UNIT III – WEB SEARCH ENGINE – INTRODUCTION AND CRAWLING

Part A – Question Bank

1. **Define web server.**

   Web server is a computer connected to the internet that runs a program that takes responsibility for storing, retrieving and distributing some of the web files.

2. **What is web Browsers?**

   A web browser is a program. Web browser is used to communicate with web servers on the Internet, Which enables it to download and display the web pages. Netscape Navigator and Microsoft Internet Explorer are the most popular browser software's available in market.

3. **Explain paid submission of search service.**

   In paid submission user submit website for review by a search service for a preset fee with the expectation that the site will be accepted and include d in that company's search engine, provided it meets the stated guidelines for submission. Yahoo! is the major search engine that accepts this type of submission. While paid submissions guarantee a timely review of the submitted site and notice of acceptance or rejection, you're not guaranteed inclusion or a particular placement order in the listings.

4. **Explain paid inclusion programs of search services.**

   Paid inclusion programs allow you to submit your website for guaranteed inclusion in a search engines database of listings for a set period of time. While paid inclusion guarantees indexing of submitted pages or sites in a search database, you're not guaranteed that the pages will rank well for particular queries.

5. **Explain in pay-for-placement of search services.**

   In pay-for-placement, you can guarantee a ranking in a search listing for the terms of your choice. Also known as paid placement, paid listing, or sponsored listings, this program guarantees placement in search results. The leaders in pay-for-placement are Google, Yahoo! and Bing.

6. **Define Search Engine Optimization.**

   Search Engine Optimization is the act of modifying a website to increase its ranking in organic, crawler-based listing of search engines. There are several ways to increase the visibility of your website through the major search engines on the internet

today. The two most common forms of internet marketing paid placement and natural placement.

7. Describe benefit of SEO.

- Increase your search engine visibility
- Generate more traffic from the major search engines.
- Make sure your website and business get NOTICED and VISITED.
- Grow your client base and increase business revenue.

8. Explain the difference between SEO and Pay-per-click

SEO	Pay-Per-click
SEO results take 2 weeks to 4 months	It results in 1-2 days
It is very difficult to control flow of traffic	It has ability to turn on and at any moment
Requires ongoing learning and experience to reap results	Easier for a novice
It is more difficult to target local markets	Ability to target "local" markets
Better for long-term and lower margin campaigns	Better for short-term and high-margin campaigns.
Generally more cost-effective , does not penalize for more traffic	Generally more costly per visitor and per conversion

9. What is web crawler?

A web crawler is a program which browses the world web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to p[provide fast searches.

10. Define focused crawler.

A focused crawler or topical crawler is a web crawler that attempts to download only pages that are relevant to a pre-defined topic or set of topic.

11. What is hard and soft focused crawling?

In **hard focused crawling** the classifier is invoked on a newly crawled document in a standard manner. When it returns the best matching category path, the out-neighbors of the page are checked into the database if and only if some node on the best matching category path is marked as good.

In **soft focused crawling** all out-neighbors of a visited page are checked into DB2, but their crawl priority is based on the relevance of the current page.

12. What is the Near-duplicate detection?

Near-duplicate is the task of identifying documents with almost identical content. Near- duplicate web documents are abundant. Two such documents differ from each other in a very small portion that displays advertisements, for example. Such differences are irrelevant and for web search.

13. What are requirements of XML information retrieval systems?

- Query language that allows users to specify the nature of relevant components, in particular with respect to their structure.
- Representation strategies providing a description not only of the content of XML documents, but also their structure.
- Ranking strategies that determine the most relevant elements and rank these appropriately for a given query.

Part – B

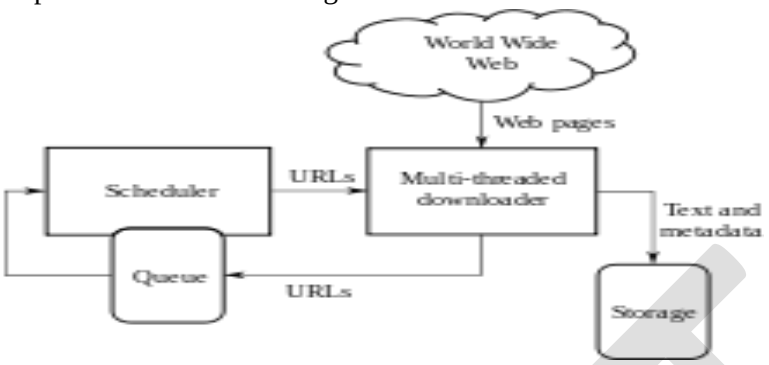
1. In what way is the signature approach advantageous over other text retrieval methods? Classify the signature based methods.
2. How is a Patricia tree constructed? Construct a Patricia tree for 01100100010111. Discuss the algorithms on PAT tree.
3. What is a use of PAT tree? How can PAT trees be represented as arrays?
4. List the technical issues for file system. Explain WOBT with three nodes having C,D, F, G and H as records.
5. Discuss the procedure to implement a lexical analyzer.
6. What is a stop list? Give some examples of stop words. How can it be used along with lexical analyzer.

Part – B

1. Explain about web Search Architecture

A **web search engine** is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

2. Explain about Web crawling architecture.



A **Web crawler**, sometimes called a **spider**, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (*web spidering*).

Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently.

3. Write short notes on XML retrieval .

Queries

Exploiting XML structure

Ranking

Existing XML search engines

Traditional XML query languages

Databases

Information retrieval

Data-Centric XML Datasets

4. Explain search engine optimization

Search engine optimization (**SEO**) is the process of affecting the visibility of a website or a web page in a web search engine's unpaid results—often referred to as "natural", "organic", or "earned" results. In general, the earlier (or higher ranked on the search results page), and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users; these visitors can then be converted into customers.[1] SEO may target different kinds of search, including image search, video search, academic search,[2] news search, and industry-specific vertical search engines. SEO differs from local search engine optimization in that the latter is focused on optimizing a business' online presence so that its web pages will be displayed by search engines when a user enters a local search for its products or services. The former instead is more focused on national searches.