DWM Assignment-6

Q.1) Apply your knowledge of a web crawler in content mining. Describe the basic components of a web crawler & explain how it decides which pages to visit first what are 2 common techniques it uses to avoid visiting the same page multiple times)

→ A web crawler, also known as a spider or bot, is an automated program that systematically browses the web to collect and index information from web pages for various purposes, such as search engines & data analysis

Basic components:

- URL Frontier: A queue that manages URLs to be visited, prioritizing based on criteria like relevance or freshness.
- Downloader: fetches web page content from the URLs in the URL frontier.
- Parser: Analyzes pages to extract text links and metadata while identifying new URLs to add to the frontier.
- Storage: Saves crawled data in a structured format, usually in a database.

Page Prioritization: Crawlers use strategies like Breadth-First Search (BFS) or Priority Queues (based on relevance or freshness) to decide which pages to visit first.

Avoiding Page Revisit:-

- URL Deduplication: Keeps a list of visited URLs to avoid revisiting.
- Cononicalization: Ensures different versions of the same page are identified and only the cononical version is crawled.

Q.2) Explain what web usage mining is and how it can be beneficial for a website. Provide an example of how web usage mining can help improve user experience on an e-commerce site. What type of data would you analyze to gather insights from user behaviour?

→ Web usage mining analyzes user behaviour from web logs to identify patterns that improves user experience and site performance.

* Benefits:

- Personalization: Adapts content and recommendations based on user preferences.
- Performance Improvement: Helps resolve issues like slow-loading pages that affect engagement.

E-Commerce Example: Analyzing user navigation and click patterns can identify bottlenecks in the purchase process, such as users abanding their carts at the shipping stage. This insight can lead to optimizing the checkout experience.

*Key Data Types:- ① Access Logs:- show visited pages.
② Clickstream Data:- Detailed records of user clicks
Analyzing these helps e-commerce sites enhances user experience and boost conversion rates.