A MODIFIED ALGORITHM FOR DATA CLEANING OF LOG FILE USING FILE EXTENSIONS IN WEB USAGE MINING

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Software Engineering

Submitted By
Surbhi Anand
(801031031)

Under the supervision of
Dr. Rinkle Rani
Assistant Professor
CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2012

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*A Modified Algorithm for Data Cleaning of Log File Using File Extensions in Web Usage Mining*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Rinkle Rani* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Surbhi

(Surbhi Anand)

801031031

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Rinkle 07/06/12

(Dr. Rinkle Rani)

Assistant Professor,
Computer Science and Engineering Department,
Thapar University,
Patiala.

Countersigned by

(Dr. Maninder Singh)
Head,
Computer Science and Engineering Department,
Thapar University,
Patiala.

(Dr. S. K. Mohapatra)
Dean (Academic Affairs),
Thapar University,
Patiala.

Acknowledgement

I express my gratitude and appreciation to all those who have helped me throughout the duration of my research work. It would not have been possible to complete this research without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

At this moment of accomplishment, first of all I pay homage to my guide, Dr. Rinkle Rani for her personal support and great patience at all times. It gives me immense pleasure to pay my gratitude for her valuable advice, constructive criticism, and expensive discussion around my work.

I am also thankful to Dr. Maninder Singh, Head, Computer Science and Engineering Department, and cooperation and Mr. Karun Verma, P.G. Coordinator for their constant support and encouragement.

I would also like to thank all the staff members of Computer Science and Engineering Department for providing me all the facilities required for the completion of my thesis work.

I would like to articulate thanks for support of my classmates and friends for their kind support and help. Last but not the least; I want to acknowledge the contributions of my family, for their constant motivation, inspirations and for supporting me spiritually throughout my life, and the one above all of us, the omnipresent God, for giving me the strength to complete this work.

Surbhi Anand

(801031031)

Abstract

Web pages typically contain a large amount of information that is not part of the main content of the pages, e.g. banner ads, navigation bars, copyright notices, etc. Such noise on web pages usually leads to poor results in Web Mining which mainly depends upon the web page content. This thesis focuses on the problem of web cleaning i.e. the preprocessing of web pages to automatically detect and eliminate noises for Web mining.

Web usage mining is the subject field of Web Mining which deals with the discovery and analysis of usage patterns from web data specifically web logs in order to improve the web based applications. The Web usage mining process consists of three phases: Data Preprocessing, Pattern Discovery and Pattern Analysis.

Preprocessing cleans up the data (server log file) in order to filter out from the data set the automatic requests generated by the web page, which were not specifically requested by the user. In addition to elimination of the irrelevant automatic requests, it is required to remove nonhuman access behaviour (e.g. spiders, crawlers, and automatic web bots) from the web log file. Another type of inaccuracy that may be present and must be removed from the log file is the entries related to the error requests. Preprocessing results also strongly influences the later phases of Web Usage Mining. This makes the preprocessing of server log files a significant step in Web Usage Mining. The data is preprocessed to improve the efficiency and ease of the mining process.

Three algorithms have been discussed in this thesis. The aim of first algorithm is to separate the data fields from the server log entries. The second algorithm stores these separated fields in a relational database. The cleaning technique is discussed in the third algorithm. This algorithm extracts data from the relational database created by second algorithm and then filters the data by eliminating the extraneous and irrelevant entries. The output of this algorithm gives the clean log file consisting of data necessary from the Web usage mining perspective.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Data Mining

The arrival of the computer technology has contributed the ability to produce and store the massive amounts of data. Now the world is not confined only to manually generated files or reports, but has become a giant store where vast amounts of data are collected and exchanged daily.

Today in every field whether it is some organization or university, there are automatic tools that generate large datasets. The businesses these days world widely produce huge sets of data such as sales transactions, finance record, stock trading records, company performance records, product descriptions, product manuals, marketing descriptions, etc.

The rapid advancement in networking has enabled global telecommunication networks to carry tons of data traffic on a daily basis. With the advent of Internet the World Wide Web has became an important place for information dissemination. There are billions of web search engines (such as Google, Yahoo, Bing, etc.), that processes large amount of data. Today the social media (e.g. Facebook, Orkut, Twitter, etc.) has become an essential element in people's life and hence an important source of data of producing numerous pictures, videos, blogs, etc. The list of sources that generate huge amounts of data is endless.

This explosive growth in the volume of available data as a result of the computerization and the fast development of powerful data collection and storage tools makes the analysis of such data very necessary. Therefore, it becomes very essential to extract information from the bulks of data and structure them into useful knowledge that will be helpful for some type of understanding. This leads to the birth of data mining. The field of data mining is young and promises to make great strides in the journey from the data age toward the coming information age. The information gained can be used for various applications which can range from market analysis to production control and science researches.

According to Frawley, data mining is defined as "*The nontrivial extraction of implicit, previously unknown, and potentially useful information from data*" [22]. Data mining can also be defined as the process of discovering meaningful new correlation, patterns and trends by analysing the large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. In simple terms, data mining is the process of discovering new patterns in large data sets.

Depending upon the final goal, data mining techniques can be considered to be descriptive or predictive. Descriptive data mining finds patterns in the data that provide some information about what the data contains. Predictive data mining uses historical data to infer information about the future events. Descriptive data mining aims to summarize data and to highlight their interesting properties, while predictive data mining aims to build models to forecast future behaviour. Figure 1 shows the subtasks that come under descriptive and predictive tasks.

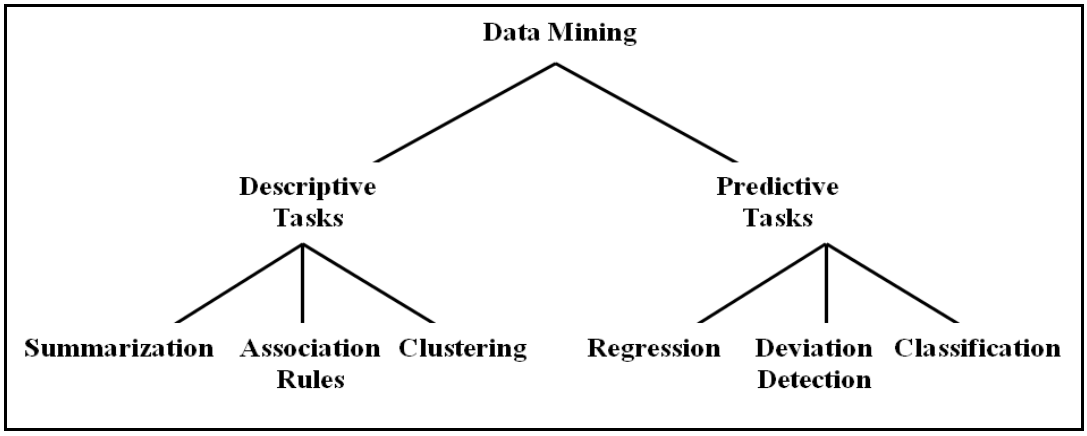


Figure 1: Data Mining Tasks Classification

Data mining is an emerging and interdisciplinary field of computer science. Data mining employs methods at the intersection of artificial intelligence, machine learning, statistics and database systems.

1.2 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases or KDD is the process of extracting previously unknown and potentially constructive information from large sets of data. The term knowledge discovery in databases was coined at the first KDD workshop in 1989 [20] to emphasize that knowledge is the end product of a data-driven discovery [20] [22].

Many researchers consider data mining and knowledge discovery in databases as comparable but data mining is one of the important steps in knowledge discovery process. KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in this process [20]. In data mining specific algorithms are applied for the extraction of data patterns. The other steps of KDD process such as data preparation, data integration, data cleaning, data selection, pattern evaluation, etc. are equally important to perform in order to guarantee that useful knowledge will be extracted from the data. The direct application of data mining techniques can lead to the discovery of insignificant or worthless patterns.

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data [20]. The knowledge discovery process (see figure 2) is an iterative sequence of the following steps:

a) **Data Cleaning and Data integration:** In this step missing data, incoherent data and errors from various sources (such as databases or flat files) are handled. Once the data from various sources have been cleaned then the multiple data sources are integrated into one. The resulting data is stored in a data warehouse.

b) **Data Selection and Transformation:** In data selection step data which is appropriate for the analysis task is retrieved from the database and then during transformation the retrieved data is consolidated and changed into forms relevant for mining purpose.

c) **Data Mining:** In this step algorithms are applied to extract data patterns.

d) **Pattern Evaluation and Knowledge Presentation:** In this step interesting patterns signifying knowledge are identified and then represented in a form such that the user may be able to understand the mined knowledge.

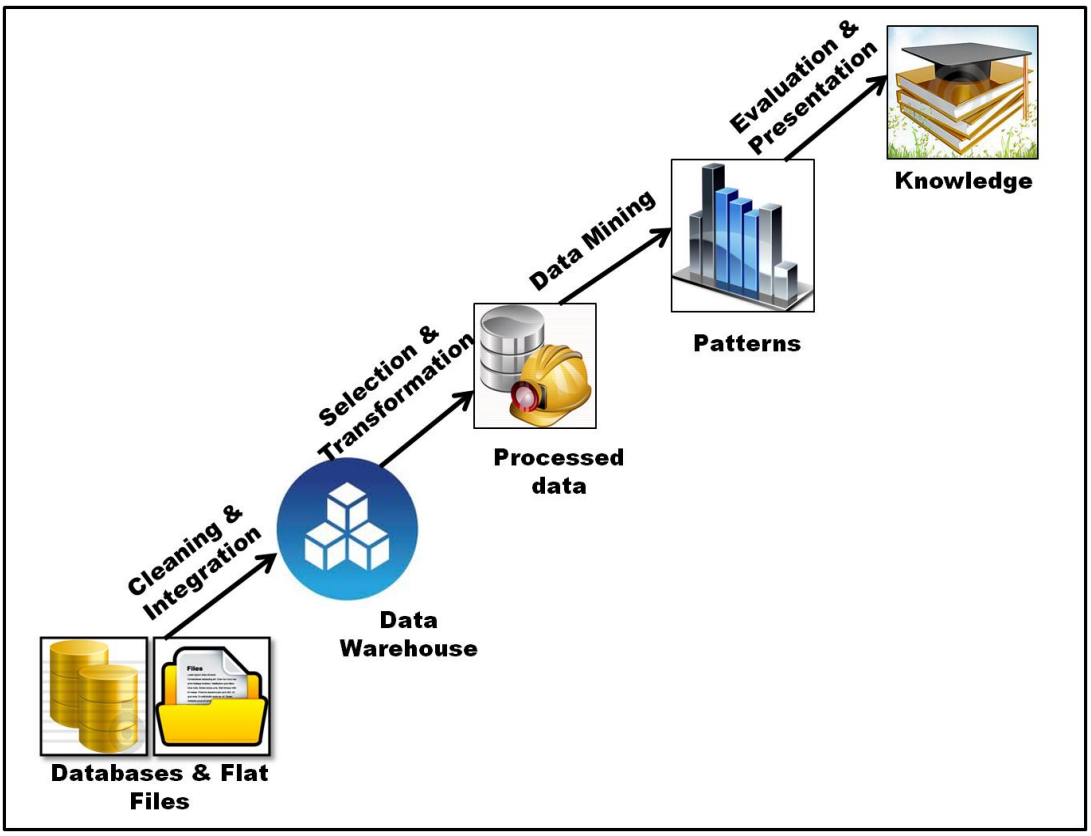


Figure 2: Knowledge Discovery Process

1.3 Types of Data Used for Mining

Today there is a myriad collection of data ranging from simple mathematical measurements and text documents to more complex information such as spatial data, multimedia channels, and web documents. Data mining can be applied to any kind of data repository from flat files to databases as long as data is meaningful and useful for the application. However, the approach may differ from data to data. The type of data on which data mining can be applied is the following subsections.

1.3.1 Relational Databases

A relational database is a collection of data items structured as a set of tables from which data can be accessed. Tables have columns and rows where columns represent attributes and rows represent tuples. Each column contains one or more data categories while each row contains a unique instance of data for the categories defined by the columns. For example a simple employee database would consist of a table with column names as employee's name, address, designation, id, etc. Figure 3 shows an example relational database

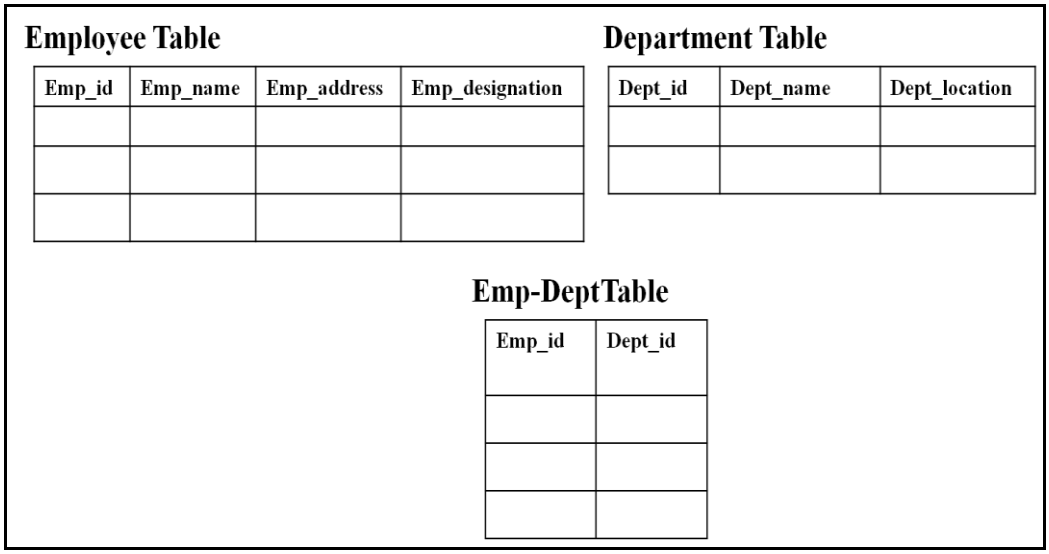


Figure 3: Example of Relational Database

Relational data can be accessed by database query language such as SQL query. Data mining takes advantages from SQL for selection, transformation and consolidation. Data mining also provides applications such as predictions, comparisons and detecting deviations with respect to the relational databases.

1.3.2 Data warehouse

A data warehouse is a heterogeneous repository of data collected from multiple data sources into a single accessible format. It usually contains historical data derived from transaction data. Data warehouses are constructed through a series of steps involving data cleaning, data integration, data transformation, data loading and data refreshing. This is also known as ETL approach where ETL stands for Extraction Transformation and Loading.

A data warehouse is usually modelled by a multidimensional data structure which is known as data cube. In a data cube, each dimension corresponds to an attribute or a set of attributes in the schema and each cell stores the value of some aggregate measure, e.g. avg (age), sum (amount), etc. The data warehouse tools provide support for data analysis but for data mining applications additional tools are needed for comprehensive analysis.

