# UNIT 4:

# Web data mining:

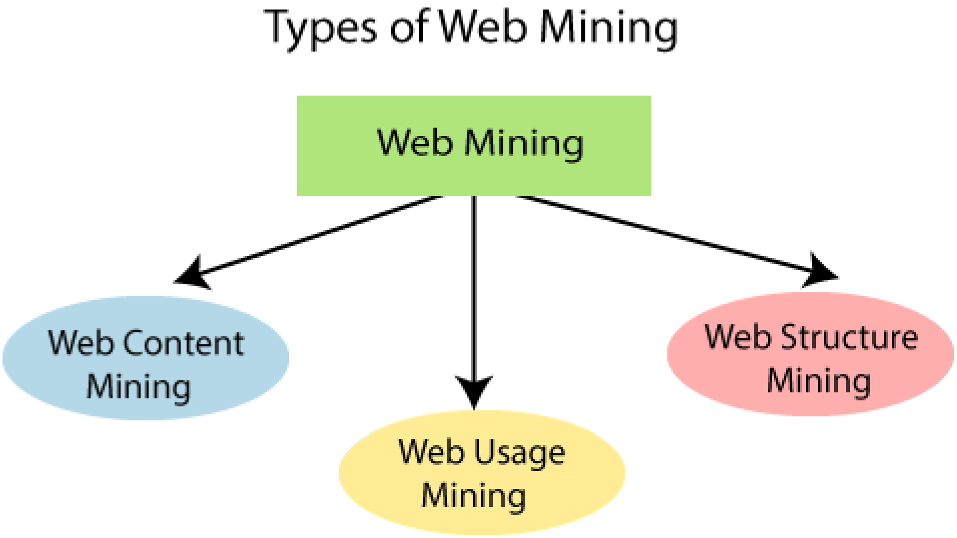
## Data Mining- World Wide Web

Over the last few years, the **World Wide Web** has become a significant source of information and simultaneously a popular platform for business. 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.

## What is Web Mining?

Web mining can widely be seen as the application of adapted data mining techniques to the web, whereas data mining is defined as the application of the algorithm to discover patterns on mostly structured data embedded into a **knowledge discovery process**. Web mining has a distinctive property to provide a set of various data types. The web has multiple aspects that yield different approaches for the mining process, such as web pages consist of text, web pages are linked via hyperlinks, and user activity can be monitored via web server logs. These three features lead to the differentiation between the three areas are web content mining, web structure mining, web usage mining.

## There are three types of data mining:



### 1. Web Content Mining:

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.

### 2. Web Structured Mining:

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.

### 3. Web Usage Mining:

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.

Some of the methods to identify and analyze the web usage patterns are given below:

#### I. Session and visitor analysis:

The analysis of preprocessed data can be accomplished in session analysis, which incorporates the guest records, days, time, sessions, etc. This data can be utilized to analyze the visitor's behavior.

The document is created after this analysis, which contains the details of repeatedly visited web pages, common entry, and exit.

#### II. OLAP (Online Analytical Processing):

OLAP accomplishes a multidimensional analysis of advanced data.

OLAP can be accomplished on various parts of log related data in a specific period.

OLAP tools can be used to infer important business intelligence metrics

## Challenges in Web Mining:

The web pretends incredible challenges for resources, and knowledge discovery based on the following observations:

- **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.

## Mining the Web's Link Structures to recognize Authoritative Web Pages:

The web comprises of pages as well as hyperlinks indicating from one to another page. When a creator of a Web page creates a hyperlink showing another Web page, this can be considered as the creator's authorization of the other page. The unified authorization of a given page by various creators on the web may indicate the significance of the page and may naturally prompt the discovery of authoritative web pages. The web linkage data provide rich data about the relevance, the quality, and structure of the web's content, and thus is a rich source of web mining.

## Application of Web Mining:

Web mining has an extensive application because of various uses of the web. The list of some applications of web mining is given below.

- Marketing and conversion tool
- Data analysis on website and application accomplishment.

- Audience behavior analysis
- Advertising and campaign accomplishment analysis.
- Testing and analysis of a site.

## Comparison Between Data mining and Web mining:

| Points | 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 access data privately. | Web Mining 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. |
| Problem Type | Clustering, classification, regression, prediction, optimization and control. | Web content mining, Web structure 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. Statistics and probability. | It includes application level knowledge, data engineering with mathematical modules like statistics and probability. |

**Web Terminology and Characteristics**

If you are someone who has to deal with computer and internet on daily basis, then you should have good grasp on computer basics knowledge and web terminology as well as network terminology. However, I will only focus on **web terminology** in this article.

Complete tutorial is available here:

- **Computer Basics tutorial**
- **Web Terminology tutorial**

If you do not have any idea about these and you are an absolute beginner, I recommend you to **go through the above tutorials. It's completely free**.

In case you already possess some knowledge then my article will help you access your basics on these topics.

I will point out the keyword related to *web terminologies* and *network terminologies* and would request you to give a thought on it and see if you can explain it to yourself with complete satisfaction.

Here goes the important ***list of web terminology***:

1. WWW or World Wide Web
2. Internet
3. Online & Offline
4. Internet Service Provider or ISP
5. Website
6. Webpage
7. Home Page or Index page
8. Static and Dynamic website
9. Web Browser
10. Web Server
11. URL
12. Domain name
13. DNS
14. IP address
15. Firewall
16. Cache
17. FTP
18. HTTP
19. HTML
20. Login
21. Logout
22. Session

Given above are the *list of web terminologies* that you will encounter on daily basis.

Having a good knowledge about the above topics will be beneficial for you.

