Approach for Developing Business Statistics Using Data Web Usage Mining

1G V S CH S L V Prasad, 2Malapati Sri Rama Lakshmi Reddy, 3Kuntam Babu Rao & 4Chodagam Suresh Kumar

Abstract – Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems process of knowledge discovery in databases also known as KDD. Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Predicting of user's browsing behavior is an important technology of E-commerce application. The prediction results can be used for personalization, building proper web site, improving marketing strategy, promotion, customer likes, product supply, getting marketing information, forecasting market trends, changes according to interests, and increasing the competitive strength of enterprises etc. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining is usually an automated process whereby Web servers collect and report user access patterns in server access logs. The navigation datasets which are sequential in nature. Clustering web data is finding the groups which share common interests and behavior by analyzing the data collected in the web servers, this improves clustering on web data efficiently using proposed robust algorithm. In the proposed work a new technique to enhance the learning capabilities and reduce the computation intensity of a competitive learning multi-layered neural network using the K-means clustering algorithm. The proposed model use multi-layered network architecture with a back propagation learning mechanism to discover and analyze useful knowledge from the available Web log data.

Keywords – Data mining, Web usage mining, E-commerce, Web Data, Clustering.

I. INTRODUCTION

A. Data Mining

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information about the characteristics of the collected data, using a number of techniques (e.g., association rule mining). Forecasting and predictive modeling provide predictions of future events, and the processes may range from the transparent (e.g., rule-based approaches) through to the opaque (e.g., neural networks).Metadata, (data about the characteristics of a data set), are often expressed in a condensed data-minable format, or one that facilitates the practice of data mining. Common examples include executive summaries and scientific abstracts.

A primary reason for using data mining is to assist in the analysis of collections of observations of behaviour. Such data are vulnerable to collinearity because of unknown interrelations. An unavoidable fact of data mining is that the (sub-) set(s) of data being analysed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviours that exist across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other mining approaches, such as Choice Modelling for human-generated data. In these situations, inherent correlations can be either controlled for, or removed
altogether, during the construction of the experimental design.

A. Motivation

Most classification algorithms perform batch classification, that is, they work on the dataset directly in memory. More complex methods often build large data structures which give them a larger memory footprint. This larger footprint prevents them from being applied to many of the larger datasets. Even when these datasets can be used, the complexity of the algorithm can make the classification task take an inordinately long time. There are solutions to these problems in the form of adding more memory or waiting longer for experiments to finish. However both have a cost in money and time and both may not ultimately solve the problem if the dataset cannot in memory when the memory is at a maximum. A feature of most large datasets is the large number of data points that contain similar information. So one solution is to remove redundant information from the dataset to allow a classifier to concentrate on data points that represent a larger group of points. This allows the construction of a classifier that correctly models the relationship expressed by the data in the dataset. Random sampling is a simple way to remove redundancy and can be very effective on uniformly distributed data.

B. Advantages of Data Mining

Data mining software allows users to analyze large databases to solve business decision problems. Data mining helps analysis recognize significant facts, relationships, trends, patterns exceptions and anomalies that might otherwise go unnoticed. For business, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Specific uses of data mining include:

- Market segmentation – Identify the common characteristics of customers who buy the same products from your company.
- Fraud detection – Identify which transactions are most likely to be fraudulent.
- Direct marketing – Identify which prospects should be included in a mailing list to obtain the highest response rate.
- Market basket analysis – Understand what products or services are commonly purchased together; e.g., beer and diapers.
- Market sale details – Identify those products which are below sales and discount those products to clear room.

C. Tasks Solved by Data Mining

Predicting

A task of learning a pattern from examples and using the developed model to predict future values of the target variable.

Classification

A task of finding a function that maps records into one of several discrete classes.

Detection of relations

A task of searching for the most influential independent variables for a selected target variable.

Explicit modeling

A task of finding explicit formulae describing dependencies between various variables.

Clustering

A task of identifying groups of records those are similar between themselves but different from the rest of the data. Often, the variables the best clustering should be identified as well.

Market Basket Analysis

Processing transactional data in order to find those groups of products that are sold together well. One also searches for directed association rules identifying the best product to be offered with a current selection of purchases products.

