Page No.
Date

# Module 6

g1) Illustrate the page rank algorithm in detail

=> (1) Page rank algorithm is an algorithm used to rank web pages based on their importance.

(2) It is widely used in web search engines like Google to display the most relevant results for a query

(3) Page rank assigns a numerical score (rank) to every page, measuring its importance.

(4) It is designed to prevent website owners from unfairly manipulating their page's position in search results.

(5) Working

(i) The algorithm treats the web as a graph where the nodes represent the web pages and the edges represent the hyperlinks between web pages.

(ii) The idea is that pages with more backlinks are more important.

(iii) Each webpage's rank is calculated using its backlinks (links from other pages)

(iv) Pages with high-quality backlinks get a higher rank compared to pages with backlinks from unknown or private websites.

(6) The formula for page rank algorithm is given as: $R(u) = \sum_{v \in Bu} \frac{R(v)}{Nv}$

where

$R(u)$ is Rank of page $u$

$Bu$ is set of pages linking to page $u$

$R(v)$ is Rank of page $v$ linking to $u$
$N_v$ is number of links on page $v$

(7) This means rank of page $u$ is the sum of the ranks of pages linking to it, divided by the number of links on those pages.

(8) Example

(i) Suppose Page A has backlinks from Page B and Page C.

(ii) Page B has a rank of 6 and links to 3 pages.

(iii) Page C has a rank of 4 and links to 2 pages

Using the formula,

$$R(A) = \frac{6}{3} + \frac{4}{2} = 2 + 2 = 4$$

so, Page A gets a rank of 4

(9) Algorithm:

(i) Convert URLs to unique numbers for faster processing.

(ii) Save hyperlinks in a database using these numbers.

(iii) Sort link structures and remove dangling links.

(iv) Start the iterative calculation of ranks using the formula.

(v) Reinsert the dangling links and recalculate the final rankings

Page No.
Date

Q2) Web usage mining. State any two applications.

⇒ (1) Web usage mining is the process of analyzing user activities on a website to extract useful patterns and information.

(2) It helps the website owners to understand how the user interacts with their site, such as which pages they visit, how long they stay, etc.

(3) Steps in web usage mining:

(i) Data collection:

(a) Data is collected from server logs, cookies or user sessions.

(b) Eg: Recording which pages user visits and at what time.

(ii) Data pre-processing:

(a) The raw data is cleaned and organized to remove irrelevant and duplicate entries.

(b) Eg: Removing entries for images and bots.

(iii) Pattern discovery:

(a) Techniques like clustering, association rules, and sequential pattern mining are used to find patterns in data.

(b) Eg: Identifying the most visited pages or common paths.

(iv) Pattern analysis:

(a) The discovered patterns are analyzed to extract insight and make decisions.

(b) eg: understanding which product pages lead to more purchases.

(4) Application of web usage mining:

(i) personalized recommendations.

(a) Based on user behavior, website can suggest

products, movies or articles.

(b) Eg: Netflix recommending shows based on your watch history.

(ii) Website optimization:

(a) Helps to improve the design and structure of website by analyzing user navigation patterns.

(b) Eg: Moving frequently visited links to the homepage.

(iii) Targeted advertising:

(a) Analyze user interests and display ads accordingly.

(b) Eg: Google ads showing ads for products you have searched for.

Q3) Web content mining

=> (1) Web content mining is the process of extracting useful and relevant information from the content of web pages.
(2) This content can include texts, images, videos, audio and structured data like tables or meta-data.
(3) It helps users and organizations analyze large amounts of online data for various purposes such as decision-making or web-analysis.
(4) Steps in web content mining:
(i) Data collection:
(a) Collect data from a website using web crawlers or scraping tools.
(b) Eg: Crawling news website to collect articles about a specific topic.
(ii) Data Pre-processing:
(a) Clean and organize the raw data to remove unnecessary elements like advertisements or duplicate content.
(b) Eg: Removing HTML tags or irrelevant sections of a webpage.
(iii) Pattern discovery:
(a) Use techniques like natural language processing (NLP), clustering or classification to find patterns or insights.
(b) Eg: Identifying trending topics from social media posts.
(iv) Pattern analysis
(a) Analyze the extracted pattern to derive meaningful conclusions or predictions.

(b) Eg: Using keyword analysis to optimize search engine rankings.

(5) Types of web content mining:

(i) Text data mining

(a) Extracting information from articles, blogs, and other text-based content.

(b) Eg: Analyzing product reviews to understand customer requirements.

(ii) Multimedia data mining

(a) Extracting and analyzing images, videos, or audio files.

(b) Eg: Detecting faces in photos or recognizing speech in audio files.

(iii) Structured data mining

(a) Extracting data from structured content like web forms, tables or metadata.