1.3.3 Spatial Databases

Spatial data provides information about the physical location and shape of geometric objects. Spatial data are in the form of graphic primitives that are in form of points, lines,

polygons or pixels. A spatial database is a database optimized to store and query data related to objects in space, including lines, points and polygons. In order to apply data mining algorithms to spatial databases additional functionalities are needed to be added to process spatial data types.

1.3.4 Multimedia Databases

A multimedia is a combination of different media (i.e., text, pictures, sounds, video, animations, etc.) used to present multimodal information in conjunction with computer technology [46]. A multimedia database is a collection of related multimedia data. Therefore, a multimedia database consists of data types such text, graphics (such as drawings, sketches, and illustrations), images, animations, videos, audios, etc. Data mining from multimedia databases entails methods from image processing, computer vision, computer graphics, and natural language processing.

1.3.5 Temporal Databases

The temporal data changes over time and it stores history of how data changed over time. A temporal database has associated with it the built-in time aspects. Data mining in temporal databases involves study of trends and correlations between different variables and prediction of trends and changes in the variable with respect to time.

1.3.6 Web data

World Wide Web is one of the most interactive and popular medium for dissemination of the information today. World Wide Web organizes data in the form interrelated documents, which are also known as web pages in web terminology. Data mining when applied to web data is termed as Web mining.

Web data can be of three types:

a) Content of web, which includes text documents, audio-video content, graphics, images, etc.
b) Structure of web, which embraces hyperlinks between web pages.
c) Usage of web, which elucidates the use of resources.

1.4 Evolution of Web Mining

In 1950's IBM started using magnetic tapes to store data as they were capable of storing higher amounts of data. With the evolution of storage media, the need to manage data on these devices emerged. In the early 1960's the term database captured the sense that the information stored within a computer could be conceptualized and structured, independently of the specific machine on which it is stored. The main aim of developing the databases was to achieve data independence.

At the beginning of the 1970's IBM started a severe research on the relational databases and an IBM researcher E. F. Codd published a paper which introduced the concept of relational data model in which data organization consisted of tables linked by keys identifying specific data records [15]. In 1979 the first commercial relational database system named Oracle was available using the structured query language SQL. Relational databases became the core technology for data management. One important reason is the uniform and standardized interface provided through SQL.

With the advancement of storage devices and the growing storage capacities, it became convenient to manage larger amounts of data. In the early 1970s researchers started to differentiate between operational and analytical systems. It was then the fundamental idea for data warehousing was born and until the 1990's data warehousing started to make its way into business companies.

With increasing processing power data warehousing became the basis for data mining. The data mining tools performs analysis on the data stored in data warehouse. Due to the large amount of data stored in a data warehouse it becomes very difficult for analysts to draw conclusions. Therefore, data mining algorithms were designed to discover patterns in large amounts of data in order to predict trends. The advent of WWW motivated the researcher's attention towards the application of data mining techniques to web data and gave birth to the concept of web mining.

The following table (see table 1) shows the evolutionary steps and supporting technologies in the evolutions of data mining and web mining.

Table 1: Evolution of Data Mining and Web Mining

Evolutionary Step	Technologies Used
Data Collection And Storage	Computer, Magnetic Tapes, Disks.
Data Access	Relational Database (RDBMS), Structure Query language (SQL), ODBC.
Data Warehousing	On-Line Analytic Processing (OLAP), Multidimensional Databases, Data warehouses.
Data Mining	Advanced algorithms, Multiprocessor Computers, Massive Databases.
Web Mining	WWW, Internet, Web databases.

1.5 Web Mining

The advent of the World Wide Web has made the data present on the web as a gigantic source of information. The World Wide Web, having over 350 million pages, continues to grow rapidly at a million pages per day [7]. This increase in the mass of data has turned researcher's attention towards the use of data mining techniques to extract useful information from the web data. Moreover, data mining when applied to the web is potentially beneficiary.

In the data mining communities, there are three types of mining: data mining, web mining, and text mining [49]. Data mining when applied on web data is described as the application of data mining algorithms and techniques to the variety of data forms, structures and usage patterns that comprise the World Wide Web. This method in mining terminology is known as web mining.

Web mining combines the two of the activated research areas i.e. Data Mining and World Wide Web. Therefore, Web mining can be defined as the application of data mining

techniques in order to discover patterns from the Web data. Web mining can be classified into three types [40] [44] as shown in figure 4:

a) Web Content Mining

b) Web Structure Mining

c) Web Usage Mining

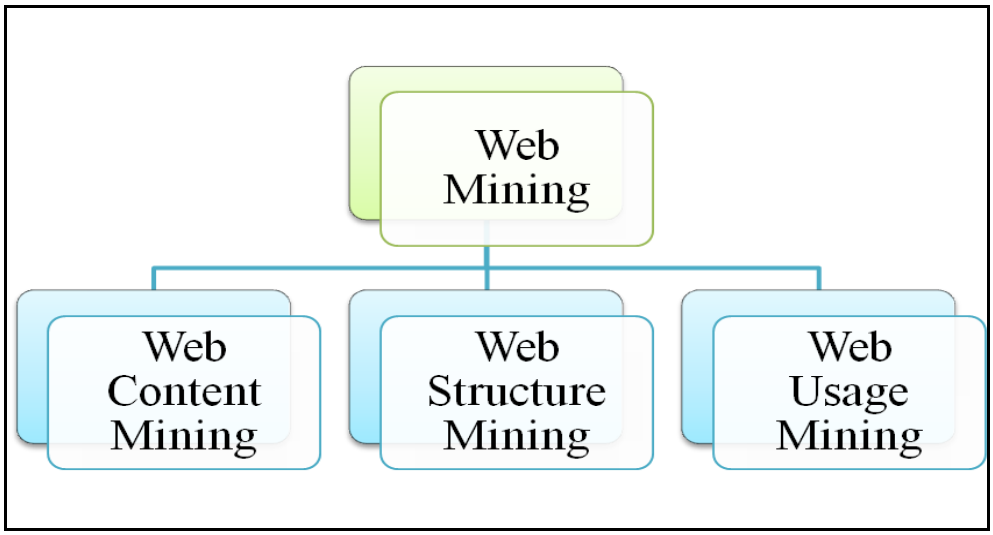


Figure 4: Web Mining Types

a) Web Content Mining: The aim of web content mining is to analyse the web content (such as text, multimedia data, and structured data) that may be located within the same web pages or linked across web pages. Web content mining helps a researcher to understand the content of web pages, provide keyword-based page indexing, web page relevance, web page ranking, web page content summaries, and information related to web search and analysis.

b) Web Structure Mining: Web structure mining is the process of using graph and network mining theories to comprehend the nodes and hyperlink structures on the Web. It can mine the document structure within a web page or across the different web pages.

c) Web Usage Mining: Web usage mining focuses on the extraction of useful information from server logs. It tries to discover the patterns that are related to some general or a particular group of users, understand user search patterns and envisage what users are looking for on the Internet.

Chapter 2

Literature Survey

2.1 World Wide Web (WWW)

The World Wide Web abbreviated as WWW and commonly known as the Web is a system of interlinked hypertext documents which can be accessed through the Internet. The World Wide Web is officially defined as a "*wide area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents*" [31]. In simpler terms it can be said that the web is an Internet based computer network that consent to users of one computer to access information stored on another computer through means of the Internet.

It provides the facility of connecting documents to other documents by means of hypertext links and enables the user to search for information by moving from one document to another. With the help of a web browser, a user can view web pages that may contain text, images, videos and other multimedia, and navigate between them using the hyperlinks.

The operation of World Wide Web depends on the structure of its hypertext documents. The hypertext enables the authors of web pages to link their documents to other related documents. To view these documents, users have to follow the links, which in web terminology are known as hyperlinks. The idea of hypertext was given by Ted Nelson in 1965. The concept of hypertext is considered as a milestone in the history of World Wide Web. In 1989 Tim Berners Lee, who was a consultant with the CERN (European Organization for Nuclear Research), invented the World Wide Web. He coined the term World Wide Web [31].

2.1.1 Challenges Associated with the World Wide Web

World Wide Web is one of the most interactive and popular medium to spread the information. The increasing popularity and size growth of WWW has overwhelmed us with an immense amount of widely dispersed interconnected and dynamic information. It has profoundly influenced many aspects of human lives and changed the way of communication, conducting businesses, etc. Additionally the web data is also gigantic, diverse and dynamic in

nature due to which users could encounter problems while interacting with the web. Following are the challenges associated with the WWW:

- **Finding Relevant Information:** The amount of information available on the web is vast and still emergent. Moreover, coverage of the information is very broad and diverse. Users use the search service to find out the specific information on the web. These days search tools have certain problems like low accuracy due to insignificance of many of the search results. This results in difficulty in finding the relevant information.

- **Heterogeneity:** Information available on Web is heterogeneous in nature. The reason behind is that authors of different web pages vary due to which multiple pages may refer to the same or similar piece of information but at variance in language or words and formats. This makes the integration of information from different pages a problem.

- **Semi Structured or Unstructured Data:** Generally, there are two types of data available on the web. First is structured and second one is unstructured. Structured data are arranged in a database like structure in which specific information is stored based on a methodology of columns and rows. While unstructured data refers to any data that has no identifiable structure. Images, videos, vector graphics, documents and text are considered to be unstructured data within a dataset. Both semi structured and unstructured data does not have a compact or precise description of data items.

- **Noisy:** Information on web is noisy because of two reasons.

    i. Firstly, a typical web page comprises of many different piece of information (e.g. main page content, advertisements, navigational links, etc). For a particular user search only part of information is important and rest is ineffectual.

    ii. Secondly, there is no control over the quality of the information available on the web because there is no restriction on what one writes.

  Therefore, much of information available on web is sometimes misleading and ambiguous.

- **Dynamic:** Web is very dynamic in nature since the information available on it changes frequently because of changing requirements or for some other reasons like in order to improve readability. Therefore, it is needed to maintain these changes.

- **Redundancy:** Data available on web may be redundant since same segment of information or its variants may appear in many pages.

- **Personalization of Information:** Users of internet differs in their experience intended for the contents they search for and the presentation of their search result. It leads to the problem of personalization.

2.2 Web Mining

World Wide Web is a monolithic repository of web pages that provides the Internet users with heaps of information. With the brisk growth of the World Wide Web, the web has become an imperative medium of information dissemination. Therefore, the information available on the Web is a vital source of information for the users of the internet. Due to these reasons there is an increase in the number and size of websites available on internet which makes the World Wide Web remarkably gigantic. However, the abundant information on the web is not stored in any systematically structured way, which poses great challenges to those looking for high quality information underlying in web pages. The growth in the mass of data present on web has engrossed the attention of the scholars and researchers towards the application of data mining techniques on the data available on the web in order to extract useful information. Web mining has therefore become an important subject matter in data mining.

In 1996 Etzioni was first who coined the term web mining [19]. According to him Web mining is the use of data mining techniques to automatically extract information from World Wide Web documents and web services. The aim of Web Mining is to discover constructive information from the hyperlink structure of web pages, webpage content, and web usage data. Therefore, Web mining can be defined as the application of data mining techniques in order to discover patterns from the web data, comprising web documents, hyperlinks between documents, and usage logs of the websites [41].

There were two different approaches proposed for defining Web mining. First approach is a process centric view and second approach is a data centric view. The data centric definition has become more acceptable [9] [32].

i. From the process centric view, web mining is defined as a sequence of ordered tasks [19].

ii. From the data-centric view, web mining is defined with respect to the types of web data that was used in the mining process [16].

Though web mining takes its foundations from the data mining techniques but it does not solely depends on data mining. The reason is that data mining is applied on structured data while web mining is applied on web data which is heterogeneous and unstructured or semi structured in nature. Data mining is work upon offline whereas web mining is work upon online [41]. Data mining is performed on data stored in database or data warehouse while web mining is performed on data stored in server database and web log.

2.3 Web Mining Tasks

Web mining can be decomposed into the following subtasks [8] [19]:

a) Information Retrieval (or Resource Discovery): Search is probably the one of the prime application of the Web having its roots in information retrieval. Information retrieval (IR) helps the users to find the required information available from a large collection of text documents. Given a query which expresses the user's information necessity, an IR system finds a set of documents from its underlying collection matching the stated query. IR has the primary goals of indexing text and searching for useful documents in a collection. Nowadays research in IR includes modelling, document classification, document categorization, data visualization, filtering, etc.

b) Information Extraction (Selection and Preprocessing): This task deals with the transformation of the data retrieved during information retrieval process into a form that can be easily analysed. The goal of information extraction (IE) is to transform a collection of documents into information that is more readily digested and analysed [17]. Information extraction aims to select relevant facts from the documents while information retrieval aims to select relevant documents. Information extraction is concerned with the structure of a document while information retrieval considers the text in a document as a bag of unordered words. Thus information extraction is at a more granular level that information retrieval.

c) **Generalization (Pattern Recognition and Machine Learning):** It automatically generates general patterns from both the individual web sites as well as across multiple sites. Machine learning methods or data mining techniques are generally used for the generalization purpose.

d) **Analysis (Validation and Interpretation):** Once the patterns have been identified it is necessary to explore and confirm those mined pattern. The aim of this task is to analyse and validate the mined patterns.

Based on the above mentioned subtasks, web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services [41].

2.4 Web Data

The most essential step in the knowledge discovery process is to create a suitable target set of dataset for the mining tasks. In web mining data can be collected at the server side, client-side, and proxy servers or can be obtained from an organization's database. The type of data collection differs not only in terms of the location of the data source but also the kinds of data available, the segment of population from which the data was collected and its method of implementation [42]. The data that can be used in web mining is classified as [42]:

- **Content:** It is the data that a web page is designed to convey to the users. It consists of free text, semi-structured data like HTML pages and more structured data like automatically generated HTML pages, XML documents or data in tables related to web content. The data types such as text, image, audio and video also comes under this category.

- **Structure:** It is the data that describes the organization of the content. It can be of two types: intra page structure information and inter page structure information. The intra page structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure where the HTML tag becomes the root of the tree. The inter page structure consists of hyper-links connecting one page to another.

- **Usage Data:** Usage data includes web log data obtained from the web server access logs, proxy server logs, browser logs, cookies and any other data generated as the

results of web user interactions with web servers. It is the data that describes the pattern of usage of web pages, such as IP addresses, page references and the date and time of accesses.

- **User Profile**: It is the data that provides demographic information such as registration data and customer profile information about users of the Web site.

2.5 Classification of Web Mining

The most standard categories of the web mining (see figure 5) are web content mining, web structure mining, and web usage mining [40][44]. This classification is based on the type of web data to be mined.

