Inverted index

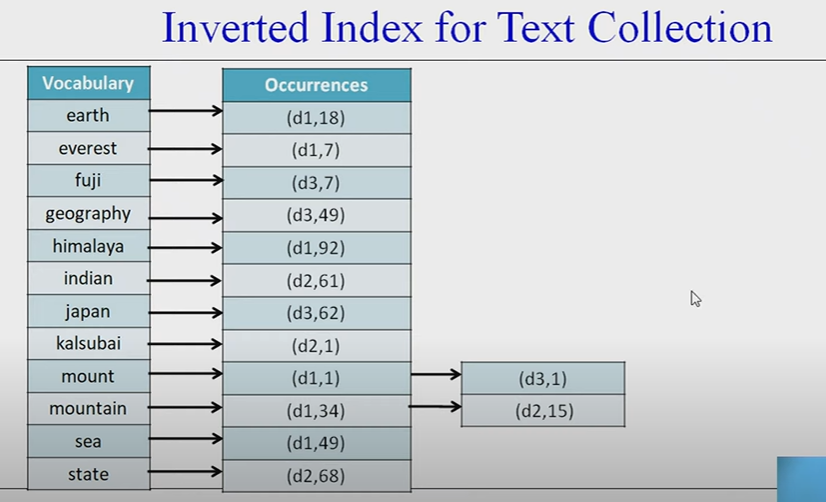
Components

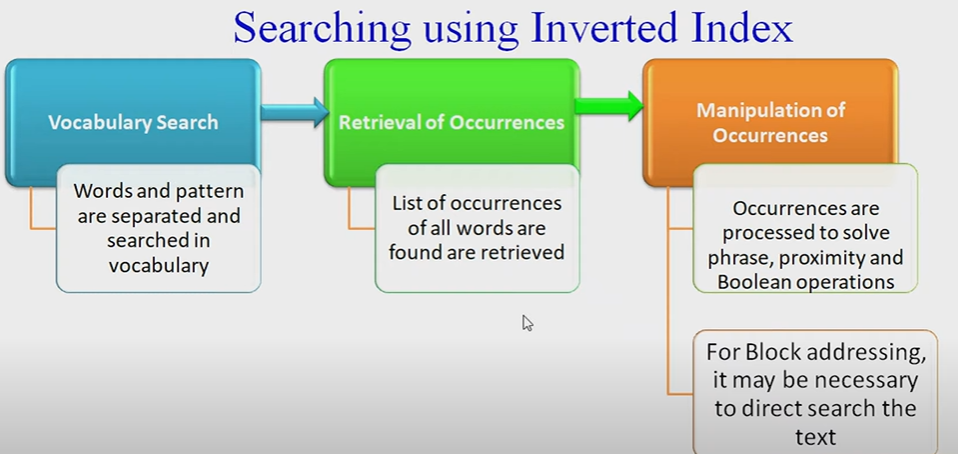
* Vocabulary
* Occurences (word or character positions)

