이학석사학위 논문

웹 크롤링(Web Crawling)에 관한 연구
– 그 원리와 법적 책임에 관하여

A Study on Web Crawling
– Focused on its working mechanism and legal responsibility

2023년 2월

서울대학교 융합과학기술대학원

수리정보과학과 디지털포렌식학 전공

김 태 균

웹 크롤링(Web Crawling)에 관한 연구

- 그 원리와 법적 책임에 관하여 -

지도교수 이 상 원

이 논문을 이학석사 학위논문으로 제출함

2022년 12월

서울대학교 융합과학기술대학원
수리정보과학과 디지털포렌식학 전공

김 태 균

김태균의 석사 학위논문을 인준함

2023년 1월

위 원 장 국 웅 (인)

부위원장 이 상 원 (인)

위 원 엄 현 상 (인)

[국문초록]

웹 크롤링(Web Crawling)에 관한 연구
- 그 원리와 법적 책임에 관하여

웹 크롤링은 크롤러 또는 스파이더라는 프로그램을 사용하여 웹 데이터를 수집하는 방법을 의미한다. 크롤러의 기본적 원리는 주어진 시드 URL에서 출발하여 그 URL과 연결된 웹 페이지를 다운로드하고, 여기에 포함된 하이퍼링크를 추출하고, 이러한 하이퍼링크로 식별되는 웹 페이지를 재귀적으로 계속 다운로드하는 것이다. 웹 크롤링은 이제 데이터가 핵심 요소가 되는 모든 곳에서 데이터를 수집하는 가장 효과적이고 유용한 방법 중 하나로 널리 사용되고 있다. 특히 빅데이터 또는 인공지능의 등장으로 인하여, 마케팅 또는 비즈니스 전략에 있어 경영 판단 또는 의사 결정 과정에서 웹 크롤링은 이제 필수 불가결한 것이 되었다.

웹 크롤링이 점점 중요해지고 있지만 이를 사용하는 것의 법적 책임에 대한 연구는 거의 없었다. 이 논문은 웹 크롤러의 작동 메커니즘과 그것을 사용한 행위의 법적 책임을 중심으로 검토한다. 최근 대법원은 피고인들이 숙박정보제공업체의 직원이 경쟁업체의 모바일 애플리케이션 서버에 접속해 자신의 크롤링 프로그램을 통해 숙박시설 목록 등 데이터베이스를 복사한 사건에서 피고인들을 무죄로 판단한 바 있다. 그 판결은 ①서비스 제공자가 네트워크에 대한 접근권한을 제한하는지 여부는 보호조치나 이용약관 등의 대상에 의하여 결정되어야 하며, ②데이터베이스의 상당 부분은 양과 질 모두를 기초로 판단하여야 한다고 하였다. 이 같은 법리에 기초하여 정보통신망침입, 데이터베이스 도용으로 인한 저작권법위반, 업무방해 혐의 모두에 대해서는 무죄를 선고하였다. 이것은 크롤링을 통한 데이터 수집에 대한 최초의 대법원 판결이다.

이 글에서는 위 대법원 판결을 기초로 3가지 측면 즉 정보통신망침입, 데이터베이스 도용으로 인한 저작권법위반, 형법상 업무방해에 대해 구체적으로 검토하고, 그 외 부정경쟁방지법, 개인정보보호법 기타 경쟁법적 측면에서도 검토한다. 그 결론은 다음과 같다. 웹 크롤러 사용의 법적 책임은 정보통신망 접근 범위, 데이터베이스제작자 권리 침해 여부 그리고 장애업무방해 여부를 기준으로 판단되어야 하는데, 그와 같은 법률적 평가는 웹 크롤러가 사용되는 상황, 사용자의 의도 그리고 사용으로 인하여 발생한 결과와 같은 사정을 기반으로 하여야 한다.

주요어: 웹 크롤링, 웹 크롤러, 데이터, 데이터 수집, 정보통신망 침입, 데이터베이스제작자의 권리침해, 업무방해, 부정경쟁 행위, 불공정거래행위, 시장지배적지위 남용

학번: 2021-24496

<차례>

<표 차례>

<그림 차례>

Ⅰ. 서론

1. 문제의식

공학은 최적의 것(the optimum)의 추구하고, 법학은 가장 정당한 것(the just)를 추구한다. 최적의 것은 사실적 개념으로 최소의 시간과 최소의 자원을 투입하여 최대의 목표를 달성하는 어떤 것이다. 시간과 자원이라는 비용의 한계값, 즉 한계비용이 체증하기 시작하는 어느 지점에서 최적의 것을 찾을 수 있다. 최적의 것은 효율성과 그 의미가 상당 부분 겹쳐진다. 반면 가장 정당한 것은 원칙적으로 효율성과는 무관하다. 정당하다는 것은 개념상 가치적이고 평가적인 것이어서 윤리, 가치, 법규와 같은 규범에 근거한다. 주어진 규범에 비추어 그 행위가 허용될 수 있는지가 정당성의 문제가 된다. 최적의 것과 가장 정당한 것은 그러한 개념상 차이와 판단 기준의 상이성으로 인하여 그것들의 발견과정에서 서로 영향을 미치거나 만나는 일은 거의 없다.