---

All the *web terminology definitions and explanations* given below are taken from our ***web terminology tutorial*** whose link is provided at the beginning of this article.

**WWW or World Wide Web**

Full form of *WWW* is *World Wide Web*.

WWW is the system consisting of interlinked hypertext documents that can be accessed on the internet.

*World Wide Web is a collection of documents or web pages which are connected to multiple document or web pages through hypertext links. These documents are accessible over internet and anyone can search for information by navigating from one document to the other documents easily.*

## Internet

Internet is popularly known as *network of networks*.

Internet helps any computer system/mobile to connect with any other computer system globally using TCP/IP protocol. TCP/IP protocol is also known as Internet protocol.

Internet identifies each system in the network through a unique address known as IP address. Each computer system has a unique IP to distinguish from other computer on the network just like voter id of human beings.

## Online & Offline

When you are connected to the internet with your computer, laptop or mobile device you are said to be **online**.

Once your device or system gets disconnected with internet, you are said to be **offline**.

Suppose you want to upload your photo on facebook.

When you are online (connected to internet) photo gets easily uploaded and you get lots of likes and comments.

But if you are not connected to internet (offline), your photo doesn't get uploaded and you get message - 'You seem to be offline. We will try uploading your photo once you go online'.

## Internet Service Provider or ISP

Internet Service Provider is full form of ISP.

ISP is company or organization that provides access to internet services to an individual or family or company or organization using dial-up or other means of data telecommunications.

ISP provide you an Internet account for a monthly or yearly fee, using which you can manage your account.

It also provides other services such as website hosting and building.

## URL

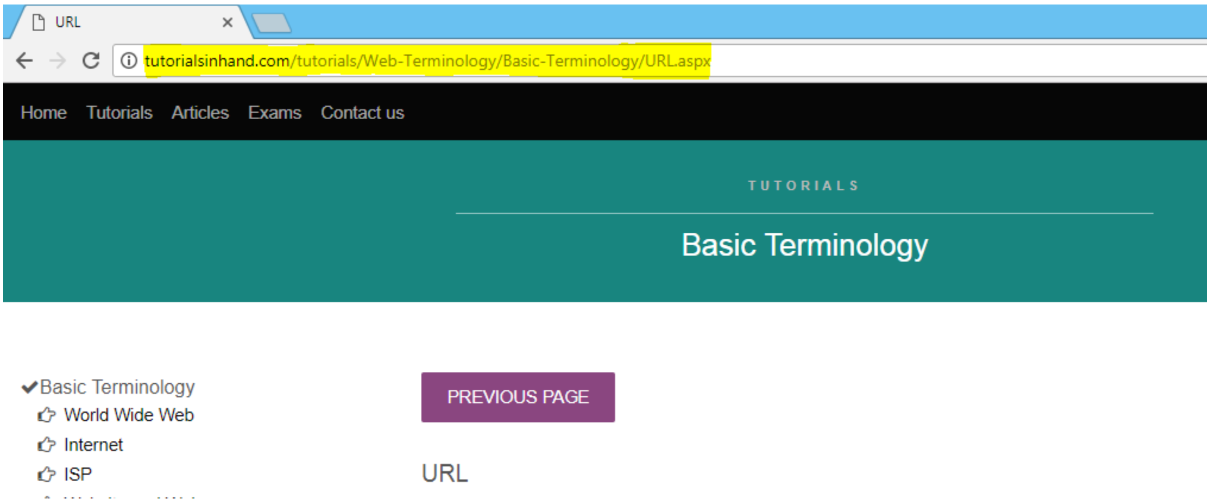
URL stands for ***Uniform Resource Locator***.

It is also known as ***URI*** or ***Uniform Resource Identifier***.

To visit any website, you need to type its URL or URI on web browser.

Suppose you need to visit, Google so you need to type its URL - www.google.com

See the *example of URL* highlighted in yellow in the below image of this website:



## Webpage and Website

There are number of html documents present on World Wide Web. These html documents contain lot of information which can be accessed using URL via web browser.

These *html* documents are referred to as ***Web Pages***.

A web page may consist of texts, images, audio, video, graphics, hyperlinks, etc.

Web pages are placed on the server.

Collection of interlinked web pages with related information is referred to as **website**.

*The page you are currently on is a* ***webpage****.*

*All the pages or webpages of tutorialsinhand domain combined together is a e-learning* ***website****.*

## Home Page or Index page

Every website has a landing page.

It is the page where a user is redirected to first when user visits the website. The landing page of a Website is commonly referred to as **Home Page** or **index page**.

For example,

When you visit amazon.com or flipkart.com, you are first redirected to their home page.

From there you can search different products based on categories, signup or login to their website, sell product, purchase product, etc.

## Signup, Login and Logout

Consider visiting an online banking website for *example*.

Given below are the activities you need to perform in sequence to access an online banking:

- Create account for online banking on bank's website by **signup**.
- After signup, you need to **login** to the website to access restricted products (online fund transfer, book fixed deposit, apply for loan, etc.) of the banking website with your **username** and **password** provided during signup.
- Once you have completed your tasks on website, you need to **logout**.

**Signup** is a *one time activity*. Once you have signup with the website, your account is created with unique id also known as ***customer id***. During signup, you may be asked to provide certain informations related to you like your name, username, password, email, etc. The information required varies from website to website.

You can **login** to the website any number of time on any day using your account details provided during signup. Once you login, your ***session*** begins with the website. All your activities like transfering fund, booking FD, etc. will be tracked under your account with timestamp. This ***enhances the security*** as no one else except you will be able to access your account and perform any activity. You will have complete control of the activities.

To make sure there is no unauthorised access to your account, you need to end the session everytime when you are done with your task on website. To end session you need to ***logout*** of the website.

