UNIT IV : **Web Analytics, Web Mining:**

TOPIC -1

BASIC CONCEPTS IN MINING DATA STREAMS: MINING TIME SERIES DATA

Mining Time series data

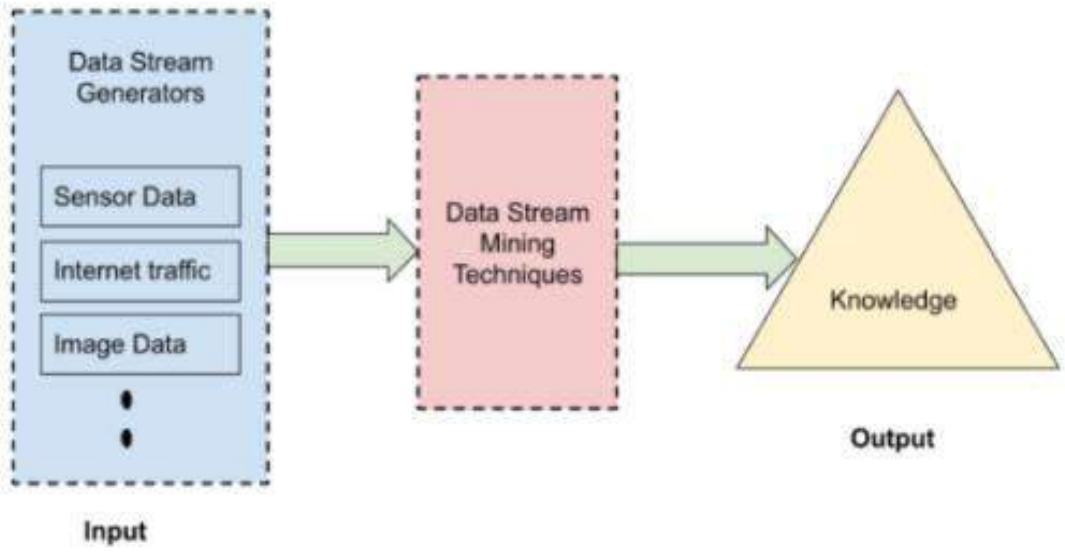
1. Data Stream mining:

- Data Stream is a continuous, fast-changing, and ordered chain of data transmitted at a very high speed. It is an ordered sequence of information for a specific interval. The sender's data is transferred from the sender's side and immediately shows in data streaming at the receiver's side. Streaming does not mean downloading the data or storing the information on storage devices.

Sources of Data Stream

- There are so many sources of the data stream, and a few widely used sources are listed below:
- Internet traffic
- Sensors data
- Live event data
- Satellite data
- Audio listening
- Watching videos
- Real-time surveillance systems

- By applying data mining to data streams, knowledge and insights are extracted from continuous, real-time data using specialized processing techniques.
- **Data mining** is the process of extracting knowledge or insights from large amounts of data using various statistical and computational techniques



2. Time series database

- Time-series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are popular in many applications, such as stock market analysis, economic and sales forecasting , observation of natural phenomena (such as atmosphere, temperature, wind, earthquake),and medical treatments.

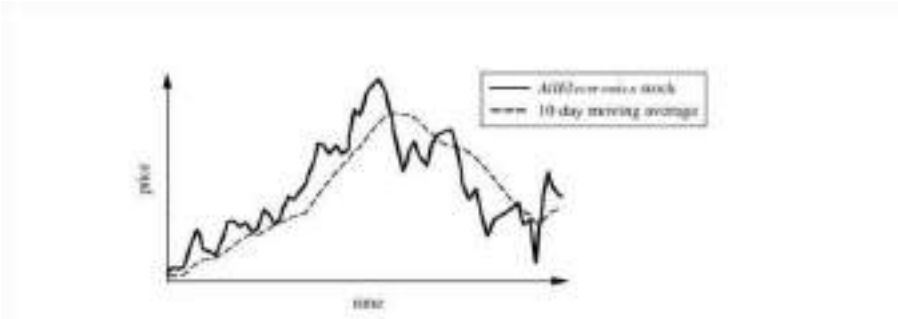


Figure 8.4 Time-series data of the stock price of *AllElectronics* over time. The *trend* is shown with a dashed curve, calculated by a moving average.

- A time-series database is also a sequence database. However, a sequence database is any database that consists of sequences of **ordered events**, with or without concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

- With the growing deployment of a large number of sensors, telemetry devices, and other on-line data collection tools, the amount of time-series data is increasing rapidly, often in the order of gigabytes per day (such as in stock trading) or even per minute (such as from NASA space programs).

3. Methods for Mining Time-Series Databases

- Two common methods for mining time-series databases are Trend Analysis and Similarity Search.
  - Trend Analysis
  - Similarity Search
- **Trend Analysis**:

  In this ,we Identify patterns or trends in time-series data over time.In this , techniques like moving averages, exponential smoothing, and regression models can be used to detect long-term trends.It is Used for forecasting, anomaly detection, and understanding the underlying patterns in the data.

- **Anomaly detection** is the process of identifying data points, events, or patterns that deviate significantly from the normal or expected behavior. In heartbeat analysis, **anomaly detection** identifies irregular heart rhythms or patterns that deviate from the normal heartbeat pattern.

- **Similarity Search**:
  - It is aimed at finding patterns or subsequences in a time-series database that are similar to a given query sequence.Measures like Euclidean distance,discrete Fourier transform , singular value decomposition or cosine similarity are often used for this.Helpful in applications like anomaly detection, pattern matching, and clustering time-series data.

2.1. Trend Analysis

- A time series involving a variable Y, representing the daily closing price of a share in a stock market, can be viewed as a function of time t, that is, Y = F(t). Such a function can be illustrated as a time-series graph, as shown in Figure 8.4, which describes a point moving with the passage of time.
- In general, there are two goals in time-series analysis:
- (1) modeling time series (i.e., **to gain insight** into the mechanisms or underlying forces that generate the time series), and
- (2) forecasting time series (i.e., to predict the future values of the time-series variables).
- Trend analysis consists of the following four major components or movements for characterizing time-series data:
  - Trend or long-term movements
  - Cyclic movements or cyclic variations
  - Seasonal movements or seasonal variations
  - Irregular or random movements

1. Trend or long-term movements:

These indicate the general direction in which a time series graph is moving over a long interval of time. This movement is displayed by a trend curve, or a trend line. For example, the trend curve of Figure 8.4 is indicated by a dashed curve. Example: A company wants to determine whether its annual revenue is increasing over time. They analyze the revenue data over the past decade and fit a regression line to identify a positive trend.

2. Cyclic movements or cyclic variations:

- These refer to the cycles, that is, the long-term oscillations about a trend line or curve, which may or may not be periodic. That is, the cycles need not necessarily follow exactly

similar patterns after equal intervals of time. The retail store's sales fluctuate every few years due to economic recessions .

3. **Seasonal movements or seasonal variations:**

   - **These are systematic or calendar related.**

   - seasonal movements are the identical or nearly identical patterns that a time series appears to follow during corresponding months of successive years.Peak demand for electricity during **summer months** (April to June) due to air conditioning.The observed increase in water consumption in summer due to warm weather is another example.

4. **Irregular or random movements:**

   - **These characterize Irregular motion of time** series due to random events, such as labor disputes, floods, or announced personnel changes within companies .The store's sales dropped unexpectedly in March due to a sudden supply chain disruption. (A **sudden supply chain disruption** refers to an unexpected interruption in the flow of goods, materials, or services within the network that impacts production or delivery.)

Techniques like moving averages, exponential smoothing, and regression models can be used to detect long-term trends.

1. Moving Averages:

- This helps to see the **overall upward or downward trend**

- **Use**: This helps to see the **overall upward or downward trend** in sales while ignoring short-term variations or outliers.

- **Example**: Monthly Sales Data

- You have monthly sales data for the past two years. To smooth out short-term fluctuations and highlight long-term trends, you calculate a **3-month moving average**.

- **Calculation**:

  - For April, the 3-month moving average is the average of sales for January, February, and March.

  - For May, it's the average of sales for February, March, and April, and so on.

2. Exponential Smoothing:

- This helps detect **recent trends**

- **Use**: This helps detect **recent trends** in traffic, making it useful for **forecasting** website traffic for the next few days or weeks based on recent patterns.
- **Example**: Forecasting Website Traffic
- You have daily website traffic data and want to forecast future traffic based on past trends. Using **simple exponential smoothing**, you give more weight to recent data points while smoothing out older data.
- $\hat{y}_t = \alpha \cdot y_{t-1} + (1 - \alpha) \cdot \hat{y}_{t-1}$, where α is the smoothing constant.

3. Regression Models:

- This helps identify whether there is an overall **upward or downward trend** and can also provide a **forecast** for future.
- **Use**: This helps identify whether there is an overall **upward or downward trend** in housing prices over time, and can also provide a **forecast** for future prices based on the linear trend.
- **Example**: Predicting Housing Prices
- You have historical data of housing prices over several years. A **linear regression model** can be applied where the dependent variable is the housing price, and the independent variable is time (e.g., the number of years).

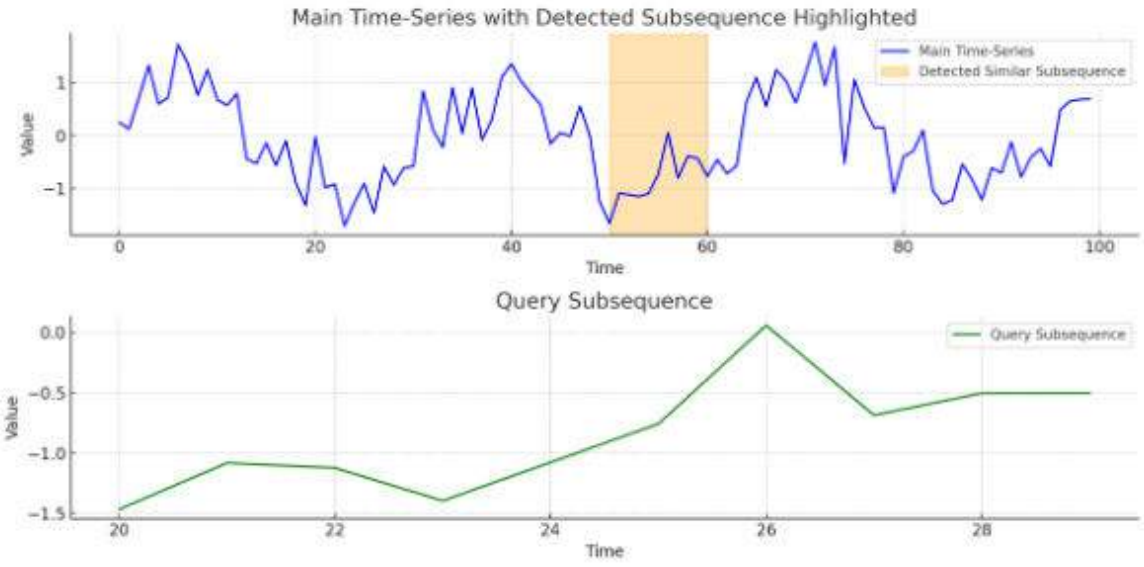
$$\text{Model: Price} = \beta_0 + \beta_1 \cdot \text{Time}$$

2.2Similarity Search

- **Similarity search** in time-series analysis involves finding sequences in a dataset that are **similar** to a given query sequence.This is useful when data points are expected to have slight variations or noise but still follow a recognizable pattern.A **similarity search** differs from normal database queries by finding data that **approximate** the given query, instead of matching it exactly. This is particularly useful in time-series analysis where slight variations may exist in patterns, but the overall trend or shape of the data sequence is similar.
- **Two Types of Similarity Searches:**
    - **Subsequence Matching**:

    - **Whole Sequence Matching**

- **Subsequence Matching**:
    - In **subsequence matching**, the goal is to find subsequences within a larger time-series sequence that are similar to a given query subsequence. The query sequence may be just a part of the larger dataset, and you're looking for parts of the data that resemble the query.
    - **Example**: In **stock market analysis**, if you're looking for days where stock prices increased followed by a slight dip (a pattern you identify as important), subsequence matching will help you find **similar patterns** in other parts of the dataset, even if the exact values of stock prices vary.



