[Project Review] 데이터 크롤링의 세계

날짜 : 2020.10.21

진행 : 유상진 컨설턴트

인터넷에서 데이터를 수집하는 방법

- OpenAPI 등을 사용해 공개된 데이터를 얻는 방법(ex. 공공데이터포털, naver Developers)
- HTTP Get Method를 사용해 HTML을 가져오는 방법
- Selenium Web Driver 등을 사용해 사람이 하는 것과 유사한 자동화 방법
- 사람이 수작업으로 데이터를 수집하는 방법

HTTP Get Method

- 정보가 게시되어 있는 대상 웹사이트를 HTTP Get을 사용하여 html 코드를 얻고 Text Parsing 해서 사용
- Java / Go / Python / C/C++ / Perl / Node.js 등등 거의 대부분의 언어로 구현 가능
- 클라이언트 사이트나 SPA 등의 최근에 나온 웹사이트는 가져오기 힘듬

Selenium Web Driver

- 웹브라우저 인스턴스를 생성해 실행 시킨 후 해당 인스턴스를 컨트롤
- 웹사이트 테스트 자동화 목적으로 개발
- 가상의 브라우저를 실행시키는 Headless mode 등이 있음
- HTTP Get 방식에 비해 느리고 불안정적이나 보다 많은 웹사이트를 스크래핑 가능

크롤러와 스크래퍼

- 크롤러
  - 조직적, 자동화 된 방법으로 웹을 탐색 / 수집하는 프로그램
    - ex) 구글, 네이버 등의 검색엔진 결과 데이터를 수집하기 위한 봇(Bot)
- 스크래퍼
  - 웹 사이트에서 정보를 추출하는 프로그램
    - ex) 상품별 가격을 알기 위해 해당 상품을 파는 페이지들의 가격을 추출
- 크롤러 보다는 대부분 단순 스크래퍼 개발 수요가 많음
- 우리나라에서는 많은 기업들이 같은 의미로 혼용

구글은 어떻게 데이터를 수집하는가?

- 사용자가 검색하기 전, 수천억 개에 달하는 웹페이지에서 정보를 수집
- 수집한 정보를 바탕으로 검색 색인에 정리
- 과거 정보 수집으로 만들어진 웹URL 목록과 웹사이트 소유자가 제공한 사이트맵에서 대상 페이지 목록 수집 후 페이지의 내용 수집
- 사이트에 있는 링크를 사용하여 다른 페이지를 색인
- 해당 페이지를 수집하는 동안 새로운 사이트, 기존 사이트의 변경사항, 깨진 링크를 주의 깊게 확인

크롤링은 불법인가?

- 합법 / 불법 크롤링 구분은?
    - 홈디렉토리에 위치한 robots.txt 파일에 포괄적인 크롤링 금지 또는 특정 검색엔진의 크롤링 금지
    - 특정 디렉토리에 대한 크롤링 금지 등을 표시하였음에도 불구하고, 그 표시를 무시라고 크롤링을 하였다면 이는 사이트 운영자의 의사에 반한 크롤링에 해당함
    - 운영자는 robots.txt 외에 페이지 하단 약관등에 크롤링 금지를 표시할 수 도 있다

검색 로봇의 매너

- robots.txt 열어보고 서버의 로봇 배제 표준을 준수 할 것
- UserAgent를 속이지 않을 것

robots.txt의 로봇 배제 표준 규칙

- User-agent: * Disallow: / : 모든 검색엔지이 긁어가는 것 모두 막기
- User-agent: * Disallow: : 모든 허용하기
- User-agent: * Disallow: /cgi-bin/ Disallow: /tmp/ : cgi-bin, tmp 디렉토리 긁어가는 것만 막기
- User-agent: BadBot Disallow: / : 배드봇 검색로봇만 긁어가기 제외
- User-agent: WebCrawler Disallow: : 웹크롤러 검색로봇만 긁어가기 허락

크롤링 법적 분쟁 사례(국내외)

- 사람인 VS 잡코리아
    - 사람인에서 경쟁사인 잡코리아 홈페이지에서 채용정보를 크롤링하고 상업적을 활용
    - 크롤링을 이용해 확보한 콘텐츠를 자신의 영업에 무단 사용하는 것은 DB권 침해 행위, 불법
- 여기어때 VS 야놀자
    - 여기어때에서 경쟁사인 야놀자의 숙박업체 정보를 크롤링 & 활용
    - 경쟁관계에서 우위를 점하기 위해 상당 기간 크롤링 프로그램을 이용해 서버에 침입, 숙박 업소에 관한 각종 정보를 복제한 것은 불법
- HiQ Labs VS Linkedin
    - 로그인 필요없는 정보는 공공재, 크롤링 불법 아니다
    - 퍼블릭 정보를 크롤링 하는 것은 불법이 아니다
- 리그베다위키 VS 엔하위키(악용사례)
    - 엔하위키 미러는 리그베다위키를 지속적으로 크롤링해 새로운 콘텐츠가 올라오면 이를 자신의 사이트에 자동으로 반영하는 형태로 운영을 지속

    - 데이터베이스에 해당하는 원고 사이트를 제작하기 위하여 인적또는 물적으로 상당한 투자를 하였기 때문에 원고 사이트에 대한 데이터베이스 제작자에 해당한다고 봄이 타당

Web Abusing

- 데이터 크롤링을 응용해 데이터 수집이 아닌 다른 용도로 사용
- 검색엔지과 쇼핑몰 등은 Web Abusing과의 전쟁 중
- 일반 사용자와 Abuser를 판단하는데 많은 자원이 투입

User Agent

- 인간이 조작하는 웹 브라우저 형식
    - ex) Mozilla/5.0 (iPad; U; CPU OS 3_2_1 like Mac OS X; en-us) AppleWebKit /531.21.10 (KHTML, like Gecko) Mobile /7B405
- 자동화된 에이전트(봇)의 형식
    - ex) Googlebot /2.1 (+http://www.google.com/bot.html)
- User Agent의 종류
    - SK telecom, KT, LG U+ 별로 모바일 IP가 다름

정리

- 구글은 어떻게 데이터를 수집하는가?
- 데이터 크롤링 / 스크래핑
- OpenAPI / HTTP Get / Selenium Web Driver
- robots.txt
- 크롤링은 불법인가? - 국내 / 해외 판례 및 악용사례
- Web Abusing

자유 질문 소통 시간

- 상업용으로 사용하지 않으면 불법 아닌가요?
    - 서버에 트래픽 유발에 따른 부하에 대한 문제 제기하기 때문에 가져가지 말라고 할 것 같습니다.
    - robots.txt에 명시 되어 있는 것은 가져가면 안된다
- 페이스북은 태그 네임에 난수를 만들어 크롤링하기 힘들게 만듬
- Open API 서버에서 타 도메인 서버의 접근을 막아서 cors가 발생했을 때 어떻게 해결해야 할까요?
    - Open API인데 cors가 발생하는게 이상하기 때문에 확인이 필요할 것 같다
- 네이버 검색 순위를 크롤링을 해봤는데 태그가 보이는데 안나옵니다
    - 특정 아이피가 여러번 크롤링 및 접근을 하면 블록을 한다고 합니다.
- 웹 메크로를 개인적으로 만들어도 불법인가요?
    - 불법입니다.
- 합법적으로 스크래핑 하기 좋은 사이트가 있나요?
    - robots.txt가 별 내용 없는 사이트

Quiz

- 웹크롤러가 크롤링을 하기전에 서버에서 반드시 확인해야되는 파일 이름은?
    - robots.txt