure 2.2. The goal of web usage mining is to get into the records of servers (log files) that store the transactions performed in the web in order to find patterns revealing the usage of the customers [31] [32]. WUM has become an active area of research in the field of data mining due to its vital importance [33].

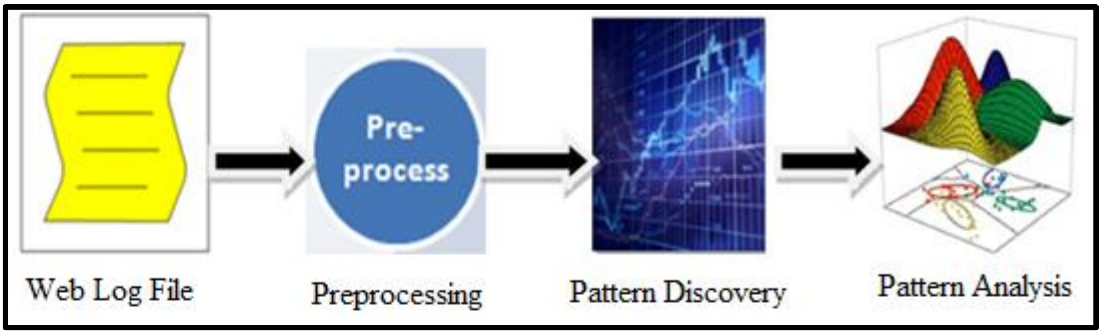


Figure 2.2: Phases of Web Usage Mining

2.4 WEB ACCESS PATTERNS

Web access pattern mining is an application of sequence mining on web log data to generate interesting user access behaviours on World Wide Web. Mining of web access patterns generated by the users' interaction with the World Wide Web is thrust area of research [34].

2.5 THE LOG FILE: WHAT IS IT AND HOW DO WE STORE INFORMATION ON IT?

2.5.1 Log File Definition

A log file is defined as "a file that lists actions that have occurred" [35]. Such files are generated by servers – a computer or a device on a network that manages network resources and contains a list of all requests made to the server by the network's users.

A Web log file records activity information when a Web user submits a request to a Web Server [36]. The main source of raw data is the web access log which we shall refer to as the log file [37].

2.5.2 Storage of Information on A Log File

As it is the rule for every file, information in the log file has to be written in a specific format; that is in a specific sequence and in a certain way that will facilitate the analysis of the file and 'instruct' the computer as to how to read and use.

Log files can be located in three places [38] [39] [40].

- **Web Servers-** A web server dispenses the web pages as they are requested
- **Proxy Server**- A proxy server is an intermediary computer that acts as a computer hub through which user requests are processed.
- **Web Client**- A Web client is a computer application, such as a web browser, that runs on a user local computer or workstation and connects to a server as necessary.

2.6 WEB SERVER LOG FILE

A web server log file is a log file that automatically creates and maintains the activities performed in it [41]. This file is used to record each and every hit to a web site. It maintains a history of page requests, also helps us in understanding how and when your website pages and application are being accessed by the web browser. These log files contain information such as an IP address of a remote host, content requested, and time of request [42] [43].

2.7 NCSA COMBINED LOG FORMAT

Stores all common log information with two additional fields referrer and user agent.

Syntax: Host IP address, Proprietor, Username, date: time, request method, status code, byte size, referrer and user agent [44] [45].

Descriptions of the access that were utilized to generate the data sets are provided below, each row of the log contains the information is shown in Table 2.1.

Table 2.1: Description of Log Access used to Generate Data Sets

Access name	Description
IP Address	Remote IP address
Proprietor	The name of the owner making an http request
User name	Username and password if the server requires user authentication
Date / Time	date/time of the transaction
Method	Modes of request
URL	URL requested by the client
Protocol	HTTP protocol
Statues code	HTTP return code
Byte Size	Size in bytes of the response sent to the client
Referred	The site from which the visitor came
Agent	User agent

Log proprietor- The name of the owner making an http request is recorded through this field. They do not expose this information for security purpose. When they are not exposed they are denoted by (-).

Username- This field records the name of the user when it gets a http request. They do not expose this information for security purpose. When they are not exposed they are denoted by (-).Figure 2.3 shows a sample of a single entry log file a common transfer log extract from SUST log file.

```
41.209.88.192 - - [07/Nov/2008:00:46:51 +0300] "GET /j_images/sar.jpg HTTP/1.1" 200 14292 "http://jst.sustech.edu/" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.3) Gecko/2008092417 Firefox/3.0.3"
```

Figure 2.3: Single Entries of Log File from SUST

Here, 41.209.88.192 is the IP address of the client, 07/Nov/2008:00:46:51 is the date/time of transaction; GET is the method of transaction, j_images/sar.jpg is URL requested by client, HTTP/1.1 is the HTTP protocol,200 is HTTP return code (200 means OK), 14292 is the size in bytes of the response sent to the client, http://jst.sustech.edu is the URL referring to the request one, “Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.3) Gecko/2008092417 Firefox/3.0.3" is the user agent.

2.8 DATA PRE-PROCESSING ON WEB SERVER LOGS

Web usage mining is the application of data mining techniques to usage logs of large data repositories. Usually, the data collected in web log file is incomplete and not suitable for mining directly. Therefore, pre-processing is necessary to convert the data into a suitable form for pattern discovery [46]. We begin this phase by data extraction then data cleaning and finally data filtering, because the origin web logs data sources are blended with irrelevant information. Data pre-processing plays an important role in Web usage mining. It uses to filter and organize only appropriate information before using Web mining algorithms on the Web server logs [47].

The original server logs are cleaned, formatted, and then grouped into meaningful sessions before being utilized by WUM. This phase contains three sub-steps: Data Cleaning, User Identification, and Session Identification [48], as shown in Figure 2.4.

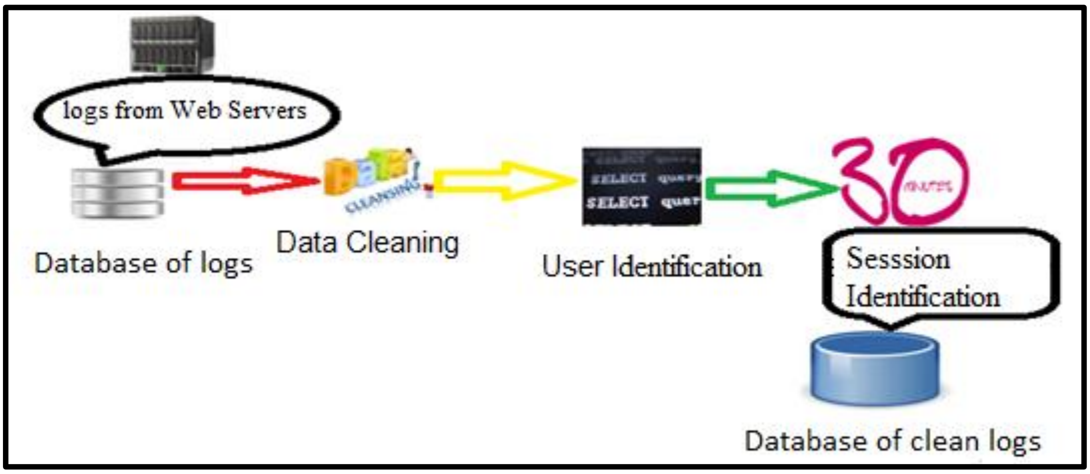


Figure 2.4: Pre-Processing Steps

2.8.1 Data Cleaning

The data cleaning process removes the data tracked in Web logs that are useless or irrelevant for mining purposes [49]. The request processed by auto search engines, such as Crawler, Spider, and Robot, and requests for graphical page content. Thus the data cleaning step removes the following entries from the original log file [50] [51].

- The entries having suffixes like .jpg, .jpeg, .css, .mapetc.,
- Entries having status code failure.
- Remove all record which do not contain method” GET”.
- Remove navigation sessions performed by Crawler, Spider, and Robot.

2.8.2 User Identification

User identification is the process of identifying each different user accessing Web site. Goal of user identification is to mine every user's access characteristic, and then make user clustering and provide personal service for the users [52]. Each user has unique IP address and each IP address represents one user. However, in fact there are three conditions: **(l)** Some users have unique IP address. **(2**) Some user has two or more IP addresses. **(3)** Due to proxy server, some user may share one IP address. Rules for user identification are [53]:

- Different IP addresses refer to different users.
- The same IP with different operating systems or different browsers should be considered as different users.
- While the IP address, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before, according to the topology of the site.
- Users are uniquely identified by combination of referrer URL and user agent.

2.8.3 User Session

After identifying users, we need to identify sessions. To do this, we can divide access of the same users into sessions. It is difficult to detect when one session is finished and start another. To detect sessions is common use of time between requests; if two requests are called in of time frame, we can suppose that these requests are in the same session; in another way below of time frame, we can consider two different sessions. A good time frame is 30 minutes [54].

2.9 PATTERN DISCOVERY

After data pre-processing phase, the pattern discovery method should be applied [55]. This phase consists of different techniques derived from various fields such as statistics, machine learning method mainly have Association Rules, pattern recognition, etc. applied to the web domain and to the available data [56]. Several methods and techniques have already been developed for this step [57]. Some of the frequently used solutions are statistical analysis, clustering, association rules and classification [58].

2.9.1 Statistical Analysis

Statistical analysis is the most common method to extract knowledge about visitors to a web site [59]. We can compute various kinds of descriptive statistics measurements like (frequency, mean, and median) on variables such as page views, viewing time, or length of the navigation path [60]. Although the statistical analysis useful for improving system performance, enhancing system security, or facilitating site modification. For example, we can detect unauthorized entry points to our web site.

2.9.2 Clustering

Clustering has been widely used in WUM to group together similar sessions among large amounts of data based on a general idea of distance function which computes the similarity between groups [61] [62].Clustering means the act of partitioning an unlabelled dataset into groups of similar objects. Each group, called a 'cluster', consists of objects that are similar between themselves and dissimilar to objects of other groups. In the past few decades, cluster analysis has played a central role in diverse domains of science and engineering [63] [64]. Two types of clusters can be found in WUM: user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content [65]. Here we will briefly describe some techniques to discover patterns from processed data. Commonly used clustering algorithms are: K-means and Density based clustering.

2.9.2.1 K-Means Clustering

The k-means method partitions the data set to classify objects based on attributes into positive k cluster in which each observation belongs to the cluster with the nearest mean [66] .The clustering is done by minimizing the sum of squared distance in each cluster. Thus, the strength of K-means algorithm lies in its computational efficiency and the nature of easy to use. The procedure follows a simple way to classify a log file dataset.

The basic step of k-means clustering as shown in Figure 2.5 is simple. In the beginning we determine number of clusters k and assume the centroid or center of clusters. We can take random objects as the initial centroid or the first k object which serves as an initial centroid. Then the k-means algorithm will carry out its steps until convergence. Iterate until stable (no move group) to group the objects based on minimum distance.

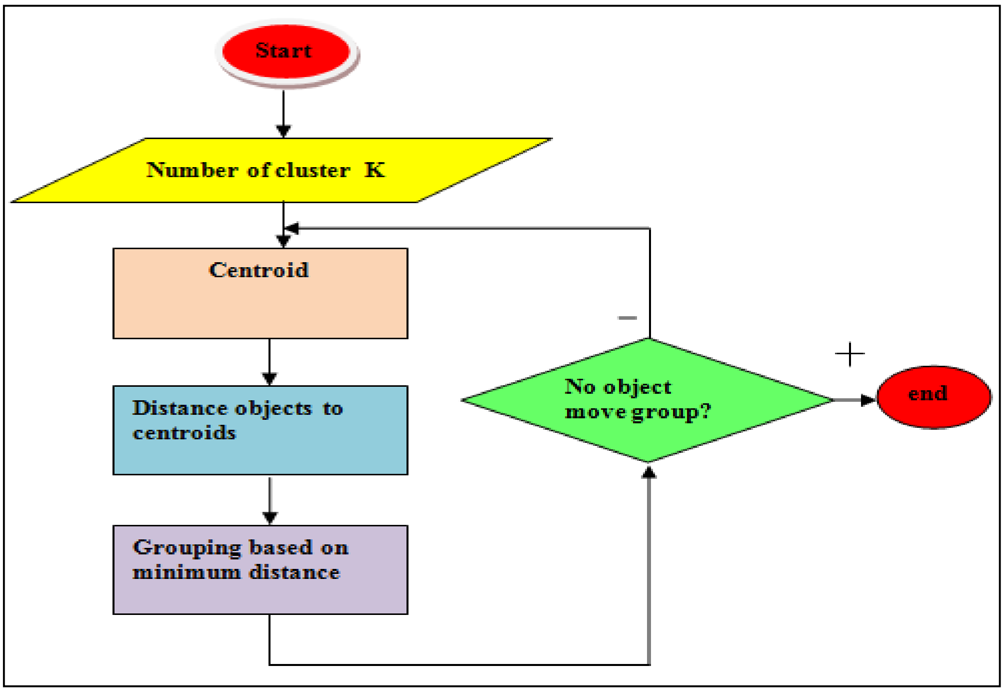


Figure 2.5: K-Means Clustering Steps

2.9.2.2 Density Based Clustering

The basic idea of density-based clustering is that clusters are dense regions in the data space, separated by regions of lower object density [67]. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts) [68]. Figure 2.6 shows the flowchart for Density based algorithm.

2.9.2.3 Feature Selection

Feature Selection is a term commonly used in data mining to describe the techniques available for reducing inputs to a manageable size for processing and analysis. Correlation-based Feature Selection (CFS) measures correlations between nominal features, so numeric features are first discretized. It is also an effective dimensionality reduction technique and is an essential pre-processing method to remove noise features [69]. The basic idea of feature selection algorithms is searching through all possible combinations of features in the data to find which subset of features works best for prediction. The selection is performed by reducing the number of features of the feature vectors, keeping the most meaningful discriminating ones, and removing irrelevant or redundant ones [70].

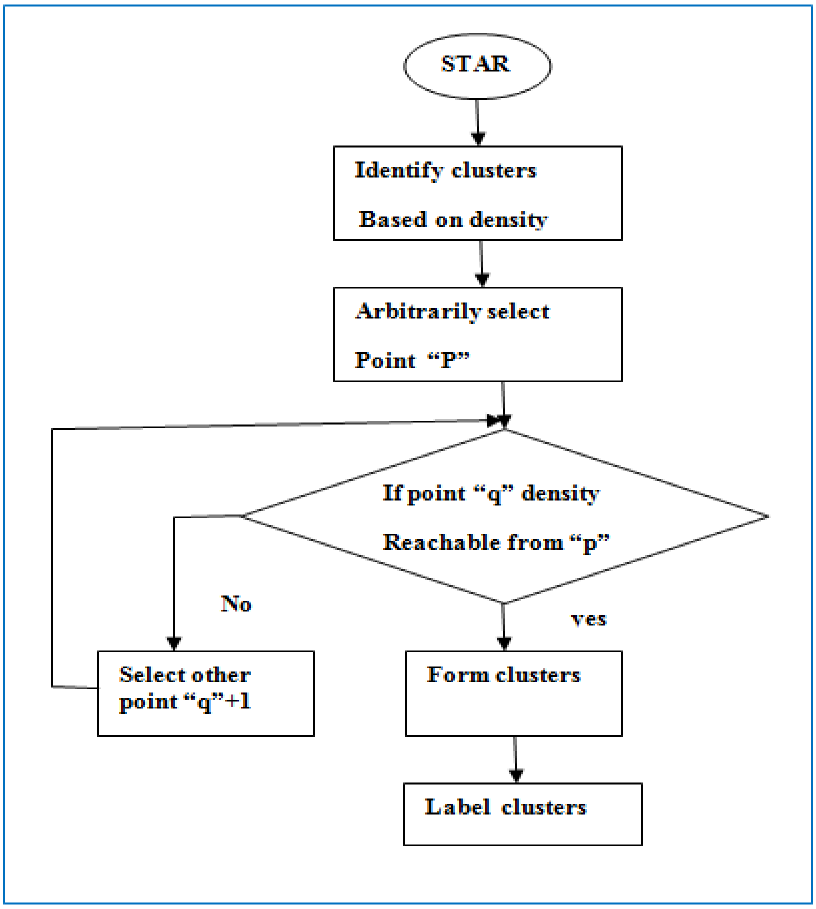


Figure 2.6: Flowchart for Density Based Algorithm

2.9.3 Association Rule Mining

Association rule mining is one of the major techniques of data mining and it is the most common form of pattern discovery in unsupervised learning systems. It serves as a useful tool for finding correlations between items in large database [71].Most common approaches to association discovery are based on the Apriori algorithm. This algorithm finds groups of items (namely; page-views appearing in the pre-processed log) occurring frequently together in many transactions (i.e. satisfying a user specified minimum support threshold) [72]. It finds rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Two measurements in association rule mining are support and confidence. The support corresponds to the frequency of the pattern while confidence indicates rule's strength [73].

Support of a rule A → B = no. of instances with A and B / no. of all instances

Confidence of a rule A → B = no. of instances with A and B / no. of instances with A
= support (A & B) / support (A).

The goal of association rule mining is to find all rules having: support ≥ minsup threshold and confidence ≥ minconf threshold.

2.9.3.1 Lift

Lift is an interestingness measure of an association rule that compares the rule confidence to the expected rule confidence.

Lift of a rule A → B = support (A & B) / [support (A) * support (B)]

2.9.3.2 Large Item Set

A large item set is an item set whose number of occurrences is above a threshold or support. The minimum support requirement dictates the efficiency of association rule mining. One major motivation for using the support factor comes from the fact that we are usually interested only in rules with certain popularity [74]. The minimum support threshold parameter needs to be set to the value that gives optimal results. If the support threshold is set too low, too many potentially not truly interesting rules are generated, cluttering the rule set and making it hard to understand for the final user. On the other hand, if the support threshold is set too high, there is a chance that too many potentially interesting rules are missed from the rule set. Eliminating Redundant rules and Clustering decreased the size of the generated rule set for obtain Interestingness rules.

2.9.3.3 Redundant rules

Deleting redundant rules from the result set: If you have A $\rightarrow$ B and A & C $\rightarrow$ B, the second rule is redundant.

2.9.3.4 Page cluster

Let us suppose that the set of all rules R contains the following rules:

$a \rightarrow b$, $conf(a \rightarrow b) \approx 1$, $b \rightarrow a$, $conf(a \rightarrow b) \approx 1$, where a and b are items $a \in I$, $b \in I$.

We define a cluster $C_{ab} = \{a, b\}$.

2.9.4 Classification

Given a training data set, the classification model was used to categorize the given training data set into attributes and the attributes were referred to as class [75]. In web log data time stamp, users, etc. were considered as attributes or class. Classification can be performed using different techniques. The goal was to predict the target class based on our source data (web log data). Our model takes into consideration the category type of classification in which the target attribute has only two possible variations: forenoon and afternoon.

2.9.4.1 Base Classifiers

Base classifiers refer to individual classifiers used to construct the ensemble classifiers. J48, k-NN, NBand BN classifiers are some of the commonly used base classifiers. However, the proposed technique is a very general approach and its performance may further improve depending on the choice and/or the number of classifiers as well as the use of more complex features.

2.9.4.1.1 Decision Tree

Decision tree is one of the most popular approaches for both classification and predictions. It is the predictive machine-learning model that classifies the required information from the data. Each internal node of a tree is considered as attributes and branches between the nodes are possible values [76].Building algorithms may initially build the tree and then prune it for more effective classification. With pruning technique, portions of the tree may be removed or combined to reduce the overall size of the tree. The time and space complexity of constructing a decision tree depends on the size of the data set, the number of attributes in the data set, and the shape of the resulting tree [77].Decision tree classifier has limitations as it is computationally expensive because at each node, each candidate splitting field must be sorted before its best split can be found [78] .

2.9.4.1.2 K-Nearest Neighbor

Nearest Neighbor (also known as Collaborative Filtering or Instance-based Learning) is a useful data mining technique that allows using the past data instances, with known output values, to predict an unknown output value of a new data instance [79]. Hence, at this point, this description should sound similar to both regression and classification. Many researchers have found that the k-nearest neighbors (KNN) algorithm achieves very good performance in their experiments on different data sets [80].The general principle is to find the k training samples to determine the k-nearest neighbors based on a distance measure. Next, the majority of k nearest neighbors decides the category of the next instance.

2.9.4.1.3 Naive Bayes

A Naive Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. It can handle an arbitrary number of independent variables whether continuous or categorical [81]. The final classification is done by calculating the posterior probability of the object by multiplying the prior probability and likelihood. Based on the posterior probability, it takes the decision. The performance of Naive Bayes depends on the reality of data set [82].

2.9.4.1.4 Bayes Net

Bayes Net (BN) is based on the Bayes' theorem. So, conditional probability on each node is calculated and formed a Bayesian Network. Bayesian Network is a directed acyclic graph. In BN, it is assumed that all attributes are nominal and there are no missing values. Different types of algorithms are used to estimate the conditional probability such as Genetic Search, Hill Climbing, Simulated Annealing, Tabu Search, Repeated Hill Climbing and K2 [83]. The output of the BN can be visualized in terms of graph.

Figure 2.7 shows the visualized graph of the BN for a SUST web data set. Visualize graph is formed by using the children attribute of the web data set. In this graph, each node represents the probability distribution table within it. A new neural network architecture referred to as BAYESNET (Bayesian network) is capable of learning the probability density functions (PDFs) of individual pattern classes from a collection of learning samples, and designed for pattern classification based on the Bayesian decision rule. Bayes nets are often used as classifier to predict the probability of a target class label given features [84] .

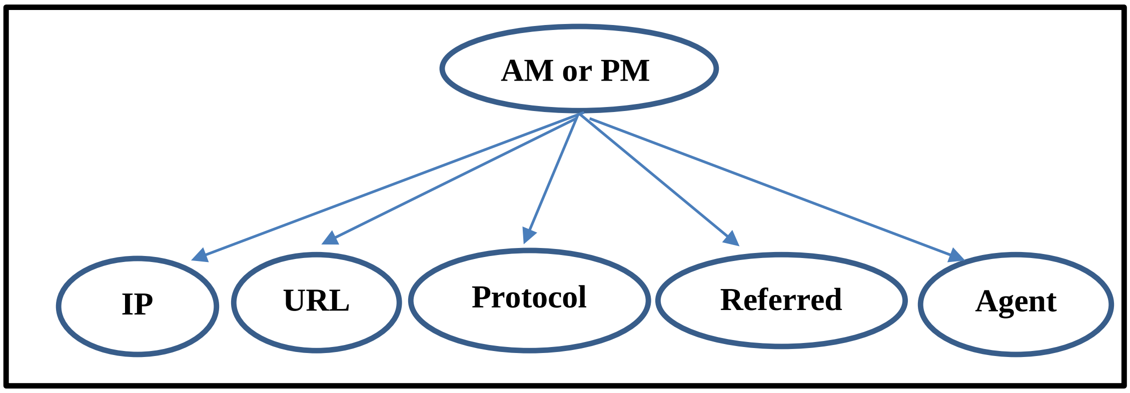


Figure 2.7: Visualize Graph of the Bayes Net for a Web Dataset

2.9.4.2 Meta Classifiers

Meta-learning means learning from the classifiers produced by the inducers and from the classifications of these classifiers on training data. The following sections describe the most well-known meta-combining methods: Stacking and Voting.

