Machine Learning Algorithms for Big Data Analytics: Introduction, Estimating the relationships, Outliers, Variances, Probability Distributions, and Correlations, Regression analysis, Finding Similar Items, Similarity of Sets and Collaborative Filtering, Frequent Itemsets and Association Rule Mining. Text, Web Content, Link, and Social Network

Analytics: Introduction, Text mining, Web Mining, Web Content and Web Usage Analytics, Page Rank, Structure of Web and analyzing a Web Graph, Social Network as Graphs and Social Network Analytics:

Module-5

Q. 09	a	What is Machine Learning? Explain different types of Regression Analysis.
	b	Explain with neat diagram K-means clustering.
	c	Explain Naïve Bayes Theorem with example.
		OR
Q. 10	a	Explain five phases in a process pipeline text mining.
	b	Explain Web Usage Mining.

1. What is Machine Learning ? Explain different types of Regression Analysis

What is Machine Learning (ML)?

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that allows systems to learn patterns and make decisions or predictions based on data without being explicitly programmed. ML models improve their performance as they are exposed to more data over time.

Key ML tasks include:

1. Prediction: Forecasting outcomes based on input data.
2. Classification: Assigning data points to predefined categories.
3. Clustering: Grouping similar data points together.
4. Optimization: Finding the best solution among many alternatives.

Types of Regression Analysis in ML

Regression analysis is a supervised learning technique used to understand relationships between variables, specifically to predict a dependent variable based on one or more independent variables. Here are the common types explained in detail:

1. Simple Linear Regression

- **Description**: Models the relationship between a single independent variable (x) and a dependent variable (y) using a straight line.
- **Equation**:

$$y = \beta_0 + \beta_1 x + \epsilon$$

  Where:

  - y: Dependent variable (output/predicted value).
  - x: Independent variable (input feature).
  - β_0, β_1: Coefficients (intercept and slope).
  - ϵ: Error term (accounts for deviation).
- **Example**: Predicting house prices based on square footage.

2. Multiple Linear Regression

- **Description**: Extends linear regression by involving multiple independent variables ($x_1, x_2, ..., x_n$).
- **Equation**:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

  Where $x_1, x_2, ..., x_n$ are the independent variables.
- **Example**: Predicting house prices based on multiple factors like square footage, location, and number of bedrooms.

3. Polynomial Regression

- **Description**: Captures non-linear relationships between independent and dependent variables by including polynomial terms of x.
- **Equation**:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_n x^n + \epsilon$$

  - Degree of the polynomial (n) determines the model's complexity.
- **Example**: Predicting growth rate over time where the relationship isn't linear.

4. Logistic Regression

- **Description**: Used for classification problems where the output is binary (e.g., 0 or 1, True or False). It predicts the probability that a given input belongs to a specific category.
- **Equation**:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

  - Outputs probabilities between 0 and 1 using the sigmoid function.
- **Example**: Determining whether an email is spam or not spam.

5. Ridge Regression (L2 Regularization)

- **Description**: Adds a penalty term proportional to the square of the coefficients (β_j^2) to reduce overfitting.
- **Modified Loss Function**:

$$\text{Loss} = \sum (y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2$$

  - λ: Regularization parameter that controls the penalty's strength.
- **Example**: Predicting sales while avoiding overfitting due to noisy data.

7. Non-Linear Regression

- **Description**: Models relationships where the change in the dependent variable isn't proportional to the independent variables. The relationship may be modeled using curves or higher-order equations.
- **Equation Example**:

