# Google Distributed Search Servers

## Team members

1.Ali Saleh Ali

2.Eslam Yasser Younes

3.Amr Hany Ali

4.Ahmed Mohamed Arabi

5.Ahmed Abdelrahman

6.Yousef Reda

7. Yousef Mohamed Ahmed Eisa

Under supervision of / DR

Mohamed Tahon Azab

# What is A Search Engine?

## Overview

Search Engine refers to a huge database of internet resources such as web pages, newsgroups, programs, images, etc. It helps to find information on the World Wide Web.

Users can search for any information by passing queries in the form of keywords or phrases. It then searches for relevant information in its database and returns it to the user.

An Ideal Search Engine

How good the architecture of a search engine is determined by two requirements.

- First, effectiveness (the quality of results and how precise they are)
- Second, efficiency (how fast is response time and throughput)

### Main Search Engine Components

1. Web Crawler
2. Database
3. Search Interfaces

- Web crawler

It is also known as a spider or bot. It is a software component that traverses the web to gather information.

- Database

All the information on the web is stored in a database. It consists of huge web resources.

- Search Interfaces

This component is an interface between the user and the database. It helps the user to search through the database.

## How Search Engine Works

Search engines make use of Boolean expressions AND, OR, NOT on the user query and the given key words to restrict and widen the results of a search. Making the search results as realvant as possible steps.

## performed by the search engine:

- The search engine looks for the keyword in the index for a predefined database instead of going directly to the web to search for the keyword. It then uses software to search for the information in the database.

- Once the web crawler finds the pages, the search engine then shows the relevant web pages as a result. These retrieved web pages include the title of the page, size of the text portion, first several sentences, etc.
- These search criteria may vary from one search engine to the other. The retrieved information is ranked according to several factors such as frequency of keywords, relevancy of information, links, etc.

## Examples of search engines:

| Search Engine | Description |
|---|---|
| Google | It was originally called Backrub. It is the most popular search engine globally. |
| Bing | It was launched in 2009 by Microsoft. It is the latest web-based search engine that also delivers Yahoo's results. |
| Ask | It was launched in 1996 and was originally known as Ask Jeeves. It includes support for match, dictionary, and conversation question. |
| AltaVista | It was launched by Digital Equipment Corporation in 1995. Since 2003, it is powered by Yahoo technology. |
| AOL Search | It is powered by Google. |
| LYCOS | It is top 5 internet portal and 13th largest online property according to Media Matrix. |
| Alexa | It is subsidiary of Amazon and used for providing website traffic information. |

# Distributed system name

## Google Web Server

To a normal user, distributed computing systems appear as a single system whereas internally distributed systems are connected to several nodes which perform the designated computing tasks. Let us consider the Google web server from the user's point of view. When users submit a search query, they believe that the Google web server is a single system where they need to log in to Google.com and search for the required term. What happens is that underneath is a Distributed Computing technology where Google develops several servers and distributes them in different geographical locations to provide the search result in seconds or at time milliseconds.

## Distributed Search Engine

- Distributed search engine is a search engine where there is no central server. Unlike traditional centralized search engines, work such as crawling, data mining, indexing, and query processing is distributed among several peers in a decentralized manner where there is no single point of control
- Distributed searching is the capability to search across multiple computers. A typical search system indexes the data on a single computer. Users will search on the computer and respond with results from one computer. In distributed search, users still search the same way. The search system sends the request to all the computers which are part of the distributed search system. The computers send the result back to the

main computer and the main computer compiles a single result and sends it back to the user. The whole search experience is the same for the end user. Distributed searching ensures that a collection of independent computers appears like one single solution.

## Google search server charcteristcs

Reliability: All search servers are independent of each other. If one computer crashes the system will still survive. Distributed web server systems improve reliability through higher availability.

Cost Effective: The bigger the computer system the higher the cost and complexity. With a distributed search system, low-cost computer servers are added together, which is overall cost-effective.

