دعاء أشرف حسن
CS
IR
Solution Lectures Assignments

Solution Assignments 1 of Lecture 3

How to add your website to Google index?

To add your website to Google's index, you can follow these steps:
1. Sign into your Google account and go to the Google Search Console (formerly known as Google Webmaster Tools).
2. Click on "Add a property" and enter your website's URL.
3. Verify ownership of your website by following the instructions provided by Google. This may involve uploading an HTML file to your website, adding a meta tag to your website's header, or verifying through your domain name provider.
4. Once ownership is verified, go to the "Crawl" tab in the Search Console and click on "Fetch as Google".
5. Submit your website's URL and click on "Fetch and Render".
6. After the URL is fetched, click on "Request Indexing".
7. Choose whether to crawl only that URL or the URL and its direct links.
8. Click on "Go" to submit the request.

It may take some time for Google to crawl and index your website, but this process should help speed up the process. Additionally, you can create and submit a sitemap to Google to help them better understand the structure and content of your website.

Solution Assignments 2 of Lecture 3

How can we reduce index size?

Reducing index size can help improve the performance of your search engine and can make it easier to manage and maintain. Here are some ways to reduce the index size:
1. Remove unnecessary content.
2. Use meta tags.
3. Control crawl rate.
4. Optimize content.
5. Compress and optimize images.
6. Use pagination.

By implementing these strategies, you can help reduce the size of your search engine index, improve search engine performance, and provide a better user experience.

Solution Assignments 1 of Lecture 4

How web page duplication detection?

Web page duplication, also known as "duplicate content," occurs when identical or very similar content appears on multiple web pages. Duplicate content can negatively impact search engine rankings and user experience. Here are some methods for detecting web page duplication:
1. Manual checking.

2. Google Search Console.
3. Copy scape.
4. SEO tools.
5. Canonical tags.
By detecting and addressing web page duplication, you can help improve your website's search engine rankings, avoid penalties, and provide a better user experience.

Solution Assignments 2 of Lecture 4

Explain in detail Crawling policies, including:

1- a selection policy, which pages to download.
2- A re-visit policy which states when to check for changes to the pages.
3- a politeness policy that states how to avoid overloading Web sites.
4- a parallelization policy that states how to coordinate distributed web crawlers.

What are the methods that prevent web pages from being indexed by traditional search engines?

Crawling Policies:

Crawling policies are sets of rules and guidelines that web crawlers use to navigate through websites, collect data, and index web pages. The following are the main crawling policies:
1. Selection policy: This policy determines which web pages should be downloaded by the crawler. Some factors that determine this include the relevance of the page, its popularity, its freshness, and its value to users. The selection policy is usually determined by the search engine and can change over time.
2. Re-visit policy: This policy determines how often a crawler should revisit a web page to check for changes. The re-visit policy is based on the frequency of updates to the web page and its importance to users. Pages that are frequently updated or are more important to users may be revisited more often.
3. Politeness policy: This policy ensures that web crawlers do not overload websites with requests, which can cause a decrease in website performance. It includes rules such as the maximum number of requests per second, the minimum time between requests, and the maximum depth of a crawl. Politeness policies help maintain a positive relationship between search engines and website owners.
4. Parallelization policy: This policy determines how distributed web crawlers coordinate to avoid duplicating requests or creating excessive traffic. It includes rules such as the use of shared databases, the distribution of crawlers, and the frequency of updates.

Methods to Prevent Web Pages from being Indexed by Traditional Search Engines:

1. Robots.txt file: This file is placed in the root directory of a website and includes instructions that tell search engines which pages not to crawl or index.
2. No index Meta tag: This tag can be added to the header of a webpage to instruct search engines not to index it.
3. Password protection: Websites can be protected with a password, which prevents search engines from indexing the pages.
4. Canonical tags: These tags are used to specify the preferred version of a web page, which can prevent duplicate content issues and ensure that search engines index the correct version of a page.
5. JavaScript: JavaScript can be used to dynamically generate web

pages, which can make it difficult for search engines to crawl and index them.
6. Content Delivery Networks (CDNs): CDNs can be used to deliver website content, which can make it difficult for search engines to identify the original source of the content.