

Chapter 01:

Introduction to Data Mining and Web Mining

Overview

Chapter 1: Introduction: Data Mining and Web Mining

- ▶ 1. Data Mining
- ▶ 2. Objectives
- ▶ 3. Importance of Data Mining
- ▶ 4. Web Mining
- ▶ 5. Web Mining Context
- ▶ 6. Web Mining Categories
- ▶ 7. Data Mining Process Steps

Foundational Web Concepts: Internet & Protocols

The Underlying Infrastructure

- **Internet:**
 - A global, interconnected network of networks.
 - The underlying infrastructure for billions of communicating devices.
- **TCP/IP (Transmission Control Protocol/Internet Protocol):**
 - Foundational suite of communication protocols for data transfer.
 - **IP:** Addresses and routes data packets.
 - **TCP:** Ensures reliable, ordered, error-checked data delivery.
- **HTTP (Hypertext Transfer Protocol):**
 - Application-layer protocol for distributed, collaborative, hypermedia systems.
 - Defines how web servers and browsers communicate.

Foundational Web Concepts:

Web Resources

The Information Layer

- **Web (World Wide Web):**
 - An information system built *on top* of the Internet.
 - Allows access to docts and resources via URLs, interconnected by hyperlinks.
- **URL (Uniform Resource Locator):**
 - A specific address for locating a resource on the Internet (e.g., <https://www.example.com/page.html>).
- **Hyperlinks:**
 - References to data that users can directly follow (e.g., clicking text/images).
 - The "glue" connecting documents, forming the web's graph structure.
- **HTML (Hypertext Markup Language):**
 - Standard markup language for creating web pages.
 - Uses tags to structure content and embed resources, including hyperlinks.

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.

Goal : The main goal of data mining is to transform raw data into actionable information by identifying meaningful relationships, predicting future outcomes, and providing a deeper understanding that helps organizations make informed, strategic decisions.

Key Characteristics of Data Mining

- **Hidden Pattern Discovery:** Finds non-obvious relationships and trends not visible through simple queries.
- **Predictive Power:** Builds models to forecast future behaviors, risks, or opportunities.
- **Descriptive Power:** Summarizes and explains data by identifying typical behaviors and deviations.
- **Large Dataset Handling:** Efficiently processes massive volumes of data from gigabytes to petabytes.
- **Iterative Process:** Involves repeated cycles of preparation, modeling, and evaluation for refinement.

Data Characteristics in Data Mining: Overview

Structuring Information for Analysis

- **Data Table:** A collection of related data in a structured format (like a spreadsheet).
 - Organized into rows and columns.
- **Records (Rows/Instances/Observations):**
 - Each row represents a single, complete entity or observation (e.g., a unique customer).
- **Attributes/Features (Columns/Variables):**
 - Each column represents a specific characteristic, property, or measurement of the records (e.g., Age, Income).
- **Target (Dependent Variable/Class Label):**
 - The specific attribute the data mining model aims to predict or classify (e.g., Will_Respond_to_Offer).

Data Characteristics: Numerical Data

Measurable Quantities

- **Numerical (Quantitative) Data:** Represents measurable quantities, expressed as numbers.
 - **Discrete:**
 - Values can only take on specific, distinct integer values.
 - Often results from counting.
 - **Examples:** Number of clicks on a webpage, count of items in a shopping cart.
 - **Continuous:**
 - Values can take on any value within a given range.
 - Often results from measurements.
 - **Examples:** Person's height, temperature readings, duration of a web session.

Data Characteristics: Categorical Data

Labels and Categories

- **Categorical (Qualitative) Data:** Represents categories or labels, rather than numerical measurements.
 - **Nominal:**
 - Categories without any inherent order or ranking.
 - You cannot meaningfully say one is "greater" than another.
 - **Examples:** Colors (red, blue), marital status (single, married), operating systems (Windows, macOS).
 - **Ordinal:**
 - Categories with a meaningful order or ranking.
 - Differences between categories may not be uniform or precisely measurable.
 - **Examples:** Education levels (high school, bachelor's), customer satisfaction (poor, fair, good), clothing sizes (S, M, L).

