Search Research

(pun intended)

Notes

General

- Before search engines can results for a query, it needs to know where to get possible results - do this by crawling the Internet [1]
- Steps:
    - crawl
    - render
    - index

Crawling

- Done with spiders → bots that crawl the web to find as many webpages as possible, and what they're each about [1]
    - Diff search engines have diff spiders
    - Google's spiders
        - crawl websites geographically close to them, to minimise bandwidth usage [15]
        - are specialised for diff media types (ex. images, videos, news stories) [4]
- Spider:
    - Starts from a seed → list of known URLs [1]
        - Might include URLs that the spider has already crawled [3]
            - *Makes sense for re-crawling to update pages*
    - Iterate through the URLs:
        - Check robots.txt to see if allowed to crawl
            - No → move on to next page [1]
                - Doesn't mean page won't show up in search though – engine just won't know what's on it [13]
        - Fetch page
            - Not available/is redirected → move on to next page
        - Check headers for indexing permission [14]
            - No → just crawl it
        - Go through page to look for hyperlinks
            - For each hyperlink:
                - Check if noindex tag used
                - Add it to list of pages to crawl
                    - Pages not necessarily crawled in order of addition – diff search algos prioritise differently based on [1]:
                        - Popularity → number of links that link to that URL, from other pages
                        - Popularity → amount of visitors
                    - Google crawls in depth-first order [2]

    - Stop once got enough
    - Repeat again to make sure have up-to-date info, bc web is always changing [1, 12]
        - But will control crawl speed to not overload sites [3, 8]
        - Speed unique to sites based on [8]:
            - Reaction speed to Googlebot's reqs
            - Content quality
            - Potential server errors
- Have diff parts
    - According to [10, 7:23], have *three* diff parts:
        - Fetcher → download sth off the internet
        - Controller → merges links from HTML of fetched pages with links from sitemaps, gives them to fetcher to fetch
        - Scheduler → tells fetcher when to fetch
    - According to [12], have *four* diff parts:
        - Selection policy → how it chooses what to crawl
        - Revisit policy → how often the spider comes back to a resource to see if it's been updated
        - Politeness policy → how it responds to server reqs to not overload it
        - Parallelisation policy → how to have multiple crawls going at the same time, without re-crawling URLs
- Not just search engines who crawl [10]:
    - Chatbots
    - SEO services

Indexing

- Recording what a webpage is about, using its text content (no images or videos) and metadata [1, 5]
    - Text has to be part of the DOM – stuff in a canvas or added with CSS doesn't count [5]
        - Will render it before processing though, so JS counts [6]
    - Ignores filler words [1]
    - Metadata isn't just the HTML tags, it's also the JSON schema [1]
- Also detects whether page is a duplicate (duplicate clustering) [9]
    - Diff language versions of a page count as duplicates if main content is unchanged (they're localised versions of the same page) [6]
        - Indicate them by listing the alternate URLs, inc. itself in link tags in the document head [6]
        - Or lay sitemap out correctly [6]
    - Only shows the canonical page [3]
        - Determine which page is canonical by:
- Stored in database of some sort (index selection) [9]
    - Google uses distributed database called the Google index [3]
    - Pages not indexed if don't meet a quality threshold [8]

Ranking

- Happens as part of search
- Pages ranked according to:
    - Reliability (static) [2]
    - Relevance to query (dynamic) [2]
        - Location [3]
        - Language [3]
        - Device [3]
        - Significance of query (wrt. life or death matters, sensitive topics, etc) [15]

Sources

[1] "What is a web crawler? | How web spiders work". *Cloudflare*. Available: https://www.cloudflare.com/learning/bots/what-is-a-web-crawler/. [Accessed Aug. 14, 2024].

[2] K. Dearie. "Website Crawlers: What They Are & How to Use Them". *Semrush Blog*, Dec. 21, 2023. Available: https://www.semrush.com/blog/website-crawler/. [Accessed Aug. 14, 2024].

[3] "In-depth guide to how Google Search works". *Google Search Central*. Available: https://developers.google.com/search/docs/fundamentals/how-search-works. [Accessed Aug. 14, 2024].

[4] "Overview of Google crawlers and fetchers (user agents)". *Google Search Central*. Available: https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers. [Accessed Aug. 14, 2024].

[5] "Get started with Search: a developer's guide". *Google Search Central*. Available: https://developers.google.com/search/docs/fundamentals/get-started-developers. [Accessed Aug. 14, 2024].

[6] "Tell Google about localized versions of your page". *Google Search Central*. Available: https://developers.google.com/search/docs/specialty/international/localized-versions. [Accessed Aug. 14, 2024].

[7] "How to specify a canonical with rel="canonical" and other methods". *Google Search Central*. Available: https://developers.google.com/search/docs/crawling-indexing/consolidate-duplicate-urls. [Accessed Aug. 14, 2024].

[8] Google Search Central, "How Google Search crawls pages". *YouTube*. Available: https://www.youtube.com/watch?v=JuK7NnfyEuc&list=PLKoqnv2vTMUN83JWBNM6MoBuBcyqhFNY3&index=2. [Accessed Aug. 14, 2024].

[9] Google Search Central, "How Google Search indexes pages". *YouTube*. Available: https://www.youtube.com/watch?v=pe-NSvBTg2o&list=PLKoqnv2vTMUN83JWBNM6MoBuBcyqhFNY3&index=3. [Accessed Aug. 14, 2024].

[10] Google Search Central, "What is a web crawler, really?" *YouTube*. Available: https://www.youtube.com/watch?v=xVg9LcrSwyQ. [Accessed Aug. 14, 2024].

[11] Google Search Central, "Let's talk ranking updates". *YouTube*. Available: https://www.youtube.com/watch?v=bjELEAeIQyY. [Accessed Aug. 14, 2024].

[12] S.S. Dhenakaran and K.T. Sambanthan. "Web Crawler – An Overview". *International Journal of Computer Science and Communication*, Jan-Jun 2011, vol. 2, no. 1, pp. 265-267. Available: https://www.csjournals.com/IJCSC/PDF2-1/Article_49.pdf. [Accessed Aug. 21, 2024].

[13] "Crawlablity". *Yoast Academy*. Available: https://academy.yoast.com/topic/crawlability/. [Accessed Aug. 21, 2024].

[14] "Robots meta tag, data-nosnippet, and X-Robots-Tag specifications". *Google Search Central*. Available: https://developers.google.com/search/docs/crawling-indexing/robots-meta-tag. [Accessed Aug. 22, 2024].

[15] Google. "A Google documentary | Trillions of questions, no easy answers". *YouTube*. Available: https://www.youtube.com/watch?v=tFq6Q_muwG0. [Accessed Aug. 14, 2024].