CGB1201 – JAVA PROGRAMMING
PROJECT REVIEW-2

Department of Information Technology
Academic Year: 2024 – 2025 (Odd Semester)

Register Number	**: 927623BIT095**
Name	**: RASMITHA.R**
Year	**: II**
Semester	**: III**
Section	**: B**
Date	**:**

Title of the Project

SEARCH ENGINE

Abstract

- Search engines play important roles in the success of the Web, search engines helps any Internet user to rapidly find relevant information. The search engines should be more useful and efficient for searching the relevant Web information. Search Engine Optimization is a process of increasing the chances of a webpage to appear in the first page of the search result. Since, whenever the consumer searches for information, they provide a particular phrase or a keyword instead of the complete web address, then the search engine use that keyword to find the relevant web pages.

Abstract with CO/PO Mapping

Abstract	CO	POs	PSO
Search engines play important roles in the success of the Web, search engines helps any Internet user to rapidly find relevant information. The search engines should be more useful and efficient for searching the relevant Web information. Search Engine Optimization is a process of increasing the chances of a webpage to appear in the first page of the search result. Since, whenever the consumer searches for information, they provide a particular phrase or a keyword instead of the complete web address, then the search engine use that keyword to find the relevant web pages	CO1 CO4	PO-1 PO-2 PO-5 PO-9 PO-10 PO-12	PSO-1 PSO-2

Introduction

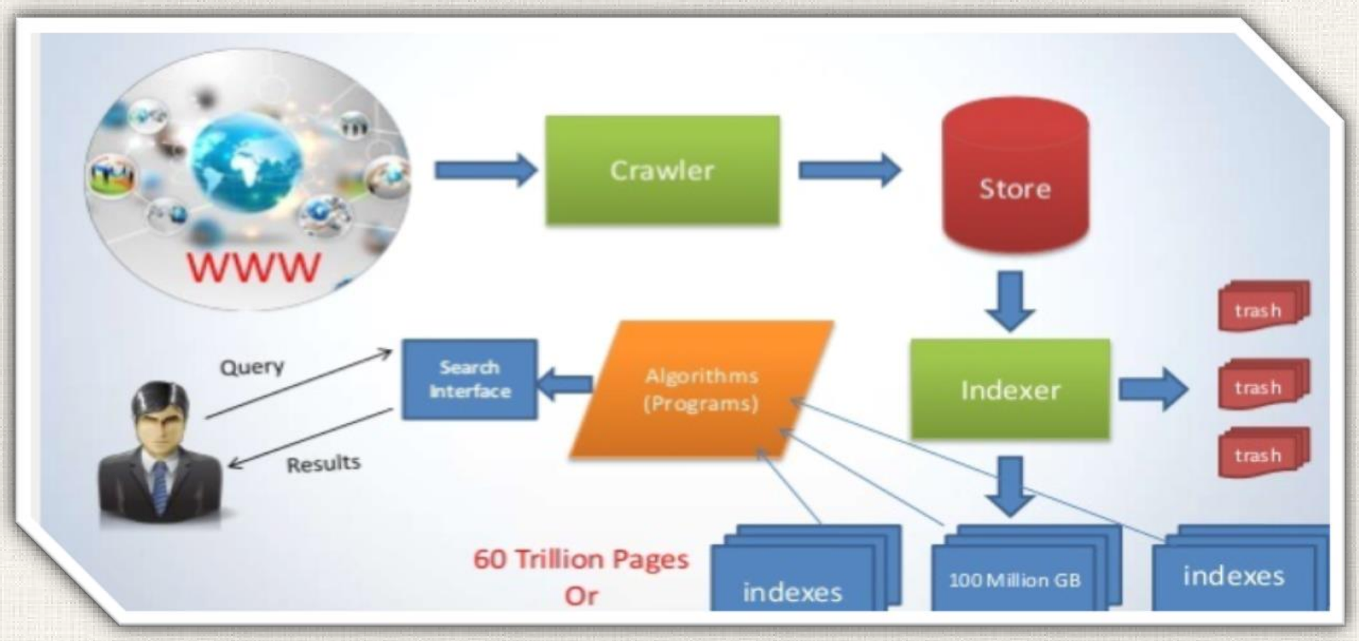
- A search engine is a software application or platform that retrieves and ranks relevant information based on user queries from a large database or the internet.
- This software system that provides hyperlinks to web pages and other relevant information and provides response to the user. The search engine assumes that the title contains all of the important words that define the topic of the piece.
- They use algorithms to index and rank web pages based on relevance to a user's query, providing a list of results for users to explore.
- They may also search specifically for images, videos, phrases, questions, or news articles or for names of websites.