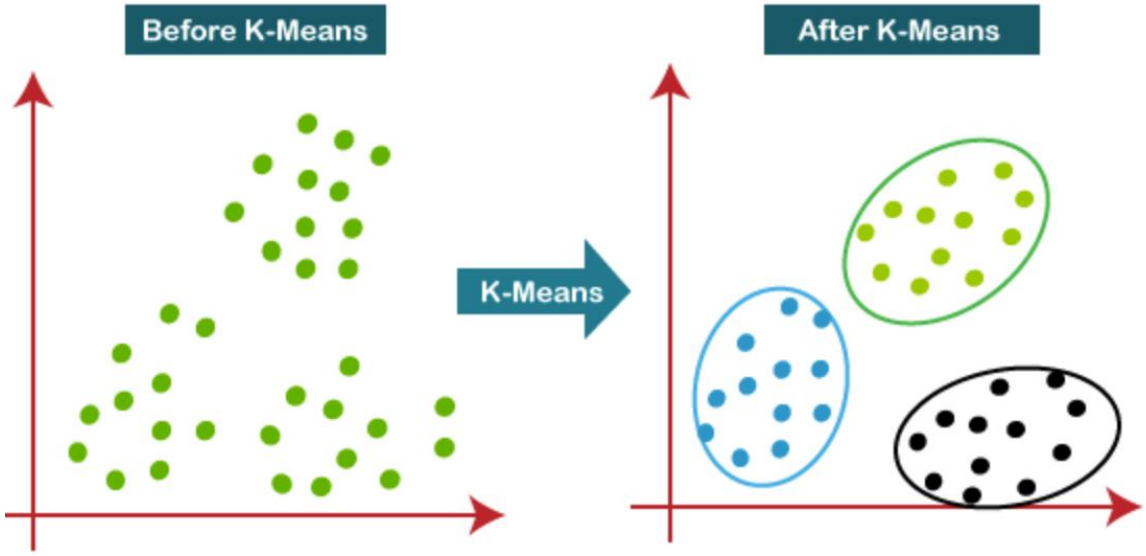
$$y = a + b \cdot x + c \cdot x^2 + d \cdot \log(x) + \epsilon$$

- **Example**: Modeling the growth of a population, where growth rate decreases over time.

8. Stepwise Regression

- **Description**: Automatically selects the most significant independent variables by adding or removing predictors based on statistical criteria like p-values.
- **Example**: Building models for medical diagnosis by identifying the most impactful symptoms.

2. Explain With neat diagram K-means Clustering



K-Means Clustering is an **unsupervised machine learning algorithm** used to group a dataset into a predefined number of clusters (k) based on similarity. It minimizes the variance within each cluster, making data points within the same cluster more similar to each other than to those in other clusters.

How K-Means Works

1. **Initialization**:
    - Choose the number of clusters, k
    - Randomly initialize k cluster centroids (points in the dataset).
2. **Assignment Step**:
    - For each data point, compute the distance to each centroid.
    - Assign each data point to the nearest centroid. This forms k clusters.
3. **Update Step**:
    - Calculate the mean (average) of all data points in each cluster.
    - Update the cluster centroid to this mean.
4. **Repeat**:
    - Repeat steps 2 and 3 until either:
        - The centroids no longer change significantly.

- A maximum number of iterations is reached.
- Cluster assignments stabilize.

Advantages

1. **Simplicity**: Easy to implement and computationally efficient.
2. **Scalability**: Works well with large datasets.
3. **Interpretability**: Clusters can often be visually interpreted in 2D or 3D.

Disadvantages

1. **Choosing kkk**: Requires the number of clusters to be specified beforehand.
2. **Sensitive to Initialization**: Random initialization of centroids can lead to different results.
3. **Non-convex Data**: Assumes clusters are spherical, so it struggles with non-linear or irregularly shaped clusters.
4. **Outliers**: Sensitive to outliers, as they can distort the centroids.

Applications

1. **Customer Segmentation**:
    - Grouping customers by purchasing behavior for targeted marketing.
2. **Image Compression**:
    - Reducing the number of colors in an image by clustering similar colors.
3. **Anomaly Detection**:
    - Identifying points that don’t belong to any cluster.
4. **Document Clustering**:
    - Grouping similar documents or articles based on content.

Mathematics Behind K-Means

1. **Distance Metric**:
   - Commonly uses **Euclidean Distance** to measure similarity:

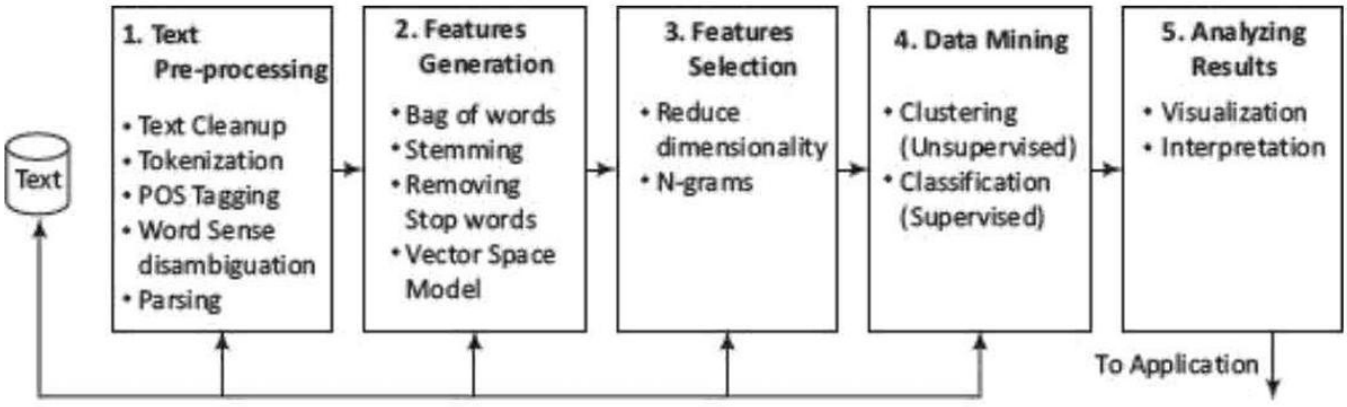
$$d(x, c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}$$

