1. **Short note on Page Ranking Algorithm.**

**Ans:-**

1. The PageRank algorithm helps search engines like Google decide which websites should appear first.
2. Websites that have more links from other good websites are given a higher rank.
3. If a website is linked by many trusted sites, it is considered important.
4. A website with no links or bad links gets a lower rank.
5. This algorithm helps users find useful information quickly.
6. It also checks if a website is updated regularly.
7. This system works automatically using mathematical formulas.
8. PageRank makes sure to show the most relevant websites to users.
9. Without this algorithm**,** search engines wouldn’t work well.
10. The **Damping Factor (d)** in the PageRank algorithm helps prevent a website from gaining too much importance just because it has many links. It ensures that a user doesn’t stay on one website forever but keeps exploring different pages.

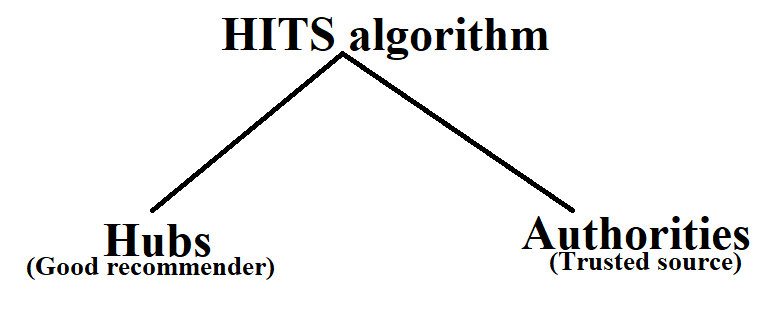
**Example:** If the **damping factor is 0.85**, it means that a user will follow **85% of the links** on a webpage instead of **100%**. The remaining **15%** represents the chance that the user will randomly visit an **irrelevant or unlinked page**.

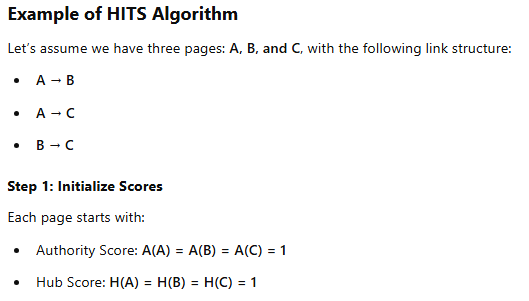
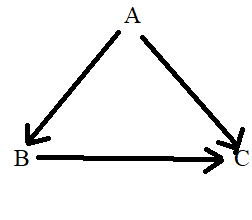
1. Weighted PageRank gives more importance to links from trusted and highly ranked websites instead of treating all links equally. A link from a famous site carries more value than a random link.
2. By using Damping Factor (0.85) and Weighted PageRank, search engines show the best and most useful websites at the top of search results.
3. **Explain HITS algorithm in detail?**

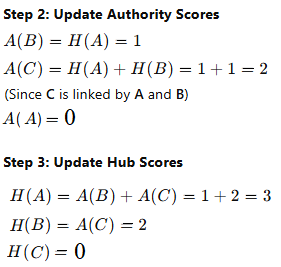
**Ans:-**

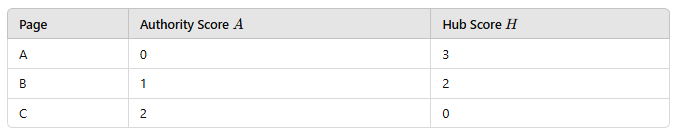
The HITS Algorithm, also known as Hubs and Authorities, was developed by Jon Kleinberg in 1999. It is used to rank web pages based on link analysis, just like Google's PageRank. However, unlike PageRank, HITS focuses on identifying two types of pages:

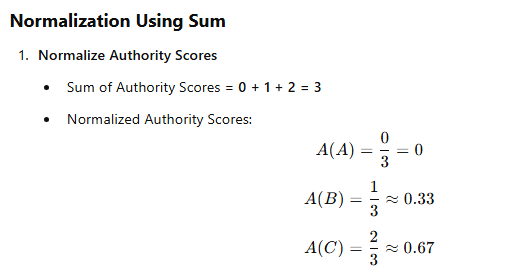
1. Hubs – Pages that link to many other important pages.
2. Authorities – Pages that are linked by many good hubs.