2.9.4.2.1 Stacking

The first method that we employ for classifier combination is stacking, where the rule-based classifier is applied on the output produced by the based Classifier. Stacked generalization (or stacking) is a different way of combining multiple models that introduces the concept of Meta learner [85].

Stacking procedure as follows [86]:

1) Split the training set into two disjoint sets.

2) Train several base learners on the first part.

3) Test the base learners on the second part.

4) Using the predictions from 3) as the inputs, and the correct responses as the outputs, train a higher- level learner.

2.9.4.2.2 Voting

In the voting framework for combining classifiers, the predictions of the base-level classifiers are combined according to a static voting scheme, which does not change with training data set [84].Voting does use a simple combination scheme of the base-classifier predictions to derive the final ensemble prediction. There are several types of voting schemes, which differ by the number of votes required for an ensemble prediction. Alternately, often a more powerful voting technique is to use a sum of each classifier's probability distribution for the classes and predict the class with the highest value.

2.10 PATTERN ANALYSIS

This is the final step in the WUM process. It helps to filter insignificant information to obtain the valuable information. The pattern analysis phase means applying data mining techniques on the pattern discovery data. The patterns are analyzed using several techniques. The most common form of pattern analysis consists of Structured Query Language (SQL), Online Analytical Processing (OLAP) [87]. In OLAP techniques, the result of pattern discovery is loaded into data cube and then OLAP operations are performed. After this, to interpret the results, visualization techniques are used [46], such as graphing patterns or assigning colours to different values, can often highlight overall patterns or trends in the data. The result of pattern analysis helps to improve the system performance and to modify the web site. It helps to attract the visitors and to give the personalized services to regular user [88]. The result of such analysis might include: most recent visit per page, who is visiting which page, the frequency of use of each hyperlink, and most recent use of hyperlinks [89].

2.10.1 Data Warehouse Construction

We can define Data warehouse as “a huge repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making” [90]. Once the Data warehouse is constructed we apply intelligent methods called data mining techniques to extract data patterns. Generally data warehouse is modelled by a multidimensional database structure called data cubes. A data warehouse provides the data source for online analytical processing and data mining. A well-designed data warehouse would feed business with the right information at the right time in order to make the right decisions in Server log file system [91].

2.10.2 OLAP

The most common form of pattern analysis consists of a knowledge query mechanism such as SQL (Structured Query Language), which needs end user access tools. OLAP (Online analytical processing) as an end-user access tool than can support advanced quarry by using a strong methodology called data cube [92]. OLAP can simplify the analysis of usage statistics of the server access logs. It pre-calculate summary information to enable roll-up or aggregation, which allows the user to move to the higher aggregation level, drilling, which is the reverse of a roll-up and represents the situation when the user moves down the hierarchy of aggregation, applying a more detailed grouping, pivoting, which changes the perspective in presenting the data to the user, slicing, which is based on selecting one dimension and focusing on a portion of a cube, and dicing, which creates a sub-cube by focusing on two or more dimensions [93].

2.11 WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS (WEKA)

WEKA includes several machine learning algorithms for data mining tasks. The algorithms can either be called from the users own Java code or be applied directly to the ready dataset [94]. WEKA contains general-purpose environment tools for data pre-processing, regression, classification, association rules, clustering, feature selection and visualization [95].WEKA provides an attribute selection tool. The process is separated into two parts, Attribute Evaluator and Search Method.

2.11.1 Evaluators

- CfsSubsetEval - Evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them; subsets of features that are highly correlated with the class while having low inter-correlation are preferred.
- ConsistencySubsetEval - Evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of features.
- PCA - Performs a principal components analysis and transformation of the data.
- Wrapper Subset Eval- Wrapper attributes subset evaluator.

2.11.2 Search Methods

- Best First - Searches the space of feature subsets by greedy hill-climbing augmented with a backtracking facility.
- Genetic Search - Performs a search using the simple genetic algorithm
- Ranker - Ranks features by their individual evaluations. Use in conjunction with feature evaluators (Relief F, Gain Ratio, Entropy).
- Exhaustive search –Performs an exhaustive search over all the features.
- Forward Selection –Performs selection of an attribute one by one.

2.12 RELATED WORK

The important concepts of Web usage mining and its various practical applications presents by [96]. Further a novel approach called “intelligent-miner" (i-Miner) is presented. I-Miner could optimize the concurrent architecture of a fuzzy clustering algorithm (to discover web data clusters) and a fuzzy inference system to analyze the Web site visitor trends. A hybrid evolutionary fuzzy clustering algorithm is proposed to optimally segregate similar user interests. The clustered data is then used to analyze the trends using a Takagi-Sugeno fuzzy inference system learned using a combination of evolutionary algorithm and neural network learning. Proposed approach is compared with self-organizing maps to discover patterns and several function approximation techniques like neural networks, linear genetic programming and Takagi-Sugeno fuzzy inference system to analyze the clusters. I-Miner framework gave the best results with the lowest RMSE on test error and the highest correlation coefficient. When optimal performance is required (in terms of accuracy and smaller structure) such algorithms might prove to be useful as evident from the empirical results. An important disadvantage of I-Miner is the computational complexity of the algorithm.

The research entailed the development of a ’Say account’ field classification system introduced by [97]. MLP networks were used to classify caller interactions. Binary coded and real coded GAs that utilized ranking as well as tournament selection functions were also employed to optimize the classifier architecture.

The development methodology utilized for creating all the networks involved, initially, pre-processing the data sets. This ensured that the classifiers would interpret the inputs proficiently. Thereafter, the numbers of hidden nodes were optimized utilizing the GA algorithm. This resulted in creating acceptable network architecture. GA results were compared in terms of computational efficient, repeatability and the quality of the solution. As a result, it can be concluded that this GA is most suited to this application in terms of optimal solution; however it is not the most computational efficient algorithm. In [98] a new method to extract navigational patterns from web logs is proposed. Ant-based clustering has been used for this purpose. It needs a neighbourhood function to be defined for. After the clustering is completed, alignment processing has been applied to the extracted sequences in each cluster and extract the representative for each cluster. The advantage is that the total numbers of cluster is generated automatically, and the disadvantage is that its cluster result is random and its result is influenced by the input data and the parameters, which leads low quality of its cluster result.

[99] Describe a relational OLAP (ROLAP) approach for creating a web-log warehouse. This is populated both from web logs, as well as the results of mining web logs. They also present a web based ad-hoc tool for analytic queries on the warehouse. They discuss the design criteria that influenced their choice of dimensions, facts and data granularity, and present the results from analyzing and mining the logs. This study solve the problems of many existing tools that generate fixed reports from web logs, they typically do not allow ad-hoc analysis queries. Moreover, such tools cannot

discover hidden patterns of access embedded in the access logs. However, the researchers have populated the fact table directly from the web logs instead of loading the data into the transactional database, and then populate the fact tables from it.

By creating such a transactional database, the pre-processing step will be automated to a large extent, and more dynamic "monitoring" can be done by the system automatically. Features like alerts and warnings can be easily incorporated in such architecture.

A descriptive study of Knowledge Discovery from Web Usage Mining is presented in [100]. The Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from web logs, in order to improve web based applications. This study is useful for researcher exclusively for doing research on web mining. However the work focuses only on descriptive study and ignores experimental design.

A novel approach called Growing Neural Gas (GNG) is introduced by [101]. A neural network is used in the process of Web Usage Mining to detect user's patterns. The process details the transformations necessaries to modify the data storage in the Web Servers Log files to an input of GNG. The result showed that the Growing Neural Gas Algorithm is better than K-Means and SOM for identify common patterns in Web. Also GNG has a better group of users. The salient disadvantage of growing neural gas (GNG) is the permanent increase in the number of nodes.

A new ensemble of decision tree classifiers that ensembles ID3 classifier for mining web data streams is introduced by [102]. It is an efficient mining method to obtain a proper set of rules for extracting knowledge from a large amount of web data streams. They built a web server using Model 2 Architecture to collect the web data streams and applied the ensemble classifier for generating decision rules using several decision tree learning models. Experimental results demonstrate that the proposed method performs well in decision making and predicting the class value of new web data streams. However, in this study the researchers have only using ID3 classifier in ensemble, although ID3 classifier suffer from number of problem such as: Only one attribute at a time is tested for making a decision and can only handle nominal values.

Goel and Jha [103] present a log analyzer tool called Web Log Expert for ascertaining the behavior of users who access an astrology website. It also provides a comparative study between a few log analyzer tools available. Web Log Expert tool gives information about our site's visitors: activity statistics, accessed files, information about referring pages, search engines, browsers and operating systems. It is, however, not apparent which WUM algorithms are used for this analysis and only descriptive statistics are provided.

Dhillon and Kaur [104] present a frame for web usage mining based on classification algorithms including their features and limitations. They analyze the performance of some classification algorithms such as Decision Tree Classifier (DTC), Naïve Bayesian Classifier (NBC), Support Vector Machine (SVM), Neural Networks (NNs), Rule Based Classifier (RBC) and K-Nearest Neighbor Classifier (KNN) on the bases of some factors like accuracy, precision, session based timing, recall. The results show that Naive Bayesian performed well with respect to all the factors and Decision Tree classifier and SVM also perform well as compared to others. The advantage of this study is represented in using various number of classification algorithms and comparing their results. However, the study doesn’t benefit from using a combination of these algorithms.

The Table 2.2 analyze the related works and their approaches and finding out the limitations of the related works.

Table 2.2: Related Work

Investigator	Research\Approach	Strengths	Limitations
Joshi, Yesha and Krishnapurm (2010)	Warehousing and Mining Web Logs	This study solve the problems of many existing tools that generate fixed reports from web logs, they typically do not allow ad-hoc analysis queries.	The researchers have populated the fact table directly from the web logs instead of loading the data into the transactional database, and then populate the fact tables from it
H. Yogish, D. Raju, and T. Manjunath (2011)	The Descriptive Study of Knowledge Discovery from Web Usage Mining	This study is useful for researcher exclusively for doing research on web mining.	The work focuses only on descriptive study and ignores experimental design.
Tani, Farid and Zahidur (2012)	Ensemble of Decision Tree Classifiers	Experimental results demonstrate that the proposed method performs well in decision making and predicting the class value of new web data streams.	In this study the researchers have only using ID3 classifier in ensemble, although ID3 classifier suffer from number of problem such as: Only one attribute at a time is tested for making a decision and can only handle nominal values.
Goel and Jha (2013)	Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool	Web Log Expert tool gives information about our site's visitors: activity statistics, accessed files, information about referring pages, search engines, browsers and operating systems	This study not apparent which WUM algorithms are used for this analysis and only descriptive statistics are provided.
Dhillon and Kamaljit (2014)	Comparative Study of Classification Algorithms for Web Usage Mining.	The advantage of this study is represented in using various number of classification algorithms and comparing their results.	The study doesn’t benefit from using a combination of these algorithms.

2.13 OPEN ISSUES

The related works indicate that the current Web usage mining needs an improvement to be useful such as:

- How do design an efficient hybrid system to discover and analyze the patterns?
- The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop system for extract and mining knowledge.

2.14 CHAPTER SUMMARY

This chapter discusses WUM as well as the various data mining algorithms which can be used for pattern discovery and analysis from log file data. Log file format is used to record all user accesses. In order to determine the structure of the log file, a formal specification of the log file format was provided. The fields from the log file identified as relevant for WUM are date, time, ip address, username, and URL and user agent. The origin web logs data sources are blended with irrelevant information. Therefore, pre-processing is necessary to convert the data into a suitable form for pattern discovery. This phase contains three sub-steps: Data Cleaning, User Identification, and Session Identification. In addition, this chapter gives information to the various data mining methods and techniques used (statistical analysis, clustering, association rules and classification). Clustering algorithms are: K-means and Density based clustering, Classifier Algorithms namely: J48, KNN, NB and BN and data warehouse and OLAP are described in detail. This chapter evaluated related systems by investigating the WUM algorithms and the data mining models. None of these systems are able to satisfy the objectives that were established for this research.

CHAPTER THREE

RESARCH METHODOLOY

3.1 INTRODUCTION

This chapter provides information on the research methodology of this thesis including data collection, performing pre-processing operation, applying data mining techniques for pattern discovery and data warehouse and OLAP technique for patterns analysis. The methodology is depicted in Figure 3.1 and described below. This is a four steps process.

1. The beginning is basically collecting of SUST web log file and process of separating out different data fields from single server log entry is identified as data field extraction. After field extraction, the read logs records will be stored in Staging area to facilities data transformation, and mapping.
2. In the transformation, the pre-processing step will be conducted. This contains three sub steps: Data Cleaning, User Identification, and Session Identification.
3. In the next step, data mining technique like clustering, association rule mining and classification will be applied for pattern discovery.
4. Finally in step 4, Data warehouse and OLAP technique will be used to find patterns analysis.

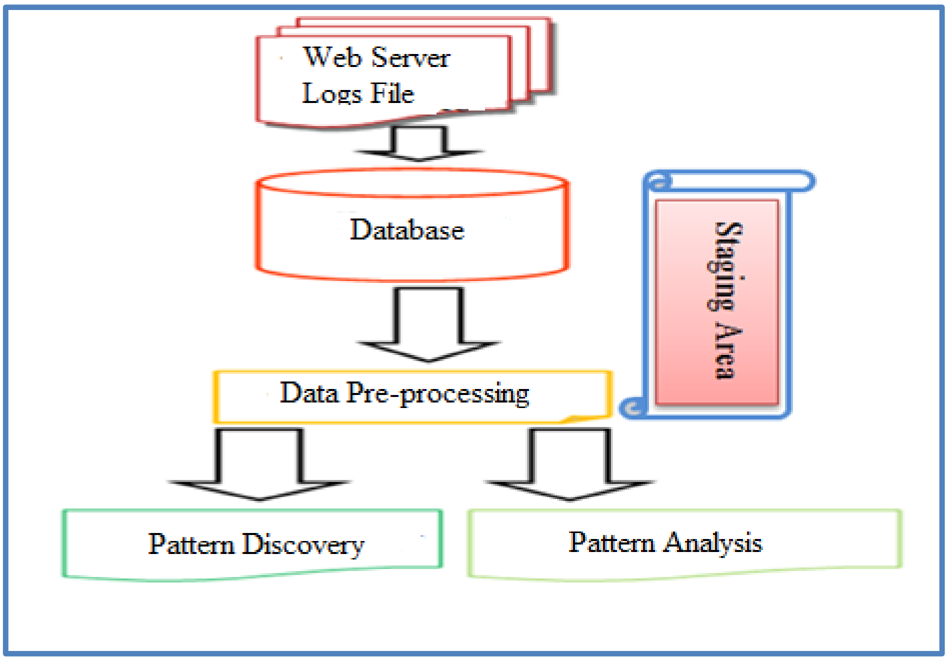


Figure 3.1: Diagram of the Methodology Steps

3.2 WEB LOG DATA

As the developed system is to be used to identify trends of visitor website behavior within the SUST web site applications from 7/Nov/2008 through 20/Aug/2009.A portion of the log file used for the experimentation is illustrated in Figure 3.2.

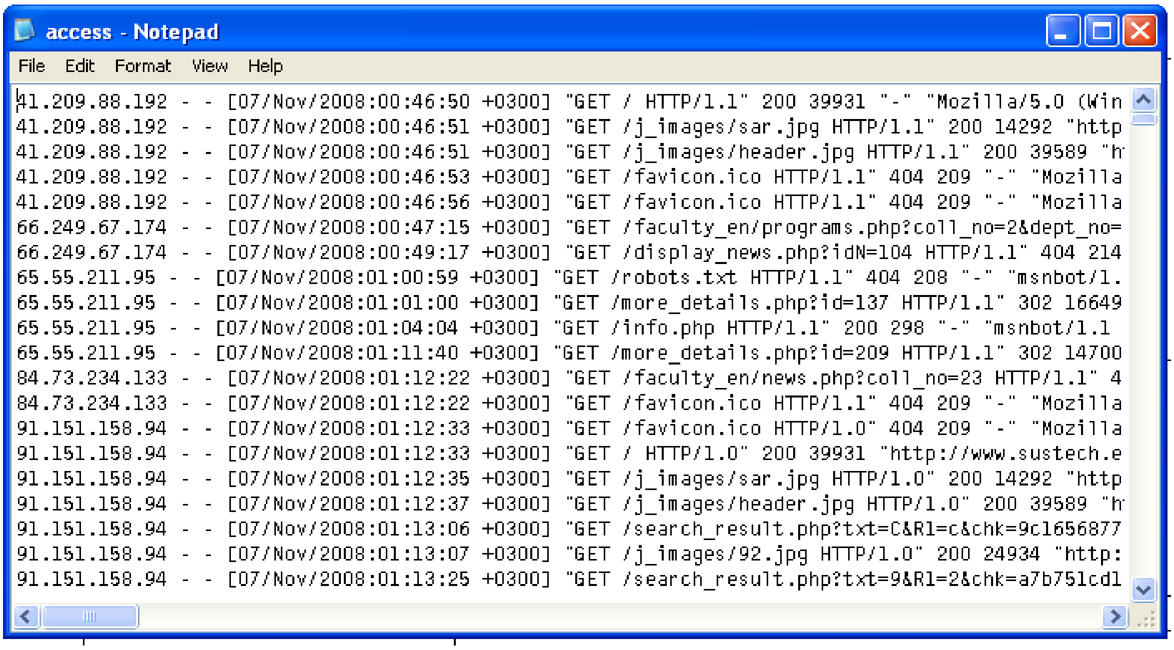


Figure 3.2: A Portion of SUST Log File

3.3 DATA FIELD EXTRACTION AND TRANSFER SERVER LOGS TO DATABASE

A server log file consists of various data fields that should be separated before applying any cleaning procedure. The process of separating out different data fields from single server log entry is identified as data field extraction. A server uses different characters such as a comma or a space character which works as separators.

To analyze the log file data, we need first to deal with text file using regular expression so as to separate each line in the text file into different fields and then we need database object for loading these data in a database table. The whole log file can be read in one variable and then we can move this variable line by line using a loop statement. After reading the log files, several attributes are considered important for the analysis. The read logs records will be stored in a database. Figure 3.3 shows the database to store the data.

IPADD	Proj	User	BRDatetime	methodr	webextension	wprotocol	statuscode	ByteSize	wurl	
213.185.116.12	-	-	[16/Nov/2008:13:27:57+0300]	GET	/j_images/header.jpg	HTTP/1.0	200	39589	3c64f7a197182fe930f6d579	.1; SV1; Mozilla/
213.185.116.11	-	-	[16/Nov/2008:13:28:00+0300]	GET	/j_images/92.jpg	HTTP/1.0	200	24934	3c64f7a197182fe930f6d579	.1; SV1; Mozilla/
213.185.116.12	-	-	[16/Nov/2008:13:28:46+0300]	GET	/info.php	HTTP/1.0	200	298	3c64f7a197182fe930f6d579	.1; SV1; Mozilla/
213.185.116.12	-	-	[16/Nov/2008:13:28:48+0300]	GET	/j_images/noaccess.jpg	HTTP/1.0	200	32230	'/jst.sustech.edu/info.php	.1; SV1; Mozilla/
65.55.211.90	-	-	[16/Nov/2008:13:44:53+0300]	GET	346fcbb25bccac086472dee	HTTP/1.1	200	16649	-	
65.55.211.90	-	-	[16/Nov/2008:13:45:46+0300]	GET	cf4a1f8cb42a8ed6fc6603e	HTTP/1.1	200	15725	-	
196.1.209.67	-	-	[16/Nov/2008:13:46:05+0300]	GET	/	HTTP/1.0	200	39931	sustech.edu/sudannewar/	Mozilla/4
196.1.209.67	-	-	[16/Nov/2008:13:46:06+0300]	GET	/j_images/sar.jpg	HTTP/1.0	200	14292	http://jst.sustech.edu/	Mozilla/4
196.1.209.67	-	-	[16/Nov/2008:13:46:06+0300]	GET	/j_images/header.jpg	HTTP/1.0	200	39589	http://jst.sustech.edu/	Mozilla/4
65.55.211.100	-	-	[16/Nov/2008:14:17:09+0300]	GET	)ff80e28fbe42d65b40bbc1	HTTP/1.1	200	17685	-	
212.118.149.34	-	-	[16/Nov/2008:14:30:20+0300]	GET	0a2e2ec81736903d89babf	HTTP/1.1	200	17243	2%D8%B1%D8%B6&meta=	.0 (compatible; I
212.118.149.34	-	-	[16/Nov/2008:14:30:20+0300]	GET	0a2e2ec81736903d89babf	HTTP/1.1	200	17243	2%D8%B1%D8%B6&meta=	.0 (compatible; I

Figure 3.3: The Data after Transferred to a Database

Figure 3.3 shows the server log data after transferring to database and note that all attributes are shown in this figure due to the space restrictions. Several attributes are interesting fields are included in the database.

3.4 DATA PRE-PROCESSING

The data collected in web log file is not suitable for mining directly. Pre-processing is necessary to convert the data into suitable form for pattern discovery. It use to filter and organize only appropriate information before using web mining algorithms on the server logs. We begin this phase by data cleaning, because the origin web logs data sources are blended with irrelevant information. This phase contains three sub steps: Data Cleaning, User Identification, and Session Identification.