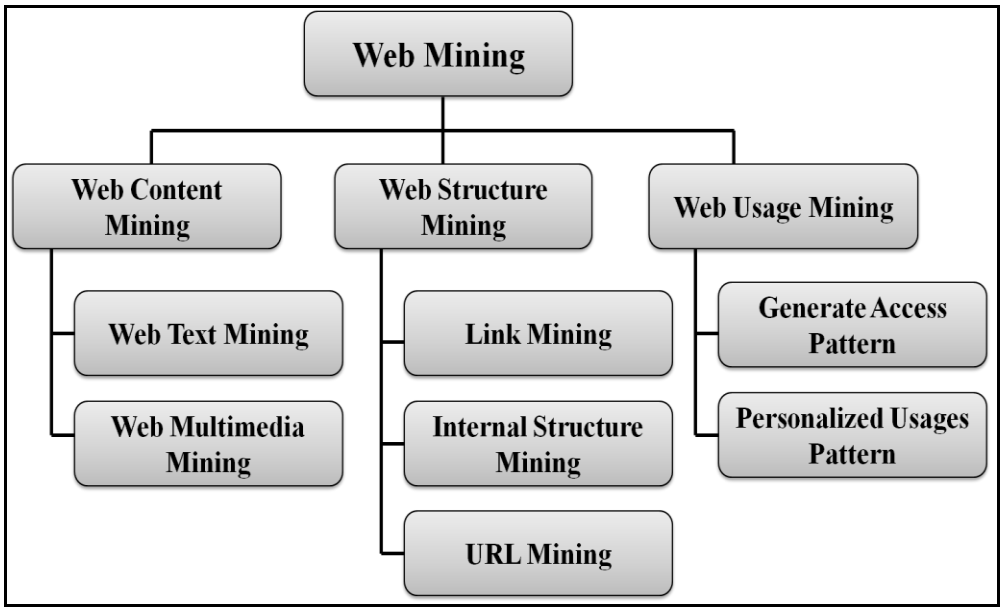


Figure 5: Web Mining Classifications

A brief overview of the above three categories are given as follows:

a) Web Content Mining: deals with the extraction of useful information from the contents (e.g. text, images, etc) of Web documents.

b) Web Structure Mining: deals with the discovery and modelling of the hyperlink structure of websites.

c) Web Usage Mining: deals with the discovery of user's access pattern from log records of web pages.

Table 2 shows the web mining categories and a summary of each type of web mining.

Table 2: Web Mining Categories [8]

	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR View	**DB View**		
View of Data	▪ Unstructured ▪ Semi Structured	▪ Semi Structured ▪ Web site as DB	Links Structure	Interactivity
Main data	▪ Text documents ▪ Hypertext documents	▪ Hypertext documents	Links Structure	▪ Server Logs ▪ Browser Logs
Representation	▪ Bag of words ▪ Terms or phrases ▪ Concepts or Ontology ▪ Relational	▪ Edge Labelled Graph ▪ Relational	Graph	▪ Relational Table ▪ Graph
Method	▪ Machine Learning ▪ Statistical	▪ Proprietary Algorithms ▪ Association rules	Proprietary Algorithms	▪ Machine Learning ▪ Statistical ▪ Association rules
Application Categories	▪ Categorization ▪ Clustering ▪ Finding Extraction Rules ▪ Finding Patterns in Text ▪ User Modeling	▪ Finding Frequent Sub Structures ▪ Web site schema discovery	Categorization and Clustering	▪ Site Construction, Adaptation and Management ▪ Marketing ▪ User Modeling

2.6 Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of web documents. Web content mining describes the automatic search of information resource available online and involves mining the contents of web data [32]. The main source of web content mining are the web documents which may consists of text, images, videos, audio clippings, abstract graphics, hyperlinks etc. Therefore, a web document can be semi structured such as HTML documents or a more structured data like the data in the relational database tables generated through HTML pages but most of the data is unstructured text data. However, much of the web content data is unstructured text data [19]. The unstructured attribute of Web data makes the web content mining process to be a more complex approach. The Web content mining is differentiated from two different points of view [8] [42]:

a) Information Retrieval View

b) Database View

The research work done for unstructured data and semi-structured data from information retrieval view is summarized in [8]. For the unstructured data, the researchers use a bag of words. In this approach words are isolated from each other to represent unstructured text. Then a single word, which is found in the training mass, is taken as feature for retrieving the information. On the other hand, for the semi structured data a different approach is followed. For semi-structured data either the HTML structures inside the documents or the hyperlink structure between the documents are utilized for document representation.

The application of database techniques on the web is related to the problems of managing and querying the information on the web. There are three classes of tasks related to those problems: a) modelling and querying the web, b) information extraction and integration, c) Website construction and restructuring [21]. The first two problems are related to web content mining but not all the work there comes under its scope. This is due to the lack of machine learning and data mining techniques in web content mining process. Therefore from the database view, the mining process tries to comprehend the structure of the web site in order to transform a web site to become a database. This helps in improved information management and querying on the Web. The two main categories are multilevel databases and web based query systems [5].

The main idea behind having multilevel databases is that the lowest level of the database contains semi structured information which is stored in various web repositories such as hypertext document and metadata are extracted from lower levels and organized in structured collections, i.e. relational or object-oriented databases. For example, the ARANEUS system extracts relevant information from hypertext documents and integrates these into higher level derived web hypertexts which are generalizations of the notion of database views [34].

Web based query systems and languages use standard database query languages (such as SQL), structural information about Web documents and natural language processing for the queries that are used in World Wide Web searches. W3QL combines structure queries which are based on the organization of hypertext documents and content queries which are based on information retrieval techniques [30].

A major research area in the domain of Web content mining is related to multimedia databases. Multimedia mining is a part of the content mining which is engaged to mine the multifaceted information and knowledge from large online multimedia data sources. Multimedia data mining on the Web has recently gained attention from many researchers. Although multimedia data has received the major attention from many researchers and many techniques for multimedia information retrieval and extraction have been proposed yet the multimedia mining is still in infancy [48]. Information retrieval is one of the research areas that provide a range of popular and effective, mostly statistical methods for web content mining [40]. Web multimedia mining is related to various research areas such as computer vision, multimedia processing, multimedia retrieval, data mining, machine learning, database and artificial intelligence.

Another field of research that comes under the web content mining is text mining. Text mining is the data analysis of text resources so that novel and previously unknown knowledge can be discovered from the text documents [25]. Text documents are different from the information stored in database systems since they are unstructured in nature. The field of text mining has received plenty of attention owing to the necessity of managing the information that resides in the immense amount of available text documents.

2.7 Web Structure Mining

Web structure mining focuses on discovering and modelling the link structure of websites [40]. There are many web information retrieval tools available that ignores the link information of the web pages that could be beneficiary for the researchers and the users. The goal of Web structure mining is to produce the structural summary of the websites by using the hyperlinks present within the web page or across the web pages. The difference between web content mining and web structure mining is that former mainly focuses on the structure of the intra-document, while the latter tries to discover the link structure of the hyperlinks at the inter document level. The main focus of web structure mining is on hyperlink information.

Web structure mining categorizes the web pages and generates the information from the topology of the hyperlinks between the web pages to illustrate the similarities and relationships between different web sites. The web structure mining has two levels:

a) **Intra document level structure mining:** In intra document level structure mining the organisation of hyperlinks within a webpage that connects to a different part of the same page is examined.

b) **Inter document structure mining:** The inter document level structure mining can be used to reveal the structure (schema) of Web pages. It solves the purpose of the navigation and helps in comparison and integration of the Web page schemas

When a Web page is linked to another web page directly or the web pages are neighbours, it is desirable to discover the relationships among those Web pages. These web pages may either be related by synonyms or ontology [44]. When the web pages are synonyms it means they have similar contents. When the relationship between the web pages is ontology it means that they reside in the Web server created by the same person.

Some algorithms that were proposed for web structure mining are: HITS [29] and PageRank [10] and improvements of HITS by adding content information to the links structure [12]. These algorithms are applied to evaluate the quality rank and relevance of each web page. HITS was first time used in the Clever search engine [12], and PageRank was used by Google [10].

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates Web pages. HITS introduce the concepts of hubs (i.e. pages that refer to many other pages) and authorities (i.e. pages that are referred by many other pages). The idea behind Hubs and Authorities originated from the creation of web pages. According to Kleinberg, "*Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs*" [29].

During the design of web pages certain web pages were stated as hubs. These pages did not carry any authoritative information but were used to serve as large directories of information that led users directly to other pages. The pages to which the hubs led are known as authoritative pages.

A good hub represents a page that points to many other pages, and a good authority represents a page that was linked by many different hubs. Hubs and Authorities can be viewed as fans and centers in a bipartite core of a web graph as shown in figure. The nodes on the left represent the hubs and the nodes on the right represent the authorities as shown in the figure 6.

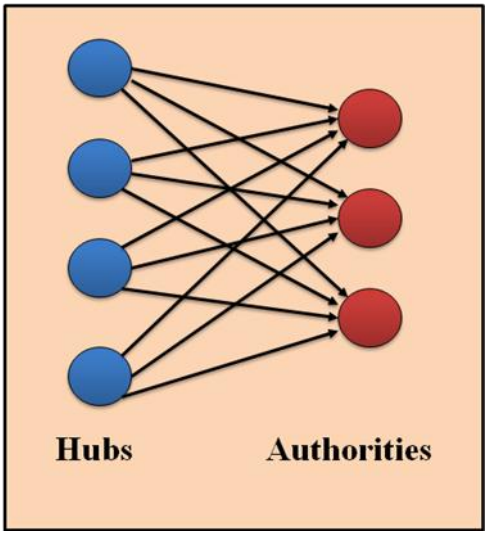


Figure 6: Bipartite Core

Another important research done in the field of web structure mining was related to PageRank calculation [10]. PageRank is a link analysis algorithm used by Google search engine. It assigns a numerical weight to each element of a hyperlinked set of web documents with the purpose of measuring its relative importance within the set.

L. Page and S. Brin [10] proposed the PageRank algorithm to calculate the importance of web pages using the link structure of the web. In their approach Brin and Page extended the idea of simply counting inlinks equally by normalizing the number of links on a page. Let a web page A has pages T1...Tn that points to A. The parameter d is a damping factor which can be set between 0 and 1. The damping factor d is usually set to 0.85. C (A) is the number of links going out of page A. Then the Page Rank of a page A is:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it (i.e. incoming links). A page that is linked to by many pages with high PageRank receives a high rank itself but if there is a web page with no links pointing to it, implies that there is no support for that page.

The web is viewed as a directed graph whose nodes are the documents and the edges are the hyperlinks between them and the graph structure of the World Wide Web can be exploited for improved retrieval performance and classification accuracy [23]. Many search engines utilizes the concept of viewing the web as directed graph and uses the graph properties in ranking their query results. The web as a directed graph is shown in figure 7. The figure 7 clearly shows that the edges represent the hyperlinks and nodes represent web documents.

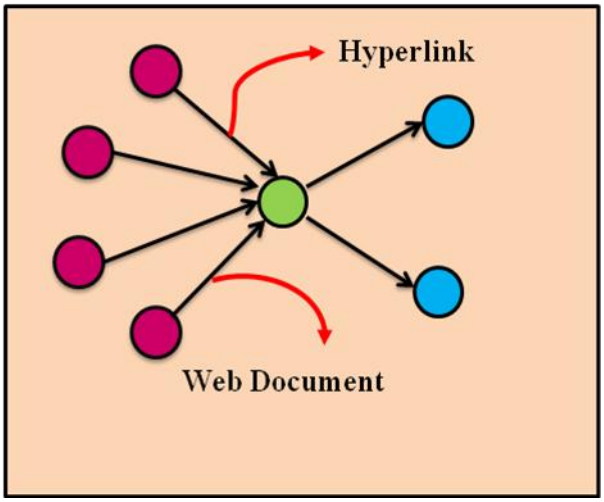


Figure 7: Web Graph Structures

Apart from search ranking, hyperlinks are also useful for finding web communities [40]. A web community is a collection of web pages which focuses on a particular topic or

theme. Many graph clustering algorithms may be used for mining the community structure of a graph as they adopt the same assumption. They assume that a cluster is a vertex subset such that for all of its vertices, the number of links connecting a vertex to its cluster is higher than the number of links connecting the vertex outside its cluster.

2.8 Web Usage Mining

Web usage mining is process of discovering usage patterns from Web data, in order to better understand the needs of web based applications. Web usage mining deals with the extraction of knowledge from web server log files. The main source of data for Web usage mining mainly consists of the (textual) logs, which are collected when users access web servers. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries:

i. General Access Pattern tracking

ii. Customized Usage Tracking

A high level web usage mining process is shown in figure 8. The Web usage mining is parsed into three distinctive phases [42]:

a) Data Preprocessing- It performs a series of steps covering data cleaning, user identification, session identification, path completion and transaction identification.

b) Pattern Discovery- It involves application of various data mining techniques to processed data like statistical analysis, association, clustering and pattern matching.

c) Pattern Analysis- It filters out the irrelevant patterns from the identified patterns generated in pattern discovery phase.

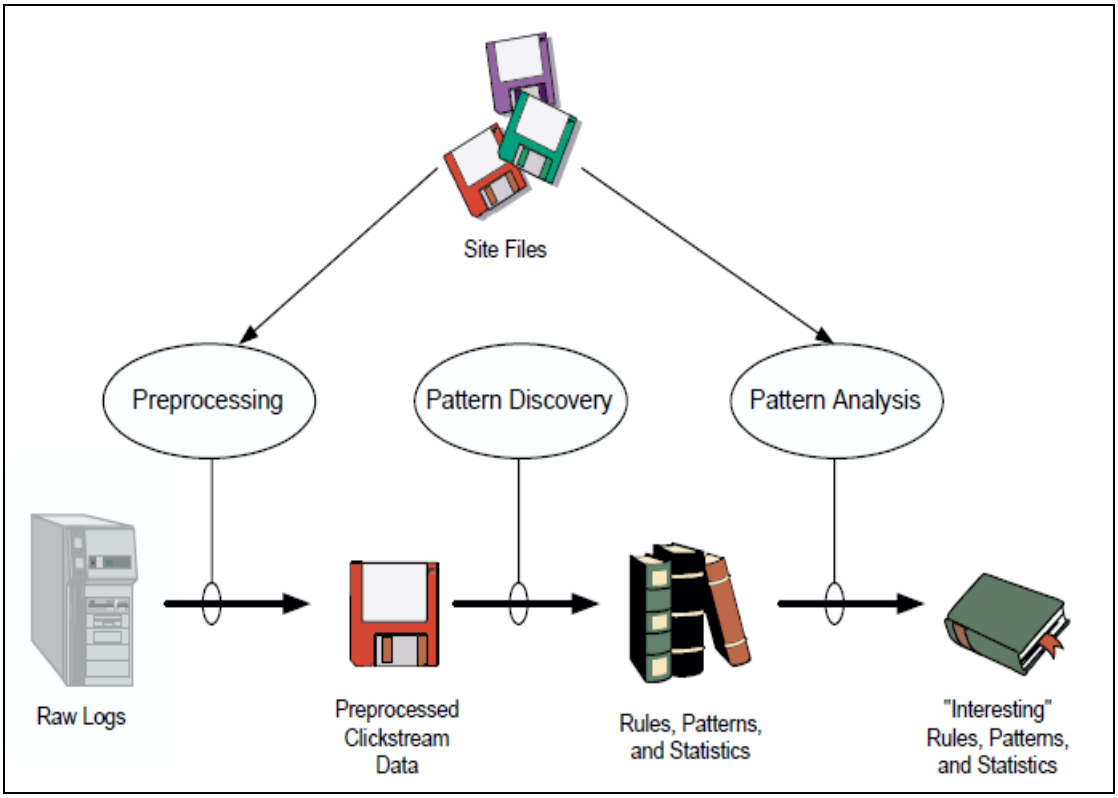


Figure 8: A High Level Web Usage Mining Process [42]

The Web usage mining process can be classified into two commonly used approaches; first approach maps the usage data of the Web server into relational tables before an adapted data mining technique is performed while second approach uses the log data directly by utilizing special preprocessing techniques [9].