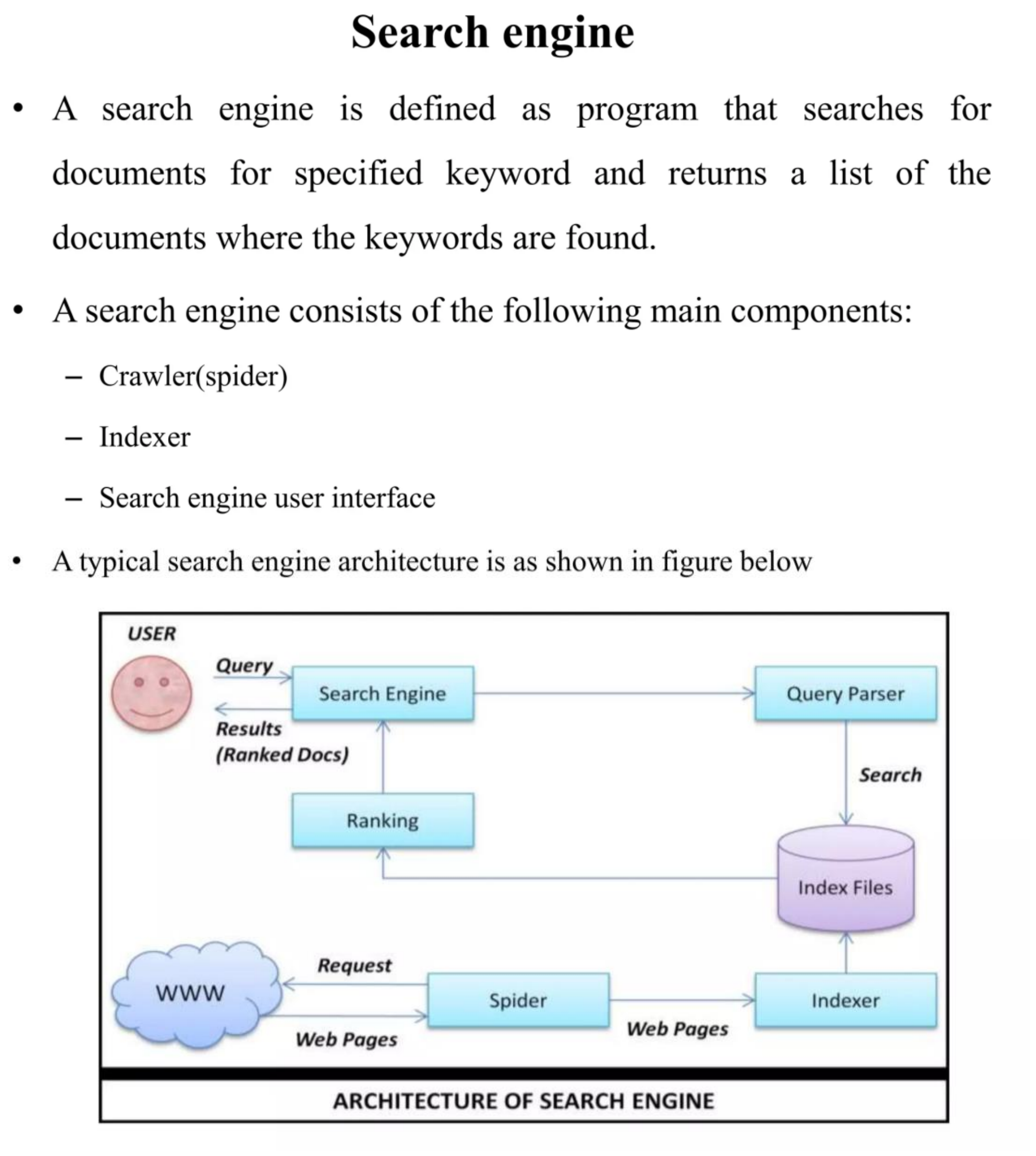
Incremental Growth: The system can grow by aligning itself with processing power requirements and storage needs. Distributed servers all over the world provide the ability to have a modular system and grow incrementally.

Resource Sharing: One of the biggest benefits of the distributed search system is resource sharing. The work is shared between the independent server nodes scattered over the globe, which keeps the overall performance of the distributed system optimal.