3.4.1 Data Cleaning

A proposed algorithm is used for data cleaning as shown in Figure 3.4 . The VB.Net is used to implement this algorithm. After data cleaning only 122122 entries out of 291642 are left in the log. The results are shown in Figure 3.5.

```
1.  Define variables (method, status code, agent and web extension) As string
2.  Check method:
        If method = "GET" Then
            method = 1
        Else
            method = 0
        End If
3.  Check status code
        If status code = "200" Then
            status code = 1
        Else
            status code= 0
        End If
4.  Check agent
       If agent contain the Spider or Robot or Crawler Then
            agent = 0
        Else
            agent = 1
        End If
5.  check web extension
        If web extension .jpgi Or .jpegi Or .jsi Or .cssi Or .gifi   Then
            web extension= 0
        Else
            web extension = 1
        End If
6.  Add data
        If method = 1 and web extension = 1 and status code = 1 and agent = 1
        Then
                "INSERT INTO Web data After Filtering (all fields)
        End If
        Next i
7.  Close db
```

Figure 3.4: A Proposed Data Cleaning Algorithm

WebdataAfterFiltering

IPADD	Prc	Use	BRDatetime	methodı	webextension	wprotocol	statuscode	ByteSize	wurl	agent
41.209.88.192	-	-	[07/Nov/2008:00:46:50+0300]	GET	/	HTTP/1.1	200	39931	-	o/2008092417 Firefox/3.0
65.55.211.95	-	-	[07/Nov/2008:01:04:04+0300]	GET	/info.php	HTTP/1.1	200	298	-	rch.msn.com/msnbot.htm
91.151.158.94	-	-	[07/Nov/2008:01:12:33+0300]	GET	/	HTTP/1.0	200	39931	ewAR/index.php	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:13:06+0300]	GET	ea86c65330f3f39e6f463305	HTTP/1.0	200	15393	'jst.sustech.edu/	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:13:25+0300]	GET	'20cb01857b5738b3f497ccf	HTTP/1.0	200	21033	0f3f39e6f463305	6.0; Windows NT 5.1; SV
65.55.211.95	-	-	[07/Nov/2008:01:13:26+0300]	GET	/info.php	HTTP/1.1	200	298	-	rch.msn.com/msnbot.htm
41.221.17.5	-	-	[07/Nov/2008:01:33:25+0300]	GET	/info.php	HTTP/1.1	200	298	mg&fr=yfp-t-501	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:38:08+0300]	GET	/	HTTP/1.0	200	39931	ewAR/index.php	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:38:28+0300]	GET	/index.php	HTTP/1.0	200	39931	'jst.sustech.edu/	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:38:51+0300]	GET	/info.php	HTTP/1.0	200	298	ı.edu/index.php	6.0; Windows NT 5.1; SV
41.221.17.5	-	-	[07/Nov/2008:01:43:07+0300]	GET	/info.php	HTTP/1.1	200	298	mg&fr=yfp-t-501	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:47:59+0300]	GET	/	HTTP/1.0	200	39931	ewAR/index.php	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:49:45+0300]	GET	627fca2c280e70196c3a4d6	HTTP/1.0	200	10943	'jst.sustech.edu/	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:49:58+0300]	GET	'20cb01857b5738b3f497ccf	HTTP/1.0	200	21033	280e70196c3a4d6	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:50:17+0300]	GET	b2f38abc39fb4bc3e8a3eb6	HTTP/1.0	200	23498	7b5738b3f497ccf	6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:55:31+0300]	GET	'20cb01857b5738b3f497ccf	HTTP/1.0	200	21033	280e70196c3a4d6	6.0; Windows NT 5.1; SV

Figure 3.5: Data after Cleaning

3.4.2 User Identification

Compute the unique User by combination of Referred and Agent.

a- Distinct IP addresses refer to different users.

b- Combine Referred and Agent.

c- The same IP with different combined felid should be considered as different users.

3.4.3 User Session

After we specify the number of unique users in the previous step. In this step we need to get the users sessions. To achieve this we can divide the access of the same users in sessions. The time spent within time limit of 30 minutes for same user will be considers user session. The following is a proposed session identification algorithm as shown in Figure 3.6. Rules for session identification are:

- Different IP addresses refer to different session.
- The same user with time exceeds a certain limit (30 minutes) should be considered as different session.

Begin
Start session=Current time of current session
Session =1
While not eof (LogFile) Do
LogRecord=Read (LogFile)
In the same IP Address
If (the time of current Record –start session) <= 30 min
Then
We are in same the session
Move to next record
Else
Increment session = session +1
Start time = time of current Record
Move to next record
End If
End While
End
Result

Figure 3.6: A Proposed Session Identification Algorithm

A fraction of log files with user sessions identification is shown in Figure 3.7. The figure shows the session length distribution for the SUST dataset after the session is identified.

ID	"IP Address"	"Date/Time"	"URL request by the client"
1	109.82.134.76	10/30/2009 10:13:41 PM	/iepngfix.htc
1	109.82.134.76	10/30/2009 10:13:34 PM	/content_details.php?id=168&chk=29a4
2	109.82.36.41	10/28/2009 5:32:26 AM	/search_result.php?search_words=\xcc\
2	109.82.36.41	10/28/2009 5:32:09 AM	/
3	109.82.78.127	11/1/2009 7:55:19 PM	/index.php?target=f012124b9e00f16e36
3	109.82.78.127	11/1/2009 7:58:50 PM	/staff_details.php?no=9000229&chk=28
3	109.82.78.127	11/1/2009 7:52:37 PM	/search_result.php?txt=10&ver=1&chk=
3	109.82.78.127	11/1/2009 7:41:24 PM	/
3	109.82.78.127	11/1/2009 7:45:22 PM	/vols.php
3	109.82.78.127	11/1/2009 7:52:03 PM	/author_result.php?search_words=sam
3	109.82.78.127	11/1/2009 7:44:09 PM	/vols.php
3	109.82.78.127	11/1/2009 7:54:26 PM	/index.php?target=5f70eeec24504a29dc
4	110.37.30.237	11/6/2009 9:52:30 AM	/search_result.php?txt=F&R1=f&chk=45
4	110.37.30.237	11/6/2009 9:52:38 AM	/iepngfix.htc
4	110.37.30.237	11/6/2009 9:52:48 AM	/index.php?target=7bfe58de0ad9dd370
4	110.37.30.237	11/6/2009 9:52:42 AM	/iepngfix.htc
5	110.37.63.23	11/7/2009 11:33:47 AM	/iepngfix.htc
5	110.37.63.23	11/7/2009 11:35:15 AM	/index.php
5	110.37.63.23	11/7/2009 11:35:38 AM	/index.php?target=70bc0b4dc4cafc98b0
5	110.37.63.23	11/7/2009 11:33:38 AM	/
6	110.8.8.18	6/4/2009 5:00:16 PM	/search_result.php?jour_no=http://212.
7	110.8.8.22	6/18/2009 9:56:51 AM	/search_result.php?jour_no=http://212.
8	112.104.4.185	10/24/2009 2:20:46 AM	/iepngfix.htc
8	112.104.4.185	10/24/2009 2:20:39 AM	/search_result.php?txt=R&R1=r&chk=6d

Figure 3.7: A Fragment from Users Session Result

3.5 PATTERN DISCOVERY

Various data mining techniques have been investigated for mining web usage logs. They are statistical analysis, clustering, association rule mining and classification.

3.5.1 Statistical Approach

The useful statistical information discovered from web logs are usually generated periodically in reports and used by administrators for improving the system performance, facilitating the site modification task, enhancing the security of the system, and providing support for marketing decisions.

3.5.2 Clustering

The web log file of SUST is taken as the input dataset. Clustering of web logs was based on the two types of clusters that can be found in web usage mining: user clusters and page clusters. User clusters will discover users having the same browsing patterns whereas page clusters will discover pages possessing similar content. IP or Agent represented attributes were used for user clusters, where a Requested Page was an attribute used for page clusters. Feature extraction or selection is one of the most important steps in pattern clustering. It is also an effective dimensionality reduction technique and an essential pre-processing method to remove noise features. In this experiment, we performed cfsSubsetEval feature selection method over log file dataset to select relevant features. The clustering algorithms were compared according to these factors: Cluster Instances, number of clusters, time taken to form clusters, incorrect cluster instances, number of iterations and accuracy. The proposed Working Scheme for Clustering and Association Rule Mining is shown in Figure 3.8.

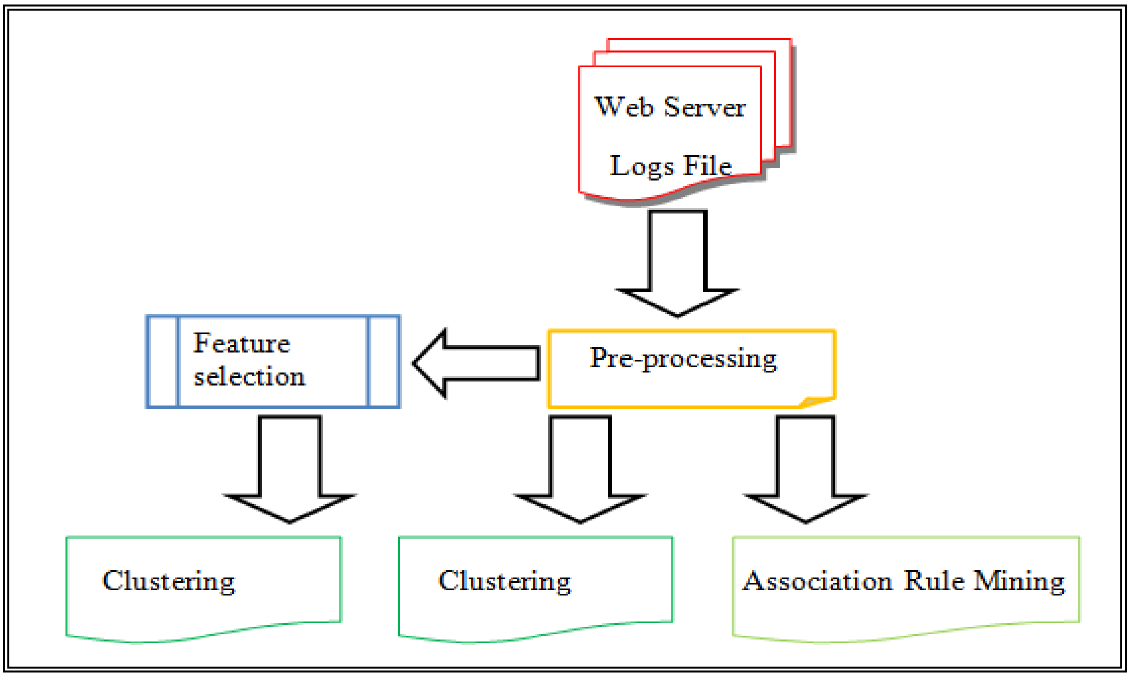


Figure 3.8: Architectural of Clustering and Association Rule Mining Working Scheme

3.5.3 Association Rule Mining

Then the A priori algorithm was applied on the log dataset. This algorithm was suitable for finding correlations between items and frequent patterns in large database. Setting parameter values in right way and Eliminating redundant rules and Page Clusters lead to interesting rule, which it is useful for analysis.

3.5.3.1 Setting Parameter Values

While conducting the experiments, we noticed that a lot of interesting rules contained item sets with support of less than 0.1, which is a default value in Weka tool. Based on our empirical research, we chose to set the minimum support of an item set to 0.07.

3.5.4 Classification Model

Classification was defined as the automated process of assigning a class label and mapping a user-based on the browsing history. The data were classified according to the predefined attributes. In this paper we consider four algorithms namely; J48, KNN, NB and BN. Combination of Multiple Classifiers (CMC) can be considered as a general solution method for the session classification.

The inputs of the CMC are results of separate classifiers and output of the CMC is their combined decisions [13,14]. Since the generalization ability of an ensemble could be significantly better than a single classifier, combinational methods have been a hot topic during the past years [15]. By combining classifiers, we intended to increase the performance of classification. There are several ways of combining classifiers. This work was done using voting majority method, which is the simplest way to find the best classifier as shown in Figure 3.9.

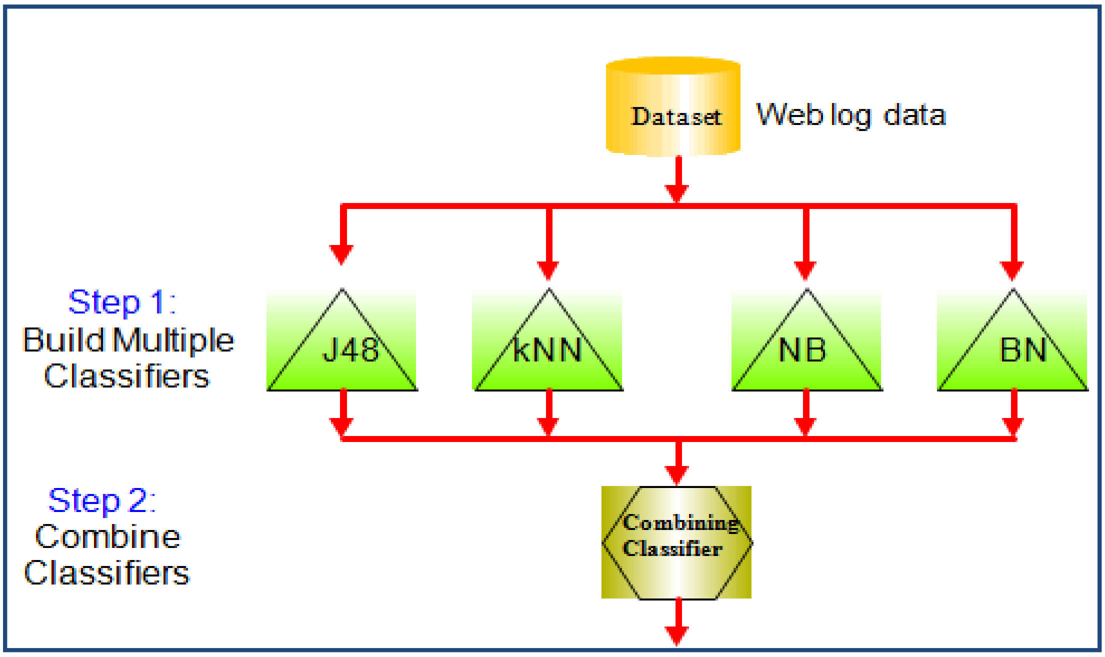


Figure 3.9: Majority Vote

In order to gauge the performance of ensemble techniques in the domain of web usage mining, we set up classification accuracy tests to compare ensembles against base classifier. Here we first compare the performance of base and Meta classifiers on training set. Then select the best classifier, we combine those classifiers to generate ensembles using the best Meta classifier method. If ensemble techniques were useful in this domain, then we would expect a higher level of classification accuracy. If classification accuracy does not increase, then the added complexity and computational overhead of using an ensemble of classifiers would outweigh the benefit.

3.1.1.1 Performance Measures

The performance of the classifiers is evaluated using the 10-fold cross-validation. In this research, we compared different classifiers, based on the measures of performance evaluation. According to Confusion matrix for two possible outcomes P (Positive) and N (Negative), as shown in Figure 3.10, many concepts are often used:

		Actual		
		P	N	Total
Predicted	P	True Positive (TP)	False Positive (FP)	P
	N	False Negative (FN)	True Negative (TN)	N
	Total	P	N	

Figure 3.10: Confusion Matrix for Two Possible Outcomes

i- Precision: Means the positive predictive value in information retrieved, which can be defined as:

$$\text{Precision} = TP / TP + FP \qquad \text{Eq. (1)}$$

ii- Recall: Proportion of actual positives, which are predicted positive.

$$\text{Recall} = TP / TP + FN \qquad \text{Eq. (2)}$$

iii- Accuracy: The Accuracy of a classifier on a given set is the percentage of test set tuples that are correctly classified by the classifier. Technically it can be defined as:

$$\text{Accuracy} = TP + TN / P + N \qquad \text{Eq. (3)}$$

iv- F-Measure: Other performance measures because the accuracy determined using equation 3 may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases.F-Measure is defined in equation 4.

$$\text{F-Measure} = 2\,(\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \qquad \text{Eq. (4)}$$

v- MCC: The Matthews correlation coefficient is used in machine learning as a measure of the quality of binary (two-class) classifications.

$$MCC = (TP*TN - FP*FN) / ((TP+FP)(TP+FN) + (TP+FP)(TN+FN))^{1/2} \qquad \text{Eq.(5)}$$

vi- ROC graphs: Are another way besides confusion matrices to examine the performance of classifiers. A ROC graph is a plot with the false positive rate on the X axis and the true positive rate on the Y axis. The point (0, 1) is the perfect classifier: it classifies all positive cases and negative cases correctly.

3.6 PATTERN ANALYSIS

Performing systematic analysis on such a huge amount of data is time consuming. Online Analytical Processing (OLAP) can be used for this purpose. The primary requirement in the construction of multidimensional data cube is the identification of dimensions and measures. In this research the web usage mining is analyzed by applying the pattern Analysis techniques on web log data. First, the dimensions and measures in web usage data warehouse are nominated and then a technique on how to apply (OLAP) on web usage data warehouse is proposed.

3.6.1 Data Warehouse Construction

To construct the data warehouse first we nominated 6 dimensions, these dimensions are time, Protocol type, Users, Agent, IP address and Pages. Each dimension will have primary key and other fields that can be used in the analysis. Second, we determined 2 facts; these facts are the number of visits, and the document size. The facts are used as fields in the fact table. The other fields of the fact table are foreign keys that can be used in constructing the relations with the dimension tables to yield a schema called the star schema as shown in Figure 3.11. The time dimension is designed to contain the hierarchies (Year, Month, Day, Hour, Minute, Second). Once the Data warehouse is constructed we can apply intelligent methods called OLAP or data mining techniques to extract data patterns. Generally OLAP is modelled by a multidimensional database structure called data cubes. Our constructed data warehouse can provide the data source for OLAP and if needed for the data mining techniques.

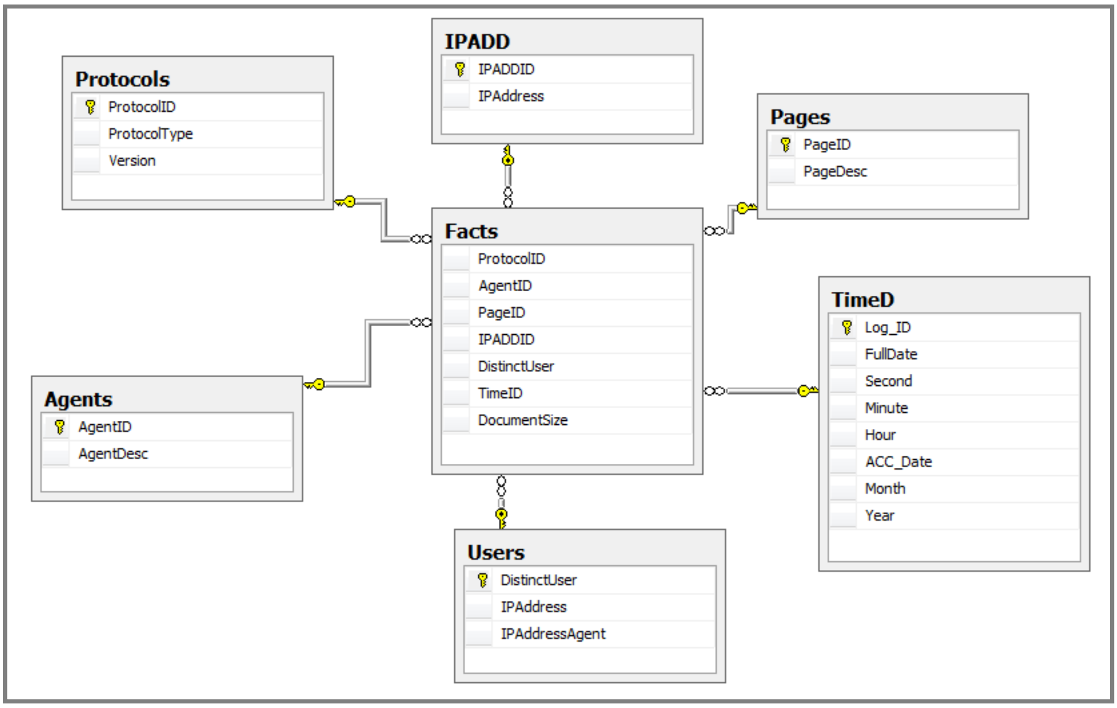


Figure 3.11: Web log Warehouse Schema

3.6.2 Filling the Dimension Table

Our dimension tables contain descriptive attributes, which are textual. These attributes are designed to for query constraint or filtering. Also they can be used to label the results in the OLAP cube. They are filled directly from the attributes of the log file (i.e. the time dimension is filled from the time attribute in the log file). To accomplish the filling task, we used SQL statements within a VB.net program connected to both the log file and SQL Server. The time attribute in the log file is divided into hours, minutes, seconds, Years, months, and days. The attributes of the other dimension tables are taken as they are from the log file (i.e the IP Address attribute is used to fill the IP address field in the IP address dimension).

3.6.3 Filling the Fact Table

To fill the fact table we need to reference our dimension tables in the query. We used a temporary table as main driver of the query and then we look up the resulting ID based on the primary keys of the dimension tables. The lookups are accomplished using the LEFT OUTER joins, which implies that the relationship may not exist in which case NULL value will go into the fact table.

3.6.4 OLAP

OLAP describes a set of technologies that allows analysts to quickly gain answers to the 'who' and 'what' questions premised on a, usually large, set of data. OLAP applications typically achieve this through multidimensional views of aggregate data derived from the data set. OLAP also answers tougher questions such as 'what if' and 'why' and this will be the emphasis of this paper. Some of the important questions are:

- Which are the top pages visited by user over the time?
- Which IP address accessed which site using which protocol and how many times?
- What is the distribution of network traffic over time (hour of the day, day of the week, month of the year)?

Answering the above questions require the inclusion of the time, IP address, Page dimensions and it requires the cube to render facts such as, the number of visitors, the document size by users or by IP address as shown in Figure 3.12.

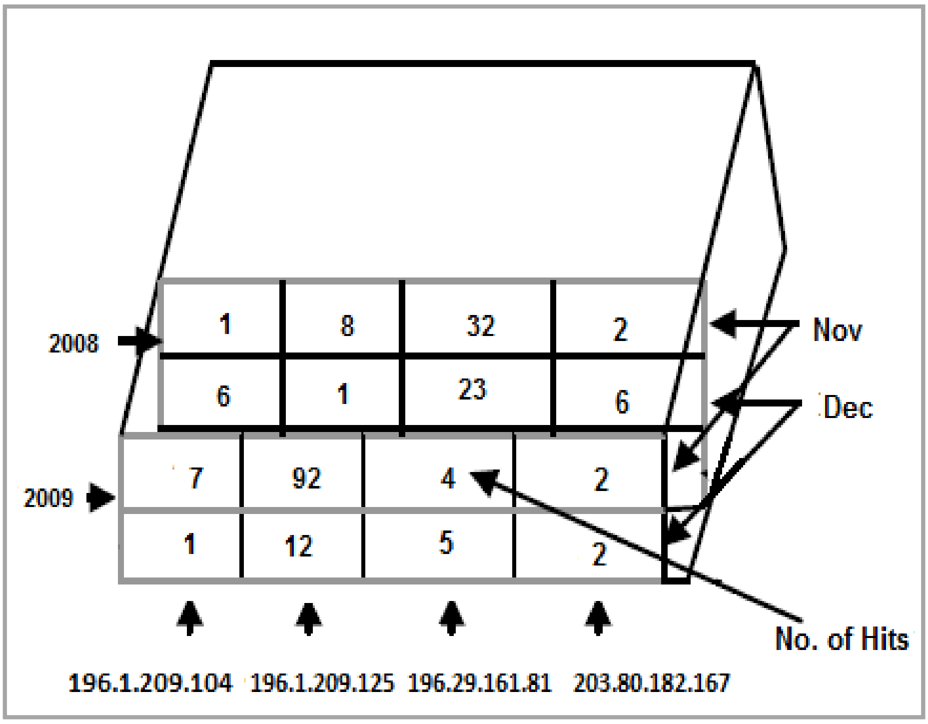


Figure 3.12: Data Cube.

3.6.5 WEKA Data Mining Software

We used WEKA software as the tool for clustering, feature selection, association rules and classification.