As the preprocessing or data cleaning is important in data mining applications, similarly it is important in web usage mining applications also. The applications of Web usage mining could be classified into two main categories: learning a user profile or user modelling adaptive interfaces (personalized) and learning user navigation patterns (impersonalized) [8].

The web usage mining is one of the major areas of research in the domain of Web mining. A World Wide Web usage mining examination tool called SpeedTracer was created in order to realize user browsing pattern by investigating the Web server log files with data mining procedures [45].

A Web usage mining system, called SUGGEST, was proposed in 2002 [4]. SUGGEST creates the connections to Web pages of probable importance for a user.

Web based recommender systems are very helpful in directing the users to the target pages in particular web sites. Therefore, it appears to be beneficiary to have Web usage mining recommender systems that can predict targets and navigation behaviour of a user. A web usage mining technique based on LCS (Least Common Sequence) algorithm for online predicting recommendation systems was suggested in [27]. The authors proposed a model for online predicting through web usage mining system and recommended an approach for classifying user navigation patterns to predict user's future intention. The approach used LCS algorithm for categorizing user navigation behaviour for forecasting user's future requests. For providing the online prediction effectively, architecture for online recommendation for predicting in Web Usage Mining System was described in [39]. The author described the architecture of On Line Recommendation in Web Usage Mining (OLRWMS) for improving the accuracy in prediction.

Web usage mining applications are based on data collected from three main sources [42] shown in figure 9:

a) Web Servers

b) Proxy Servers

c) Web Clients

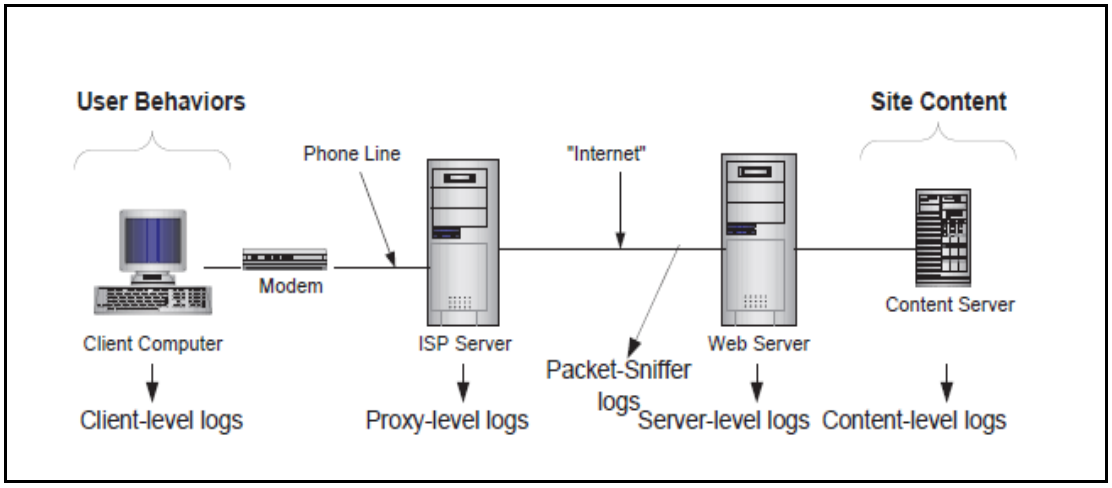


Figure 9: Simplified Web Access Diagram [16]

a) The Server Side: Web servers are one of the most common sources of data. They collect large amount of information in their log files. A web server records and accumulates data about the user's interactions whenever user sends request for resources to the web server. This piece of information is named as web logs. A web

log is a file in which Web server records information about the user's request for a resource from a particular website. Web server logs are plain text ASCII files that are independent of server platform. A log entry records the traversal from one page to another and stores the details like user IP number or domain name, time and type of access method (GET, POST, etc.) and address of the page being accessed.

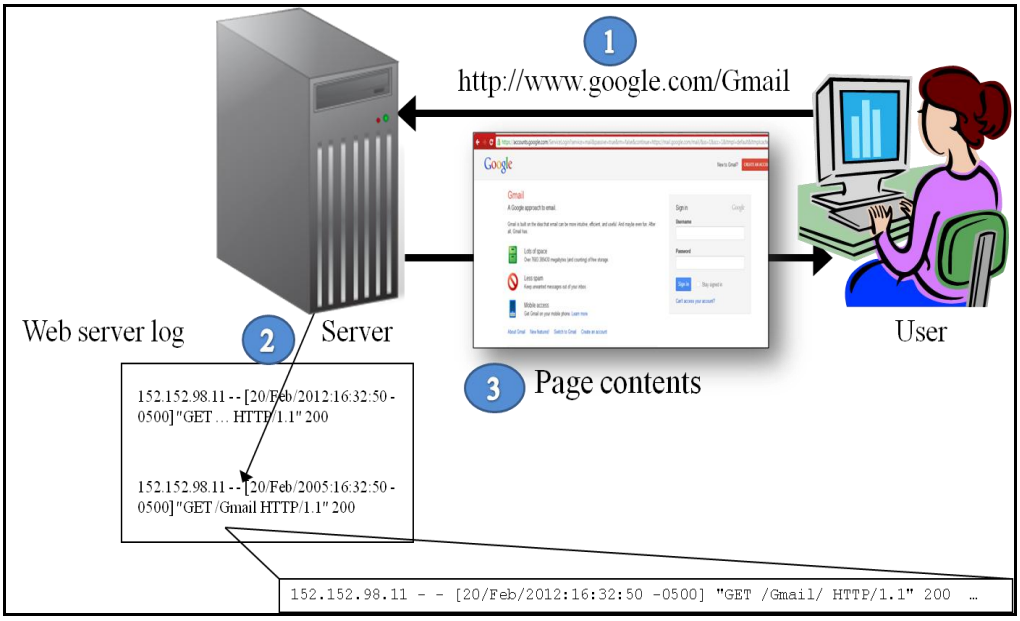


Figure 10: Working of Web Server in Creating Log Entry for a Request

Figure 10 explains the working of a web server in creating a log entry for a user request. It is a three step process:

1. Client requests for some resource from the web server. In this example client is trying to access the gmail account.

2. When the web server receives the user request, it records the details such IP address of the client, timestamp when request was made, type of request, etc. Web server stores these details in the log file.

3. After creating the log entry for the client request, the web server sends the requested resource to the client.

The analysis of the web access logs from different web sites can help to understand the user behaviour and the web structure, thereby improving the design of collection of web resources. These logs usually contain basic information e.g. name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc.

The log files can be stored in various formats such as Common Log Format (CLF) or Extended Common Log Format (ECLF). The main data fields of any log files are: client IP address, user id, hostname, timestamp, offset, method, path, protocol, status sand bytes. Figure 11 shows a single log entry which may be a part of a large server log file.

```
213.135.131.79 - - [15/May/2002:19:21:49 -0400] "GET /features.htm HTTP/1.1" 200 9955
```

Figure 11: A Single Log Entry

- Client IP address indicates the internet address of the machine from which request has been made,
- User ID and hostname are filled only when the authentication is required.
- Timestamp consists of time and date when the request was made.
- Method specifies the type of request (i.e. GET, POST or HEAD). GET method requests an object from the Web server, POST method sends information to the web server and HEAD method request the HTTP header for an object.
- Path indicates the actual path or URL of the requested resource.
- Protocol specifies the protocol used for sending the request.
- Status field indicates the action taken in response to a request. Not all browser requests succeed. Codes from 200 to 299 indicate success. A common error code is 404, which indicates that the client's request cannot be fulfilled, due to incorrect syntax or a missing file.
- Bytes state the number of bytes that have been transferred while processing the request. Only GET requests that have been completed successfully (Status = 200) will have a positive value in the transfer volume field. Otherwise, the field will consist of a hyphen or a value of zero.

b) The Proxy Side: An internet service provider presents proxy server services to improve the navigation speed by enabling the use of caching. A proxy server act as an intermediate between client browser and the web server. The use of proxy serve not only improves the navigation speediness but also reduces the network traffic load at both the client and server side. Collecting navigation data at the proxy level is

basically the same as collecting data at the server level in many respects. Proxy caching not only reduces the loading time of a web page experienced by users but also the network traffic at the server and client sides.

c) **The Client Side:** Data can be collected at client side by using a remote agent such as javascript or java applets as well as by changing the source code of existing browser to facilitate improvement in its data collection properties [16]. These techniques avoid the problems of user sessions identification and the problems caused by caching as the data is located at the user's site. It also provides detailed information about the actual user behaviours. However, these approaches depend on the user's cooperation and can cause many issues concerning the privacy laws which are quite strict.

The information provided by the above data sources is used in the construction and identification of the several data abstractions like users, server sessions, user session, clickstreams and page views [16].

In order to provide some consistency in the way these terms are defined, the W3C Web Characterization Activity (WCA) has published a draft of Web term definitions relevant to analyzing Web usage [16].

The definitions of these terms as described in the drafts are as follows:

- **User:** A user can be an individual who accesses or requests for some resources from the web servers through a browser.
- **Page View:** A page view consists of every file (webpage) that is displayed on a user's browser at a specific point of time.
- **Click-stream:** A click-stream is a sequential series of page view requests.
- **User Session:** A user session consists of the click-stream of page views for a single user across the web.
- **Server Session:** A server session is a set of page views in a user session for a particular web site.

2.9 Web Usage Mining Process

A detailed web usage mining process with its sub phases is given in figure 12. The three steps invloved in Web usage mining process are as follows:

a) **Data Preprocessing-** It performs a series of steps covering:
   - Data Cleaning
   - User Identification
   - Session Identification
   - Path Completion
   - Transaction Identification

b) **Pattern Discovery-** It involves application of various data mining techniques to processed data like,
   - Statistical Analysis
   - Association
   - Clustering
   - Pattern Matching

c) **Pattern Analysis**- It filters out the irrelevant patterns from the identified patterns generated in pattern discovery phase.

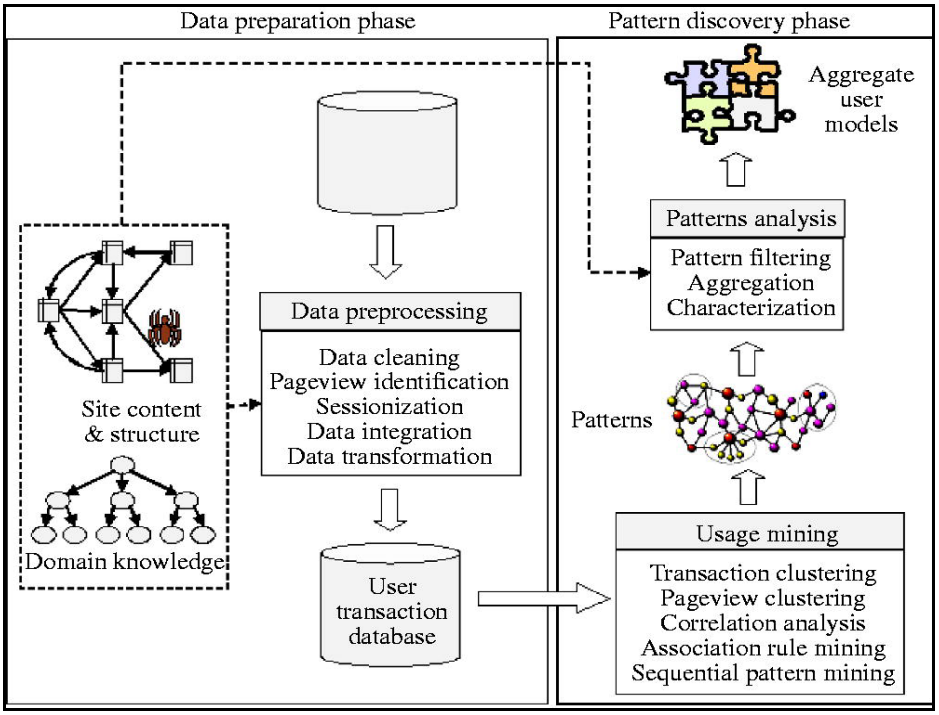


Figure 12: Detailed Web Usage Mining Process [31]

2.9.1 Data Preprocessing

The information that can be accessed through web is heterogeneous and semi structured or unstructured in nature. Due to this heterogeneity a web log file may consists of some undesirable log entries whose presence does not matters from the web usage mining point of view. This makes the preprocessing of log file an important precondition for discovering the knowledgeable patterns. The purpose of performing preprocessing is to transform the raw click stream data into a set of user profiles.

Data preprocessing tasks are carried out former to the application of mining algorithms. Preprocessing enables to translate the unprocessed data which is composed from server log files into constructive data abstraction. The appropriate analysis of a web server log proves to be beneficiary to manage the websites efficiently from the administrative and user's perspective. Preprocessing results strongly influences the later phases of web usage mining. This makes the preprocessing of server log files a significant step in Web Usage Mining. Data preprocessing is one of the most complex phase of the Web Usage Mining process.

Data preprocessing presents a number of unique challenges which led to a variety of algorithms and heuristic techniques for preprocessing tasks such as merging and cleaning, user and session identification, and path completion. Many researchers have explored the area of data preprocessing extensively in order to overcome the challenges related with this field. A summary of research done in data preprocessing area is given in the table 3. It summarizes the preprocessing techniques based on the parameters such as source of log file, format log file, preprocessing technique, and algorithm applied to the preprocessing phase.

The table shows that in most of the preprocessing techniques, a server log file was used. A method for user identification, session identification, page view identification, path completion, and episode identification was discussed by R. Cooley, J. Srivastava and their co researchers in [42].

Pabarskaite worked on advance data cleaning and filtering techniques on web server log [36]. In 2003 Yuan and his fellow researchers proposed an algorithm that was able to perform the complete preprocessing on server log [47]. His algorithm consisted of element for data cleaning, user identification, session identification and path completion. The

proposed approach uses the IP address, browsing software and operating system characteristics of server log.

Khasawneh in 2006 proposed an algorithm that was able to identify individual users and their sessions [28]. His approach also used the IP address, browsing software and operating system attributes of server log for user identification and session construction.

An algorithm for data cleaning, user identification and session identification was proposed by Sunneetha and Krishnamoorthi. The approach was based on the concept of using access the usage pattern of preprocessed data using snow flake schema for easy retrieval [43].

Castellano and his co-researchers [11] worked on filtering out the least visited pages. In addition to filtering techniques he also worked on individual user identification. Wahab and his fellow researchers proposed an algorithm for applying data cleaning techniques on IIS Web Server Logs ranging from the raw log file until before mining process can be performed.

Raju and Satyanarayana described the user sessions identification as a complex process [38] due to the problem of caching. In most of the session identification techniques, 30 minute timeout was taken and transactions made by user with web site are in 30 minutes are grouped as session.

