Module 4 : Web Usage Mining

With the rapid expansion of **e-commerce, Web services, and Web-based systems**, the volume of **clickstream, transaction, and user profile data** collected has reached immense levels. Analyzing this data helps organizations improve **customer value, marketing strategies, campaign effectiveness, Web application performance, personalized content delivery, and website structure**.

Web Usage Mining (WUM) is the process of **automatically discovering patterns** in user interactions with Web resources. It focuses on **modeling and analyzing user behavior** through frequently accessed pages, objects, and resources.

The Web usage mining process follows the **standard data mining methodology**, consisting of three key stages:

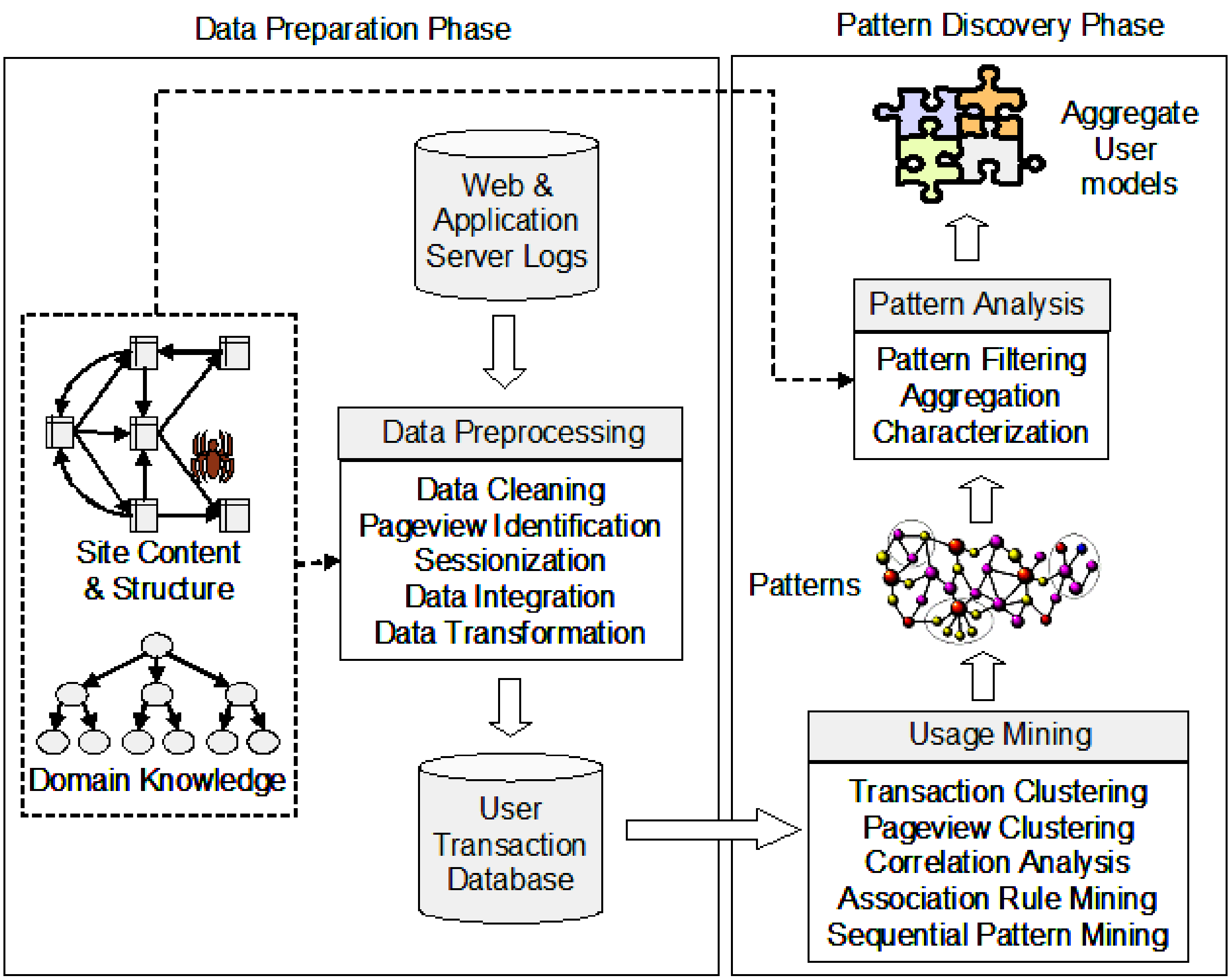
1. **Data Collection and Preprocessing**

   ○ Clickstream data is **cleaned, segmented, and structured** into user transactions.
   ○ Additional knowledge from **site content, structure, and ontologies** can be integrated.
2. **Pattern Discovery**

   ○ Techniques such as **statistical analysis, machine learning, and database operations** identify user behavior patterns.
   ○ Insights include **common browsing paths, session trends, and resource usage statistics**.
3. **Pattern Analysis and Applications**

   ○ The discovered patterns are **filtered, processed, and modeled** into **user profiles**.
   ○ Applications include **recommendation engines, analytics, visualization tools, and reporting systems**.

