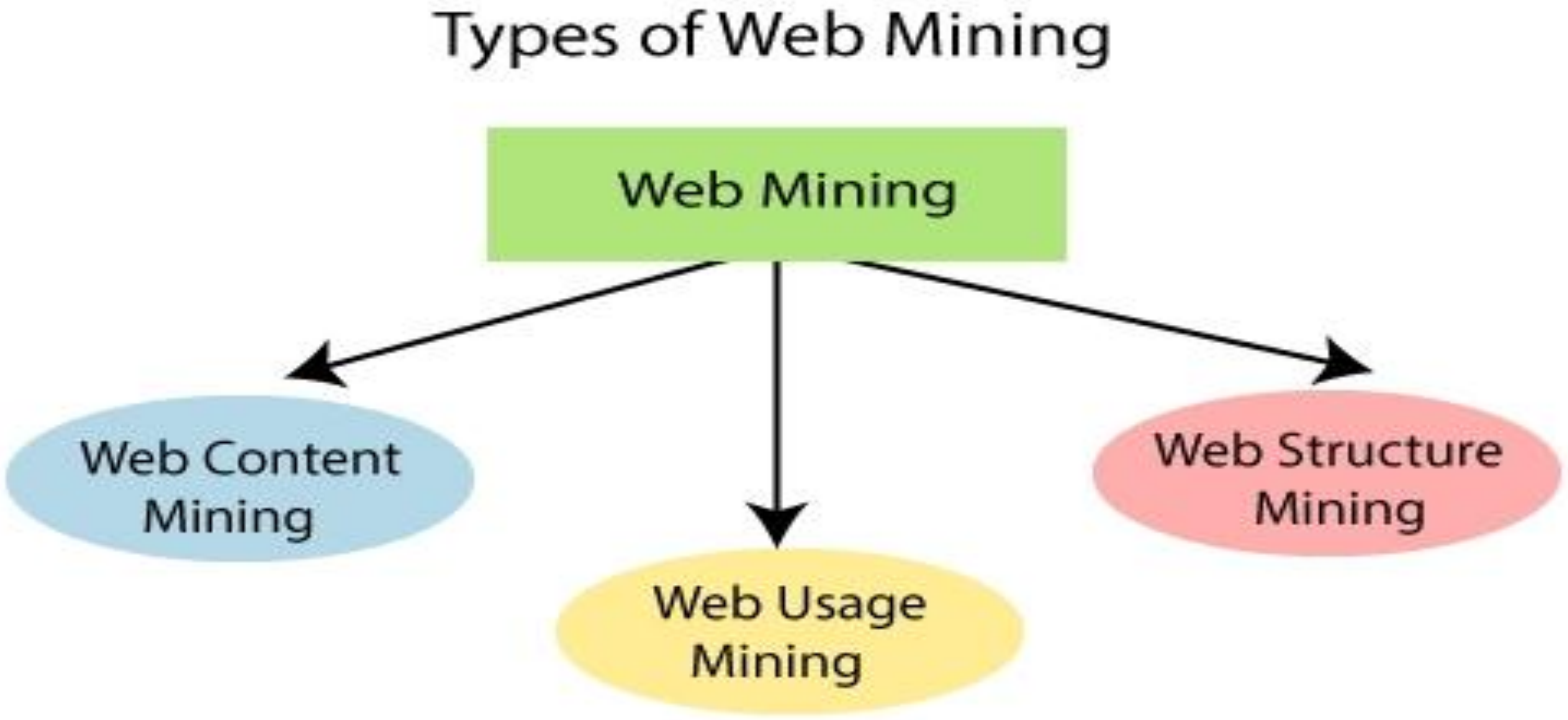
Unit 5

Web Data Mining

Web Data 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:


Types of Web 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.

1. 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.
- It is also quite different from data mining because web 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.
- In the past few years, there has been a rapid expansion of activities in the web content mining area. This is not surprising because of the phenomenal growth of web content and the significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of web data, automated discovery of targeted or unexpected knowledge information still present many challenging

1. Web Content Mining:

Web content mining could be differentiated from 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.

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.

2. Web Structured 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.

2. Web Structured Mining:

- 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..

2. Web Structured Mining:

- 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:

2. Web Structured Mining:

- **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.
- **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.
- **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.
- **Link Strength:** Links could be associated with weights.
- **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

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

3. 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, discovering 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:

3. Web Usage Mining:

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.

