WEB MINING

Subject Incharge: Priya Sachdeva, Assistant Professor (CSE)

Web Mining

- Web mining can define as the method of utilizing data mining techniques and algorithms to extract useful information directly from the web, such as Web documents and services, hyperlinks, Web content, and server logs.
- The World Wide Web contains a large amount of data that provides a rich source to data mining. The objective of Web mining is to look for patterns in Web data by collecting and examining data in order to gain insights.

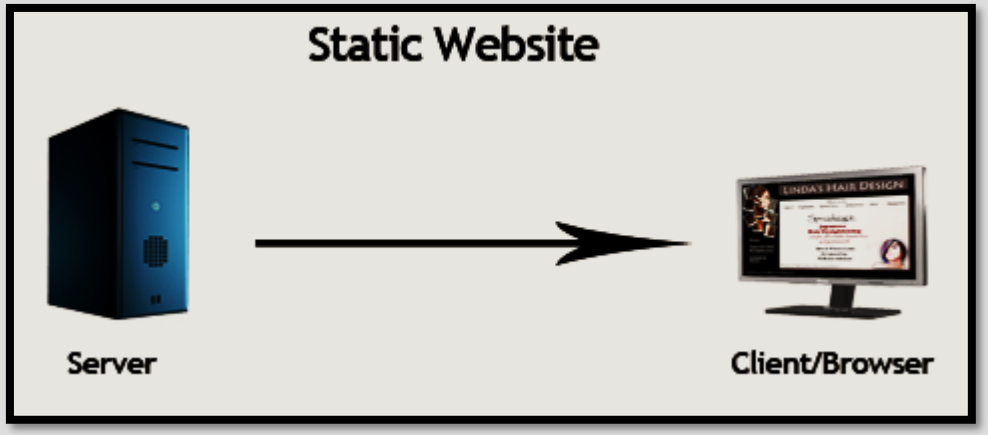
Uses of Web Mining

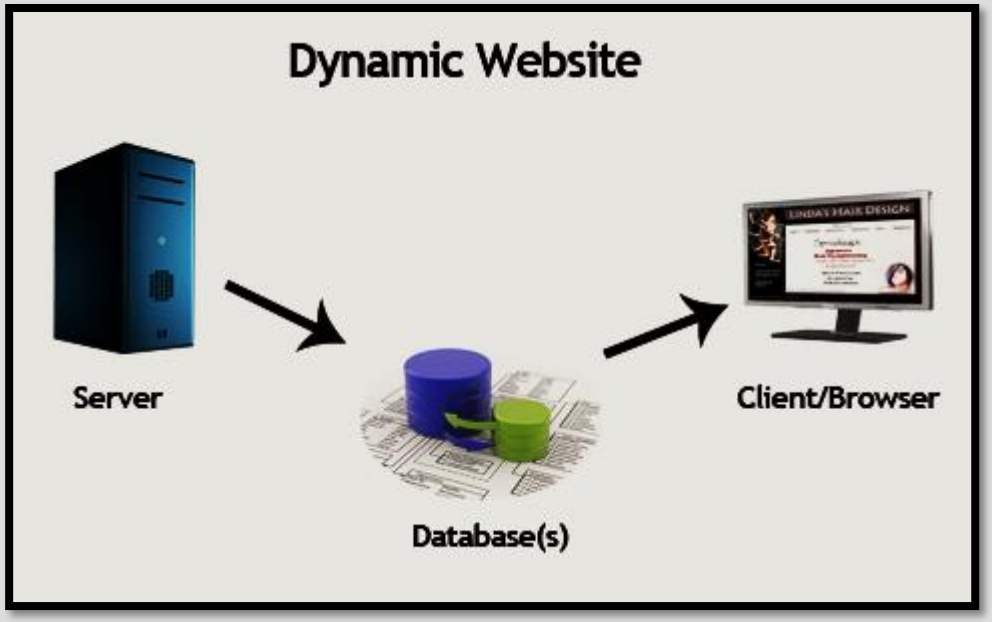
- Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
- It is used for Web Searching e.g., Google, Yahoo etc.
- Web mining is used to predict user behavior.
- Web mining is useful of a particular Website and e-service e.g., landing page optimization. Page optimization refers to all measures that can be taken directly within the website in order to improve its position in the search rankings.