(b) Eg: Scraping prices and product specifications from e-commerce website.

(6) Applications of web content mining:

(i) Search engine optimization (SEO)

(a) Analyzing content keywords to improve website visibility in search engine results.

(ii) Sentiment analysis

(a) Analyzing user opinions for reviews, blogs, or social media

(iii) Research and trend analysis

(a) Mining articles, research papers and blogs to identify new trends or developments in a field.

Q4) Web structure mining.

=> (1) Web structure mining is the process of analyzing the structure of hyperlinks within websites to understand the relationships between web pages.

(2) It treats the web page as a graph, where:

(i) nodes represent web pages.

(ii) edges represent hyperlinks connecting those pages.

(3) The goal is to extract patterns and insights about how pages are connected.

(4) Steps in web structure mining:

(i) Data collection:

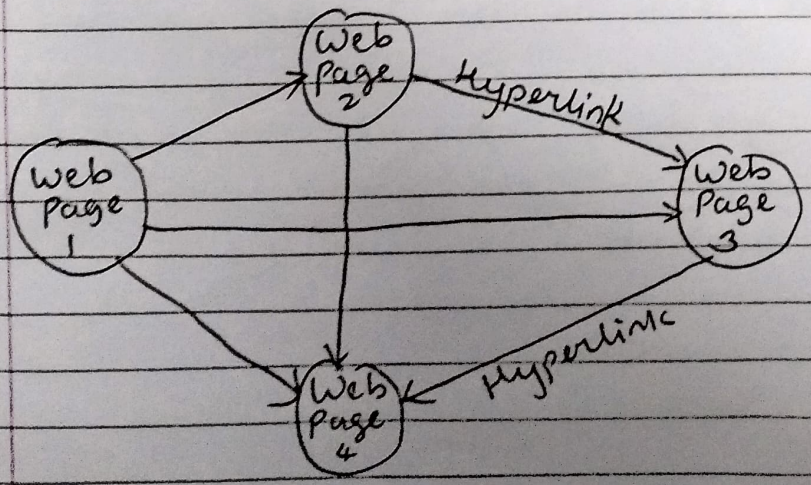
(a) Collect hyperlink data using web crawlers.

(b) eg: Crawl an academic website to map how research papers are linked.

(ii) Graph construction:

(a) Represent web pages and links as a graph.

(b) eg: Create a graph where nodes are pages and edges are hyperlinks




(iii) Analysis:

(a) Apply algorithms like PageRank or HITS to

rank or categorize pages.

(5) Application of web structure mining:

(i) Spam detection:

(a) Detects spam websites by analyzing link patterns.

(b) eg: Identifying 'link farms' that artificially boost rankings.

(ii) Recommendation system:

(a) Suggests related web pages based on link structure

(b) Eg: Suggesting related news articles on a news website based on hyperlink connections

(iii) Search engine optimization (SEO)

(a) Helps search engines to rank web pages based on their importance in the web graph.

(b) Eg: Google uses page rank algorithm to prioritize relevant pages in search results.

Page No.
Date

(Q5) Web mining and its types

=> (1) Web mining is the process of discovering useful information and patterns from the vast amount of data available from the web.
(2) It involves applying data mining techniques to extract knowledge from web content, structure and usage data.
(3) Web mining helps in understanding user behaviour, improving website performance and delivering personalized experience.
(4) Types of web mining:
(i) Web usage mining
Refer Q2)
(ii) Web content mining
Refer Q3)
(iii) Web structure mining
Refer Q4)

Page No.
Date

q6) K-medoids clustering.

⇒ (1) K-medoid clustering is a partitioning technique used to group data into K-clusters.
(2) It is similar to K-means but is more robust because it minimizes the effect of outliers.
(3) Unlike K-means, which uses the mean of data points as the center, K-medoid chooses actual data points as cluster centres.
(4) Working
(i) Initialization
(a) Select k random data points from the dataset as initial medoids (cluster centres)
(ii) Assign data points to clusters
(a) Assign each data point to the cluster with the closest medoid based eucleidan distance.
(iii) Update medoids
(a) For each cluster, calculate the total distance between all points in the cluster and the medoid
(b) Replace the medoid with another point in the cluster if it reduces the total distance.
(iv) Repeat
(a) Repeat step 2 and 3 until the medoids stop changing or a maximum number of iterations is reached.
(5) Applications
(i) Market segmentation
(a) Grouping customers based on purchasing behavior.
(b) eg: A retail store clustering customer for personalized offers.

(ii) Image segmentation

(a) Segmenting regions in an image for analysis

(b) Eg: Dividing a satellite image into land, water and vegetation regions.

(iii) Healthcare

(a) Grouping patients based on symptoms for better treatment plans.

(iv) Fraud detection

(a) Identifying unusual patters in financial transactions.