Additionally, specialized areas like **recommendation systems, query log mining (QLM), and Ad click mining** are explored for more targeted insights



4.1 Data Collection and Pre-Processing

1. Introduction

- The first step in web usage mining is **data collection and preparation** to create a structured dataset for analysis.
- Web data comes from **multiple sources** (clickstream logs, databases, external APIs) and must be **pre-processed** before applying data mining techniques.
- Pre-processing is **time-consuming and computationally expensive** but essential for extracting meaningful patterns.
- This process includes:
  - **Cleaning raw data**

  - **Integrating multiple data sources**
  - **Transforming data into structured formats**

---

2. Key Pre-Processing Tasks

Web usage mining requires several **pre-processing** tasks to improve data quality:

a) Data Cleaning & Fusion

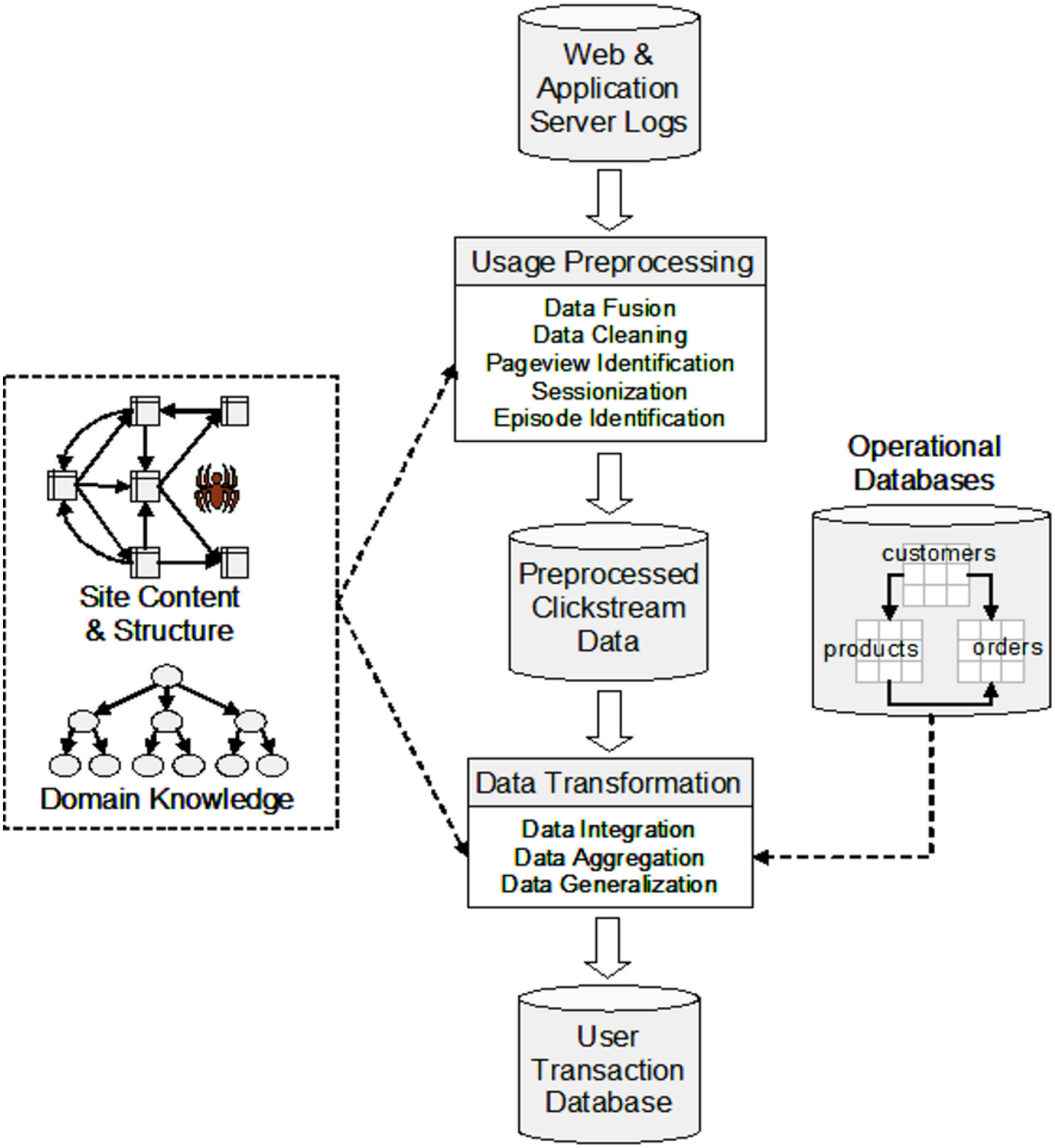
- Removes unnecessary entries like:
  - Requests for images, CSS, JavaScript
  - Duplicate entries
  - Corrupt or incomplete data
- **Example:** Filtering out non-relevant HTTP requests from web logs.

b) User and Session Identification

- **Goal:** Identify unique users and their sessions.
- Methods:
  - Using **IP address, cookies, browser agent**
  - Identifying session timeouts (e.g., **30 minutes of inactivity starts a new session**)
- **Example:**
  - User **U001** visits **/home** at **10:00 AM**, then visits **/product1** at **10:05 AM**.
  - If there is no activity for **35 minutes**, a **new session** starts.

c) Pageview Identification

- Groups multiple HTTP requests into **meaningful pageviews** (i.e., a single user action like opening a product page).
- **Example Transformation:**
  - Raw URLs → Categorized views
  - `/product/123?ref=ad` → **Product Page**
  - `/cart` → **Shopping Cart**
  - `/checkout?user=xyz` → **Checkout Page**



4.1.1 Sources and Types of Data

Web mining data comes from **multiple sources**, categorized into four types:

1. Usage Data (Web Logs)

- Captured in **server access logs** from web & application servers.