# System Objective / Aim

## Toward a Distributed Search Engine

In this invited talk we address the algorithmic problems behind a truly distributed Web search engine. The main goal is to reduce the cost of a Web search engine while keeping all the benefits of a centralized search engine despite the intrinsic network latency imposed by the Internet. The key ideas to achieve this goal are layered caching, online prediction mechanisms, and exploiting the locality and distribution of queries.



## Search engine

- A search engine is defined as program that searches for documents for specified keyword and returns a list of the documents where the keywords are found.
- A search engine consists of the following main components:
  - Crawler(spider)
  - Indexer
  - Search engine user interface
- A typical search engine architecture is as shown in figure below

USER
Query
Results
(Ranked Docs)
Search Engine
Query Parser
Search
Ranking
Index Files
WWW
Request
Web Pages
Spider
Web Pages
Indexer
ARCHITECTURE OF SEARCH ENGINE

- How search engine works?

A search engine operates in the following order:

1. Web crawling

2. Indexing

3. Searching

- **Web Crawling:**
  - Search engine has a huge databases of web pages . Such databases are built and updated automatically by the web crawler.
  - The web crawler performs web crawling as follows:
    - The crawler begins with one or more URLs that constitute a URL set.
    - It picks a URL from this URL set, and then fetches the web page at that URL.
    - The fetched page is then parsed to extract both the text and the links from the page.
    - The extracted links (URLs) are then added to a URL set.
    - The extracted text is fed to a text indexer.
- **Indexing:**
  - The indexer module of the search engine is responsible for indexing the extracted text supplied by the web crawler.
  - Most commonly used indexing is the inverted indexing

- **Searching:**
  - When a user enters a query to the search engine, user is not searching the entire web. Instead user is only searching the database that has been compiled by the search engine.
  - The user's query is parsed into the words by the query parser.
  - Such parsed words are matched with the words in the inverted list of indexed documents.
  - The matched list of documents are returned to the user with ranking.

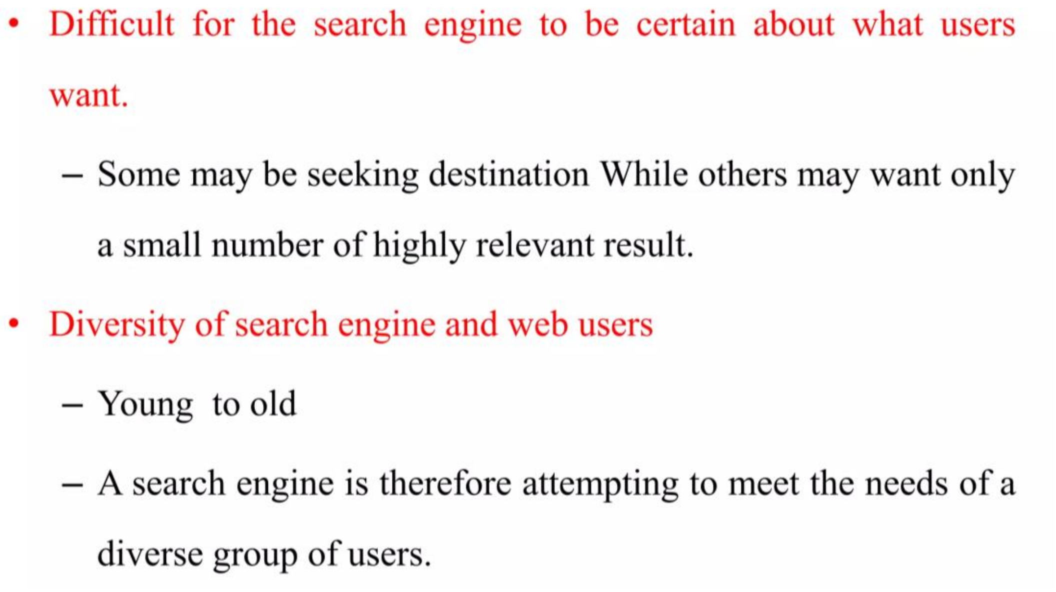
## Characteristics of search engines

- **Features a search engine must provide:**
  - **Robustness:**
    - search engine must be distributed over large number of machine to deal search engine failure due to the machine failure.
  - **Politeness:**
    - Web servers have policies regulating the rate at which a search engine can visit them. These politeness policies must be respected.

