**Study Experiment**

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:

1. Web Content Mining:  
   Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

**Unstructured Text WEB CONTENT Mining**: Web content data is much of unstructured text data. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence, one could consider text mining as an instance of Web content mining. To provide effectively exploitable results, preprocessing steps for any structured data is done by means of information extraction, text categorization, or applying NLP techniques.

**Semi-Structured** **Structured WEB CONTENT Mining**: Structured data on the Web are often very important as they represent their host pages, due to this reason it is important and popular. Structured data is also easier to extract compared to unstructured texts. Semi-structured data is a point of convergence for the Web and database communities: the former deals with documents, the latter with data. The form of that data is evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real- world objects like books, papers, movies, etc., without sending the application writer into contortions. Emergent representations for semi-structured data (such as XML) are variations on the Object Exchange Model (OEM). In OEM, data is in the form of atomic or compound objects: atomic objects may be integers or strings; compound objects refer to other objects through labeled edges. HTML is a special case of such intra-document structure

1. Web Structure Mining:  
   Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.
2. Web Usage Mining:  
   Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviours or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

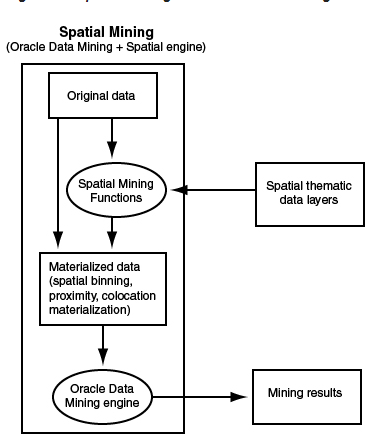
Comparison Between Data mining and Web mining:

|  |  |  |
| --- | --- | --- |
| 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 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. |
| 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. |

A2.)

Spatial Data Mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. This chapter provides an overview on the unique features that distinguish spatial data mining from classical Data Mining, and presents major accomplishments of spatial Data Mining research.

To perform spatial data mining, you materialize spatial predicates and relationships for a set of spatial data using thematic layers. Each layer contains data about a specific kind of spatial data (that is, having a specific "theme"), for example, parks and recreation areas, or demographic income data. The spatial materialization could be performed as a preprocessing step before the application of data mining techniques, or it could be performed as an intermediate step in spatial mining, as shown here



The original data, which included spatial and nonspatial data, is processed to produce materialized data.

Spatial data in the original data is processed by spatial mining functions to produce materialized data. The processing includes such operations as spatial binning, proximity, and colocation materialization.

The following are examples of the kinds of data mining applications that could benefit from including spatial information in their processing:

Business prospecting: Determine if colocation of a business with another franchise (such as colocation of a Pizza Hut restaurant with a Blockbuster video store) might improve its sales.

Store prospecting: Find a good store location that is within 50 miles of a major city and inside a state with no sales tax. (Although 50 miles is probably too far to drive to avoid a sales tax, many customers may live near the edge of the 50-mile radius and thus be near the state with no sales tax.)

Hospital prospecting: Identify the best locations for opening new hospitals based on the population of patients who live in each neighborhood.

Spatial region-based classification or personalization: Determine if southeastern United States customers in a certain age or income category are more likely to prefer "soft" or "hard" rock music.

Automobile insurance: Given a customer's home or work location, determine if it is in an area with high or low rates of accident claims or auto thefts.

Property analysis: Use colocation rules to find hidden associations between proximity to a highway and either the price of a house or the sales volume of a store.

Property assessment: In assessing the value of a house, examine the values of similar houses in a neighborhood, and derive an estimate based on variations and spatial correlation.

A3.)

Temporal data mining refers to the extraction of implicit, non-trivial, and potentially useful abstract information from large collections of temporal data. Temporal data are sequences of a primary data type, most commonly numerical or categorical values and sometimes multivariate or composite information. Examples of temporal data are regular time series (e.g., stock ticks, EEG), event sequences (e.g., sensor readings, packet traces, medical records, weblog data), and temporal databases (e.g., relations with timestamped tuples, databases with versioning). The common factor of all these sequence types is the total ordering of their elements. They differ on the type of primary information, the regularity of the elements in the sequence, and on whether there is explicit temporal information associated to each element (e.g., timestamps).

Temporal data mining is concerned with data mining of large sequential data sets. By

sequential data, we mean data that is ordered with respect to some index. For example, time

series constitute a popular class of sequential data, where records are indexed by time. Other

examples of sequential data could be text, gene sequences, protein sequences, lists of moves

in a chess game etc. Here, although there is no notion of time as such, the ordering among

the records is very important and is central to the data description/modelling

A4.)

Association rules are sentences of type X ◊ Y (X implies Y, X is called the antecedent and Y the consequent), where X, Y are sets of frequent items in a given database such that X ∩ Y = Φ. If rule X ◊ Y is true, then it has valid support, P(X U Y) and confidence, P(Y/X), where the support of the rule represents the percentage of transactions in the database that contain both X and Y and, the confidence of the rule is the percentage of transactions in the database containing X that also contain Y. In the previous step each transaction is made up of the items that appear before a pre-set antecedent, within a time window with a set width and a lag also set by the analyst. The association rules obtained from this database will already include time constraints, because of the way in which the previous database is created.

A5.)

Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. A [data stream](https://en.wikipedia.org/wiki/Data_stream) is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities.

In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Machine learning techniques can be used to learn this prediction task from labelled examples in an automated fashion. Often, concepts from the field of [incremental learning](https://en.wikipedia.org/wiki/Incremental_learning) are applied to cope with structural changes, [on-line learning](https://en.wikipedia.org/wiki/Online_Machine_Learning) and real-time demands. In many applications, especially operating within non-stationary environments, the distribution underlying the instances or the rules underlying their labelling may change over time, i.e. the goal of the prediction, the class to be predicted or the target value to be predicted, may change over time. This problem is referred to as [concept drift](https://en.wikipedia.org/wiki/Concept_drift). Detecting [concept drift](https://en.wikipedia.org/wiki/Concept_drift) is a central issue to data stream mining. Other challenges that arise when applying machine learning to streaming data include: partially and delayed labelled data, recovery from concept drifts, and temporal dependencies.