Process

1. Identify the index words
2. Sort them alphabetically
3. Associate occurrences

Block addressing - reduce space requirements

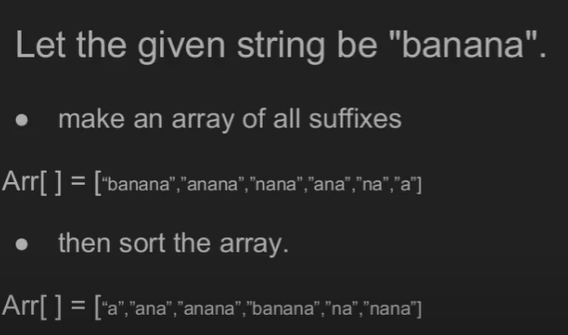


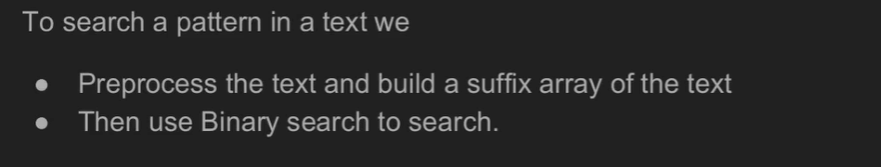


Type of queries

1. Single word queries
2. Prefix queries
3. Range queries
4. Proximity(context) queries

Suffix arrays

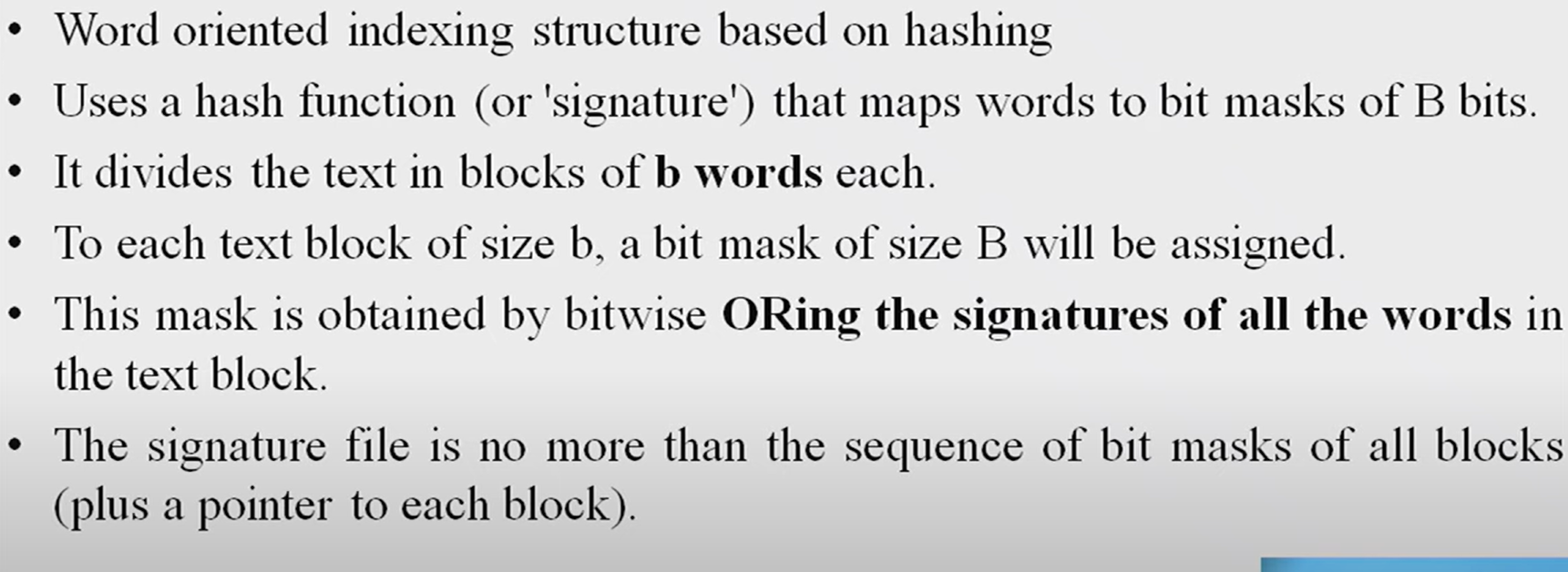


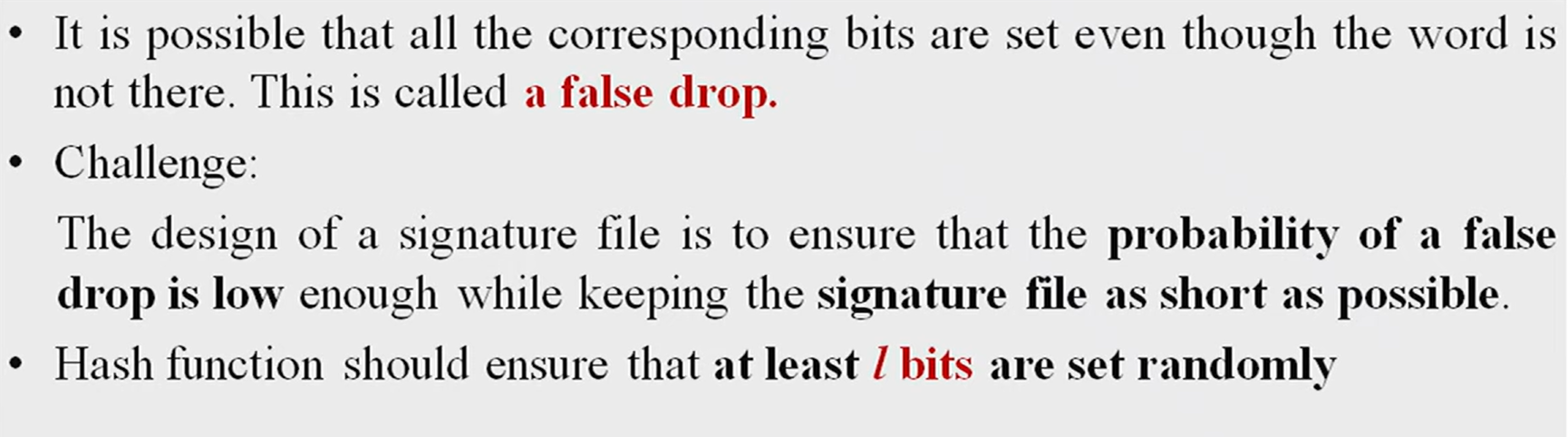


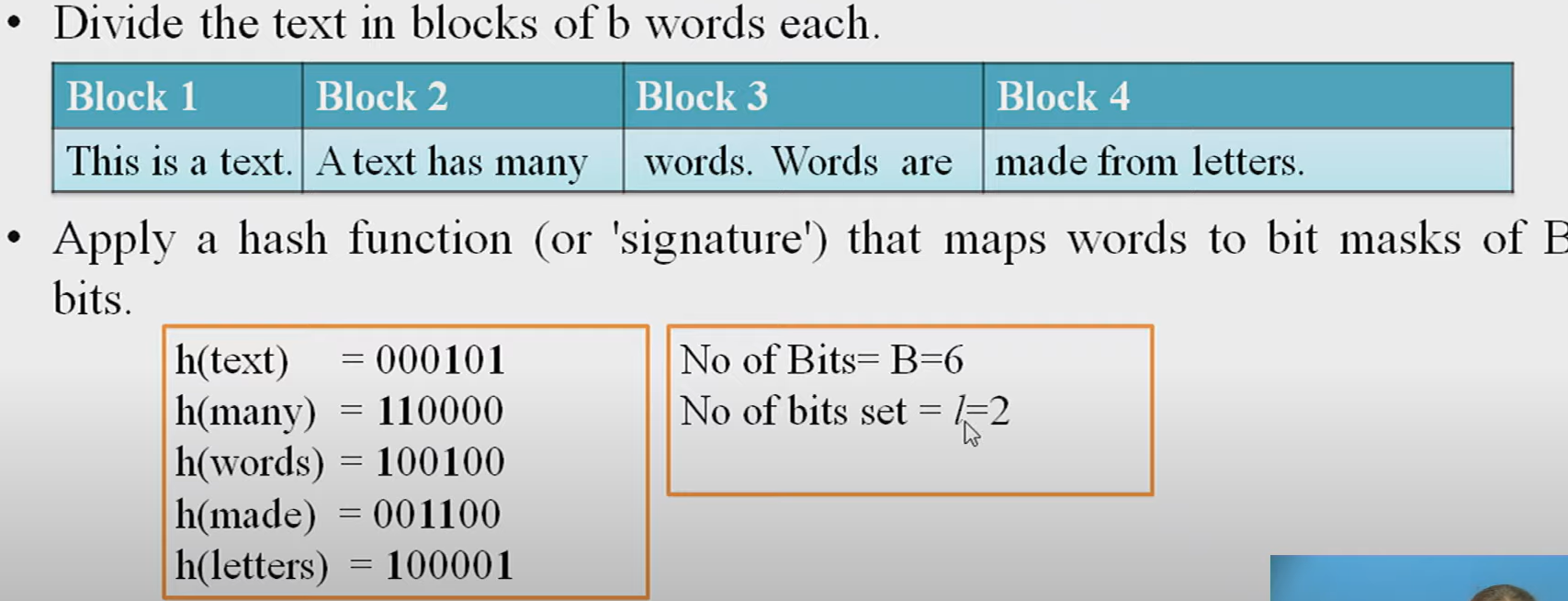
Time complexity = O(MlogN)