Java Programming - Concepts Used

key concepts such as**:**

- **Multithreading**: For fetching multiple web pages concurrently. Networking Use of classes like Http URL Connection for making HTTP requests.
- **Indexer Modules** : search engine is responsible for processing, organizing, and storing information extracted by the crawler.
- **String Manipulation**: For processing user queries using classes like String, StringBuilder , that helps in the search.
- **HashMap and Tree** : This concepts used in faster lookups of the searched content.

Proposed Architecture



Proposed Architecture - Description

- **Crawling :** Accessing a website and obtaining data through a software program. This is used by almost all search engines like Google, Yahoo. It indexes content from all over the internet and searches in-depth and width for hyperlinks to extract.
- **Storing :** In a search engine, documents or web pages are stored in a structured format like a database or flat files. Java can use databases like MySQL or NoSQL to store documents.
- **Indexer :** Processes documents, extracts relevant terms (tokens), and creates an index for fast retrieval. This allows quick lookup during the search, enabling faster search results for queries.
- **Interface**: The search interface interacts with the back-end index to retrieve documents that match the query and then displays the results in an organized manner.

Proposed Architecture - Description (Cont..)

- **Pattern searching :** It is the feature where a user-entered input is searched against text files. The number of occurrences of the input keyboard will be counted.

  The **Boyer-Moore** algorithm is used for pattern searching feature implementation.
- **Spell Check :** It is a widely used feature in search engines. The user's input keyword is considered to be spelled correct if it is found in the long list of valid words called **dictionary**. If the keyword is not found in the dictionary, **edit distance** algorithm will suggest a similar word.
- **Data structures** like trees, hash maps, and graphs are used to organize and manage the vast amounts of data indexed by search engines.

List of Modules

1. Crawler (Web Scraping) Module
2. Indexer Module
3. Query Processor Module
4. Ranking Module
5. Storage Module
6. User Interface (UI) Module

Module Description

1. Crawler (Web Scraping) Module: This module is responsible for automatically collecting data from websites by visiting and extracting the required content such as text, images, and links.

2. Indexer Module: The Indexer organizes and stores the collected data in an inverted index format, mapping keywords to documents. This helps in quickly locating and retrieving relevant information based on queries.

3. Query Processor Module: This module processes user queries to find relevant search results by leveraging the inverted index. It parses the query, matches keywords, and ranks documents based on relevance and other criteria.

Module Description (Cont..)

4. Ranking Module: The Ranking Module evaluates and assigns scores to search results based on relevance, quality, and additional factors like link popularity. It helps in ranking the most relevant results first.