이 글은 최적의 것과 가장 정당한 것의 접점을 찾고자 하는 시도이다. 이론상으로는 서로 교섭할 일이 없다고 하여도 현실에서 사람들은 항상 최적의 것을 통하여 활동하려고 하고, 사람들의 모든 활동은 일차적으로 규범적 평가의 대상이 될 수 있기 때문이다.[1] 그 대상으로서 웹 크롤링(Web Crawling)을 선택하였다. 공학적 입장에서 웹 크롤링은 최소의 비용, 최소한 한계비용이 증가하지 않는 한도에서 최선

1) 해당 규범이 적용될 수 없다는 것도 해당 규범에 관한 판단이다. 가령 윤리의 문제는 법적 평가 대상이 아니라고 하였을 때, 그 판단은 이미 법적 판단을 거쳐서 나온 것이므로, 여기서 해당 규범이 적용 여부는 중요하지 않다.

의 결과를 가져올 수 있도록 설계되어야 한다. 웹 크롤러의 작동 원리와 좋은 웹 크롤러의 요건을 최적의 것을 찾는 공학자의 입장에서 검토하였다. 그 후 그 최적의 웹 크롤러를 대상으로 하여 그것을 사용하는 것을 정당화할 수 있는지와 정당화할 수 있는 요건은 무엇인지를 법학자의 입장에서 검토하였다.

웹 크롤링을 주제로 선택한 나의 문제의식은 이런 것이었다. 사람이 일일이 손으로 하는 일이 법적으로 허용되는 것이라면 그것을 컴퓨터 프로그램과 같은 자동화 도구를 사용하는 것도 마찬가지로 허용되어야 하는 것이 아닐까? 만약 손으로 하는 일은 허용되지만 그것을 자동화 도구를 사용하여 하는 것은 무제한 허용될 수 없다고 한다면, 양자의 사이에 무슨 차이가 있는 것인가? 자동화 도구는 최적의 것을 찾으려는 노력의 결과인데, 여기에 대해서 제한을 가하는 이유는 무엇인가? 손으로 하는 일과 자동화 도구를 사용하는 일 사이에 법적 평가를 달리해야만 하는 필연적인 차이는 무엇인가? 그 차이는 양적인 것인가 질적인 것인가? 자동화에도 여러 단계가 있을 터인데, 과연 자동화의 어느 지점에서 법적 허용 여부가 갈리는 것인가? 정보통신기술의 발달과 보급으로 인해 인터넷의 사용뿐만 아니라 업무와 생활의 다방면에서 자동화 도구의 사용이 증가되고 일반화될 터인데, 문제가 될 때마다 매번 그와 같은 도구의 사용이 정당화될 수 있는지를 따져 물어야 하는가? 자동화 도구 사용의 일반적인 허용 기준을 설정할 수는 없는 것인가?

나는 이와 같은 문제의식에 가장 적합한 주제가 웹 크롤링이라고 보았다. 인터넷이나 웹에는 무수한 정보들이 넘쳐나는데, 사람이 손으

로 일일이 웹 사이트를 돌아다니며 정보를 수집하는 행위 자체에 대해서는 (타인의 비밀을 침해한다든지 하는 별도의 수단을 사용하지 않는 이상) 아무런 법적 제한이 없다. 반면 이미 보편화되어 사용되고 있는 웹 크롤링 기술을 사용한 정보의 수집에는 일정한 법적 책임을 지우려는 시도들이 계속되고 있다. 손으로 웹 사이트를 둘러보는 것과 웹 크롤링을 사용하는 것의 차이가 무엇인지 명시적으로 밝힌 연구도 찾을 수 없었다. 최근에 선고된 이른바 '야놀자 판결'[2]은 위와 같은 호기심에 내가 충분히 수긍할 수 있는 해결책을 제시하였다. 반면 '야놀자 판결'의 관련자들을 당사자로 하는 민사소송[3]에서는 내가 쉽사리 수긍할 수 없는 정반대의 결과가 나왔다. 이처럼 상반된 법원의 판단이 웹 크롤링을 이 글의 주제로 정한 직접적인 계기이다.