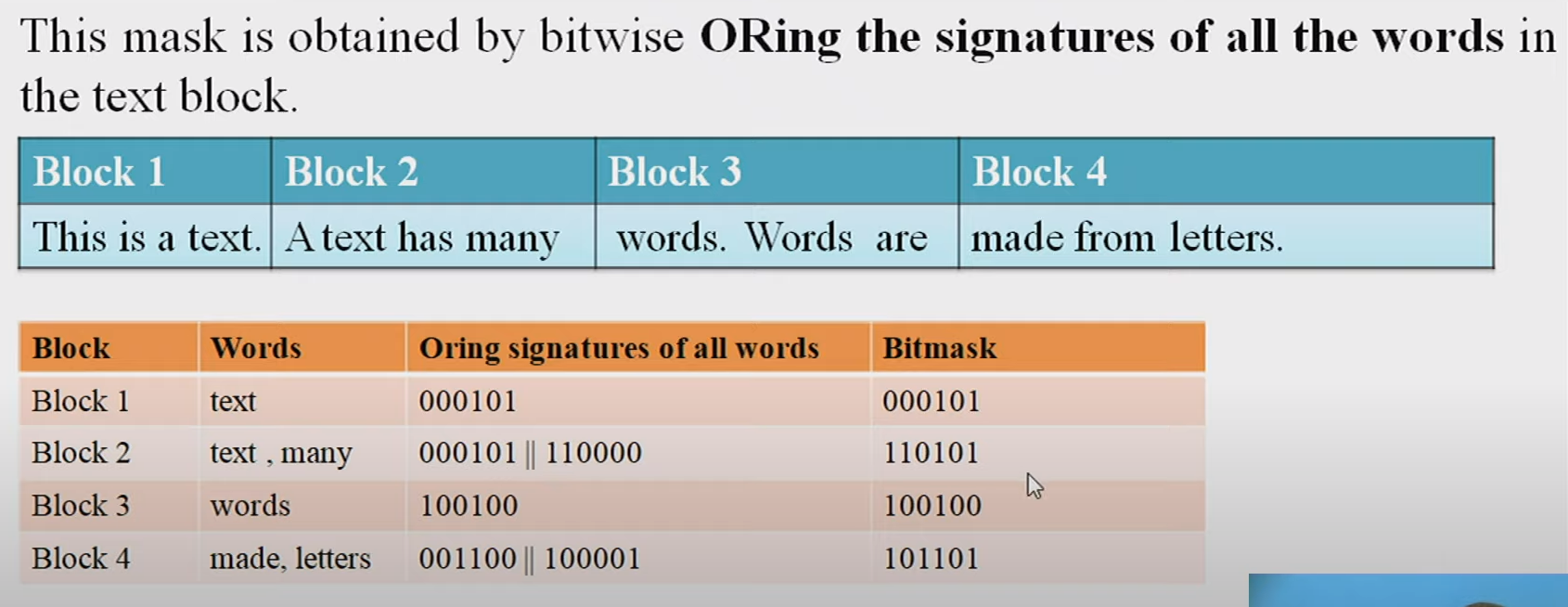
Overhead of M

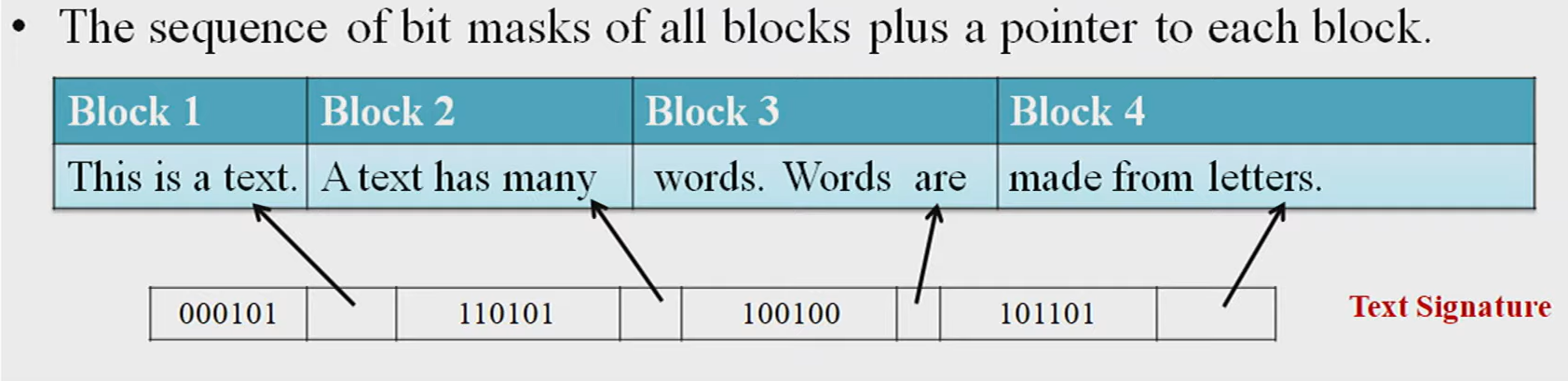
Signature files

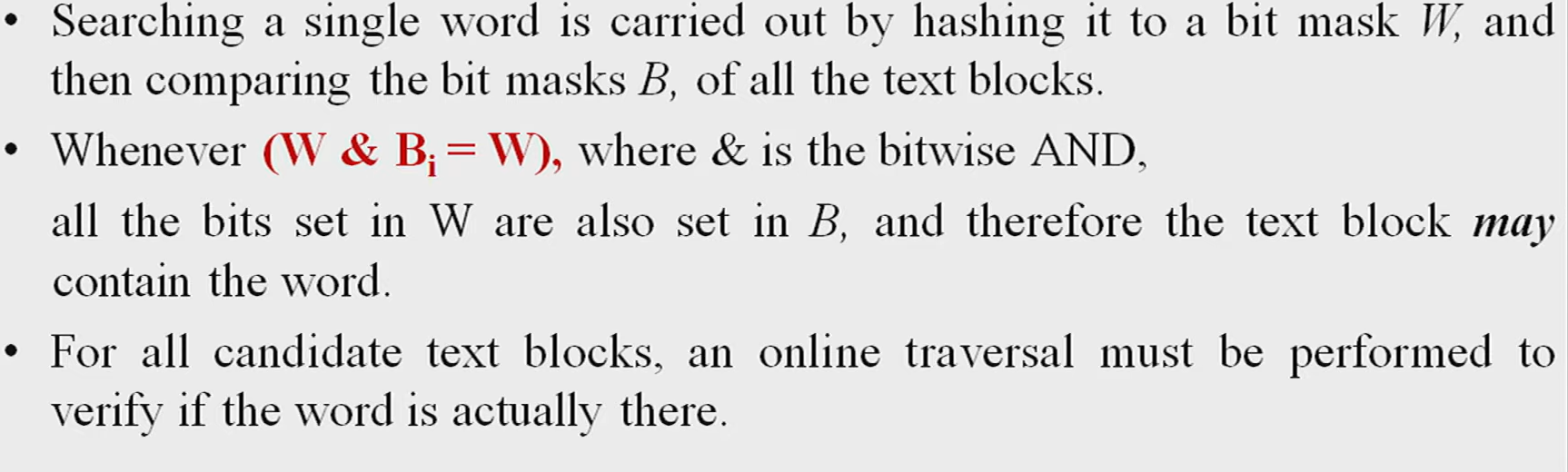












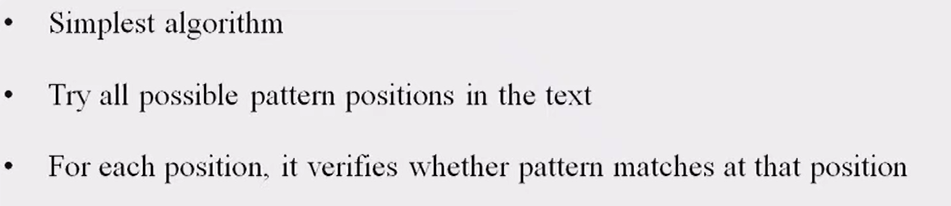
SEQUENTIAL SEARCHING

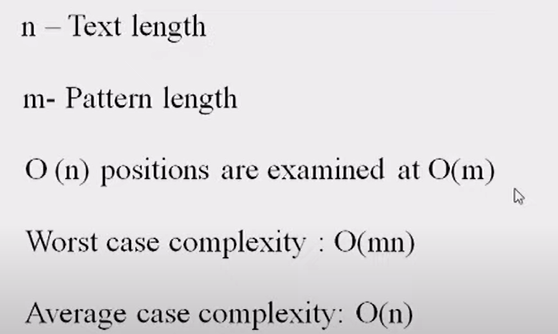
Appropriate when text is

* Not processed
* Small
* Volatile

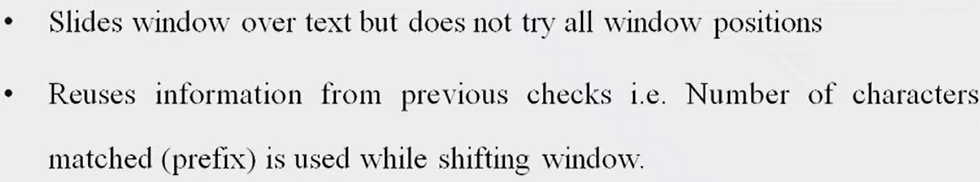
Index space overhead is not affordable

1. Brute Force Approach





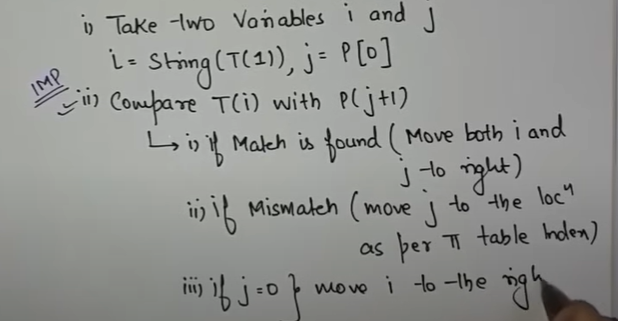
1. Knuth Moriss Pratt



Works on proper prefix and proper suffix

O(m+n)

1. Find longest proper prefix
2. Follow these steps



Challenges in web searching are a result of the vast and dynamic nature of the internet. Here's an explanation of the points you've mentioned, as well as some additional challenges:

1. Distributed Data:

- Explanation: The web is a distributed network of servers and websites that are located across the globe. Searching for information often requires querying multiple sources, which can be geographically dispersed and have different data structures.

- Additional Challenges: Dealing with data stored on various servers and in different formats can be challenging. This also raises issues related to data consistency, latency, and access control.

2. High Percentage of Volatile Data:

- Explanation: Web content is constantly changing, with information being updated, added, or removed. This volatility makes it difficult to maintain up-to-date search results.