---

**Static and Dynamic website**

There are two different types of website:

Static website → *Static Website* displays same content or information to all the visitors. Static websites are not interactive in nature as their content remains same irrespective of the number of times you visit it.

For example, Consider this web page of our website that you are currently reading. No matter how many times you or someone else visits this page from any device, the content of this page remains same until updated by admin. It doesn't change from user to user. Same is the case with different pages of this tutorial. So this makes it static in nature.

Dynamic website → *Dynamic website* is the one which displays content created on go by considering the information entered by the user. They do not show the same information every time you visit the page or refresh it. Dynamic websites are *interactive in nature*.

For example, think of a Facebook page. You do not see the same post every time you login. Every time you get fresh posts from your friend or page you follow as they update it from their end. As soon as you like or comment on any post it is visible to the world on press of the enter button. At the same time Facebook, doesn't show same profile information for every individual.

If you go to online exam section of this website, the questions would be different for different users or on every visit. This can be also viewed as dynamic nature of the website.

---

**Web Browser and Web Server**

A *web browser* helps us send request to the ***server***.

A web browser also helps receive response as HTML document from the server, converts them to a form that user can read and finally displays them on computer screen.

Popular web browsers used around the globe are Internet Explorer, Google, Firefox, Yahoo, Bing, Safari, UC Browser, etc.

A *web server* receives the request from the user with help of a browser and then it process the request, prepares the necessary response and sends it back to the browser.

Example on working of Browser and Server

Suppose your results for B.Tech is published by your university.

Now you go to browser (say google), open university url where result is available, enter your roll number and press enter. So you have send the ***request*** from **browser** to **server** asking to send back your result as response.

**Server** contains lots of data. Assume it has result of B.Tech students from various streams, MCA students, BCA students and so on.

Now servers reads your request data, extracts your requirement (roll number) and then finds information related to that extracted data(roll number), and sends back the prepared information as response back to you. And you get to see your result.

## Domain name

*Domain Name* is the way to identify and locate computers connected to the internet.
Two websites cannot have same domain name along with top level domain.
For example, consider our website tutorialsinhand.com

- tutorialsinhand is the **domain name** of this website and **.com** is the top level domain.
- google is the domain name and .com is the top level domain.

## IP address

Full form of *IP* is **Internet Protocol**.
*IP Address* is a unique logical address provided to each computer system on the internet network.

To communicate, send files, send emails, share informations, etc with other systems it is necessary to know where that computer is. IP address helps identify the different systems uniquely.
IP address is an identifier for a particular computer on a particular network.

There are *two types of IP address*:

- **IPv4**: Example is 190.167.48.160
- **IPv6**: Example is 2003:0eb8:75b3:0000:0000:8c2d:0371:7434.

## Firewall

*Firewall* is a kind of security device for computers accessing informations via internet.
Firewall protects the computer and network by restricting the access of outsiders or intruders. It also sets up the criteria that must be met before access to the network or system is allowed to anyone.
Firewall is hardware or software or both that helps protect your system connected on the network from untrusted sites that may contain viruses or other malwares.

## Cache

Cache stores data of the recently or frequently visited websites.
Cache helps to speed up the serving of the web pages faster as the stored data is not required to be fetched from server again which is time consuming task.
Browser cache is used for purposes to store data of the frequently visited websites.

Many ad serving websites use the cache to find out the activity or searches that you do online and then serve ad according to your recent activities. That is why you start seeing ad related to footwear on every website you visit after you have searched anything related to footwear recently from your browser.

## FTP

FTP is an abbreviation for ***File Transfer Protocol***.
FTP is a network protocol used to transfer data from one computer to another through a network.
FTP helps in exchanging and manipulating files over any TCP-based computer network. A FTP client may connect to a FTP server to manipulate files on that server.

## HTTP

HTTP is an abbreviation for ***Hypertext Transfer Protocol***.
HTTP is a request / response standard between a client and a server

HTTP is a communication protocol that helps in transfer of information on the internet and the WWW or World Wide Web. Original purpose od HTTP was to provide a protocol to publish and retrieve hypertext pages over the internet.

Hypertext pages are specially coded using HTML or hypertext markup language. HTML pages may contain text, sound, animations, images, or link to another hypertext pages. When user clicks on any hyperlink the client program on the computer uses HTTP to contact server and ask the server to provide response based on clients request. Server responds back after processing the request over HTTP.

## HTML

HTML stands for ***Hyper Text Markup Language***.
HTML was the first language to be used to design the web pages. Those web pages were static in nature.
HTML designed web page can contain texts, images, audio, videos, etc.
HTML along with CSS can be used to design attractive websites. You can view HTML as a plain design on a white paper whereas CSS is a paint that can fill up the design with beautiful colors.

| Servlet Terminology | Description |
|---|---|
| Website: static vs dynamic | It is a collection of related web pages that may contain text, images, audio and video. |
| HTTP | It is the data communication protocol used to establish communication between client and server. |
| HTTP Requests | It is the request send by the computer to a web server that contains all sorts of potentially interesting information. |
| Get vs Post | It gives the difference between GET and POST request. |
| Container | It is used in java for dynamically generating the web pages on the server side. |
| Server: Web vs Application | It is used to manage the network resources and for running the program or software that provides services. |
| Content Type | It is HTTP header that provides the description about what are you sending to the browser. |

## Locality and Hierarchy in the web

Most social structures tend to organize themselves as hierarchies. The web shows a strong hierarchical structure.