	Data Mining	Web Mining
Definition	Data Mining is the process that attempts to discover pattern and hidden knowledge in large data sets in any system.	Web Mining is the process of data mining techniques to automatically discover and extract information from web documents.
Application	Data Mining is very useful for web page analysis.	Web Mining is very useful for a particular website and e-service.
Target Users	Data scientist and data engineers.	Data scientists along with data analysts.
Access	Data Mining is access data privately.	Web Mining is access data publicly.
Structure	In Data Mining get the information from explicit structure.	In Web Mining get the information from structured, unstructured and semi-structured web pages.
Type	Clustering, classification, regression, prediction, optimization and control.	Web content mining, Web structure mining, Web Usage mining.
Tools	It includes tools like machine learning algorithms.	Special tools for web mining are Scrapy, PageRank and Apache logs.
Skills	It includes approaches for data cleansing, machine learning algorithms.	It includes application level knowledge, data engineering with mathematical modules like statistics and probability.

Website

- Website is a collection of related web pages that may contain text, images, audio and video. The first page of a website is called home page. Each website has specific internet address (URL) that you need to enter in your browser to access a website.
- Website is hosted on one or more servers and can be accessed by visiting its homepage using a computer network.
- A website is managed by its owner that can be an individual, company or an organization.
- A website can be of two types:
  - ✓Static Website
  - ✓Dynamic Website





Static website

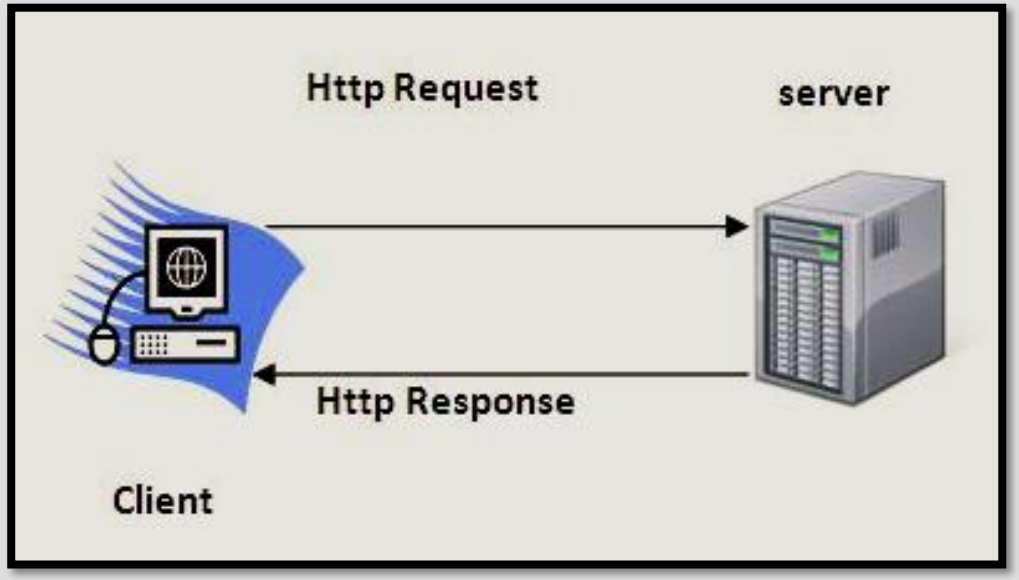
- Static website is the basic type of website that is easy to create. You don't need the knowledge of web programming and database design to create a static website. Its web pages are coded in HTML.
- The codes are fixed for each page so the information contained in the page does not change and it looks like a printed page.

Dynamic website

- Dynamic website is a collection of dynamic web pages whose content changes dynamically. It accesses content from a database or Content Management System (CMS). Therefore, when you alter or update the content of the database, the content of the website is also altered or updated.
- Dynamic website uses client-side scripting or server-side scripting, or both to generate dynamic content.
- Client side scripting generates content at the client computer on the basis of user input. The web browser downloads the web page from the server and processes the code within the page to render information to the user.
- In server side scripting, the software runs on the server and processing is completed in the server then plain pages are sent to the user.

