해외 개인정보보호 동향 보고서

최신동향 보고서 2019년 9월 4주

KISA 한국인터넷진흥원

웹 크롤링 관련 개인정보보호·저작권 침해 분쟁 사례 및 시사점 검토

< 목 차 >

1. 개요 및 배경

▶ 웹 크롤링(web crawling) 혹은 스크랩핑(web scraping)은 웹 페이지를 그대로 가져와 데이터를 추출해 내는 행위로서, 월드와이드웹을 탐색하는 웹 크롤러(web crawler)[1]를 활용하여 웹에 존재하는 콘텐츠를 자동으로 수집

- 웹 크롤링은 웹 페이지의 데이터를 자동으로 다운로드하여 여기에 포함된 하이퍼링크를 추출한 후 수행하는 작업으로, 다운로드 된 데이터는 쉽게 검색 가능하도록 색인 또는 데이터베이스에 저장
- 웹 스크랩핑은 웹 페이지의 데이터를 자동으로 다운로드하고 구체적인 정보를 추출하는 것으로, 추출된 정보는 데이터베이스와 파일 등 어느 곳에나 저장 가능
- 방대한 웹페이지의 정보를 모으기 위해서는 자동화된 크롤링 기술이 필요하지만[2], 이로

1 웹 크롤러(Web Crawler)의 유사 용어로는 앤트(ants), 자동 인덱서(automatic indexers), 봇(bots),웜(worms),웹 스파이더(web spider), 웹 로봇(webrobot) 등이 있음

2 검색 엔진, 링크 체크, HTML 소스코드 검증, 자동 이메일 수집 등 다양한 형태로 사용되어 웹 상에서 방대한 데이터 수집 작업을 자동으로 수행

인해 무분별한 개인정보 수집이나 저작권 침해 위험을 야기

- 공개 데이터(public data)에 대한 웹 크롤링과 데이터 스크랩은 합법적이지만, 특히 개인정보가 포함되고 정보주체의 동의 여부가 모호할 경우 개인정보침해 논란 발생 가능
  - 국내의 판례에서도 공개된 개인정보에 대한 영리적 목적의 수집 문제[3]와 고유의 목적에서 벗어나 영리를 목적으로 한 개인정보 수집의 문제[4] 등에 대해 언급
- 웹 사이트 운영자의 의사에 반하여 웹 크롤링이 이루어지는 경우 저작권 침해 등을 둘러싼 분쟁 발생 가능

▶ 웹 크롤링[5] 자체가 불법은 아니지만 특히 다음과 같은 경우 불법성을 다툴 여지가 있으며, 지난 몇 년 동안 웹 크롤링에 대한 비판이 제기된 것도 이러한 상황을 반영[6]

- 타인의 상당한 노력으로 만들어진 성과를 무단으로 이용해 이익을 취하는 등 불공정한 비즈니스 행위를 하는 경우
- 저작권 보호 원칙과 웹 사이트의 서비스약관(Terms of Service, ToS) 내용을 무시하고 크롤링 행위를 하는 경우
- 웹 크롤링 과정에서 과도한 트래픽을 야기해 웹 사이트에 예기치 않은 로드를 발생시키고, 크롤링 주체의 신원을 감추거나, 데이터 다운로드를 목적으로 웹 사이트의 보안 조치를 우회하는 등의 편법을 동원하는 경우
- 기타 사이트 운영자의 의사에 반하거나 실정법과 충돌하는 크롤링 행위는 불법으로 간주될 수 있으나[7] 구체적 사안에 따른 판단이 필요

2. 해외 분쟁 사례를 통해 본 주요 쟁점

(1) LinkedIn 對 HiQ Labs 사례

▶ 웹사이트 소유자가 명시적으로 웹 크롤링 중지를 요청한 경우에도 웹 크롤링이 해킹 행위로 간주되지는 않는 것으로 판결 [8] (2017년)

- 미국의 비즈니스 분석 스타트업 HiQ Labs는 Microsoft 소유의 비즈니스 중심 소셜

3 대법원 2014다235080 참조
4 서울고등법원 2013나49885 참조
5 웹 크롤링과 웹 스크랩핑은 엄밀히 구분하면 차이가 있으나, 이하의 내용에서는 일반적으로 통칭되는 웹 크롤링으로 묶어서 표현하기로 함
6 https://benbernardblog.com/web-scraping-and-crawling-are-perfectly-legal-right/
7 http://www.ddaily.co.kr/news/article/?no=151940
8 https://arstechnica.com/tech-policy/2017/08/court-rejects-linkedin-claim-that-unauthorized-scraping-is-hacking/