2. 연구의 내용과 방법

이 글에서 내가 다루는 내용은 다음과 같다. **「Ⅱ. 웹 크롤링 (Web Crawling)」**에서는 최적의 것을 찾는 공학자의 입장에서 웹 크롤링 일반에 대해서 다룬다. 주된 초점은 웹 크롤링의 기술적 원리와 성능 좋은 웹 크롤러를 위한 크롤링 정책(crawling policies)에 관한 것이다. 여기까지는 웹 크롤링을 통한 정보[4]를 수집하려는 사람의 입장이다. 여기에 더하여 웹 크롤링을 통한 정보 유출을 막으려는 사람의 입장에서 웹 크롤링 방지 기술도 검토한다. **「Ⅲ. 웹 크롤링 사용**

2) 대법원 2022. 5. 12. 선고 2021도1533 판결
3) 서울고등법원 2022. 8. 25. 선고 2021나2034740 판결
4) 웹 크롤링은 '웹 페이지에 실려 있는 내용'을 대상으로 한다. 이 글에서 웹 크롤링의 대상이 되는 것을 문맥에 따라 콘텐츠, 정보, 데이터 등으로 부를 것이나, 어느 것이나 그 의미의 대차는 없다.

의 형사법적 책임」에서는 웹 크롤링을 사용하여 정보를 수집하는 행위의 형사법적 책임을 검토한다. 검토의 대상이 되는 구성요건은 3가지이다. 정보통신망 이용촉진 및 정보보호 등에 관한 법률상 정보통신망 침입, 저작권법상 데이터베이스제작자의 권리 침해, 형법상 컴퓨터 등장애업무방해가 그것이다. 여기서는 법원의 판결례는 물론 필요에 따라 외국(대부분의 경우 미국)의 입법례와 사례도 함께 볼 것이다. 법원의 판결례에서 가장 중요한 것은 앞서 말한 '야놀자 판결'이 될 것이다. **「Ⅳ. 웹 크롤링 사용의 기타 법적 책임」**에서는 웹 크롤링을 통한 정보수집행위의 형사적 책임 이외의 법적 책임을 검토한다. 주로 부정경쟁방지 및 영업비밀보호에 관한 법률상 부정경쟁행위가 되는지와 독점규제 및 공정거래에 관한 법률상 시장지배적 사업자의 남용행위 또는 불공정거래행위에 해당하는지가 주된 논의의 대상이다. 여기서는 앞서 말한 '야놀자 판결'과 정반대의 결론을 제시한 민사판결과 데이터의 부정사용을 새로 규정한 개정 부정경쟁방지 및 영업비밀보호에 관한 법률의 내용도 다뤄질 것이다. **「Ⅴ. 결론」**은 마무리와 전망이다.

이 글이 웹 크롤링의 기술적 원리라는 공학적 입장에서 출발을 하지만 종국적으로 지향하는 것은 웹 크롤링의 사용을 정당화할 수 있는 상황과 요건을 확인하려는 것이다. 웹 크롤링은 법률에 규정된 개념이 아니다. 웹 크롤링 자체가 역사가 오래되었고, 여러 방면에서 각각의 목적에 맞게 설계되어 사용되고 있다. 단일한 웹 크롤링이나 웹 크롤러의 모습을 찾는 것은 불가능하다. 그래서 웹 크롤러 사용의 법적 책임을 따지는 것은 웹 크롤러의 개념 즉 원리와 기능만으로 해결되지 않는다. 웹 크롤러 사용의 전후 맥락, 사용자의 의도, 크롤링의

대상이 된 정보의 성격, 웹 크롤링의 결과 등 충분한 정황과 간접사실들이 주어져야만 법적 책임을 판단할 수 있다. 그것이 이 글에서 판결례들을 소개할 때 사실관계를 비교적 소상히 소개하는 이유이다.