HTTP

- The Hypertext Transfer Protocol (HTTP) is application-level protocol for collaborative, distributed, hypermedia information systems. It is the data communication protocol used to establish communication between client and server.
- HTTP is TCP/IP based communication protocol, which is used to deliver the data like image files, query results, HTML files etc. on the World Wide Web (WWW). It provides the standardized way for computers to communicate with each other.



- It is the protocol that allows web servers and browsers to exchange data over the web.
- It is a request response protocol.
- HTTP is media independent: It specifies that any type of media content can be sent by HTTP as long as both the server and the client can handle the data content.
- HTTP is connectionless: It is a connectionless approach in which HTTP client i.e., a browser initiates the HTTP request and after the request is sent the client disconnects from server and waits for the response.
- HTTP is stateless: It is stateless means each request is considered as the new request. In other words, server doesn't recognize the user by default. The client and server are aware of each other during a current request only. Afterwards, both of them forget each other. Due to the stateless nature of protocol, neither the client nor the server can retain the information about different request across the web pages.

HTTP Requests

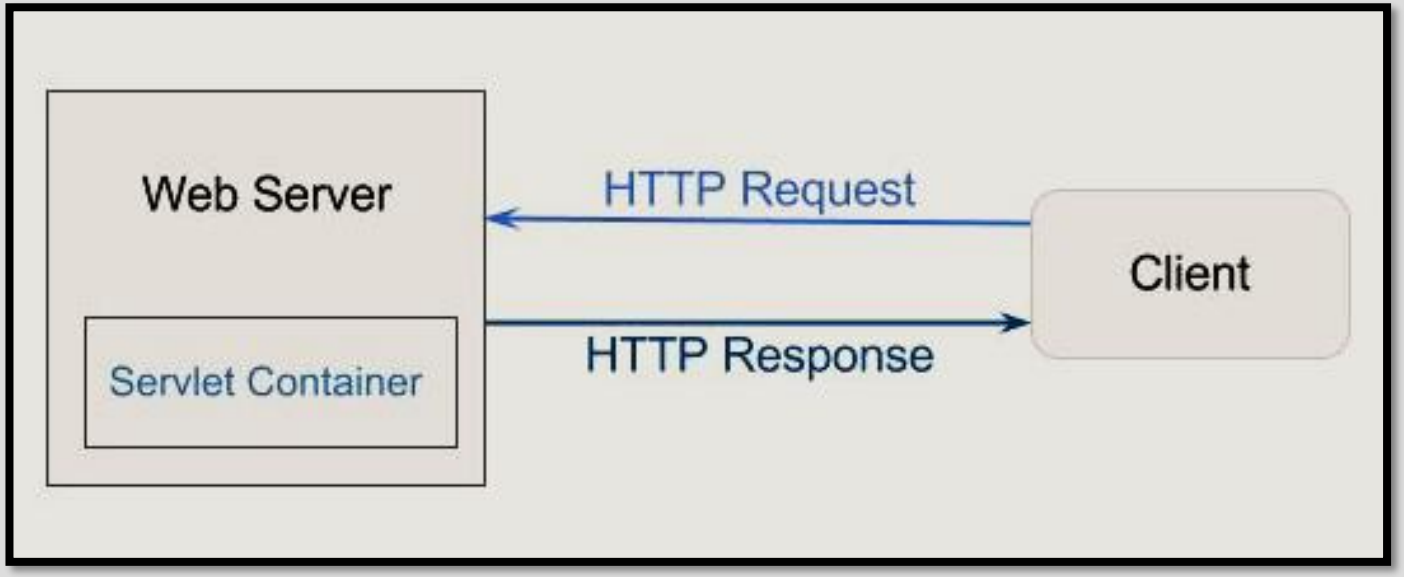
- The request sent by the computer to a web server, contains all sorts of potentially interesting information; it is known as HTTP requests.
- An HTTP request is made by a client, to a named host, which is located on a server. The aim of the request is to access a resource on the server. To make the request, the client uses components of a URL (Uniform Resource Locator), which includes the information needed to access the resource.
- The HTTP request method indicates the method to be performed on the resource identified by the Requested URI (Uniform Resource Identifier). This method is case-sensitive and should be used in uppercase.