3.6.6 Business Intelligence Development Studio

Business Intelligence Development Studio is Microsoft Visual Studio 2008 with additional project types that are specific to SQL Server business intelligence. Business Intelligence Development Studio is the primary environment that will used to develop business solutions that include Analysis Services, Integration Services, and Reporting Services projects [105]. Each project type supplies templates for creating the objects required for business intelligence solutions, and provides a variety of designers, tools, and wizards to work with the objects. We used visual studio to construct the data cubes. This accomplished by: linking database, determined dimension, determined facts table and then running the cube.

3.7 CHAPTER SUMMARY

This chapter has detailed the thesis's theoretical and practical approach and rationalizes the different decisions and processes undertaken throughout the research journey. Also the chapter provides a detailed discussion of a host of activities and techniques used at different stages of this cycle. Firstly the pre-processing of data from SUST is necessary to convert the data into suitable form for pattern discovery. This phase contains three sub steps: Data Cleaning, User Identification, and Session Identification. Secondly pattern discovery and analysis techniques that are typically applied to this cleaned data. The Method we have detailed show how pattern discovery techniques such as clustering, association rule mining, and classification algorithms used, data warehouse and OLAP performed on Web usage data. Feature extraction or selection is one of the most important steps in pattern clustering. It is also an effective dimensionality reduction technique and an essential pre-processing method to remove noise features. To provide the most useful and effective result, ensemble method need to incorporate classification algorithms.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter presents experimental results using all the algorithms described earlier. The first section describes the details of the information in the log file has to be written in a specific format; that is in a specific sequence and in a certain way that will facilitate the analysis of the file. The whole log file can be read in one variable and then can move this variable line by line using a loop statement and stored in a database. This is followed by experimental results of data Pre-Processing, data clustering, an association rule mining and classification techniques. Finally, the last sections present result of an Online Analytical Processing (OLAP) was used to analyze the data in the data warehouse. In the chapter we summarize the results and explain the experiments, we have conducted to measure the effectiveness of the proposed method.

4.2 SOURCE OF THE DATA

Our experiments were performed on a 2.8GHz Pentium CPU, 2GB of main memory, Windows 7 Ultimate, SQL Server 2008 and Microsoft Visual Studio 2010. As the developed system is to be used to identify trends of visitor website behaviour within the SUST web site applications from 7/Nov/2008 through 20/Dec/2009. A portion of the log file used for the experimentation, which has been shown in Figure 3.2 in section 3.1.

4.3 DATA PRE-PROCESSING

After reading the log files, several attributes are considered important for the analysis. The read logs records will be stored in a database. After data cleaning only 122122 entries are left in the log. Table 4.1 shows the comparison between size and number of records before and after data cleaning. Figure 4.1 and Figure 4.2 illustrate the change in the number of records and log file size, respectively.

Table 4.1: Size and Number of Records before and after Cleaning

	Size(MB)	Number of records
Before	567	291642
After	143	122122
Percentage in Reduction	74.77%	58.13%

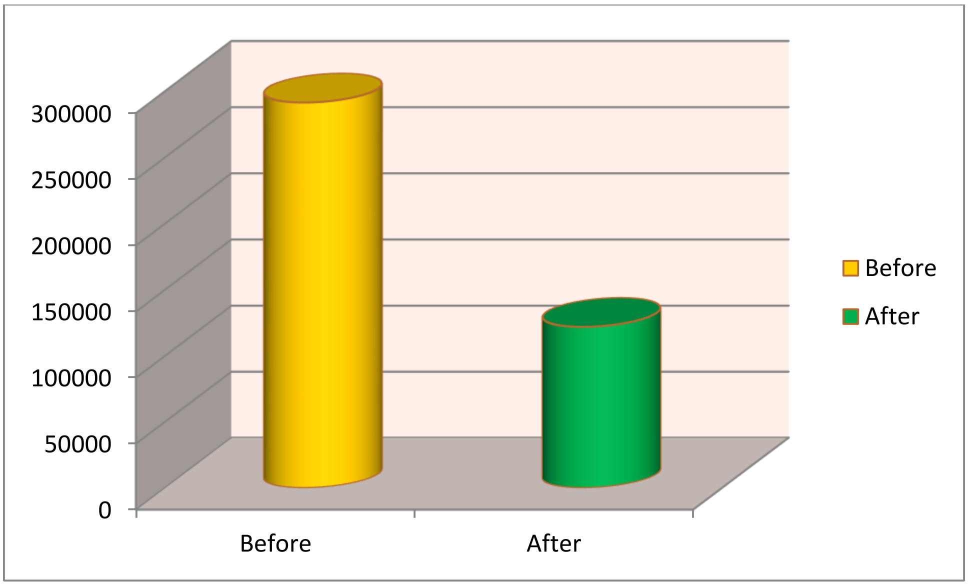


Figure 4.1: Bar Chart Showing the Change in Number of Records

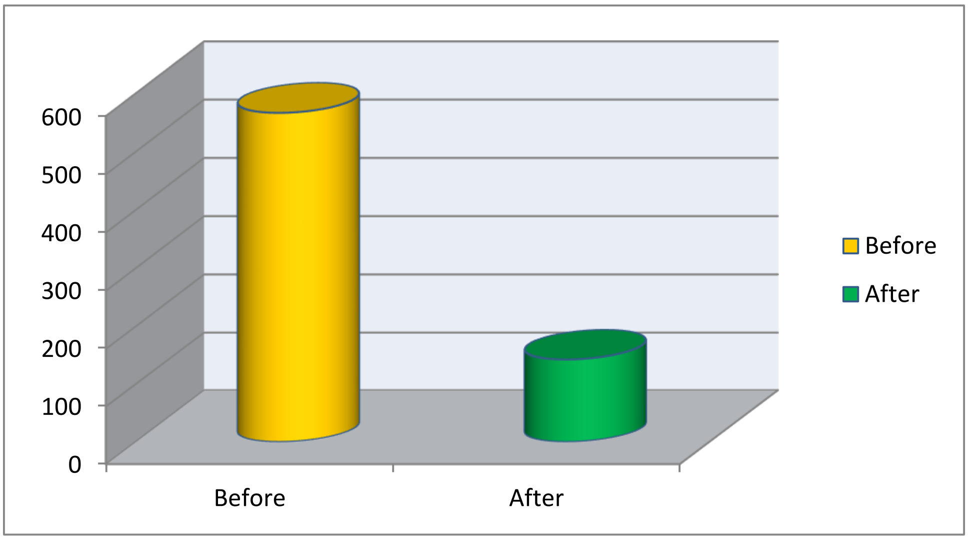


Figure 4.2: Bar Chart Showing the Change in the Size of Log File.

Table 4.2 shows the details of our data during the User and session's identification process. There were a total of 23200 unique visiting IP addresses, 8861 unique pages and 13869 sessions.

Table 4.2: Statistical Summery

Number of unique Users	23200
Number of Unique IP address	11030
Number of unique pages	8861
Number of sessions	13869

The result in Figure 4.3 shows that a significant number of sessions only consist of one or two request. In addition, there are not too many web user sessions, which extend over five visits. The vertical axis stands for the percentage of occurrence of the number of session length. The horizontal axis is marked with the length of session.

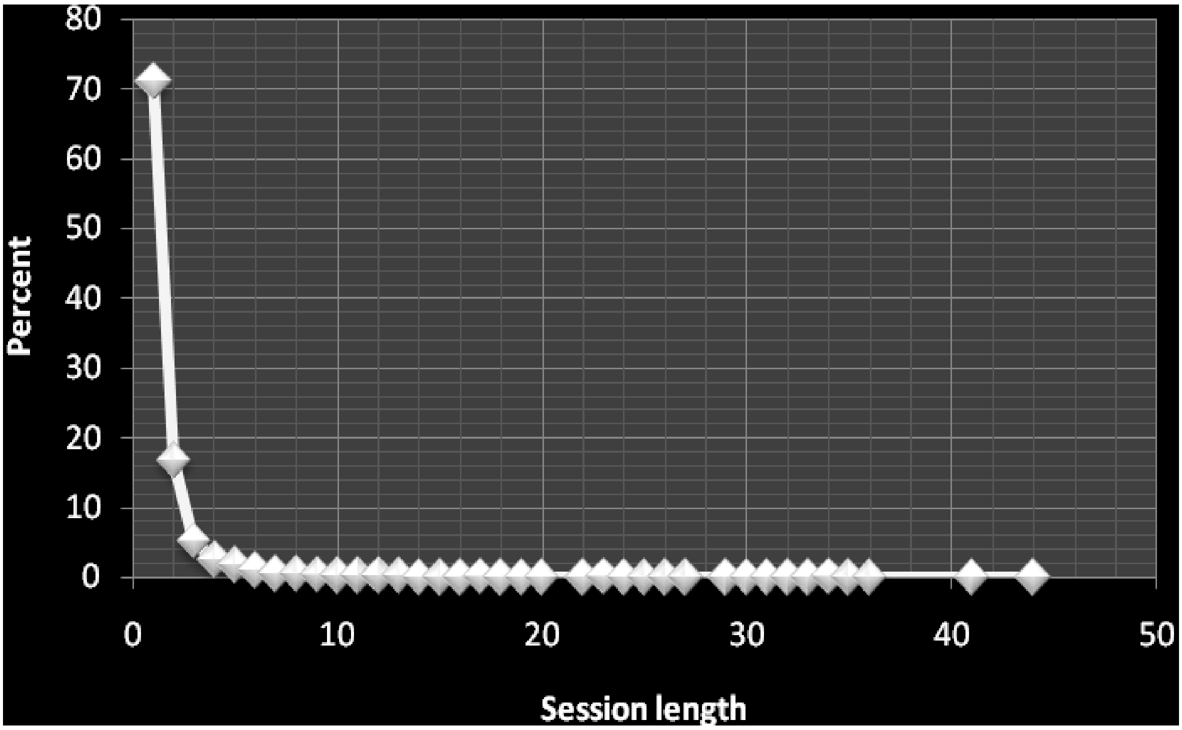


Figure 4.3: The Session Length Distribution of Dataset

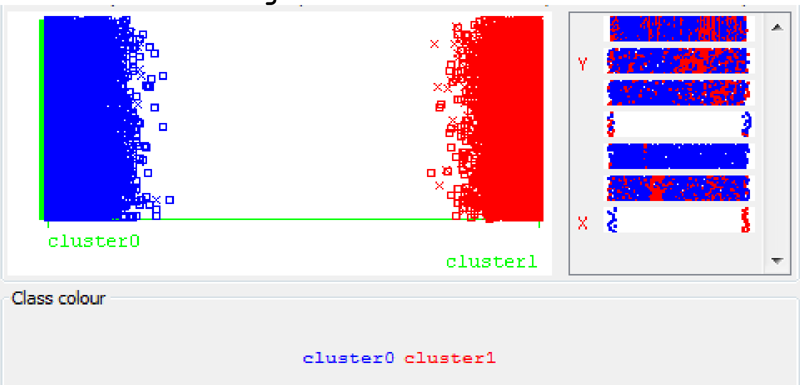
4.4 PATTERN DISCOVERY

4.4.1 Clustering

The K-mean algorithm and Density-based clustering were used to obtain the clusters using WEKA Clustering Tool on a set of Pre-processed log file. The output for the data set for user cluster with two clusters using K-mean is shown in Figure 4.4. Figure 4.5 shows output for the same data set with two clusters using Density-based clustering. Figure 4.6 and Figure 4.7 show the content of each cluster.

```
Time taken to build model (full training data): 0.58 seconds
=== Model and evaluation on training set ===
Clustered Instances

0        16153 (69%)
1         7089 (31%)
```




```
Class attribute: agent
Classes to Clusters:

 0    1 <-- assigned to cluster
 0    1 | Mozilla/2.0 (compatible; AOL 3.0; Mac_PowerPC)
 0    1 | Mozilla/2.0 (compatible; MSIE 3.0B; Win32)
 1    0 | Mozilla/3.01 (compatible; AmigaVoyager/2.95; AmigaOS/MC680x0)
 0    1 | Mozilla/3.x (I-Opener 1.1; Netpliance)
 0    1 | Mozilla/4.0 (compatible; MSIE 5.0; Windows NT; Girafabot; girafabot at girafa dot com
 1    0 | Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0; NetCaptor 6.5.0RC1)
 0    1 | Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; SAFEXPLORER TL)
 0    1 | Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x 4.90; MSIECrawler)
 0    1 | Mozilla/4.0 (MobilePhone PM-8200/US/1.0) NetFront/3.x MMP/2.0
 0    1 | Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.2b) Gecko/20021007 Phoenix/0.3

Cluster 0 <-- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
Cluster 1 <-- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET
CLR 1.1.4322; .NET CLR 2.0.50727)

Incorrectly clustered instances :   17781.0     76.5073 %
```

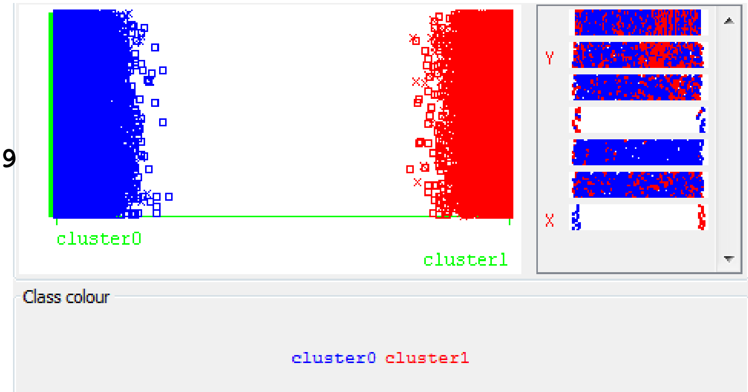
Figure 4.4: User Cluster using K-Means Clustering

```
Time taken to build model (full training data) : 0.37 seconds
=== Model and evaluation on training set ===
Clustered Instances

0        16269 (70%)
1         6973 (30%)

Log likelihood: -18.67379

Class attribute: agent
Classes to Clusters:
```




```
0   1 <-- assigned to cluster

0   1 | Mozilla/2.0 (compatible; AOL 3.0; Mac_PowerPC)
0   1 | Mozilla/2.0 (compatible; MSIE 3.0B; Win32)
1   0 | Mozilla/3.01 (compatible; AmigaVoyager/2.95; AmigaOS/MC680x0)
0   1 | Mozilla/3.x (I-Opener 1.1; Netpliance)
0   1 | Mozilla/4.0 (compatible; MSIE 5.0; Windows NT; Girafabot; girafabot at girafa dot com
1   0 | Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0; NetCaptor 6.5.0RC1)
0   1 | Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; SAFEXPLORER TL)
0   1 | Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x 4.90; MSIECrawler)
0   1 | Mozilla/4.0 (MobilePhone PM-8200/US/1.0) NetFront/3.x MMP/2.0
0   1 | Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.2b) Gecko/20021007 Phoenix/0.3
Cluster 0 <-- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
Cluster 1 <-- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET
CLR 1.1.4322; .NET CLR 2.0.50727)

Incorrectly clustered instances :  17775.0     76.4779 %
```

Figure 4.5: User Cluster using Density Based Algorithm

Cluster 0 <-- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)

- **Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; InfoPath.2)**
- **Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)**
- **Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 2.0.50727)**
- **Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)**
- **Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.10) Gecko/2009042316 Firefox/3.0.10**
- **Opera/9.80 (Windows NT 5.1; U; en) Presto/2.2.15 Version/10.00**
- **Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 2.0.5072**
- **Mozilla/5.0 (compatible; heritrix/1.14.3 +http://www.accelobot.com)**

Figure 4.6: Content of Cluster 0

Cluster 1 <-- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)

- **Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.9) Gecko/2009040821 Firefox/3.0.9**
- **Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;1813)**
- **Mozilla/3.0 (compatible; Indy Library)**
- **Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB6; .NET CLR 1.1.4322; .NET CLR 2.0.50727; InfoPath.1)**
- **msnbot/2.0b (+http://search.msn.com/msnbot.htm)**

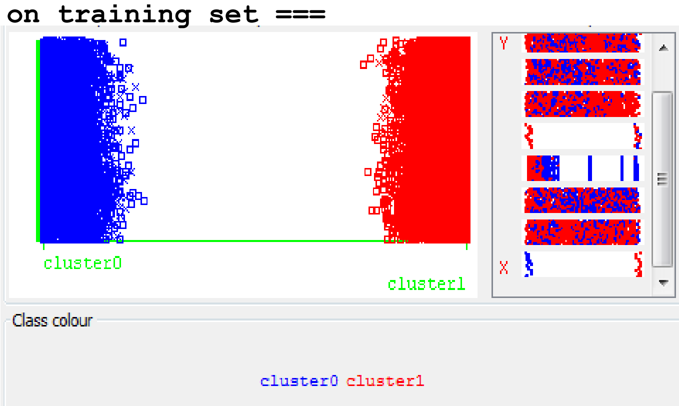
Figure 4.7: Content of Cluster 1

Figure 4.8, Figure 4.9 and Figure 4.10 show the clusters according to the page request using K-mean algorithm with 2, 3 and 4 clusters respectively. Figure 4.11 shows the Page Clustering using Density-based Algorithm with 2 Clusters.

```
Time taken to build model (full training data): 0.38 seconds
=== Model and evaluation on training set ===
Clustered Instances

0        7764 (33%)
1       15478 (67%)

Class attribute: URL
Classes to Clusters:
```




```
 0    1 <-- assigned to cluster
 0    1 | /guide.php?target=http://217.218.225.2:2082/index.html?
 0    1 | /index.php?target=http://amyru.h18.ru/images/cs.txt?
 0    1 | /staff_details.php?staff_no=http://mypregnancy.orgfree.com/index.html?
 0    1 | /search_result.php?txt=C&R1=c&chk=9c...9ea86c65330f3f39e6f463305
 0    1 | /search_result.php?txt=S&R1=s&chk=78...f0b4eaa49cb5188f53feeae43
 0    1 | /search_result.php?txt=A&R1=a&chk=37...19d1990bd4300e4f6e5e90e6e
 0    1 | /index.php?target=http://xishisniceplace.chat.ru/images?
 0    1 | /index.php?target=http://ninaru.hut2.ru/images/cs.txt?
Cluster 0 <-- /
Cluster 1 <-- /info.php
Incorrectly clustered instances :   17764.0      76.4306 %
```

Figure 4.8: Page Cluster using K-Mean Algorithm with 2 Clusters

```
Time taken to build model (full training data): 0.52 seconds
=== Model and evaluation on training set ===
Clustered Instances

0        6677 (29%)
1       12528 (54%)
2        4037 (17%)

Class attribute: URL
Classes to Clusters:
```


```
0    1    2 <-- assigned to cluster
 0    1    0 | /index.php?target=http://amyru.h18.ru/images/cs.txt?
 0    1    0 | /staff_details.php?staff_no=http://mypregnancy.orgfree.com/index.html?
 0    1    0 | /search_result.php?txt=C&R1=c&chk=9c...9ea86c65330f3f39e6f463305
 0    1    0 | /search_result.php?txt=S&R1=s&chk=78...f0b4eaa49cb5188f53feeae43
 0    1    0 | /search_result.php?txt=A&R1=a&chk=37...19d1990bd4300e4f6e5e90e6e
 0    1    0 | /index.php?target=http://xishisniceplace.chat.ru/images?
Cluster 0 <-- /
Cluster 1 <-- /info.php
Cluster 2 <-- /index.php?jour_no=1
Incorrectly clustered instances:  17988.0     77.3944 %
```

Figure 4.9: Page Cluster using K-Mean Algorithm with 3 Clusters

```
Time taken to build model (full training data): 0.50 seconds
=== Model and evaluation on training set ===
Clustered Instances

0       7588 (33%)
1       9005 (39%)
2       3921 (17%)
3       2728 (12%)

Class attribute: URL
Classes to Clusters:
```


```
 0    1    2    3 <-- assigned to cluster
 1    0    0    0 | /search_result.php?jour_no=1&R1=v2&txt=sail&B2=search
 0    1    0    0 | /index.php?target=http://babycaleb.fortunecity.co.uk/index.htm?
 0    0    1    0 | /search_result.php?jour_no=1&R1=v1&txt=%26%231581%3B%26%231580%
 0    1    0    0 | /staff_details.php?jour_no=1&staff_no=9000125
 0    1    0    0 | /search_result.php?txt=M&R1=m&chk=6764879999733c187c5111bddf5ae
 0    1    0    0 | /search_result.php?jour_no=http://google.com
Cluster 0 <-- /
Cluster 1 <-- /info.php
Cluster 2 <-- /index.php?jour_no=1
Cluster 3 <-- /iepngfix.htc
Incorrectly clustered instances: 18369.0     79.0336 %
```

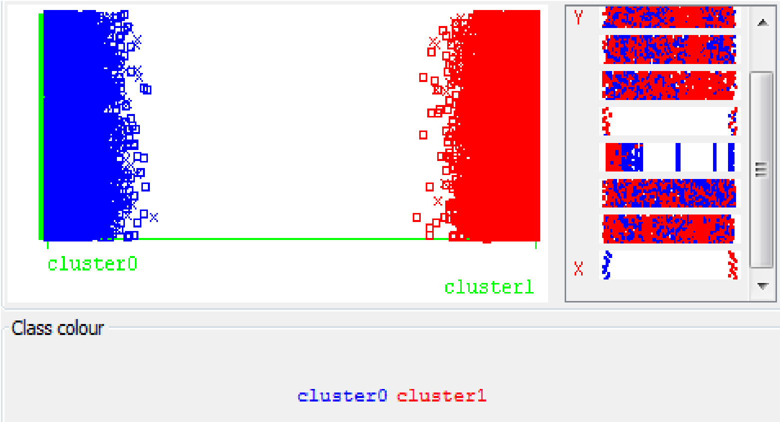
Figure 4.10: Page Cluster using K-mean Algorithm with 4 Clusters

```
Time taken to build model (full training data): 0.42 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       7318 (31%)
1      15924 (69%)
```




```
Class attribute: agent
Classes to Clusters:

 0    1 <-- assigned to cluster
 1    0 | /more_details.php?jour_no=1&id=190&chk=0189f5f4f6ede1d0557fd4f299c9ac73
 0    1 | /search_result.php?jour_no=&R1=v1&txt=%26%231575%3B%26%231604%3B%26%231602%3
 0    1 | /guide.php?target=http://217.218.225.2:2082/index.html?
 0    1 | /index.php?target=http://amyru.h18.ru/images/cs.txt?
 0    1 | /staff_details.php?staff_no=http://mypregnancy.orgfree.com/index.html?
 0    1 | /search_result.php?txt=C&R1=c&chk=9c...9ea86c65330f3f39e6f463305
 0    1 | /search_result.php?txt=S&R1=s&chk=78...f0b4eaa49cb5188f53feeae43
 0    1 | /search_result.php?txt=A&R1=a&chk=37...19d1990bd4300e4f6e5e90e6e
 0    1 | /index.php?target=http://xishisniceplace.chat.ru/images?
 0    1 | /index.php?target=http://ninaru.hut2.ru/images/cs.txt?

Cluster 0 <-- /
Cluster 1 <-- /info.php

Incorrectly clustered instances :  17748.0     76.3618 %
```

Figure 4.11: Page Cluster using Density Based Algorithm with 2 Clusters.

