Introduction:

Overview of data mining and web mining.

Web communities and social networks.

Reminders

The internet comprises various protocols and components that allow it to function seamlessly. The core components include:

Internet:	TCP/IP	HTTP
URL	Hyperlinks	HTML

What is the World Wide Web?

The World Wide Web (WWW or Web) is an application of the internet that enables users to browse and share documents via web browsers. Initially designed for document sharing, the web has evolved into a platform for developing new technologies. The core technologies of the web are HTTP protocols and HTML/URL document formats.

The web is one of the many applications of the internet, alongside email, video conferencing, and peer-to-peer file sharing.

Client-Server Model: This model is fundamental to the web, where a client (user's browser) requests resources or services from a server (web server).

Web Data Characteristics

Web data is characterized by its:

- **Enormous Size**: The web's volume grows exponentially.
- **Heterogeneity**: It includes diverse data types such as text, images, audio, video files, and programs within a single page.
- **Distribution**: Data is dispersed geographically across various computers and platforms.
- **Unstructured Nature**: The web comprises structured data (databases), semi-structured data (HTML/XML documents), and unstructured data (free text).
- **Dynamicity**: Continuous addition, updating, and deletion of data and links.

What is Data Mining?

Data mining, also known as knowledge discovery from data, involves developing methods and tools to automate the process of extracting knowledge and discovering patterns from large datasets. These datasets can include databases, texts, images, and web data.

Definition: Data mining is a non-trivial process of extracting valid, comprehensible, previously unknown, and potentially useful information from large datasets (Fayyad et al., 1996). It is a multidisciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization.

Data Table

In data mining, a data table (also known as a data set or data matrix) is the fundamental format in which data is organized for analysis. Each table typically consists of rows and columns, where:

- **Rows** (records) represent individual instances, observations, or transactions in the dataset
- **Columns** (attributes/features) represent the variables or characteristics of the data being collected.
- **Target Variable**: In supervised learning, there is typically a column representing the target or label that the model will predict

Types of Data

- Numerical Data: Continuous Discrete
- Categorical Data: Nominal Ordinal
- Binary Data

Data Mining tasks 1/4

Data mining tasks can be broadly categorized into:

Supervised learning : Supervised learning tasks require labeled datasets where the input data is paired with the correct output (label or target). The model learns from this training data to make accurate predictions on new, unseen data.

Classification: Assign a label or category to input data.

Examples:

- Email spam detection (spam vs. not spam)
- Image recognition (dog vs. cat vs. other)
- Sentiment analysis (positive vs. negative review)
- Disease diagnosis (e.g., classify medical images as cancerous or non-cancerous)

Data Mining tasks 2/4

Regression (Estimation): Predict a continuous numerical value based on input data.

Examples:

- Predicting house prices based on features (e.g., size, location)
- Forecasting sales based on historical sales data
- Predicting temperature or stock prices

Sequence Prediction : Predict the next item in a sequence or a sequence of values.

Examples:

- Predicting the next word in a sentence
- Predicting the next frame in a video sequence

Data Mining tasks 3/4

Unsupervised learning : unsupervised learning tasks are focused on discovering underlying patterns or structures in data without explicit labels or supervision.

Clustering: It allows the partitioning or grouping of similar data into groups, which can reveal hidden structures or similarities between the data.

Examples :

- Customer segmentation (grouping customers by purchasing behavior)
- Document clustering (grouping similar documents or articles)
- Image segmentation (grouping pixels in images without labeled data)

Data Mining tasks 4/4

Anomaly detection : Identify data points that deviate significantly from the norm or general pattern.

Examples :

- Fraud detection in financial transactions
- Detecting manufacturing defects in quality control
- Identifying unusual network traffic in cybersecurity

Association Rule: Discover relationships or patterns between items in large datasets, such as finding frequent itemsets.

Examples :

- Market basket analysis (identifying items frequently bought together, e.g., "customers who buy bread often buy butter")
- Recommendation systems (e.g., "customers who watched this movie also watched…")

Steps of applying data Minig

1. Understanding the scope of application (identifying the Objectives)
2. Data preparation
    1. Data collection
    2. Cleaning
    3. Integration
    4. Selection
    5. Transformation
3. Data mining
    1. Defining the task
    2. Choice of algorithms
    3. Performing the mining
4. Analysis of results
    1. Presentation
    2. interpretation
    3. Evaluation and validation
5. Exploitation of results

Web Mining

Web mining is the process of extracting knowledge from the web. That is, applying data mining techniques to discover patterns from the web.

It can be used for a variety of purposes, including:

- Decision Support: Enhancing decision-making processes.
- Information Retrieval: Improving the efficiency of retrieving information.
- Personalization: Tailoring content to user preferences.
- Strategic Monitoring: Monitoring web activity for strategic insights.
- Fraud and Cyber Threat Detection: Identifying fraudulent activities and cyber threats.

Categories of Web Mining

Depending on the analysis objectives, Web Mining could be classified into three categories.

1. **Web Content Mining**:

   Extracting valuable information from web content (e.g., web documents), often referred to as text mining.