The HTTP client sends the request to the server in the form of request message which includes following information:

HTTP Request	Description
GET	Asks to get the resource at the requested URL.
POST	Asks the server to accept the body info attached. It is like GET request with extra info sent with the request.
HEAD	Asks for only the header part of whatever a GET would return. Just like GET but with no body.
TRACE	Asks for the loopback of the request message, for testing or troubleshooting.
PUT	Says to put the enclosed info (the body) at the requested URL.
DELETE	Says to delete the resource at the requested URL.
OPTIONS	Asks for a list of the HTTP methods to which the thing at the request URL can respond.
CONNECT	Establishes a tunnel to the server identified by a given URI.

Servlet Container

It provides the runtime environment for Java applications. The client/user can request only a static Web Pages from the server. If the user wants to read the web pages as per input then the servlet container is used in java.
The servlet container is the part of web server which can be run in a separate process. We can classify the servlet container states in three types:



Server

- Server is a device or a computer program that accepts and responds to the request made by other program, known as client. It is used to manage the network resources and for running the program or software that provides services.
- There are two types of servers:
  - ✓Web Server
  - ✓Application Server
- **Web Server:** Web server contains only web or servlet container. It is a computer where the web content can be stored. In general web server can be used to host the web sites but there also used some other web servers also such as FTP, email, storage, gaming etc. Examples of Web Servers are: Apache Tomcat and Resin.
- **Application Server:** It is a component based product that lies in the middle-tier of a server centric architecture. It provides the middleware services for state maintenance and security, along with persistence and data access. It is a type of server designed to install, operate and host associated services and applications for the IT services, end users and organizations.
  -

Challenges in Web Mining

- **The complexity of web pages:** The site pages don't have a unifying structure. They are extremely complicated as compared to traditional text documents. There are enormous amounts of documents in the digital library of the web. These libraries are not organized according to a specific order.
- **The web is a dynamic data source:** The data on the internet is quickly updated. For example, news, climate, shopping, financial news, sports, and so on.
- **Diversity of client networks:** The client network on the web is quickly expanding. These clients have different interests, backgrounds, and usage purposes. There are over a hundred million workstations that are associated with the internet and still increasing tremendously.
- **Relevancy of data:** It is considered that a specific person is generally concerned about a small portion of the web, while the rest of the segment of the web contains the data that is not familiar to the user and may lead to unwanted results.
- **The web is too broad:** The size of the web is tremendous and rapidly increasing. It appears that the web is too huge for data warehousing and data mining.