Where:

- x: A data point.
- c: A cluster centroid.
- n: Number of dimensions.

3. Explain Naïve Bayes Theorem With example

4. Explain five phases in a process pipeline text mining



1. Text Pre-processing

This is the foundational step that prepares raw text data for analysis by cleaning and structuring it. Key activities include:

- **Text Cleanup**:
  - Removing noise such as punctuation, special characters, HTML tags, and stop words (common words like "and," "the," "is").
  - Correcting typos or inconsistencies in the text.
- **Tokenization**:
  - Splitting the text into individual units (tokens), such as words or phrases.

- Example: *"Text mining is useful." → [Text, mining, is, useful]*

- **Part-of-Speech (POS) Tagging**:
  - Labeling words in a sentence with their grammatical roles (e.g., noun, verb, adjective).
  - Example: *"Text mining" → Text/Noun, mining/Noun*.
- **Stemming and Lemmatization**:
  - Reducing words to their root or base forms.
  - Example: *"Running" → "Run."*
- **Word Sense Disambiguation**:
  - Resolving ambiguity in words with multiple meanings based on context.
  - Example: *"Bank" → financial institution or riverbank.*

---

2. Feature Generation

This phase converts pre-processed text into structured data (features) that can be analyzed.

- **Bag-of-Words Model**:
  - Representing text as a collection of unique words, ignoring order.
  - Example: *"Text mining is useful" → {Text: 1, Mining: 1, Useful: 1}*
- **TF-IDF (Term Frequency-Inverse Document Frequency)**:
  - Assigning importance to words based on their frequency in a document relative to the entire dataset.
- **N-Grams**:
  - Capturing sequences of nnn consecutive words to preserve context.
  - Example: *"Text mining is useful" → Bigrams: [Text mining, Mining is, Is useful]*.
- **Named Entity Recognition (NER)**:
  - Identifying entities such as names, dates, and locations in the text.
  - Example: *"Barack Obama was born in Hawaii" → [Barack Obama: Person, Hawaii: Location]*.

---

3. Text Transformation

This phase involves transforming the structured text data into a representation suitable for machine learning or statistical analysis.

- **Vectorization**:
    - Representing text as numerical vectors (e.g., word embeddings, TF-IDF scores).
- **Dimensionality Reduction**:
    - Reducing the number of features while retaining meaningful information (e.g., PCA or Singular Value Decomposition).
- **Clustering or Classification Labels**:
    - Assigning labels to text for supervised or unsupervised learning.

---

4. Data Mining/Analysis

This phase applies machine learning algorithms or statistical techniques to derive insights or make predictions.

- **Text Clustering**:
    - Grouping similar documents into clusters based on features.
    - Example: News articles about "sports," "politics," and "technology."
- **Classification**:
    - Categorizing documents into predefined categories.
    - Example: Identifying emails as "spam" or "not spam."
- **Sentiment Analysis**:
    - Detecting the sentiment (positive, negative, neutral) expressed in text.
    - Example: *"The movie was fantastic"* → *Positive.*
- **Topic Modeling**:
    - Discovering hidden themes in text using techniques like Latent Dirichlet Allocation (LDA).