According to the previous implementation of the data clustering techniques, the two clustering algorithms are compared according to these factors: Cluster Instances, Number of clusters, Time taken to form clusters, Incorrect cluster Instances, Number of Iterations and Accuracy. It is useful to summarize the results and present some comparison of performances. A summary of the best-achieved results for each of the two techniques is presented in Table 4.3.

Table 4.3: Performance Results Comparison

Algorithm	No. of clusters	Cluster Instances	Time taken to build model	Incorrect cluster instances	No. of iterations	Within cluster some of squared errors
K-Means	2	23242	0.38	17764.0 (76.4306%)	8	59696.305
	3	23242	0.52	17988.0 (77.3944%)	10	55395.901
	4	23242	0.50	18369.0 (79.0336%)	12	54661.442
Density based clustering	2	23242	0.42	17748.0 (76.30618%)	8	59696.305

From this comparison we can conclude that Density-based clustering with 2 clusters produces fairly higher accuracy than K-means technique with 2, 3 and 4 clusters and requires significant computation and RMSE.

4.4.2 Clustering with and without Feature Selection

In this experiment we performed cfsSubsetEval feature selection evaluator and over log file dataset to select relevant features. It evaluates a subset of attributes which are more relevant for the requested page (URL) attribute. It selected only two attributes: IP address (IPADD) and Referred (WURL) form 6 attributes (see Figure 4.12). Then, we performed K-means, and Density-based clustering methods on this subset (see Figure 4.13 and Figure 4.14). Then we compared the result of clustering method with and without feature selection, as shown in Table 4.4

```
Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     FldCompLast-weka.filters.unsupervised.attribute.Remove-R1-weka.filters
Instances:    23242
Attributes:   6
              IPADD
              webextension
              wprotocol
              ByteSize
              wurl
              agent
Evaluation mode:    evaluate on all training data

 === Attribute Selection on all input data ===

 Search Method:
         Best first.
         Start set: no attributes
         Search direction: forward
         Stale search after 5 node expansions
         Total number of subsets evaluated: 19
         Merit of best subset found:    0.616

 Attribute Subset Evaluator (supervised, Class (nominal): 2 webextension):
         CFS Subset Evaluator
         Including locally predictive attributes

 Selected attributes: 1,5 : 2
                      IPADD
                      wurl
```

Figure 4.12: Features selected using Filtering Technique in WEKA.

```
Time taken to build model (full training data): 0.21 seconds
=== Model and evaluation on training set ===
Clustered Instances

0       15898  (68%)
1        7344  (32%)


Class attribute: URL
Classes to Clusters:
```


```
    0    1   <--assigned
    1    0 | /author_result.php?search_words=sami&option=Author
    1    0 | /search_result.php?jour_no=1&R1=v8&txt=&B2=search
    1    0 | /staff_details.php?no=9000107&chk=4ea1d018808aa24c7b565f99213ad14e
    1    0 | /search_result.php?search_words=Tomato&option=Title
    0    1 | /search_result.php?search_words=journal%20scince%20in%20foods&option=Title
    0    1 | /search_result.php?search_words=tomatos&option=Title
    1    0 | /content_details.php?id=111&chk=562052e41c499bc9cc23215b5e0a6fb7
    1    0 | /staff_details.php?staff_no=9000056

Cluster 0 <-- /
Cluster 1 <-- /info.php
Incorrectly clustered instances :   16925.0      72.8208 %
```

Figure 4.13: K-means Method with Feature Selection on Log File Dataset

```
Time taken to build model (full training data): 0.16 seconds
=== Model and evaluation on training set ===
Clustered Instances

0       15993 (69%)
1        7249 (31%)
```


```
Log likelihood: -13.87602
Class attribute: URL
Classes to Clusters:
0    1 <-- assigned to cluster
 0    1 | /index.php?target=http://amyru.h18.ru/images/cs.txt?
 0    1 | /staff_details.php?staff_no=http://mypregnancy.orgfree.com/index.html?
 0    1 | /search_result.php?txt=C&R1=c&chk=9c...9ea86c65330f3f39e6f463305
 0    1 | /search_result.php?txt=S&R1=s&chk=78...f0b4eaa49cb5188f53feeae43
 0    1 | /search_result.php?txt=A&R1=a&chk=37...19d1990bd4300e4f6e5e90e6e
 0    1 | /index.php?target=http://xishisniceplace.chat.ru/images?
Cluster 0 <-- /
Cluster 1 <-- /info.php
Incorrectly clustered instances :  16887.0     72.6573 %
```

Figure 4.14: Density Based with Feature Selection on Log File Dataset.

Table 4.4: Clustering Method with and without Feature Selection.

Algorithm / Factor	K-means without feature selection.	K-means with feature selection.	Density based without feature selection.	Density based with feature selection.
Incorrectly clustered instance	17764.0 (76.4306%)	16925.0 (72.8208%)	17748.0 (76.3618%)	16887.0 (72.6537%)
Time taken to build model (seconds)	0.52	0.21	0.42	0.16
Number of iteration	8	3	8	3
Within cluster sum of squared errors	59696.305	37504.0	59696.305	37504.0

The accuracy of clustering algorithms in terms of correctly classified instances for Dataset is shown in Figure 4.15.

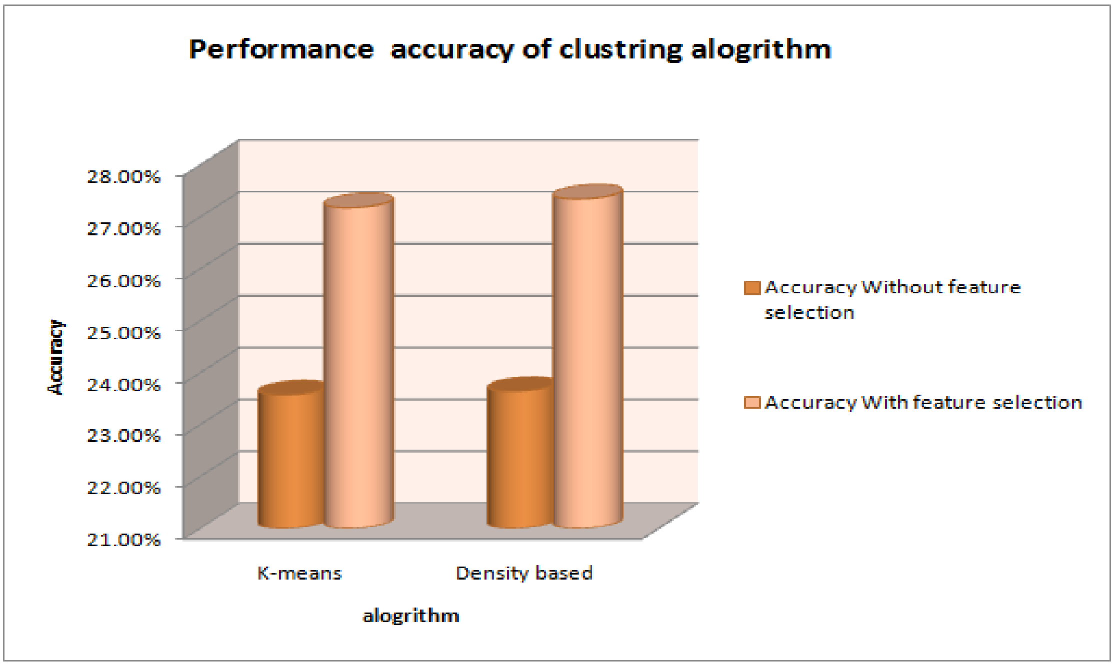


Figure 4.15: Clustering Performance

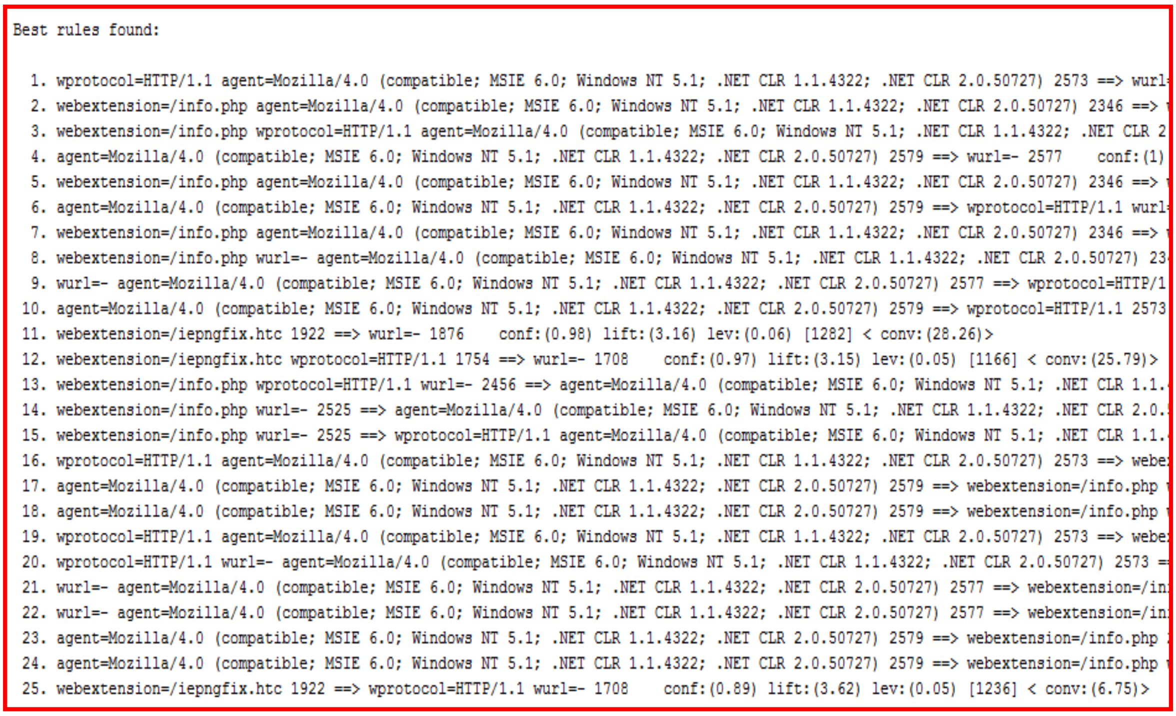
For K-means and Density-based clustering, we observed that time was reduced to 0.21 and 0.16 and accuracy increased to 27.18% and 27.35% respectively. K-means and Density-based clustering method, within cluster sum of square errors, was reduced to 37504.0. Also for two algorithms, the number of iterations was reduced to 3.

4.4.3 Association Rule Mining

Association rule mining aims to extract interesting correlations, frequent patterns and associations or casual structures among sets of items in the SUST log file. Based on our empirical research we chose to set the minimum support of an item set to 0.07. Eliminating redundant rules and identify page Clusters lead to an interesting rule, useful for analysis.

4.4.3.1 The Generated Rule Set

In accordance with our expectations, the initially generated association rule set contained many rules that had very high confidence. There were 10 (out of 50) rules with confidence equal to 1.0, while 29 (out of 50) rules had confidence greater than 0.85. This can be explained by the fact that many web pages are strongly correlated due to the link structure of the website. Figure 4.16 shows some results of association rule mining.



```
Best rules found:

 1. wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> wurl
 2. webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==>
 3. webextension=/info.php wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2
 4. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> wurl=- 2577    conf:(1)
 5. webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==>
 6. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> wprotocol=HTTP/1.1 wurl
 7. webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==>
 8. webextension=/info.php wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 23
 9. wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2577 ==> wprotocol=HTTP/1
10. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> wprotocol=HTTP/1.1 2573
11. webextension=/iepngfix.htc 1922 ==> wurl=- 1876    conf:(0.98) lift:(3.16) lev:(0.06) [1282] < conv:(28.26)>
12. webextension=/iepngfix.htc wprotocol=HTTP/1.1 1754 ==> wurl=- 1708    conf:(0.97) lift:(3.15) lev:(0.05) [1166] < conv:(25.79)>
13. webextension=/info.php wprotocol=HTTP/1.1 wurl=- 2456 ==> agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.
14. webextension=/info.php wurl=- 2525 ==> agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.
15. webextension=/info.php wurl=- 2525 ==> wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.
16. wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> webe
17. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
18. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
19. wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> webe
20. wprotocol=HTTP/1.1 wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 =
21. wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2577 ==> webextension=/in
22. wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2577 ==> webextension=/in
23. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
24. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
25. webextension=/iepngfix.htc 1922 ==> wprotocol=HTTP/1.1 wurl=- 1708    conf:(0.89) lift:(3.62) lev:(0.05) [1236] < conv:(6.75)>
```

Figure 4.16: Some Results of Association Rule Mining

4.4.3.2 Eliminating Redundant Rules

As a first step, and after removing redundant rules, our rule set contained 43 rules out of the 50 rules generated originally. Some redundant rules are selected in Table 4.5.

Table 4.5: Some Redundant Rules.

The Rule	**Redundant Rule**
/iepngfix.htc ==> wurl=-	/iepngfix.htc ,HTTP/1.1 ==> wurl=-
/iepngfix.htc ==> HTTP/1.1	/iepngfix.htc, wurl=- ==> HTTP/1.1
/info.php ==> wurl=-	/info.php ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> wurl=-
wurl=- ==> /info.php	wurl=- ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php
wurl=- ==> /info.php ,HTTP/1.1	wurl=- ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php ,HTTP/1.1

4.4.3.3 Identifying Page Clusters

We eliminated 10 rules and introduced 5 rules as their cluster presentation (1 for each cluster), thus decreasing the size of the rule set by 38 rules (out of 43). For example, we eliminated four rules and introduced their cluster representatives, as shown in Table 4.6. The confidence of all eliminated rules was close to 1.

Table 4.6: Rules and Clusters

Number of Cluster	Rules and Cluster
1	info.php, HTTP/1.1 ==> Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.8) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php ,HTTP/1.1 conf:(0.91)
2	/info.php ,HTTP/1.1, wurl=- ==> Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.95) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php, HTTP/1.1, wurl=- conf:(0.91)
3	/info.php, wurl=- ==> Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.93) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php, wurl=- conf:(0.91)
4	/info.php ,wurl=- ==> HTTP/1.1 ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.93) HTTP/1.1,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php ,wurl=- conf:(0.91).

4.4.3.4 Interestingness of the Resulting Association Rules

Pruning our rule set according to redundant rules and using Clustering to decrease the size of the rule set from 50 to only 38 rules, we identified a webmaster to enhance the website structure and improve its browsing experience for the visitors. We identified 8 truly interesting rules out of the 38 rules in the rule set (21%). Some of the interesting rules are shown in Table 4.7.

Table 4.7: Some Association Rules

Number	association rule of homepage
1	webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==> wurl=- 2346 <conf:(1)> lift:(3.24)
2	wurl=http://www.sustech.edu/sudannewar/staff_publicationsAR.php 169 ==> web extension=/ 169 <conf:(1)> lift:(5.06)
3	web extension=/iepngfix.htc agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) 282 ==> web url=- 274 <conf:(0.97)> lift:(3.15)
4	HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> web extension=/info.php 2343 <conf:(0.91)> lift:(6.89)

The first interesting rule found was that the information page is accessed with the most using agents: *Mozilla/4.0 + Windows NT 5.1* and the referrer was *‘-‘*. This indicates that the request made to the information page was from regular visitors who know the website well. The second association rule shows that if a user referrer is *http://www.sustech.edu/sudannewar/ staff_publicationsAR.php*, then they will very likely request *web extension=/*. The third association rule was found by a priori algorithm. It is an interesting rule which can be stated as: if visitors visit the “/*iepngfix.htc*" page with platform *Mozilla/4.0*, then they will be referrer‘-‘. This means that the request made to the “*/iepngfix.htc*” page is from the regular visitors who use *Mozilla/4.0* as agent.

The fourth association rule shows that, the number of requests made to” */info.php*” page was from web protocol=*HTTP/1.1*, agent=*Mozilla/4.0*. This indicates that these users visited the information home page almost use platforms *Mozilla/4.0* and *Windows NT 5.1* and also used the *HTTP/1.1* protocol.

4.4.4 Classification

A log file data with approximately 23242 entries was classified according to the predefined attributes, such as the pages visited by each user categorized into two sessions namely; forenoon (form 00:00:00 to 11:59:59) and afternoon (form 12:00:00 to 23:59:59). Figure 4.17 explains the number of entries classified into forenoon and afternoon.

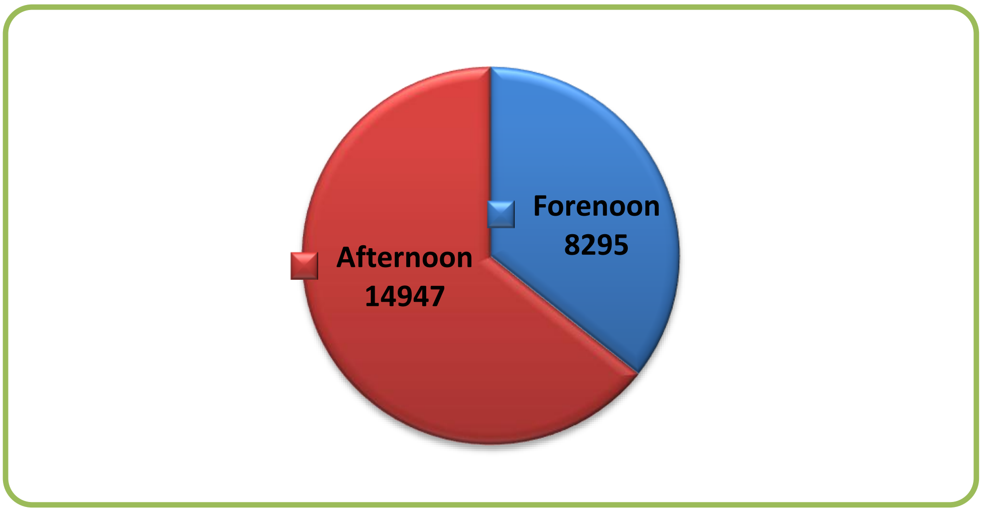


Figure 4.17: Users Count in Each Session

We compared the performance of Decision Tree Classifier (J48), K-Nearest Neighbor Classifier (KNN), Naïve Bayesian Classifier (NB), K-Nearest Neighbor Classifier (KNN) and BayesNet classifier (BN). The results were displayed in form of tables. The comparison of accuracy, time and kappa statistic is presented in Table 4.8. Table 4.9 shows the mean absolute error (MAE) and the root relative squared error (RMSE).Meanwhile, Table 4.10 shows the result based on recall, precision, F-measure, MCC, Roc Area and Error Rate. Figure 4.18 shows the obtained accuracy using different classification techniques. Figure 4.19 shows the performance metrics on balance-scale. The result inferred is that BayesNet classifier outperformed the others: base and Meta classifiers with MAE = 0.3218 and 73.4274 % correctly classified. The Stacking Meta classifiers had the same results with Voting, but it will take longer time to build model.

Table 4.8: Comparison of Different Classifiers for Base and Meta Classifiers.

Algorithm	Correctly Classified Instances (% Value)	Incorrectly Classified Instances (% Value)	Time Taken to build model (in seconds)	Kappa Statistic
J48	14947 (64.3103 %)	8295 (35.6897%)	5.14	0
KNN	16192 (69.667 %)	7050 (30.333 %)	0.04	0.2661
NB	16895 (72.6917 %)	6347 (27.3083 %)	0.09	0.4038
BN	**17066 (73.4274 %)**	**6176 (26.5726 %)**	**0.04**	**0.4379**
Stacking	14947 (64.3103 %)	8295 (35.6897%)	0.16	0
Voting	14947 (64.3103 %)	8295 (35.6897%)	0.01	0

Table 4.9: The MAE and RMSE for each Base and Meta Classifier.

Base and Meta Classifier.	MAE	RMSE
J48	0.459	0.4791
KNN	0.373	0.4432
NB	0.3373	0.4137
BN	**0.3218**	**0.4106**
Meta Classifiers	0.459	0.4791

Table 4.10 : The Classification Performance of each Base and Meta Classifier

Parameters / Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Error Rate
J48	0.643	0.643	0.414	0.643	0.503	0.000	0.500	0.541	0.357
KNN	0.697	0.457	0.685	0.697	0.670	0.289	0.723	0.741	0.303
NB	0.727	0.324	0.726	0.727	0.727	0.404	0.799	0.810	0.273
BN	**0.734**	**0.283**	**0.744**	**0.734**	**0.737**	**0.440**	**0.814**	**0.825**	**0.266**
Meta Classifiers	0.643	0.643	0.414	0.643	0.503	0.000	0.500	0.541	0.357

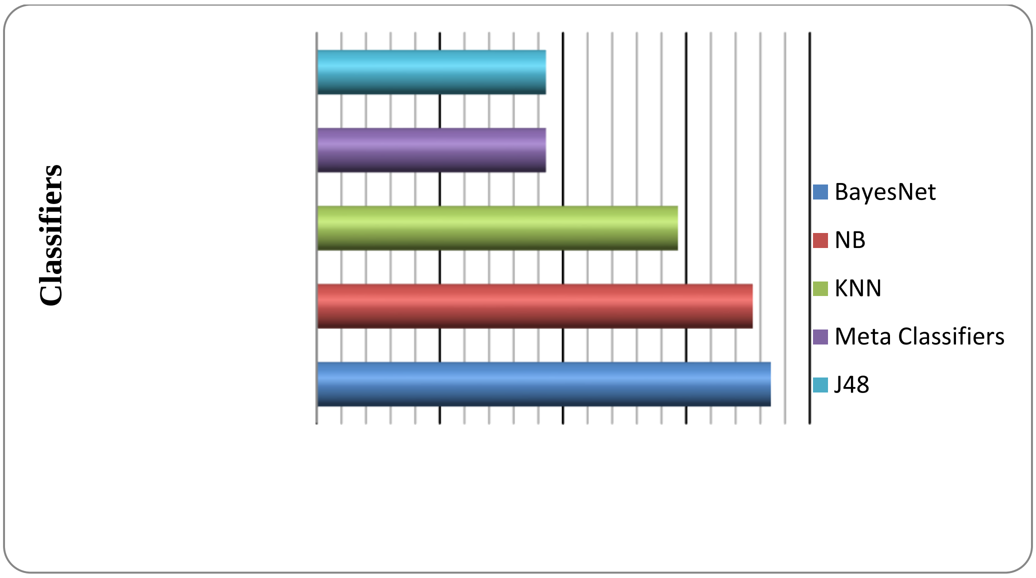


Figure 4.18: Comparison between Accuracy using Different Classification Techniques

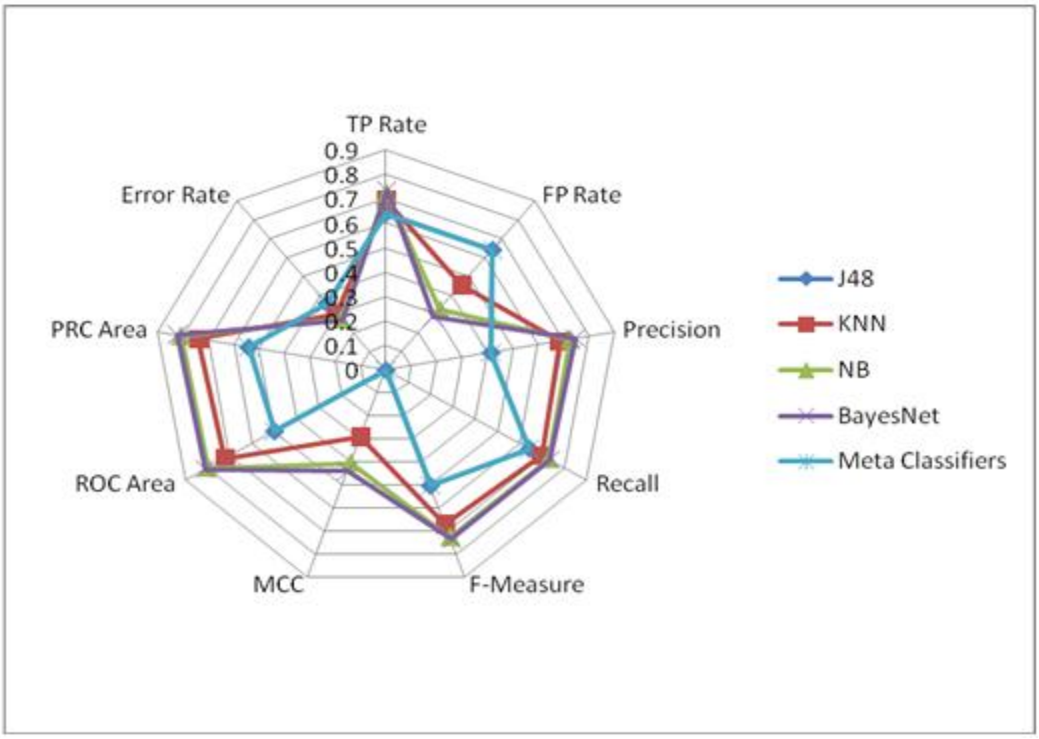


Figure 4.19: Performance Metrics on Balance-Scale

Table 4.11 shows the classifier performance using Ensemble Model of Meta Voting Classifiers combining with KNN, NB and BN classifiers. Voting combining two classifiers named 2 classifiers with vote. Voting combining three classifiers named 3 classifiers with vote.