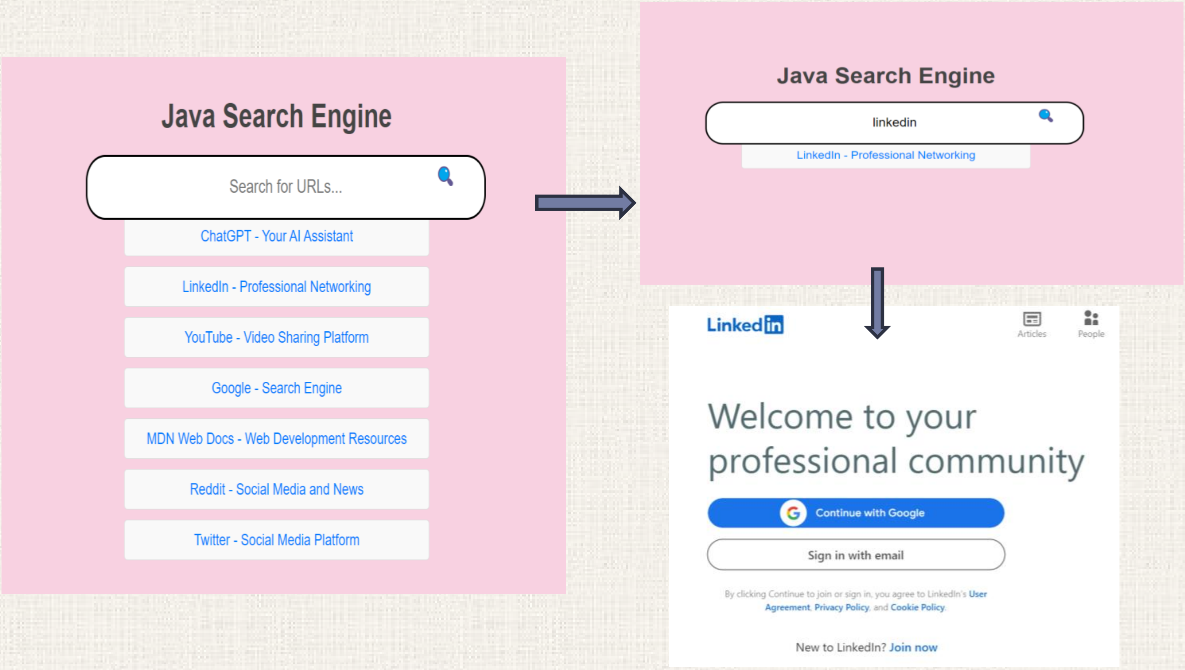
5. Storage Module: This module handles the storage of web data, including the inverted index, documents, and any metadata. It ensures data integrity, reliability, and efficient retrieval.

6. User Interface (UI) Module: The UI module provides an interface for users to interact with the search engine, allowing them to enter queries, view search results, and navigate through content in a user-friendly manner.

Results and Discussion

- Each module plays a distinct role in ensuring that the engine can crawl, index, process, rank, and display results efficiently.
- The **Crawler** and **Indexer** modules ensure that content is continuously collected and structured in a way that allows for efficient searching.
- The **Query Processor** and **Ranking** modules work together to interpret and return the most relevant search results.
- **User Interface (UI) Module**: The UI module provides an interface for users to interact with the search engine, allowing them to enter queries, view search results, and navigate through content in a user-friendly manner.

Results and Discussion (Cont..)


Java Search Engine
Search for URLs...
ChatGPT - Your AI Assistant
LinkedIn - Professional Networking
YouTube - Video Sharing Platform
Google - Search Engine
MDN Web Docs - Web Development Resources
Reddit - Social Media and News
Twitter - Social Media Platform
Java Search Engine
linkedin
LinkedIn - Professional Networking
Linked in
Articles
People
Continue with Google
Sign in with email

Welcome to your professional community

By clicking Continue to join or sign in, you agree to LinkedIn's User Agreement, Privacy Policy, and Cookie Policy.

New to LinkedIn? Join now

Conclusion

- Overall, while the search engine performs well at its core functions, there are opportunities for optimization and expansion, such as improving the crawler's speed, refining the ranking algorithm, or enhancing the user interface for better interaction. Additionally, incorporating machine learning models for personalized search results and query understanding could further enhance the system's accuracy and relevance. Continuous updates and maintenance of the crawler and indexer are necessary to keep up with the dynamic nature of the web and ensure the freshness of search results.

- **Reference:** 1.Google Chrome
  2.Youtube
  3.Chatgpt

Thank You

ANY QUERIES???