- Contains:

    - **Timestamp**
    - **IP Address**
    - **Requested Resource (URL)**
    - **HTTP Status Code**
    - **Browser & OS Information**
    - **Referring URL (Previous Page Visited)**

Example Web Log Entry:

```
2025-03-08 10:00:00 1.2.3.4 - GET /product/123 - 200
HTTP/1.1 mywebsite.com Mozilla/5.0 (Windows NT 10.0; Win64; x64)
Chrome/110.0
http://google.com/search?q=best+laptops
```

-
    - **1.2.3.4** → User's IP address
    - **GET /product/123** → Requested product page
    - **http://google.com/...** → Came from a Google search

---

2. Content Data

- Represents **the actual content of the website** (HTML, images, text, videos).

- Includes:

    - Static pages (HTML/XML)
    - Dynamic content from scripts
    - Metadata (keywords, semantic tags, document attributes)
    - Ontologies (product categories, relationships)
- **Example:**

    - A **product page** containing a **product image, description, and price**.

---

3. Structure Data

- Represents **how pages are linked together** through hyperlinks.
- Includes:
  - **Inter-page structure** (site navigation, menus)
  - **Intra-page structure** (HTML tree structure, XML elements)
- Helps understand **how users navigate the site**.
- **Example:**

Site Map:

```
Home → Products → Laptops → Dell XPS 15
```

  -

---

4. User Data (Profiles & Demographics)

- Contains information about **users and their preferences**.
- Sources:
  - **User registration data (name, age, location)**
  - **Purchase history**
  - **Clickstream behavior**
  - **Cookies (anonymous tracking)**
- **Example:**
  - A user profile for **John Doe** may include:
    - **Age:** 25
    - **Location:** USA
    - **Interest:** Laptops, Gaming

Conclusion

- **Data collection and pre-processing** form the backbone of **Web Usage Mining**.
- Without proper **cleaning, session identification, and integration**, the quality of insights decreases.
- **Pre-processed data** helps in:
    - **User behavior analysis**
    - **Personalized recommendations**
    - **Marketing insights & website optimization**

Example: Server Log Entry (Raw Data) → Processed Data

Raw Server Log Entry (Before Cleaning)

```
2025-03-08 10:00:00 1.2.3.4 - GET /product/123 - 200
HTTP/1.1 mywebsite.com Mozilla/5.0 (Windows NT 10.0; Win64; x64)
Chrome/110.0
http://google.com/search?q=best+laptops
```

- **1.2.3.4** → User's **IP Address**
- **GET /product/123** → User **requested product page**
- http://google.com/search → User came from **Google Search**
- **Mozilla/5.0 (Windows NT 10.0; Win64; x64)** → **Browser & OS**

Processed Data (After Cleaning & Transformation)

Timestamp	IP Address	Session ID	Pageview	Referrer	Device
2025-03-08 10:00:00	1.2.3.4	S1001	Product Page	Google Search	Windows 10
2025-03-08 10:02:15	1.2.3.4	S1001	Add to Cart	Product Page	Windows 10
2025-03-08 10:04:30	1.2.3.4	S1001	Checkout	Add to Cart	Windows 10

🔹 **Cleaning Removes:**
✔ Irrelevant hits (CSS, JS files)
✔ Duplicate entries
✔ Bot traffic

🔹 **Session ID Assignment:**

- If the same **IP and user agent** interact within **30 minutes**, it's part of **one session**.
- If inactive for **30+ minutes**, it starts a **new session**.

---

3. Data Types in Web Usage Mining (Real-World Example)

Data Type	Description	Example
Usage Data	User behavior from server logs	Clickstream logs, timestamps, URLs visited
Content Data	Page content, keywords, metadata	Product descriptions, blog articles
Structure Data	Website hierarchy, navigation paths	Sitemap, menu structure
User Data	User profile, preferences	Name, age, purchase history

Example Use Case: E-commerce Website

If an **online shopping website** wants to understand **customer behavior**, they can analyze:
✔ **Usage Data:** Which products users click on the most?
✔ **Content Data:** What keywords (metadata) attract customers?
✔ **Structure Data:** Do users follow recommended product links?
✔ **User Data:** What demographic shops the most?

By **combining these insights**, the website can:
✔ Improve **recommendations**
✔ Optimize **page layouts**
✔ Target **specific user segments**

4.1.2 Data Modeling for Web Usage Mining –

1. Introduction to Data Modeling for Web Usage Mining

Web Usage Mining (WUM) is the process of extracting useful patterns from web access logs, user interactions, and browsing behavior. **Data modeling** plays a crucial role in structuring and representing this data for analysis.

Data modeling helps:
✔ Identify **user navigation patterns**
✔ Understand **user behavior**
✔ Improve **website structure and recommendations**

---

2. Components of Web Usage Data

Before applying mining techniques, web usage data is structured into:

(a) Pageviews (P)

A **pageview** represents a unique web entity (e.g., a webpage, product page, or action).

$P=\{p1,p2,...,pn\}$

Each pageview is an element in the dataset.

✔ Example:
If a website has **3 pages** (Home, Products, Contact), then:

$P=\{Home,Products,Contact\}$

(b) User Transactions (T)

A **user transaction** is a collection of pageviews visited in a session.

$T=\{t1,t2,...,tm\}$

Each transaction contains **pageviews and their weights**:

$t=\{(p1,w(p1)),(p2,w(p2)),...,(pl,w(pl))$

where **w(p)** represents the weight/significance of a pageview in a session.

✔ Example Transaction:
A user browses **Home → Products → Contact**, spending **5, 10, and 3 seconds** on each page.

t1={(Home,5),(Products,10),(Contact,3)}

---

3. Weighting Methods for Pageviews

The importance of a pageview **w(p)** is determined by different weighting schemes:

(a) Binary Weighting

- 1 → If the page is visited
- 0 → If the page is not visited

✔ Example:
If a user visits **Home and Products but not Contact**, then:

t1={(Home,1),(Products,1),(Contact,0)}

(b) Time-Based Weighting

- **w(p) = Time spent on a page (normalized)**
- If **last page duration is missing**, use **mean time** across sessions.