Table 4.11: Comparison of the Ensemble of Different Classifiers.

Ensemble	Correctly Classified Instances (% Value)	Incorrectly Classified Instances (% Value)	Time Taken to build model (in seconds)	Kappa Statistic
KNN and NB with vote	16939 (72.881%)	6303 (27.119%)	0.04	0.3648
KNN and BN with Vote	**17133 (73.7157%)**	**6109 (26.2843 %)**	**0.03**	**0.4217**
NB and BN with vote	17036 (73.2983 %)	6206 (26.7017 %)	0.08	0.427
3 classifiers with vote	17114 (73.6339 %)	6128 (26.3661%)	0.07	0.4212

Table 4.12 shows the mean absolute errors (MAE) and root mean squared error (RMSE) of the ensemble of different classifiers.

Table 4.12: MAE and RMSE of the Ensemble of Different Classifiers.

Ensemble	**MAE**	**RMSE**
KNN and NB with vote	0.3552	0.415
KNN and BN with vote	0.3474	0.409
NB and BN with vote	0.3295	0.4109
3 classifiers with vote	0.344	0.4078

It was inferred from Table 4.11 and Table 4.12, that the ensemble, 3 classifiers with vote had the least RMES than ensemble 2 classifiers with vote, but will take longer time to build model. It was inferred from Table 4.8 and Table 4.11, that ensemble of ***KNN and BN with Vote*** had the best correctly classified than all individual Base and Meta Classifiers. Table 4.13 shows the classification performance of each Ensemble model in term of recall, precision, f- measure, MCC and Roc Area for Forenoon and Afternoon class.

Table 4.13: The Classification Performance of Each Ensemble Model.

Parameters / Ensemble	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	**Class**
KNN and NB with Vote	0.463	0.124	0.675	0.463	0.549	0.378	0.798	0.690	Forenoon
	0.876	0.537	0.746	0.876	0.806	0.378	0.798	0.873	Afternoon
KNN and BN with Vote	**0.610**	**0.192**	**0.638**	**0.610**	**0.623**	**0.422**	**0.811**	**0.705**	**Forenoon**
	0.808	**0.390**	**0.789**	**0.808**	**0.798**	**0.422**	**0.811**	**0.883**	**Afternoon**
NB and BN with Vote	0.660	0.227	0.618	0.660	0.638	0.428	0.808	0.706	Forenoon
	0.773	0.340	0.804	0.773	0.788	0.428	0.808	0.882	Afternoon
3 classifiers with Vote	0.613	0.195	0.635	0.613	0.624	0.421	0.812	0.707	Forenoon
	0.805	0.387	0.789	0.805	0.797	0.421	0.812	0.885	Afternoon

Table 4.14 shows the overall Ensembles, Base and Meta classifiers performance ranked by accuracy and error rate. It was inferred from Table 4.14 that ensemble ***KNN and BN with Vote*** classifier had the highest accuracy. The Base classifiers J48 and Meta classifiers had the lowest accuracy and greater error rate.

Table 4.14: Overall Ensembles, Base and Meta Classifiers Performance Ranked by: Accuracy and Error Rate.

Models	Accuracy	Error Rate
KNN and BN with vote	73.7157	0.263
3 classifiers with vote	73.6339	0.264
BN	73.4274	0.266
NB and BN with vote	73.2983	0.267
KNN and NB with vote	72.881	0.271
NB	72.6917	0.273
KNN	69.667	0.303
Meta Classifiers	64.3103	0.357
J48	64.3103	0.357

In this work, we evaluated the performance in terms of classification accuracy of J48, KNN, NB, BN, Stacking and Vote meta classifiers using various accuracy measures on log file dataset like TP rate, FP rate, Precision, Recall, F-measure and ROC Area.

- It was observed from results that an error rate of ***KNN and BN with Vote*** classifier was the lowest i.e. 0.263 and it will take shorter time to build model (0.03 seconds) in comparison with the others classifier, which was the most desirable.
- Accuracy of KNN ***and BN with Vote*** classifier was the highest i.e. 73.7157% in comparison with the others classifier, which was highly required. This investigation suggests that, the ***KNN and BN with Vote*** classifier is the optimum ensemble since it gives more classification accuracy for class session in web log file dataset having two values forenoon and afternoon.
- J48 was slightly bad algorithm. Thus we found that J48 was bad algorithm in most of performance measures.

KNN and BN with Vote classifier had the highest accuracy, followed by the ***three*** classifiers together with ***Voting***, followed by ***BN***, followed by ***NB***, followed by ***NB and BN with voting*** , followed by ***KNN and NB with voting*** , followed by ***NB***, followed by ***KNN***, followed by ***Meta Classifiers*** , followed by ***J48***.

4.5 PATTERN ANALYSIS

Users accessed each web page different number of times. Since each web page was not of the same interest. The top of the most frequently visited pages are illustrated in Figure 4.20 below and the graphical representation of the top 7 visited pages are illustrated in Figure 4.21.

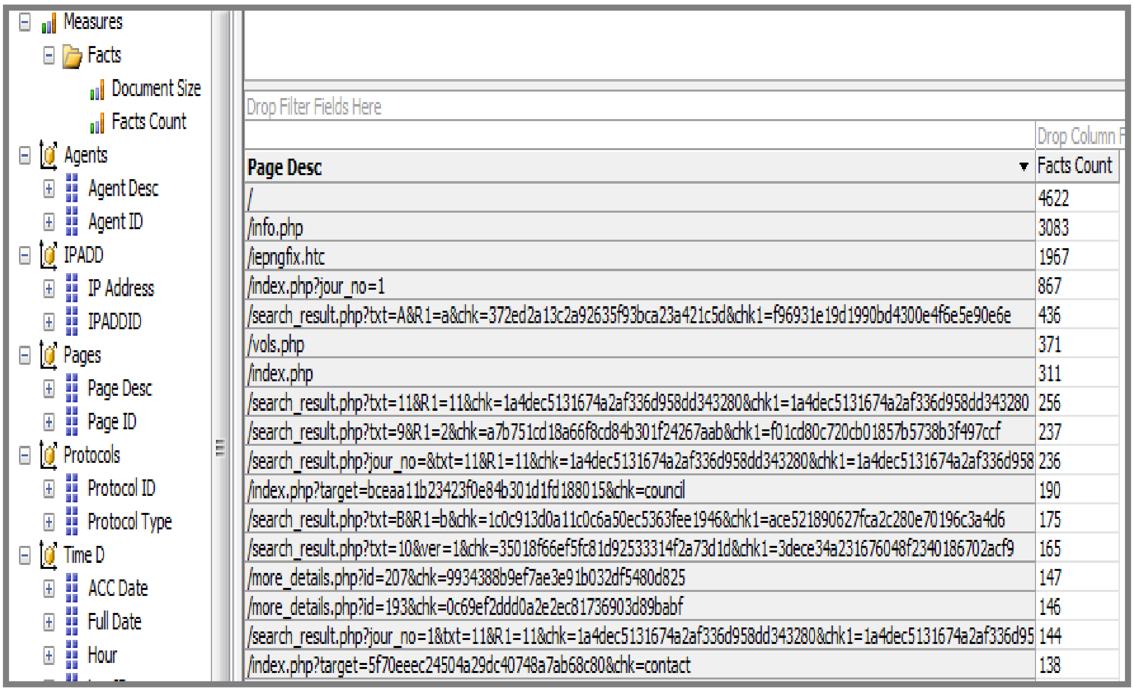
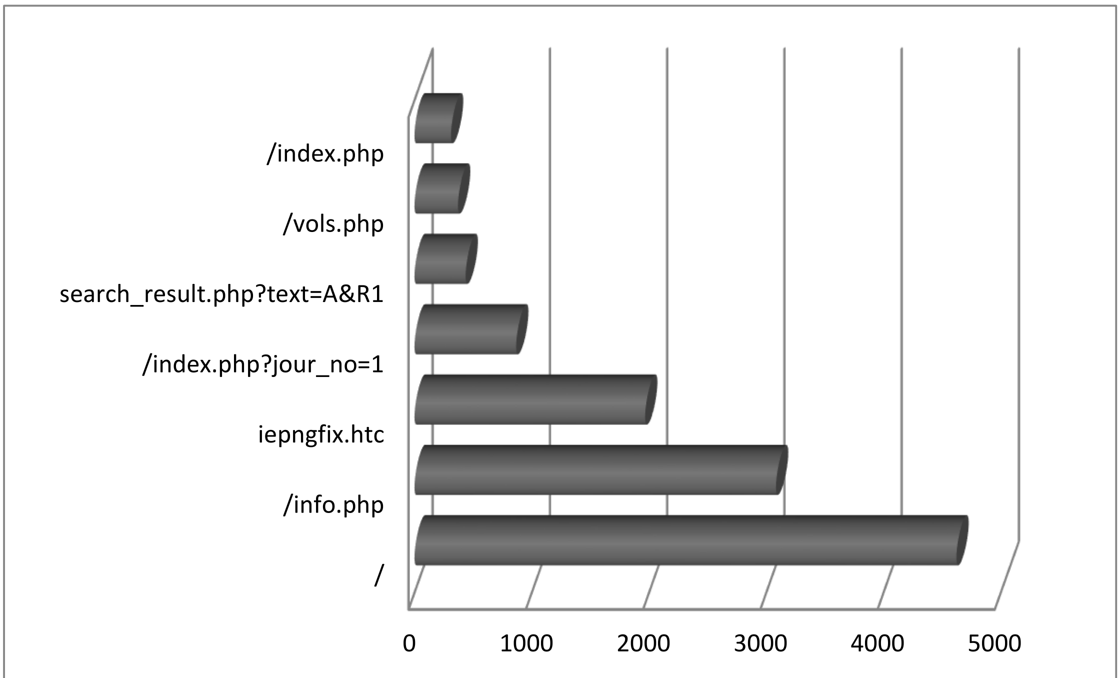


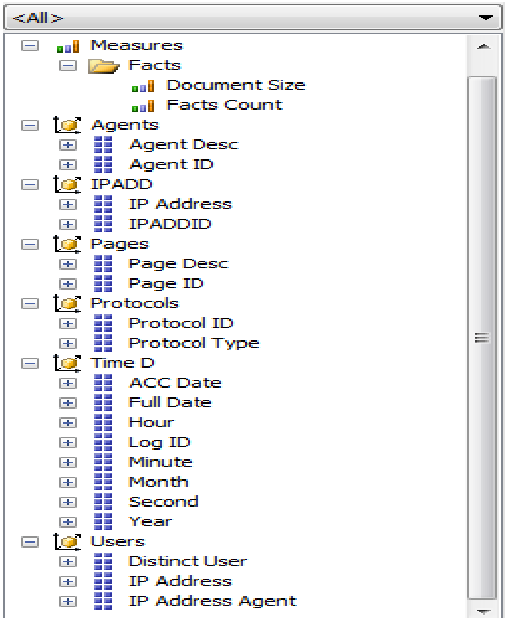
Figure 4.20: The Top of the Most Frequently Visited Pages



	/	/info.php	iepngfix.htc	/index.php?jour_no=1	search_result.php?text=A&R1	/vols.php	/index.php
Visitors	4622	3083	1967	867	436	371	311

Figure 4.21: Graphical Representation of the Top 7 Pages shown in Figure 4

Figure 4.22 below answers the question: Which IP address has accessed the website using which Protocol and how many times?



<Select dimension>

Drop Filter Fields Here

	Protocol Type		
	HTTP/1.0	HTTP/1.1	Grand Total
IP Address	Facts Count	Facts Count	Facts Count
109.82.134.76		2	2
109.82.36.41		2	2
109.82.78.127		8	8
110.37.30.237		4	4
110.37.63.23	4		4
110.8.8.18		1	1
110.8.8.22		1	1
112.104.4.185		2	2
112.200.14.123		1	1
112.200.227.124		1	1
112.202.140.96		1	1
112.206.149.0		1	1
113.254.166.119		1	1
113.254.172.103		1	1
113.254.43.160		1	1
113.254.91.161		1	1
113.92.43.113	1		1
114.164.16.230		1	1
114.189.244.196		1	1
114.198.187.163		1	1
114.48.60.33		1	1
114.58.10.91		2	2
114.58.253.116		2	2
114.59.190.112		1	1
114.72.248.225		1	1

Figure 4.22: IP Address Accessed the Web Site using HTTP Protocol

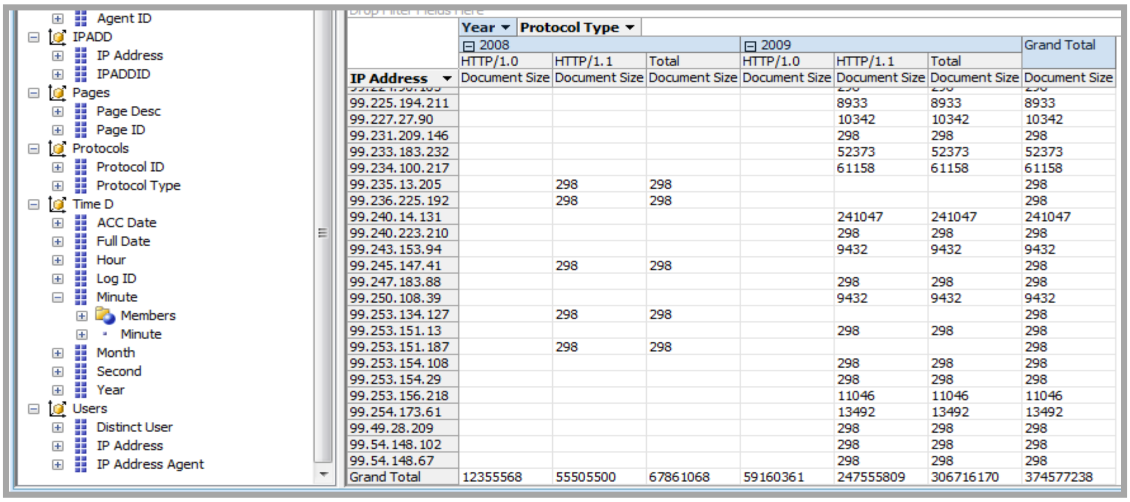
Figure 4.23 shows the number of bytes transferred on month 3 was greater than number of bytes transferred on month 10, although the number of users was equal. Also the number of bytes transferred slightly increased from month 4 through month 5 until month 6, although the number of users in these months decreased rapidly.

Drop Filter Fields Here

Year	Month	Facts Count	Document Size
2008	11	1520	32687965
	12	1780	35173103
	Total	3300	67861068
2009	1	1858	35407984
	10	2109	24277705
	11	1675	19895494
	12	557	6730123
	2	1836	35827284
	3	2109	32370546
	4	1790	25664183
	5	1619	27915992
	6	1491	36847786
	7	1799	21546580
	8	1669	19618279
	9	1712	20614214
	Total	20224	306716170
Grand Total		23524	374577238

Figure 4.23: Number of Bytes Transferred by Users

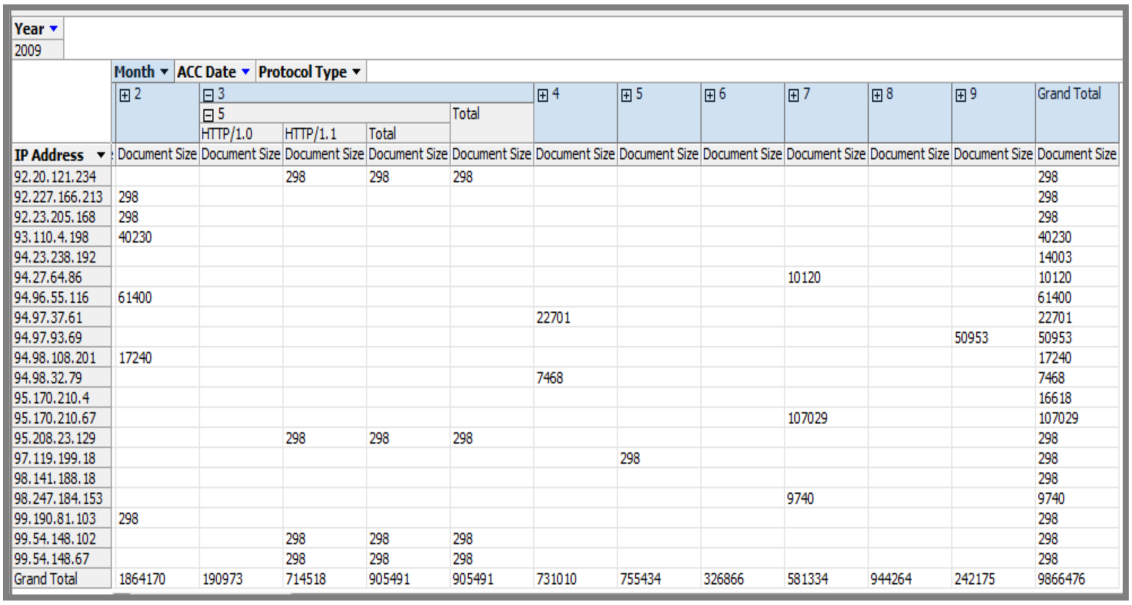
Figure 4.24 shows part of the Web log cube with 3 dimensions: Time, User IP Address and Protocol Type, where time was at level Year. Document size was numeric codes. For example in 2009, IP Address 99.243.153.94 used protocol HTTP/1.1 to download a document with a size 9432 MB.



	Year / Protocol Type: 2008			2009			Grand Total
	HTTP/1.0	HTTP/1.1	Total	HTTP/1.0	HTTP/1.1	Total	
IP Address	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size
99.225.194.211					8933	8933	8933
99.227.27.90					10342	10342	10342
99.231.209.146					298	298	298
99.233.183.232					52373	52373	52373
99.234.100.217					61158	61158	61158
99.235.13.205		298	298				298
99.236.225.192		298	298				298
99.240.14.131					241047	241047	241047
99.240.223.210					298	298	298
99.243.153.94					9432	9432	9432
99.245.147.41		298	298				298
99.247.183.88					298	298	298
99.250.108.39					9432	9432	9432
99.253.134.127		298	298				298
99.253.151.13					298	298	298
99.253.151.187		298	298				298
99.253.154.108					298	298	298
99.253.154.29					298	298	298
99.253.156.218					11046	11046	11046
99.254.173.61					13492	13492	13492
99.49.28.209					298	298	298
99.54.148.102					298	298	298
99.54.148.67					298	298	298
Grand Total	12355568	55505500	67861068	59160361	247555809	306716170	374577238

Figure 4.24: Example Data Cube Created having Time (Year), User IP Address, Protocol Type as Dimensions and Document Size Transferred as a Measure.

Figure 4.25 shows the drill down operations. Here we drilled down the data cube shown in Figure 4.24 into months and access date in the time dimension.



Year: 2009

	Month / ACC Date / Protocol Type: 2	3				4	5	6	7	8	9	Grand Total
		5			Total							
		HTTP/1.0	HTTP/1.1	Total								
IP Address	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size
92.20.121.234			298	298	298							298
92.227.166.213	298											298
92.23.205.168	298											298
93.110.4.198	40230											40230
94.23.238.192												14003
94.27.64.86									10120			10120
94.96.55.116	61400											61400
94.97.37.61						22701						22701
94.97.93.69											50953	50953
94.98.108.201	17240											17240
94.98.32.79						7468						7468
95.170.210.4												16618
95.170.210.67									107029			107029
95.208.23.129			298	298	298							298
97.119.199.18							298					298
98.141.188.18												298
98.247.184.153									9740			9740
99.190.81.103	298											298
99.54.148.102			298	298	298							298
99.54.148.67			298	298	298							298
Grand Total	1864170	190973	714518	905491	905491	731010	755434	326866	581334	944264	242175	9866476

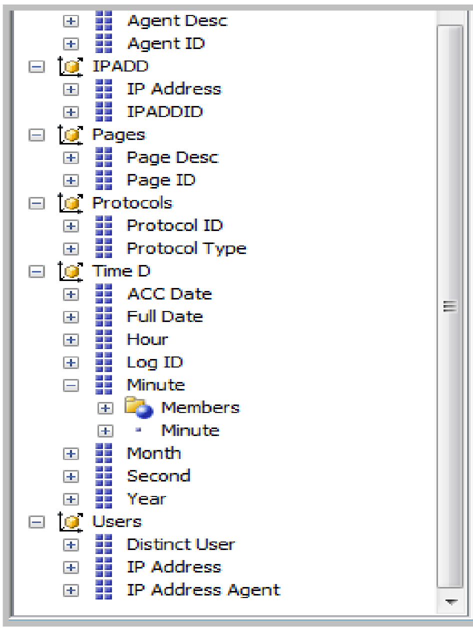
Figure 4.25: Resultant Data Cube after Drill down to Month and Access Date in the Time Dimension in the Data Cube given in Figure 4.24

Figure 4.26 illustrates the roll up operation over the data cube shown in Figure 4.24. The figure allows the user to move to month 2, which was a higher aggregation level through the minutes of the hour 17 on the days 3 and 5.