이 글은 기본적으로 다양한 문헌을 바탕으로 관련 내용을 추려내고, 거기서 결론을 이끌어내려고 시도한다. 거듭 말하거니와 웹 크롤링에 대한 일의적이고 통일된 개념을 끌어낼 수 없는 상황에서 첨단의 웹 크롤링 기술을 하나하나 추적하는 것은 불가능할 뿐만 아니라 무의미하다. 이 글에서는 웹 크롤링에 대해서 최대한 공통적인 것을 추려낼 수 있는 관련 분야의 대표적 문헌을 기초로 하였다. 법률적 문제에 대해서도 관련 문헌을 기초로 하였는데, 웹 크롤링이 이미 폭넓게 사용되고 있음에 반하여, 법률의 영역에서 관심을 갖고 이를 포섭하려는 시도는 데이터에 대한 논의가 충분히 심화된 비교적 최근의 일이다. 법학 중 데이터와 관련 기술을 적극적으로 그 논의의 대상으로 삼으로는 경쟁법학계에서도 그러한 시도는 길게 잡아도 10년을 넘기 어렵다.[5] 그런 만큼 법률적 검토에 인용하고 있는 문헌의 상당수는 웹 크롤링을 관련 법률에 직접 적용한 것은 아니고, 관련 분야의 일반적인 내용을 소개한 것들이다. 그것을 바탕으로 웹 크롤링에 적용했을 때의 모습을 그리는 것이 이 글이 만들어내고자 하는 결과이다.

3. 일러두기

5) 빅데이터 또는 데이터의 수집에서 발생하는 문제들에 대해서 소비자 후생(consumer's welfare)의 관점에서 위법성 판단을 해야 한다는 논의가 그 시초라고 보이는데, 그 중 대표적이자 가장 모범적인 것은 다음과 같다. 강정희, "빅데이터를 기반으로 하는 배제남용 행위의 위법성 판단기준 연구 - 소비자 선택 기준의 적용을 중심으로", 서강대학교 박사학위 논문, 2015년

가. 이 글의 화자로서 '나'

이미 앞서서 나는 이 글의 화자를 '나'라고 표현하였다. 학술적인 글에서는 이야기를 풀어나가는 화자를 '필자', '저자', '작성자' 등으로 표현한다. 그 소이(所以)는 그 글의 객관성을 확보하고자 하는 시도라고 나는 생각한다. 상술하면 이렇다. 그리 멀지 않은 과거의 어느 시점부터 학문에서 확실성(certainty)을 미덕으로 생각하게 되었고, 확실성은 객관성의 확보로서 이루어진다고 생각했다. 수학을 기본 언어로 하는 자연과학에서는 그와 같은 객관성과 확실성의 확보가 상대적으로 수월하였고, 그런 연유로 사회과학[6]을 업으로 하는 사람들 사이에서는 불안감을 느끼기 시작했다. 그들은 자연과학에 대한 사회과학의 상대적 약세를 '학문의 위기'[7]라고 생각하였다. 그 위기를 타도하기 위한 방법으로 사회과학에서도 객관성을 담보하려는 시도가 계속되었는데,[8] 그 중 하나가 화자로서 '나'를 숨기는 것이었다. 수치와 계산으로 표현되는 자연과학에 비하여 주관적인 '나'가 이야기하는 사회과학은 덜 객관적으로 보였던 모양이다. 그러나 나는 '학술적'이자 자연과학적 성격뿐만 아니라 '사회과학'적 성격도 다분한 이 글에서 화자로서 '나'를 적극적으로 들어내는 데에 아무런 주저함이 없다.

---

6) 우리식 용어법으로는 정확히 '인문·사회과학'이라고 할 것이나 여기서는 단순히 '사회과학'이라고만 한다. 어차피 여기서는 자연과학에 대비되는 학문분과를 지칭하는 것이므로 사회과학이라고만 하여도 그 의미가 충분히 전달되기 때문이다.

7) 예컨대, 이에 대한 대표적인 것은 에드문트 후설(Edmund Husserl)의 저서 '유럽학문의 위기와 선험적 현상학(Die Krisis der europaischen Wissenschaften und die transzendentale Phänomenologie)'을 들 수 있다.

8) 사회과학으 분야에서 그러한 시도에 가장 성공한 것은 심리학이라고 생각한다. 오늘날 심리학, 특히 미국이나 기타 영미권에서 나오는 심리학 문헌은 심리학의 창시자라고 일컬어지는 프로이트의 그것과는 전혀 성격이 다른 그 무엇으로 보인다. 프로이트의 문헌이 다분히 주관적이고, 비논리적이지만 그 가운데에서 수긍할 수 있는 그 무엇을 이야기했다면, 오늘날 (미국의) 심리학은 실험과 통계 외에는 아무것도 찾을 수 없기 때문이다.