- **Whole Sequence Matching**:
    - In **whole sequence matching**, the focus is on finding time-series sequences that are **similar to each other as a whole**, meaning you're comparing the entire length of the sequence to find similar patterns across different time periods.
    - **Example**: In **medical diagnosis**, you might have a sequence of heartbeats from a healthy patient and want to find other patients' heartbeat data that shows a **similar rhythm** across the entire sequence.

UNIT -IV

TOPIC -2

MINING SEQUENCE PATTERNS IN TRANSACTIONAL DATABASES

Mining sequence patterns in Transactional databases

- A sequence database consists of sequences of ordered elements or events, recorded with or without a clear concept of time.Typical examples include customer shopping sequences, Web clickstreams, biological sequences, sequences of events in science and engineering. A **transactional database** stores data in the form of individual transactions, each with a set of items (or events), and is commonly used for analyzing purchasing behavior, event logs, or any data where the sequence of actions matters .

What is sequential pattern mining?"

- *Sequential pattern mining is the mining of frequently* occurring ordered events or subsequences as patterns. An example of a sequential pattern is "*Customers who buy a Canon digital camera are likely to buy an HP color printer within a month."*

- *For retail data, sequential patterns are useful for shelf placement* and promotions.Other areas in which sequential patterns can be applied include Web access pattern analysis, weather prediction, production processes, and network intrusion detection.

Techniques for Mining Sequential Patterns

- **Scalable methods** are techniques used to efficiently mine sequential patterns from large datasets (e.g., GSP, SPADE, PrefixSpan).

- **Constraint-based mining** refers to the process of mining sequential patterns while applying user-defined constraints (e.g., frequency, length, time) to focus on specific patterns.

Scalable Methods for Mining Sequential Patterns

- These are methods designed to work efficiently with large datasets. They reduce the time and memory needed to find patterns by organizing the data smartly and limiting unnecessary calculations.**Where it's used**: In businesses like retail and e-commerce to analyze customer buying habits or web browsing behavior.

- Techniques: GSP, SPADE, PrefixSpan.

1.GSP

GSP (Generalized Sequential Pattern):

It is an extension of the Apriori algorithm, GSP generates candidates iteratively and remove infrequent patterns based on a minimum support threshold.

GSP Process:

i. First, GSP generates candidate sequences (like {milk, butter} and {bread, butter}).

ii. Then, it checks how often these candidate patterns appear in the dataset. If a candidate appears more than the minimum support threshold, it is retained as a frequent sequential pattern.

iii. The algorithm tests and validates these candidates until it identifies frequent subsequences like {milk, butter} and {bread, butter}.

2. SPADE

SPADE (Sequential Pattern Discovery using Equivalence Classes):

a. **SPADE Process**:
   i. In SPADE, patterns are mined based on the equivalence classes (groups of items that occur together in sequences). SPADE efficiently discovers frequent patterns by applying **set intersection** to find common occurrences between sequences.
   ii. For example, the subsequence {milk, bread} appears in all four sequences, so it would be considered a frequent sequential pattern.

Example: Using the same customer purchase sequences, SPADE stores the dataset in a **vertical format**:

b. **Vertical Format**:

i. {milk} → [1, 2, 3, 4]

ii. {bread} → [1, 2, 3, 4]

iii. {butter} → [1, 2, 3, 4]

iv. {cheese} → [4]

Step	GSP	SPADE
1. Input Sequences	S1: (milk), (bread), (butter)	T1: (milk), (bread), (butter)
	S2: (milk), (butter), (bread)	T2: (milk), (butter), (bread)
	S3: (bread), (butter), (milk)	T3: (bread), (butter), (milk)
	S4: (milk), (bread), (butter), (cheese)	T4: (milk), (bread), (butter), (cheese)
2. 1-item Sequences	(milk): 4, (bread): 4, (butter): 4,	(milk) - TIDs: (T1, T2, T3, T4)
		(bread) - TIDs: (T1, T2, T3, T4)
		(butter) - TIDs: (T1, T2, T3, T4)
3. Generate 2-item Sequences	(milk -> bread), (milk -> butter), (bread -> butter)	(milk -> bread): (T1, T2, T4)
		(milk -> butter): (T1, T2, T4)
		(bread -> butter): (T1, T3, T4)
4. Count Support of 2-item Sequences	(milk -> bread): 3 (milk -> butter): 3 (bread -> butter): 3	(milk -> bread): TIDs = {T1, T2, T4} -> Support = 3 (milk -> butter): TIDs = {T1, T2, T4} -> Support = 3 (bread -> butter): TIDs = {T1, T3, T4} -> Support = 3

5. Generate 3-item Sequences	**(milk -> bread -> butter)**	**(milk -> bread -> butter): (T1,T4)**
6. Count Support of 3-item Sequences	**(milk -> bread -> butter): 2**	**(milk -> bread -> butter): TIDs = {T1, T4} -> Support = 2**
7. Final Frequent Sequences	**(milk), (bread), (butter), (milk -> bread), (milk -> butter), (bread -> butter), (milk -> bread -> butter)**	**(milk), (bread), (butter), (milk -> bread), (milk -> butter), (bread -> butter), (milk -> bread -> butter)**

3.PrefixSpan

- **PrefixSpan (Prefix-Projected Sequential Pattern Growth)**:
- Given the same purchase sequences, PrefixSpan focuses on **projecting** the dataset based on the frequent prefixes.
    - For example, starting with the frequent prefix {milk}, the algorithm projects the dataset to sub-sequences that start with {milk} (i.e., Sequence 1: [{milk}, {bread}, {butter}],
    - Sequence 2: [{milk}, {butter}, {bread}],
    - Sequence 4: [{milk}, {bread}, {butter}, {cheese}]...).
    - It then recursively mines the projected sub-sequences, growing the pattern incrementally.The patterns discovered would include {milk, bread} and {milk, butter}, which are frequent across the sequences.

2. Constraint-Based Mining of Sequential Patterns

- Without user-defined constraints, mining can generate irrelevant patterns, reducing efficiency and usability. Constraint-based mining incorporates user-specified constraints to focus the search on patterns of interest, improving both efficiency and the relevance of results.Suppose we are analyzing customer shopping behavior, where each sequence represents a customer's purchase history.
- **Constraints:**
- **Duration Constraint (T)**:
  - **Example**: We want to find sequential patterns of products bought by customers only in the year 2023.
  - **Constraint**: T = 2023.
  - This constraint will limit the sequences being mined to only those that occurred during 2023, ignoring sequences from other years.
- **Monotonic Duration Constraint**:
- **Example**: We want to find customers who bought at least 5 items in any sequence.
- **Constraint**: $T \geq 5$.
- Once we find a sequence with 5 items, all subsequences with more than 5 items are also considered valid, ensuring we capture longer patterns.

3. Periodicity Analysis of Time Sequence Data

- **Description**: Detects repeating patterns or periodic behaviors in time-series data.
- **Example**:
  - Electricity usage data shows a pattern:
    - High usage at 8:00 AM daily (morning routine).
    - Low usage at 2:00 PM daily (afternoon downtime).

- **Key Feature**: Useful for detecting regular cycles in data, such as seasonal trends or recurring events.

Applications

1. Retail and E-commerce: Discover purchase patterns, e.g., "Customers buying shoes often buy socks."

2. Web Usage Mining: Analyze browsing patterns to improve website navigation.(Sequential pattern: *Homepage → Product Category → Product Details → Add to Cart → Checkout*)

3. Medical Data Analysis: Study sequences of symptoms, diagnoses, or treatments to identify patterns in disease progression or treatment outcomes.

4. Social Media: Track user activities and interactions to personalize content delivery and improve ad targeting.

UNIT -IV

TOPIC -3

SPATIAL DATA MINING

1.Introduction

- **Spatial database?**
- A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data. Spatial databases have many features distinguishing them from relational databases.
- **Online Analytical Processing (OLAP)** is a category of data processing that enables users to interactively analyze and explore large amounts of data from different perspectives. It is primarily used in business intelligence (BI) systems to support decision-making.
- **Spatial data mining**
- Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologiesIt has wide applications in geographic information systems, geomarketing , remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used.Spatial data mining is used to analyze geographic information and spatial data.It explores patterns and relationships in geographical spaces.
- **Geostatistics vs. Spatial Statistics**:
- **Geostatistics** is associated with continuous geographic space (like elevation or temperature).
- **Spatial statistics** is related to discrete spaces (such as counts of objects in a certain area).

2. Spatial Data Cube Construction and Spatial OLAP

- Can we construct a spatial data warehouse?”
- Yes, as with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining.

- **Example 10.5 Spatial Data Cube and Spatial OLAP:**
  In British Columbia, 3,000 weather probes record daily temperature and precipitation, sending data to a weather station.

- Weather probes record daily temperature and precipitation

STEPS

Step 1: Data Integration: Collect spatial and non-spatial weather data from BC probes, integrating it into a centralized data warehouse.

Step 2: Multidimensional Model: Use a star schema to organize dimensions like region, time, and weather variables, with measures like temperature and precipitation.

Step 3: OLAP Analysis: Build the OLAP cube and perform operations (drill-down, roll-up, slice, dice) to uncover regional weather trends and patterns in BC.