Year																
2009																
	Month / ACC Date / Hour / Minute / Protocol Type															
	2															Grand Total
	3								5						Total	
	17							Total	17					Total		
	4		22		33		Total		11		44		Total			
	HTTP/1.0	Total	HTTP/1.1	Total	HTTP/1.1	Total			HTTP/1.0	Total	HTTP/1.1	Total				
IP Address	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen
196.1.209.119									39932	39932			39932	39932	39932	39932
66.198.41.20	22948	22948					22948	22948							22948	22948
67.71.40.113											298	298	298	298	298	298
88.153.249.1					298	298	298	298							298	298
91.188.4.110			24906	24906			24906	24906							24906	24906
Grand Total	22948	22948	24906	24906	298	298	48152	48152	39932	39932	298	298	40230	40230	88382	88382

Figure 4.26: Shows the Rollup Operation in the Data Cube shown in Figure 4.24.

Figure 4.27 shows the dicing operation on the cube shown in Figure 4.24. This diced cube contains only two dimensions: Time (Year) and User IP Address.



Drop Filter Fields Here

	Year		
	2008	2009	Grand Total
IP Address	Document Size	Document Size	Document Size
[illegible]	[illegible]		[illegible]
99.224.90.105		296	296
99.225.194.211		8933	8933
99.227.27.90		10342	10342
99.231.209.146		298	298
99.233.183.232		52373	52373
99.234.100.217		61158	61158
99.235.13.205	298		298
99.236.225.192	298		298
99.240.14.131		241047	241047
99.240.223.210		298	298
99.243.153.94		9432	9432
99.245.147.41	298		298
99.247.183.88		298	298
99.250.108.39		9432	9432
99.253.134.127	298		298
99.253.151.13		298	298
99.253.151.187	298		298
99.253.154.108		298	298
99.253.154.29		298	298
99.253.156.218		11046	11046
99.254.173.61		13492	13492
99.49.28.209		298	298
99.54.148.102		298	298
99.54.148.67		298	298
Grand Total	67861068	306716170	374577238

Figure 4.27: Dicing Data Cube shown in Figure 4.24 Contains two Dimensions Time and IP Address.

Using slicing operation (as shown in the Figure 4.28), we were able to focus on the values of the specific cells. In the Figure 4.28 we sliced the data cube for day 1. We can easily see users who accessed the website on day 1/1/2009 of each hour and minute. Note the absence of the user in a few hours like 3, 4 and 5.

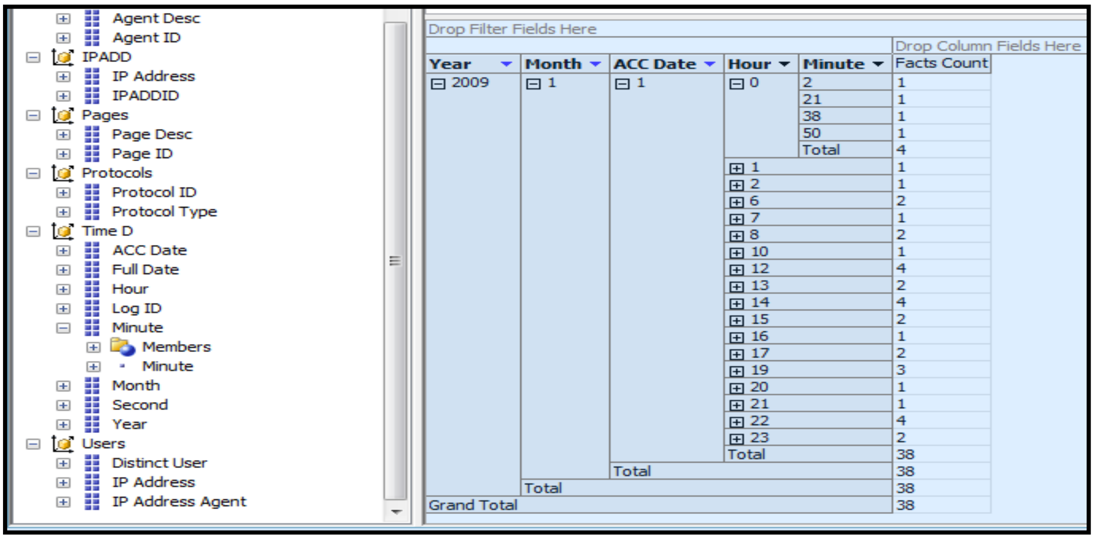


Figure 4.28: Slicing, Data Cube on the Time Dimension for the Day 1/1/2009

4.6 DISCUSSION

Mining web log start with collecting of SUST web log file and process of separating out different data fields from single server log entry is identified as data field extraction. After field extraction, the read logs records will be stored in staging area to facilities data transformation, and mapping. In the transformation, the pre-processing step will be conducted. Using the staging area this transformation can be done easily to a large extent, and more dynamic "monitoring" can be done by the system. Features like alerts and warnings can be easily incorporated. In data mapping, each filed of the staging area will be mapped easily to its equivalent in the data warehouse. An attribute will be mapped to zero or more columns in a relational database (The time attribute in the log file was divided into hours, minutes and seconds; years, months, and days). The attributes of the other dimension tables were taken, as found, from the log file (i.e. the IP Address attribute was used to fill the IP address field in the IP address dimension. The staging area, make it simple for us to divide the sessions in a day into two classes before noon and after noon and store this new divisions in a new attribute or field. In this investigation the log file data entries is classified into forenoon and afternoon using four algorithms namely; J48, KNN, NB and BN, and then combining them in order to decide which of the ensembles, if any, performs better. The new ensemble approach aims to obtain better accuracy. The performance using various accuracy measures like TP rate, FP rate, Precision, Recall, F-measure and ROC Area was evaluated. It was observed from results that an error rate of combining the ***KNN and BN with Vote*** classifier was the lowest and it will take shorter time to build model in comparison with the others classifier, which was the most desirable. Accuracy of combining the ***KNN and BN with Vote*** classifier was the highest in comparison with the others classifier, which was highly required. This investigation shows that, ensemble learning-techniques (***KNN and BN with Vote*** classifier) can increase classification accuracy in the domain of web usage mining, therefore obtains better classification performance than could be obtained from any of the constituent learning algorithms.

4.7 RESULTS SUMMARY

The results obtained after pre-processing contained valuable information about the log files. The results showed (58.13%) a reduction in the number of records in the log file. the data size is reduced to 143 MB that is (74.77 %)by eliminating unnecessary data and hence increase the quality of the available data. From the cleaned data 122122 records were considered and from which, we obtained 23200 unique users of 13869 sessions.

K-mean algorithm and Density-based clustering were used to obtain the clusters, the Performance of two clustering algorithms with and without feature selection are compared according to accuracy factors. From this comparison we can conclude that Density-based clustering clusters with and without feature selection produces fairly higher accuracy, lower RMSE and requires significant computation time than K-means clustering algorithm.

A priori algorithm is used to discover relationship among data. Eliminating redundant rules and clustering decreased the size of the generated rule set to obtain Interestingness rules. Some of the interesting rules are interpret to show the desired relation between items within log file. Analysis results show that using an association rules in WUM can model the rules for managing and optimizing the website structure and advised to be used by users. This helps the web designers to improve website usability by determining related link connections in the website.

J48, KNN, NB and BN algorithms and combination of them are applied on the log file data entries is classified into entries are classified into forenoon and afternoon. In order to evaluate the performance of various accuracy measures was using. The result shows that, combination can increase classification accuracy. An analytical tool for finding relevant information easily and precisely is used in this research, it allows analysts to quickly gain answers to the ‘who’ and ‘what’ questions premised on a usually large set of data. This tool can simplify the analysis of usage statistics of the server access logs. It pre-calculates summary information to enable roll-up or aggregation, drilling, grouping, pivoting, slicing and dicing. The tool allows users to perform ad-hoc analysis of both the web log warehouse and the mining results. A large number of analysis queries were given to the tool and it produce correct results.

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

In this dissertation, the theoretical and experimental studies have shown that the proposed model is effective and applicable for web usage mining. As deduction of this research, a conclusion and future work have been presented.

5.1 CONCLUSION

The log file contains a huge amount of information that needs to be organized, cleaned and analyzed. There for, in order to achieve that a new approach has been introduced to mine and analyze the web log file through different phases as follows:

- The cleaning process phase was achieved by removing irrelevant data like image access, failed entries. Many interesting patterns are available in the raw web log file. However, it is very complicated to extract the interesting patterns without pre-processing. The results obtained after pre-processing were satisfactory and contained valuable information about the log files, has shown that (58.13%) a reduction in the number of records and in the log file size and hence increases the quality of the available data.
- In the pattern discovery phase, the clustering technique and association rule mining were implemented. Two clustering techniques were used in this work, namely: K-means clustering and Density-based clustering. The clustering solved the problem of categorizing data by partitioning a data set into a number of clusters based on some similarity measure so that the similarity in each cluster was larger than among clusters. Clustering algorithms applied with and without feature selection for SUST log file dataset using WEKA tools. Performance of the clustering method was measured by the percentage of the incorrectly classified instances. Density-based clustering gave better performance compared to k-means clustering without feature selection. Density-based Clustering recognized characters with higher accuracy and minimum amount of time compared to k-means algorithm. Clustering algorithms was applied with feature selection. It was concluded that choice of

a good feature can contribute a lot to clustering techniques. Also with feature selection the performance of Density-based clustering was better than K-Means algorithm.

- Implementation of a system for pattern discovery using association rules was discussed as a method for Web Usage Mining. Pruning our rule set according to redundant rules and Clustering decreased the size of the rule set from 50 to only 38 rules. We identified 8 truly interesting rules out of the 38 rules in the rule set (21%). Analysis results show that using an association rules in WUM can model the rules for managing and optimizing the website structure and advised to be used by users. This helps the web designers to improve website usability by determining related link connections in the website.
- The classification is one of the most pattern discovery techniques used to extract knowledge from pre-processed data. There are different methods used to classify users' session. One of these is to classify them into “forenoon” and “afternoon”. J48, KNN, NB and BN algorithms have been used and evaluated. The ensemble of KNN and BN with vote Meta classifier introduced higher classification accuracy for SUST web log file dataset having two values “forenoon” and “afternoon.
- In the last phase of this research, the pre-processed data was uploaded in a data warehouse in form of dimension tables and fact table. The organization of the log file was achieved by grouping the data into unique users, unique IP address, protocol, pages, and agent. Cleaned and organized data were presented in the form of a cube, the basic structure that can be used by the Online Analytical Process (OLAP). The results achieved have proved that the data warehouse can be implemented successfully to analyze the log files to make appropriate decisions.
- Finally the result of this research can help SUST in analyzing log files for the generating ad-hoc queries for derives indicators about when, how, and by whom a web server is visited. Also the important reports are usually generated on demand as easy as possible.
- In addition, this research can be useful in other area such ass e-commerce. E-commerce is any type of business or commercial transaction that involves the transfer of information across the internet. In this situation a huge amount

of information is generated and stored in the web services. This information overhead leads to difficulty in finding relevant and useful knowledge, therefore this research may help to discover and extract pattern from the web to mine customer behaviour. Customer behaviour pattern is analyzed to improve e-commerce websites. Also to cluster customer segments by using clustering algorithms in which input data comes from web log of various e-commerce websites. Hence, discover the relationship between different web pages within a web site. When you perform deeper analysis on the clickstream data and examine user behaviour on the site, patterns are bound to emerge. Besides, the security issues are the most precious problems in every electronic commercial process, therefore by loading the data into the transactional database, Features like alerts and warnings can be easily incorporated in this architecture.

5.2 FUTURE WORK

Results from our research have uncovered a number of additional areas that warrant further study. As future work, there is a need to solve problems related to parallel processing, especially for huge amount of data that resulted from the growing usage of the web. This growing usage is due to the large volumes of data stored in servers, which resulted in an increasing amount of data and thus growing in the size of log file. Also due to the complexity of the dataset and the difficulty in understanding them, a visualization tools are needed to render the information related to these complex dataset in an easy and understandable way.

Large log files that are generated from the web servers need an efficient way to analyze and handle them. This efficient way needs a development of algorithms that work in a parallel, because sequential algorithms suit computers which are basically performs operations in a sequential fashion. Although there is an improvement in the speed of the sequential machine, this improvement is coming at a greater cost. As a consequence there is a need to work and improve parallel algorithm in a cost effective way. Such algorithms can handle and analyze large log files in parallel machine in an effective and efficient way.

The large log files normally contains a huge data, this resulted in a significant challenge to understanding the dataset. This challenge can be addressed by rendering the data in a way that the user can see it in visual form. Traditional two-dimensional presentation cannot work effectively with the current volumes of data. A scientific investigation should be carried out to develop new tools that can display and visualize the data in an understandable and dynamic ways. Visualizing complex data can help researchers or practitioners explore patterns and trends within the data. In Web log files such tool can provide graphical reports that show hits for web pages, user's activity, in which part of website users are interested, traffic source, etc.

REFERENCES

[1] Chauhan, Abhishek ; Tarar, Sandhaya ;, "Prediction of User Browsing Behavior Using Web Log Data," *IJSRSET,* vol. 2, no. 1, pp. 419-422, January-February 2016.

[2] Singhal, Vidhu ; Pandey, Gopal ;, "A Web Based Recommendation Using Association Rule and Clustering," *International Journal of Computer & Communication Engineering Research (IJCCER),* vol. 1, no. 1, pp. 1-5, 2013.

[3] Shirgave, Suresh ; Kulkarni, Prakash, "Semanticlly Enriched Web Usage Mining For Predicting User Future Movements," *International Journal of Web & Semantic Technology* , vol. 4, no. 4, pp. 59-72, October 2013.

[4] Gupta, Ravindra ; Gupta, Prateek ;, "Application specific web log pre-processing," *Int.J.Computer Techology & Applications,* vol. 3, no. 1, pp. 160-162, Jan-Feb 2012.

[5] Sheware, Seema ; Nikose, A. A.;, "A Review on Clustering Techniques used in Web Usage," *International Journal of Modern Trends in Engineering and Research,* vol. 2, no. 2, pp. 475-479, February 2015.

[6] Kumar C. U., OM ; Bhargavi, P.;, "Analysis of Web Server Log by Web Usage Mining for Extracting," *International Journal of Computer Science Engineering,* vol. 3, no. 2, pp. 123-136, Jun 2013.

[7] Kaur, Jaswinder ; Garg, Dr. Kanwal ;, "Analyzing the Different Attributes of Web Log Files To Have An Effective Web Mining," *International Journal of Advanced Scientific and Technical Research,* vol. 3, no. 5, pp. 127-134, May-June 2015.

[8] Katkar, Dr. Girish S.; Kasliwal, Amit Dipchandji ;, "Use of Log Data for Predictive Analytics through Data Mining," *Current Trends in Technology and Science,* vol. 3, no. 3, pp. 217-222, May 2014.

[9] Davamani, Dr. Antony Selvdoss ; , V.Chitraa;, "A Survey on Preprocessing Methods for Web Usage Data," *(IJCSIS) International Journal of Computer*

Science and Info, vol. 7, no. 3, pp. 78-83, 2010.

[10] Abd Wahab, Mohd Helmy ; Haji Mohd, Mohd Norzali ; Hanafi, Hafizul Fahri ; Mohamad Mohsin, Mohamad Farhan;, "Data Pre-processing on Web Server Logs for," *Proceedings of World Academy of Science, Engineering and Technology,* vol. 36, pp. 970-977, December 2008.

[11] Vishwakarma, Amit ; Singh, Kedar Nath ;, "A Survey on Web Log Mining Pattern Discovery," *(IJCSIT) International Journal of Computer Science and Information Technologies,* vol. 5, no. 6, pp. 7022-7031, 2014.

[12] G. Kaur, "Accurate Analysis of Weblog Server File by Using Clustering," *nternational Journal of Advanced Research in Computer Engineering & Technology,* vol. 2, no. 7, pp. 2341-2343, July 2013.

[13] Nithya, P.; Sumathi, P.;, "An Effective Web Usage Analysis using Fuzzy Clustering," *ARPN Journal of Science and Technology,* vol. 3, pp. 693-698, July 2013.

[14] Sahu, Shashi ; Sahu, Leena ;, "A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* vol. 4, no. 3, pp. 825-829, March 2015.

[15] Upadhyay, Mr. Akshay ; Purswani, Mr. Balram;, "Web Usage Mining has Pattern Discovery," *International Journal of Scientific and Research Publications,* vol. 3, no. 2, pp. 1-4, February 2013.

[16] Jafari, Maryam ; Sabzchi, Farzad Soleymani; Irani, Amir Jalili;, "Applying Web Usage Mining Techeniques to Design Effective Web Recommendation Systems:A Case Study," *A dvances in Computer Science: an International Journal,* vol. 3, no. 2, pp. 78-90, March 2014.

[17] , linHuaXu; , HongLiu;, "Web User Clustering Analysis based on KMeans Algorithm," *International Conference on Information, Networking and Automation (ICINA),* vol. 2, pp. 6-9, 2010 .

[18] Mangai, J. Alamelu ; D. Kothari, Dipti ; Kumar, V. Santhosh ;, "A Novel Approach for Automatic Web Page Classification using Feature Intervals," *International Journal of Computer Science,* vol. 9, no. 5, pp. 282-287,

September 2012.

[19] D. M. Tank, "Improved Apriori Algorithm for Association Rules," *I.J. Information Technology and Computer Science* , vol. 7, no. 3, pp. 15-23, 2014.

[20] V. Vidyapriya and S. Kalaivani, "An Efficient Clustering Technique for Weblogs," *International Journal of Innovative Science, Engineering & Technology,* vol. 2, no. 7, pp. 518-524, July 2015.

[21] Kaur, Chintandeep ; Aggarwal, Rinkle Rani ;, "Web Mining Tasks and Types: A Survey," *International Journal of Research in IT & Management,* vol. 2, no. 2, pp. 547-558, February 2012.

[22] Rathi, Ankit ; Raipurkar, Abhijeet ;, "Web Usage Mining - A Review," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 5, no. 2, pp. 496-498, February 2016.

[23] B. Patel, Ketul ; A. Chauhan, Jignesh ; D. Patel, Jigar ;, "Web Mining in E-Commerce: Pattern Discovery,Issues and Applications," *International Journal of P2P Network Trends and Technology,* vol. 11, no. 3, pp. 40-45, 2011.

[24] D. Satokar, Kavita ; Gawali, Prof..S.Z.;, "Web Search Result Personalization using Web Mining," *International Journal of Computer Applications,* vol. 2, no. 5, pp. 29-32, June 2010.

[25] Dohare, Mahendra Pratap; Arya, Premnarayan ; Bajpai, Aruna ;, "Novel Web Usage Mining for Web Mining Techniques," *International Journal of Emerging Technology and Advanced Engineering,* vol. 2, no. 1, pp. 253-262, January 2012.

[26] K. Pani, S.; Panigrahy, L.; Ratha, Bikram Keshari ;, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs," *International Journal of Instrumentation, Control & Automation (IJICA),* vol. 1, no. 1, pp. 15-23, 2011.

[27] Mansur, M. Iqbal ; Kavitha, C. ; Thangadurai, K. ;, "Web Prediction Method on Social Network Analysis," *Journal of Engineering and Applied Sciences,* vol. 10, no. 7, pp. 2855-2860, April 2015.

[28] Gupta, Mr. Ravindra ; Gupta, Prateek ;, "Fast Processing of Web Usage Mining with Customized Web Log Pre-processing and modified Frequent Pattern Tree," *International Journal of Computer Science & Communication Networks,* vol. 1,

no. 3, pp. 277-279, 2011.

[29] Singh, Arun ; Pathak, Avinav ; Sharma, Dheeraj ;, "Web Usage Mining : Discovery Of Mined Data Patterns and their Applications," *International Journal of Computer Science and Management Research,* vol. 2, no. 5, pp. 2423 -2429, May 2013.

[30] S. Kamat, Mona; Bakal, Dr. J. W.; Nashipudi, Madhu ;, "Improved Data Preparation Technique in Web Usage Mining," *International Journal of Computer Networks and Communications Security,* vol. 1, no. 7, p. 284–291, December 2013.

[31] Bhawsar, Sawan ; Pathak, Kshitij ; Mariya, Sourabh ; Parihar, Sunil ;, "Extraction of Business Rules from Web logs to Improve Web Usage Mining," *International Journal of Emerging Technology and Advanced Engineering,* vol. 2, no. 8, pp. 333-340, August 2012.

[32] G. T. Wei, K. Shirly , W. Husain and Z. Zainol, "A Study of Customer Behaviour Through Web Mining," *Journal of Information Sciences and Computing Technologies,* vol. 2, no. 1, pp. 103-107, February 2015.

[33] Nithya, P.; Sumathi, Dr. P.;, "A Survey on Web Usage Mining: Theory and Applications," *Int.J.Computer Technology & Applications,* vol. 3, no. 4, pp. 1625-1629, July-August 2012.

[34] Mekala, T. ; Nandhini, P. ;, "Modified Agglomerative Clustering for Web Users Navigation Behavior," *Int. J. Advanced Networking and Applications,* vol. 5, no. 1, pp. 1842-1846, 2013.

[35] Sendre, Rupesh ;, "A Survey on Research trends & approaches for structuring Web server log files data," *International Journal of Computer Trends an,* vol. 24, no. 1, pp. 11-16, June 2015.

[36] Meghwal, Arjun Ram ; Sharma, Arvind K ;, "Identifying System Errors through Web Server Log Files in Web Log Mining," *IJCST,* vol. 7, no. 1, pp. 57-61, Jan 2016.

[37] Chandel, Gajendra Singh ; Patidar, Kailash ; Mali, Man Singh ;, "Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm," *International Journal of Computer Science and Network Security,*

vol. 16, no. 1, pp. 135-140, January 2016.

[38] Anand, Neetu; Hilal, Saba;, "Identifying the User Access Pattern in Web Log Data," *(IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 3, no. 2, pp. 3536-3539, 2012.

[39] G.Langhnoja, Shaily ; Barot, Mehul P.; Mehta, Darshak B. ;, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery," *International Journal of Data Mining Techniques and Applications,* vol. 2, no. 1, pp. 141-150, June 2013.

[40] Suneetha, K. R.; Krishnamoorthi, R. ;, "Identifying User Behavior by Analyzing Web Server Access Log File," *International Journal of Computer Science and Network Security,* vol. 9, no. 4, pp. 327-332, April 2009.

[41] Kundu , Shakti ;, "Cluster Diagnostics and Verification Tool for Effective and Scalable Web Log Analysis," *International Journal of Computer Science and Information Technologies,* vol. 3, no. 3, pp. 4097-4100, 2012.

[42] Adhikari, Siddharth ; Saraf, Devesh ; Revanwar, Mahesh ; Ankam, Nikhil ;, "Analysis of Log Data and Statistics Report Generation Using Hadoop," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 2, no. 4, pp. 4054 - 4058, April 2014.