그 이유는 이렇다. 자연과학이든 사회과학이든 모든 학문 분야에서 연구자로부터 벗어나 완전히 객관적인 것은 없다. 무릇 학문은 연구 주제의 선정으로부터 연구자의 주관성을 벗어날 수 없기 때문이다. 수치와 계산으로 표현한다든가 '나' 대신 '필자'라고 쓴다고 해서 객관성을 담보할 수 있다는 생각 자체가 유치한 것이다. 좀 더 근본적인 이유를 보자면, '인식이 세계'이고, '인식은 존재에 앞서기' 때문이다. 존재하는 물자체(物自體, Das Ding an sich)에 대한 그 어떤 것의 개입도 없는 직접적인 인식은 불가능하므로 세계에 대한 객관적이고 온전한 인식은 애초에 불가능하다. 인간의 인식은 감각에서 비롯되므로 인식의 한계는 감각의 한계만큼이나 명확하다. 신과 같은 자리에 위치하지 않는 이상 세계는 그것을 인식한 수단에 종속되어 표현된다. 그러니 세계에 대한 물리적·수학적 인식이라고 하더라도 그것은 법률적 인식만큼이나 주관적이고 오류의 위험을 내포하게 된다. 그 어느 경우에도 세계에 대한 객관성을 확보하지 못한다면, 차라리 그 세계를 인식하고 해석하는 주체를 명시적으로 표현하는 것이 온당한 태도라고 생각한다. '나'를 적극적으로 드러냄으로써 '나'의 기본 생각들과 내가 처한 입장이 무엇인지를 밝히고, 독자는 '나'의 입장을 감안하여 글을 읽는 비판적인 작업이 더 용이해질 것이기 때문이다.

나. 인더스트리4.0(Industrie4.0)

내가 웹 크롤링을 주제로 이 글을 쓰게 된 계기 중 하나로, 각종 정보통신기술과 빅데이터 또는 인공지능의 부상 등을 상징하는 소위 '4차 산업혁명'적인 요사이 분위기였음은 부인하지 못한다. 그러나 나

는 ‘4차 산업혁명’이라는 용어법에 동의하지 않는데, 그 이유는 이렇다. 첫째, 국내에는 위 용어를 클라우스 슈밥이 2016년 다보스 포럼에서 최초로 사용한 것으로 알려져 있으나, 이는 정확하지 않다. 위 용어는 그 전에 이미 독일에서 국가경쟁력 발전을 위한 사업으로 진행하여 왔던 것을 ‘인더스트리 4.0(Industrie 4.0)’으로 불렀던 것을 클라우스 슈밥이 다보스 포럼에서 차용하였다고 봄이 사실에 더 가까운 관찰이다. 둘째, 위 용어는 바른 개념이라고 할 수 없다. 인터스트리 4.0은 정보기술, 인공지능, 빅데이터 등의 등장에 적응하여 국가경쟁력을 살리자는 의미에 더하여 그로 인하여 노동자들의 처지가 열악해지는 것을 인식하고 그것을 막으려는 노력을 포함하는 것이다. 그러나 위 용어가 현재 사용되고 있는 용례를 보면 그와 같은 노동자 및 사회적 약자를 보호하자는 의미를 거의 상실하였다. 셋째, 위 용어는 사실과 가치를 의도적으로 혼동하였다. 새로운 기술의 도입과 그로 인한 사회 변화는 사실(Sein) 그 자체일 뿐 그것이 바로 당위(Sollen)나 가치가 될 수 없다.[9] 그와 같은 상황에서 우리는 얼마든지 새로운 가치를 발견하고 약자를 보호하면서 사회를 구성하려는 노력을 할 수 있으며, 그것이야말로 우리가 마땅히 해야 하는 당위이자 가치이다. 그런데 위 용어는 정보기술의 발전으로 인한 사회의 변화를 마치 당연한 가치로 받아들이면서 그에 저항하고 그 속도를 늦추려는 노력 또는 거기서 오는 부작용을 제거하거나 최소화하려는 노력은 시대착오적인 반동적이라는 인식을 저변에 깔고 있다.[10] 넷째, 위 용어는 현재의 상황이 네 번째 산업혁명이라는 것인데, 도대체 2차, 3차 산업혁명의 시작과 끝은 어디였고, 그 영향은 무엇이었는지에 대한 사회적으로

9) 최소한 칸트의 도덕철학이나 법철학에서는 사실에서 당위가 도출될 수 없다.