✔ Example:
A user spends **5, 10, and 3 seconds** on pages, but last page time is missing.
If the average time spent on "Contact" (last page) is **4 seconds**, then:

t1={(Home,5),(Products,10),(Contact,4)}

(c) Collaborative Filtering Weighting

- Uses **user ratings** or interaction scores.
- Example: Amazon/Netflix recommendation systems.

✔ Example:
A user rates products on a scale of 1-5:

t1={(Laptop,5),(Phone,4),(Tablet,3)}

---

4. User-Pageview Matrix (UPM)

A **User-Pageview Matrix (UPM)** is an **m×n matrix** where:

- **Rows** represent users (or sessions)
- **Columns** represent pageviews
- **Values** represent time spent (or other weights)

✔ Example: User-Pageview Matrix (UPM)

Users	Home	Products	Contact
user1	15	5	0
user2	12	0	56
user3	9	47	0

📌 **Use Cases:**
✔ Clustering similar users
✔ Identifying high-traffic pages

---

5. Content-Enhanced Transactions (Integrating Content & Usage Data)

To improve analysis, **semantic content** (page topics) is integrated with usage data.

(a) Pageview-Feature Matrix (PFM)

Each pageview is represented as a **vector of features** (topics, keywords).

✔ Example: Term-Pageview Matrix (PFM)

Terms	Home	Products	Contact
Shopping	1	1	0
Support	0	0	1
Offers	1	1	0

This shows:
✔ **Home & Products pages** contain **shopping-related content**
✔ **Contact page** contains **support-related content**

(b) Creating a Content-Enhanced Transaction Matrix (TFM)

By **multiplying UPM and PFM**, we get:

✔ **Example: Transaction-Feature Matrix (TFM)**

Users	Shopping	Support	Offers
user1	20	0	15
user2	12	56	12
user3	47	0	9

📌 **Now, transactions reflect user interests based on topics, not just pages**.

---

6. Applications of Data Modeling in Web Usage Mining

🔹 Clustering Users for Segmentation

- **K-means clustering** groups users based on browsing patterns.
- Example: Identifying **"Tech Enthusiasts"**, **"Casual Shoppers"**, etc.

✔ **Example: Clustering Users by Behavior**

Users	Shopping Score	Support Score	Cluster
user1	20	0	Shopper
user2	12	56	Support Seeker
user3	47	0	Shopper

🔹 Association Rule Mining (Recommendations)

- **Example Rule:**
  - **{“Laptop”, “Phone”} → {“Tablet”}**
  - Suggests that users who buy laptops & phones may also buy a tablet.

✔ Example: Amazon’s **“Customers who bought this also bought”**.

🔹 Sequential Pattern Mining (User Journeys)

- Identifies **common navigation paths**
- Example: Users often visit **"Homepage → Product Page → Checkout"**

✔ Example: Pattern extracted from logs:

(Home)→(Products)→(Checkout)

7. Conclusion

✔ **User-pageview matrices** help in analyzing web behavior.
✔ **Weighting techniques** improve personalization & clustering.
✔ **Content-enhanced transactions** provide **semantic insights**.
✔ **Clustering, pattern mining, and recommendations** improve user experience.

4.1.3 Session and Visitor Analysis

🔹 Overview

Session and Visitor Analysis involves statistical and analytical techniques applied to **aggregated web session data**. These insights help website administrators understand visitor behavior, optimize performance, and support business strategies.

✔ **Key Aggregation Units**:

- **Days** (Daily traffic trends)
- **Sessions** (A single browsing instance)
- **Visitors** (Unique users over time)
- **Domains** (Organization-level analysis)

🔹 Common Metrics in Session & Visitor Analysis

Commercial **Web Log Analysis Tools** (e.g., Google Analytics, Adobe Analytics) generate reports on the following key metrics:

Metric	Definition
Most Frequently Accessed Pages	Pages with the highest visits.
Average Page View Time	Time spent on a single page.
Average Session Duration	Total time spent per session.
Common Entry Points	First page accessed in a session.
Common Exit Points	Last page visited before leaving.
Bounce Rate	Percentage of single-page visits.
Path Analysis	Common navigation routes taken by users.

📌 **Example:**
E-commerce site analysis might reveal that:
✔ The **"Homepage** → **Product Page** → **Cart"** is the most common path.

✔ Users spend **5 minutes on average** on the "Product Page."
✔ The **"Checkout Page"** has a **high exit rate**, indicating **potential issues**.

🔹 Importance of Session Analysis

Although statistical session analysis lacks deep insights, it provides valuable knowledge for:

✔ **Optimizing website performance**
✔ **Improving user experience** (UX)
✔ **Marketing decision-making** (e.g., ad placement)
✔ **Identifying usability issues** (e.g., high bounce rates)

📌 **Example Use Case:**
A **news website** finds that most users **drop off after reading headlines**. The company **redesigns the homepage** to encourage deeper engagement.

Online Analytical Processing (OLAP)

🔹 Overview

OLAP enables **multidimensional analysis** of web usage data using an **integrated data warehouse**. It allows analysts to **drill down, roll up, slice, and dice** data across different dimensions.