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

[44] Lokeshkumar, R. ; Sindhuja, R. ; Sengottuvelan, P. ;, "A Survey on Preprocessing of Web Log File in Web Usage Mining to Improve the Quality of Data," *International Journal of Emerging Technology and Advanced Engineering,* vol. 4, no. 8, pp. 229-234, August 2014.

[45] Chavan, AlgorithmMukul B.; Patil, Sarita ;, "Analysis of Web Log from Database System utilizing E-web Miner Algorithm," *International Journal on Recent and Innovation Trends in Computing and Communication,* vol. 3, no. 7, pp. 4730 - 4734, July 2015.

[46] Patel, Ketul B.; Patel, Dr. A. R. ;, "Process of Web Usage Mining to find

Interesting Patterns," *International Journal of Computers & Technology,* vol. 3, no. 1, pp. 144-148, Aug 2012.

[47] K, Savitha ; MS, Vijaya ;, "Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies," *International Journal of Advanced Computer Science and Applications,* vol. 5, no. 1, pp. 137-142, 2014.

[48] kaur, Harmit ; singh, Hardeep ;, "A Survey of Preprocessing Method for Web Usage Mining Process," *International Journal of Computer Trends and Technology,* vol. 9, no. 2, pp. 62-66, Mar 2014.

[49] Sriram, 1Ranjena ; Mallika, R. ;, "Innovative Pre-Processing Technique and Efficient Unique User Identification Algorithm for Web Usage Mining," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 6, no. 2, pp. 85-91, February 2016.

[50] Anand, Surbhi ; Aggarwal, Rinkle Rani ;, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions," *International Journal of Computer Applications,* vol. 48, no. 2, pp. 13-18, June 2012.

[51] Deepa, A.; Raajan, P. ;, "An Efficient Preprocessing Methodology of Log File for Web Usage Mining," *National Conference on Research Issues in Image Analysis and Mining Intelligence,* vol. 2, no. 2, pp. 13-16, 2015.

[52] Sagar, Payal ; Nimavat, A. V.;, "Web Usage Mining: Survey on Process and Methods," *International Multidisciplinary Research Journal,* vol. 2, no. 5, pp. 1-4, May 2015.

[53] Elhebir, Mohammed ; Abraham, Ajith ;, "Data Pre-Processing of Web Server Logs for Mining users Access Patterns," *International Journal of Engineering Sciences Paradigms and Researches,* vol. 3, no. 1, pp. 23-31, August 2015.

[54] Jarkad, Megha P.; Bhonsle, Mansi;, "Improved Web Prediction Algorithm Using Web Log Data," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 3, no. 5, pp. 4902-4907, May 2015.

[55] Aldekhail, M. ;, "Application and Significance of Web Usage Mining in the 21st Century: A Literature Review," *International Journal of Computer Theory and Engineering,* vol. 8, no. 1, pp. 41-47, February 2016.

[56] Sundari, M. Rekha ; Srinivas, Y.; Reddy, Prasad;, "A Review on Pattern Discovery Techniques of Web Usage Mining," *Int. Journal of Engineering Research and Applications,* vol. 4, no. 9, pp. 131-136, September 2014.

[57] Devipriyaa, K. ; Kalpana, B.;, "Users' Navigation Pattern Discovery using Ant Based Clustering and LCS," *Journal of Global Research in Computer Science,* vol. 1, no. 1, pp. 1-5, August 2010.

[58] P. Suthar and B. Oza, "A Survey of Web Usage Mining Techniques," *International Journal of Computer Science and Information Technologies,* vol. 6, no. 6, pp. 5073-5076, 2015.

[59] Garg, Tamanna ; Dhawan, Sanjeev ;, "Web Usage Mining in Online Social Network," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 5, no. 3, pp. 818- 828, March 2015.

[60] Mathur, Abhishek ; Agrawal, Trapti ;, "A Survey: Access Patterns Mining Techniques and ACO," *International Journal of Engineering and Advanced Technology,* vol. 2, no. 5, p. 2249 – 8958, June 2013.

[61] Ivancsy, Renata; Kovacs, Ferenc;, "Clustering Techniques Utilized in Web Usage Mining," in *Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, Madrid, 2006.

[62] Padmaja, S.; Sheshasaayee, Ananthi ;, "Clustering of User Behaviour based on Web Log data using Improved K-Means Clustering Algorithm," *International Journal of Engineering and Technology,* vol. 8, no. 1, pp. 305-310, Feb-Marh 2016.

[63] Das, S. ; Abraham, a. ; Konar, a. ;, "Automatic Clustering Using an Improved Differential Evolution Algorithm," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 38, no. 1, pp. 218-237, jan 2008.

[64] Izakian, Hesam ; Abraham, Ajith ;, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Systems with Applications,* vol. 38, no. 3, pp. 1835-1838, March 2011.

[65] R. R. Patil and A. Khan, "Bisecting K-Means Web Log data," *International Journal of Computer Applications,* vol. 116, no. 19, pp. 36-41, April 2015.

[66] Ansari, Zahid ; Ahmed, Waseem ; Azeem, M. F.; Babu, A. Vinaya ;, "Discovery of Web Usage Profiles Using Various Clustering Techniques," *International Journal of Computer Information Systems,* vol. 1, no. 3, pp. 18-69, 2011.

[67] Katariya, Karuna ; Aluvalu, Rajanikanth;, "Agglomerative Clustering in Web Usage Mining: A Survey," *International Journal of ComputerApplications* , vol. 89, no. 8, pp. 24-27, March 2014.

[68] Chaudhari, Bharat; Parikh, Manan ;, "A Comparative Study of Clustering Algorithms Using Weka Tools," *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* vol. 1, no. 2, pp. 154-158, October 2012.

[69] Selvakuberan, K. ; Indradevi, M. ; Rajaram, Dr. R. ;, "Combined Feature Selection and Classification – A novel Approach for the Categorization of Web Pages," *Journal of Information and Computing Science,* vol. 3, no. 2, pp. 083-089, April 2008.

[70] Elhebir, Mohammed Hamed ; Abraham, Ajith ;, "Discovering Web Server Logs Patterns Using Clustering and Association Rules Mining," *Journal of Network and Innovative Computing,* vol. 3, no. 1, pp. 159-167, 2015.

[71] Serasiya, Shilpa Dhanjibhai ; Chaudhary , Neeraj ;, "Simulation of Various Classifications Results using WEKA," *International Journal of Recent Technology and Engi,* vol. 1, no. 3, pp. 155-162, August 2012.

[72] Rahul Jadhav, Kiruthika M; Dixit, Dipa ; J, Rashmi ; Nehete, Anjali ; Khodkar, Trupti ;, "Pattern Discovery Using Association Rules," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 2, no. 12, pp. 69-74, 2011.

[73] Veeramalai, S.; Jaisankar, N.; Kannan, A.;, "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy," *International journal of computer science & information Technology,* vol. 2, no. 4, pp. 60-74, August 2010.

[74] Singhal, Vidhu; Pandey, Gopal ;, "A Web Based Recommendation Using Association Rule and Clustering," *International Journal of Computer & Communication Engineering Research (IJCCER),* vol. 1, no. 1, pp. 1-5, May

2013.

[75] Elhebir, Mohammed Hamed ; Abraham, Ajith ;, "A Novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification," *International Journal of Computer Information Systems and Industrial Management Applications,* vol. 7, no. 1, pp. 189-195, 2015.

[76] Jameela, A.; Revathy, P.;, "Comparison of Decision and Random Tree Algorithms on A Web Log Data for Finding Frequent Patterns," *International Journal of Research in Engineering and Technology,* vol. 3, no. 7, pp. 155-161, May 2014.

[77] Ramdas, Shruthi ; P, Rithesh Pakkala ; R, Akhila Thejaswi ;, "Determination and Classification of Interesting Visitors of Websites using Web Logs," *International Journal of Computer Science and Mobile Computing,* vol. 5, no. 1, pp. 1-9, January 2016.

[78] Phyu, Thair Nu ;, "Survey of Classification Techniques in Data Mining," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2009.

[79] Prasanth, Anupama ;, "Web Personalization using Web Usage Mining Techniques," *International Journal of Current Engineering and Scientific Research,* vol. 3, no. 3, pp. 45- 49, 2016.

[80] Baoli, Li ; Shiwen, Yu ; Qin, Lu ;, "An Improved k-Nearest Neighbor Algorithm for Text Categorization," in *International Conference on Computer Processing of Oriental Languages*, Shenyang, 1-7.

[81] D. K. Tiwary, "A Comparative Study of Classification Algorithms for Credit Card Approval using Weka," *International Interdisciplinary Research Journal,* vol. 2, no. 3, pp. 165-174, March 2014.

[82] Sarangi, Samir Kumar ; Jaglan , Dr. Vivek ;, "Performance Comparison of Machine Learning Algorithms on Integration of Clustering and Classification Techniques," *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS),* vol. 4, no. 3, pp. 251-257, March 2013.

[83] Vaithiyanatha, V.; Rajeswari, K. ; Tajane, Kapil ; Pitale, Rahul ;, "Comparison of Different Classification Techniques using Different Datasets," *International*

Journal of Advances in Engineering & Technology, vol. 6, no. 2, pp. 764-768, May 2013.

[84] Roos , Teemu ; Wettig, Hannes ; Grunwald, Peter ; Myllym, Petri ; Tirri, Henry, "On Discriminative Bayesian Network Classifiers and Logistic Regression," *Machine Learning,* vol. 59, no. 3, p. 267–296, 2005.

[85] Sigletos , Georgios ; Hatzopoulos , Michalis ;, "Combining Information Extraction Systems Using Voting and Stacked Generalization," *Journal of Machine Learning Research,* vol. 6, pp. 1751-1782, 2005.

[86] Rani, M. Usha ; Kumari , G.T. Prasanna ;, "A Study of Meta-Learning in Ensemble Based Classifier," *Engineering Science and Technology: An International Journal ,* vol. 2, no. 1, pp. 36-41, 2012.

[87] Reddy, G.Satyanarayana ; Srinivasu, Rallabandi ; Rao, M. Poorna Chander ; Rikkula, Srikanth Reddy, "Data Mining, Olap and Oltp Technologies are Essential Elements to Support Decision-Making Process in Industries," *International Journal on Computer Science and Engineering,* vol. 2, no. 9, pp. 2865-2873, 2010.

[88] R. Pandya, "Web Usage Mining with Personalization on Social Web," *International Journal of Engineering Trends and Technology,* vol. 29, no. 6, pp. 325-328, November 2015.

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

[90] Reddy, G. Satyanarayana ; et. al.;, "Data Warehousing, Data Mining, OLAP and OLTP Technologies are Essential Elements to Support Decision-Making Process in Industries," *International Journal on Computer Science and Engineering,* vol. 2, no. 9, pp. 2865-2873, 2010.

[91] Elhebir, Mohammed Hamed; Elfaki, Murtada Khalafallah; Abraham, Ajith ;, "Web Log Data Analysis Using a Data Warehouse and OLAP," *Journal of Network and Innovative Computing,* vol. 2, pp. 359-365, 2014.

[92] Grace, L.K. Joshila ; Maheswari, V.Maheswari; Nagamalai, Dhinaharan ;, "Analysis of Web Logs and Web User in Web Mining," *International Journal of*

Network Security & Its Applications, vol. 3, no. 1, pp. 99-110, January 2011.

[93] Jain, Ratnesh Kumar ; Jain, Dr. Suresh ; Kasana, Dr. R. S.;, "On Line Analytical Mining of Web Usage Data Warehouse," *International Journal of Comput er Science & Emerging Technologies (,* vol. 1, no. 1, pp. 15-24, June 2010.

[94] Bhan, Namita ; Mehrotra, Deepti ;, "Compartive Study of EM and K-MEANS Clustring techniques in WEKA Interface," *International Journal of Advanced Technology & Engineering Research,* vol. 3, no. 4, pp. 40-44, July 2013.

[95] B. Jagtap, Dr. Sudhir; B. G., Dr. Kodge ;, "Census Data Mining and Data Analysis using WEKA," *International Conference in "Emerging Trends in Science, Technology and Management*, pp. 35-40, 2013.

[96] A. Abraham, "Business Intelligence from Web Usage Mining," *ournal of Information & Knowledge Management,* vol. 2, no. 4, pp. 375-390, 2003.

[97] P. B. Patel, "An IVR Call Performance Classification System using Computational Intelligent Techniques," 2009.

[98] Etminani, Kobra ; Akbarzadeh-T., Mohammad-R.; Yanehsari, Noorali Raeeji ;, "Web Usage Mining: users' navigational patterns extraction from web logs," *IFSA-EUSFLAT,* pp. 396-401, 2009.

[99] Joshi, Karuna P ; Joshi, Anupam ; Yesha, Yelena ; Krishnapuram, Raghu ;, "Warehousing and Mining Web Logs," *International Journal of Computer Science & Emerging Technologies,* vol. 1, no. 1, pp. 15-24, 2010.

[100] K, Yogish H ; Raju, Dr. G T ; N, Manjunath T ;, "The Descriptive Study of Knowledge Discovery from Web Usage Mining," *IJCSI International Journal of Computer Science Issues,* vol. 8, no. 5, pp. 225-230, September 2011.

[101] Sharma, Anshuman, "Web Usage Mining Using Neural Network," *International Journal of Reviews in Computing,* vol. 9, no. 1, pp. 72-78, April 2012.

[102] Tani, Fauzia Yasmeen ; Farid, Dewan Md. ; Rahman, Mohammad Zahidur;, "Ensemble of Decision Tree Classifiers," *International Journal of Applied Information Systems (IJAIS),* vol. 1, no. 2, pp. 30-36, January 2012.

[103] Goel, Neha ; Jha, C.K;, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool," *International Journal of Computer Applications,* vol. 62, no. 2, pp. 29-33, January 2013.

[104] Dhillon , Supreet ; Kaur, Kamaljit ;, "Comparative Study of Classification Algorithms for Web Usage Mining," *nternational Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 7, pp. 137-140, July 2014.

[105] Chopra, Kamal Nain;, "Modeling and Technical Analysis of Electronics Commerce and Predictive Analytics," *Journal of Internet Banking and Commerce*, vol. 19, no. 2, pp. 1-10, August 2014.

APPENDIX

FILLING THE DIMENSION AND FACT TABLE SAMPLE CODE:

```
Dim sqlcon As New SqlClient.SqlConnection
Dim Excelcon As New
System.Data.OleDb.OleDbConnection("Provider=Microsoft.Jet.OLEDB.4.0;" & "data
source=D:\FldCompLast.xls; Extended Properties=Excel 8.0;")
        sqlcon.ConnectionString = "server=SUNCOM-PC; Database=WebLog;
Trusted_connection=True;"
        sqlcon.Open()
        MsgBox("connectuon succeded")
        Excelcon.Open()
        MsgBox("connectuon succeded")


        Dim cmdexcel As New OleDbCommand
        cmdexcel = New OleDbCommand("select * from [FldCompLast$]", Excelcon)
        Dim drexcel As OleDbDataReader
        drexcel = cmdexcel.ExecuteReader
        Dim sqlcmd As New SqlCommand
        '=================================================
        Dim strind As Integer
        Dim agentstring As String
        Do While drexcel.Read()
            If String.IsNullOrEmpty(drexcel.GetValue(9).ToString) = False Then
                agentstring = drexcel.GetValue(9).ToString
                strind = agentstring.IndexOf("Windows NT")
            Else
                agentstring = "-"
            End If
            'MsgBox(agentstring)
            'MsgBox(strind)

Dim sqlcon As New SqlClient.SqlConnection
        sqlcon.ConnectionString = "server=SUNCOM-PC; Database=WebLog;
Trusted_connection=True; MultipleActiveResultSets=True"
        sqlcon.Open()
        Dim cmdsqltemp As New SqlCommand("SELECT DISTINCT Agentname FROM
AgentsTemp", sqlcon)
        Dim sqlRtemp As SqlDataReader
        sqlRtemp = cmdsqltemp.ExecuteReader
        Dim sqlcmd As New SqlCommand
        Dim rcount As Integer
        rcount = 1
        Do While sqlRtemp.Read
            sqlcmd.CommandText = "INSERT INTO Agents VALUES(" & rcount & ",'" &
sqlRtemp.GetValue(0) & "')"
            sqlcmd.Connection = sqlcon
            sqlcmd.ExecuteNonQuery()
            rcount = rcount + 1
        Loop
        MsgBox("connection succeded")

        Dim cmdexcel As New OleDbCommand
        cmdexcel = New OleDbCommand("select * from [FldCompLast$]", Excelcon)

        Dim drexcel As OleDbDataReader
        drexcel = cmdexcel.ExecuteReader
```

```
        Dim sqlcmd As New SqlCommand
        '==================================================
        Dim agentstring As String
        Do While drexcel.Read()
            If String.IsNullOrEmpty(drexcel.GetValue(4).ToString) = False Then
                agentstring = drexcel.GetValue(4).ToString
            Else
                agentstring = "-"
            End If
            'MsgBox(agentstring)
            'MsgBox(strind)

            sqlcmd.CommandText = "INSERT INTO PageTemp VALUES('" & agentstring
& "')"
            sqlcmd.Connection = sqlcon
            sqlcmd.ExecuteNonQuery()
        Loop
        MsgBox("Seccessful")

Dim cmdsqltemp As New SqlCommand("SELECT DISTINCT Page FROM PageTemp", sqlcon)
        Dim sqlRtemp As SqlDataReader
        sqlRtemp = cmdsqltemp.ExecuteReader
        Dim sqlcmd As New SqlCommand
        Dim rcount As Integer
        rcount = 1
        Do While sqlRtemp.Read

            sqlcmd.CommandText = "INSERT INTO Pages VALUES(" & rcount & ",'" &
sqlRtemp.GetValue(0) & "')"
            sqlcmd.Connection = sqlcon
            sqlcmd.ExecuteNonQuery()
            rcount = rcount + 1
        Loop
        MsgBox("succeeded")

Dim cmdexcel As New OleDbCommand
        cmdexcel = New OleDbCommand("select * from [FldCompLast1$]", Excelcon)

        Dim drexcel As OleDbDataReader
        drexcel = cmdexcel.ExecuteReader

        'Dim sqlcmd As New SqlCommand
        '==================================================

        Dim Datetstring As String
        Dim date1 As Date
        Dim date2 As String
        Dim Time1 As String

        Dim dcount As Integer = 0
        Do While drexcel.Read()
            MsgBox(drexcel.GetValue(2).ToString())

            Datetstring = drexcel.GetValue(2).ToString().Remove(0, 1)

            MsgBox(Datetstring)

            Datetstring = Datetstring.Remove(Datetstring.Length - 1, 1)
            MsgBox(Datetstring)

            Datetstring = Datetstring.Substring(0, Datetstring.IndexOf("+"))
```

```
            MsgBox(Datetstring)
            date1 = Datetstring.Substring(0, Datetstring.IndexOf(":"))
            date2 = date1
            MsgBox("Date is:    " & date1)
            Dim Dates() As String = date2.Split("/")
            MsgBox(Dates(0) & "   mm   " & Int(Dates(1)) & "   mm   " & Dates(2))
            Time1    =    Datetstring.Substring(Datetstring.IndexOf(":")    +    1,
Datetstring.Length - (Datetstring.IndexOf(":") + 1))
            MsgBox("Time is:     " & Time1)
            Dim Times() As String = Time1.Split(":")
            MsgBox(Times(0) & "   mm   " & Int(Times(1)) & "   mm   " & Times(2))
            dcount = dcount + 1
        Loop
        MsgBox("Seccessful")

Dim sqlcon As New SqlClient.SqlConnection
        Dim Excelcon As New
System.Data.OleDb.OleDbConnection("Provider=Microsoft.Jet.OLEDB.4.0;" & "data
source=D:\FldCompLast.xlsx; Extended Properties=Excel 8.0;")
        sqlcon.ConnectionString    =    "server=SUNCOM-PC;    Database=WebLog;
Trusted_connection=True;"
        sqlcon.Open()
        MsgBox("connectuon succeeded")
        Excelcon.Open()
        MsgBox("connectuon succeeded")

        Dim cmdexcel As New OleDbCommand
        cmdexcel = New OleDbCommand("select * from [FldCompLast$]", Excelcon)
        Dim drexcel As OleDbDataReader
        drexcel = cmdexcel.ExecuteReader
        Dim sqlcmd As New SqlCommand
        Dim i As Integer = 0
        Do While drexcel.Read()
            i = i + 1
            sqlcmd.CommandText  =  "INSERT  INTO  IPADD  VALUES("  &  i  &  ",'"  &
drexcel.GetValue(1) & "')"
            sqlcmd.Connection = sqlcon
            sqlcmd.ExecuteNonQuery()
Dim cmdexcel As New OleDbCommand
        cmdexcel = New OleDbCommand("select * from [FldCompLast$]", Excelcon)
        Dim drexcel As OleDbDataReader
        drexcel = cmdexcel.ExecuteReader
        Dim sqlcmd As New SqlCommand
        Dim agentstring, pagestring As String
        Dim Datetstring As String
        Dim date1 As Date
        Dim date2 As String
        Dim Time1 As String
        Dim bytesize As String
        Do While drexcel.Read()
            If String.IsNullOrEmpty(drexcel.GetValue(9).ToString) = False Then
                agentstring = drexcel.GetValue(9).ToString
            Else
                agentstring = "-"
            End If
            If String.IsNullOrEmpty(drexcel.GetValue(4).ToString) = False Then
                pagestring = drexcel.GetValue(4).ToString
            Else
                pagestring = "-"
            End If
```

```
            bytesize = drexcel.GetValue(7)
           ' MsgBox(bytesize)

            '===============================================================
            Datetstring = drexcel.GetValue(2).ToString().Remove(0, 1)
            Datetstring = Datetstring.Remove(Datetstring.Length - 1, 1)
            Datetstring = Datetstring.Substring(0, Datetstring.IndexOf("+"))
            date1 = Datetstring.Substring(0, Datetstring.IndexOf(":"))
            date2 = date1
            Dim Dates() As String = date2.Split("/")
            Time1   =   Datetstring.Substring(Datetstring.IndexOf(":")   +   1,
Datetstring.Length - (Datetstring.IndexOf(":") + 1))
            Dim Times() As String = Time1.Split(":")
            '===============================================================

            sqlcmd.CommandText = "INSERT INTO
Temp(AgentDesc,IPAddressofUsers,IPAddressAgent,PageDesc,Second,Minute,Hour,
Month, ACC_Date,Year,ProtocolType,Bytesize,FullDate) VALUES('" & agentstring &
"','" & drexcel.GetValue(1) & "','" & drexcel.GetValue(0) & "','" & pagestring
& "'," & _
                                   Times(2) & "," & Times(1) & "," & Times(0) &
","   &   Dates(0)   &   ","   &   Dates(1)   &   ","   &   Dates(2)   &   ",'"   &
drexcel.GetValue(5).ToString & "'," & bytesize & ",'" & Datetstring & "')"
            sqlcmd.Connection = sqlcon
            sqlcmd.ExecuteNonQuery()
        Loop
        MsgBox("Success")
```