Step 1: Data Integration:

- **Data Integration:** Collect spatial and non-spatial weather data from BC probes, integrating it into a centralized data warehouse.There are three types of *dimensions in a spatial data cube:*

- Non-spatial dimension

- spatial-to-non-spatial dimension

- spatial-to-spatial dimension

- A **non-spatial dimension** contains only non-spatial data, such as temperature and precipitation, which can be generalized into non-spatial categories like "hot" or "wet." These dimensions do not involve geographic or spatial information.

- A **spatial-to-non spatial dimension** starts with spatial data but generalizes to non-spatial categories. For example, a city like Seattle may be generalized to the non-spatial region "Pacific Northwest.“(string)

- A **spatial-to-spatial dimension** contains spatial data at both the primitive and generalized levels. For example, temperature regions like 0-5°C and 5-10°C remain spatial at all levels.(dimension *equi temperature region* contains spatial data)(one being a spatial region (geographical area) and the other being a temperature range assigned to a spatial region).Spatial-to-spatial refers to mapping or connecting geographic regions at different levels of detail, such as looking at both specific districts and broader states.

Step 2: Multidimensional Model:

- A **spatial data cube** essentially organizes data into a **multidimensional structure**, but underlying it is a collection of related tables—much like traditional database schemas, such as **star schemas.** In a star schema, dimensions like **region**, **temperature**, **time**, and **precipitation** are linked to measures such as **region map**, **area**, and **count**.Numerical measures like **area** and **count** can be computed similarly to non-spatial data cubes, while the **region map** is a spatial measure containing pointers to spatial objects.

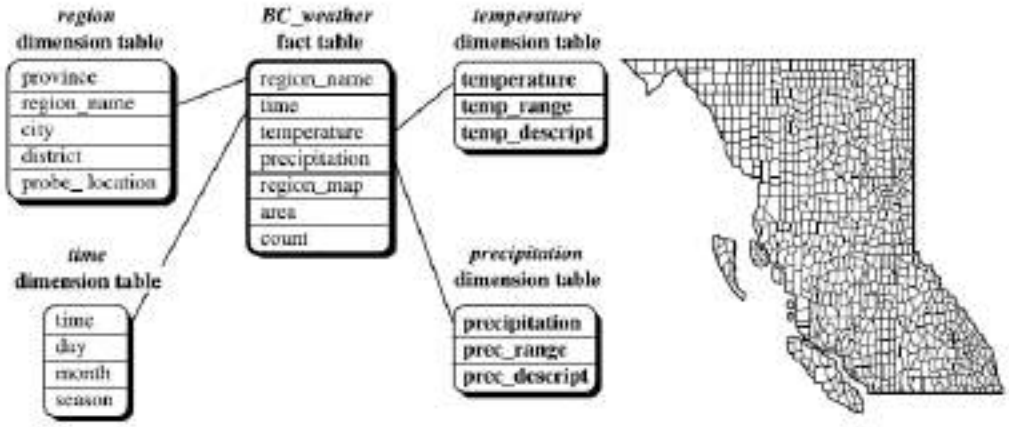


Figure 10.2 A star schema of the *BC_weather* spatial data warehouse and corresponding BC weather probes map.

- For example, two different roll-ups on the BC weather map data (Figure 10.2) may produce two different generalized region maps, as shown in Figure 10.4, each being the result of merging a large number of small (probe) regions from Figure 10.2.

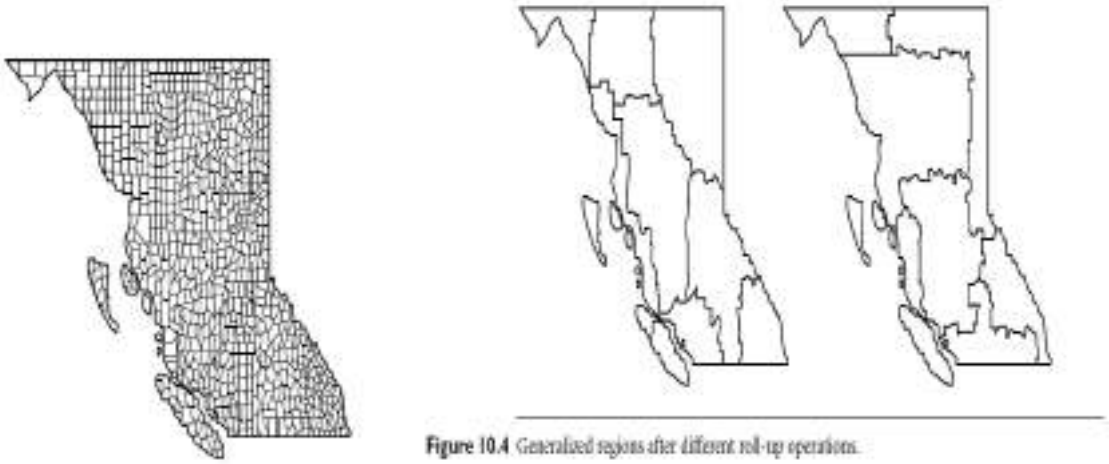
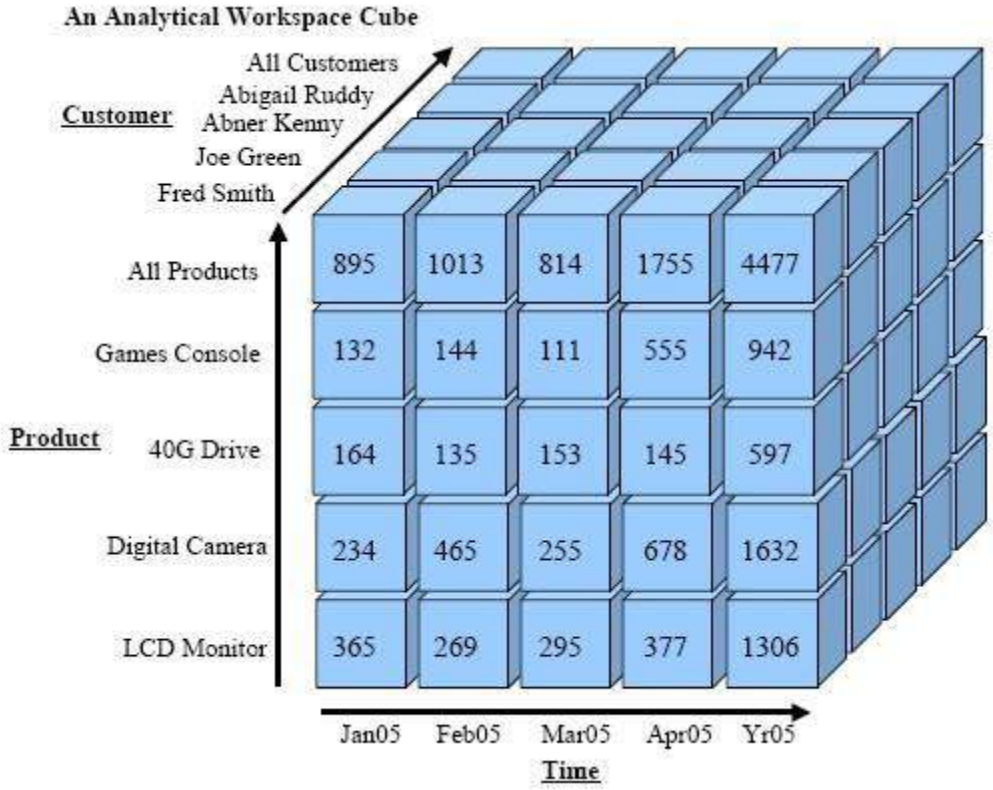


Figure 10.4 Generalized regions after different roll-up operations.

- A **spatial data cube** allows you to analyze data across both traditional dimensions (like time) and spatial dimensions (like geographic locations).
- **Spatial OLAP** lets you perform operations such as drill-down, roll-up, and slice on spatial data, helping to gain insights from large and complex spatial datasets.
- Image below highlights the MultiDimensional nature of the OLAP Cube showing a list of Products purchased by a number of Customers and the Cost implication (Measure)across a Time Period



- **Integration of heterogeneous spatial data** from multiple sources and formats is one of the main challenges in constructing a spatial data warehouse.
- The use of the **star schema** for organizing spatial data helps facilitate OLAP operations, though spatial indexing and efficient querying mechanisms are required to ensure fast and scalable performance.

Step 3: OLAP Analysis:

- Build the OLAP cube and perform operations (drill-down, roll-up, slice, dice) to uncover regional weather trends and patterns in BC.

3. **Spatial Clustering Methods**

- Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set. Some common **spatial data clustering** techniques include
- **K-means Clustering**: Groups data points into K clusters based on distance to the centroid.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Identifies clusters of arbitrary shape by grouping closely packed points and marking outliers.
- **K-medoids**: Similar to K-means but uses actual data points as cluster centers (medoids) instead of means.

4. **Spatial Classification and Spatial Trend Analysis**

- **Spatial classification** analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as the *neighborhood of a district, highway, or river.* **Spatial trend analysis** deals with another issue: the detection of changes and trends along a spatial dimension. Typically, trend analysis detects changes with time, such as the changes of temporal patterns in time-series data. Spatial trend analysis replaces time with space and studies the trend of non spatial or spatial data changing with space.
- Spatial trend analysis studies how non-spatial or spatial data change across space, such as observing the economic situation moving away from a city center or the climate and vegetation changes with increasing distance from an ocean.For such analyses, regression and correlation analysis methods are often applied by utilization of spatial data structures and spatial access methods.

UNIT -IV

TOPIC -4 : MULTIMEDIA DATA MINING

What is a multimedia database?"

Multimedia database is a collection of multimedia data which includes text, images, graphics (drawings, sketches), animations, audio, videos,etc.

Similarity Search in Multimedia Data

For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems:

(1)description-based retrieval systems: which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation

Description-based retrieval is labor-intensive if performed manually. If automated, the results are typically of poor quality. For example, the assignment of keywords to images can be a tricky and arbitrary task.

(2) content-based retrieval systems: which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image.

Content-based retrieval uses visual features to index images and promotes object retrieval based on feature similarity, which is highly desirable in many applications

In a content-based image retrieval system, there are often two kinds of queries: *image sample- based queries* and *image feature specification queries*

Image-sample-based queries find all of the images that are similar to the given image sample. This search compares the feature vector (or signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database. Based on this comparison, images that are close to the sample image are returned. You upload an actual image, and the system extracts its feature vector to find similar images in the database. It's like asking, "Show me images that look like this one."

- A **feature vector** (or **signature**) is a numerical representation of an image's key characteristics, such as its colors, textures, shapes, and patterns.

Image feature specification queries specify image features like color, texture, or shape, which are translated into a feature vector to be matched with the feature vectors of the images in the

database.**In this,** Instead of uploading an image, you describe specific features (like color, texture, or shape). The system then looks for images that match your description, even if the images aren't exactly the same.