□ Web pages can be classified into several types :

1 Home page or the head page : represents an entry point for the web site of an enterprise

2 Index page : assists the user to navigate through the enterprise’s web site

3 Reference page : provides some basic information that is used by a number of pages . For ex., link to a page that provides enterprise’s privac policy

4 Content page : provides content and are often the leaf nodes of a tree

**Web Content Mining –**

Web Content Mining can be used for the mining of useful data, information, and knowledge from web page content. Web content mining performs scanning and mining of the text, images, and group of web pages according to the content of the input, by displaying the list in search engines.

There are two approaches that are used for Web Content Mining

- **(i) Agent-based approach :** This approach involves intelligent systems. It usually relies on autonomous agents, that can identify websites that are relevant.
- **(ii) Data-based approach :** Data-Based approach is used to organize semi-structured data present on the internet into structured data.

Web content mining is referred to as text mining. Content mining is the browsing and mining of text, images, and graphs of a Web page to decide the relevance of the content to the search query.

This browsing is done after the clustering of web pages through structure mining and supports the results depending upon the method of relevance to the suggested query.

With a large amount of data that is available on the World Wide Web, content mining supports the results lists to search engines in order of largest applicability to the keywords in the query.

It can be defined as the phase of extracting essential data from standard language text. Some data that it can generate via text messages, files, emails, documents are written in common language text. Text mining can draw beneficial insights or patterns from such data.

Text mining is an automatic procedure that facilitates natural language processing to derive valuable insights from unstructured text. By changing data into information that devices can learn, text mining automates the phase of classifying texts by sentiment, subjects, and intent.

Text mining is directed toward specific data supported by the user search data in search engines. This enables the browsing of the entire Web to fetch the cluster content triggering the scanning of definite web pages within those clusters.

The results are pages transmitted to the search engines through the largest level of applicability to the lowest. Though the search engines can support connection to Web pages by the hundreds about the search content, this kind of web mining allows the reduction of irrelevant data. Web text mining is efficient when used in a content database dealing with definite subjects.

For instance, online universities need a library system to recall articles related to their frequent areas of study. This definite content database allows to pull only the data within those subjects, supporting the most specific outcomes of search queries in search engines.

This allowance of only the most relevant data being supported gives a larger quality of results. This increase in productivity is direct to the need for content mining of text and visuals. The need for this type of data mining is to gather, classify, organize and support the best possible data accessible on the WWW to the user requesting the data.

This tool is imperative to browsing the several HTML files, images, and text supported on Web pages. The resulting data is supported by the search engines in order of relevance giving higher productive results of every search.

**Web content mining algorithm:**

This deals with discovering useful information from the web □

The algorithm proposed is called Dual Iterative Pattern Relation Extraction (DIPRE). It works as follows:

1 Sample: Start with a sample provided by the user

2 Occurrences: Find occurrences of tuples starting with those in S. Once tuples are found, the context of every occurrence is save. Let these be O. O→S

3 Pattern: Generate patterns based on the set of occurrences O. This requires generating patterns with similar contexts. P →O

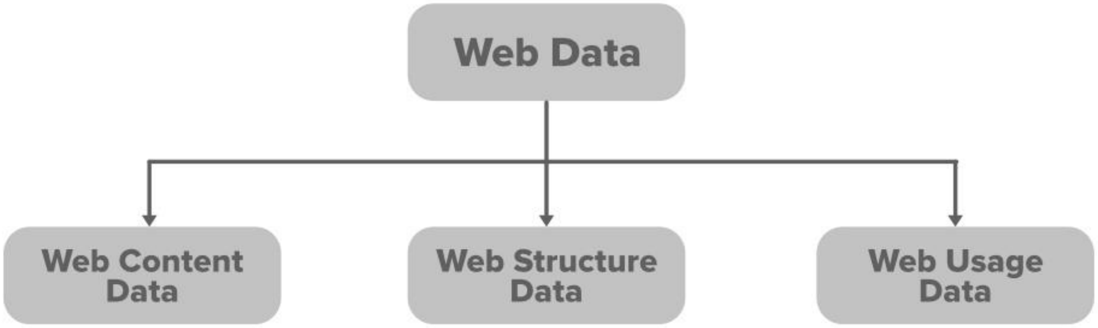
4 Match patterns: The web is now searched for the patterns

5 Stop if enough matches are found. Else, go to Step 2

## Web Usage Mining

Web usage mining, a subset of Data Mining, is basically the extraction of various types of interesting data that is readily available and accessible in the ocean of huge web pages, Internet- or formally known as World Wide Web (WWW). Being one of the applications of data mining technique, it has helped to analyze user activities on different web pages and track them over a period of time. Basically, Web Usage Mining can be divided into 2 major subcategories based on web usage data.

**There are 3 main types of web data:**



**1. Web Content Data:** The common forms of web content data are HTML, web pages, images audio-video, etc. The main being the HTML format. Though it may differ from browser to browser the common basic layout/structure would be the same everywhere. Since it's the most popular in web content data. XML and dynamic server pages like JSP, PHP, etc. are also various forms of web content data.

**2. Web Structure Data:** On a web page, there is content arranged according to HTML tags (which are known as intrapage structure information). The web pages usually have hyperlinks that connect the main webpage to the sub-web pages. This is called Inter-page structure information. So basically relationship/links describing the connection between webpages is web structure data.

**3. Web Usage Data:** The main source of data here is-Web Server and Application Server. It involves log data which is collected by the main above two mentioned sources. Log files are created when a user/customer interacts with a web page. The data in this type can be mainly categorized into three types based on the source it comes from:

- Server-side
- Client-side
- Proxy side.

There are other additional data sources also which include cookies, demographics, etc.

**Types of Web Usage Mining based upon the Usage Data:**

**1. Web Server Data:** The web server data generally includes the IP address, browser logs, proxy server logs, user profiles, etc. The user logs are being collected by the web server data.