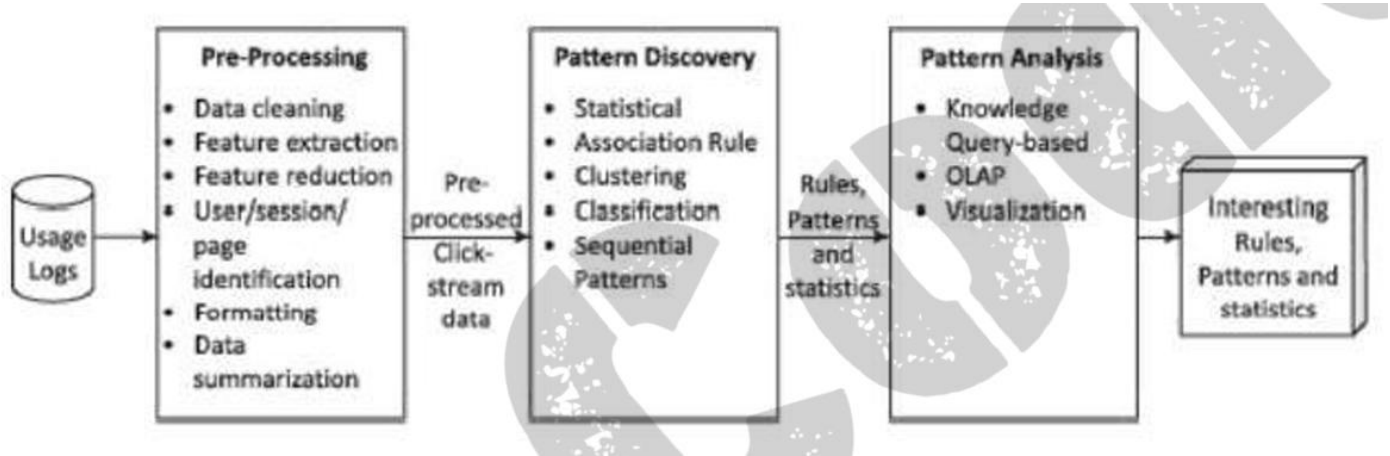
---

5. Interpretation and Evaluation

The final phase involves interpreting and validating the results to ensure they align with the objectives.

- **Visualization**:
    - Presenting results using graphs, word clouds, or charts for better understanding.
    - Example: Word frequency visualized as a word cloud.
- **Evaluation Metrics**:
    - Measuring the performance of the models used (e.g., precision, recall, F1 score for classification tasks).
- **Business Insights**:
    - Translating results into actionable decisions or strategies.
    - Example: Using customer reviews to improve product quality.

5. Explain web usage mining



Definition

Web Usage Mining involves discovering meaningful patterns from **web usage data** to understand user interactions, preferences, and trends. It helps in enhancing web applications, improving user experiences, and providing insights for decision-making.

Types of Data in Web Usage Mining

1. **Web Server Logs**:
    - Logs maintained by web servers that capture HTTP requests made by users.
    - Contains details like IP address, timestamps, requested URLs, and status codes.
2. **Browser Data**:
    - User activity data captured from browsers, including cookies, cache, and browsing history.
3. **Clickstream Data**:
    - Sequence of clicks made by users while navigating a website, showing page views and interactions.
4. **User Profiles**:
    - Data related to registered users, such as demographics, interests, and behavior patterns.

Phases of Web Usage Mining

1. **Data Collection**:
    - Collect raw web usage data from sources like server logs, cookies, and user sessions.
2. **Preprocessing**:
    - Clean and transform raw data to remove irrelevant entries.
    - Steps include removing bots, session identification, and data formatting.
3. **Pattern Discovery**:
    - Use techniques such as **clustering**, **classification**, and **association rule mining** to identify patterns.

        - Example: Discover which pages are often viewed together (association).

4. **Pattern Analysis**:
    - Analyze the discovered patterns to derive insights.
    - Visualize patterns using tools like graphs, charts, and heatmaps.

Techniques in Web Usage Mining

1. **Association Rule Mining**:
    - Identify relationships between web pages that are frequently visited together.
        - Example: Users visiting Page A often visit Page B.
2. **Clustering**:
    - Group users with similar browsing behavior into clusters.
        - Example: Classifying users into "shoppers," "readers," or "casual browsers."
3. **Classification**:
    - Predict user behavior based on predefined classes.
        - Example: Classify users as new or returning visitors based on patterns.
4. **Sequential Pattern Mining**:
    - Discover the order in which web pages are visited.
        - Example: User path: Home → Product → Cart → Checkout.

Applications of Web Usage Mining

1. **Personalization**:
    - Recommend products or content based on user behavior.

- Example: Amazon's product recommendations.

2. **Improving Website Design**:
   - Optimize navigation and layout to improve user experience.
3. **Business Intelligence**:
   - Analyze customer behavior to improve marketing and decision-making.
4. **E-commerce**:
   - Understand purchasing patterns to boost sales.
5. **Web Performance Improvement**:
   - Identify slow or unused pages for optimization.
6. **Fraud Detection**:
   - Detect unusual patterns of behavior that may indicate fraudulent activity.

