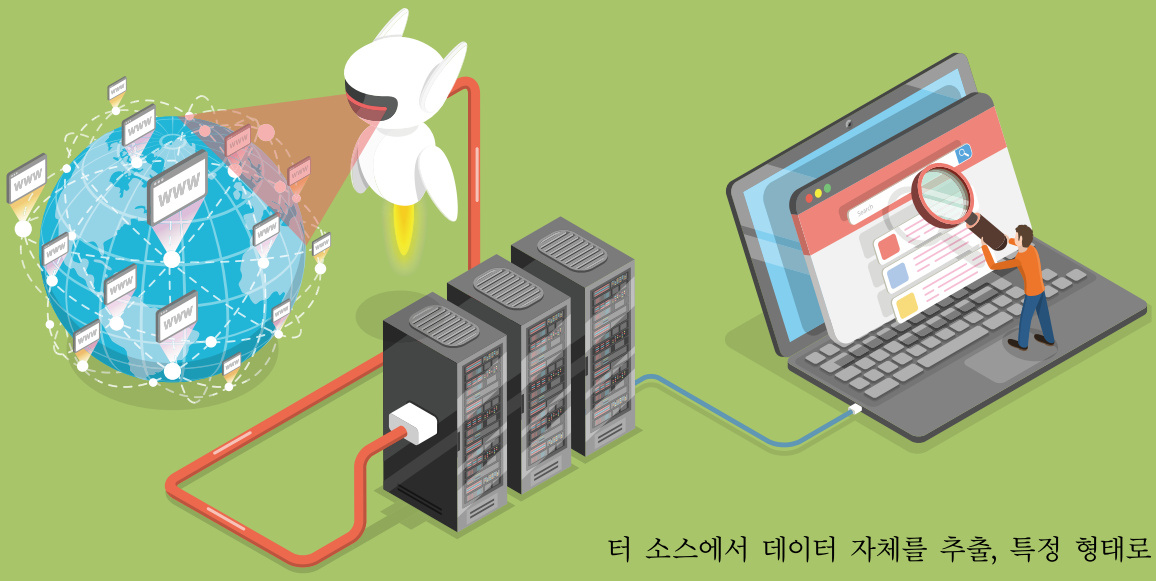
인터넷의 바다에서 정보를 긁어 들이는 데이터 노동, 크롤링

김원배 전자신문 부장, 정보통신용어표준화위원회(WORDSTD) 위원



크롤링(crawling)은 컴퓨터에 분산·저장되어 있는 다양한 정보 자원을 검색 엔진 로봇을 이용해 자동화된 방법으로 수집해 분류 및 저장하고 검색 대상의 색인으로 포함시키는 기술이다.

통상 인터넷 웹페이지를 수집해 분류하고 저장하는 것을 의미한다. 크롤링은 단순한 데이터 수집보다는 여러 웹페이지를 돌아다닌다는 뜻으로, 데이터가 저장된 위치에 대한 분류 작업이 크롤링의 주요 목적이라 할 수 있다. 크롤링의 결과물은 머신러닝 등 빅데이터 처리의 재료가 된다.

크롤링의 주요 대상은 다양한 형태로 존재하는 데이터다. 크롤링은 데이터를 대량 수집하는 기법으로, 웹 크롤링(web crawling) 또는 데이터 크롤링(data crawling)으로도 불린다.

유사 개념으로 소프트웨어(SW)를 통해 데이터 소스에서 데이터 자체를 추출, 특정 형태로 저장하는 스크레이핑(scraping)이 있다

크롤링을 수행하는 크롤러는 웹페이지를 돌아다니며 어떤 데이터가 어디에 있는지 색인(index)을 만들어 데이터베이스(DB)에 저장한다. HTML 페이지에서 관련 하이퍼링크를 찾아 데이터를 분류하고 저장하는 작업을 반복한다.

크롤링이 가장 많이 사용되는 분야는 검색 엔진이다. 인터넷 사용자가 검색 엔진을 통해 키워드를 검색하면 검색 엔진 로봇은 크롤러를 이용해 인터넷의 수많은 정보를 수집한 후 그 결과를 인터넷 사용자에게 노출한다.

구글 등 검색 엔진 로봇은 크롤링으로 데이터를 수집한 다음 이를 색인한 검색결과를 네티즌에게 보여준다.

