

Module-4:

Data Collection and Pre-processing, Sources and types of Data, Data Modelling, Session and Visitor Analysis, Cluster Analysis and Visitor segmentation, Association and Correlation Analysis, Analysis of Sequential and Navigational Patterns, Classification and Prediction based on Web User Transactions.

Web Analytics: Key Concepts and Techniques

Web analytics involves the collection, measurement, analysis, and reporting of web data to understand and optimize web usage. Here are the critical components and techniques in web analytics:

Data Collection and Pre-processing

Data collection in web analytics involves gathering information from various sources such as web servers, application servers, and web clients. Techniques include using log files, which record every request made to the server, JavaScript tags embedded in web pages to collect data when the page is loaded in a browser, and API integration to collect data directly from various platforms like social media and email marketing tools. Pre-processing this data is crucial to ensure accuracy and relevance. It involves cleaning the data to remove irrelevant or redundant information and correct errors, transforming raw data into a usable format such as parsing log files or normalizing fields, and integrating data from different sources to provide a comprehensive view.

Sources and Types of Data

The sources of web data include web logs, which capture HTTP requests, user profiles generated from registrations and interactions, and third-party data from external sources like social media platforms and advertising networks. The types of data can be categorized into structured data, which has a defined format such as user profiles or transaction records; unstructured data, which is free-form text data like comments or reviews; and semi-structured data, which has some level of organization, such as JSON or XML files. Each type of data requires different methods of processing and analysis to extract meaningful insights.

Data Modelling

Data modeling is the process of representing and organizing data in a structured format to facilitate analysis. Common techniques in data modeling include creating Entity-Relationship (ER) diagrams to visually represent data models and relationships, normalizing data to reduce redundancy and improve integrity, and using Online Analytical Processing (OLAP) for multi-dimensional analysis. These techniques help in organizing the data efficiently, making it easier to retrieve and analyze information to support decision-making processes.

Session and Visitor Analysis

Session analysis involves examining groups of user interactions with a website that occur within a given timeframe. Key metrics include session duration, pages per session, and bounce rate. Visitor analysis, on the other hand, focuses on understanding unique visitors, returning visitors, and visitor behavior patterns. Unique visitors are distinct individuals visiting the site,

returning visitors are those who visit more than once, and visitor behavior involves patterns in how visitors navigate and interact with the site. Analyzing these aspects helps in understanding user engagement and identifying areas for improvement.

Cluster Analysis and Visitor Segmentation

Cluster analysis is used to group users into clusters based on similar behaviors or attributes. Techniques like K-Means clustering, which partitions data into K clusters where each data point belongs to the cluster with the nearest mean, and hierarchical clustering, which builds a hierarchy of clusters using either a bottom-up or top-down approach, are commonly used. Visitor segmentation aims to identify and categorize different types of visitors for targeted marketing and personalization. Segments can be based on demographics, behavior, geography, and other attributes. These analyses help in creating personalized experiences for different user groups.

Association and Correlation Analysis

Association analysis identifies relationships between variables in large datasets. Techniques like the Apriori algorithm, which finds frequent itemsets and generates association rules, and the FP-Growth (Frequent Pattern Growth) method, a faster alternative to Apriori, are commonly used. Correlation analysis measures the strength and direction of the relationship between two variables using metrics like the Pearson correlation coefficient and Spearman rank correlation. These analyses help in uncovering patterns and relationships within the data that can inform strategic decisions.

Analysis of Sequential and Navigational Patterns

Sequential pattern analysis identifies sequences of events or actions performed by users over time. Techniques include sequential pattern mining, which identifies frequent sequences in data, and Markov chains, which model the probability of transitioning from one state (page) to another. Navigational pattern analysis focuses on understanding how users navigate through a website, using techniques like path analysis to examine the paths users take and heatmaps to visually represent where users click on a page. These analyses provide insights into user behavior and site usability.

Classification and Prediction Based on Web User Transactions

Classification in web analytics categorizes data into predefined classes or groups. Techniques include decision trees, which classify data by splitting it into branches based on feature values; naive Bayes, a probabilistic classifier based on Bayes' theorem; and support vector machines (SVM), which find the hyperplane that best separates different classes. Prediction aims to forecast future user behavior based on historical data using techniques like regression analysis, which predicts a continuous outcome based on one or more predictors; time series analysis, which analyzes data points collected or recorded at specific time intervals; and neural networks, which model complex relationships between inputs and outputs. These methods provide valuable insights for optimizing website content, enhancing user experience, and driving business decisions.