# Module 6 --> Web Mining

**Web Mining** is the process of [Data Mining](https://www.geeksforgeeks.org/data-mining/) techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.  It involves using data mining techniques to analyze web data and extract valuable insights.

**Applications of Web Mining:**

The applications of web mining are wide-ranging and include:

**Personalized marketing:**

Web mining can be used to analyze customer behavior on websites and social media platforms. This information can be used to create personalized marketing campaigns that target customers based on their interests and preferences.

**E-commerce**

Web mining can be used to analyze customer behavior on e-commerce websites. This information can be used to improve the user experience and increase sales by recommending products based on customer preferences.

**Search engine optimization:**

Web mining can be used to analyze search engine queries and search engine results pages (SERPs). This information can be used to improve the visibility of websites in search engine results and increase traffic to the website.

**Fraud detection:**

Web mining can be used to detect fraudulent activity on websites. This information can be used to prevent financial fraud, identity theft, and other types of online fraud.

**Sentiment analysis:**

 Web mining can be used to analyze social media data and extract sentiment from posts, comments, and reviews. This information can be used to understand customer sentiment towards products and services and make informed business decisions.

**Web content analysis:**

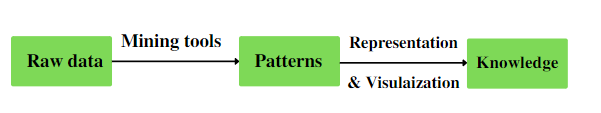
Web mining can be used to analyze web content and extract valuable information such as keywords, topics, and themes. This information can be used to improve the relevance of web content and optimize search engine rankings.

**Customer service:**

Web mining can be used to analyze customer service interactions on websites and social media platforms. This information can be used to improve the quality of customer service and identify areas for improvement.

**Healthcare:**Web mining can be used to analyze health-related websites and extract valuable information about diseases, treatments, and medications. This information can be used to improve the quality of healthcare and inform medical research.

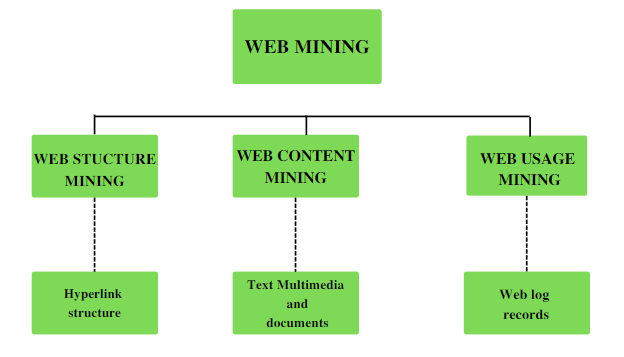
**Process of Web Mining:**



***Web Mining Process***

| **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 mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.



### What is Web Content Mining?

Web Content Mining is 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.

It is also quite different from data mining because web mining data are mainly semi-structured or unstructured, while data mining deals primarily with structured data.

Web content mining is also different from text mining because of the semi-structured nature of the web, while text mining focuses on unstructured texts.

Thus, Web content mining requires creative applications of data mining and text mining techniques and its own unique approaches.

Web content mining could be done in from of two approaches, such as:

****1. Agent-based Approach****

This approach involves intelligent systems. It aims to improve information finding and filtering. It usually relies on autonomous agents that can identify relevant websites. And it could be placed into the following three categories, such as:

* ****Intelligent Search Agents:**** These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
* ****Information Filtering or Categorization:**** These agents use information retrieval techniques and characteristics of open hypertext Web documents to retrieve automatically, filter, and categorize them.
* ****Personalized Web Agents:**** These agents learn user preferences and discover Web information based on other users' preferences with similar interests.

****2. Data based approach****

Data based approach is used to organize semi-structured data present on the internet into structured data. It aims to model the web data into a more structured form to apply standard database querying mechanisms and data mining applications to analyze it.

### Web Content Mining Challenges

Web content mining has the following problems or challenges also with their solutions, such as:

* ****Data Extraction:**** Extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction, are used to solve this problem.
* ****Web Information Integration and Schema Matching:**** Although the Web contains a huge amount of data, each website (or even page) represents similar information differently. Identifying or matching semantically similar data is an important problem with many practical applications.
* ****Opinion extraction from online sources:**** There are many online opinion sources, e.g., customer reviews of products, forums, blogs, and chat rooms. Mining opinions are of great importance for marketing intelligence and product bench marking.
* ****Segmenting Web pages and detecting noise:**** In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web pages to extract the pages' main content is an interesting problem.