10) 독일의 인더스트리 4.0 중 노동 분야에서 소외되는 노동자 또는 노동을 위한 독일의 논의에 관해서는 김경래, “독일 Industrie 4.0의 특징: 노동 4.0을 중심으로”, 한독사회과학논총 제28권 제2호, 2018년, 3-26쪽 참조

나 학문적으로나 합의된 인식을 찾을 수 없는 상황에서 ‘4차’를 논한다는 것 자체가 어불성설이다. 그래서 나는 (굳이 표현해야 한다면) 위 용어에 대신하여 나는 ‘인더스트리 4.0’을 사용한다.[11]

11) 보다 온당한 용어법으로 ‘정보화 사회’정도가 무난하다고 생각한다. 그러나 ‘정보화 사회’만으로는 작금의 상황을 보다 선명하게 전달하는 데에 한계가 있다고 생각하여 ‘인더스트리4.0’이라고 쓴다. 만약 ‘4차 산업혁명’이라는 용어가 사멸하는 때가 오면 나도 ‘인더스트리4.0’이라는 용어를 폐기할 것이다.

Ⅱ. 웹 크롤링(Web Crawling)[12]

1. 도입

웹에서 데이터 수집은 선정, 수집, 정리의 3단계를 거친다. 선정은 원하는 데이터가 있는 위치를 확인하는 것이다. 수집은 URL을 입력하여 대상 웹 페이지를 열고 데이터를 추출하는 것이다. 정리는 추출한 데이터를 기록하여 정리하는 것이다. 이 과정을 초기 URL을 지정해주면 자동을 해주는 것이 웹 크롤러이다. 종래부터 웹 크롤러는 존재하였는데 비교적 최근 이를 활용한 데이터 수집이 문제가 되는 것은 데이터가 중요한 기업의 의사 결정 수단이 되거나 그 자체로 경제적 가치를 지니게 되었기 때문이다.

초기 웹 크롤링은 대부분의 경우 인터넷에서 검색 서비스를 위하여 활용되었다. 이것이 확대되어 가격 또는 상품 비교, 주소추출, 소셜미디어 모니터링, 반복 업무 자동화 등 다방면에 활용되고 있다.[13] 대부분의 웹 사이트 소유자는 검색 엔진에서 강력한 존재를 나타내기 위해 웹 크롤러가 가능한 광범위하게 페이지를 인덱싱하기를 원하지만, 반면 크롤링을 통하여 의도하지 않게 검색 엔진이 인덱싱 하지 않

12) 이 부분은 특별한 언급이 없는 이상 다음의 자료들에 기초하여 정리한 것이다. Marc Najork, “Web Crawler Architecture”(https://marc.najork.org/pdfs/eds2009a.pdf 2022. 8. 5. 방문); Christopher Olston, Marc Najork, 『Web Crawling』, Foundations and Trends® in Information Retrieval, Vol. 4, No. 3, 2010 (http://infolab.stanford.edu/~olston/publications/crawling_survey.pdf 2022. 8. 5. 방문); Md. abu Kausar, V.S. Dhaka, Sanjeev Kumar Singh, “Web Crawler: A Review”, 2013 (https://research.ijcaonline.org/volume63/number2/pxc3885125.pdf: 2022. 11. 1. 방문)

13) 이주호, “크롤링에 의한 민감 정보 침해에 대응하는 로그인 강화 연구”, 숭실대학교 석사학위논문, 2020년, 6-8쪽

아야 하는 정보를 인덱싱하는 경우에는 데이터 유출로 이어질 수도 이는 위험도 있다.

이하에서는 웹 크롤링의 작동 원리와 효율적 웹 크롤러 구성을 위한 크롤링 정책(Crawling policies) 및 웹 크롤링 방지 기술 등을 중심으로 살펴본다.

2. 웹 크롤링의 개요

가. 웹 크롤링과 콘텐츠 수집

웹 크롤링은 웹 페이지를 효율적·자동적으로 다운로드 받는 프로세스이다. 하나 이상의 시드(seed) URL에서 그와 관련된 웹 페이지를 다운로드 하고, 여기에 포함된 하이퍼링크를 추출한 후 하이퍼링크를 따라 웹 페이지를 재귀적으로 계속 다운로드한다. 출발점이 되는 URL을 주면 웹 크롤러는 그 URL상의 모든 웹 페이지를 다운로드 하는데, 그 웹 페이지에서 다른 웹 페이지를 연결해 놓은 하이퍼링크들을 다시 추출하여 하이퍼링크로 연결된 웹 페이지들을 다시 다운로드하는 것이다. 웹 크롤러는 그 과정을 반복 수행하면서 웹 페이지에 있는 콘텐츠를 수집한다.