**2. Application Server Data:** An added feature on the commercial application servers is to build applications on it. Tracking various business events and logging them into application server logs is mainly what application server data consists of.

**3. Application-level data:** There are various new kinds of events that can be there in an application. The logging feature enabled in them helps us get the past record of the events.

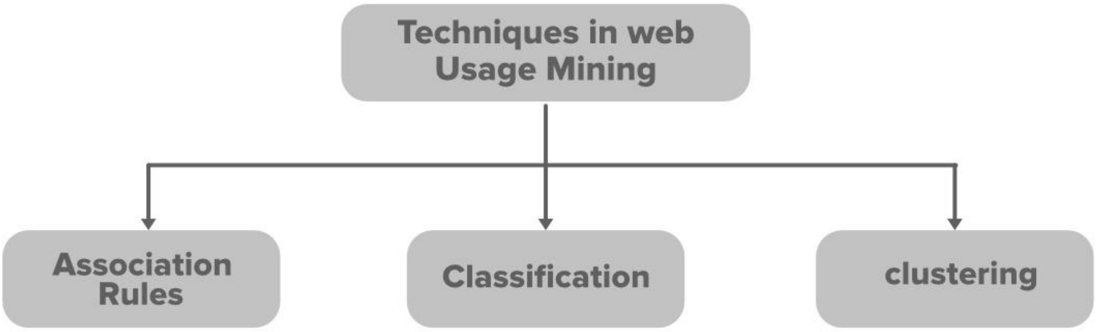
### Advantages of Web Usage Mining

- Government agencies are benefited from this technology to overcome terrorism.
- Predictive capabilities of mining tools have helped identify various criminal activities.
- Customer Relationship is being better understood by the company with the aid of these mining tools. It helps them to satisfy the needs of the customer faster and efficiently.

### Disadvantages of Web Usage Mining

- Privacy stands out as a major issue. Analyzing data for the benefit of customers is good. But using the same data for something else can be dangerous. Using it within the individual's knowledge can pose a big threat to the company.
- Having no high ethical standards in a data mining company, two or more attributes can be combined to get some personal information of the user which again is not respectable.

### Some Techniques in Web Usage Mining



**1. Association Rules:**The most used technique in Web usage mining is Association Rules. Basically, this technique focuses on relations among the web pages that frequently appear together in users' sessions. The pages accessed together are always put together into a single server session. Association Rules help in the reconstruction of websites using the access logs. Access logs generally contain information about requests which are approaching the webserver. The major drawback of this technique is that having so many sets of rules produced together may result in some of the rules being completely inconsequential. They may not be used for future use too.

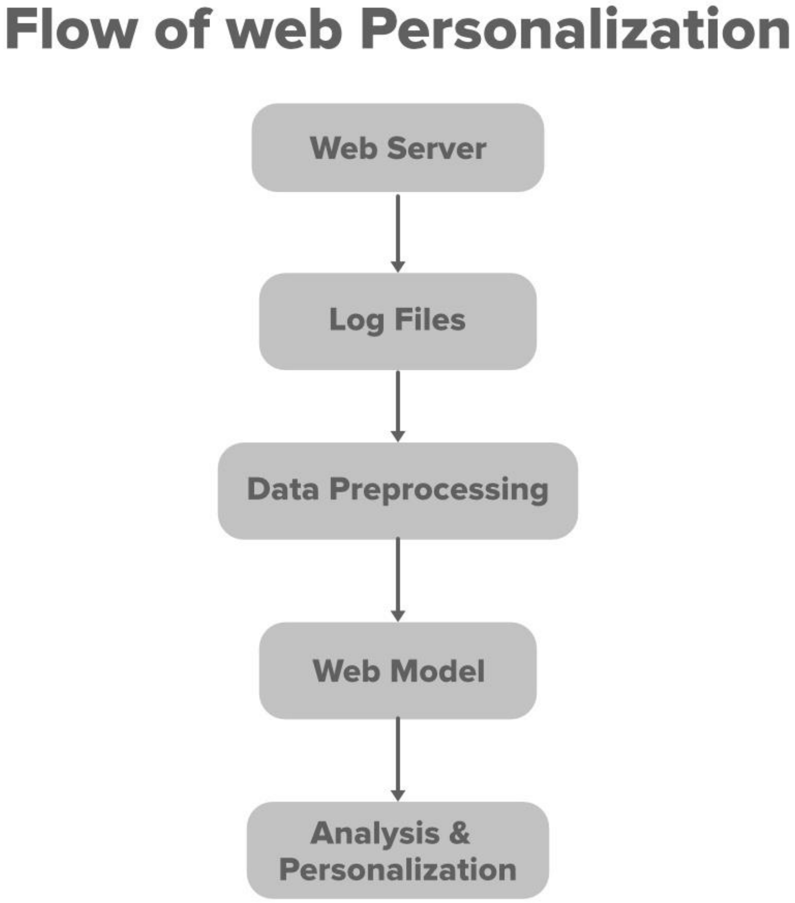
**2. Classification:** Classification is mainly to map a particular record to multiple predefined classes. The main target here in web usage mining is to develop that kind of profile of users/customers that are associated with a particular class/category. For this exact thing, one requires to extract the best features that will be best suitable for the associated class. Classification can be implemented by various algorithms – some of them include- Support vector machines, K-Nearest Neighbors, Logistic Regression, Decision Trees, etc. For example, having a track record of data of customers regarding their purchase history in the last 6 months the customer can be classified into frequent and non-frequent classes/categories. There can be multiclass also in other cases too.