Market Sale Details

A task of identifying those products which are out of sale.

D. Deviation Detection

A task of determining the most significant changes in some key measures of data from previous or expected values.

The most commonly used techniques in data mining are:

- Artificial neural networks - Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision Trees - Tree-shaped structure that represent sets of decisions. These decisions generate rules for the classification of a datasets. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- Genetic Algorithms - Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design bases on the concepts of evolution.
- Nearest neighbor method - A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k*1). Sometimes called the k-nearest neighbor technique.
- Rule induction - The extraction of useful if-then rules from data based on statistical significance.

Many of these techniques have been in use for more
than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

E. Definition

With the rapid development of Internet, web-sites and web applications are getting more and more important in people’s daily life. Web log mining is the technology to find out the hidden user patterns by analyzing the information that include in web logs. So that capability, structure and individual functions can be optimized and provided by using this statistics. There are three ways to record user visit information: web server, web proxy server and user client. In the web server logs, one web-site’s visit information of multi users is recorded. In the web proxy server logs, multi web-sites’ visit information of multi users is recorded. In the client logs, multi website’s visit information of one user is recorded. During above three logs, the first and the second ones are recorded automatically, and the third one should be recorded by appropriate software. Generally speaking, web server log has the highest structured form. We pay special attention to the web server logs, through the research and analysis of web server logs; the pages which are outlying, but related actually are discovered in order to optimized web-sites structure in a more reasonable way. And these mining statistics are finally used to establish fast and direct information channels. World Wide Web is a global village and rich source of information. Day by day number of web sites and its users are increasing rapidly. While surfing the web sites, users’ interactions with web sites are recorded in web log file. There are three main sources to get the raw web log file such as 1.) Client Log File .2) Proxy Log File and .3) Server Log File as in Fig. 1 Architecture.

![Fig. 1 : Architecture](image)

Usage of these sources has its own pros and cons but their importance to collect the data for WUM is invaluable. The true user behavior can be portrayed from client log file. Client log files are most authentic and accurate to depict the user behavior but it is difficult task to modify the browser for each client and requires users’ essence and collaboration as well. It is in the form of one-to-many relationships of client and web sites visited by that particular user. Proxy log file is also used to capture user access data. Proxy server log files are most complex and more vulnerable to user access data in log file. To unleash the true picture of user behavior is difficult. Same IP address is used by many users but on the other hand we can have unique user login. Proxy server is in many-to-many relationships. One user can access many sites and many users can visit one site. To capture the real user and users’ browsing behavior is difficult. Server log files are in relationship of many to one. Many users may visit one web site and user behavior about that particular web site can be captured accordingly. In this survey, we observed that most of the researcher considered web server log file as most reliable and accurate for WUM process.

Web mining involves a wide range of applications which aim to discover and extract knowledge from data. Web mining also provides an important mechanism to make data access more efficient and adequate to users. In web usage (log files) mining, for instance, the main task is to discover the activities (preferences) of the users while they are browsing and navigating through the web. The goal of understanding the preferences of the users is to enhance the quality of the service and to personalize this service. In this area of data mining, several techniques can be applied such as classification and clustering. The main aim of clustering algorithms, for instance, is to use attribute information to group instances (records) that have similar attribute values.

Web usage mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs.

World Wide Web (WWW) is expanding tremendously everyday in the number of websites and also the population of users. The basic purpose of website is to deliver useful information to its users efficiently and timely; at the same time websites are competing to acquire their own shares of visitors. Websites are striving to improve themselves by offering personalized contents and services that supposedly will match best of the users’ tastes or needs. In addition to explicit methods like questionnaires and options of setting users’ profiles, one subtle approach is to probe into web log files for revealing patterns of users’ interests on the websites. It is well known that users’ online interactions with the website are recorded in server web log files that serve as a valuable pool of information. By applying the data mining techniques on web log file, we obtain good insights about the users’ behaviors; thereby we can customize the contents and services on the website to better suit the users. We can analyze the web log files for various aspects of website enhancements. Furthermore, proper analysis of web log unleashes useful information for webmaster or
administrator for numerous advantages such as web personalization, website schema modification, user surfing behaviors, website structure modification and we can tackle the issue of web server performance as well.