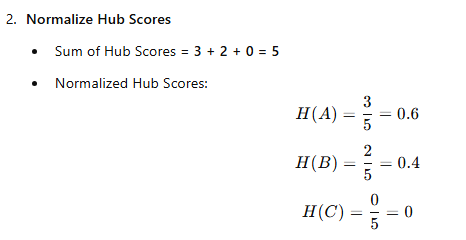


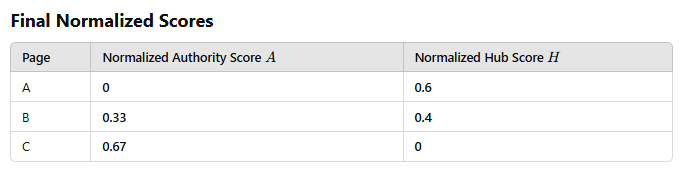


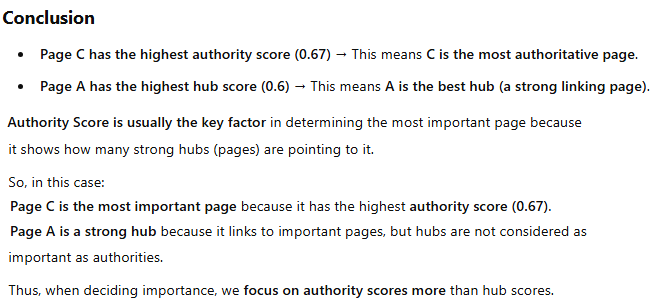


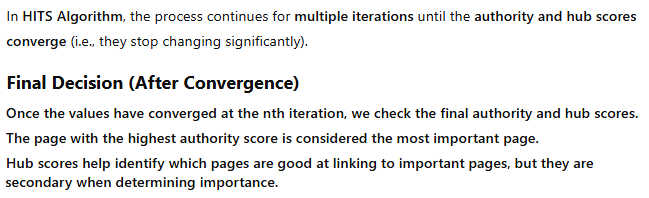












1. **Short note on web crawlers?**

**Ans:-**

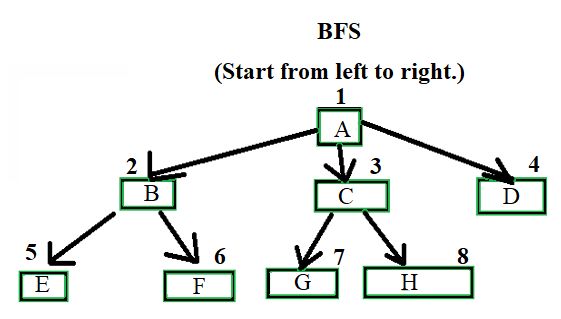
A Web Crawler (also called a spider or bot) is a program used to visit, read, and automatically extract information from web pages. It follows links from one page to another, just like a person browsing websites.

It uses several techniques for extracting links from web pages:

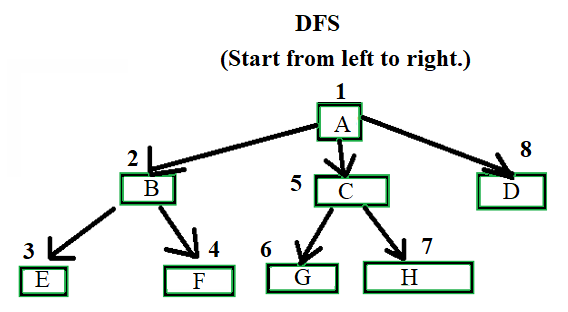
**Types of Crawler Algorithms:**

**1. Breadth-First Search (BFS) Crawler-**

* Visits all links on one page first, then moves to the next level.
* **Example:** Like checking all the rooms on one floor before going to the next floor.
* **Used for:** Finding wide information (e.g., news, social media).

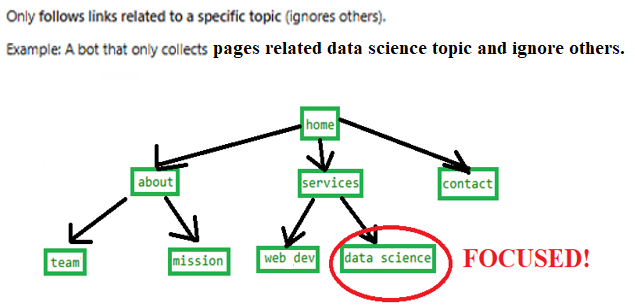


**2. Depth-First Search (DFS) Crawler-**

* Follows one link deeply before coming back.
* **Example:** Like following one tunnel deep underground before checking another.
* **Used for:** Exploring specific topics deeply.