**3. Clustering:** Clustering is a technique to group together a set of things having similar features/traits. There are mainly 2 types of clusters- the first one is the usage cluster and the second one is the page cluster. The clustering of pages can be readily performed based on the usage data.

In usage-based clustering, items that are commonly accessed /purchased together can be automatically organized into groups. The clustering of users tends to establish groups of users exhibiting similar browsing patterns. In page clustering, the basic concept is to get information quickly over the web pages.

**Applications of Web Usage Mining**

**1. Personalization of Web Content:** The World Wide Web has a lot of information and is expanding very rapidly day by day. The big problem is that on an everyday basis the specific needs of people are increasing and they quite often don't get that query result. So, a solution to this is web personalization. Web personalization may be defined as catering to the user's need-based upon its navigational behavior tracking and their interests. Web Personalization includes recommender systems, check-box customization, etc. Recommender systems are popular and are used by many companies.



**2. E-commerce:** Web-usage Mining plays a very vital role in web-based companies. Since their ultimate focus is on Customer attraction, customer retention, cross-sales, etc. To build a strong relationship with the customer it is very necessary for the web-based company to rely on web usage mining where they can get a lot of insights about customer's interests. Also, it tells the company about improving its web-design in some aspects.

**3. Prefetching and Catching:** Prefetching basically means loading of data before it is required to decrease the time waiting for that data hence the term 'prefetch'. All the results which we get from web usage mining can be used to produce prefetching and caching strategies which in turn can highly reduce the server response time.

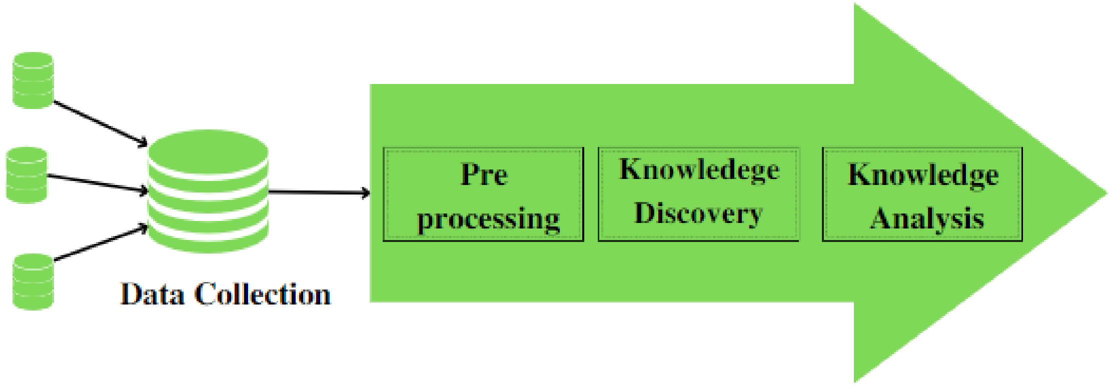
## Difference between Web Content, Web Structure, and Web Usage Mining

**Here are the following difference between web content, web structure, and web usage mining, such as:**

| **Terms** | **Web Content** | | **Web Structure** | **Web Usage** |
|---|---|---|---|---|
| | **IR View** | **DB View** | | |
| View of data | ○ Unstructured<br>○ Structured | ○ Semi-structured<br>○ Website as DB | Link structure | Interactivity |
| Main data | ○ Text documents<br>○ Hypertext documents | Hypertext documents | Link structure | ○ Server logs<br>○ Browser logs |
| Method | ○ Machine Learning<br>○ Statistical (Including NLP) | ○ Proprietary algorithm<br>○ Association rules | Proprietary algorithm | ○ Machine learning<br>○ Statistical<br>○ Association Rules |
| Representation | ○ Bag of words, n-gram terms<br>○ Phrases, concepts, or ontology<br>○ Relational | ○ Edged labeled graph<br>○ Relational | Graph | ○ Relational Table<br>○ Graph |
| Application Categories | ○ Categorization<br>○ Clustering<br>○ Finding Extract rules<br>○ Finding Patterns in text | ○ Finding frequent substructures<br>○ Web site schema discovery | ○ Categorization<br>○ Clustering | ○ Site construction<br>○ Adaptation an management |

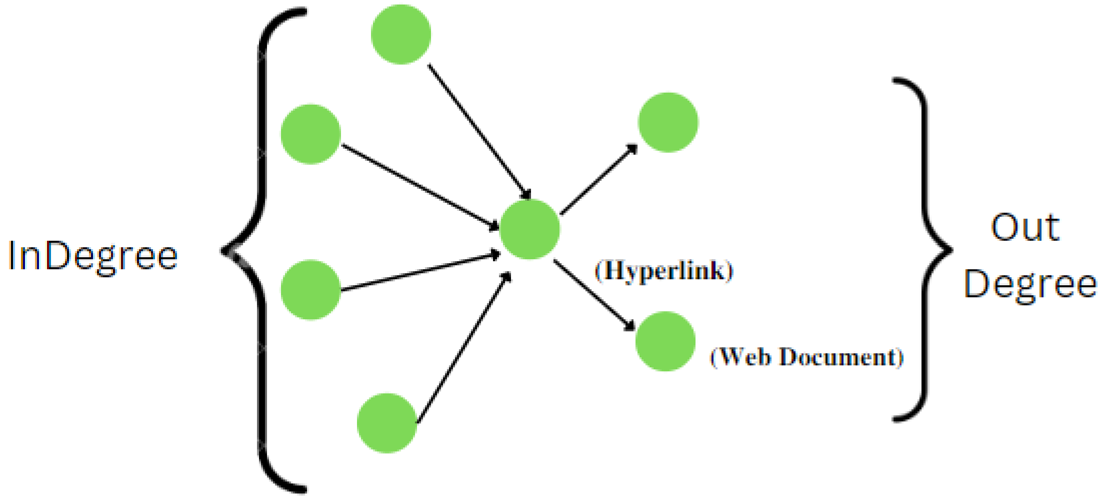
# Web Structure Mining

Web Structure Mining is one of the three different types of techniques in Web Mining. In this article, we will purely discuss about the Web Structure Mining. Web Structure Mining is the technique of discovering structure information from the web. It uses graph theory to analyze the nodes and connections in the structure of a website.