Areas and Applications of Text Mining

1. **Natural Language Processing (NLP):**
   - **Definition:** NLP enables computers to analyze, understand, and derive meaning from human language.
   - **How it Works:** NLP algorithms, based on machine learning (ML), analyze a collection of sentences to infer rules statistically.
   - **Applications:**
     - Automatic text summarization
     - Sentiment analysis
     - Topic extraction
     - Named entity recognition
     - Parts-of-speech tagging
     - Relationship extraction

        - Stemming
    - **Common Uses:** Text mining, machine translation, and automated question answering.
2. **Information Retrieval (IR):**
    - **Definition:** The process of searching and retrieving a subset of documents from a large collection.
    - **How it Works:** IR uses metadata or full-text indexing to extract required information, often leveraging database technology.
    - **Applications:** Search engines use IR techniques like crawlers to retrieve information from diverse data sources.
3. **Information Extraction (IE):**
    - **Definition:** Extracting structured information from unstructured or semi-structured documents.
    - **How it Works:** Relies on machine learning and NLP. IE uses extraction patterns and semantic lexicons to derive relationships or desired content from text.
    - **Example:** Extracting information from images, audio, or video.
4. **Document Clustering:**
    - **Definition:** Groups text documents into clusters for organization, topic extraction, or fast information retrieval.
    - **Example:** Web document clustering helps users perform easier searches.
5. **Document Classification:**
    - **Definition:** Categorizes text documents into predefined classes or categories.
    - **Applications:** Useful for publishers, news sites, blogs, and areas with a large volume of content.
6. **Web Mining:**

- **Definition:** Application of data mining techniques to discover patterns from web data.
- **Applications:**
  - Analyzing web usage for understanding browsing behavior.
  - Enhancing web-based applications.

7. **Concept Extraction:**
   - **Definition:** Extracts concepts from textual data by grouping words and phrases into semantically similar categories.
   - **Application:** Used in text classification to improve understanding of textual content.