3. Web Usage Mining:

2. 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.

3. Web Usage Mining:

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.

3. Web Usage Mining:

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.

Challenges in Web Mining:

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

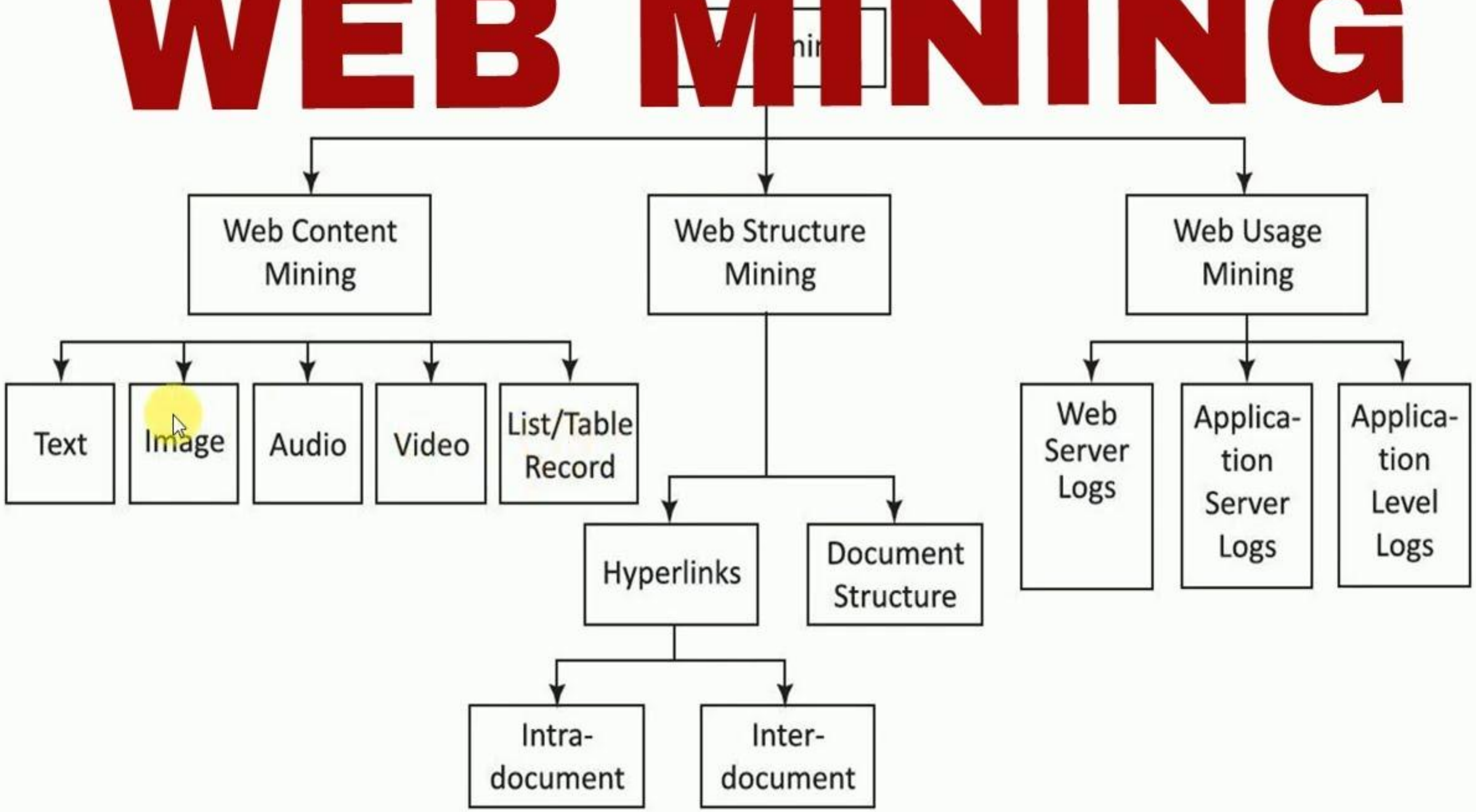
Challenges in Web Mining:

- 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.