Depending upon the type of Web Structural data, Web Structure Mining can be categorised into two types:

**1.Extracting patterns from the hyperlink in the Web:** The Web works through a system of hyperlinks using the hyper text transfer protocol (http). Hyperlink is a structural component that connects the web page according to different location. Any page can create a hyperlink of any other page and that page can also be linked to some other page. the intertwined or self-referral nature of web lends itself to some unique network analytical algorithms. The structure of Web pages could also be analyzed to examine the pattern of hyperlinks among pages.



**2. Mining the document structure.** It is the analysis of tree like structure of web page to describe HTML or XML usage or the tags usage . There are different terms associated with Web Structure Mining :

- **Web Graph**: Web Graph is the directed graph representing Web.
- **Node**: Node represents the web page in the graph.
- **Edge(s)**: Edge represents the hyperlinks of the web page in the graph (Web graph)
- **In degree(s)**: It is the number of hyperlinks pointing to a particular node in the graph.

- **Degree(s)**: Degree is the number of links generated from a particular node. These are also called the Out Degrees.

All these terminologies will be more clear by looking at the following diagram of Web Graph:

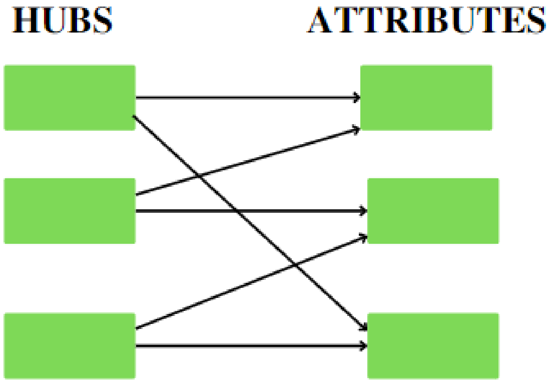
**Example of Web Structure Mining:**

One of the techniques is the **Page rank Algorithm** that the **Google** uses to rank its web pages. The rank of a page is dependent on the number of pages and the quality of links pointing to the target node.

So, we can say that the Web Structure Mining is the type of Mining that can be performed either at the **document level** (intra-page) or at the **hyperlink level** (inter-page). The research done at the hyperlink level is called as Hyperlink Analysis. the Hyperlink Structure can be used to retrieve useful information on the Web.

Web structure Mining basically has two main approaches or there are two basic strategic models for successful websites:

- Page rank : refer Page Rank
- Hubs and Authorities



*Hubs And Attributes*

- **Hubs:** These are pages with large number of interesting links. They serve as a hub or a gathering point, where people visit to access a variety of information. More focused sites can aspire to become a hub for the new emerging areas. The pages on website themselves could be analyzed for quality of content that attracts most users.
- **Authorities:** People usually gravitate towards pages that provide the most complete and authentic information on a particular subject. This could be factual information, news, advice, etc. these websites would have the most number of inbound links from other websites.

**Applications of Web Structure Mining:**

- Information retrieval in social networks.
- To find out the relevance of each web page.
- Measuring the completeness of Websites.
- Used in Search engines to find the relevant information.

# Web mining Software:

Here are the 10 most popular Web mining software & tools.

- Data Miner
- SimilarWeb
- Google Analytics
- Scrapy
- Bixo
- Oracle Data Mining
- Majestic
- WebScraper.io
- Tableau
- Wekan

### 1. Data Miner (Web content mining tool)

Data Miner

Web Mining Software and Data Mining Tool that is effective in scraping data from a website. It provides the scraped data into Excel & CSV format.

Data Miner delivers more than 40,000 solutions for numerous famous websites.

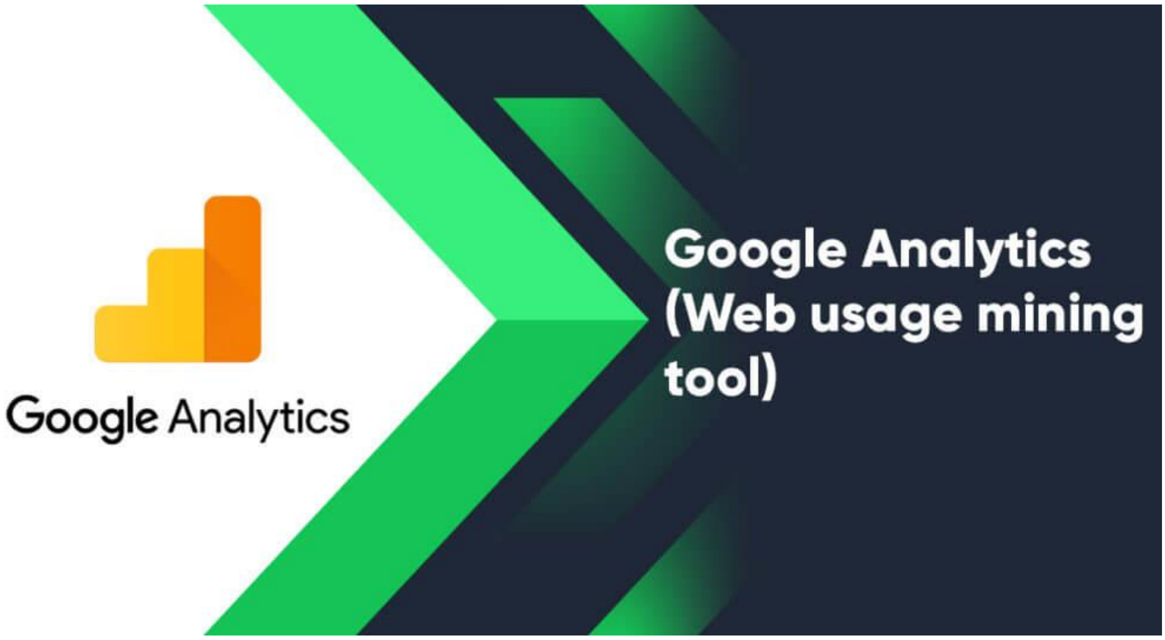
With these agendas, you can effortlessly acquire the organized data as per your need.

