CS 8080 - Information Retrieval techniques

312217104035

S.Dhinesh

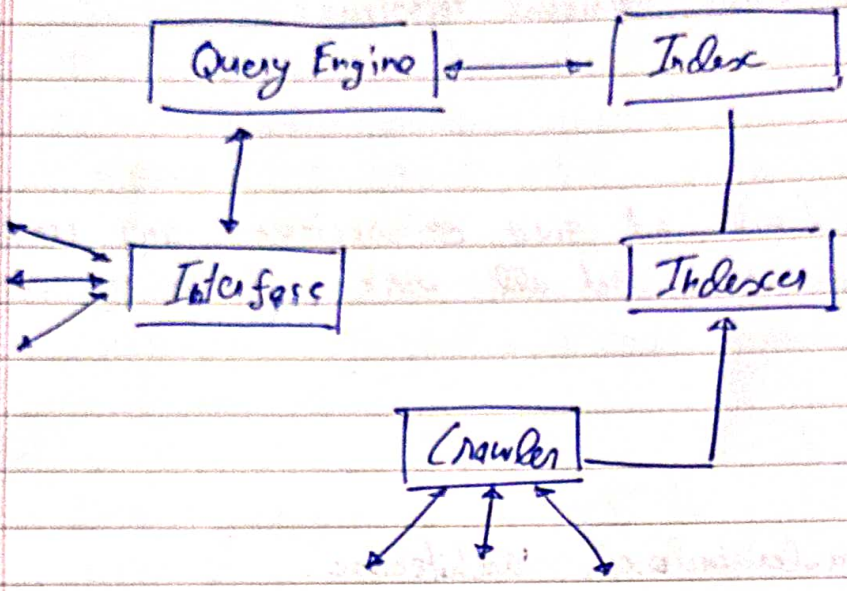
I affirm that I will not give or recieve any unauthorised help on this exam, that all work will be my own

S.Dhinesh

1.

a. Centralized crawler-indexer architecture.

The most common type of search engines is centralized crawler-indexer architecture. Crawlers are programs (software-agents) that traverse the Web sending new or updated pages to a main server where they are indexed. Crawlers are also called robots, spiders, wanders, walkers, and know bots. In spite of their name, a crawler does not actually have to and run on remote machines, rather the crawler runs on a local system and sends requests to remote web servers. The index is used in a centralized fashion to answer queries submitted from different places in the web. The following figure shows the software architecture of a search engine based on the users, consisting of the user interface and the query engine and another that consists of the crawler and indexer modules.



Problems

1. The main problem faced by this architecture is the gathering of the data, because of the highly dynamic nature of the web, the saturated communication links, and the high load of Web servers.