- Content-based retrieval has wide applications, including medical diagnosis, weather prediction, TV production, Web search engines for images, and e-commerce.Some systems, such as *QBIC* (*Query By Image Content*), support both sample-based and image feature specification queries. There are also systems that support both content based and description-based retrieval.
- Several approaches have been proposed and studied for similarity-based retrieval in image databases, based on image signature:

1. Color histogram–based signature

2. Multi-feature composed signature

Color histogram–based signature:

- In this approach, the signature of an image includes color histograms based on the color composition of an image regardless of its scale or orientation.
- This method does not contain any information about shape, image topology, or texture.
- Thus, two images with similar color composition but that contain very different shapes or textures may be identified as similar, although they could be completely unrelated semantically.
- A signature refers to a representation or descriptor of an image, typically based on features like color histograms, used for comparison or identification.

Multi feature composed signature:

- In this approach, the signature of an image includes a composition of multiple features: color histogram, shape, image topology, and texture.
- The extracted image features are stored as metadata, and images are indexed based on such metadata. Often, separate distance functions can be defined for each feature and subsequently combined to derive the overall results.
- Multidimensional content-based search often uses one or a few key features to search for images containing such (similar) features. It can therefore be used to search for similar images. This is the most popularly used approach in practice.

2.2 Classification and Prediction Analysis of Multimedia Data

- Classification and predictive modeling are used in scientific fields like astronomy, seismology, and geosciences to analyze multimedia data.
- **Example: Astronomy Data :** Sky images, classified by astronomers, are used to train models to recognize celestial bodies such as galaxies and stars, based on attributes like magnitudes, areas, and orientations.
- **Some of the Key Components in Classification and Prediction Analysis of Multimedia Data are**
    - Data Preprocessing
    - Image Data Processing
    - Web Multimedia Data
    - Web Mining and Data Linkage
- **Data Preprocessing :** Key steps include **data cleaning**, **data transformation**, and **feature extraction** using techniques like edge detection, eigenvector decomposition, and probabilistic models.
- **Image Data Processing:** Handling large image datasets requires **parallel and distributed processing** due to their high volume and processing power needs.
- **Web Multimedia Data:** The **World Wide Web** serves as a vast source of multimedia data (images, videos, etc.) embedded in web pages, serving various purposes such as content, advertisements, or suggestions.
- **Web Mining and Data Linkage:** Mining data by utilizing the **relative locations** and **linkages** among text, images, and web page elements aids in better understanding and classifying multimedia content on the web.

2.3 Mining Associations in Multimedia Data

- "What kinds of associations can be mined in multimedia data?" Association rules involving multimedia objects can be mined in image and video databases. At least three categories can be observed:
- Associations between image content and nonimage content features: A rule like "If at least 50% of the upper part of the picture is blue, then it is likely to represent sky" belongs to this category since it links the image content to the keyword sky.

- Associations among image contents that are not related to spatial relationships: A rule like “If a picture contains two blue squares, then it is likely to contain one red circle as well” belongs to this category since the associations are all regarding image contents.
- Associations among image contents related to spatial relationships: A rule like “*If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath” belongs to this category since it associates objects in the image with* spatial relationships.
- To mine associations among multimedia objects, we can treat each image as a transaction and find frequently occurring patterns among different images.

UNIT IV:

TOPIC -5:

SECURITY FIRST INSURANCE DEEPENS CONNECTION WITH POLICYHOLDERS

Security First Insurance Deepens Connection with Policyholders

1. Security First Insurance is one of the largest homeowners' insurance companies in Florida.

2. Headquartered in Ormond Beach, it employs more than 80 insurance professionals to serve its nearly 190,000 customers.

I. CHALLENGES

Following are the challenges faced by Security First Insurance

1. Florida's Hurricane Exposure
2. Security First's Commitment
3. Surge in Claims After Hurricanes
4. Evolving Communication Channels
5. Need for Proactive and Integrated Approach

1. Florida's Hurricane Exposure: Florida faces the highest exposure to hurricanes in the U.S., with an average of 12 named storms and 9 hurricanes annually, impacting property and people.

2. Security First's Commitment: The company is financially strong enough to withstand multiple natural disasters and promises to support customers "storm after storm, year after year."

3. Surge in Claims After Hurricanes: While Security First processes 700 claims monthly, this number can rise to tens of thousands after a hurricane, creating significant challenges in managing the influx.

4. Evolving Communication Channels: Customers now contact the company through various means, including phone, email, and **social media platforms like Facebook and Twitter, necessitating a more responsive approach.**

5. Need for Proactive and Integrated Approach: Concerned about response times, Security First recognized the need to integrate social media responses into the claims process and document them for regulatory compliance.

II. SOLUTIONS:

1. Providing Responsive Service No Matter How Customers Get in Touch

2.Prioritizing Communications with Access to Smarter Content

1. Providing Responsive Service No Matter How Customers Get in Touch

- **Collaboration with IBM Business Partner:** Security First partnered with a leading IBM Business Partner to enhance customer experience by utilizing social media.
- **SMC4 Solution:** The IBM Business Partner implemented a solution called Social Media Capture (SMC4), built on "IBM Enterprise Content Management software", providing four essential capabilities: capture, control, compliance(regulation) and communication.For example, the SMC4 solution logs all social networking interaction for Security First, captures content, monitors incoming and outgoing messages and archives(retains) all communication for compliance(regulation) review.Content Collector for Email software automatically captures email content and attachments and sends an email back to the policyholder acknowledging receipt
- **Content Analytics and Claims Integration**: "IBM Content Analytics with Enterprise Search" analyzes customer posts and emails, captures relevant information, and integrates it directly into claims documents to initiate the claims process.
- **Centralized Communication Repository:** All incoming communications from the company's website, emails, and the Internet are collected into a "central FileNet Content Manager repository" to maintain, control, and link to the appropriate workflow.
- **Seamless Integration with Company Systems:** The solution integrates easily with Security First's existing applications, databases, and processes, enhancing overall claims management efficiency.

2.Prioritizing Communications with Access to Smarter Content

- **Emergency Support**: After a hurricane, people whose homes are damaged or destroyed are often forced to leave homes quickly with little more than the clothes on their backs. They rely on their insurance companies to promptly provide the necessary support.
- **Prioritizing Requests**: When tens of thousands of policyholders need assistance in a short time, Security First must quickly sort and prioritize requests. The Content Analytics with Enterprise Search software helps identify the most critical cases.
- **Smarter Content Analysis**: The software uses text mining, text analytics, natural language processing, and sentiment analytics to analyze emails, social media posts, tweets, and comments, detecting words and tones that signal significant property damage or distress.
- **Prioritization and Routing**: This analysis enables Security First to prioritize messages and direct them to the appropriate personnel to offer reassurance, address complaints, or process claims.
- **Faster, Personalized Responses**: With access to smarter content, Security First can respond more quickly, efficiently, and personally to customers, ensuring that those in distress receive the appropriate level of assistance.

III. Result

- **Meeting Regulatory Requirements**: Security First uses "IBM software's text analysis" capabilities to filter inappropriate incoming communications and monitor outgoing messages, ensuring they meet industry regulations.
- **Controlled Employee Responses**: The system allows Security First to designate specific employees or roles to create and submit responses, ensuring the responses follow company policies and industry standards.
- **Automated Verification and Message Review**: The system automatically verifies the designated personnel and analyzes outgoing message content, flagging ineffective or questionable communications for further review.
- **Tracking and Maintaining Control**: All communication interactions are recorded, enabling Security First to track and manage the process, including controlling which employees respond, their authority level, and message content.
- **Expanded Use of Social Media:** With regulatory concerns addressed, Security First confidently expands its use of social media, allowing the company to directly engage with customers and enhance communication opportunities.

UNIT IV

TOPIC -6

WEB MINING OVERVIEW

- **Global Reach and Increased Competition:** The Internet has expanded opportunities for businesses to reach new customers and markets, while also intensifying competition in a global, fast-changing marketplace.
- **Necessity of an Online Presence for Companies :** Companies must have an engaged and active presence on the Internet, as it is no longer optional. Customers expect businesses to offer products and services online, alongside other digital interactions.

Challenges in Web Mining: Overcoming the Complexities of the Web

- Because of its sheer size and complexity, mining the Web is not an easy undertaking by any means.
- The Web also poses great challenges for effective and efficient knowledge discovery:

1.The Web is too big for effective data mining.

2.The Web is too complex.

3.The Web is too dynamic

4.The Web is not specific to a domain

5.The Web has everything

1.The Web is too big for effective data mining.

The Web is so large and growing so rapidly that it is difficult to even quantify its size. Because of the sheer size of the Web, it is not feasible to set up a data warehouse to replicate, store, and integrate all of the data on the Web, making data collection and integration a challenge.

2.***The Web is too complex.*** The complexity of a Web page is far greater than that of a page in a traditional text document collection. Web pages lack a unified

structure.They contain far more authoring style and content variation than any set of books, articles, or other traditional text-based document.

3.***The Web is too dynamic.*** The Web is a highly dynamic information source. Not only does the Web grow rapidly, but also its content is constantly being updated.Blogs, news stories, stock market results, weather reports, sports scores, prices, company advertisements, and numerous other types of information are updated regularly on the Web.

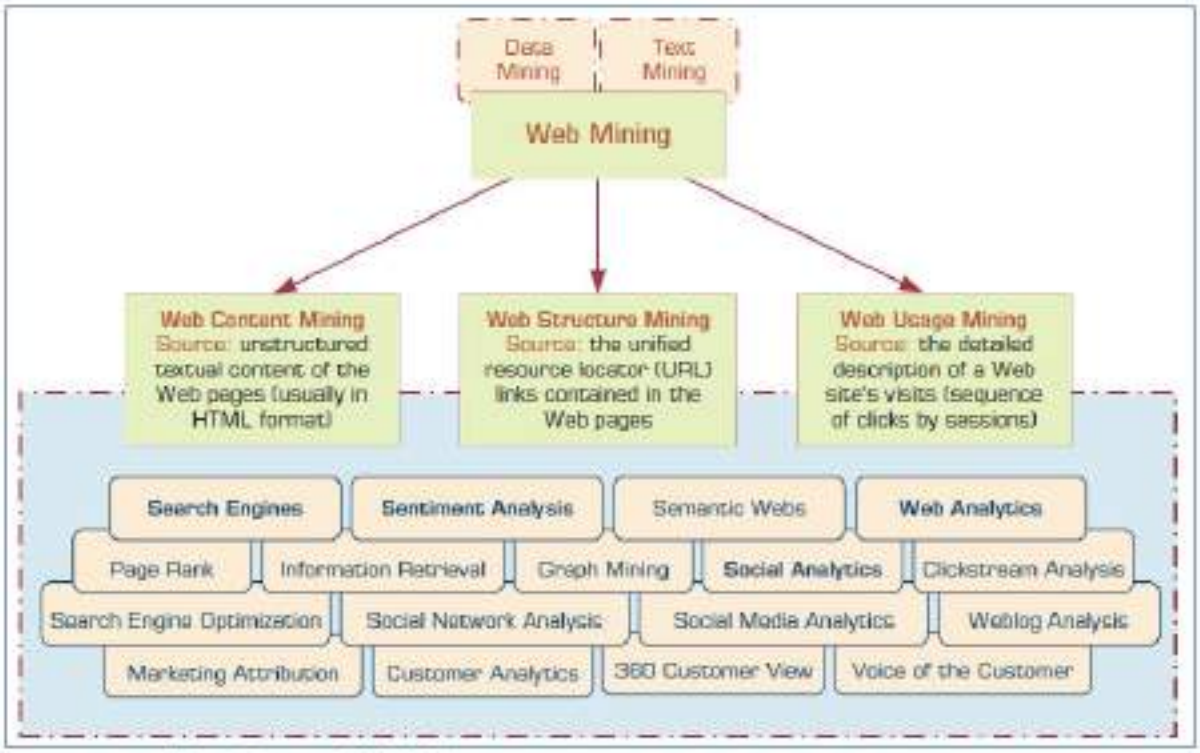
4. ***The Web is not specific to a domain.*** The Web serves a broad diversity of communities and connects billions of workstations. Web users have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search that they perform.