### What is Web Structure Mining?

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis has increased. These efforts resulted in a newly emerging research area called **Link Mining**, which is located at the intersection of the work in link analysis, hypertext, web mining, relational learning, inductive logic programming, and graph mining.

Web structure mining uses graph theory to analyze a website's node and connection structure. According to the type of web structural data, web structure mining can be divided into two kinds:

* ****Extracting patterns from hyperlinks in the web:**** a hyperlink is a structural component that connects the web page to a different location.
* ****Mining the document structure:**** analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

The web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in, out, and co-citation (two pages linked to by the same page). Attributes include HTML tags, word appearances, and anchor texts. Web structure mining includes the following terminology, such as:

* ****Web graph:**** directed graph representing web.
* ****Node:**** web page in the graph.
* ****Edge:**** hyperlinks.
* ****In degree:**** the number of links pointing to a particular node.
* ****Out degree:**** number of links generated from a particular node.

An example of a technique of web structure mining is the **PageRank** algorithm used by Google to rank search results. A page's rank is decided by the number and quality of links pointing to the target node.

Link mining had produced some agitation on some traditional data mining tasks. Below we summarize some of these possible tasks of link mining which are applicable in Web structure mining, such as:

1. ****Link-based Classification:**** The most recent upgrade of a classic data mining task to linked Domains. The task is to predict the category of a web page based on words that occur on the page, links between pages, anchor text, html tags, and other possible attributes found on the web page.
2. ****Link-based Cluster Analysis:**** The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Unlike the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.
3. ****Link Type:**** There is a wide range of tasks concerning predicting the existence of links, such as predicting the type of link between two entities or predicting the purpose of a link.
4. ****Link Strength:**** Links could be associated with weights.
5. ****Link Cardinality:**** The main task is to predict the number of links between objects. page categorization used to
   * Finding related pages.
   * Finding duplicated websites and finding out the similarity between them.

### What is Web Usage Mining?

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW.

Web usage mining, discovers user navigation patterns from web data, trying to discover useful information from the secondary data derived from users' interactions while surfing the web.

Web usage mining collects the data from Weblog records to discover user access patterns of web pages. Several available research projects and commercial tools analyze those patterns for different purposes. The insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence, and usage characterization.

The only information left behind by many users visiting a Web site is the path through the pages they have accessed. Most of the Web information retrieval tools only use textual information, while they ignore the link information that could be very valuable. In general, there are mainly four kinds of data mining techniques applied to the web mining domain to discover the user navigation pattern, such as:

****1. Association Rule Mining****

Association rule is the most basic rule of data mining methods which is used more than other methods in web usage mining. This method enables the website for more efficient content organization or provides recommendations for an effective cross-selling product.

These rules are statements in the form X => Y where (X) and (Y) are the set of available items in a series of transactions. The rule of X => Y states that transactions that contain items in X may also include items in Y. Association rules in the web usage mining are used to find relationships between pages that frequently appear next to one another in user sessions.

1. ****Sequential Patterns****

Sequential patterns are used to discover the subsequence in a large volume of sequential data. In web usage mining, sequential patterns are used to find user navigation patterns that frequently appear at meetings. The sequential patterns may seem to be association rules. But the sequential patterns are included the time, which means that the sequence of events that occurred is defined in sequential patterns. Algorithms that are used to extract association rules can also be used to generate sequential patterns. Two types of algorithms are used for sequential mining patterns.

* The first type of algorithm is based on association rules mining. Many common algorithms of sequential mining patterns have been changed for mining association rules. For example, GSP and AprioriAll are two developed species of Apriori algorithms that are used to extract association rules.
* The second type of sequential patterns mining algorithms has been introduced in which the tree structure and Markov chain are used to represent survey patterns. For example, in one of these algorithms called WAP-mine, the tree structure called WAP-tree is used to explore access patterns to the web. Evaluation results show that its performance is higher than an algorithm such as GSP.

****3. Clustering****

Clustering techniques diagnose groups of similar items among high volumes of data. This is done based on distance functions which measure the degree of similarity between different items. Clustering in web usage mining is used for grouping similar meetings. What is important in this type of search is the contrast between the user and individual groups. Two types of interesting clustering can be found in this area: user clustering and page clustering.