Table 3: Summary of Research Done in Data Preprocesssing

Author	**Source of Log File**	**Preprocessing Technique**	**Algorithm Applied**
Robert Cooley, Bamshad Mobasher and Jaidep Srivastava (2000)	Server Log	▪ User Identification, ▪ Session Identification, ▪ Page View Identification, ▪ Path Completion, ▪ Episode Identification	NA
Pabarskaite (2002)	Server Log	▪ Advance Data Cleaning ▪ Filtering ▪ Data Visualization	NA

Yuan and others (2003)	Server Log	▪ Path Completion ▪ Data Cleaning ▪ User Identification ▪ Session Identification	Proposed
Khasawneh, (2006)	Server Log	▪ User Identification ▪ Session Identification	Proposed
Stermsek, (2007)	Server Log	▪ Data Cleaning ▪ User Identification ▪ Session Identification	NA
Castellano and others, (2007)	Server Log	▪ Data Cleaning ▪ User Sessions Identification ▪ Data Filtering	NA
Wahab (2008)	Server Log	▪ File Reading ▪ Data Cleaning ▪ Data Filtering	Proposed
Raju and Satyanarayana, (2008)	Server Log	▪ Data Merging ▪ Data Cleaning ▪ User Identification ▪ Session Identification	NA
Suneetha and Krishnamoorthi (2009)	Server Log	▪ Data Cleaning ▪ User Identification	NA

Data preprocessing consists of four sub-phases: data cleaning, user identification, session identification, path completion. The usage preprocessing architecture is shown in figure 13.

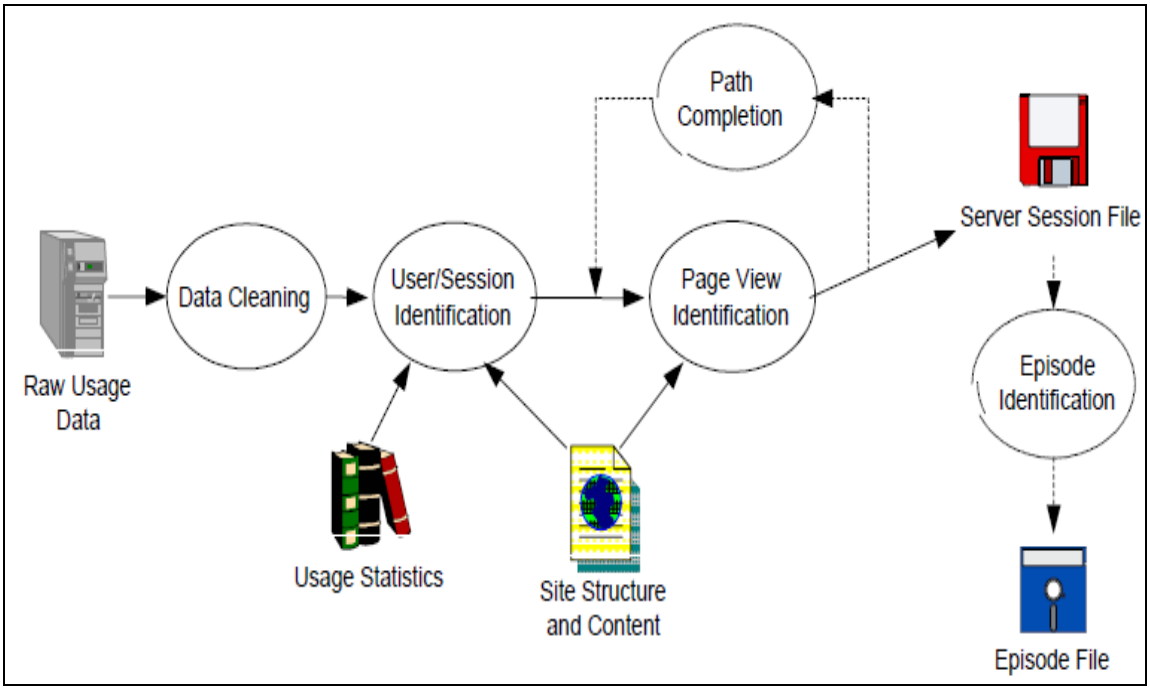


Figure 13: Usage Preprocessing Architecture [16]

a) Data Cleaning: With the important entries, a web log file may consist of certain undesirable rather useless data which has nothing to do with the mining procedure. Therefore, it is imperative to remove those irrelevant entries from the log file. There are three kinds of irrelevant or redundant data needed to clean [13]:

- Accessorial Resources Embedded in HTML File
- Robots' Requests
- Error Requests

Accessorial resources are graphics, scripts or some videos, that may be present in an html page but may not have any relationship with the content of the html page on which they are embedded. They may be part of some advertisement. When a user requests for a particular webpage, these may also get downloaded along with the HTML file and forms several log entries. As discussed earlier the objective of web usage mining is to capture the user's behaviour, therefore, these entries for graphics, images and scripts are useless. Due to this reason the removal of these irrelevant items seems to be essential.

Web robots are software tools that are used to automatically extract contents of a website by following all the hyperlinks from a web page. A search engine uses spiders to

grab all the pages from a web site to update their search indexes. These are also not important from the mining perspective and hence must be removed.

The third type of inaccuracy that has to be removed is the entries related to the error request. When interacting with the web there are certain errors that came around. The errors on web are represented by means of status codes. For example the status code 404 represents the error that the requested resource does not exist. But this type of errors also occupies entry in the log file. Error requests are useless for mining process and should be removed. Error request can be removed from the log file by checking the status of request. The status code less than 200 and greater than 299 are also error codes.

Therefore, it becomes necessary to eliminate these extraneous entries so that the performance of web usage mining process can be improved. Data cleaning is one of the most important steps in the web usage mining. The reason is that accuracy of results of this step affects the results of later phases. Therefore, data cleaning has always remained an important topic of research and attracted much of the researchers' attention. In 2011, T.T. Aye proposed an algorithm which was to perform data cleaning technique on log file [3]. His approach was able to remove the .gif, .jpg and .css entries but robots request were not considered.

b) User Identification: In this step different users who contact the web server and request for some resource on the web are identified. Different methods are suggested for user identification. The simplest one is to assign different user id to different IP address. If the IP address is same, but the agent log shows a difference in the browser's software or operating system, it means that the IP address represents a different user. If both IP address and user agent are same then the referrer URL and site topology must be checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user but with the same IP address. During user identification, problem due to caching may occur. The solution to caching problem is to assign a short expiration time to HTML pages enforcing the browser to retrieve every page from the server.

c) Session Identification: After the user has been identified the pages accessed by each user are divided into the individual session. This is known as session identification. The objective of the session identification step is to come across each user's access pattern

and frequently accessed path. The simplest method for identifying the session is using a timeout. In this method it is assumed that if the time between page requests exceeds a certain limit then it implies that the user is beginning a new session. Usually, a timeout of 30 minutes is taken as a default timeout.

Many researchers have worked on methods or approaches that enable the effective reconstruction of sessions. A simple algorithm was devised by Baoyao Zhou in which an access session is created as a pair of URL and the requested time in a sequence of requests with a timestamp [50]. Smart Miner is a new method devised by Murat Ali and team [6]. The framework was a part of Web Analytics Software in which sessions were constructed by SMART-SRA. The sessions thus produced contains sequential pages accessed from server-side worked in two stages and follows timestamp ordering rule and topology rule. A new method using integer programming was proposed by Robert F.Dell [18]. The proposed method enables the construction of all sessions simultaneously. In this approach each web log was considered as a register, and registers from the same IP address and agent were grouped to form a session. Another algorithm proposed by Junjie Chen and Wei Liu in which data cleaning and session identification is combined [14].

d) Path Completion: There may be chances of having missing pages after constructing transactions because of proxy servers and caching problems. In such a condition it becomes necessary to identify the user's access path, and add the missing paths. The procedure of adding missing page involves checking whether the page request is directly linked to last page. If there exists no link to the last page, then check the history. If the page exists in the recent history, it means back button was used for caching for reaching on that page. In case there are many pages linked to the requested page then the closest page is the source of new request and that page is added to the session.

An optimal algorithm is devised by G.Arumugam and S.Suguna to generate accurate path sequences by using two way hashed structures based access history list to frame a complete path with optimal time [2].

2.9.2 Pattern Discovery

The second phase of web usage mining is pattern discovery which is the key component of the web mining. Pattern discovery utilizes the algorithms and techniques from several

research areas such as data mining, machine learning, statistics and pattern recognition [44]. The methods of pattern discovery are as follows:

a) **Statistical Analysis:** Statistical techniques are the most common method to extract knowledge about visitors of a web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path [42].

b) **Association Rules:** In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks.

c) **Clustering:** Clustering analysis is a technique to group together users or pages with the similar characteristics. In the web usage mining domain, there are two kinds of interesting clusters to be discovered:
   i. **Usage clusters:** Usage clusters are used to find groups of users with similar browsing preference and habit.
   ii. **Page clusters:** Page clusters aims to discover groups of pages that seem to be conceptually related according to the user's perception.

d) **Classification:** Classification maps a data item into one of several predefined classes. In the Web domain, this technique is used to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category.

e) **Sequential Pattern:** It intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes.

f) **Dependency Modelling:** It tries to create a model that represents significant dependencies among various variables found in the Web domain.

A complete idea for the pattern discovery of web usage mining that web site creators must have clear knowledge of user's profile and site intentions and also emphasized information of the approach users will browse web site was suggested [35].

2.9.3 Pattern Analysis

Pattern Analysis is the final stage of the Web usage mining. The goal of pattern analysis is to eliminate the irrelevant rules or patterns from the output of the pattern discovery process. There are two most common approaches for the pattern analysis [42]:

a) SQL query mechanism

b) Construction of multidimensional data cube to perform OLAP operations

The discovery of usage patterns are not helpful unless there were mechanisms and tools that may help the researchers or an analyst to better grab knowledge from those patterns. Therefore besides having techniques for mining usage patterns from web logs it is necessary to develop techniques and tools for enabling the analysis of discovered patterns. These techniques are expected to draw from a number of fields including statistics, graphics and visualization, usability analysis and database querying [33]. Usage analysis of Web access behaviour being a very new area, therefore a very little work is done in this field.

A system for visualizing WWW access patterns was illustrated in which sets of server log entries were used to extract sub sequences of web traversal patterns called web paths [37]. On Line Analytical Processing (OLAP) is emerging as a powerful paradigm for strategic analysis of databases in business settings [33]. Many researchers have demonstrated that the functional and performance needs of OLAP require a new information structures to be designed. This has led to the development of the data cube information model [24] and techniques for its efficient implementation [1] [26] has shown that the analysis needs of web usage data have much in common with those of a data warehouse and hence OLAP techniques are quite applicable.

Chapter 3
Problem Statement

The most important issue in Web usage mining is related to the preprocessing of log file as the raw web log files cannot be used directly for modeling purposes. Therefore, it is needed to preprocess the log file before moving forward in the Web usage mining process.

Preprocessing results also strongly influences the later phases of Web Usage Mining. This makes the preprocessing of server log files a significant step in Web Usage Mining. The data should be preprocessed to improve the efficiency and ease of the mining process. So it is important to perform preprocessing before applying data mining techniques to discover user access patterns from web log. The main task of data preprocessing is to prune noisy and irrelevant data and to reduce data volume for the pattern discovery phase.

Preprocessing enables to translate the data which is collected from server log files into constructive data abstraction. Preprocessing cleans up the data (server log file) to filter out from the data set the automatic requests generated by the web page, which were not specifically requested by the user. In addition to elimination of the irrelevant automatic requests, it is required to remove nonhuman access behaviour (e.g., spiders, crawlers and automatic web bots) from the web log file. The behaviour of the bots differs qualitatively from human behaviour and is not considered interesting from a web usage mining standpoint. There is a third type of inaccuracy that has to be removed from the log file is the entries related to the error request. Sometimes when interacting with the web there are certain error codes that come across. These error codes are in form of numbers. For example the status code 404 represents an error that the requested resource does not exist. The error requests also occupy space in the log file. Error requests are useless from mining perspective and should be removed. Error request can be removed from the log file by checking the status of request. The status code less than 200 and greater than 299 are also error codes.

The algorithms that were proposed earlier in this field were able to handle one or two of these issues. But to have accuracy in results of later phases it is needed that the result should be near to perfection and all of these three discrepancies should be removed.

To solve these issues "**A Modified Algorithm for Data Cleaning of Log File Using File Extensions in Web Usage Mining**" has been proposed in this thesis. The proposed algorithm performs the cleaning of web log file. The algorithm removes irrelevant records such as records with gif, jpg, css and txt as suffixes as these have no importance from the perspective of the web usage mining. Therefore, the main objectives of thesis include:

i. To remove the log entries from the server log file having status code other than 200.

ii. To remove the log entries from the server log file having file extensions as .gif, jpg, .css.

iii. To remove the log entries from the server log file for robot requests.

Chapter 4
Proposed Algorithm

4.1 Introduction

Data preprocesssing converts the raw data into data abstractions that are significantly necessary for the pattern discovery. The log file preprocessing consists of data cleansing, user identification, session identification and path completion. The purpose of data preprocessing on log files is to improve the quality of data and increase the precision of mining results. The significance of preprocessing lies in the accuracy of data cleaning results.

Data cleaning is basically a two step process in which firstly data fields of the log file are extracted and separated and then stored in a relational database table, and in second step actual data cleaning is performed on the database table where the data fields are stored. In data cleaning irrelevant records such as records with gif, jpg, css and txt as suffixes are eliminated. These files have no importance from the web usage mining viewpoint.

The main focus of the purposed solution is on cleaning the raw web log files and inserting the processed data into a relational database so that it is becomes appropriate to apply the mining techniques in the later phases.

Outline of the problem solution:

1. Extract the web log and separate its data fields.
2. Store the separated fields in relational database table.
3. Remove irrelevant data by applying data cleaning algorithm on the table where data of log file is stored after separation.
4. Store the relevant data (or entries) in a relational database table.

4.2 Data Field Extraction

A server log file consists of various data fields (such as IP address, status code, method, path, etc) must be separated out before applying cleaning procedure. This 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 a separator for separating the fields in a single log entry. The algorithm proposed here for data field extraction uses the space character as a separator to separate the fields of the log file.