5.***The Web has everything.*** Only a small portion of the information on the Web is truly relevant or useful to someone (or some task). It is said that 99% of the information on the Web is useless to 99% of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, whereas the rest of the Web contains information that is uninteresting to the user and may swamp desired results. Finding the portion of the Web that is truly relevant to a person and the task being performed is a prominent issue in Web related research.

The challenges of Web data discovery have led to research aimed at improving efficiency.While index-based search engines help locate documents using specific keywords, they often return irrelevant or too many results, and may miss highly relevant documents that don't match the exact search terms. Web mining offers a more effective solution

Web mining

- **Web mining (or Web data mining)** is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information.
- Web mining is essentially the same as data mining that uses data generated over the Web.
- The goal is to turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information (i.e., knowledge) to promote better decision making throughout the enterprise.
- For example, an e-commerce platform can analyze user browsing and purchase history to recommend items like clothes, electronics, or accessories based on individual preferences, boosting customer engagement and sales.
- Because of the increased popularity of the term *analytics, nowadays* many have started to refer to Web mining as *Web analytics*
- However, these two terms are not the same. Web analytics is primarily Web site usage data focused(which includes how visitors interact with a **specific website**), Web mining is inclusive of all data generated via the Internet including transaction, social, and usage data.
- Figure presents a simple taxonomy of Web mining, where it is divided into three main areas: Web content mining, Web structure mining, and Web usage mining.
- In the figure, the data sources used in these three main areas are also specified.
- Although these three areas are shown separately, as you will see in the following section, they are often used collectively and synergistically to address business problems and opportunities.



A Simple Taxonomy of Web Mining

- **Web content mining** refers to the extraction of useful information from Web pages.
- **Web structure mining** is the process of extracting useful information from the links embedded in Web documents.
- **Web usage mining** (also called Web analytics) is the extraction of useful information from data generated through Web page visits and transactions.

UNIT -IV

TOPIC -7

WEB CONTENT AND WEB STRUCTURE MINING

Web Content Mining

- **Web content mining** refers to the extraction of useful information from Web pages. The documents may be extracted in some machine-readable format so that automated techniques can extract some information from these Web pages. A company wants to understand customer sentiment about its product.
- **Web crawlers** (also called spiders) are used to read through the content of a Web site automatically. The information gathered may include document characteristics similar to what is used in text mining.Such an automated (or semi automated) process of collecting and mining of Web content **can be used for** competitive intelligence (collecting intelligence about competitors' products, services, and customers). **It can also be used** for information/news/opinion collection and summarization, sentiment analysis, and automated data collection and structuring for predictive modeling.
- **Illustrative example:** Drs. Sharda and Delen developed models to predict Hollywood movie financial success before release, using data from various websites with different page structures.Manually collecting data on thousands of movies was time-consuming and error-prone, so they used Web content mining and spiders to automate the process.The automated system collects, verifies, and stores data in a relational database, ensuring data quality while saving valuable time.

AUTHORITATIVE PAGE AND HUB:

- Web pages have hyperlinks that connect to other pages. These links can reveal which pages are more important or central. Pages with more links are often considered more **authoritative.**
- When a Web developer includes a link to another page, it acts as an endorsement, with collective endorsements indicating the importance of the page.The vast Web linkage information offers valuable insights into the relevance, quality, and structure of Web content, making it a rich resource for Web mining.
- Example of **authoritative page :**

- Websites like **Indian Institutes of Technology (IITs)**, **Indian Council of Medical Research (ICMR)**, or **UNICEF India** might link to the government health portal to provide further information, research, or news updates. These endorsing pages highlight the relevance and authority of the health portal.In this case, the Indian government's official health portal is considered the **central or authoritative page**, as it is endorsed by trusted, reputable sources, and is likely to rank higher in search results.
- The structure of Web hyperlinks has led to another important category of Web pages called **a hub.** A hub is one or more Web pages that provide a collection of links to authoritative pages.Wikipedia is a **hub** because each page provides links to authoritative sources, such as academic papers, books, and government or organizational websites.
- For instance, the page on **Climate Change** links to authoritative research papers, studies, and data from institutions like **NASA.**

Web structure mining

- Web structure mining is the process of extracting useful information from the links embedded in Web documents.
- It is used to identify authoritative pages and hubs, which are the cornerstones of the contemporary page-rank algorithms that are central to popular search engines such as Google and Yahoo!
- In search engines like Google and Yahoo, links to a website help determine its authority, meaning the more links pointing to a page, the more popular or trusted it is.
- In Google's PageRank algorithm, **www.mohfw.gov.in** is considered an authoritative page due to links from reputable sources like WHO and Indian medical institutes, providing official health-related information.
- A news portal like **India Today** functions as a hub, linking to authoritative pages such as government websites and international news outlets across various topics like politics, health, and technology.
- Analysis of links is very important in understanding the interrelationships among large numbers of Web pages, leading to a better understanding of a specific Web community.

UNIT -IV

TOPIC -8

MINING THE WORLD WIDE WEB

- The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services.
- The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining.
- However, based on the following observations, **the Web also poses great challenges for effective resource and knowledge discovery:**

- ❑ *The Web seems to be too huge for effective data ware housing and data mining*
- ❑ *The complexity of Web pages is far greater than that of any traditional text document collection.*
- ❑ *The Web is a highly dynamic information source*
- ❑ *The Web serves a broad diversity of user communities*
- ❑ *Only a small portion of the information on the Web is truly relevant or useful*

Types of Web Mining:

- Web mining enhances search engines by identifying authoritative pages and classifying documents, categorized into Web Content Mining, Web Structure Mining, and Web Usage Mining.

Web Content Mining

- Focuses on extracting useful information from the content of web pages, such as text, images, and videos.
- It involves techniques like text mining, natural language processing, and multimedia data mining to identify patterns and insights.

Web Structure Mining

- Analyzes the structure of hyperlinks between web pages to identify relationships and important connections.
- Techniques like the HITS algorithm and PageRank are used to discover hubs, authoritative pages, and the overall web structure.

Web Usage Mining

- Examines user interaction data, such as browsing behavior, click patterns, and server logs, to understand user preferences.
- Helps in creating personalized recommendations, improving website design, and enhancing user experience.

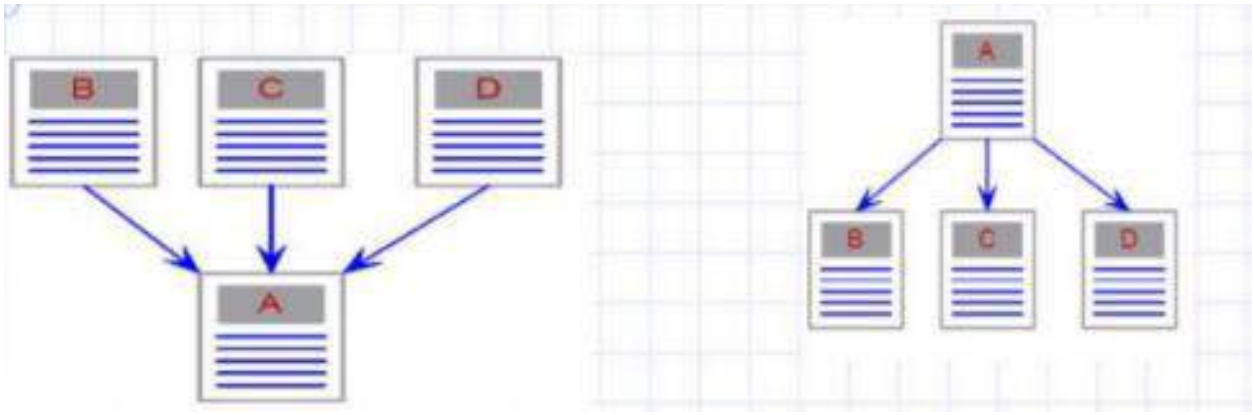
Mining the Web's Link Structures to Identify Authoritative Web Pages

- **Authoritative Web Pages**: These are important and trustworthy pages about a specific topic. They are identified by looking at how other pages link to them. Example: authoritative website in India is the Income Tax Department of India. It is widely linked by financial institutions, government portals, and professional advisory sites for accurate information about income tax rules and services.
- **Web Linkages and Information**: Hyperlinks provide valuable information about the relevance and importance of pages, although some links serve other purposes like navigation or advertisements.
- **Hubs and Authorities**: Hubs are pages that link to many authoritative pages on a specific topic, while authoritative pages are those linked to by many good hubs.
- For example, a popular news website like NDTV can be a hub if it links to many trusted news sources, while a trusted website like NPTEL is an authoritative page, as it provides reliable educational content in India.

HITS(Hyperlink-Induced Topic Search) Algorithm:

- This method begins with a root set of pages from a search engine, expands it to a base set by adding linked and linking pages, and assigns hub and authority weights through an iterative process.
- Links within the same domain, often used for navigation, are excluded from the analysis to focus on meaningful connections between different pages.
- HITS identifies good authorities and hubs for a topic by assigning two numbers to a page: an authority and a hub weight.