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...")

Importance of Data Mining

Driving Modern Advancements

- **Competitive Advantage:** Gaining an edge by understanding customers and optimizing operations.
- **Informed Decision-Making:** Shifting from intuition to data-driven strategies.
- **Cost Reduction:** Identifying inefficiencies, predicting failures, detecting fraud.
- **Revenue Growth:** Targeted marketing, personalized recommendations, increased sales.
- **Risk Management:** Detecting fraud, cyber threats, assessing credit risk.
- **Scientific Discovery & Research:** Accelerating breakthroughs in various fields.
- **Personalization & User Experience:** Creating engaging and relevant digital interactions.

Web Mining

Definition :

Web mining is the application of data mining techniques to analyze and extract valuable knowledge from web data, which includes web page content, hyperlink structures, and user interaction logs. It aims to handle the complexity, scale, and heterogeneity of information available on the World Wide Web.

Goal :

The primary goal of web mining is to uncover patterns and insights from the vast amount of web data in order to improve search results, personalize user experiences, optimize website design, support business intelligence, and detect fraudulent or malicious activities online.

Web Mining Objectives (Part 1)

Enhancing Web Experience and Business

- **Improving Web Search:**

- Enhance relevance, quality, and speed of search engine results.
- Understanding user queries, effective content indexing, and page ranking.

- **Personalization:**

- Tailoring web content, advertisements, or recommendations to individual users.
- Based on inferred preferences, browsing history, demographics.

- **Website Design & Optimization:**

- Improve usability, information architecture, and overall design.
- Analyzing user navigation patterns, popular pages, and abandonment points.

Web Mining Objectives (Part 2)

Security, Intelligence, and Knowledge

- **Business Intelligence:**

- Gain deeper understanding of market trends, competitive landscapes, customer sentiments.
- Analyzing public web data (social media, competitor sites).

- **Security & Fraud Detection:**

- Identify malicious activities, spam, phishing attempts, or fraudulent transactions.
- Analyzing patterns in website content, link structures, and user behavior.

- **Knowledge Discovery:**

- Extract structured knowledge from unstructured web text and multimedia.
- Building ontologies, populating knowledge bases, discovering factual relationships.