Web mining involves a wide range of applications which aim to discover and extract knowledge from data. Web mining also provides an important mechanism to make data access more efficient and adequate to users. In web usage (log files) mining, for instance, the main task is to discover the activities (preferences) of the users while they are browsing and navigating through the web. The goal of understanding the preferences of the users is to enhance the quality of the service and to personalize this service. In this area of data mining, several techniques can be applied such as classification and clustering. The main aim of clustering algorithms, for instance, is to use attribute information to group instances (records) that have similar attribute values.

Web usage mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs.

Web mining can be categorized into three areas of interest based on which part of the web to mine (Web mining research lines):

1. Web content mining – discovery of useful information from the web contents/data/documents (or) is the application of data mining techniques to content published on the Internet. The web contains many kinds and types of data. Basically, the web content consists of several types of data such as plain text (unstructured), image, audio, video, meta data as well as HTML (semi-structured), or XML (structured documents), dynamic documents, multimedia documents. Recent research on mining multi types of data is termed multimedia data mining. Thus we could consider multimedia data mining as an instance of web content mining. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts/ text data mining/ text mining. Hence we could consider text mining as an instance as an instance of web content mining. Research issues addressed in text mining are: topic discovery, extracting association patterns, clustering of web documents and classification of web pages.

Issues in Web content Mining:

i) developing intelligent tools for information retrieval
ii) finding keywords and key phases
iii) discovering grammatical rules collections
iv) hypertext classification/categorization
v) extracting key phrases from text documents
vi) learning extraction rules
vii) predicting relationships

Web content mining approaches: Agent based and Data base approaches Agent based approaches: Involves AI systems that can "act autonomously or semi autonomously on behalf of a particular user, to discover and organize web based information". Agent Based approaches focus on intelligent and autonomous web mining tools based on agent technology. i. Some intelligent web agents can use a user profile to search for relevant information, then organize and interpret the discovered information. Example: Harvest. ii) Some use various information retrieval techniques and the characteristics of open hypertext documents to organize and filter retrieved information. Example: Hypursuit. iii) Learn user preferences and use those preferences to discover information sources for those particular user. Example: Xpert Rule Rminer.

Data base approach: focuses on "integrating and organizing the heterogeneous and semi-structured data on the web into more structured and high level collections of resources". These organized resources can then be accessed and analyzed. These “metadata, or generalization are then organized into structured collections and can be analyzed.

2. Web Structure Mining: operates on the web’s hyperlink structure. This graph structure can provide information about page ranking or authoritativeness and enhance search results through filtering i.e., tries to discover the model underlying the link structures of the web. This model is used to analyze the similarity and relationship between different web sites. Uses the hyperlink structure of the web as an additional information source. This type of mining can be further divided into 2 kinds based on the kind of structural data used. a) Hyperlinks: A hyperlink is a structural unit that connects a web page to different location, either within the same web page (intra_document hyperlink) or to a different web page (inter_document hyperlink). b) Document structure: In addition, the content within a web page can also be organized in a tree structured format, based on various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

Web link analysis used for:

i) ordering documents matching a user query (ranking)
ii) deciding what pages to add to a collection page categorization
iii) finding related pages
vi) finding duplicated web sites and also to find out similarity between them

3. Web Usage Mining: Web usage mining is the application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based
Applications. It tries to make sense of the data generated by the web surfer’s sessions/behaviors. While the web content and structure mining utilize the primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web. The web usage data includes the data from web server logs, proxy server logs, browser logs, and user profiles. (The usage data can also be split into 3 different kinds on the basis of the source of its collection: on the server side (there is an aggregate picture of the usage of a service by all users), the client side (while on the client side there is complete picture of usage of all services by a particular client), and the proxy side (with the proxy side being somewhere in the middle). Registration data, user sessions, cookies, user queries, mouse clicks, and any other data as the results of interactions. Web usage mining analyzes results of user interactions with a web server, including web logs, click streams, and database transactions at a web site of a group of related sites.

Web usage mining also known as web log mining.