웹 크롤러는 검색 엔진 또는 웹 페이지에 있는 콘텐츠나 데이터의 수집에 광범위하게 사용된다. 웹 크롤러는 World Wide Web의 시작과 거의 동시에 등장하였다. 일반적으로 Mattew Gray가 1993년 통계 목적으로 작성하여 사용했던 것이 최초의 웹 크롤링 프로그램으로 알려

져 있다. 웹 크롤러는 검색 엔진의 중요한 구성 요소로 기능하는데, 검색 엔진에서 크롤러는 웹 페이지의 말뭉치(corpus)를 수집하고, 이를 인덱싱 하는 데 사용된다.[14] 또한 웹 데이터 마이닝, 가격 비교와 같이 다수의 웹 페이지의 정보를 처리하는 응용 프로그램에도 사용된다. 예컨대 구글의 뉴스 서비스 방식은 검색 엔진을 이용하여 다른 웹 사이트에 게재된 뉴스 기사를 크롤링한 후 그 결과를 이용하여 뉴스 기사 페이지(기사제목과 기사의 첫 단락 2~3줄 제공)를 구성하여 링크 정보를 제공해 주고 있다.[15]

나. 용어의 정리

웹 크롤링 기술이 현재 웹이 있는 곳에서는 대부분 사용되고 있고, 그 개발도 곳곳에서 이루어지고 있어서 웹 크롤링과 웹 크롤러를 가리키는 용어는 무수히 많다. 앤트(ants), 자동 인덱서(automatic indexers), 봇(bots), 웜(worms), 웹 스파이더(web spider), 웹 로봇(web robot) 등이 크롤러와 거의 같은 뜻으로 사용되고 있다.[16] 웹 크롤링과 유사한 기술로 웹 스크레이핑(web scraping)이 있다. 웹 스크레이핑은 웹 브라우저 화면에 표시되는 다양한 정보 중 사용자가 지정하거나 필요한 정보만 추출한 다음 이를 가공·저장하여 사용자에게 제공하는 기술임에 반하여 웹 크롤링은 특정 형태의 정보를 웹에서 자동으로 수집하여 정보를 최신으로 유지하는 소프트웨어 기술이

14) 구체적으로 크롤러에서 키워드 색인기, 색인어 추출, 파일 색인기, 조회수 랭킹 등에 관한 정보를 추출하여 데이터베이스화하는 경우에 이를 검색 엔진으로 확장할 수 있다(장현호, 전경식, 이후기, “분산형 병렬 크롤러 설계 및 구현”, 융합보안논문지 제19권 제3호, 2019. 9., 23쪽).

15) 유대종, “웹검색 서비스와 ISP 책임에 관한 소고”, 장작과 권리, 2007년, 103쪽

16) 최대우, 장영재, 이석호, 『데이터과학입문』, 한국방송통신대학교출판문화원, 2020년, 76쪽

라고 할 수 있다. 웹 스크레이핑과 웹 크롤링을 엄밀하게 구분하면, 웹 스크레이핑은 웹 문서를 가공하여 정보를 추출하는 과정으로서 웹 크롤링을 하지 않고도 웹 스크레이핑을 할 수 있음에 반하여 웹 크롤링은 기초가 되는 시드 URL을 저장한 뒤 웹 페이지의 하이퍼링크를 인식하여 URL을 갱신하며 반복적으로 웹 링크(web link)를 찾는 과정이라고 할 수 있다. 다시 말해 웹 스크레이핑은 웹으로부터 내용을 추출하는 것(pulling content from a page)이고, 웹 스크롤링은 수많은 웹 페이지에 이르기 위해 웹 링크를 따라가는 과정(following links to reach numerous pages)을 의미하는 것으로 구분할 수 있다.[17] 이를 검색 엔진에서의 기능을 중심으로 보면 웹 크롤링은 검색 엔진에서 가장 최신의 검색 결과를 보여주기 위해서 웹 페이지들을 돌아다니면서 새로운 페이지와 콘텐츠를 추출하는 것을 의미하고, 웹 스크레이핑은 대상이 된 사이트를 위해서 특별히 고안된 수단에 의하여 웹 페이지로부터 구조화된 정보(structured information)를 추출하는 과정을 말한다. 웹 스크레이핑이 각각의 페이지에서 필요한 정보를 빼내는 것이고, 웹 크롤링은 자동으로 정보 수집을 반복하는 프로그램이라고 할 수 있다.[18] 콘텐츠 혹은 데이터 수집의 관점에서는 웹 스크레이핑이 웹 페이지의 내용 전체를 웹 코드까지 가져오는 것이라면, 웹 크롤링은 이에 더하여 웹에서 공개된 정보를 데이터화하는 것까지 포함한다.[19] 이러한 개념상의 차이는 있으나 실제에서는 웹 스크레이핑이나 웹 크롤링 모두 웹에 존재하는 데이터의 수집을 위한 목적으로 사용되고 있어서, 양자를 엄밀하게 구분하지 않고 사용되고 있는 것으

