Assignment 6

Web crawling

Crawling is where it all begins: the acquisition of data about a website.

This involves scanning sites and collecting details about each page: titles, images, keywords, other linked pages, etc. Different crawlers may also look for different details, like page layouts, where advertisements are placed, whether links are crammed in, etc.

**But how is a website crawled?** An automated bot (called a “spider”) visits page after page as quickly as possible, using page links to find where to go next. Even in the earliest days, Google’s spiders could read several hundred pages per second. Nowadays, it’s in the thousands.

When a web crawler visits a page, it collects every link on the page and adds them to its list of next pages to visit. It goes to the next page in its list, collects the links on that page, and repeats. Web crawlers also revisit past pages once in a while to see if any changes happened.

This means any site that’s linked from an indexed site will eventually be crawled. Some sites are crawled more frequently, and some are crawled to greater depths, but sometimes a crawler may give up if a site’s page hierarchy is too complex.

One way to understand how a web crawler works is to build one yourself. We’ve written a tutorial on **creating a basic web crawler in PHP**, so check that out if you have any programming experience.

Note that pages can be marked as “no index,” which is like asking search engines to skip its indexing

Indexing

Indexing is when the data from a crawl is processed and placed in a database.

Imagine making a list of all the books you own, their publishers, their authors, their genres, their page counts, etc. Crawling is when you comb through each book while indexing is when you log them to your list.

**Now imagine it’s not just a room full of books, but every library in the world.** That’s a small-scale version of what Google does, who stores all of this data in vast data centers with [**thousands of petabytes worth of drives**](https://www.makeuseof.com/tag/memory-sizes-gigabytes-terabytes-petabytes/).