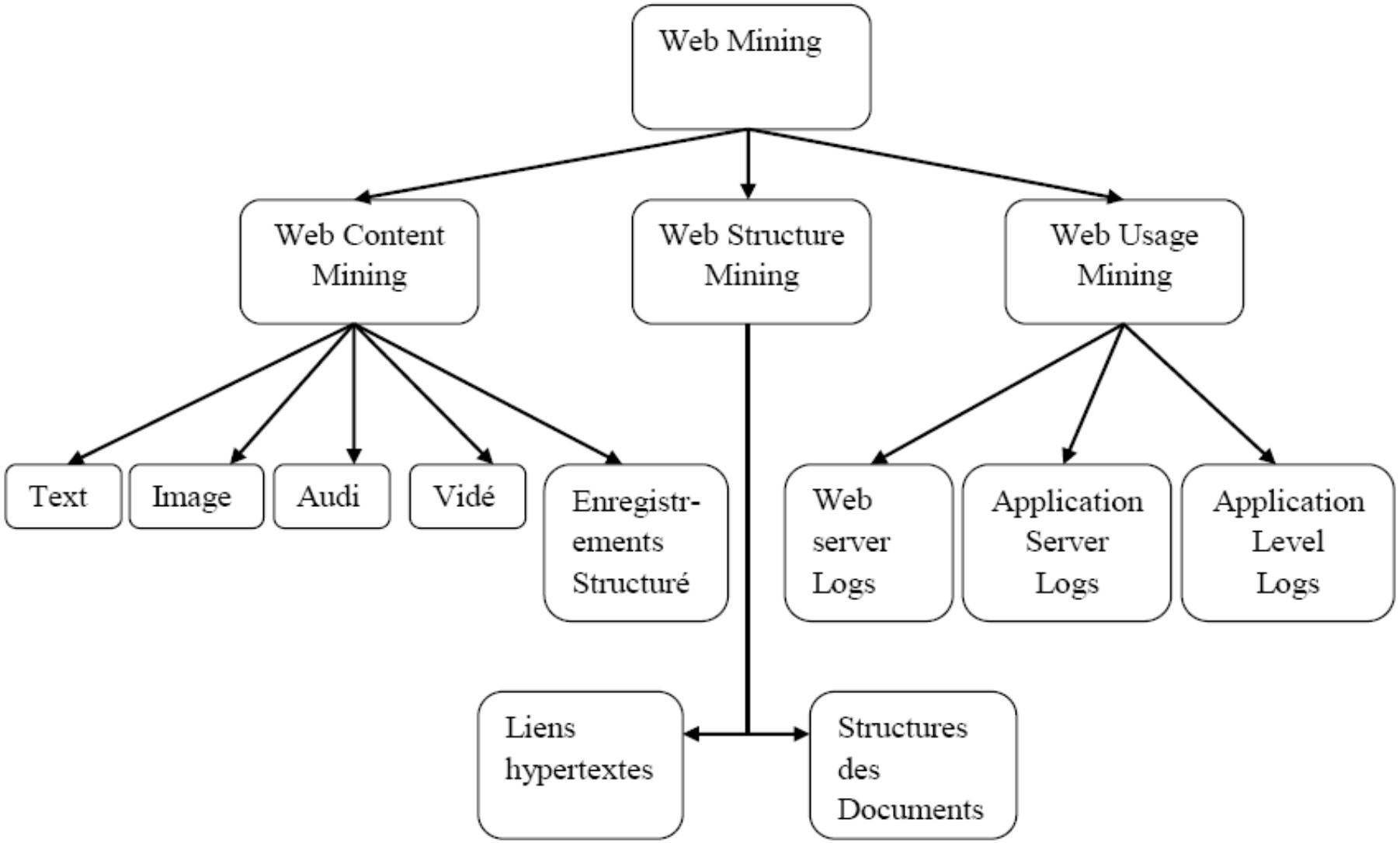
2. **Web Structure Mining**:

   Modeling websites based on their link structures to construct web communities or find relevant pages based on similarity or connectivity.

3. **Web Usage Mining**:

   Analyzing user access patterns from web transaction data or user session data recorded in web log files.

Categories of Web Mining



Web Community

- A web community is an aggregation of web objects, such as web pages or users, where each object is linked to others within a defined distance or relationship. Unlike traditional database management, which relies on predefined data models and schemas, web communities provide a dynamic and flexible method of organizing web objects. This approach enhances information retrieval and supports various applications.
- Analyzing web communities has several practical uses, including improving information retrieval, search engine optimization (SEO), spam detection, website categorization, and understanding web structure.
- **Example**: A collection of web pages where all members share a similar hyperlink topology that connects them to a specific page.

Social Networks

- With the advent of Web 2.0, an increasing number of advanced web services and applications have emerged, allowing users to easily create and distribute content while sharing information within a collaborative environment. Web 2.0's core components include web communities and hosted services such as social networking sites, wikis, and folksonomies. These platforms are characterized by open communication, decentralized authority, and the freedom to share and manage content independently.
- These enhanced Web 2.0 features empower users to effortlessly share and locate content, collaborate, engage socially, and manage knowledge in a flexible, self-organized manner.

Social Network Analysis (SNA)

- SNA involves studying and analyzing relationships, interactions, and communications within social networks. It can be implemented through visualization (sociograms) and mathematical analysis (graph theory). SNA helps in understanding and leveraging the social structure for various applications such as information retrieval, search engine optimization, spam detection, website categorization, and web structure analysis.

End of chapter 01