CRAWLER DOCUMENTATION

Nguyễn Ngọc Sơn

Hà Nội – 2025

Mục lục

1 LÝ THUYẾT VỀ CRAWLER **4**

1 Khái niệm về Crawler. 4

2 Kiến trúc Crawler. 4

3 Các dạng đối tượng để Crawler. 5

4 Kỹ thuật Crawler. 5

5 Crawler nâng cao. 6

6 Các công cụ Crawler. 6

7 Khía cạnh pháp lý. 7

2 ỨNG DỤNG CRAWLER **8**

Danh sách hình vẽ

1.1 Hệ thống Crawler 4
2.1 Các thành phần của hệ thống Crawler 4

0.1 Sơ đồ ứng dụng Crawler 8

Danh sách bảng

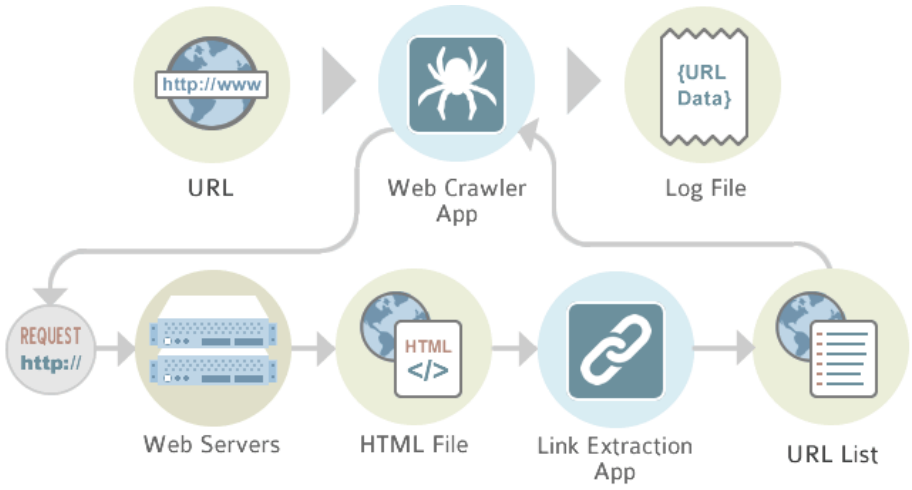
3.1 Phân loại trang web và phương pháp thu thập dữ liệu phù hợp 5
6.1 Các công cụ phổ biến trong thu thập dữ liệu trực tuyến 6
7.1 Các quy tắc và nguyên tắc khi thu thập dữ liệu trực tuyến 7

LÝ THUYẾT VỀ CRAWLER

1. Khái niệm về Crawler.

Crawler (hoặc web spider, bot) là chương trình tự động thu thập dữ liệu từ Internet...

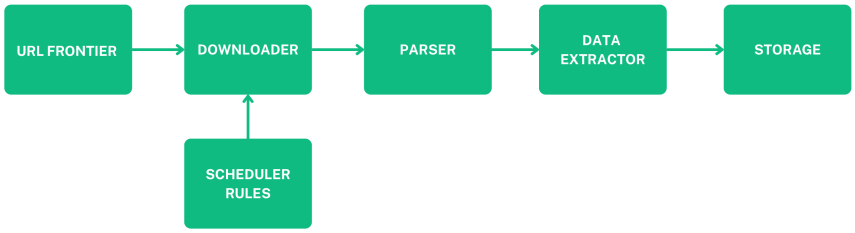
Ví dụ:



Hình 1.1: Hệ thống Crawler

2. Kiến trúc Crawler.

Một hệ thống crawler cơ bản gồm các thành phần: URL Frontier, Downloader, Parser, Scheduler, Storage.



Hình 2.1: Các thành phần của hệ thống Crawler

1. URL Frontier
   - Đây là danh sách các URL chúng ta cần để thu thập thông tin.

     Ví Dụ:
   - Chúng ta có thể thiết lập độ ưu tiên theo domain, theo thời gian cập nhật, theo nội dung...
2. Downloader
   - Khi mà thực hiện thao tác gửi HTTP Request để có thể lấy nội dung (HTML, JSON, v.v...)

- Có thể sử dụng requests, aiohttp hoặc Playwright/Selenium nếu trang dùng JavaScript.

3. Parser
   - Phân tích HTML/XML bằng thư viện như BeautifulSoup, lxml hoặc parsel.
   - Tìm phần tử chứa dữ liệu cần.
4. Data Extractor
   - Áp dụng selector (CSS, XPath, regex) để trích thông tin cụ thể.
   - Có thể kèm thêm ML model nếu cần phát hiện nội dung phức tạp (ví dụ: bài đăng leak data).
5. Storage
   - Lưu dữ liệu đã trích xuất vào database (SQLite, MongoDB, Elasticsearch, . . .)
   - Hoặc ghi ra file (CSV, JSONL,. . .).
6. Scheduler & Rules
   - Điều phối luồng crawl, tránh trùng lặp, giới hạn tốc độ, kiểm soát domain, obey robots.txt.
   - Đặt thời gian refresh (recrawl interval).