Web Mining Process

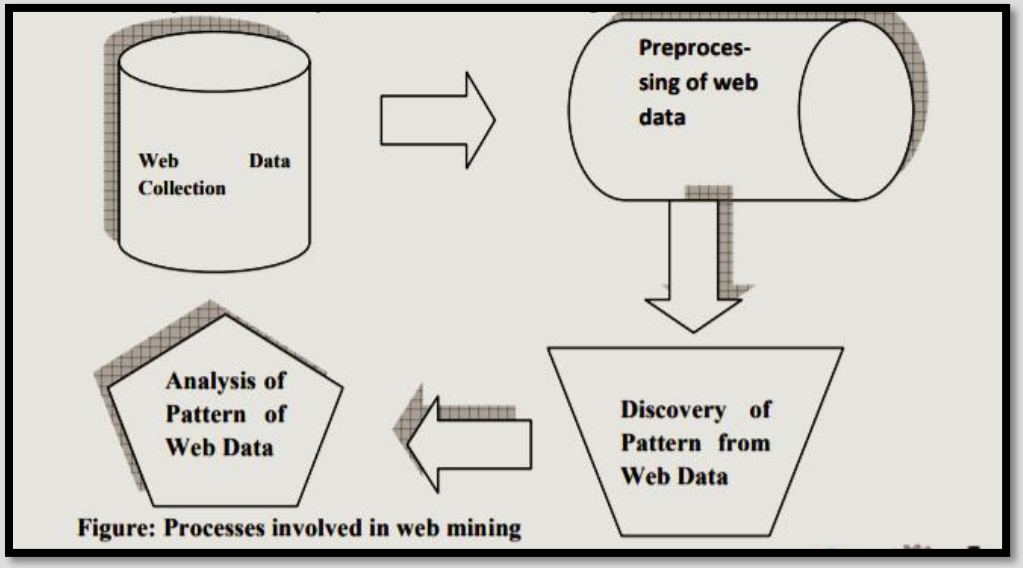
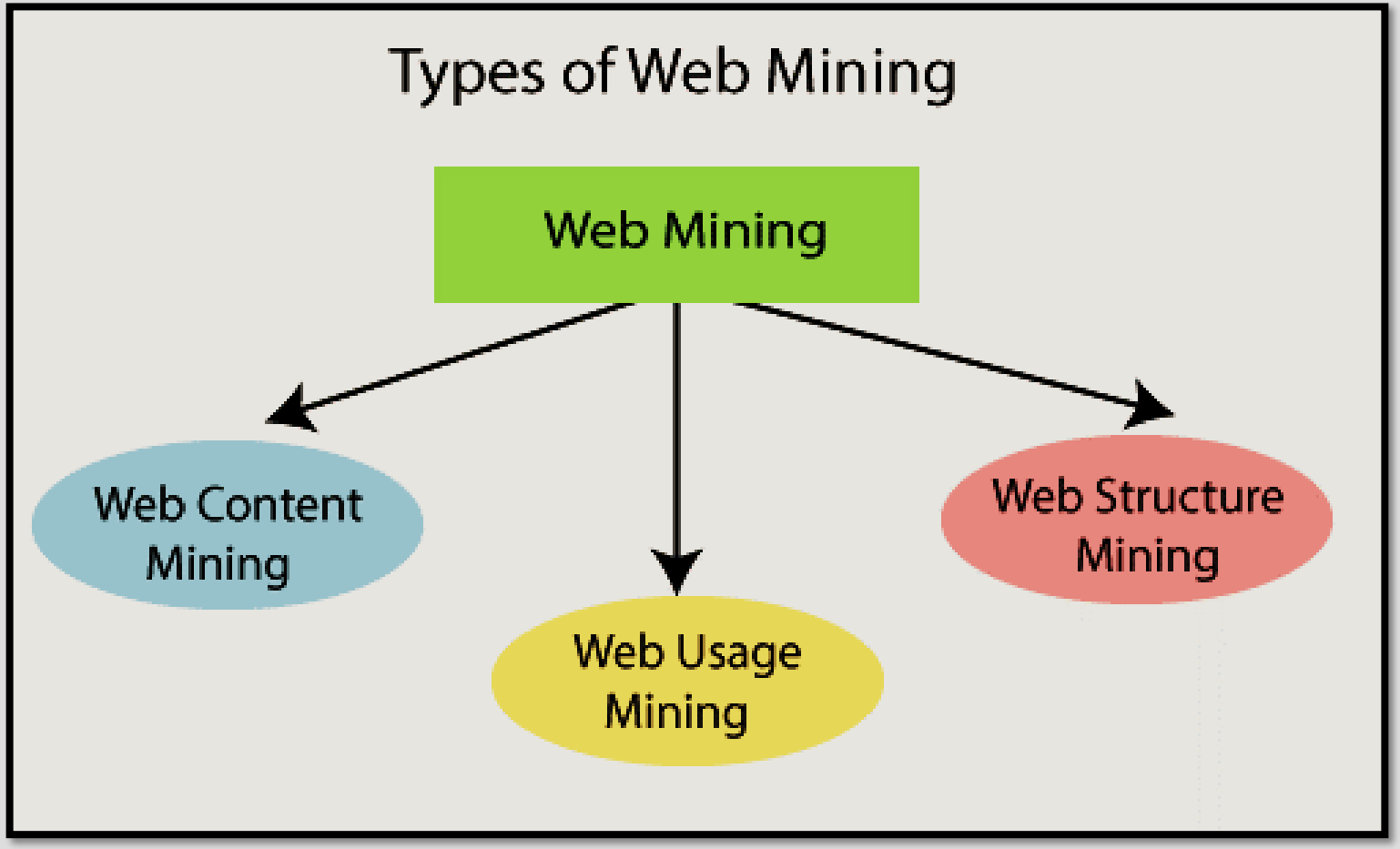


Figure: Processes involved in web mining

Web mining is defined as the "process of studying and discovering web user behavior from web log data." Generally the web data collection is done over a long period of time (one day, one month, one year, etc). Preprocessing of Web Data, Discovery of Pattern from Web Data and Analysis of Pattern of Web Data are being indexed. Pre-processing of web data is the process of transformation of the raw data into a usable data model. Pattern discovery uses several data mining algorithms is used to extract the user patterns. Pattern analysis from web data uncovers useful and interesting user patterns and trends. These steps are normally executed after the web log data is collected.