네트워크 서비스 LinkedIn의 웹 사이트에 공개된 개인 프로필 페이지를 크롤링한 후 새 직장을 찾는 직원들에 대한 보고서를 작성하여 고용주에게 판매

- LinkedIn은 이 같은 행위가 계속될 경우 해킹 방지를 위해 제정된 CFAA(Computer Fraud and Abuse Act) 위반이 될 수 있다며 HiQ Labs에 대해 웹 크롤링 중단을 요청하는 경고 공문을 발송
  - CFAA는 "승인 없이 컴퓨터에 액세스하거나 승인된 범위를 초과하여 액세스하는"[9] 것을 범죄로 간주
  - 따라서, LinkedIn이 HiQ Labs에게 △웹 크롤링 중단을 요구하는 공문을 발송하고 △robots.txt[10] 파일 게시 및 IP 차단 등의 기술적 조치를 취했음에도 불구하고 HiQ Labs가 계속 LinkedIn의 서버에 액세스하는 것은 위법이라는 주장
- HiQ Labs는 LinkedIn이 공개된 정보에 자유롭게 접근할 권리를 침해하고 웹 크롤링 금지를 통해 반경쟁적 행위를 했다며 맞고소
  - 일반적으로 모든 웹 크롤링이 사이트 소유자의 허가를 통해 이루어지지 않는 상황에서, 공개된 정보를 크롤링하는 일반적인 행위에 대해 CFAA를 적용한 LinkedIn의 해석이 과도하다는 지적도 제기
- 이에 대해 미 캘리포니아 법원은 비밀번호 없이 웹 사이트를 공개적으로 게시할 때 웹 사이트에 대한 액세스 권한을 암시적으로 부여한 것이며, HiQ Labs의 웹 크롤링을 해킹으로 간주할 수 없다고 판결
- 이와 함께, HiQ Labs가 크롤링을 통해 사용자 데이터를 수집하지 못하도록 IP 차단 등 기술적 수단을 사용하는 LinkedIn 행위 역시 정당한 것으로 판단
- 한편, 이 사례에서는 정보주체의 데이터 자체가 아닌 데이터 변경 사항에 대한 접근 권한이 쟁점이 된 점에 유의
  - LinkedIn 사이트에서는 사용자가 프로필을 공개하더라도 특정 변경 사항은 공유하지 않도록 설정할 수 있으나 HiQ Labs는 웹 크롤링을 통해 프로필 변경 사항을 탐지

(2) Field 對 Google 사례[11]

▶ 웹 크롤링을 통한 저작권 침해 행위를 방지하기 위해서는 웹 사이트 운영자가 적절한 조치를 취해야 하는 것으로 판결(2006년)

9 "access a computer without authorization or exceed authorized access"

10 로봇 배제 표준 (Robots exclusion standard, Robots.txt): 웹 사이트에 로봇 혹은 봇(bot)이 접근하는 것을 방지하기 위한 규약으로, 봇이 robots.txt 파일을 읽고 접근을 중지하도록 접근 제한에 대한 설명을 robots.txt에 기술함

11 https://www.lawinsociety.org/legal-perspectives-on-scraping-data-from-the-modern-web

- 미 네바다주의 Blake Field 변호사는 Google의 검색 엔진이 크롤링을 통해 자신의 개인 홈페이지에 게시된 내용을 무단으로 복제 및 배포하여 저작권을 침해했다며 Google을 상대로 소송을 제기
- Google은 웹 사이트 크롤링한 후 현재 페이지를 사용할 수 없는 경우에 대비해 백업으로 각 페이지의 캐시 버전을 만들었으며, 웹 페이지에 게시되어 있는 출판물 내용이 캐시되는 과정에서 저작권 침해가 발생했다는 것이 원고의 주장
- Google의 검색 사이트 색인을 위한 웹 크롤링 과정에는 로봇 혹은 봇(bot)이 이용되며, 웹 운영자는 robots.txt를 통해 봇의 접근에 대한 거부 의사를 밝히는 것이 가능
- 이 사례에서 원고는 Google의 사이트 색인 생성 메커니즘과 이를 방지하기 위한 robot.txt의 기능을 인지하고 있었으나, robot.txt를 사용하지 않기로 결정
- 법원은 이 같은 결정이 피고에게 사이트를 캐시하고 색인을 생성할 수 있는 묵시적 라이선스를 부여한 것이라고 판단했으며 원고는 패소