A portion of the log file used for the experimentation has been shown in the figure 14. The log file used for the experimentation is in CLF (Common Log Format) form which means it consists of ten fields, namely,

- IP address
- User ID
- Hostname
- Timestamp
- GMT offset
- Method
- Path
- Protocol
- Status code
- Number of bytes

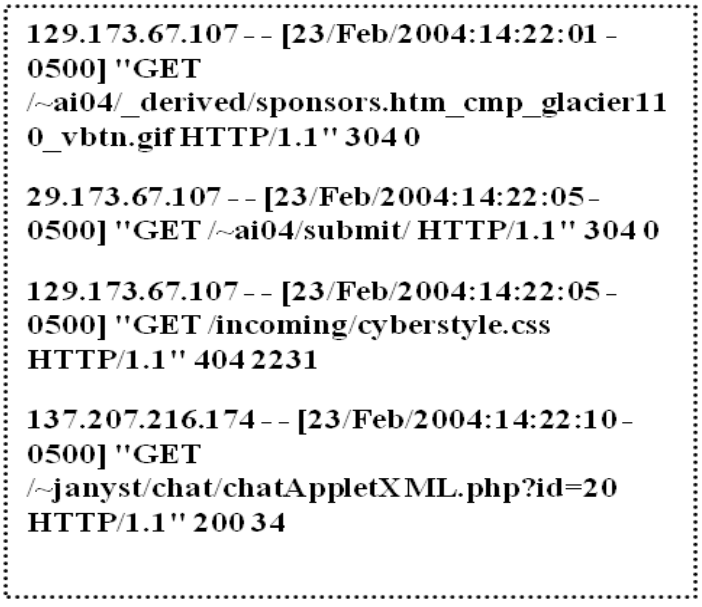


Figure 14: Sample Log File

The implementation of the algorithm is done in Java programming language. It makes use of some of Java's inbuilt classes and methods. The data field extraction is done using methods of String class. It is assumed that space character is acting as the separator. The log

file is read character by character up to the end and then by using the methods of String Tokenizer class the data fields are broken into tokens and saved in an array.

The algorithm for data field extraction is described:

Algorithm 1: Data Field Extraction

Input: Log File
Output: Separated log fields

Step 1: Initialize token: = null /* token is variable that will contain the items read from the log file */

Step 2: Initialize retTokArr: = null /* retTokArr is an array that will contain the items read from the log file after separation */

Step 3: Find location of the log file to be read.

Step 4: Open file for reading.

Step 5: token: = readFile () /* Read items from the log file character by character in the form of string */

Step 6: while (token! = null) /* Run a while loop until all the items are not read from the file*/

{

retTokArr := token.split(“ ”) /* Use space as a delimiter to separate out the fields */

}

Step 7: Close the file

Step 8: End

4.2 Data Storage

The second algorithm describes the storage of field extracted from the log file using the first algorithm. Before data storage we need to create the table named as log table in which each entry from the original log file is stored. The sample SQL query for creating the log table with the column names and data types is shown in figure 15.

```
CREATE TABLE LOGTABLE
(
        IPADDRESS VARCHAR2 (50),
        HOSTNAME VARCHAR2 (50),
        USER_NAME VARCHAR2 (50),
        TIME_STAMP VARCHAR2 (50),
        OFFSET VARCHAR2 (50),
        METHOD VARCHAR2 (50),
        PATH VARCHAR2 (250),
        PROTOCOL VARCHAR2 (50),
        STATUS VARCHAR2 (50),
        BYTES VARCHAR2 (50)
)
```

Figure 14: SQL Query for Table Creation

The algorithm for data storage is given as follows:

Algorithm 2: Data Storage

Input: Result of Algorithm 1

Output: Log table

Step 1: Open a database connection and create a statement object.

Step 2: Create a table to store the log data.

Step 3: Add field names to the log table.

Step 4: Insert the items to the appropriate field column into the log table.

Step 5: Close the database.

4.3 Data Cleaning

The third algorithm shows the data cleaning. This algorithm retains only those data entries in the log file whose status code is 200, method is GET and file type is except from gif, jpg and css, .txt.

The algorithm that removes the gif, jpg, css entries and cleans the web log file is given described as:

Algorithm 3: Data Cleaning

Input: Log table

Output: Summarized log table

Step1: Declare filename, method, ip_address, file_extension, hostname, username, timestamp, offset, protocol, bytes, status_code.

Step 2: Open a database connection.

Step 3: Create an object of PreparedStatement to read each record in log table.

Step 4: For each record read from the log table

a. Read status_code /* the status as extracted from the database.*/

b. Read method /* method as extracted from the database.*/

c. If (status_code = 200 and method = GET)

{

i. Read ip_address, hostname, username, timestamp, offset, protocol, bytes, and path.

ii. Extract file_extension from path.

iii. If file_extension != {.gif, .jpg,.css, .txt}

{

Insert data entries into summarized logtable.

}

iv. Else

{

Remove data entries.

}

}

Step 5: Close connection

Step 6: End

Chapter 5

Experimental Results And Analysis

1.1 Introduction

This section summarizes the result and the analysis of the research work done on the proposed algorithm. The proposed algorithm performs the cleaning of web log file. The algorithm removes irrelevant records such as records with gif, jpg, css, and txt as suffixes as these have no importance from the perspective of the web usage mining. The following table shows extensions of redundant file types which appear in web log data that the proposed algorithm eliminates.

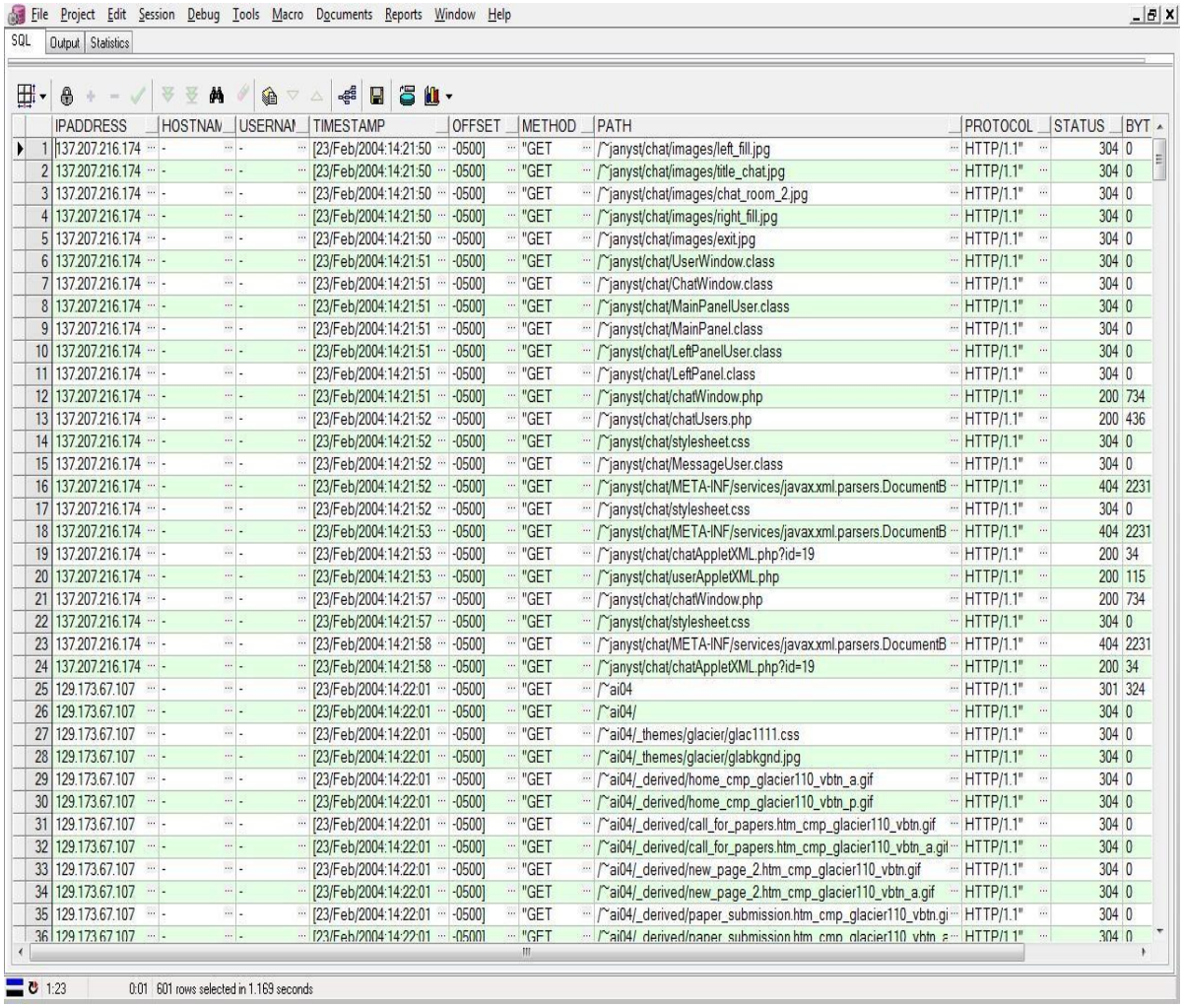
Table 4: Extensions of Redundant File Types which Appear in Web Log Data

Extension	Description of the file	Example of file
GIF	Image file, extended name is "Graphics Interchange Format".	reports.gif
JPEG	Image file, extended name is "Join Photographic Experts Group".	percent.jpg
CSS	HTML cascaded style sheet, template files which allows lots of different pages can have the same heading description, fonts and etc.	news.css
TXT	Text file used for search engines. It contains different information about the site (e. g. keywords) to make it found faster. It is related to robot request.	robots.txt

To validate the effectiveness and efficiency of our methodology mentioned above, an experiment on the sample web server log which is downloaded from the Internet has been made. The initial data source of our experiment is a file whose size is 60KB. By comparing the mining results before cleaning and after cleaning, the results show that the proposed Web page cleaning methods are able to improve the web mining results dramatically.

5.2 Experimental Results

The log file used for the work was of size 60 KB and consists of 601 entries. But after data cleaning the size of the file is reduced to 15.6 KB and only 150 entries are left in the log. The results are shown through the snapshots shown in figure 16 and figure 18. Figure 16 shows the log table formed by separating the data fields of a log file. The figure shows that there are entries for gif, jpg and css files. As discussed in previous sections, these entries are unnecessary from the web usage mining outlook. Therefore, it is essential to remove these entries. The original log table consists of 601 entries.



File Project Edit Session Debug Tools Macro Documents Reports Window Help

SQL | Output | Statistics

	IPADDRESS	HOSTNAM	USERNAM	TIMESTAMP	OFFSET	METHOD	PATH	PROTOCOL	STATUS	BYT
1	137.207.216.174	-	-	[23/Feb/2004:14:21:50	-0500]	"GET	/~janyst/chat/images/left_fill.jpg	HTTP/1.1"	304	0
2	137.207.216.174	-	-	[23/Feb/2004:14:21:50	-0500]	"GET	/~janyst/chat/images/title_chat.jpg	HTTP/1.1"	304	0
3	137.207.216.174	-	-	[23/Feb/2004:14:21:50	-0500]	"GET	/~janyst/chat/images/chat_room_2.jpg	HTTP/1.1"	304	0
4	137.207.216.174	-	-	[23/Feb/2004:14:21:50	-0500]	"GET	/~janyst/chat/images/right_fill.jpg	HTTP/1.1"	304	0
5	137.207.216.174	-	-	[23/Feb/2004:14:21:50	-0500]	"GET	/~janyst/chat/images/exit.jpg	HTTP/1.1"	304	0
6	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/UserWindow.class	HTTP/1.1"	304	0
7	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/ChatWindow.class	HTTP/1.1"	304	0
8	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/MainPanelUser.class	HTTP/1.1"	304	0
9	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/MainPanel.class	HTTP/1.1"	304	0
10	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/LeftPanelUser.class	HTTP/1.1"	304	0
11	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/LeftPanel.class	HTTP/1.1"	304	0
12	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	"GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200	734
13	137.207.216.174	-	-	[23/Feb/2004:14:21:52	-0500]	"GET	/~janyst/chat/chatUsers.php	HTTP/1.1"	200	436
14	137.207.216.174	-	-	[23/Feb/2004:14:21:52	-0500]	"GET	/~janyst/chat/stylesheet.css	HTTP/1.1"	304	0
15	137.207.216.174	-	-	[23/Feb/2004:14:21:52	-0500]	"GET	/~janyst/chat/MessageUser.class	HTTP/1.1"	304	0
16	137.207.216.174	-	-	[23/Feb/2004:14:21:52	-0500]	"GET	/~janyst/chat/META-INF/services/javax.xml.parsers.DocumentB	HTTP/1.1"	404	2231
17	137.207.216.174	-	-	[23/Feb/2004:14:21:52	-0500]	"GET	/~janyst/chat/stylesheet.css	HTTP/1.1"	304	0
18	137.207.216.174	-	-	[23/Feb/2004:14:21:53	-0500]	"GET	/~janyst/chat/META-INF/services/javax.xml.parsers.DocumentB	HTTP/1.1"	404	2231
19	137.207.216.174	-	-	[23/Feb/2004:14:21:53	-0500]	"GET	/~janyst/chat/chatAppletXML.php?id=19	HTTP/1.1"	200	34
20	137.207.216.174	-	-	[23/Feb/2004:14:21:53	-0500]	"GET	/~janyst/chat/userAppletXML.php	HTTP/1.1"	200	115
21	137.207.216.174	-	-	[23/Feb/2004:14:21:57	-0500]	"GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200	734
22	137.207.216.174	-	-	[23/Feb/2004:14:21:57	-0500]	"GET	/~janyst/chat/stylesheet.css	HTTP/1.1"	304	0
23	137.207.216.174	-	-	[23/Feb/2004:14:21:58	-0500]	"GET	/~janyst/chat/META-INF/services/javax.xml.parsers.DocumentB	HTTP/1.1"	404	2231
24	137.207.216.174	-	-	[23/Feb/2004:14:21:58	-0500]	"GET	/~janyst/chat/chatAppletXML.php?id=19	HTTP/1.1"	200	34
25	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04	HTTP/1.1"	301	324
26	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/	HTTP/1.1"	304	0
27	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_themes/glacier/glac1111.css	HTTP/1.1"	304	0
28	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_themes/glacier/glabkgnd.jpg	HTTP/1.1"	304	0
29	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/home_cmp_glacier110_vbtn_a.gif	HTTP/1.1"	304	0
30	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/home_cmp_glacier110_vbtn_p.gif	HTTP/1.1"	304	0
31	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/call_for_papers.htm_cmp_glacier110_vbtn.gif	HTTP/1.1"	304	0
32	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/call_for_papers.htm_cmp_glacier110_vbtn_a.gif	HTTP/1.1"	304	0
33	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/new_page_2.htm_cmp_glacier110_vbtn.gif	HTTP/1.1"	304	0
34	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/new_page_2.htm_cmp_glacier110_vbtn_a.gif	HTTP/1.1"	304	0
35	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/paper_submission.htm_cmp_glacier110_vbtn.gi	HTTP/1.1"	304	0
36	129.173.67.107	-	-	[23/Feb/2004:14:22:01	-0500]	"GET	/~ai04/_derived/paper_submission.htm_cmp_glacier110_vbtn_a	HTTP/1.1"	304	0

1:23 0:01 601 rows selected in 1.169 seconds

Figure 16: Snapshot of Result after Extraction of Data Fields

Figure 17 shows an excerpt of the above snapshot, i.e., figure 16. The highlighted part shows that the original log table consists of entries for images (.jpg or .gif) and templates files (.css). These are the entries which does not have any importance. Therefore, it is necessary to eliminate them.