- **Web Data Collection:** The main task is to get data from Web document, some attentions should be there that sometimes information resources is not limited to online Web documents, but also includes e-mail, electronic documents, news group, or the site through the Web log data and even the formation of transaction data in the database.
- **Preprocessing of Web Data:** The actual data that is to be collected generally have the features that incomplete, redundancy and ambiguity. Mining the knowledge more effectively, pre-processing the data collected is essential. Preprocessing provides accurate, concise data for data mining. Preprocessing of Data, includes data cleaning, user identification, user sessions identification, access path supplement and transaction identification.
- **Discovery of Pattern from Web Data:** Pattern discovery gives us effective, novel, potentially, useful and ultimately understandable information and knowledge using mining algorithm. The main methods are classification analysis, association rule discovery, sequential pattern discovery, clustering analysis, and dependency modeling.
- **Analysis of Pattern of Web Data:** Pattern analysis is mainly concerned with selecting pattern we are interested in from the pattern set found by model pattern discovery algorithm. The main aim is to find out a valuable model, namely, the rules and modes we are interested in and providing graphical user interface using visualization techniques to users.

Types of Web Data Mining



Web Content Mining

- This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. Techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.
- Web content mining can be used to extract useful data, information, knowledge from the web page content. In web content mining, each web page is considered as an individual document. The individual can take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only the layout but also logical structure.
- The primary task of content mining is data extraction, where structured data is extracted from unstructured websites. The objective is to facilitate data aggregation over various web sites by using the extracted structured data. Web content mining can be utilized to distinguish topics on the web. For Example, if any user searches for a specific task on the search engine, then the user will get a list of suggestions.
- Web structure mining helps to find useful knowledge or information pattern from the structure of hyperlinks. Due to heterogeneity and absence of structure in web data, automated discovery of new knowledge pattern can be challenging to some extent.

Web Structure Mining

- This is the process of analyzing the nodes and connection structure of a website through the use of graph theory. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected.
- The web structure mining can be used to find the link structure of hyperlink. It is used to identify that data either link the web pages or direct link network.
- In Web Structure Mining, an individual considers the web as a directed graph, with the web pages being the vertices that are associated with hyperlinks.
- The most important application in this regard is the Google search engine, which estimates the ranking of its outcomes primarily with the PageRank algorithm. It characterizes a page to be exceptionally relevant when frequently connected by other highly related pages.
- Structure and content mining methodologies are usually combined. For example, web structured mining can be beneficial to organizations to regulate the network between two commercial sites.

Web Usage Mining

- This is the process of extracting patterns and information from server logs to gain insight on user activity including where the users are from, how many clicked what item on the site and the types of activities being done on the site.
- Web usage mining is used to extract useful data, information, knowledge from the weblog records, and assists in recognizing the user access patterns for web pages. In Mining, the usage of web resources, the individual is thinking about records of requests of visitors of a website, that are often collected as web server logs. While the content and structure of the collection of web pages follow the intentions of the authors of the pages, the individual requests demonstrate how the consumers see these pages. Web usage mining may disclose relationships that were not proposed by the creator of the pages.
- Methods to identify and analyze the web usage patterns are :

  i) Session and visitor analysis: The analysis of preprocessed data can be performed in session analysis ,which includes the record of visitors, days, sessions etc. This information can be used to analyze the behavior of visitors. Report is generated after this analysis, which contains the details of frequently visited web pages, common entry and exit.

  ii) OLAP (Online Analytical Processing) performs Multidimensional analysis of complex data. OLAP can be performed on different parts of log related data in a certain interval of time. The OLAP tool can be used to derive the important business intelligence metrics.


Thank
you