[9장] 웹 크롤러 설계

가상 면접 사례로 배우는 대규모 시스템 설계 기초

이민석 / unchaptered

https://inblog.ai/monthly-cs
https://github.com/monthly-cs/2024-03-system-design-interview-1

웹 크롤러(Web Crawler)란 무엇인가?

웹 크롤러는 웹사이트를 탐색하며 데이터를 수집하는 도구이다.

웹사이트에 포함된 링크(URL)을 따라가면서 정보를 수집하고 분류하며, 이 특징은 다음과 같이 활용된다.

1. 검색 엔진 인덱싱(Search Engine Indexing)
2. 웹 아카이빙(Web Archiving)
3. 웹 마이닝(Web Mining)
4. 웹 모니터링(Web Monitoring)

검색 엔진 인덱싱(Search Engine Indexing)

- [예시]

  Google, Naver, Bing 등

- [정의]

  검색 엔진이 검색 전에 정보를 구성하여 쿼리에 대한 초고속 응답을 가능하게 하는 프로세스

- [프로세스]

  역색인은 텍스트를 저장하는 데이터베이스를 해당 텍스트 문서와 이를 가리키는 포인터와 함께 컴파일합니다.

  이후, 검색엔진은 토큰화를 통해 단어를 핵심 의미로 줄여 데이터를 저장하고 검색하는데 필요

  줄입니다.

https://www.lumar.io/learn/seo/indexability/search-engine-indexing/#:~:text=Indexing%20is%20the%20process%20by,engines%20to%20identify%20relevant%20information.

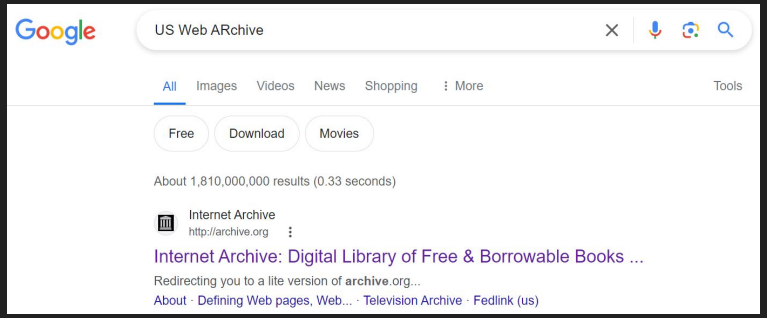
웹 아카이빙(Web Archiving)이란?

- [예시]

  US Library of Congress, EU Web Archive, Internet Archive

- [정의]

  나중에 사용할 목적으로 **장기보관**하기 위해서 웹에서 정보를 모으는 절차



https://www.google.com/search?q=US+Web+ARchive&oq=US+Web+ARchive&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIHCAEQABiABDIHCAIQABiABDIHCAMQABiABDIHCAQQABiABDIHCAUQABiABDIHCAYQABiABDIHCAcQLhiABDIHCAgQABiABDIHCAkQABiABNIBCDQwMDNqMGo5qAIAsAIB&sourceid=chrome&ie=UTF-8