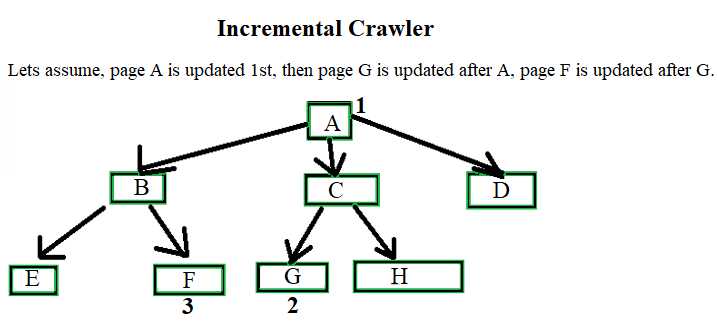
**3. Focused Crawler-**

* Only follows links related to a specific topic (ignores others).
* **Example:** A bot that only collects medical research papers.
* **Used for:** Research topic-specific websites.



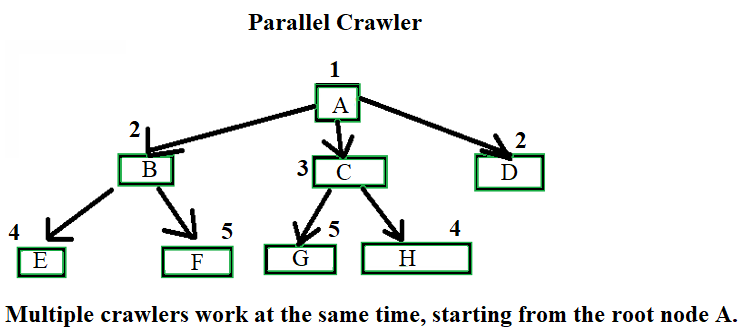
**4. Incremental Crawler**

* Instead of crawling everything again, it only updates new or changed pages.
* **Example:** Like a detective checking only new evidence in an old case.
* **Used for:** News websites, stock markets, frequently updated content.

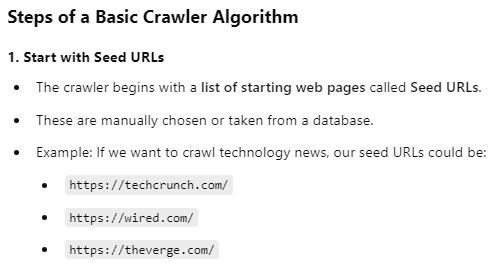
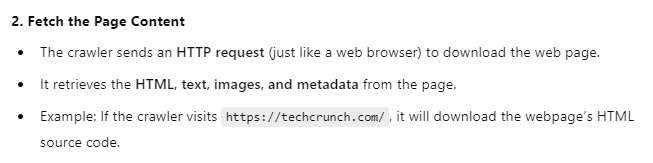


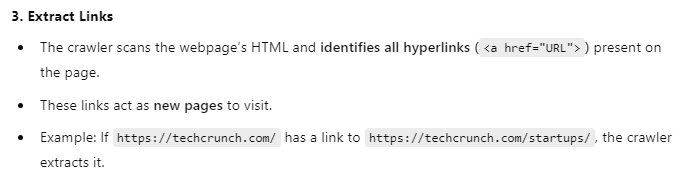
**5. Parallel Crawler**

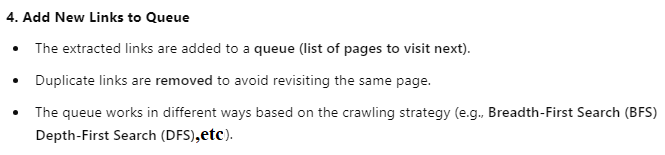
* Multiple crawlers work together to speed up searching.
* **Example:** A team of robots splitting up the work to cover more ground.
* **Used for:** Large-scale search engines like Google and Bing.

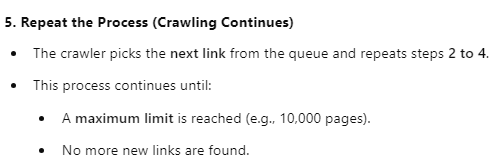


1. **Explain brief basic crawler algorithm?**

 **Ans:-**







**Example of the Basic Crawler Algorithm in Action-**

Let's say we start with **one seed URL**:  
**Seed URL:** https://example.com

The crawler follows these steps:

1️. Fetches https://example.com → Extracts links:

* https://example.com/about
* https://example.com/news
* https://example.com/contact