2. Another important problem is the volume of the data

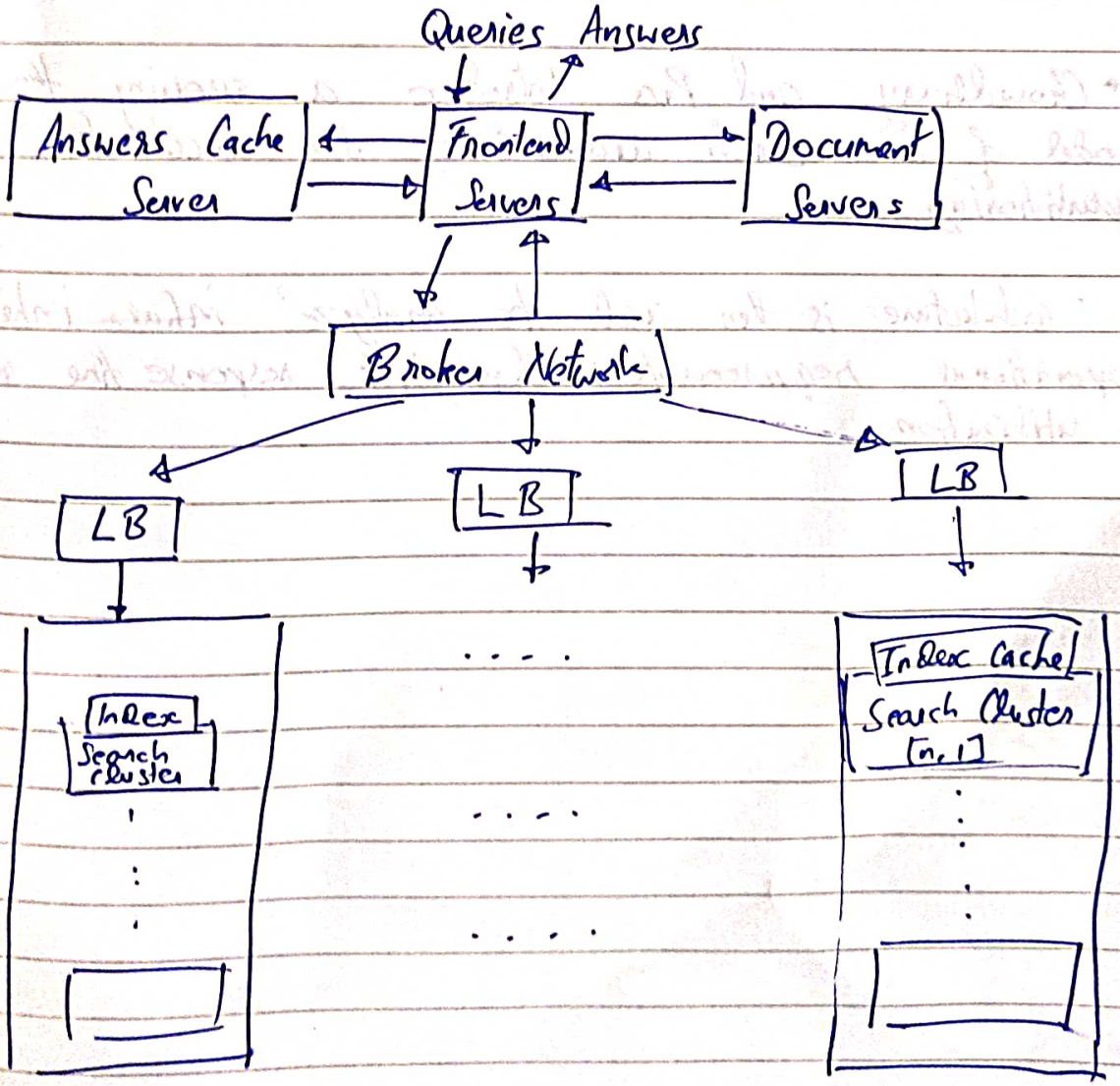
b.

Current engines adopt a massively parrallel cluster-based architecture.

- document partitioning is used.
- replicated to handle the overall query load.
- cluster replicas maintained in various geographical locations to decrease latency time.

Many crucial details need to be addressed

- good balance between the internal and external activities
- good load balancing among different clusters
- fault tolerance at software level to protect against hardware failure.



Architecture

→ Orlando et al present a parallel and distributed search engine architecture based on two strategies

- a task parallel strategy: a query is executed independently by a set of homogeneous index servers.

- a data parallel strategy: a query is processed in parallel by index servers accessing distinct partitions of the database.