✔ **Key Dimensions in OLAP for Web Usage Mining**:

- **Time** (Hourly, daily, weekly trends)
- **Domain** (User's location or organization)
- **Requested Resource** (Web pages, products, services)
- **User Agent** (Browser, device type)
- **Referrer** (Source of traffic)

🔹 Multidimensional Data Warehouse for OLAP

OLAP systems store web usage data in a **data warehouse**, integrating:
✔ **Web logs** (usage patterns)
✔ **Content data** (page structures, topics)
✔ **E-commerce transactions** (sales, cart interactions)

📌 **Example: Analyzing E-commerce Data**
An **e-commerce company** integrates web usage with sales data:

Metric	Product A	Product B	Product C
Views (Web Logs)	10,000	8,000	5,000
Add to Cart (%)	5%	8%	3%
Purchase Rate	2%	6%	1%

👉 The company sees **Product B has a high add-to-cart and purchase rate**, suggesting **better demand**. **Product C needs improvement** in conversion.

🔹 OLAP Operations in Web Usage Analysis

OLAP Operation	Description	Example Use Case
Drill Down	View data at a **detailed level**.	Zoom into hourly website traffic.
Roll Up	Aggregate data at a **higher level**.	View total weekly or monthly visits.
Slice	Extract data for a **specific dimension**.	Analyze **mobile traffic only**.
Dice	Analyze a **subset of multiple dimensions**.	Study traffic for **mobile users in India**.

📌 **Example:**
A retail website wants to **analyze holiday season trends**. Using **OLAP**, they:
✔ **Drill down** into **hourly traffic spikes on Black Friday.**

✔ **Slice** data to focus on **tablet users only.**
✔ **Dice** data to see **tablet users from the U.S. on Black Friday.**

Benefits of OLAP for Web Usage Mining

✔ **Flexible Data Exploration** – Analysts can **interactively** query data.
✔ **Multidimensional Analysis** – Insights from **time-based, user-based, and content-based** perspectives.
✔ **Integration with Business Metrics** – E-commerce data improves **customer segmentation & recommendations.**
✔ **Supports Advanced Data Mining** – OLAP output **feeds into machine learning models** for **predictive analytics**.

4.1.4 Clustering Analysis and Visitor Segmentation in Web Usage Mining

Introduction

Clustering is a fundamental data mining technique used in Web Usage Mining to group similar objects based on their characteristics. In web analytics, clustering is applied to:

- **User Clustering:** Groups visitors with similar browsing patterns.
- **Page Clustering:** Groups web pages based on common user access patterns.

These clusters help in personalizing user experiences, market segmentation, business intelligence, and improving website structure.

---

1. User Clustering (Visitor Segmentation)

User clustering involves grouping visitors into segments based on their browsing behavior. This segmentation can be used for:

- Inferring **user demographics** for targeted marketing.
- Providing **personalized content** to users with similar interests.
- Identifying **trends and preferences** of different user groups.

Clustering Algorithms for User Segmentation

(i) K-Means Clustering

- Users are represented as vectors in a **multi-dimensional space**, where each dimension represents a page view.
- The **K-means algorithm** partitions users into K clusters based on their browsing similarity.
- Each user cluster represents a **segment of visitors** with similar behavior.

✅ **Example:**
Consider a website that sells electronics. Based on user sessions, clustering may identify:

- **Cluster 1:** Users interested in smartphones and accessories.
- **Cluster 2:** Users browsing laptops and gaming devices.

- **Cluster 3:** Users interested in home appliances.
  By analyzing these segments, businesses can create personalized offers for each group.

(ii) Aggregate User Profiles

- Each cluster can be summarized using a **centroid vector**, where the value of each pageview dimension indicates its significance in the cluster.
- **Formula for weight calculation:**

$\text{weight}(p, cl)=\sum w(p,s) \mid cl \mid$

Where:

- $w(p,s)$ is the weight of page p in a transaction s.
- $\mid cl \mid$ is the total number of transactions in cluster cl

✅ **Example:**
A clustered visitor segment may have the following **aggregate profile**:

Pageview	Weight
B (Smartphones)	1.00
F (Accessories)	1.00
A (Laptops)	0.75
C (Cameras)	0.25

From this, we infer that users in this segment strongly prefer smartphones and accessories. If a new user visits **A (Laptops) and B (Smartphones)**, they may also be interested in **F (Accessories)**, which can be recommended.

---

2. Page Clustering

Page clustering is used to group similar web pages based on:

1. **User access patterns** – Pages frequently visited together.

2. **Content similarity** – Pages with similar keywords or categories.

Techniques for Page Clustering

(i) Usage-Based Page Clustering

- Uses web log data to identify **pages that are frequently accessed together**.
- Helps in **recommendation systems** and improving website navigation.
- Can be used to create **dynamic HTML pages** suggesting related links.

✅ **Example:**
If many users visiting a product page **(Smartphone X)** also visit a comparison page **(Smartphone X vs Y)**, then the website can automatically suggest related links.

(ii) Content-Based Page Clustering

- Pages are grouped based on **keywords, meta-tags, or product attributes**.
- Helps in organizing **similar topics or products**.

✅ **Example:**
A news website may use clustering to group articles into:

- **Cluster 1:** Sports News (Football, Cricket, Tennis)
- **Cluster 2:** Technology News (AI, Gadgets, Software Updates)
- **Cluster 3:** Political News (Elections, Policies, Global Relations)

---

3. Advanced Clustering Techniques

Traditional clustering methods like K-Means have limitations (e.g., sensitivity to initial conditions). To address this, **probabilistic models** are used:

(i) Mixture Models (e.g., Gaussian Mixture Model, Markov Models)

- Assume users belong to **multiple behavioral groups** with different probabilities.
- Example: A user may exhibit shopping behavior (cluster 1) and also browse blogs (cluster 2).