Clustering of user records is usually used to analyze web mining and web analytics tasks. More knowledge derived from clustering is used to partition the market in e-commerce. Different methods and techniques are used for clustering, which includes:

* Using the similarity graph and the amount of time spent viewing a page to estimate the similarity of meetings.
* Using genetic algorithms and user feedback.
* Clustering matrix.
* K -means algorithm, which is the most classic clustering method.

The repetitive patterns are first extracted from the user's sessions using association rules in other clustering methods. Then, these patterns are used to construct a graph where the nodes are the visited pages. The edges of the graph connect two or more pages. If these pages exist in a pattern extracted, the weight will be assigned to the edges that show the relationship between the nodes. Then, for clustering, this graph is recursively divided to user behavior groups are detected.

****4. Classification Mining****

Discovering classification rules allows one to develop a profile of items belonging to a particular group according to their common attributes. This profile can classify new data items added to the database. In Web Mining, classified techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients or their navigation patterns.

### Advantages

Web usage mining has many advantages, making this technology attractive to corporations, including government agencies.

* This technology has enabled e-commerce to do personalized marketing, resulting in higher trade volumes. Government agencies are using this technology to classify threats and fight against terrorism.
* Companies can establish better customer relationships by understanding the customer's needs better and reacting to customer needs faster. They can increase profitability by target pricing based on the profiles created. They can even find customers who might default to a competitor. The company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.
* More benefits of web usage mining, particularly personalization, are outlined in specific frameworks like the probabilistic latent semantic analysis model, which offers additional features to user behavior and access patterns. This is because the process provides the user with more relevant content through collaborative recommendations.
* There are also elements unique to web usage mining that show the technology's benefits. These include the way semantic knowledge is applied when interpreting, analyzing and reasoning about usage patterns during the mining phase.

### Disadvantages

Web usage mining by itself does not create issues, but when used on data of personal nature, this technology might cause concerns.

* The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without the individual's knowledge or consent. The obtained data will be analyzed, made anonymous, and then clustered to form anonymous profiles.
* These applications de-individualize users by judging them by their mouse clicks rather than by identifying information. De-individualization, in general, can be defined as a tendency to judge and treat people based on group characteristics instead of on their characteristics and merits.
* The companies collecting the data for a specific purpose might use the data for totally different purposes, violating the user's interests.

### Web Usage Mining Applications

The main objective of web usage mining is to collect data about the user's navigation patterns. This information can improve the Web sites in the user view. There are three main applications of this mining, such as:

****1. Privatization of web content****

Web usage mining techniques can be used for the personalization of web users. For example, user behavior can be immediately predicted by comparing her current survey patterns with those extracted from the log files. Recommendation systems with a real application in this area suggest links that direct the user to his favorite pages. Some sites also organize their product catalogs based on the predicted interests of a specific user and represent them.

1. ****Pre - recovery****

The results of web usage mining can be used to improve the performance of Web servers and Web-based applications. Web usage mining can be used for retrieving and caching strategies and thus reduce the response time of Web servers.

****3. Improvement of Web site design****

Usability is one of the most important issues in designing and implementing websites. The results of web usage mining can help to appropriate the design of websites. Adaptive websites are an application of this type of mining. Website content and structure are dynamically reorganized based on data derived from user behavior in these sites.

# Text Data Mining

Text data mining can be described as the process of extracting essential data from standard language text. All the data that we generate via text messages, documents, emails, files are written in common language text. Text mining is primarily used to draw useful insights or patterns from such data.

The primary source of data is e-commerce websites, social media platforms, published articles, survey, and many more. The larger part of the generated data is unstructured, which makes it challenging and expensive for the organizations to analyze with the help of the people. It is not only able to handle large volumes of text data but also helps in decision-making purposes. Text mining software empowers a user to draw useful information from a huge set of data available sources.