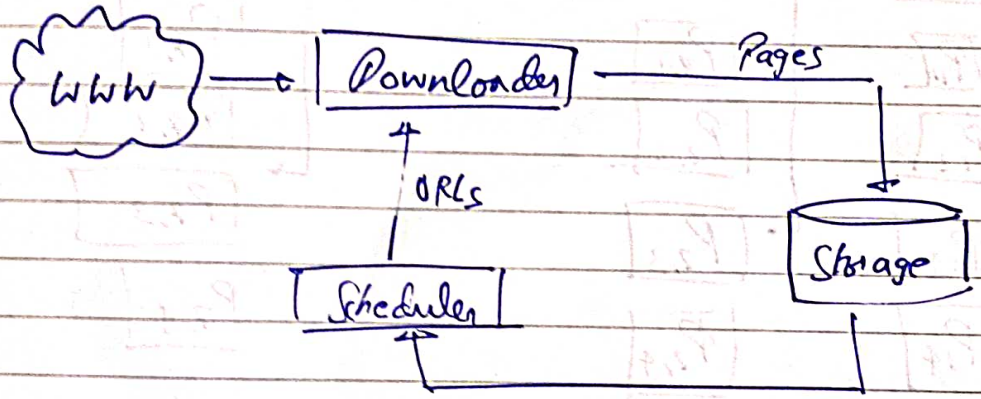
→ Chowdhury and Pass introduce a queuing theory model of a search architecture for document partitioning.

- architecture is then used to analyze inherent operational requirements: throughput, response time and utilization.

3. a. Architecture of the crawlers.

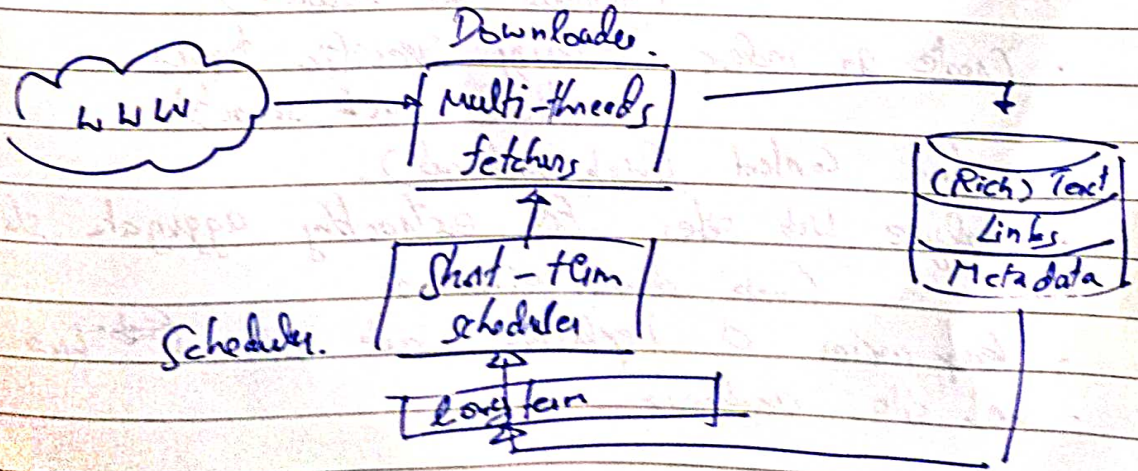
· The crawler is composed of three main modules: downloader, storage, and scheduler.

→ Scheduler : maintains a queue of URLs to visit.
→ Downloader : downloads the pages
→ Storage : makes the indexing of the pages and provides to the scheduler with metadata on the pages retrieved.