(ii) Probabilistic Latent Semantic Analysis (PLSA)

- Models hidden relationships between users and pages.
- Reduces **overfitting** issues in simpler clustering models.

- Uses **Expectation-Maximization (EM) algorithm** to estimate probabilities.

✅ **Example:**
A streaming platform uses PLSA to analyze user preferences, discovering that:

- **Cluster A:** Users who watch action movies also prefer thrillers.
- **Cluster B:** Users who watch documentaries also like science fiction.
  Using this, it can recommend relevant content to new users.

---

4. Applications of Clustering in Web Usage Mining

- **Personalization & Recommender Systems:** Suggesting relevant products or content.
- **Market Segmentation:** Identifying key customer groups for targeted marketing.
- **Website Optimization:** Improving navigation by linking frequently accessed pages.
- **Fraud Detection:** Identifying unusual browsing patterns that may indicate fraudulent activities.

Conclusion

Clustering is a powerful tool in **Web Usage Mining**, enabling businesses to **understand visitor behavior, segment users, and optimize website design**. Whether through **K-Means, probabilistic models, or PLSA**, clustering helps provide a **personalized web experience and valuable business insights**.

Detailed Example: Clustering Analysis and Visitor Segmentation in Web Usage Mining

To better understand **clustering analysis and visitor segmentation**, let’s go through a detailed real-world example of how a **retail e-commerce website** (such as Amazon or Flipkart) applies these techniques.

Scenario: Analyzing Visitor Behavior for an E-commerce Website

Objective:
The website wants to **segment visitors** based on browsing behavior to improve recommendations and marketing strategies.

Data Collected:
From web server logs, the website records:

- **User Sessions (Transactions):** Series of pages visited in one session.
- **Time Spent on Pages:** To measure user engagement.
- **Clickstream Data:** Sequence of clicks a user makes.

✅ **Example Web Log Data (Sample of 5 Users)**

User ID	Pages Visited (Sequence)	Duration (Minutes)	Product Views	Purchase?
U1	Home → Mobiles → iPhone 15 → Reviews → Checkout	15	3	Yes
U2	Home → Mobiles → Samsung S23 → iPhone 15	10	2	No
U3	Home → Laptops → Dell Inspiron → Checkout	20	1	Yes
U4	Home → Mobiles → Samsung S23 → Reviews	12	3	No
U5	Home → Laptops → MacBook Air → Checkout	18	1	Yes

Step 1: Applying Clustering Algorithm (K-Means)

We represent each user as a **feature vector** based on:

- **Number of visits to product categories (Mobiles, Laptops, Accessories, etc.)**
- **Total time spent on site**
- **Whether they made a purchase (1 = Yes, 0 = No)**

Feature Matrix Representation

User ID	Mobile Category Visits	Laptop Category Visits	Avg. Time Spent	Purchase (1/0)
U1	3	0	15	1
U2	2	0	10	0
U3	0	2	20	1
U4	3	0	12	0
U5	0	2	18	1

Now, we apply **K-Means clustering** (with **K=2**) to segment users into groups.

Step 2: Identifying Clusters

After running K-Means, we obtain two clusters:

Cluster	Description
Cluster 1 (Tech Enthusiasts)	Users who visit **laptop categories**, spend **more time**, and mostly **make a purchase**.
Cluster 2 (Window Shoppers)	Users who visit **mobile categories**, spend **less time**, and mostly **don't make a purchase**.

Step 3: Analyzing Clusters and Business Insights

Cluster 1: Tech Enthusiasts (Laptop Buyers)

- These users are **high-value customers** who browse **laptops** and make purchases.
- They tend to spend more time reading product descriptions before buying.
- **Business Strategy:**
  - Send them **personalized discounts on laptops**.
  - Show them **premium laptop accessories** (cooling pads, bags, extended warranties).

Cluster 2: Window Shoppers (Mobile Browsers)

- These users browse **mobile phones** but do not make a purchase.
- They **spend less time** compared to Cluster 1.
- **Business Strategy:**
  - Offer them **limited-time discounts on smartphones**.
  - Use **retargeting ads** on social media to bring them back.
  - Send follow-up emails with **comparisons and reviews** to help decision-making.

Step 4: Page Clustering for Recommendation System

The website can also **group pages** that are frequently visited together using **page clustering techniques**.

✅ **Example: Page Access Patterns from Web Logs**

Page	Frequent Co-Accessed Pages
iPhone 15	iPhone 15 Reviews, Samsung S23, Checkout
Samsung S23	Samsung S23 Reviews, iPhone 15
MacBook Air	Laptop Accessories, Dell Inspiron

By clustering pages, we can create **better recommendations**:

- If a user visits **Samsung S23**, suggest **Samsung S23 Reviews** and **iPhone 15** as alternatives.

- If a user adds **MacBook Air** to the cart, recommend **laptop accessories**.

Final Business Benefits of Clustering Analysis

Benefit	Explanation
Targeted Marketing	Each user segment receives personalized promotions.
Improved Recommendations	Users see relevant products based on their browsing behavior.
Increased Conversion Rates	Discounts and retargeting convert more window shoppers into buyers.
Better UX & Navigation	Page clustering helps show related content, improving user experience.