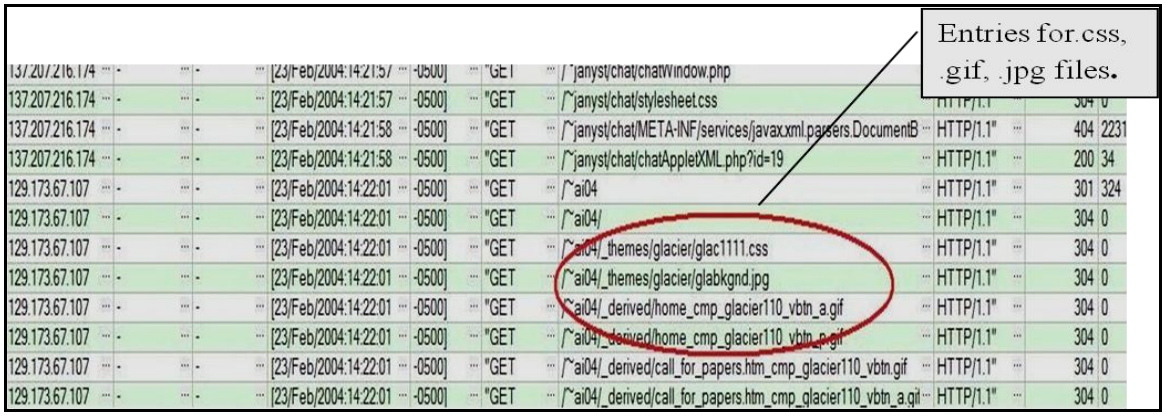
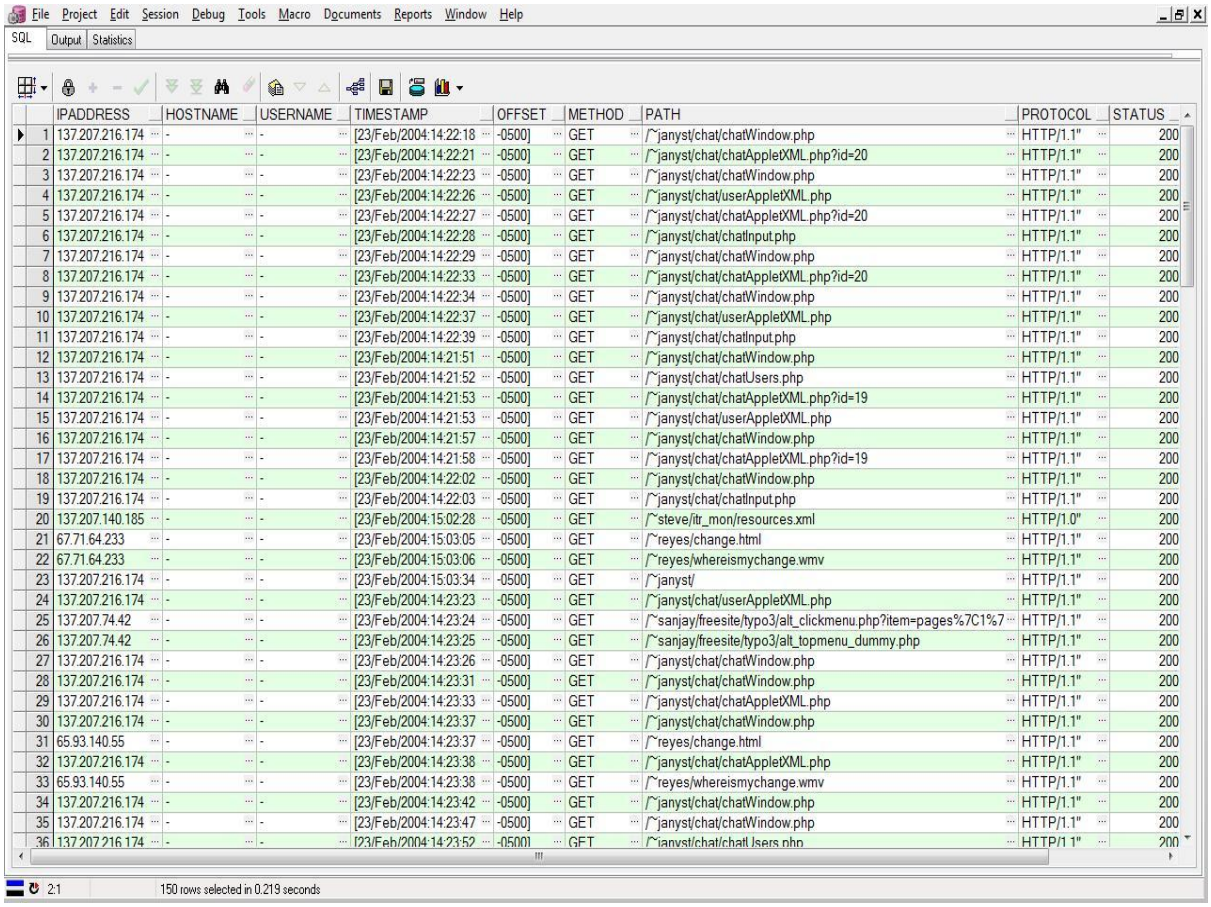


Figure 17: Snapshot of Part of Log Table Showing Entries For .css, .gif, .jpg Files

Figure 18 shows the resulting log table obtained after the application of data cleaning algorithm. In this it can be seen that only 150 entries are left instead of 601 entries. It means out of 601 entries, 451 entries are useless. Only 150 entries are valuable for web usage mining process. The reduction in size of log enhances the speed and accuracy of later phases of web usage mining i.e. pattern discovery and pattern analysis.



	IPADDRESS	HOSTNAME	USERNAME	TIMESTAMP	OFFSET	METHOD	PATH	PROTOCOL	STATUS
1	137.207.216.174	-	-	[23/Feb/2004:14:22:18	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
2	137.207.216.174	-	-	[23/Feb/2004:14:22:21	-0500]	GET	/~janyst/chat/chatAppletXML.php?id=20	HTTP/1.1"	200
3	137.207.216.174	-	-	[23/Feb/2004:14:22:23	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
4	137.207.216.174	-	-	[23/Feb/2004:14:22:26	-0500]	GET	/~janyst/chat/userAppletXML.php	HTTP/1.1"	200
5	137.207.216.174	-	-	[23/Feb/2004:14:22:27	-0500]	GET	/~janyst/chat/chatAppletXML.php?id=20	HTTP/1.1"	200
6	137.207.216.174	-	-	[23/Feb/2004:14:22:28	-0500]	GET	/~janyst/chat/chatInput.php	HTTP/1.1"	200
7	137.207.216.174	-	-	[23/Feb/2004:14:22:29	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
8	137.207.216.174	-	-	[23/Feb/2004:14:22:33	-0500]	GET	/~janyst/chat/chatAppletXML.php?id=20	HTTP/1.1"	200
9	137.207.216.174	-	-	[23/Feb/2004:14:22:34	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
10	137.207.216.174	-	-	[23/Feb/2004:14:22:37	-0500]	GET	/~janyst/chat/userAppletXML.php	HTTP/1.1"	200
11	137.207.216.174	-	-	[23/Feb/2004:14:22:39	-0500]	GET	/~janyst/chat/chatInput.php	HTTP/1.1"	200
12	137.207.216.174	-	-	[23/Feb/2004:14:21:51	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
13	137.207.216.174	-	-	[23/Feb/2004:14:21:52	-0500]	GET	/~janyst/chat/chatUsers.php	HTTP/1.1"	200
14	137.207.216.174	-	-	[23/Feb/2004:14:21:53	-0500]	GET	/~janyst/chat/chatAppletXML.php?id=19	HTTP/1.1"	200
15	137.207.216.174	-	-	[23/Feb/2004:14:21:53	-0500]	GET	/~janyst/chat/userAppletXML.php	HTTP/1.1"	200
16	137.207.216.174	-	-	[23/Feb/2004:14:21:57	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
17	137.207.216.174	-	-	[23/Feb/2004:14:21:58	-0500]	GET	/~janyst/chat/chatAppletXML.php?id=19	HTTP/1.1"	200
18	137.207.216.174	-	-	[23/Feb/2004:14:22:02	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
19	137.207.216.174	-	-	[23/Feb/2004:14:22:03	-0500]	GET	/~janyst/chat/chatInput.php	HTTP/1.1"	200
20	137.207.140.185	-	-	[23/Feb/2004:15:02:28	-0500]	GET	/~steve/itr_mon/resources.xml	HTTP/1.0"	200
21	67.71.64.233	-	-	[23/Feb/2004:15:03:05	-0500]	GET	/~reyes/change.html	HTTP/1.1"	200
22	67.71.64.233	-	-	[23/Feb/2004:15:03:06	-0500]	GET	/~reyes/whereismychange.wmv	HTTP/1.1"	200
23	137.207.216.174	-	-	[23/Feb/2004:15:03:34	-0500]	GET	/~janyst/	HTTP/1.1"	200
24	137.207.216.174	-	-	[23/Feb/2004:14:23:23	-0500]	GET	/~janyst/chat/userAppletXML.php	HTTP/1.1"	200
25	137.207.74.42	-	-	[23/Feb/2004:14:23:24	-0500]	GET	/~sanjay/freesite/typo3/alt_clickmenu.php?item=pages%7C1%7	HTTP/1.1"	200
26	137.207.74.42	-	-	[23/Feb/2004:14:23:25	-0500]	GET	/~sanjay/freesite/typo3/alt_topmenu_dummy.php	HTTP/1.1"	200
27	137.207.216.174	-	-	[23/Feb/2004:14:23:26	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
28	137.207.216.174	-	-	[23/Feb/2004:14:23:31	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
29	137.207.216.174	-	-	[23/Feb/2004:14:23:33	-0500]	GET	/~janyst/chat/chatAppletXML.php	HTTP/1.1"	200
30	137.207.216.174	-	-	[23/Feb/2004:14:23:37	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
31	65.93.140.55	-	-	[23/Feb/2004:14:23:37	-0500]	GET	/~reyes/change.html	HTTP/1.1"	200
32	137.207.216.174	-	-	[23/Feb/2004:14:23:38	-0500]	GET	/~janyst/chat/chatAppletXML.php	HTTP/1.1"	200
33	65.93.140.55	-	-	[23/Feb/2004:14:23:38	-0500]	GET	/~reyes/whereismychange.wmv	HTTP/1.1"	200
34	137.207.216.174	-	-	[23/Feb/2004:14:23:42	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
35	137.207.216.174	-	-	[23/Feb/2004:14:23:47	-0500]	GET	/~janyst/chat/chatWindow.php	HTTP/1.1"	200
36	137.207.216.174	-	-	[23/Feb/2004:14:23:52	-0500]	GET	/~janyst/chat/chatUsers.php	HTTP/1.1"	200

Figure 18: Snapshot of Result after Data Cleaning

5.3 Result Analysis

This section presents an analysis of the results of applying the proposed approach on log file. Table 5 shows the summary of general statistics of the log file. The table contains the values of total hits, visitor hits, spider hits, cached requests, total visitors, number of unique IPs, failed requests. It also presents the number of gif, jpg and css files.

Table 5: Summary Report

Total Hits	601
Visitor Hits	601
Spider Hits	0
Cached Requests	248
Total Visitors	36
Unique IPs	31
Failed Requests (404 not found)	43
Gif	48
Jpg	13
Css	52

The above table makes it obvious that a total of 601 requests were made to the server. There were a total of 36 visitors who contacted the server but only 31 unique IP addresses are identified. The log file consists of 61 entries for image files. It means out of total requests 61 requests are made for the image files. Out of these 61 requests 48 request are made for gif files while only 13 requests are made for jpg files. There were 43 requests that were failed. This is due to the reason that resource was not found. This was the case before the application of cleaning algorithm.

After the experimentation only 150 useful entries were left. Moreover, the number of unique IP addresses decreases to 16. Table 6 shows the number of accesses by the unique IP addresses before and after cleaning.

Table 6: Comparison of Number of Accesses by the Unique IP Addresses Before and After Cleaning

S.No.	IP Address	Before Cleaning	After Cleaning
1.	137.207.216.174	232	95
2.	137.207.72.42	42	13
3.	24.57.61.43	40	23
4.	137.207.76.3	16	1
5.	67.71.64.233	6	2
6.	65.95.153.179	3	2
7.	63.139.177.37	3	2
8.	65.93.140.55	3	2
9.	137.207.248.83	3	2
10.	216.191.209.234	3	2
11.	24.57.255.45	2	1
12.	64.231.21.231	2	1
13.	137.207.140.185	1	1
14.	156.153.255.243	1	1
15.	62.219.75.167	1	1
16.	156.153.255.236	1	1

Figure 19 shows a graphical representation of above comparison. A bar graph is used to represent the comparison. The green bar shows the number of accesses before cleaning while the red bar shows the number of accesses after cleaning. The bar chart is clearly showing that there is a major decrease in the number of request. For example, the IP address 137.207.216.174 had initially made more than 200 requests. But only 95 accesses were useful.

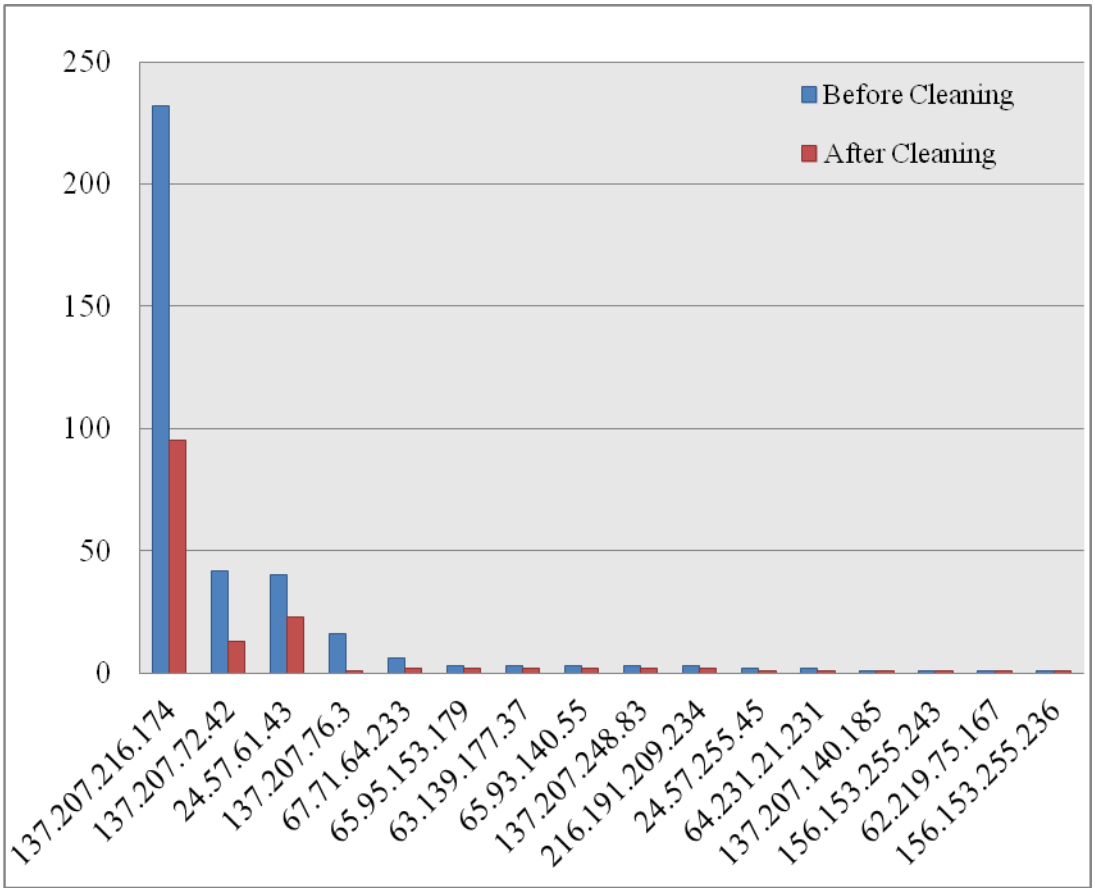


Figure 19: Bar Chart Showing Comparison in Number of Access

Due to this diminution in the number of accesses by the unique IP addresses, the size of the log also decreases. The result shows that the proposed algorithm reduces the size of the log file considerably by removing the unnecessary and irrelevant entries from the file.