Web usage mining process can be regarded as a three-phase process consisting:

1. Preprocessing/ data preparation - web log data are preprocessed in order to clean the data – removes log entries that are not needed for the mining process, data integration, identify users, sessions, and so on
2. Pattern discovery - statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, cluster and classification rules) are applied in order to detect interesting patterns.
3. pattern analysis phase - discovered patterns are analyzed here using OLAP tools, knowledge query management mechanism and Intelligent agent to filter out the uninteresting rules/patterns.

After discovering patterns from usage data, a further analysis has to be conducted. The most common ways of analyzing such patterns are either by using query or by loading the results into a data cube and then performing OLAP operations. Then, visualization techniques are used for a results interpretation. The discovered rules and patterns can then be used for improving the system performance / for making modifications to the web site. The purpose of web usage mining is to apply statistical and data mining techniques to the preprocessed web log data, in order to discover useful patterns. Usage mining tools discover and predict user behavior in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.

II. OBJECTIVES

1. Extract statistical information and discover interesting user patterns.
2. Cluster the user into groups according to their navigational behavior.
3. Discover potential correlations between web pages and user groups.
4. Identification of potential customers for ecommerce.
5. Enhance the quality and delivery of Internet information services to the end user.
6. Improve web server system performance and site design.
7. Facilitate personalization.

A requirement is a feature that the system must have or a constraint that it must satisfy to be accepted by client. Requirements engineering aims at defining the requirements for the system under construction. It includes two main activities: Requirements Elicitation and Analysis.

Requirements elicitation is about communication among developers, clients, and users for defining a new system. It focuses on describing the purpose of the system. Such a definition is called system specification. Requirement elicitation is the more challenging of the two because it requires the collaboration of several groups of participants with different backgrounds. On the one hand, the client and the users are experts in their domain and have a general idea of what the system should do, but they often have little experience in software development. On the other hand, the developers have experience in building systems, but often have little knowledge of everyday environment of the users.

Scope

- Log file should be in excel or any other database format.
- Data size should not be less than 10 instances.
- Cleaning algorithms are applied before users clustering.
- Each log file should be applied to field extraction algorithm for relevant fields.

Purpose

- This system uses preprocessing methods to retrieve “relevant information” in the web log for web usage mining (WUM) procedure.
- This system easily classifies server log data with client log data.
- Proposed system implements noise removing procedures.
- Proposed system handles Error log file like error 404 file not found, 505 server errors etc.
- Proposed system uses improved kmeans clustering algorithm with multilayered feed forward
neural network algorithm for user’s session identification.

- This system implements improved kmeans with feed forward approach in order to get effective results.

III. ANALYSIS

From the user interface, web server log files are taken. These files are saved in the form of .txt. This log file contains many fields. Unwanted fields can be removed from the log file by using some of the field extraction algorithms. After removing the fields we will get another file. By using that file we can identify the users no of clicks, no of visits to the website and users IP address by applying some of the user identification algorithms. And we can find the daily, monthly and yearly visitors to a web site by representing them in graphs, pie charts. From that log file we can also find the Session of any user by using session identification method.

Web server log file is converted into the csv dataset format. Then it is sent to the data preprocessing techniques. Data preprocessing is used to convert the raw data into processed data and is used to clean the data. On the filtered log file we apply data mining algorithms. From those algorithms we will get results.

Sample log data format:

```
"7,5511791,0,0,new,,0,,none,,1,action,""http://forum.piwik.org/read.php?2,75279"",Error: Timezone ID is invalid,478318,18392811,Sat 24 Mar 23:52:43,,Forum status,Anonymous,0,USD,$,2012-03-24,00:52:43,1332633163,1332633163,2012-03-24 23:52:43,14,14s,1,0,0,0,Italy,plugins/UserCountry/flags/it.png,Europe,IP,http://piwik.org/faq/general/#faq_52,Direct Entry,http://forum.piwik.org/list.php?5,page=2","",Windows XP,Win XP,plugins/UserSettings/images/os/WXP.gif,gecko,Gecko (Firefox),Firefox 11.0,plugins/UserSettings/images/browsers/FF.gif,wide,1680x1050,plugins/UserSettings/images/screens/wide.gif,""pdf, flash, java, quicktime, windowsmedia"",plugins/UserSettings/images/plugins/pdf.gif,plugins/UserSettings/images/plugins/flash.gif,plugins/UserSettings/images/plugins/java.gif,java,plugins/UserSettings/images/plugins/quicktime.gif,quicktime,1332633155,23:52:35,Sat 24 Mar,Sat 24 Mar,23:52:22,action,""http://forum.piwik.org/read.php?5,84797","",Piwik zählt keine Besuche!,690096,18392806,Sat 24 Mar 23:52:35,plugins/UserSettings/images/windows media.gif,windowsmedia,
```