2. Visits https://example.com/about → Extracts more links:

* https://example.com/team
* https://example.com/careers

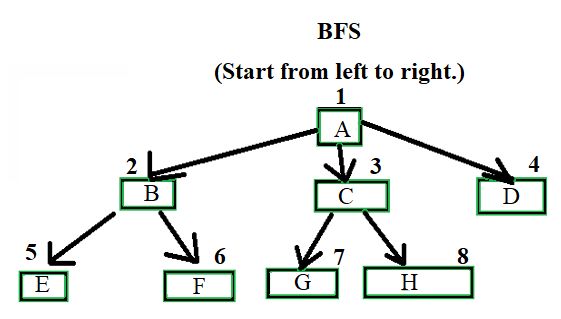
3️. Visits https://example.com/news → Extracts links:

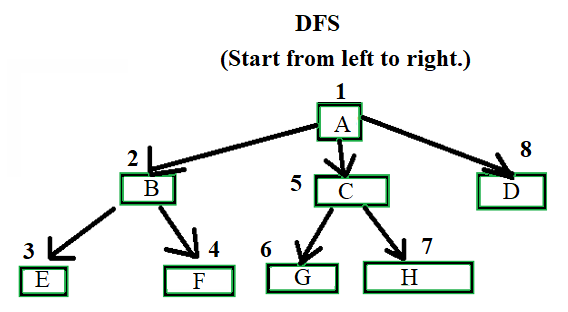
* https://example.com/news/tech
* https://example.com/news/sports

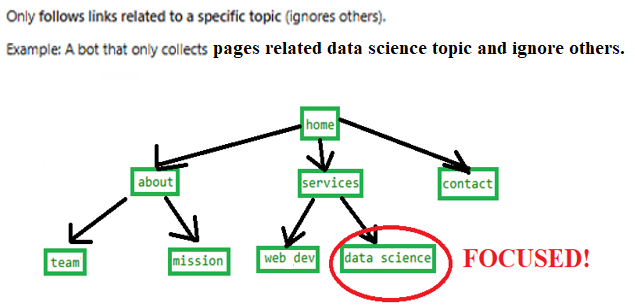
The crawler **keeps visiting new links** and adding them to the queue **until it reaches the stopping condition**.

**Types of Crawling Approaches:**

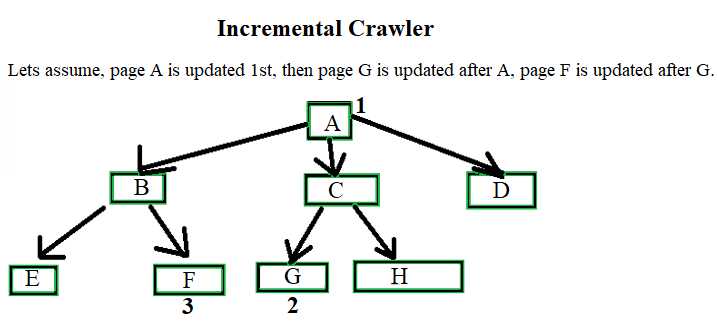
1. **Breadth-First Search (BFS) Crawler** → Visits all links at the current level before moving deeper.



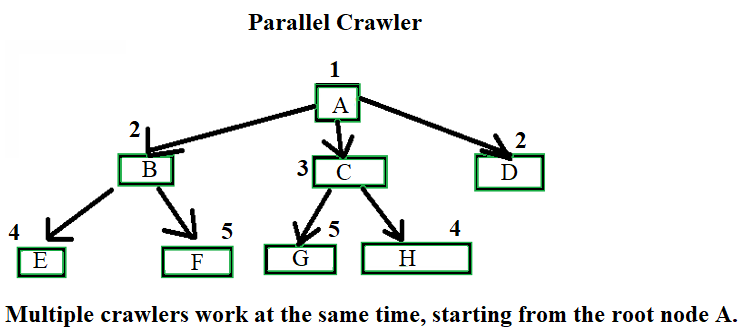
1. **Depth-First Search (DFS) Crawler** → Follows one link deeply before returning to other links.
2. **Focused Crawler** → Only follows links related to a specific topic.



1. **Incremental Crawler** → Updates only new or modified pages instead of crawling everything.



1. **Parallel Crawler** → Uses multiple crawlers to speed up the process.



1. **What is Cocitation & Bibliographic Coupling?**

**Ans:-**

Cocitation:

* When two documents (A & B) are cited together in another document (C).
* Example: If a research paper X cites both Paper A and Paper B, then A and B are cocited.
* Use: Helps identify related documents based on how often they are mentioned together.

Bibliographic Coupling:

* When two documents (A & B) cite the same reference (C).
* Example: If both Paper A and Paper B refer to Paper X, then A and B are bibliographically coupled.
* Use: Helps group documents with similar topics.