(3) Facebook 對 Power Ventures 사례

▶ 소셜 미디어 사이트에 공개된 개인 데이터를 제3자 사이트에서 수집하여 게시하는 것은 저작권 침해에 해당되는 것으로 판결(2009년)

- 비즈니스 정보 제공업체 Power Ventures는 Facebook에 게시된 사용자 프로필 데이터를 크롤링하여 통합 소셜 미디어 계정을 생성하는 서비스를 운영했으며, Facebook은 Power Ventures의 데이터 추출이 자사의 의사에 반하여 이루어졌다며 소송을 제기
- Facebook은 Power Ventures가 사용자 데이터 추출 과정에서 Facebook 플랫폼 내 웹 페이지의 사본을 생성했으며, 이를 통해 직접적 및 간접적으로 저작권 침해가 이루어졌다고 주장
- 반면 Power Ventures는 Facebook이 해당 사용자 데이터의 보호 및 관리 담당자가 아니므로 저작권을 주장할 수 없다고 반박
- 이에 대해, 법원은 △Power Ventures의 웹 크롤링 도구들이 사용자 데이터 추출에 앞서 해당 웹 페이지의 HTML 전체에 대한 캐시 작업을 수행하고 △HTML에는 Facebook 사이트의 구조에 대한 정보가 포함되므로, 결과적으로 Facebook의 저작권 침해가 이루어졌다고 판결

3. 시사점

▶ 개인정보보호 관점에서, 수집 대상 데이터에 개인식별정보가 포함되는 경우 다음과 같은

예방 조치를 취하는 것이 추천됨

- 해당 웹 사이트에서 정보주체가 제3자의 데이터 수집에 대한 동의를 제공했는지 여부에 대해 확인하고, 유효한 동의가 이루어진 경우에만 크롤링 진행
- 소셜 미디어 사이트 혹은 개인이 운영하는 웹 사이트에서 데이터 공유 및 수정을 금지하는 내용이 제시된 경우[12] 크롤링 중단
- 정보주체가 데이터 수집에 동의한 경우라도, 데이터 수집을 위한 정당한 사유가 없는 경우 크롤링이 불법일 수 있다는 점[13]에 유의

Reference

1. Ars Technica, Court rejects LinkedIn claim that unauthorized scraping is hacking, 2017. 8. 15.
2. Benoit Bernard, Web Scraping and Crawling Are Perfectly Legal, Right?, 2017.4.18
3. DataHen, Awareness Faster, 2018.5.31
4. InfoQ, LinkedIn Ordered to Allow Scraping of Public Profile Data, 2017.8.29.
5. Law In Society, LEGAL PERSPECTIVES ON SCRAPING DATA FROM THE MODERN WEB, 2019.9월 20일 접속
6. TechCrunch, LinkedIn sues anonymous data scrapers, 2016.8.15.
7. The Verge,Google fixes search issue that prevented new content from appearing, 2019.8.9

12 예컨대 웹사이트 운영자가 웹서버의 홈디렉토리에 위치한 robots.txt 파일에 포괄적인 크롤링 금지 또는 특정 검색엔진의 크롤링 금지, 특정 디렉토리에 대한 크롤링 금지 등을 표시하거나 메인페이지의 하단, 약관 등에 크롤링 금지를 표시였음에도 불구하고, 크롤링을 강행하는 경우 사이트 운영자의 의사에 반한 불법 크롤링에 해당

13 https://www.datahen.com/gdpr-data-crawling/

KISA 한국인터넷진흥원

발 행 일 2019년 9월
발 행 및 편 집 한국인터넷진흥원 개인정보보호본부 개인정보정책기획팀
주 소 전라남도 나주시 진흥길 9 빛가람동 (301-2) Tel 1544-5118