The following are commonly used attributes in all the log files:

a) Date

The date from Greenwich Mean Time (GMT x 100) is recorded for each hit. The date format is YYYY-MM-DD. The example from Fig. 1 above shows that the transaction was recorded on 2003-11-23.

b) Time

Time of transactions. The time format is HH:MM:SS. The example from Fig. 1 above shows that the transaction time was recorded at 16:00:13.

c) Client IP Address

Client IP is the number of computer who access or request the site.

d) User Authentication

Some web sites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a Website, that user’s “username” is logged in the fourth field of the log file.

e) Server Name

Name of the server. In Fig. 1 the name of the server is CSLNTSVR20.

f) Server IP Address

Server IP is a static IP provided by Internet Service Provider. This IP will be a reference for access the information from the server.

g) Server Port

Server Port is a port used for data transmission. Usually, the port used is port 80.

h) Server Method (HTTP Request)

The word request refers to an image, movie, sound, pdf, txt, HTML file and more. The above example in Fig. 1 indicates that folder.gif was the item accessed. It
is also important to note that the full path name from the document root. The GET in front of the path name specifies the way in which the server sends the requested information. Currently, there are three formats that Web servers send information [8] in GET, POST, and Head. Most HTML files are served via GET Method while most CGI functionality is served via POST.

i) URI-Stem

URI-Stem is path from the host. It represents the structure of the websites. For example:-

```
/tutor/images/icons/fold.gif
```

j) Server URI-Query

URI-Query usually appears after sign “?”. This represents the type of user request and the value usually appears in the Address Bar. For example:-

```
?q=tawaran+biasiswa&hl=en&lr=&ie=UTF-8&oe=UTF-8&start=20&sa=N
```

k) Status

This is the status code returned by the server; by definition this will be the three digit number [2]. There are four classes of codes:

i. Success (200 Series)

ii. Redirect (300 Series)

iii. Failure (400 Series)

iv. Server Error (500 Series)

A status code of 200 means the transaction was successful. Common 300-series codes are 302, for redirect from http://www.mydomain.com to http://www.mydomain.com, and 304 for a conditional GET. This occurs when server checks if the version of the file or graphics already in cache is still the current version and directs the browser to use the cached version. The most common failure codes are 401 (failed authentication), 403 (Forbidden request to a restrict subdirectory, and the dreaded 404 (file not found) messages. In the above transmission a status is 200 means that there was a successful transmission.

a) Bytes Sent

The amount of data revisited by the server, not together the header line.

b) Bytes Received

Amount of data sent by client to the server.

c) Time Stamp

This attribute is used to determine how long a visitor spent on a given page.

d) Protocol Version

HTTP protocol being used (e.g. HTTP/1.1).

e) Host

This is either the IP address or the corresponding host name (www.tutor.com.my) of the remote user requesting the page.

A. DATASET FILE FORMATS

In this project two types of file formats are used. They are CSV, ARFF

CSV: It stands for Comma Separated Value. This format is obtained using MS-Excel. KDD99 dataset is loaded into Excel and then it is saved with an extension of csv.

ARFF: It stands for Attribute Relation File Format. An ARFF is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

ARFF files have two distinct sections. The first section is the Header information, which is followed the Data in The ARFF Header Section. The ARFF Header section of the file contains the relation declaration and attribute declarations.

The @relation Declaration

The relation name is defined as the first line in the ARFF file. The format is:

```
@relation <relation-name>
```

Where <relation-name> is a string. The string must be quoted if the name includes spaces.