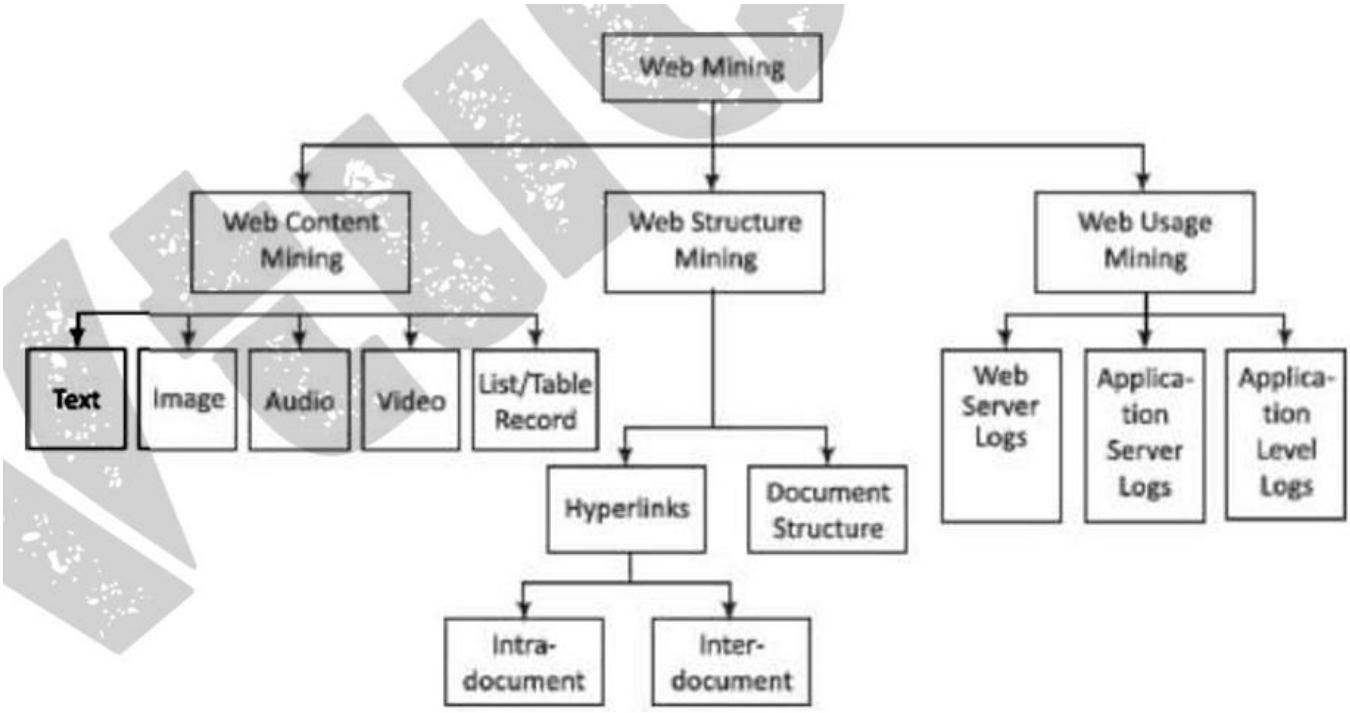


Figure 9.6 Web mining taxonomy

Web Mining: Definition

Web mining refers to the application of data mining techniques to extract useful patterns and insights from web data. This includes content, structure, and user interactions on the web. The primary goal is to improve web-based applications and understand user behavior.

Taxonomy of Web Mining

Web mining can be classified into three broad categories:

1. Web Content Mining

Web content mining focuses on extracting meaningful information from the content available on web pages, such as text, images, videos, and other multimedia.

- **Key Techniques:**
    - Natural Language Processing (NLP)
    - Information Retrieval
    - Clustering and Classification
    - Text Mining
- **Applications:**
    - Search engine optimization
    - Content-based recommendation systems
    - Sentiment analysis on social media content

2. Web Structure Mining

This type of mining examines the structure of websites, including hyperlinks and the hierarchy of web pages. It focuses on discovering relationships and patterns within the web's structural framework.

- **Key Techniques:**
    - Graph Theory (nodes and edges)
    - PageRank and HITS algorithms

- Link analysis

- **Applications:**
  - Website ranking and optimization
  - Identifying hubs and authorities in networks
  - Social network analysis

3. Web Usage Mining

Web usage mining analyzes user behavior through logs generated by web servers, such as clickstreams, browsing patterns, and session data.

- **Key Techniques:**
  - User Profiling
  - Sequence Pattern Mining
  - Collaborative Filtering
- **Applications:**
  - Personalized recommendations
  - Targeted advertising
  - Improving user experience through behavior analysis

PageRank

PageRank is an algorithm that measures the authority or importance of web pages based on the number and quality of hyperlinks pointing to them. It was introduced by Larry Page and Sergey Brin, co-founders of Google. The algorithm evaluates the significance of a page by analyzing its in-links (incoming links) and their relative authority.

Key Concepts of PageRank

1. **In-degree and Out-degree**:

- **In-degree**: The number of incoming links to a page. Pages with high in-degrees are often considered more authoritative.
- **Out-degree**: The number of outgoing links from a page.

2. **Authority Calculation**:
   - A page's rank is proportional to the authority of the pages linking to it and inversely proportional to their out-links.
   - The algorithm assumes that the link from a page distributes its rank equally among its out-links.
3. **Random Surfer Model**:
   - Models a user who randomly clicks links on the web.
   - Incorporates a "damping factor" (ddd, typically 0.85) to account for the probability of a user jumping to a random page instead of following links.
4. **Dead Ends**:
   - Pages without outgoing links reduce the flow of PageRank. This problem is mitigated by the random surfer model.

Applications of PageRank

1. **Search Engines**:
   - Determines the ranking of web pages in search results.
   - Helps prioritize high-authority content.
2. **Social Networks**:
   - Identifies influential users based on their connections.
3. **Recommendation Systems**:
   - Ranks items like products or movies based on relationships or reviews.

PageRank Implementation

PageRank uses two main algorithms for computation:

1. **Based on In-degrees**:
    - Considers the equal distribution of authority among the pages a page links to.
2. **Relative Authority of Parents over Children**:
    - Considers the authority of the parent (in-linking) pages and their weights.

Mathematical Formulation

The PageRank $PR(P)$ of a page P is defined as:

$$PR(P) = (1 - d) + d \cdot \sum_{i=1}^{n} \frac{PR(P_i)}{L(P_i)}$$

Where:

- $PR(P)$: PageRank of the page P.
- d: Damping factor (usually 0.85).
- P_i: A page linking to P.
- $L(P_i)$: The number of outgoing links on page P_i.
- n: Total number of pages linking to P.