웹사이트 보유자는 검색 엔진 로봇의 크롤러가 자신의 웹사이트 정보를 수집해 가는 것이 이득

이 되는 경우가 많을 것이다. 하지만, 후발 주자나 경쟁업체가 자신의 웹사이트 정보를 크롤링으로 수집해 이용하는 게 달갑지는 않을 것이다.

웹페이지 운영자가 수집하지 못하도록 조치한 데이터를 수집하거나, 수집한 데이터를 사용해 부당이득을 얻는 등의 행위를 할 경우엔 저작권법이나 부정경쟁방지법 등에 의해 제재를 받을 수 있다.

크롤링이 악용돼 정보를 무단 복제하게 되면 지식재산권 침해 문제가 발생할 수 있다. 크롤링으로 취득한 콘텐츠를 상업적으로 이용하는 것 또한 문제가 될 수 있다.

크롤링을 둘러싸고 경쟁업체 간 형사고소, 민사 크롤링 금지청구, 민사 손해배상 청구 등 저작권 침해 소송이 불거지는 사례도 적지 않다.

크롤링은 몇 가지 측면에서 문제가 되곤 한다. 첫째, 크롤링은 인터넷 서비스 사업자가 인터넷에 게시한 방대한 정보를 통째로 긁어가는 행위이기 때문에, 고객과 사업 정보에 대한 일종의 '침입'으로 간주되곤 한다.

또 한 가지 문제는 크롤러가 짧은 시간에 여러 번 접속을 시도함으로써 웹페이지의 서버를 과부하시킬 수 있다는 것이다. 실제로 웹페이지에 반복적으로 서버 과부하를 유도하는 것을 '디도스(DDos)'공격이라고 하는데, 크롤러가 하는 일과 크게 다르지 않다.

이에 페이스북, 네이버 등 크롤링의 대상이 되는 기업은 크롤러를 차단하기 위해 홈페이지 구조를 주기적으로 바꾸거나 페이지 호출 신호를 암호화하는 등의 방법을 사용하고 있다.

이 외에 웹페이지에 검색 로봇 배제 표준을 사용하거나 메타 태그를 사용해 크롤러로 검색 색인이 생성되는 것을 차단하는 방법도 동원한다.

그렇다고 크롤링 자체가 위법이라고 단정할 수 없다. 대부분 크롤링은 위법의 범주에 포함되지 않는다. 합법적 크롤링과 불법적 크롤링으로 구별하는 게 바람직하다. 합법적 크롤링은 웹사이트 운영자의 의사에 반하지 않은 크롤링을 의미하고, 불법적 크롤링은 웹사이트 운영자의 의사에 반하거나 또는 법률을 위반한 크롤링을 의미한다.

정보가 기업의 핵심가치라는 점을 고려하면 원하지 않은 크롤링에 대해 강력하게 대처하는 게 바람직하고 사후적인 대처보다 사전적인 예방이 바람직하다는 게 중론이다. 안티크롤링 기술 도입 등 기술적 조치뿐만 아니라 약관 등에 명확하게 크롤링이 금지됨을 적시하는 법률적 예방 조치도 필요하다.

크롤링을 계획하는 경우에는 데이터 수집 행위가 타인이 인적·물적 투자로 제작·관리하고 있는 데이터베이스 전부 혹은 상당한 부분을 복제하는 것은 아닌지 신중히 검토해야 한다.

저작권법은 데이터베이스 제작자의 권리를 명시적으로 보호하고 있다. 데이터 수집, 정리, 체계화, 관리 등에 투입된 상당한 비용과 노력 자체를 권리로서 보호하는 것이다.

또, 정보통신망 이용촉진 및 정보보호 등에 관한 법률은 '누구든지 정당한 접근권한 없이 또는 허용된 접근권한을 넘어 정보통신망에 침입하여서는 아니된다'고 규정하고 있음을 기억해야 한다. 접근권한의 부여는 서비스 제공자가 한다. 크롤러가 서비스 제공자로부터 사전에 접근권한을 부여받아야 한다는 의미다.

크롤링이 정보통신망법 위반에 해당되는 경우가 있을 수 있으며, 크롤링의 대상이 되는 프로그램 코드나 사진의 구도, 게시 방식 등이 저작권법의 보호 대상이라 크롤링이 저작권법 위반으로 이어질 수도 있다.