- Page on the left is authority page
- Page on the right is hub page



Automatic Classification of Web Documents

- In the automatic classification of Web documents, each document is assigned a class label from a set of predefined topic categories, based on a set of examples of pre-classified documents.
- For example, Yahoo!'s classification system and its associated documents can be used as training and test sets in order to derive a Web document classification scheme.
- An example of using Yahoo!'s classification system for Web document classification would be to take Yahoo!'s categorized directories, such as **Yahoo! Finance** or **Yahoo! Sports**, as labeled datasets.
- This scheme may then be used to classify new Web documents by assigning categories from the same classification system.
- Web page may contain multiple themes, advertisement, and navigation information, *block-based page content analysis may play an* important role in construction of high-quality classification models.
- **Block-based page content analysis** divides web pages into sections like the header, content, and sidebar to focus on relevant information and ignore distractions like ads. For example, it would focus on the article content in a news page while ignoring the ads in the sidebar.
- The Semantic Web is an extension of the web that organizes and links data based on its meaning, allowing machines to understand and process information more effectively.
- An example of the Semantic Web would be a system that links information about books, authors, and publishers in a way that allows a search engine to understand relationships, such as "Books by J.K. Rowling" or "Published by Penguin."

- Web mining analyzes web content to automatically extract meaningful information, such as topics or relationships between pages. This extracted data is then organized into a structured format, making it easier for machines to understand and process, which supports the creation of the Semantic Web..

UNIT -IV

TOPIC -9

SEARCH ENGINES-1

Search Engines

- Search engine is a software program that searches for documents (Internet sites or files), based on the keywords users have provided.
- The overall goal of a search engine is to return one or more documents/pages (if more than one document/page applies, then a ranked-order list is often provided) that best match the user's query.

Anatomy of a Search Engine

- At the highest level, a search engine system is composed of two main cycles: a development cycle and a responding cycle.
- While one is interfacing with the World Wide Web, the other is interfacing with the user.
- One can think of the development cycle as a production process (manufacturing and inventorying documents/pages) and the responding cycle as a retailing process (providing customers/users what they want).

1. Development Cycle

- It is the interfacing with the World Wide Web.
- The two main components of the development cycle are the Web crawler and document indexer.
- The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value.
- **Web Crawler:** A Web crawler (also called a spider or a Web spider) is a piece of software that systematically browses (crawls through) the World Wide Web for the purpose of finding and fetching Web pages. Often Web crawlers copy all the pages they visit for later processing by other functions of a search engine. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the already existing list of URLs to visit (i.e., the scheduler).

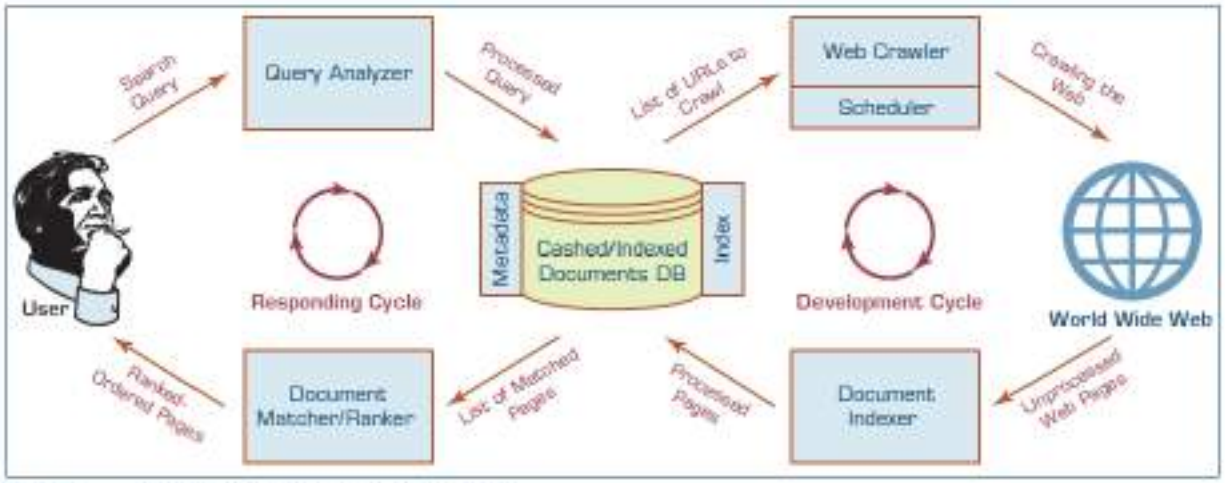


FIGURE Structure of a Typical Internet Search Engine.

- **Document Indexer**: The retrieved pages by the crawler are processed, which includes cleaning, extracting metadata, and structuring the data for storage.
- The **Document Indexer** processes documents and stores them in a structured format, often in a database-like structure where each entry links a website to relevant keywords, making it searchable.
- To convert the documents/pages into the desired, easily searchable format, the document indexer performs the following tasks.

Step 1: Pre processing the Documents

- Documents fetched by the crawler are converted into a standard format by separating and organizing different content types for easier processing.
- For example, a webpage fetched by the crawler might contain text, images, and hyperlinks. During preprocessing, the text is extracted and cleaned, images are saved separately, and hyperlinks are categorized, all converted into a uniform format for further analysis.

Step 2: Parsing the Documents

- **Parsing** refers to breaking down text into smaller components
- Parsing applies text mining techniques to extract and refine index-worthy terms by tokenizing, correcting errors, and removing stop words. Stemming and lexicons (e.g., WordNet) standardize terms, ensuring consistency before indexing.

Step 3: Creating the Term-by-Document Matrix

- This step involves creating a term-by-document matrix that maps terms to documents using weights such as term frequency or TF-IDF, which evaluates a term's importance within a document and the entire collection.

- **Metadata and Indexed Documents Database**: Stores the structured, processed pages, ensuring up-to-date and searchable information for the Responding Cycle.
- **Metadata and Indexed Documents Database**: Stores preprocessed and indexed web pages with metadata like relevance scores and content type for quick retrieval.
    - **Indexed Documents DB**:
        - A repository of web pages that have already been **crawled** and **indexed**. Contains structured information such as page titles, keywords.
    - **Metadata**:
        - Additional information associated with each document, including:
            - Page relevance scores.
            - Content type (e.g., text, images, videos).

2. Response Cycle

- It is the interfacing with the user.
- The two main components of the responding cycle are the query analyzer and document matcher/ranker.
- The two main components of the responding cycle are the query analyzer and document matcher/ranker.
- **User and Search Query**: Users enter their search query, reflecting their information need.
- **Query Analyzer**: Processes the query by extracting keywords, correcting spelling, and applying NLP for better understanding.
        - **Keyword Extraction**: Identifies key terms and phrases from the query.
        - **NLP Techniques**: Understands the query's intent, whether informational, transactional, or navigational.
        - **Spelling Corrections**: Fixes misspelled words.
        - **Stemming**: Reduces words to their root form. For example, "running," "runs," and "runner" are reduced to "run" to unify similar queries.
        - **Stopword Removal**: Eliminates common words like "the," "is," and "and" that don't affect the search intent.

- **Processed Query**: The refined query is sent to the Indexed Documents Database for matching.
- Documents that are already **crawled and indexed** are stored here for quick retrieval during search queries
- **Document Matcher/Ranker**: Compares the query against indexed documents and assigns relevance scores based on keywords, authority, and metadata.
- **Retrieved Pages (Search Results)**: Returns ranked documents as search results, completing the cycle.

UNIT -IV

TOPIC -10

SEARCH ENGINES-2, SEARCH ENGINE OPTIMIZATION

Explanation of Structure of a typical Search Engine with an example

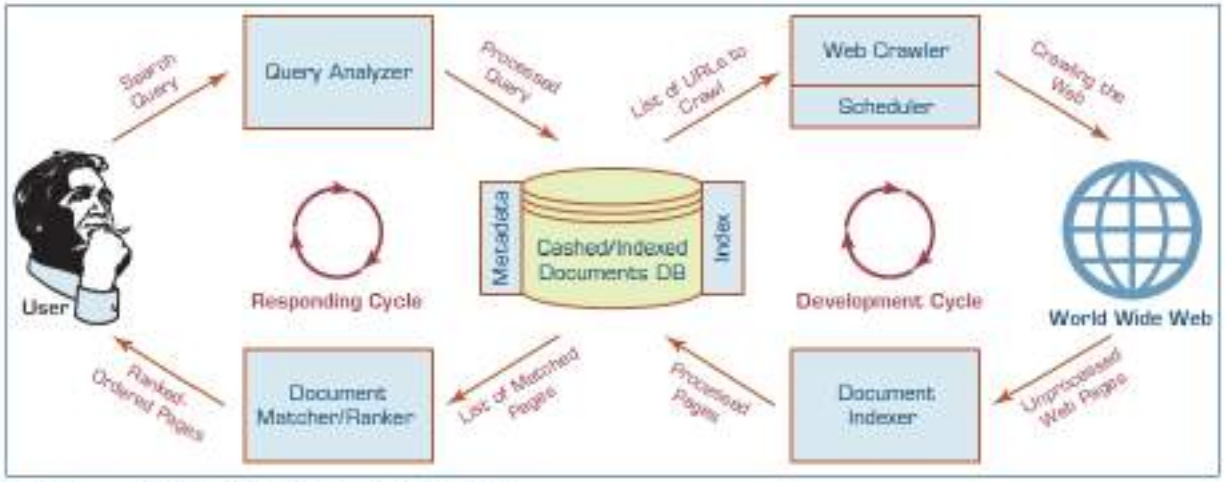


FIGURE Structure of a Typical Internet Search Engine.

User and Search Query: A user searches for "healthy breakfast recipes with oats," indicating a need for specific information about breakfast recipes using oats.

Query Analyzer: In this, The query is refined by extracting relevant keywords, correcting spelling, applying stemming, and removing stopwords for improved accuracy.

Processed Query: The cleaned query ("healthy breakfast recipe oats") is sent to the Indexed Documents Database for document matching.

Metadata and Indexed Documents Database: The database stores preprocessed pages with metadata, including keywords, content type, and relevance score for efficient retrieval.

List of URLs to Crawl: The search engine maintains a prioritized list of URLs to crawl, including new, discovered, and frequently updated URLs.

Web Crawler: Collects web page data by navigating links and downloading content like text, images, and videos.

Visits URLs like *"healthyrecipes.com/breakfast.html"* and collects data.

Downloads content such as:

Text: Recipes and descriptions.

Images: Photos of oatmeal dishes.

Links: Discovers new pages to crawl.

A Web crawler starts with a list of URLs to visit, which are listed in the scheduler and often are called the *seeds.*

As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (i.e., the scheduler).

Document indexer Removes duplicate or unnecessary content (e.g., ads).

Assigns metadata such as page title, keywords, and relevance.