The @attribute Declarations:

Attribute declarations take the form of an orderd sequence of @attribute statements. Each attribute in the data set has its own @attribute statement which uniquely defines the name of that attribute and it's data type. The order the attributes are declared indicates the column position in the data section of the file. For example, if an attribute is the third one declared then Weka expects that all that attributes values will be found in the third comma delimited column.

The format for the @attribute statement is:

```
@attribute <attribute-name> <datatype>
```

Where the <attribute-name> must start with an alphabetic character. If spaces are to be included in the name then the entire name must be quoted.

The <datatype> can be any of the four types currently (version 3.2.1) supported by Weka:

- numeric
- <nominal-specification>
- string
- date [<date-format>]

Where <nominal-specification> and <date-format> are defined below. The keywords numeric, string and date are case insensitive.
Nominal attributes
Nominal values are defined by providing an <nominal-specification> listing the possible values: \{<nominal-name1>, <nominal-name2>, <nominal-name3>, ...\}

String attributes
String attributes allow us to create attributes containing arbitrary textual values. String attributes are declared as follows:

```text
@ATTRIBUTE name string
```

Date attributes
Date attribute declarations take the form:

```text
attribute <name> date [<date-format>]
```

Where <name> is the name for the attribute and <date-format> is an optional string specifying how date values should be parsed and printed (this is the same format used by SimpleDateFormat). The default format string accepts the ISO-8601 combined date and time format: "yyyy-MM-dd'T'HH:mm:ss".

ARFF Data Section
The ARFF Data section of the file contains the data declaration line and the actual instance lines.

The @data Declaration
The @data declaration is a single line denoting the start of the data segment in the file. The format is:

```text
@data
```

The Instance data Each instance is represented on a single line, with carriage returns denoting the end of the instance. Attribute values for each instance are delimited by commas. They must appear in the order that they were declared in the header section (i.e. the data corresponding to the nth @attribute declaration is always the nth field of the attribute).