Here’s a real-time scenario to understand cocitation and bibliographic coupling better:

Scenario 1:

Imagine you are reading a news article about Artificial Intelligence (AI). The article talks about two famous research papers:

1. Paper A: "AI in Healthcare"
2. Paper B: "AI for Self-Driving Cars"

Since both Paper A and Paper B are mentioned together in the same article, they are cocited.

Scenario 2:

Now, imagine two different students writing research papers:

1. Student X writes about AI in Healthcare.
2. Student Y writes about AI in Self-Driving Cars.

Both students mention the same reference paper:

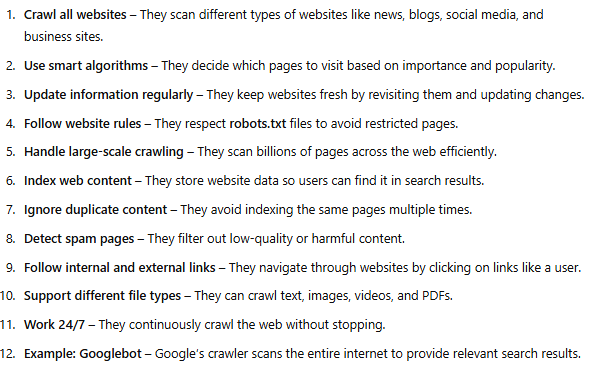
3. Paper C: "Basics of Machine Learning"

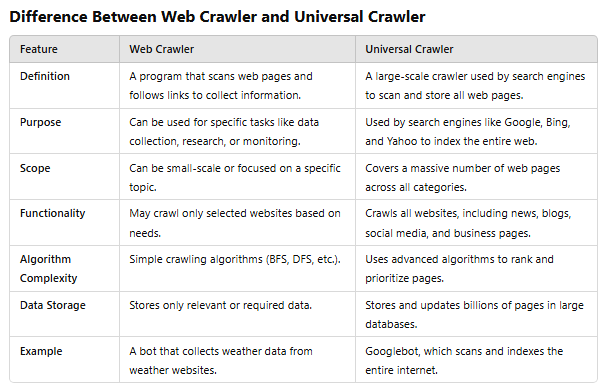
Since both Student X and Student Y refer to Paper C, their papers are bibliographically coupled.

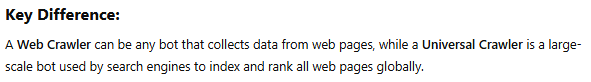
1. **What Are the Implication Issues in Web Crawling?**

**Ans:-**

1. Bandwidth Consumption – Crawlers send frequent requests, which can slow down websites.
2. Duplicate Data – Crawlers may index the same page multiple times, leading to redundancy.
3. Robot Exclusion Rules – Some websites use robots.txt to restrict or block crawling.
4. Legal & Ethical Concerns – Unauthorized crawling can violate privacy laws and website terms.
5. Storage & Processing – Crawling millions of pages requires large storage and computing power.
6. Crawl Delay Issues – Some websites limit how fast a bot can crawl, slowing down data collection.
7. Dynamic Content Crawling – Pages using JavaScript or AJAX may not load properly for crawlers.
8. Spam & Fake Content – Crawlers may collect misleading or irrelevant information from spam websites.
9. **What is a Universal Crawler?**

 **Ans:-**





1. **What are the crawlers' ethics & conflicts?**

**Ans:-**

Ethics of Web Crawlers:-

1. Respecting Robots.txt – Crawlers should follow website rules about which pages can be visited.
2. Not Overloading Servers – Crawlers must avoid sending too many requests, which can slow down or crash websites.
3. Protecting User Privacy – Crawlers should not collect sensitive data like passwords, personal emails, or bank details.
4. Following Legal Guidelines – Crawlers must follow internet privacy laws and avoid unauthorized access to data.
5. Fair Use of Data – Crawled data should not be misused for spam, fake news, or harmful activities.
6. Avoiding Copyright Violations – Crawlers should not copy or distribute content without permission from the website owner.

Conflicts in Web Crawling:-

1. Website Owners vs. Search Engines – Some websites do not want to be crawled, but search engines need to index them for users.
2. Public vs. Private Data – Crawlers aim to collect public data, but sometimes they may accidentally access restricted or private content.
3. Competition Among Companies – Some companies use crawlers to collect data from competitors, leading to legal and ethical disputes.
4. Fake Crawlers and Cybersecurity – Some harmful bots pretend to be ethical crawlers but actually steal data or launch attacks.
5. Data Ownership Issues – Websites may claim ownership of their data, but crawlers collect and use it in search results, leading to conflicts.
6. Use of Crawled Data – Governments, businesses, and researchers may use web crawled data differently, leading to ethical concerns over its purpose and impact.