Application of Web Mining:

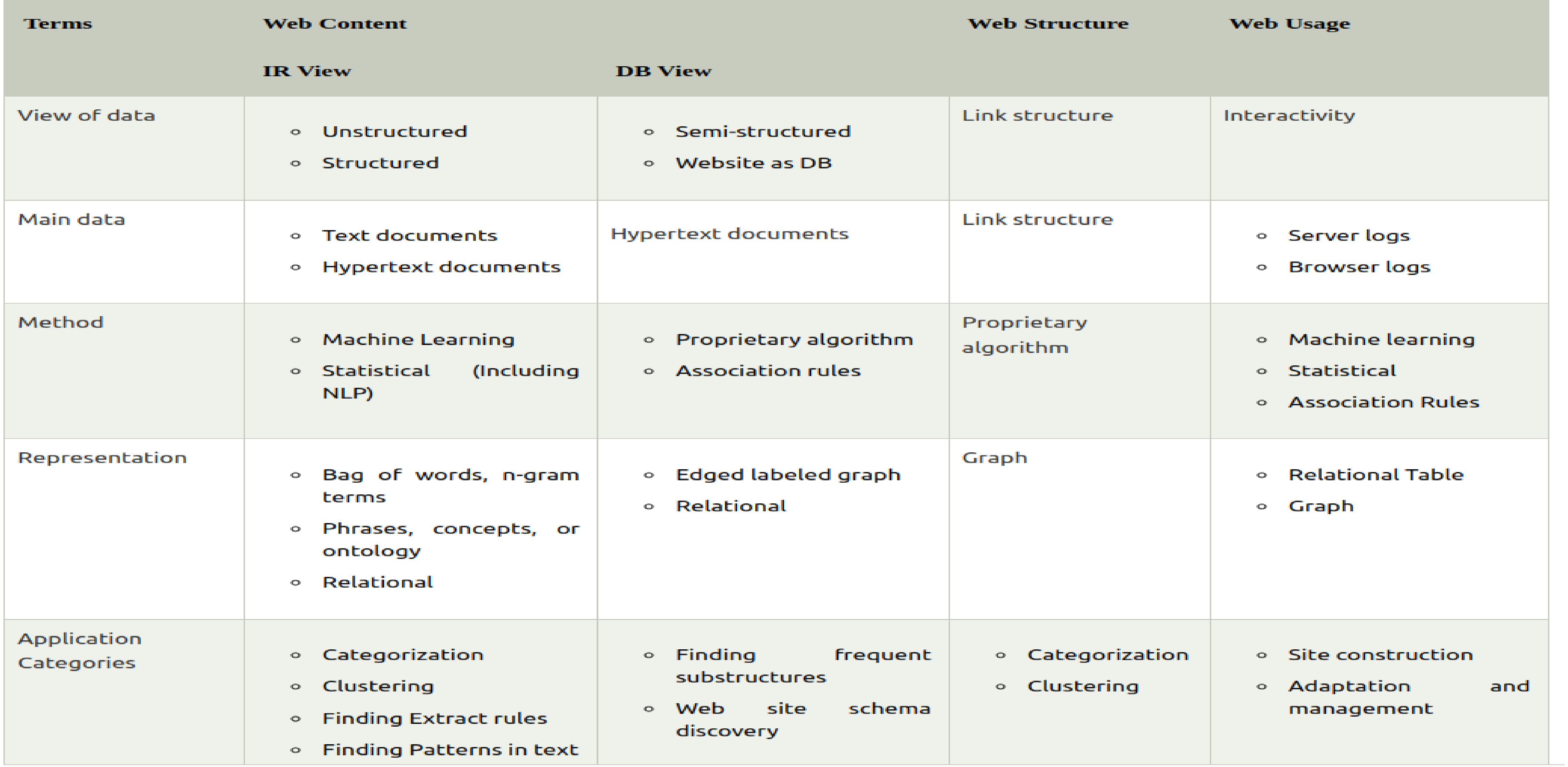
- Web mining has an extensive application because of various uses of the web. The list of some applications of web mining is given below.
- Marketing and conversion tool
- Data analysis on website and application accomplishment.
- Audience behavior analysis
- Advertising and campaign accomplishment analysis.
- Testing and analysis of a site.

WEB MINING



- Web Content Mining
  - Text
  - Image
  - Audio
  - Video
  - List/Table Record
- Web Structure Mining
  - Hyperlinks
    - Intra-document
    - Inter-document
  - Document Structure
- Web Usage Mining
  - Web Server Logs
  - Application Server Logs
  - Application Level Logs

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



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 management