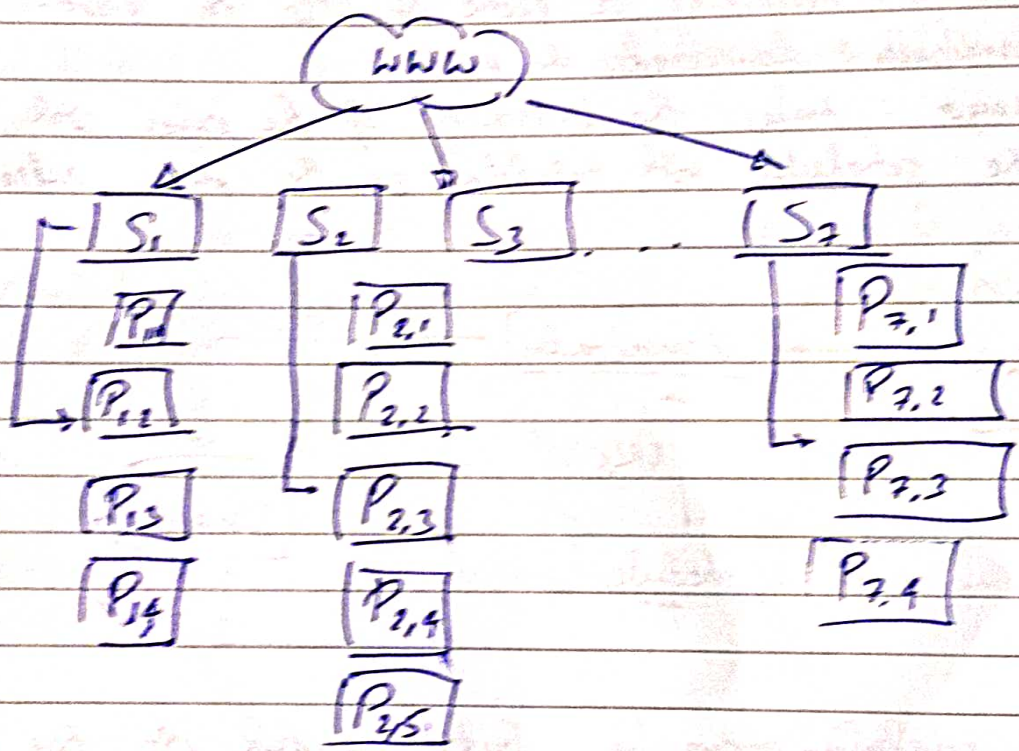


· The scheduling can be further divided into two parts.
long term :- decide which pages to visit next.
short term :- re-arrange pages to fulfill politeness.

· The storage can also be further subdivided into three parts: (rich) text, metadata and links.



In the short-term scheduler, enforcement of the politeness policy requires maintaining several queues, one for each site, and a list of pages to download in each queue



b. Applications of web crawler

A web crawler is used for

- Create an index covering broad topics (general web search)
- Create an index covering specific topics. (vertical web search).
- Caching content (Web archival).
- analyze web sites for extracting aggregate statistics (web characterization)
- keep copies of replicate web sites (Web mirroring)
- Web site analysis.

5.

Given Alice has not seen Alice.
→ find a set of users who liked the same item a Alice.
→ Use the average of their ratings to predict, if Alice will like item i.

Pearson's correlation.

$$Sim(a,b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

a, b → Users
$r_{a,p}$ → rating of user a for item p.
P → set of items, rated both by a and b.

$$\vec{r}_{Tim} = \frac{12}{5} = 2.4$$

$$\vec{r}_{Dom} = \frac{19}{5} = 3.8$$

$$\vec{r}_{John} = \frac{16}{5} = 3.2$$

$$\vec{r}_{bob} = \frac{19}{5} = 2.8.$$

$$Sim(Alice, Tim) = \frac{(1.8 \times 0.6 + 0.2 \times 1.4 + 0.4 \times 0.8 + 0.6 \times 0.2 + 0)}{\sqrt{4.56 \times 3.12}}$$

$$= 0.86$$

$$Sim(Alice, Dom) = \frac{0}{\sqrt{4.56 \times 2.78}} = 0$$

$$Sim(Alice, John) = \frac{(1.8 \times 0.2) + (1.4 \times 0.2) + (0.4 \times 2.2) + (2.2 \times 0.6)}{\sqrt{4.56 \times 3.76}}$$

$$= 0.70$$

$$Sim(Alice, Bob) = \frac{(-1.8 \times 2.7) + (2.2 \times 1.4) + (2.2 \times 0.4) - (0.8 + 0.6)}{\sqrt{4.56 \times 2.64}}$$

$$= -0.79$$

Majority of the similarities are positive, => Alice would like gladiator.