Example: A page about *"Healthy Breakfast Recipes"* is indexed with metadata like:

Keywords: *"breakfast," "oats," "healthy," "recipes."*

Content Type: *Text and Image.*

Example: A page about *"Top 10 Healthy oatmeal Recipes"* is indexed with metadata like:

Keywords: *"breakfast," "oats," "healthy," "recipes."*

Content Type: *Text and Image.*

URL	Title	Keywords	Content Type	Last Updated
`healthyrecipes.com/oats.html`	`Top 10 Healthy Oatmeal Recipes`	healthy, oats, recipes, breakfast	Text, Image	November 2024
`nutrition.com/breakfast.html`	`Healthy Breakfast Ideas with Oats`	healthy, oats, breakfast, ideas	Text	November 2024
`recipes.com/oats.html`	`Oats for Breakfast: Quick Recipes`	oats, breakfast, quick, recipes	Text, Image	November 2024

Metadata and Indexed Documents Database: Stores the structured, processed pages with metadata (keywords, content type, relevance) for efficient retrieval during search queries.

The database stores processed documents along with their **metadata**:**Example:**

Page Title: "Top 10 Healthy Breakfast Recipes with Oats."

Keywords: "healthy," "breakfast," "oats," "recipes."

Content Type: Text, Image.

Document Matcher/Ranker:

The system matches the processed query with documents:

Relevance Score is Calculated based on keyword matches (e.g., "healthy" and "oats").

Other factors like **page rank** and **freshness** are considered.

Example: The page "Top 10 Healthy Breakfast Recipes with Oats" gets a high relevance score of 95 and ranks higher.

Retrieved Pages (Search Results):

The system displays ranked results to the user:

1st Result: "Top 10 Healthy Breakfast Recipes with Oats."

2nd Result: "5 Quick Breakfast Ideas for a Healthy Start."

Search Engine Optimization (SEO):

- Purpose: SEO aims to improve the visibility of a website in unpaid (organic) search engine results, increasing traffic by ranking higher in search results.
- Methods: SEO involves optimizing website content, HTML, and coding to match relevant keywords and remove barriers(Poor Website Structure, Slow Page Load Speed)for search engine indexing.
- **Backlinks**
- Definition: Backlinks are links from other websites that point to your site. To improve search engine rankings, websites aim to get other trusted sites to link back to their content, which signals to search engines that the site is valuable and relevant.
- Search Engine Crawling: Crawlers (spiders) automatically visit websites, gather content like text, images, links, and metadata, and index it. This process happens continuously, without needing manual URL submissions, allowing search engines to rank content based on relevance to search queries.
- Ranking Tactics: SEO tactics include using relevant keywords, updating content regularly, and optimizing metadata (title tags, descriptions) to boost search rankings and drive traffic.

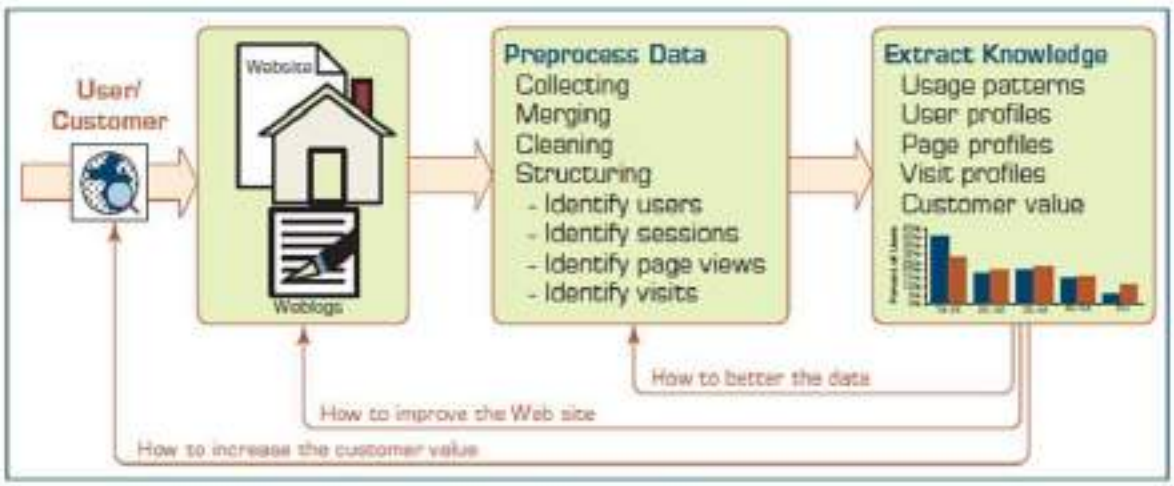
- SEO techniques can be broadly divided into two categories: white-hat SEO and black-hat SEO.
- **White-Hat SEO:**
    - These techniques align with search engine guidelines and involve no deception.
    - White-hat SEO focuses on creating quality content for users, not just for search engines, and making it easily accessible to crawlers.
    - It is about ensuring that the content indexed by search engines matches what users will see.
    - White-hat methods generally lead to long-term, sustainable results.
- **Black-Hat SEO:**
    - These methods aim to improve rankings using deceptive techniques that are not approved by search engines.
    - Examples include hiding text in the same color as the background or using cloaking (showing different content to search engines and users).
    - Black-hat techniques can result in penalties from search engines, which may include ranking reductions or removal from the search engine's index.
    - The webpage has visible content for users like:*"Explore our premium collection of coffee makers designed to suit every taste."*
    - When a search engine crawler visits the website, it is served a page filled with keyword-rich content and detailed descriptions of "Best Coffee Makers.“
    - The website may temporarily rank higher using cloaking and keyword manipulation but risks severe penalties like demotion or removal from search results, if detected by search engines.

UNIT -IV

TOPIC -11

WEB USAGE MINING(Web Analytics)

- Web usage mining (also called Web analytics) is the process of extraction of useful information from data generated through Web page visits and transactions. Analysis of the information collected by Web servers can help us better understand user behavior.
- Analysis of this data is often called clickstream analysis. By using the data and text mining techniques, a company might be able to discern interesting patterns from the clickstreams.
- For example, 60% of visitors who searched for “hotels in Delhi” had searched earlier for “airfares to Delhi.” Such information could be useful in determining where to place online advertisements.
- Clickstream analysis might also be useful for knowing when visitors access a site. For example, if a company knew that 70% of software downloads from its Web site occurred between 7 and 11 p.m., it could plan for better customer support and network bandwidth during those hours.
- Figure shows the process of extracting knowledge from clickstream data and how the generated knowledge is used to improve the process, improve the Web site, and most important, increase the customer value.



Extraction of Knowledge from Web Usage Data.

This diagram illustrates the process of extracting knowledge from web usage data, which involves analyzing user interactions on a website to gain insights.

- **User Interaction (Input)**

- Users interact with the website by browsing pages, performing actions, or generating web logs (records of user activity).

Weblogs Collection

- The website records user activities (such as page views and visits) in logs, which serve as the raw data source.

Data Preprocessing

- The raw data is cleaned, structured, and merged.
- Key tasks include:
- Identifying unique users.
- Identifying sessions (a single user's series of interactions).
- Identifying specific page views.
- Identifying visits (groups of sessions).

Knowledge Extraction

- After preprocessing, analytical techniques are applied to extract meaningful information.
- Insights include:
- Usage patterns (e.g., popular times or pages).
- User profiles (grouping users based on behavior).
- Page profiles (understanding page-level engagement).
- Visit profiles (analyzing session data).
- Customer value (measuring contributions of users to the business).

Application of Insights

- The extracted insights are used to:
- Improve the website's design and functionality.
- Enhance customer value by offering a better user experience.
- Tailor marketing and business strategies based on user behavior.
- This process ultimately aims to improve user engagement and achieve business goals.

Feedback

- Extracted knowledge helps to decide how to better data , how to improve web site, how to improve customer

Web Analytics Technologies

- Web analytics tools are increasingly popular for measuring, collecting, and analyzing Internet data to optimize Web usage and transform online business. Web analytics supports e-business and market research by tracking traffic changes, visitor numbers, page views, and trends to evaluate e-commerce effectiveness. For example, after launching a discount campaign, web analytics can track a spike in website visits and purchases, helping a retail business evaluate the campaign's success.

- **There are two main categories of Web analytics**: off-site and on-site. Off-site Web analytics refers to Web measurement and analysis about you and your products that takes place outside your Web site. It includes the measurement of a Web site’s potential

audience (prospect or opportunity), share of voice (word-of-mouth), and buzz (comments or opinions) that is happening on the Internet.

- For on-site Web analytics, there are two technical ways of collecting the data.
- The **first and more traditional method** is the **server log file analysis,** where the Web server records file requests made by browsers. The **second method** is **page tagging**, Page tagging uses a small JavaScript code on a webpage to send data to an external server whenever the page loads or a click happens. This helps track user activity for analytics.
- In addition to these two, other data sources like email, direct mail campaigns, sales history, and social media can enhance website behavior data.

Web Analytics Metrics

- Web analytics provides real-time data and insights from various sources to track marketing performance, optimize strategies, and measure ROI(return on investment).
- For example, an online retailer can use web analytics to monitor real-time sales data during a flash sale, enabling them to adjust promotions or inventory quickly to maximize revenue.Web analytics provides a broad range of metrics, there are four categories of metrics that are generally actionable and can directly impact your business objectives. These categories include

A. **Web site usability**: How were they using my Web site?

B. Traffic sources: Where did they come from?

C. Visitor profiles: What do my visitors look like?

A. Web Site Usability

- **Web Site Usability** refers to the ease with which visitors can navigate, interact with, and benefit from a website.The following points help to improve Web Site Usability
  - 1. Page views.
  - 2. Time on site
  - 3. Downloads
  - 4. Click map.

1. Page views.

- The most basic of measurements, this metric is usually presented as the “average page views per visitor.”Low page views may indicate issues with website design or a mismatch between marketing messages and site content.For example, if an online ad promotes a product but the landing page doesn't feature it prominently, visitors may leave without exploring other pages.

2. Time on site.

- Time on site measures how long visitors engage with your website, suggesting they are interacting with content or considering a purchase. However, it should be analyzed alongside page views to ensure visitors aren’t struggling to find content, which could indicate navigation or accessibility issues.For example, if a visitor spends 10 minutes on a product page but only views one page, it may indicate they are having trouble finding additional product details or are stuck on a page without further direction.

3. Downloads.

- This includes PDFs, videos, and other resources you make available to your visitors. If your Web statistics, for example, reveal that 60% of the individuals who watch a demo video also make a purchase, then you’ll want to strategize to increase viewership of that video.

4. Click map.

- Most analytics programs can show you the percentage of clicks each item on your Web page received. This includes clickable photos, text links in your copy, downloads, and, of course, any navigation you may have on the page.

B. Traffic Sources

- Web analytics helps identify where **Web traffic comes from**, such as search engines, referrals, or direct visits.It can also track traffic generated by offline or online advertising campaigns with minimal effort.
- Following are some of the Traffic Sources

1. Referral Web sites.

2. Search engines

3. Direct.

4. Offline campaigns

5. Online campaigns(search engine advertising campaign,e-mail campaign)

1. Referral Web sites: Other Web sites that contain links that send visitors directly to your Web site are considered referral Web sites. Your analytics program will identify each referral site your traffic comes from.