Context of Web Mining

The Unique Environment of the Web

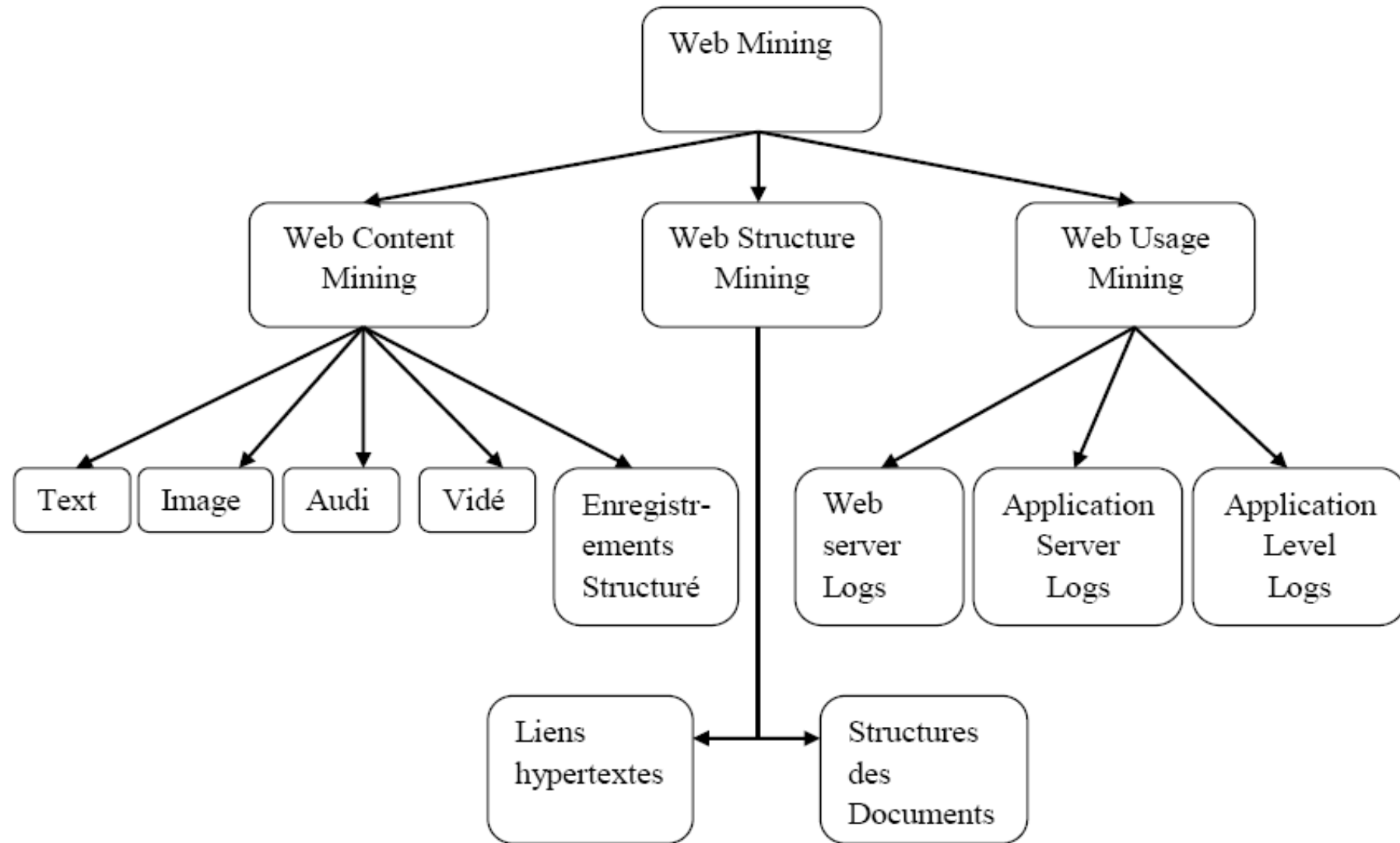
- **Massive Scale:** Billions of documents, constantly growing, challenging storage and processing.
- **Heterogeneity:** Diverse formats (HTML, XML, JSON, text, multimedia), requiring versatile tools.
- **Distributed Nature:** Data spread across countless global servers, requiring robust collection.
- **Dynamic Content:** Websites constantly updated, new pages added/removed, requiring continuous adaptation.
- **Link Structure:** Intricate hyperlink graph conveying relationships, authority, and relevance.
- **User Interaction:** Rich behavioral data (clicks, searches, form submissions) for usage analysis.

Web Data Characteristics

Specific Challenges and Opportunities

- **Semi-structured & Unstructured Nature:** Requires advanced NLP and information extraction.
- **Vastness & Dynamism:** Demands scalable and adaptive crawling/mining systems.
- **Linkage & Interconnectedness:** Provides unique relational information for link analysis.
- **Heterogeneity & Multimedia:** Requires integration of text, image, video, audio mining techniques.
- **Noise & Spam:** Significant irrelevant, low-quality, or malicious content requiring filtering.
- **Temporal Aspect:** Capturing and analyzing rapid changes in trends and relevance over time.
- **User-Generated Content (UGC):** Valuable for public opinion and trends, but challenging due to informal language and bias.

Categories of Web Mining



Categories of Web Mining

Three Main Approaches

- **Web Content Mining:**
 - **Focus:** Content of web pages (text, images, multimedia).
 - **Goal:** Extract knowledge directly from content.
 - **Examples:** Search engines, sentiment analysis, topic modeling.
- **Web Structure Mining:**
 - **Focus:** Hyperlink structure (relationships between pages).
 - **Goal:** Understand authority, relevance, and connectivity.
 - **Examples:** PageRank, HITS, web community detection.
- **Web Usage Mining:**
 - **Focus:** User interaction data (clickstreams, logs).
 - **Goal:** Discover patterns in user behavior, preferences.
 - **Examples:** Personalization, website optimization, fraud detection.

Steps of Applying Data Mining

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

End of Chapter 01

Questions & Discussion