Conclusion

By using **clustering analysis**, the e-commerce website successfully **segments visitors** into meaningful groups, helping in:

1. **Personalized marketing** for different user behaviors.
2. **Optimized recommendations** using page clustering.
3. **Better customer retention** by targeting potential buyers with customized offers

4.1.5 Association and Correlation Analysis in Web Usage Mining

Overview

Association rule discovery and correlation analysis help identify patterns in web transactions, such as frequently accessed pages or commonly purchased products. These insights assist in:

- **Organizing website content efficiently**
- **Improving cross-sale recommendations**
- **Optimizing user navigation**

Key Concepts

1. **Association Rule Discovery**
    - Finds relationships between items frequently accessed or purchased together.
    - Uses **Apriori Algorithm** to detect frequent itemsets.
    - Generates association rules based on **minimum support** and **confidence** thresholds.
2. **Association Rule Format**

    $$X \rightarrow Y[\text{sup},\text{conf}]$$

    - **Support (sup)**: Probability that **X** and **Y** appear together in transactions.
    - **Confidence (conf)**: Probability that **Y** appears given that **X** has already occurred.
    - Example Rule:

    $$\text{"special-offers/"} \rightarrow \text{"shopping-cart/"}$$

    Interpretation: Users who visit "special-offers" are likely to proceed to "shopping-cart".

3. **Correlation Analysis**
    - Examines statistical relationships between customers or visitors.
    - Helps in **personalization and recommender systems**.

Example of Association and Correlation Analysis in Web Usage Mining

Scenario: E-Commerce Website Optimization

A popular e-commerce website wants to analyze user behavior to improve product recommendations and website navigation.

---

1. Association Rule Example

Goal: Find frequently purchased product combinations

Dataset: Online Shopping Transactions

Transaction ID	Purchased Items
1	Laptop, Mouse, Keyboard
2	Phone, Charger, Earphones
3	Laptop, Mouse, USB Drive
4	Phone, Case, Screen Protector
5	Mouse, Keyboard
6	Laptop, Mouse
7	Phone, Charger
8	Laptop, Mouse, Keyboard
9	Phone, Charger, Case
10	USB Drive, Mouse

Discovered Association Rules

- **Rule 1:** If a customer buys a **Laptop**, they are **80% likely** to also buy a **Mouse**.

    - **(Support: 40%, Confidence: 80%)**
- **Rule 2:** If a customer buys a **Phone**, they are **90% likely** to also buy a **Charger**.
    - **(Support: 30%, Confidence: 90%)**
- **Rule 3:** Customers who buy a **Mouse** often buy a **Keyboard** together.
    - **(Support: 35%, Confidence: 70%)**

Business Insights:

📌 **Improve product recommendations** – Suggest **Mouse + Laptop** as a bundle.
📌 **Create combo offers** – Offer discounts on **Phone + Charger** to increase sales.

---

2. Correlation Analysis Example

Goal: Identify relationships between user behavior and website features

Dataset: User Behavior on a Shopping Website

User ID	Time Spent (minutes)	Pages Visited	Purchase Amount (₹)
101	5	3	0
102	15	7	2,000
103	20	10	3,500
104	8	5	1,000
105	25	12	5,000
106	3	2	0

Correlation Findings

🔹 **Time Spent on Site & Pages Visited → Strong Positive Correlation (0.85)**

- Users who spend more time tend to visit more pages.

🔹 **Pages Visited & Purchase Amount → Moderate Positive Correlation (0.65)**

- Users who explore more pages are likely to make a purchase.

🔹 **Time Spent & Purchase Amount → Weak Correlation (0.3)**

- Just spending time does not guarantee a purchase.

Business Insights:

✅ **Improve site engagement** – Add **personalized product recommendations** to increase page visits.
✅ **Optimize checkout process** – Reduce unnecessary steps to convert **long-time visitors into buyers**.

Conclusion

🔹 **Association Rules** identify commonly purchased items for better recommendations.
🔹 **Correlation Analysis** helps understand how user behavior impacts purchases.

📊 **Impact:** These insights help e-commerce businesses **boost sales, improve customer experience, and increase engagement**.

4.1.5 Analysis of Sequential and Navigational Patterns in Web Usage Mining

Introduction

Web usage mining aims to analyze user behavior by extracting patterns from web logs. Among various techniques, **sequential pattern mining** and **navigational pattern analysis** focus on understanding the **order of page visits** and how users navigate through a website. These insights help businesses improve website design, recommend relevant content, and enhance user experience.

---

1. Sequential Pattern Mining

What is Sequential Pattern Mining?

Sequential pattern mining identifies **frequent patterns** in which items (or pages) appear in a specific order across different sessions. Unlike association rule mining, which does not consider order, sequential pattern mining takes into account the **time-ordered sequence of events**.

Why is it Important?

- **Predicts User Behavior:** Helps websites understand how users move from one page to another.
- **Enhances Targeted Advertising:** Enables marketers to display relevant promotions based on previous navigation history.
- **Improves Web Structure:** Identifies common user paths, allowing developers to optimize page linking.

How Does It Work?

1. **Data Collection:** Web server logs record user interactions, storing page visits in time-ordered sequences.
2. **Preprocessing:** Data is cleaned by removing bot traffic, sessionizing user visits, and filtering irrelevant pages.
3. **Pattern Discovery:** Algorithms like **AprioriAll**, **GSP (Generalized Sequential Pattern Mining)**, or **PrefixSpan** extract frequently occurring sequences.

4. **Pattern Evaluation:** The extracted sequences are analyzed for support and confidence.

---

2. Navigational Pattern Analysis

What is Navigational Pattern Analysis?

Navigational pattern analysis examines how users move across a website, capturing **common paths** and **transition probabilities** between pages. This is often modeled using **Markov chains**, where each page is a state, and transitions represent the probability of moving from one page to another.

Key Applications:

- **Predicting Next Page Visit:** Used in web prefetching to reduce load times.
- **Identifying Bottlenecks:** Helps in detecting pages where users drop off frequently.
- **Personalization & Recommendations:** Suggests content or products based on previous navigation paths.