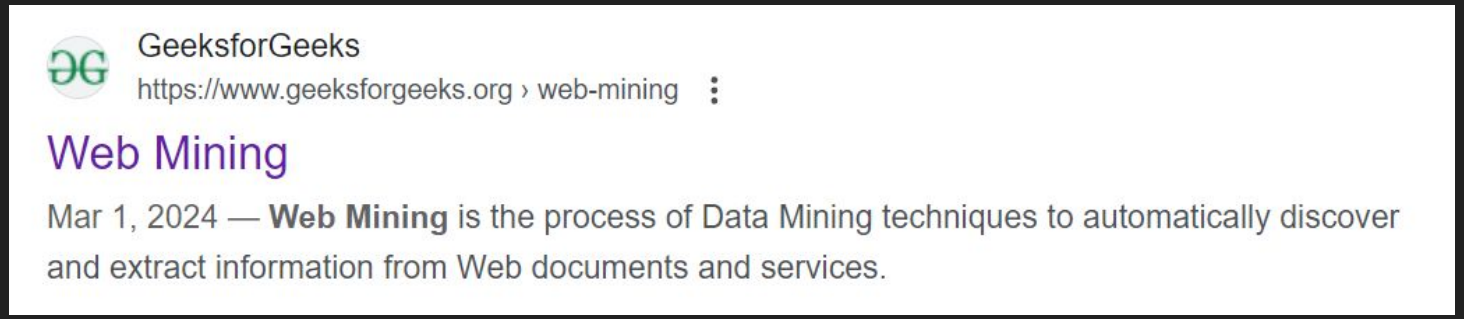
https://www.lumar.io/learn/seo/indexability/search-engine-indexing/#:~:text=Indexing%20is%20the%20process%20by,engines%20to%20identify%20relevant%20information.

웹 마이닝(Web Mining)이란?

- [정의]

  문서 및 서비스에서 정보를 **자동으로 검색하고 추출**하는 데이터 마이닝 기술 프로세스

  요구사항에 따라 관련 정보를 찾고 추출하기 위해 WWW(World Wide Web)에서 사용할 수 있는 시스템의 방대한 양의 데이터를 선별하는 가장 좋은 방법



https://www.geeksforgeeks.org/web-mining/

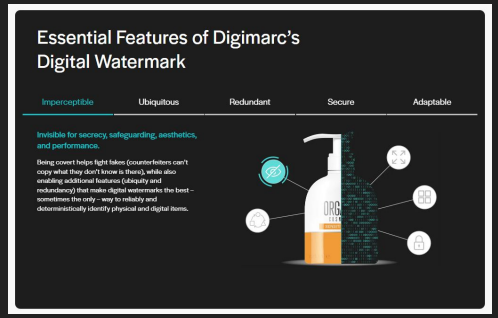
웹 모니터링(Web Monitoring)

- [정의]

  인터넷에서 저작권이나 상표권이 침해되는 사례 탐지

- [사례]

  Digimarc에서 다양한 웹 모니터링을 위한 서비스 및 기능을 지원



https://www.digimarc.com/product-digitization/data-carriers/digital-watermarks?utm_source=google+ads&utm_medium=search&utm_campaign=google+ads_search_Digital+Watermarks

결론

- 웹 크롤러(Web Cralwer)는 다음의 **기능이 반드시 포함**
  - **웹 사이트 탐색**
  - **웹 사이트 정보 저장**

요구사항 질의

- [목적]
  - 검색 엔진 인덱싱(Search Engine Indexing)
- [수량]
  - 1,000,000,000개 (10억개)
- [제한사항]
  - 새로 만들어진 웹 페이지, 수정용 웹 페이지 고려
  - 중복 콘텐츠를 가진 페이지는 무시
  - 수집한 웹 사이트는 5년 간 저장
- [추가 고려사항]
  - 규모 확장성 : 병행성(parallelism)을 활용
  - 안정성(robustness) : 잘못 작성된 HTML, 반응 없는 서버, 장애, 악성 코드가 붙은 링크
  - 예절(politeness) : 수집 대상 웹사이트에 적적량의 요청 전송
  - 확장성(extensibility) : 새로운 형태의 콘텐츠 지원이 쉬워야 함

요구사항 질의

[개요]

- 매달 10억 개의 페이지 다운로드
- 페이지당 평균 크기 = 500 KiB

[QPS]

- 기본 QPS = 1,000,000,000 / 30 / 24 / 3600 ≒ **385**
- 피크 QPS = 기본 QPS * 2 ≒ 385 * 2 ≒ 770

[용량]

- 월간 필요 용량 = 1,000,000,000 * 500 KiB = 500,000,000,000 KiB ≒ 500 TiB
- 연간 필요 용량 ≒ 500TiB * 12month * 5year ≒ 30,000TiB ≒ 30 PiB

감사합니다.