- Additional Challenges: Keeping search indexes current and ensuring that users access the most recent and accurate information is a constant challenge. Search engines must continuously crawl and index web pages to address this issue.

3. Large Volume:

- Explanation: The internet contains a massive volume of information, including text, images, videos, and more. This sheer volume can lead to scalability and performance issues for search engines.

- Additional Challenges: Processing and indexing such a vast amount of data while providing fast and relevant search results is a significant technical challenge. It requires efficient algorithms and infrastructure.

4. Unstructured and Redundant Data:

- Explanation: The web contains a mix of structured and unstructured data, and there is often a lot of redundancy and duplicate content. This can make it challenging to present users with the most relevant information.

- Additional Challenges: Search engines need to identify and filter out redundant and low-quality content to improve search result quality. They also need to understand unstructured data, such as natural language text, to provide relevant results.

5. Quality of Data:

- Explanation: Not all information on the web is accurate or reliable. Search engines must assess the quality, credibility, and trustworthiness of data sources to provide users with reliable information.

- Additional Challenges: Determining the quality and trustworthiness of web content can be difficult, and misinformation and fake news are prevalent issues. Search engines often employ various algorithms and user feedback mechanisms to combat this problem.

6. Heterogeneous Data:

- Explanation: Data on the web comes in various formats, including text, images, videos, and structured databases. This heterogeneity poses challenges in indexing and presenting diverse content types.

- Additional Challenges: Search engines need to interpret and categorize different data types to offer a diverse set of search results. They also need to handle multimedia content and provide relevant information retrieval.

Additional challenges in web searching may include:

7. Privacy Concerns: Balancing the need for personalized search results with user privacy is a challenge. Search engines must protect user data while delivering relevant content.

8. Multilingual and Multicultural Content: The web is a global platform with content in multiple languages and cultural contexts, making it challenging to provide effective search for diverse audiences.

9. User Intent Understanding: Search engines must understand the intent behind user queries and provide results that match user expectations, which can be complex due to variations in language and context.

10. Emerging Technologies: Staying up-to-date with new web technologies, such as JavaScript-heavy websites and voice search, poses challenges for search engines in adapting their crawling and indexing techniques.

11. Legal and Ethical Considerations: Search engines must navigate legal and ethical issues, including copyright, censorship, and removal requests, while ensuring access to information within legal boundaries.