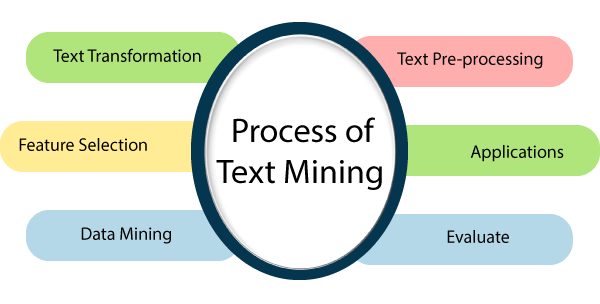
## Areas of text mining in data mining:

These are the following area of text mining :

* ****Information Extraction:****  
  The automatic extraction of structured data such as entities, entities relationships, and attributes describing entities from an unstructured source is called information extraction.
* ****Natural Language Processing:****  
  NLP stands for Natural language processing. Computer software can understand human language as same as it is spoken. NLP is primarily a component of artificial intelligence(AI). The development of the NLP application is difficult because computers generally expect humans to "Speak" to them in a programming language that is accurate, clear, and exceptionally structured. Human speech is usually not authentic so that it can depend on many complex variables, including slang, social context, and regional dialects.
* ****Data Mining:****  
  Data mining refers to the extraction of useful data, hidden patterns from large data sets. Data mining tools can predict behaviors and future trends that allow businesses to make a better data-driven decision. Data mining tools can be used to resolve many business problems that have traditionally been too time-consuming.
* ****Information Retrieval:****  
  Information retrieval deals with retrieving useful data from data that is stored in our systems. Alternately, as an analogy, we can view search engines that happen on websites such as e-commerce sites or any other sites as part of information retrieval.

## Text Mining Process:

The text mining process incorporates the following steps to extract the data from the document.



* ****Text transformation****  
  A text transformation is a technique that is used to control the capitalization of the text.  
  Here the two major way of document representation is given.
  1. Bag of words
  2. Vector Space
* ****Text Pre-processing****  
  Pre-processing is a significant task and a critical step in Text Mining, Natural Language Processing (NLP), and information retrieval(IR). In the field of text mining, data pre-processing is used for extracting useful information and knowledge from unstructured text data. Information Retrieval (IR) is a matter of choosing which documents in a collection should be retrieved to fulfill the user's need.
* ****Feature selection:****  
  Feature selection is a significant part of data mining. Feature selection can be defined as the process of reducing the input of processing or finding the essential information sources. The feature selection is also called variable selection.
* ****Data Mining:****  
  Now, in this step, the text mining procedure merges with the conventional process. Classic Data Mining procedures are used in the structural database.
* ****Evaluate:****  
  Afterward, it evaluates the results. Once the result is evaluated, the result abandon.
* ****Applications:****  
  These are the following text mining applications:
* ****Risk Management:****  
  Risk Management is a systematic and logical procedure of analyzingthe risks involved in any process in organizations. Insufficient risk analysis is usually a leading cause of disappointment. It is particularly true in the financial organizations where adoption of Risk Management Software based on text mining technology can effectively enhance the ability to diminish risk. It enables the administration of millions of sources and petabytes of text documents, and giving the ability to connect the data. It helps to access the appropriate data at the right time.
* ****Customer Care Service:****  
  Text mining methods, particularly NLP, are finding increasing significance in the field of customer care. Organizations are spending in text analytics programming to improve their overall experience by accessing the textual data from different sources such as customer feedback, surveys, customer calls, etc. The primary objective of text analysis is to reduce the response time of the organizations and help to address the complaints of the customer rapidly and productively.
* ****Business Intelligence:****  
  Companies and business firms have started to use text mining strategies as a major aspect of their business intelligence. Besides providing significant insights into customer behavior and trends, text mining strategies also support organizations to analyze the qualities and weaknesses of their opponent's so, giving them a competitive advantage in the market.
* ****Social Media Analysis:****  
  Social media analysis helps to track the online data, and there are numerous text mining tools designed particularly for performance analysis of social media sites. These tools help to monitor and interpret the text generated via the internet from the news, emails, blogs, etc. Text mining tools can precisely analyze the total no of posts, followers, and total no of likes of your brand on a social media platform that enables you to understand the response of the individuals who are interacting with your brand and content.

## Text Mining Approaches in Data Mining:

These are the following text mining approaches that are used in data mining.

****1. Keyword-based Association Analysis:****

It collects sets of keywords or terms that often happen together and afterward discover the association relationship among them. First, it preprocesses the text data, Once it pre-processed the data, then it induces association mining algorithms. Here, human effort is not required, so the number of unwanted results and the execution time is reduced.

****2. Document Classification Analysis:****

Automatic document classification:

This analysis is used for the automatic classification of the huge number of online text documents like web pages, emails, etc. Text document classification varies with the classification of relational data as document databases are not organized according to attribute values pairs.

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