Example weblog.arff file:

```text
@relation piwikk
@attribute Date {2012-02-24,2012-02-25,2012-02-26,2012-02-27,2012-02-28,2012-02-29,2012-03-01,2012-03-02,2012-03-03,2012-03-04,2012-03-05,2012-03-06,2012-03-07,2012-03-08,2012-03-09,2012-03-10,2012-03-11,2012-03-12,2012-03-13,2012-03-14,2012-03-15,2012-03-16,2012-03-17,2012-03-18,2012-03-19,2012-03-20,2012-03-21,2012-03-22,2012-03-23,2012-03-24}
@attribute 'Maximum actions in one visit' numeric
@attribute Actions numeric
@attribute 'Unique visitors' numeric
@attribute Visits numeric
@attribute 'Bounce Rate' numeric
@attribute 'Actions per Visit' numeric
@attribute 'Avg. Visit Duration (in seconds)' numeric
@attribute 'Actions by Returning Visits' numeric
@attribute 'Unique returning visitors' numeric
@attribute 'Returning Visits' numeric
@attribute 'Bounce Rate for Returning Visits' numeric
@attribute 'Avg. Actions per Returning Visit' numeric
@attribute 'Avg. Duration of a Returning Visit (in sec)' numeric
@attribute Conversions numeric
@attribute 'Visits with Conversions' numeric
@attribute 'Conversion Rate' numeric
@attribute Revenue numeric
@attribute Outlinks numeric
@attribute Pageviews numeric
@attribute 'Unique Outlinks' numeric
@attribute 'Unique Pageviews' numeric
@attribute Downloads numeric
@attribute 'Unique Downloads' numeric
@data
2012-02-24,82,3227,954,1140,63,2,8,216,1512,248,393,48,3,8,320,53,39,3,42,81,151,3051,141,2383,0,0
2012-02-25,48,1531,485,561,64,2,7,171,628,130,178,54,3,5,249,47,32,5,7,85,66,1463,63,1151,0,0
2012-02-26,63,2259,625,737,60,3,1,206,1190,176,253,41,4,7,350,55,35,4,7,91,73,2192,69,1691,0,0
2012-02-27,72,4201,1097,1298,60,3,2,234,1913,290,429,42,4,5,393,124,85,6,55,208,157,4050,143,3018,0,0
2012-02-28,27,4201,1097,1298,60,3,2,234,1913,290,429,42,4,5,393,124,85,6,55,208,157,4050,143,3018,0,0
2012-02-29,85,3556,1040,1203,63,3,2,12,1381,268,367,48,3,8,33,4,97,69,5,74,173,131,3370,121,2573,0,0
2012-03-01,85,3556,1040,1203,63,3,2,12,1381,268,367,48,3,8,33,4,97,69,5,74,173,131,3370,121,2573,0,0
2012-03-02,47,2728,887,1022,64,2,7,182,1247,251,344,50,3,6,30,1,75,48,4,7,135,115,2621,112,2050,0,0
2012-03-
IV. IMPLEMENTATION

WEBLOG Pre-processing techniques:

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process.

This system only describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process.

So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

IV. IMPLEMENTATION

WEBLOG Pre-processing techniques:

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process.

This system only describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process.

So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

IV. IMPLEMENTATION

WEBLOG Pre-processing techniques:

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process.

This system only describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process.

So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

IV. IMPLEMENTATION

WEBLOG Pre-processing techniques:

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process.

This system only describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process.

So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

IV. IMPLEMENTATION

WEBLOG Pre-processing techniques:

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process.

This system only describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process.
Data cleaning eliminates irrelevant or unnecessary items in the analyzed data. A web site can be accessed by millions of users. The records with failed HTTP status codes also may involve in log data. Data cleaning is usually site specific, and involves extraneous references to embedded objects that may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data [8]. An algorithm for cleaning the entries of server logs is presented below –

Input: Log Table (LT)
Output: Summarized Log Table (SLT)

'***' = access pages consist of embedded objects (i.e .jpg, .gif, etc)
'***' = successful status codes and requested methods (i.e 200, GET etc)

Begin
1. Read records in LT
2. For each record in LT
3. Read fields (Status code, method)
4. If Status code='***' and method='***'
Then
5. Get IP_address and URL_link
6. If suffix.URL_Link = {*.gif,*.jpg,*.css}
Then
7. Remove suffix.URL_link
8. Save IP_address and URL_Link
End if
Else
9. Next record
End if
End

3. Session Identification

A user session means a delimited set of user clicks (click stream) across one or more Web servers. The goal of session identification is to divide the page accesses of each user into individual sessions. At present, the methods to identify user session include timeout mechanism and maximal forward reference mainly. The following is the rules we use to identify user session in our experiment:

1) If there is a new user, there is a new session;
2) In one user session, if the refer page is null, there is a new session;
3) If the time between page requests exceeds a certain limit (30 or 25.5 minutes), it is assumed that the user is starting a new session.

4. User Identification

A user is defined as the principal using a client to interactively retrieve and render resources or resource manifestations. User identification is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, it’s difficult because of security and privacy. In our experiment, we use the following heuristics to identify the user:

1) Each IP address represents one user;
2) For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, an IP address represents a different user;
3) Using the access log in conjunction with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, there is another user with the same IP address.

K MEANS:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Proposed Hierarchical clustering:

- Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