3. Các dạng đối tượng để Crawler.

Bảng 3.1: Phân loại trang web và phương pháp thu thập dữ liệu phù hợp

Loại trang	**Đặc điểm**	**Phương pháp Crawl phù hợp**
HTML tĩnh	Nội dung sẵn trong mã HTML	`BeautifulSoup` + `requests`
HTML động (JS render)	Dữ liệu chỉ xuất hiện khi JavaScript được thực thi (SPA)	`Playwright`, `Selenium`, hoặc API sniffing
API trả JSON	Có endpoint JSON riêng biệt	Gửi request trực tiếp tới API
Trang có login / Cloudflare	Cần session, cookie, proxy	Giả lập session, automation, hoặc bypass hợp pháp
Forum / Telegram / Dark Web	Nội dung động, hạn chế quyền truy cập	Kết hợp API chính thức + Browser render (`Playwright`)

4. Kỹ thuật Crawler.

Khi thực hiện thu thập thông tin thì có một số kỹ thuật được sử dụng cho các bước như sau:

1. HTTP Request & Response.

   Đầu tiên Crawler gửi request với:
   - User-Agent: để định danh trình duyệt hoặc bot.
   - Cookie / Header: Để có thể mô phỏng người dùng.
   - Proxies: để tránh block hoặc bị giới hạn vùng.
2. Parsing.

- Dùng CSS selector (soup.select(".class a")) hoặc XPath (//div[@class='title']/a).
- Tách thông tin văn bản, link, hình ảnh,...

3. Scheduling.
   - Crawl định kỳ (vd: mỗi 12 giờ kiểm tra leak mới).
   - Giới hạn tốc độ (sleep, asyncio.Semaphore) để tránh bị chặn IP.
4. Error Handling.
   - Retry khi lỗi 403, 429, 500.
   - Giới hạn số lần thử per-domain.

5. Crawler nâng cao.

1. Asynchronous Crawling: dùng aiohttp + asyncio để crawl hàng trăm trang song song.
2. Distributed Crawling: dùng Kafka + Celery hoặc Scrapy Cluster.
3. Headless Browsers: Playwright, Puppeteer, Selenium để render trang có JS.
4. Smart Crawler: sử dụng AI/NLP để lọc nội dung (vd: tìm bài rò rỉ data).

6. Các công cụ Crawler.

Bảng 6.1: Các công cụ phổ biến trong thu thập dữ liệu trực tuyến

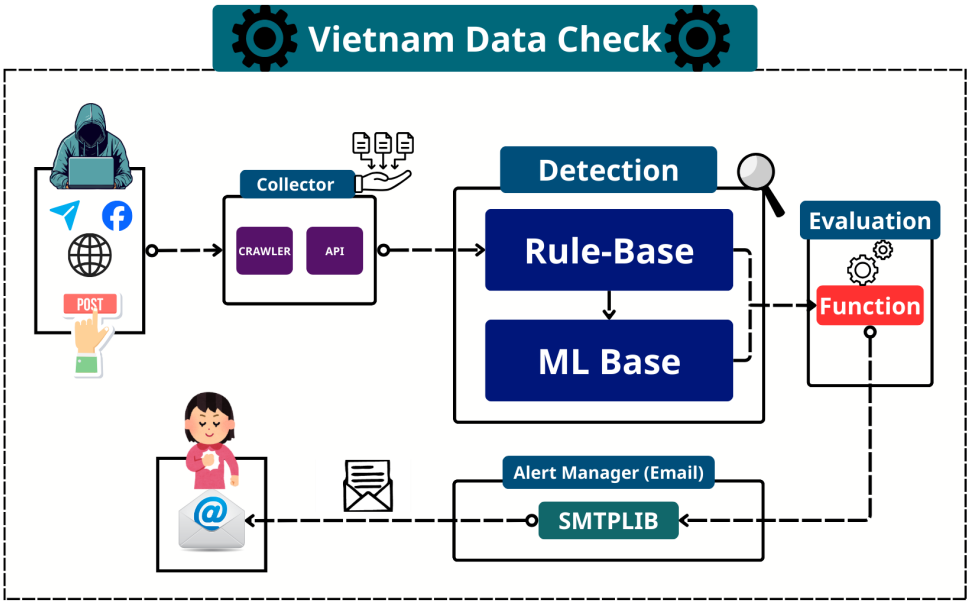
Công cụ	Mô tả	Ngôn ngữ
Scrapy	Framework mạnh mẽ cho crawl quy mô lớn, có scheduler, pipeline và hệ thống lưu trữ dữ liệu.	Python
BeautifulSoup	Thư viện đơn giản, dễ sử dụng để parse và trích xuất dữ liệu từ HTML.	Python
Playwright / Selenium	Trình duyệt headless hỗ trợ render JavaScript, tương tác với trang động, login hoặc click tự động.	Python / JS
Requests / aiohttp	Gửi HTTP request nhanh, nhẹ, thích hợp khi cần crawl API hoặc HTML tĩnh.	Python
Twint / Telethon	Công cụ crawl và truy xuất dữ liệu từ mạng xã hội như Twitter, Telegram thông qua API hoặc scraping.	Python

7. Khía cạnh pháp lý.

Bảng 7.1: Các quy tắc và nguyên tắc khi thu thập dữ liệu trực tuyến

Quy tắc	Ý nghĩa
robots.txt	Kiểm tra xem trang web có cho phép crawler truy cập hay không; tuân thủ hướng dẫn để tránh vi phạm điều khoản sử dụng.
Tôn trọng quyền riêng tư	Không thu thập hoặc lưu trữ thông tin cá nhân, dữ liệu định danh (PII) của người dùng.
Không tấn công hạ tầng	Giới hạn tốc độ truy cập (thường 1 request/giây) để tránh gây quá tải hoặc bị chặn IP.
Không lưu/chia sẻ dữ liệu nhạy cảm	Nếu phát hiện dữ liệu rò rỉ (leak data), chỉ được dùng cho mục đích cảnh báo hoặc nghiên cứu, không phổ biến lại.
Ghi rõ nguồn gốc	Khi trích dẫn hoặc báo cáo kết quả, cần nêu rõ nguồn gốc dữ liệu và thời điểm thu thập.

ỨNG DỤNG CRAWLER



Hình 0.1: Sơ đồ ứng dụng Crawler