How Does It Work?

1. **Session Identification:** User sessions are created based on timestamps to determine sequential navigation paths.
2. **State Representation:** Each page is treated as a state in a Markov model.
3. **Transition Probability Calculation:** The probability of moving from one page to another is determined by counting occurrences in past sessions.
4. **Pattern Analysis:** High-probability paths and common drop-off points are identified for website improvements.

Example: E-Commerce Website Analysis

Scenario:

An online shopping website wants to analyze user behavior to optimize recommendations and improve checkout rates.

Dataset: (User Navigation Sequences)

Session ID	Page Sequence (Visited in Order)
1	Home → Laptops → Accessories → Checkout
2	Home → Mobiles → Accessories → Checkout
3	Home → Mobiles → Checkout
4	Home → Laptops → Checkout
5	Home → Accessories → Checkout

1. Sequential Pattern Mining Analysis

Using **PrefixSpan** or **AprioriAll**, we extract frequent patterns:

- **(Home → Laptops → Checkout) [Support: 40%]**
- **(Home → Mobiles → Accessories → Checkout) [Support: 30%]**
- **(Home → Accessories → Checkout) [Support: 20%]**

📌 **Insights:**

- Users frequently visit **Laptops** or **Mobiles** before checkout, suggesting these are high-interest categories.
- **Accessories** appear frequently before checkout, meaning cross-selling might be effective.
- A **direct Home → Checkout path is missing**, indicating users explore products before purchasing.

2. Navigational Pattern Analysis using Markov Model

Transition Probabilities:

From → To	Probability
Home → Laptops	40%
Home → Mobiles	30%

Home → Accessories	20%
Laptops → Checkout	50%
Mobiles → Checkout	60%
Accessories → Checkout	70%

📌 **Insights:**

- High transition probability from **Accessories → Checkout (70%)**, meaning accessories are frequently a last-step purchase.
- Users who visit **Laptops** have a **50% chance of purchasing**, suggesting targeted discounts could improve conversion.
- **Home → Accessories (20%)** is low, meaning better homepage visibility for accessories might increase sales.

Conclusion

- **Sequential Pattern Mining** reveals common page visit sequences, helping optimize site structure and recommend frequently purchased items together.
- **Navigational Pattern Analysis** using **Markov models** predicts user movement and helps enhance the user journey by improving transitions and reducing drop-offs.
- **Actionable Insights:**
  - Improve visibility of accessories on the homepage.
  - Offer bundle discounts for laptops & accessories.
  - Optimize checkout flow for faster navigation.

By leveraging **sequential and navigational analysis**, businesses can significantly enhance user engagement and boost conversions. 🚀

Classification and Prediction Based on Web User Transactions

Classification is a process of assigning data into predefined categories based on patterns and extracted features. In web usage mining, classification helps in understanding user behavior, predicting future actions, and personalizing user experiences.

Prediction involves forecasting user actions, such as whether a visitor will make a purchase, leave the site, or return later. Machine learning models like decision trees, logistic regression, or neural networks are commonly used for these tasks.

Example: Predicting Whether a Website Visitor Will Make a Purchase

Problem Statement

An e-commerce company wants to classify visitors into **"Buyers"** and **"Non-Buyers"** based on their browsing behavior. The goal is to predict whether a visitor will complete a purchase based on various features like session duration, pages visited, previous purchases, and add-to-cart actions.

Step 1: Data Collection & Preprocessing

Data is collected from website logs, including:

- **User ID** (Unique identifier for visitors)
- **Pages Visited** (Number of pages viewed in a session)
- **Session Duration** (Time spent on the site)
- **Clicks** (Number of interactions with the site)
- **Add to Cart** (Whether the user added an item to the cart)
- **Past Purchase History** (Has the user made previous purchases?)
- **Purchase Label** (Target variable: 1 = Purchase, 0 = No Purchase)

Data is cleaned to remove missing values and normalized for better model performance.

Step 2: Exploratory Data Analysis (EDA)

- Understanding user behavior by visualizing session duration vs. purchase rate.
- Checking correlation between features (e.g., Do more pages visited increase purchase likelihood?).
- Identifying any trends in past purchasing behavior.

Step 3: Feature Engineering & Selection

- Converting categorical variables (e.g., Add-to-Cart, Past Purchases) into numerical form.
- Normalizing continuous variables like session duration and pages visited.

Step 4: Model Selection & Training

A supervised learning model like **Random Forest Classifier** or **Logistic Regression** is trained on the dataset.

Step 5: Prediction & Evaluation

- The model is tested on new users to predict whether they will make a purchase.
- Accuracy, precision, recall, and F1-score are used to measure performance.

Example Dataset (Sample)

User_ID	Pages_Visited	Session_Duration (mins)	Clicks	Add_to_Cart	Past_Purchase	Purchase (Target)
1	5	3.5	2	0	0	0
2	15	8.0	5	1	1	1
3	8	5.2	3	1	0	1

4	2	1.0	1	0	0	0
5	12	7.0	4	1	1	1

From the dataset, users who visit more pages, spend more time, and add items to the cart are more likely to make a purchase.

Key Insights from Classification & Prediction

- Users with **higher session duration and more clicks** have a higher probability of purchasing.
- **Add-to-cart behavior** is a strong indicator of purchase likelihood.
- Visitors with **past purchases** are more likely to buy again.

By using classification and prediction, businesses can:

1. **Optimize website layout** to improve user engagement.
2. **Send targeted promotions** to users likely to purchase.
3. **Personalize recommendations** based on user behavior.