1. Begin with the disjoint clustering having level
L(0) = 0 and sequence number m = 0.

2. Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)]=\text{mind[(i),(j)]} where the minimum is over all pairs of clusters in the current clustering.

3. Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to \( L(m) = d[(r),(s)] \)

4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way: \( d[(k), (r,s)] = \min \{d[(k),(r)], d[(k),(s)]\}\)

5. If all objects are in one cluster, stop. Else, go to step 2

**Fast Greedy k means:**

The fast greedy k-means algorithm uses the concepts of the greedy global k-means algorithm for a global solution. The intermediate convergence is done using the restricted k-means algorithm instead of the naïve k-means algorithm. The following 4 steps that have been illustrated in the diagram below outline the algorithm.

1. Construct an appropriate set of positions/locations which can act as good candidates for insertion of new clusters;
2. Initialize the first cluster as the mean of all the points in the dataset;
3. In the K\textsuperscript{th} iteration, assuming K-1 clusters after convergence find an appropriate position for insertion of a new cluster from the set of points created in step 1 that gives minimum distortion;
4. Run k-means with K clusters till convergence. Go back to step 3 if the required number of clusters is not yet reached.

**Backpropagation algorithm**

The logistic function is nonlinear and differentiable, allowing the backpropagation algorithm to model classification problems that are linearly inseparable. Backpropagate the error. The error is propagated backwards by updating the weights and biases to rectify the propagated errors. Weights are updated by Equations \( _{wij} \) below, where \( _{wij} \) is the change in weight \( wij \).

**Step1.** Initialize the weights and bias values

**Step2.** Pass the training values to input layer

**Step3.** Calculate the Ij values

**Step4.** For net input Ij calculate Oj values

**Step5.** Based on Oj calculate error on each node

**Step6.** Calculate the updated weights

**Fig. 3 Weblog Preprocessor** - This is the home page of weblog preprocessing. Action will be performed only, when we click the buttons one after the other.

**Fig. 4 Load Weblog Data** - Here when we click on Load weblog data button, a sample web log file is loaded into the text area. So that we can extract some of the fields that are necessary for us to make an operation on it.
Fig. 5 Field Extraction - Here we extract some of the important fields that are needed to predict the class labels. When we click on Field extraction button fields are extracted using field extraction algorithm. And those fields are displayed in the text area.

Fig. 6 User Identification - When we click on user identification button on the frame, we will get the users IP address. Users are the one who visited the web site.

Fig. 7 Session Identification - Session means nothing but users visit duration to a web server. So when we click on Session identification button users session time and date is displayed in the text area.

Fig. 8 Log Record Details - When we save a log data of a web server it is saved in the .txt file. Here when we click on Loadfulllogrecords button, the data in the text file is displayed in the text area which is in the actual format before we apply all the other algorithms or methods.
V. WEB ANALYTICAL SOFTWARES

There are different Web Analytical Softwares which are free/open source, proprietary softwares and Mixed. Free/Open Source softwares are Analog C, AWStats, CrawlfTrack, Open Web Analytics, Piwik, SnowPlow and W3Perl. Proprietary Softwares are Mint, Sawmil, Splunk, Urchin and Unica Netsight. Mixed software like LogZilla. All softwares posses with high error rate as show in Fig. 14.

VI. PITFALLS

1. Software for Data Web Usage Mining is used for one website only.

2. We are having different software’s for Web Usage Mining, but all of them exhibit noise in there capture. This should be eliminated for better use and understanding.

VII. CONCLUSION

Data preprocessing is an important task of WUM application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. Not every access to the content should be taken into consideration. So this system removes accesses to irrelevant items and failed requests in data cleaning. This project effectively used preprocessing algorithms before applying data mining algorithms. Clustering and Neural Network algorithms are applied on the cleaned log file to get high positive classified results for web analysis. After Speed up extraction time when users’ interested information is retrieved and users’ accessed pages is discovered from log data.

VIII. FUTURE SCOPE

In future this implementation gives better performance when applied on real-time website server in order to get the Web users behavior. Improved decision need to enhance in future in order to get better decision making compare to proposed decision tree approach.

IX. REFERENCES

[1] “Data mining techniques and concepts”, Jiawei Han & Micheline Kamber.
[2] “Classification Of Web log Data To Identify Interested Users Using Decision Trees”, by K. R. Suneetha, R. Krishnamoorthi.
[3] “An Overview Of Preprocessing Of Web Log Files For Web Usage Mining”, C.P. Sumathi, R. Padmaja valli,T Santanam. JATIT.
[4] “A Effective and Complete Preprocessing for Web Usage Mining”,by Mr. Sanjay Bapu Thakare. IJCSE.
[5] “Preprocessing of web logs ”, by Ms. Dipa Dixit. IJCSE.
[6] www.kdd.ics.uci.edu/databases/kddcup1999.html
[7] www.cs.waikato.ac.nz/ml/weka.
[8] http://mathworld.wolfram.com/standarddeviation.html
[9] “Fast Processing of Web Usage Mining with Customized Web Log Pre-processing and modified Frequent Pattern Tree”.
[10] “Frequent Pattern Mining in Web Log Data”, by Renáta Iváncsy, István Vajk.