Earlier there were 601 entries in the log file, but after cleaning only 150 entries have been left. The original size of the log file before cleaning was 60 KB and after cleaning the size reduces to 15.6 KB. This is shown in the table 7.

Table 7: Comparison in Size Before and After Cleaning

	Size (KB)	No. of Records
Before cleaning	60	601
After cleaning	15.6	150

The change in size and number of records for the log file is graphically represented by means of a bar chart in figure 20.

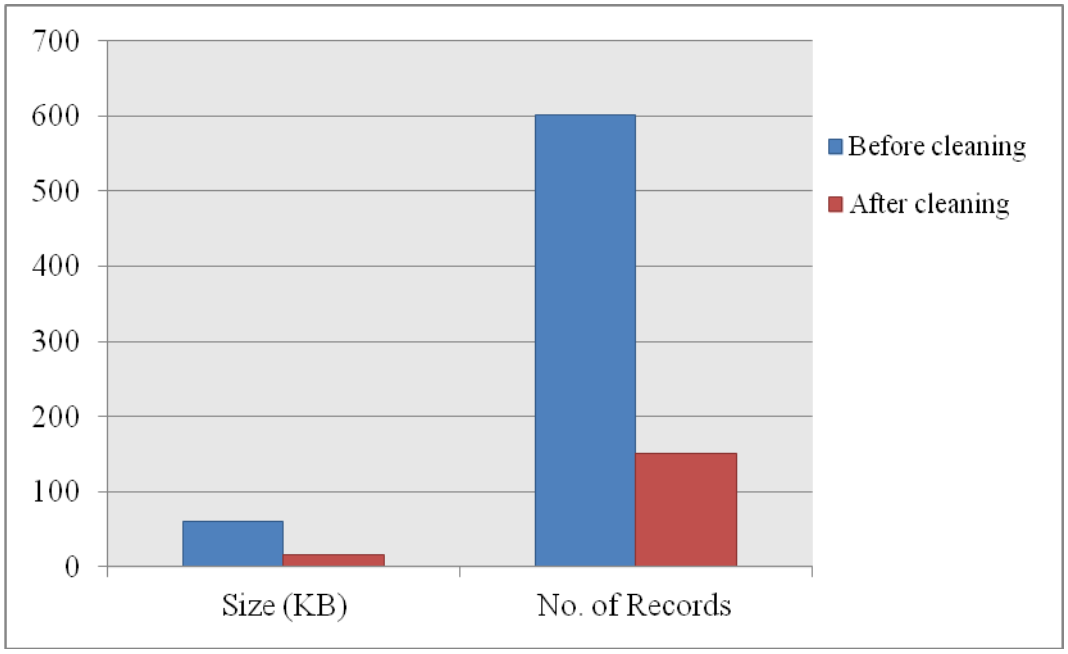


Figure 20: Bar Chart Showing Change in Size and Number of Records

The graph makes it clear that there is a severe change in both the size and number of records after data cleaning. Therefore, from this observation if calculate the percentage amount of decrease in the size of log file, it gives a reduction of 74% (see table 8) which is quite a major value.

Table 8: Results of Data Cleaning

Web Server Log File	Result
Original Size	60 KB
Reduced Size	15.6KB
Percentage in Reduction	**74.00**

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

1. The web cleaning step of data preprocessing is crucial as the result of this step have an impact on the accuracy of results of the later phases.

2. An algorithm for performing the data cleaning technique on server log was proposed.

3. The proposed algorithm was successfully tested on the log files for data cleaning. The results which were obtained after the analysis were satisfactory and contained valuable information about the log files.

4. The proposed approach showed a quite salient reduction in the number of records and in the log files size and hence increases the quality of the available data.

6.2 Future Scope

1. The research and implementation presented in this thesis is in an emerging stage. However, the subjective interpretation of the algorithm is very ingenious and can propose a lot of scope to be extended on to other problem domains.

2. The proposed algorithm was applied on a log file having a common log format. Therefore, the research can be extended to the log file of other formats such as extended common log format which consists of more fields than a common log format.

3. Many problems such as applications of user identification, session identification, and path completion are not discussed.

References

[1] Agrawal S., Agrawal R., Deshpande P.M., Gupta A., Naughton J., Ramakrishna R. and S. Sarawagi, "On the Computation of Multidimensional Aggregates", in *Proceedings of the 22nd VLDB Conference*, pp. 506-521, 1996.

[2] Arumugam G. and Suguna S, "Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs", in *Proceedings of International Conference on Network and Service Security*, pp. 1-6, 2009.

[3] Aye T.T., "Web Log Cleaning For Mining of Web Usage Patterns", in *Proceedings of 3rd International Conference on Computer Research and Development*, pp. 490-494, 2011.

[4] Baraglia R. and Palmerini P., "SUGGEST: A Web Usage Mining System", in *Proceedings International Conference on Information Technology: Coding and Computing*, pp. 282-287, 2002.

[5] Barfourosh A.A., Nezhad H.R.M., Anderson M. L. and Perlis D., "Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition", Department of Computer Science, University of Maryland, Maryland, pp. 1-45, 2002.

[6] Bayir M.A., Toroslu I.H., Cosar A. and Fidan G.,"Smart Miner: A New Framework for Mining Large Scale Web Usage Data", in *Proceedings of the 18th International Conference on World Wide Web*, pp. 161-170, 2009.

[7] Bharat K. and Broder A., "A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines", *in Proceedings of the 7th World-Wide Web Conference*, pp. 379-388, 1998.

[8] Blockeel H. and Kosala R., "Web Mining Research: A Survey", *ACM SIGKDD Explorations*, vol. 2, no. 1, pp. 1-15, June 2000.

[9] Borges J. and Levene M., "Data Mining of User Navigation Patterns", in *Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling*, pp. 31-39, August 1999.

[10] Brin S. and Page L., “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, in *Proceedings of the Seventh International World Wide Web Conference*, pp. 107-117, 1998.

[11] Castellano G., Fanelli A. M. and Torsello M.A., “LODAP: A Log Data Preprocessor for Mining Web Browsing Pattern”, in *Proceedings of the 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, pp. 12-17, February 2007.

[12] Chakrabarti S., Dom B., Gibson D., Kleinberg J., Kumar S., Rajagopalan and Tomkins A., “Mining the Link Structure of the World Wide Web”, *World Wide Web Internet And Web Information Systems*, vol. 32, no. 8, pp. 60-67, 1999.

[13] Chaofeng L., “Research and Development of Data Preprocessing in Web Usage Mining”, in *Proceedings of International Conference on Management Science and Engineering*, pp. 1311-1315, 2006.

[14] Chen J. and Liu W., “Research for Web Usage Mining Model”, in *Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation*, pp. 8-12, 2006.

[15] Codd E.F., “A Relational Model of Data for Large Shared Data Banks”, *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, June 1970.

[16] Cooley R., “Web Usage Mining: Discovery and Application of Interesting Patterns from Web data”, PhD thesis, University of Minnesota, Dept. of Computer Science, May 2000.

[17] Cowie J. and Lehnert W., “Information Extraction”, *Communications of the ACM*, vol. 39, no.1, pp. 80-91, January 1996.

[18] Dell R.F, Roman P.E. and Velasquez J.D., “Web User Session Reconstruction Using Integer Programming”, in *Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 385-388, 2008.

[19] Etzioni O., “The World Wide Web: Quagmire or Gold Mining?”, *Communications of the ACM*, vol. 39, no. 11, pp. 65-68, November 1996.

[20] Fayyad U., Piatetsky-Shapiro G. and Smyth P., “From Data Mining to Knowledge Discovery in Databases”, *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[21] Florescu D., Levy A.Y. and Mendelzon A.O., “Database techniques for the World-Wide-Web: A Survey”, *Newsletter: SIGMOD Record*, vol. 27, no. 3, pp. 59-74, 1998.

[22] Frawley W.J., Piatetsky-Shapiro G. and Matheus C.J., “Knowledge Discovery in Databases: An Overview”, *AI Magazine*, vol. 13, no. 3, pp. 57-70, 1992.

[23] Furnkranz J., “Web Structure Mining-Exploiting the Graph Structure of The World Wide Web”, *OGAI Journal*, vol. 21, no.2, pp. 17-26, 2002.

[24] Gray J., Bosworth A, Layman A. and Pirahesh H., “Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals”, in *Proceedings of IEEE 12th International Conference on Data Engineering*, pp. 152-159, 1996.

[25] Hearst M. A., "Untangling Text Data Mining", in *Proceedings of 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3-10, June 1999.

[26] Harinarayan V., Rajaraman A. and Ullman J.D., “Implementing Data Cubes Efficiently”, in *Proceedings of 1996 ACM-SIGMOD International Conference on Management of Data*, pp. 205-216, 1996.

[27] Jalali M., Mustapha N., Sulaiman N.B. and Mamat, A., "A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems", in *Proceedings of 12th International Conference Information Visualisation*, pp. 302-307, 2008.

[28] Khasawneh N. and Han C.C., “Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining”, in *Proceedings of International Conference on 2006 Web Intelligence*, pp. 3265-328, 2006.

[29] Kleinberg J. M., “Authoritative Sources in a Hyperlinked Environment”, in *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677, 1998.

[30] Konopnicki D. and Shmueli O, “W3QS: A Query System for the World Wide Web”, in *Proceedings of the 21st VLDB Conference*, pp. 54-65, 1995.

[31] Liu B., **"**Web Data Mining**:** Exploring Hyperlinks, Contents, and Usage Data", First Edition, Springer, New York, 2007.

[32] Madria S.K., Bhowmick S.S., Ng W.K., and Lim E.P., "Research Issues in Web data Mining", in *Proceedings of First International Conference Data Warehousing and Knowledge Discovery*, pp. 303-312, 1999.

[33] Menon S.P. and Hegde N.P, "Requisite for Web Usage Mining- A Survey", *International Journal of Computer Science & Informatics*, vol. 2, no.1, pp. 209-215, August 2009.

[34] Merialdo P., Atzeni P. and Mecca G, "Semi Structured and Structured Data in the Web: Going Back and Forth", *Newsletter: ACM SIGMOD Record*, vol. 26, no. 4, pp. 16-23, December 1997.

[35] Nina S.P., Rahman M., Bhuiyan K.I. and Ahmed, K., "Pattern Discovery of Web Usage Mining", in *Proceedings of International Conference on Computer Technology and Development*, pp. 499-503, 2009.

[36] Pabarskaite Z., "Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining", in *Proceedings of 24th International Conference of Information Technology Interfaces*, pp. 109-113, June 2002.

[37] Pitkow J. and. Bharat K.K., "Webviz: A Tool for World Wide Web Access Log Analysis", in *Proceedings of First International WWW Conference*, pp. 451-454, 1994.

[38] Raju G. T. and Satyanarayana P. S., "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", *International Journal of Computer Science and Network Security*, vol. 8, no. 1, January 2008.

[39] Shinde S.K. and Kulkarni U.V., "A New Approach for on Line Recommender System in Web Usage Mining", in *Proceedings of International Conference on Advanced Computer Theory and Engineering*, pp. 973- 977, 2008.

[40] Singh B., Singh H.K., "Web Data Mining Research", in *Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-10, December 2010.

[41] Srivastava G., Sharma K., and Kumar V., "Web Mining: Today and Tomorrow", in *Proceedings of 2011 3rd International Conference on Electronics Computer Technology*, pp. 399-403, April 2011.

[42] Srivastava J., and Cooley R., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM *SIGKDD Explorations*, vol. 1, no. 2, pp. 12-23, January 2000.

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

[44] Wang Y., "Web Mining and Knowledge Discovery of Usage Patterns", *CS748T Project Part I*, vol. 7, no. 1, pp. 1-25, February 2000.

[45] Wu K.L., Yu P. S., and Ballman A., "SpeedTracer: A Web Usage Mining and Analysis Tool", *IBM Systems Journal*, vol. 37, no. 1, pp. 89-105, 1998.

[46] Yu C., and Brandenburg T., "Multimedia Database Applications: Issues and Concerns for Classroom Teaching", *International Journal of Multimedia & Its Applications*, vol. 3, no. 1, February 2011.

[47] Yuan F., Wang L.J., and Yu G., "Study on Data Preprocessing Algorithm in Web Log Mining", in *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, pp. 28-32, November 2003.

[48] Zaiane O.R., Han J., Li Z.N, Chee H., and Chiang J., "MultiMediaMiner: A System Prototype for Multimedia Data Mining", in *Proceedings of International Conference on Management of Data Proceedings of ACM SIGMOD*, pp. 581-583, 1998.

[49] Zhang Q., and Segall R. S., "Web Mining: A Survey of Current Research, Techniques, and Software", *International Journal of Information Technology & Decision Making*, vol. 7, no. 4, pp. 683-720, 2008.

[50] Zhou B., Hui S.C, and Fong A.C.M, "An Effective Approach for Periodic Web Personalization", in *Proceedings of the International Conference on Web Intelligence*, pp. 284-292, 2006.

List of Publications

1. Anand Surbhi, and Aggarwal Rinkle Rani, "Data Mining Types and Techniques: A Survey", in Proceedings of *International Conference on Competitiveness and Innovativeness in Engineering, Management and Information Technology*, pp. 458-471, January 2012.

2. Anand Surbhi, and Aggarwal Rinkle Rani, "Data Mining Types and Techniques: A Survey", *International Journal of Research in IT and Management*, vol. 2, no. 2, pp. 458-471, February 2012.

3. Anand Surbhi, and Aggarwal Rinkle Rani, "A Token String Based Algorithm for Data Extraction in Web Usage Mining", in Proceedings of *International Conference on Advancements in Computing and Communications*, pp. 39-43, February 2012.

4. Anand Surbhi, and Aggarwal Rinkle Rani, "A Token String Based Algorithm for Data Extraction in Web Usage Mining", *International Journal of Computing and Business Research, special issue on computing and communication.* **[Accepted]**

5. Anand Surbhi, and Aggarwal Rinkle Rani, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions in Web Usage Mining", *International Journal of Computer Applications.* **[Communicated]**