2. Search engines: Data in the search engine category is divided between paid search and organic (or natural) search. You can review the top keywords that generated Web traffic to your site and see if they are match your products and services.

3. Direct : Direct searches occur when users bookmark a page or type the URL directly into their browser, often from sources like business cards, brochures, or ads.

4. Offline campaigns :Offline campaigns can be tracked by using special URLs in ads, like "www.mycompany.com/offer50," to see how many people visit your website.

5. Online campaigns: If you are running a search engine advertising campaign, or even e-mail campaign, you can measure individual campaign effectiveness by simply using a dedicated URL similar to the offline campaign strategy.

C. Visitor Profiles

- Segmentation in Web analytics helps you create detailed user profiles by combining data from different reports to better understand and target your audience.
- The following five points aim to analyze visitor behavior and enhance marketing strategies through segmentation and targeted insights:
- **Keywords**
- **Content Groupings**
- **Geography**
- **Time of Day**
- **Landing Page Profiles**
- **Keywords**: Identify visitor intent through search terms to tailor content that matches their understanding of your products or services. For example, search phrases reflecting your product descriptions indicate familiarity with your offerings through advertisements or brochures.
- **Geography**: Analytics permits you to see where your traffic geographically originates, including country, state, and city locations

- **Time of Day**: Web traffic generally has peaks at the beginning of the workday, during lunch, and toward the end of the workday. You can analyze this data to determine when people browse versus buy and also make decisions on what hours you should offer customer service.
- **Landing Page Profiles**: If you structure your various advertising campaigns properly, you can drive each of your targeted groups to a different landing page,which your Web analytics will capture and measure.

UNIT -IV

TOPIC -12

WEB ANALYTICS MATURITY MODEL AND WEB ANALYTICS TOOLS

WEB ANALYTICS MATURITY MODEL

- The **Web Analytics Maturity Model** is a framework that helps organizations evaluate their level of expertise in using web analytics. It outlines different stages, from basic tracking of website data to advanced data-driven decision-making.
- A **maturity model** is a formal depiction of critical(essential) dimensions and their competency(skill) levels of a business practice. Collectively, these dimensions and levels define the **maturity level of an organization** in that area of practice.
- For example, a small online retailer initially relies on guesswork (ad hoc) to decide which products to stock. As the business matures, it implements a data-driven approach by analyzing customer purchasing trends (proficiency), setting up automated inventory management systems and optimizing stock levels based on predictive analytics (optimization).
- An example is the simple business analytics maturity model, moving from simple descriptive measures to predicting future outcomes, to obtaining sophisticated decision systems (i.e., Descriptive Analytics ➔ Predictive Analytics ➔ Prescriptive Analytics).
- For Web analytics perhaps the most comprehensive model was proposed by Stephane Hamel.

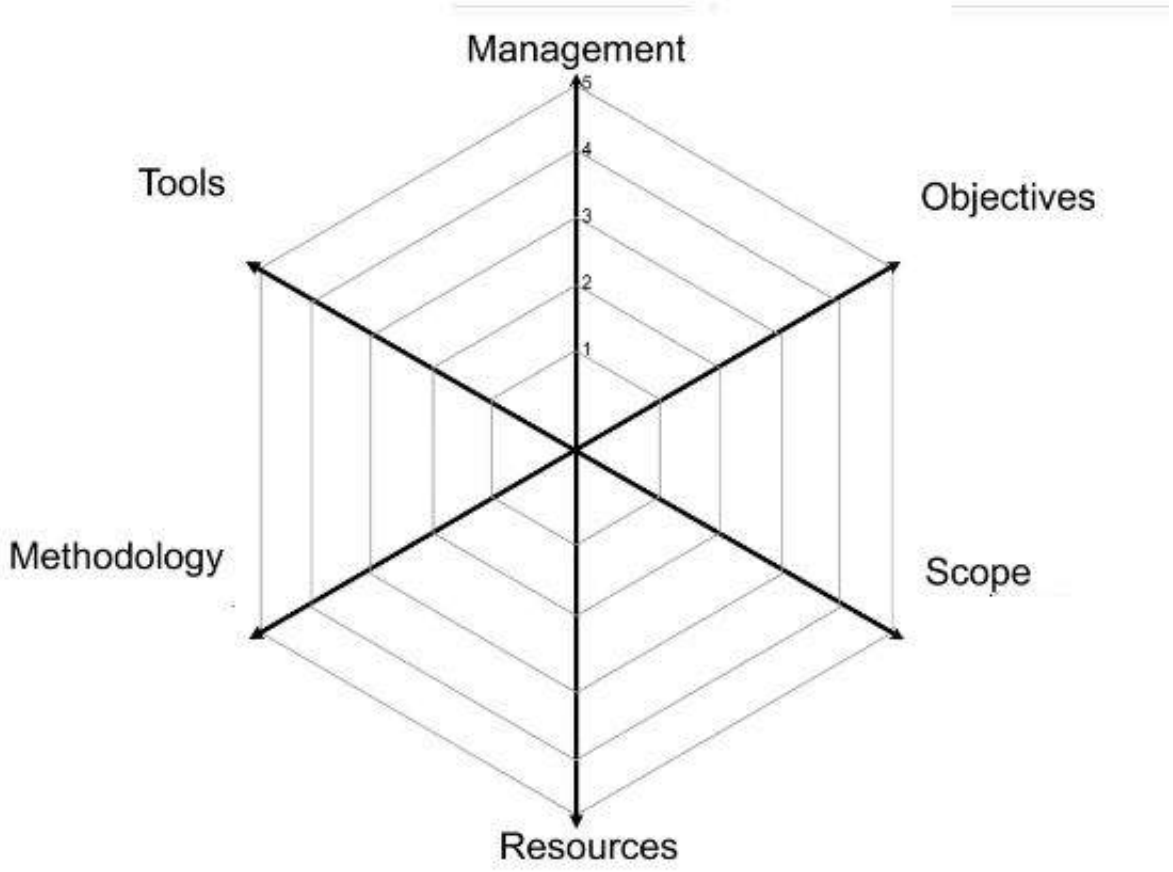


FIGURE: A Framework for Web Analytics Maturity Model

- For Web analytics , the most comprehensive model was proposed by Stephane Hamel. In this model, Hamel used **six dimensions-**

(1) Management,

(2) Objectives Definition,

(3) Scoping,

(4) The Analytics Team and Expertise,

(5) Methodology,

(6)Tools, **for each dimension he used six levels of proficiency/competence.**

The six levels are indications of analytical maturity ranging from "0-Analytically Impaired" to "5-Analytical Competitor.“

0. Impaired

1. Initiated

2. Operational

3. Integrated

4. Competitor

5. Addicted

In this model, Hamel used **six dimensions-**

1. Management:

It refers to the leadership's involvement in web analytics initiatives and their ability to create a data-driven culture.

2. Objectives : It is the clarity of goals and objectives tied to analytics efforts.

3. Scope refers to the range and extent of web analytics implementation within an organization.

4. Resources:

The availability of skilled personnel and financial investment.

5. Methodology:

The processes and frameworks used for analyzing and interpreting data.

6. Tools:

The platforms and technologies used for analytics.

- Essentially, the **six levels are indications of analytical maturity ranging** from "0-Analytically Impaired" to "5-Analytical Competitor.“

0. Impaired

1. Initiated

2. Operational

3. Integrated

4. Competitor

5. Addicted

- **Impaired**: Organizations at this level have little or no web analytics in place, resulting in limited or inaccurate data collection.

- **Initiated**: Basic analytics tools are set up, and some data is collected, but there is little analysis or action taken based on it.
- **Operational**: Web analytics are regularly used, with a focus on tracking key metrics and reporting, and some insights are used to improve decision-making.
- **Integrated**: Web analytics are integrated across multiple channels, and data is used holistically to optimize business strategies and customer experiences.
- **Competitor**: The organization uses analytics to benchmark performance against competitors, making data-driven decisions to gain a competitive edge.
- **Addicted**: Analytics are deeply embedded in every aspect of the business, with constant optimization, real-time data use, and a culture of data-driven decision-making.

Examples of **web analytics maturity levels** : **tools**

0: Impaired:

- . No web analytics tool,

1: Initiated:

- Basic tracking tools are implemented for measuring simple metrics like page views and visitors

2: Operational:

- Web analytics tools are in regular use, Reports are generated to inform decisions, though insights may still be limited.

3: Integrated:

Data from multiple platforms (website, social media, etc.) is integrated, providing a comprehensive view of user behavior across various channels.

4: Competitor:

- Web analytics tools are used to benchmark against competitors, and the organization uses comparative data to gain a competitive advantage by analyzing competitors' digital strategies.

5:Addicted :

- Web analytics tools are deeply embedded into the business strategy, with continuous optimization and real-time decision-making

Web Analytics Tools

- There are plenty of Web analytics applications (downloadable software tools and Web based /on-demand service platforms) in the market.
- The following are among the most popular free (or almost free) Web analytics tools:

1. Google Web Analytics (GOOGLE.COM/ANALYTICS)

2. Yahoo! Web Analytics (WEB.ANALYTICS.YAHOO.COM)

3. Open Web Analytics (OPENWEBANALYTICS.COM)

4. Firestat (FIRESTATS.CC)

5. Site Meter (SITEMETER.COM)

6. AWSTATS (AWSTATS.ORG)

GOOGLE WEB ANALYTICS (GOOGLE.COM/ANALYTICS)

This is a service offered by Google that generates detailed statistics about a Web site's traffic and traffic sources and measures conversions and sales.

YAHOO! WEB ANALYTICS (WEB.ANALYTICS.YAHOO.COM)

Yahoo! Web analytics is Yahoo!'s alternative to the dominant Google Analytics. It is a comprehensive Web analytics tool, it has graphs, custom-designed (and printable) reports, and real-time data tracking.

OPEN WEB ANALYTICS (OPENWEBANALYTICS.COM):

Open Web Analytics (OWA) is a popular open source Web analytics software that anyone can use to track and analyze how people use Web sites and applications.

FIRESTAT (FIRESTATS.CC)

FireStats is a simple and straightforward Web analytics application written in PHP/ MySQL. It supports numerous platforms and set-ups including C# sites, Django sites.

SITE METER (SITEMETER.COM)

Site Meter is a service that provides counter and tracking information for Web sites.

By logging IP addresses and using JavaScript or HTML to track visitor information , Site Meter provides Web site owners with information about their visitors, including how they reached the site, the date and time of their visit, and more.

AWSTATS (AWSTATS.ORG)

AWStats is an open source Web analytics reporting tool, suitable for analyzing data from Internet services such as Web, streaming media, mail, and FTP servers. AWStats parses and analyzes server log files, producing HTML reports.