- **Features a search engine should provide**
  - **Distributed:**
    - The search should have the ability to execute in a distributed fashion across multiple machines.
  - **Scalable:**
    - The search engine architecture should permit scaling up the search rate by adding extra machines.
- **Performance and efficiency**:
  - The search system should make efficient use of various system resources including processor, storage and network.
- **Quality:**
  - Given that a significant fraction of all web pages are of poor utility for serving user query needs, the search engine should be biased towards fetching "useful" pages first.

## Problems with search using search engines

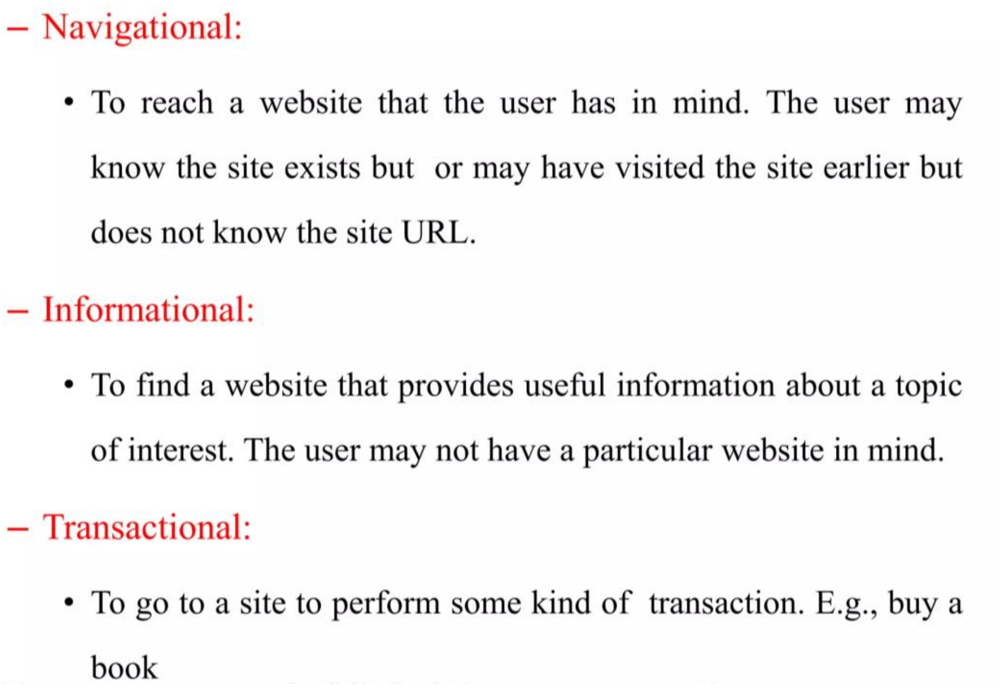
- Specifying query keywords can be challenging:
  - Search result get affected by structure of the query phrase.
  - Due to the nature of English language search result may get affected e.g., current.



- Difficult for the search engine to be certain about what users want.
  - Some may be seeking destination While others may want only a small number of highly relevant result.
- Diversity of search engine and web users
  - Young to old
  - A search engine is therefore attempting to meet the needs of a diverse group of users.

## The goals of web search

- Depending on the nature of search engine queries, the information needs of user may be divided into three classes:
  - Navigational
  - Informational
  - Transactional



- Navigational:
  - To reach a website that the user has in mind. The user may know the site exists but or may have visited the site earlier but does not know the site URL.
- Informational:
  - To find a website that provides useful information about a topic of interest. The user may not have a particular website in mind.
- Transactional:
  - To go to a site to perform some kind of transaction. E.g., buy a book

# Quality of search result

- The quality of search results from a search engine ideally should satisfy the following requirements:
  - Precision:
    - precision indicates what percentage of documents retrieved are relevant?
    - So , only relevant documents should be returned.
  - Recall:
    - means what percentage of relevant documents is retrieved from total relevant documents in the web
    - So, all relevant document should be returned
- Ranking:
  - A ranking of the documents providing some indication of the relative relevance of the results should be returned.
- First screen:
  - The first page of results should include the most relevant results.
- Speed:
  - Results should be provided quickly.