

[Advanced learner]

Q1. What are the challenges faced by search engines?

A search engine is a software system that is designed to carry out web searches. They search the world wide web in a systematic way for particular information specified in a textual web search query.

The reality is that search engines are far from perfect. They face significant challenges trying to decipher our intentions and find the pages that match with our queries.

Here are some of the problems -

1. Repetitive title tags =>

The most common is the number of repetitive title tags used. This is something Google (search engine) heavily penalizes for and something that is easily fixed.

2. Malicious URLs =>

Search engine spiders have to be very efficient, so they are leery of anything that seems like a risk to them, even though it may not be. Dynamically generated pages present a risk to them because the spiders could get caught up in an infinite loop within the site. URLs having long, complex query strings will have a harder time getting indexed with everything else being equal, than a shorter URL with no special characters.

28/2/2008 (2)

3. Spam \Rightarrow Search engines identify relevant pages by means of 'quality signals' or metrics that can be deduced from web pages by automated means. The challenge to search engines is that once black hat SEOs identify those signals, they can start to finely fake them. For humans, it is easy to spot spams but for computers it is much harder.
4. Polysemy \Rightarrow There are words that change their meaning depending on the context in which they are requested i.e. if you do a search for 'wood' you might want pages that are talking about pieces of a tree, or you might be talking about the geographical area that has many trees. Without the right context, it is hard for a human to tell. Imagine how hard it is for the search engine!
5. Poor queries \Rightarrow Many searchers don't know how to express what they want in the real world, and are even worse attempting to ask a search engine. Worst yet, they misspell words, making the problem more 'interesting' for search engine.
6. Synonymy \Rightarrow This is another challenge. There are words that have the same meaning, like 'car' and 'automobile'. When you do a search you would like to get pages that contain your exact words, and pages that contain other words that mean the same thing, as long as they are relevant to your search. These details don't pass through the minds of most searchers, but search engines

28/2/2008 (3)

have a hard time because of it.

7. Natural language searches =>

A MySQL database engine can precisely return all the relevant records given a query 'select first last from employee where last = "smith"'. There is a formal syntax and no ambiguity. A search engine on the other hand, receives 'who has smith as last name in Chicago' or 'smith last name Chicago'. There are many different ways to say the same i.e. the context, there are human idiosyncrasies. The searcher component of a search engine must disambiguate the query and translate it into a more formal manner before looking it up in the index.

8. It can't do it All.

Indexing all of the content on the web regularly is more than a daunting task, it's impossible. Google (a search engine) has indexed trillions of webpages, but it has not and likely will never index every single piece of content on the web. That's a result of the robustness of the web's content offering, and Google (a search engine) has ~~indexed~~ also admits that it cannot crawl all elements of websites effectively.

Written text is easy to crawl and index while images, videos and certain HTML elements are not easily crawl and indexed.

Q2. Compare the working of normal search engine and Meta Search Engine.

Working of normal search engine

Search engines are generally working on three parts that are crawling, indexing and ranking.

1. Crawling \Rightarrow Search engines have a number of computer programs that are responsible for finding info. that is publicly available on the internet. These programs scan the web and create a list of available websites. Then they visit each website and by reading HTML code they try to understand the structure of the page, the type of content, the meaning of the content, and when it was created or updated.
2. Indexing \Rightarrow Information identified by the crawler needs to be organized, sorted and stored so that it can be referred later by the ranking algorithm. Search engines don't store all the information in your index, but they keep things like the title and description of the page, the type of content, associated keywords, no. of incoming and outgoing links, and a lot of other parameters that are needed by the ranking algorithm.
3. Ranking \Rightarrow Ranking is the position by which your website is listed in any search engine.
 . Step 1 - Analyze user query \Rightarrow This step is to understand what kind of information user is looking for. To do that, it is break down into a number of meaningful keywords.

28/2020 B (3)

A keyword is a word that has a specific meaning and purpose.

- Step 2 : Finding matching pages \Rightarrow this step is to look into the index and find the best matching pages, for example, if you search 'dark wallpaper' then it gives you the result of images not text.
- Step 3: Present the results to the user \Rightarrow A typical search results page includes ten organic results in most cases it is enriched with other elements like paid Ads, direct answers for specific queries, etc.

Working of meta search engines

It doesn't have a repository or index of their own, they take advantage of indices created by other search engines. In fact, they present this as their strong point. A typical meta search engine pulls off the results from a number of search engines, say Google and Bing, and then apply their own algorithms in some cases to re-order the results.

A metasearch engine does not create a database of itself rather it creates a federal database that is actually an integration of the databases of various other search engines.

It develops its own algorithm and it eliminates duplicate results and calculates fresh ranking of the sites.

28/2020 8 (8)

Q3. Design a web form using HTML and also explain the process to publish it as a web page on internet.

```
<!DOCTYPE HTML>
<html>
<head>
<meta charset = "UTF-8">
<title> Web form </title>
</head>
<body>
<h2> HTML Form </h2>

<form action = "/response-page.php">
    <label for = "fname"> First name: </label> <br>
    <input type = "text" id = "fname" name = "fname"
           value = "John" > <br>
    <label for = "lname"> Last name: </label> <br>
    <input type = "text" id = "lname" name = "lname"
           value = "Doe" > <br> <br>
    <input type = "submit" value = "submit" >
</form>

<p> If you click the "Submit" button, the form
    -data will be sent to a page called
    "/response-page.php". </p>

</body>
</html>
```

28/2020 8 (2)

There are various options to publish the content on the internet -

- Getting Hosting and a domain name ⇒
Web hosting is rented file space on a hosting company's web server. You put website files on the web server. The web server provides website content to website visitors.

A domain name is the unique address where people ~~may~~ find your website like `http://www.nameof.org`. You can rent your domain name for as many years as you want from a domain registrar.

In addition, you will need a File Transfer Protocol program to actually transfer the website files over to the server. FTP programs vary widely, but generally, you have to connect to your web server using details provided by your hosting company (typically username, password, host name).

- When looking for hosting, you should consider the speed, reliability, level of security, as well as the ease of scalability to accommodate future traffic growth
- Before you hit the publish button or putting the files, you should -

28/20/2018 (8)

- proofread the content
- check the spelling
- do grammar check
- Check all the text
- Optimize images
- Follow proper semantic structure
- Test links and integrations

The other options to publish content on internet \Rightarrow

- Using an online tool like GitHub or Google App Engine
- Using a web-based IDE such as codePen
- Publishing via GitHub.