Features

- Scrape lists& tables
- One-click clutch
- Extract dynamic Ajax content
- Extract pages behind login/firewall

- Fill the form automatically
- Gather paginated outputs

**2. Google Analytics (Web usage mining tool)**



Google Analytics is treated as one of the finest business analysis tools as it can report & track site traffic.

More than 50% of users across the globe use website analysis and it assists you to carry out useful data to collect insights for your company.

Features

- Behavior & demographics analysis
- Campaign& advertising functioning analysis
- Conversion & Sales tools
- Data analysis of application performance &website
- Simply integrate Google products, like Google Tag Manager, Google Display Network, Adwords, Adsense, etc.
- Website testing & analysis

**3. SimilarWeb (Web usage mining tool)**

SimilarWeb

SimilarWeb is a robust business intelligence tool. With this tool, customers can understand the ranking, user engagement, and research of the website.

It can analyze website traffic, discover the features of identifying growth & participant website opportunities. It can assist you with traffic enhancement &track site traffic plans for different sites at a similar time.

In short, SimilarWeb is a good tool because it can assist you to track your overall company health, create effective companies' choices & track opportunities.

Features

- Audience importance
- Google Play Keyword Analysis
- Industry leader
- Interaction& Traffic metrics
- PPC keywords& Search engine optimization
- Traffic source

**4. Majestic (Web structure mining tool)**

Majestic

Majestic is a very useful business analysis tool that provides services for marketing companies, search engine optimization strategies, media analysts, and website developers. Majestic can assist you to access the world's leading index directory. You can obtain up-to-date & reliable data to examine the performance of the competitors & websites. It can also assist you to categorize the domain & each page through link mining or link analysis.

Features

- Advertising campaign
- Backlink history
- Bulk backlinks
- Comparison tool
- Keyword Checker
- Neighborhood check
- Rich plugins
- Search Explorer
- URL submission
- Website explorer

### 5. Scrapy (Web content mining tool)

Scrapy

Scrapy is one of the major open sources for web mining tools. It can assist you to scrape data from there user sessions, manage requests, website, follows handle output pipelines & redirects them.

Features

- HTTP functions such as authentication, caching, compression
- Interactive Shell console
- Processed Asynchronously& Requirements are dispatched
- Scrape data from HTML/XML& Select
- Session handling& Cookie

### 6. Bixo (Web structure mining tool)

Bixo

Bixo is an outstanding website mining open-source tool that helps you to run a sequence of followed pipelines on top of Hadoop. Building customized followed pipeline modules; you can rapidly develop web mining app for particular use-cases.

Features

- Get subassembly
- Lack of data visualization capabilities
- Parsing subcomponent

### 7. Oracle Data Mining (Web Usage Mining Tool)

Oracle Data Mining was invented by Oracle. Data mining software provides a brilliant data mining process that can assist you to make predictions, collect insights, make effective use of Oracle investments& data.

With Oracle Data Mining, guessing models are found on the Oracle website so that you can easily focus on your specific customer groups, develop customer profiles, and predict customer behavior. You easily recognize opportunities for identifying differences, prospects for fraud, and cross-selling.

Features

- Abnormal detection
- Active Data Guard
- Association
- Attribute importance
- Classification
- Clustering
- Database library
- Feature selection and extraction
- Online analysis and processing
- Return
- Space mining
- Text mining

**8. Tableau (Web usage mining tool)**

Tableau is one of the fastest-growing data & fastest used tools in business intelligence organizations. It helps you to streamline the unique data into a structured format. Visualization data can be effortlessly executed through worksheets& dashboards.

The Tableau product set contains

- Tableau Public
- Tableau Online
- Tableau desktop
- Tableau Reader
- Tableau Server

The Tableau product set contains

Here are some key features of Tableau include:

- Additional connector
- Android improvements
- Automatic query cache
- Auto-update
- Convert query to visualization
- Create a “no-code” data query
- Create interactive dashboard
- Dashboard comments
- Data-Driven Alert
- Highlight and filter data
- Import data of all ranges and sizes
- Metadata management
- PDF connector
- Shared dashboard
- Smart connection
- String in-depth understanding guidance
- Switch view and drag and drop
- Tableau Bridge
- Tableau Reader for data viewing

## 9. WebScraper.io (Web content mining tool)

Weka is a group of the machine learning process for data mining tasks. It includes tools for classification, regression, association rule mining, clustering, visualization, and data preparation.

Weka was mainly designed as a tool for investigating data from the agricultural field, but in recent times a fully Java-based version (Weka 3), which was derived in 1997, is now used in many various apps fields, research & especially for educational purposes.

Features

- Classification
- Cluster
- Data preprocessing
- Function selection
- Return
- Visualization