17) 최대우, 장영재, 이석호, 『데이터과학입문』, 한국방송통신대학교출판문화원, 2020년, 76-77쪽
18) 타쿠로 사사키, 김경록 역, 『데이터와 크롤링을 몰라도 엑셀 및 구글 스프레드시트로 쉽게 할 수 있는 웹 데이터 수집의 기술 입문편』, 2017. 8. 한빛미디어, 24쪽
19) 김현숙, "크롤링을 이용한 공개데이터 수집·활용의 법적 쟁점에 대한 비판적 검토", 강원법학 제61권, 2020년, 227쪽

로 보인다. 이 글의 최종 목적은 웹 크롤링 활동의 법적 책임을 규명하는 것이고, 거기에는 개념상 또는 기술적 차이는 큰 의미가 없다. 이하에서는 논의의 효율성 및 표현의 경제성을 위하여 별도의 표시가 없는 이상 양자를 엄밀히 구별하지 않고 웹 크롤링(또는 크롤링), 웹 크롤러(또는 크롤러나 봇)으로 통칭한다.

그러나 다음의 것들은 이 글에서 논의하는 웹 크롤링과는 개념적으로나 실제적으로 차이가 존재하고, 그 사용의 법적 책임에 있어서도 결론을 달리할 수 있는 것이다. 구체적 사안에서 웹 크롤링 사용의 법적 책임을 규명하는 것은 웹 크롤링이라고 불리는 그 활동이 앞서 본 웹 크롤링의 범위에 들어가는 것인지 아니면 그와 확연히 구별되는 다른 것인지를 먼저 확인하여야 한다. 아래의 것들은 웹 크롤링과 확연히 구별되어야 하는 것들이다. 우선 미러링에 대하여 본다. 원래 미러링(mirroring)은 크기가 작고 저가인 여러 개의 하드디스크를 묶어 하나의 기억 장치처럼 사용하는 RAID(Redundant Array of Inexpensive Disks) 시스템에서 디스크에 데이터를 저장하는 방식 중 하나로서 하나 이상의 미러 장치에 중복으로 데이터를 저장하는 방식을 말한다.[20] 현재는 이를 넘어 해킹이나 불법 접근으로 로컬 컴퓨터 시스템에 저장된 데이터가 손실되는 것을 막기 위해 데이터를 하나 이상의 장치에 저장하거나 동일한 내용을 복수의 기기에서 사용하게 하는 것까지 의미가 확장되었는데, 이것을 웹 페이지의 콘텐츠 수집에 적용해보면 미러링은 특정 웹 페이지에 담긴 자료 전부를 다른 인터넷 사이트로 복사하여 오는 것이다. 문제는 미러링이 웹 페이지의 관리자나 소유자가 아니라 권한 없는 제3자에 의하여 이루어지는 경우

20) 김형근, 손진곤, 『컴퓨터 구조』, 한국방송통신대학교출판문화원, 2021년, 266쪽

에 법적 정당성을 인정하기 어려운 경우가 많다. 웹 크롤링이 이슈가 된 사례들 중 상당수는 웹 크롤링을 통한 미러링이 문제가 된 것이다.[21]

웹 크롤링은 콘텐츠 수집을 목적으로 하여 웹 페이지에 접속하여 그곳에 있는 링크들을 통하여 콘텐츠를 수집한다는 점에서, 리버스엔지니어링, 디컴파일, 디스어셈블 등과 구별된다. 리버스엔지니어링(reverse engineering, 역공학)은 원래 어떠한 시스템의 후기 상태에서 주어지는 생성물로부터 이전 상태의 관련 생성물을 추출하는 과학적 지식의 응용을 말한다. 소프트웨어 공학 입장에서는 소프트웨어 생명주기의 마지막 단계에서 주어지는 프로그램이나 사용자 매뉴얼 등으로부터 생명주기의 초기 단계 생성물에 해당하는 기능 명세나 설계 문서 등을 생성하는 과정을 의미한다. 리버스엔지니어링은 관심 있는 시스템을 분석하여 그것의 구성 요소들과 설계 정보를 획득하는 일이다.[22]. 디컴파일(decompile, 역컴파일)은 역공학의 중요한 도구 중의 하나인데, 컴파일이 프로그램 언어로 생성된 소스 코드를 컴퓨터에서 실행할 수 있는 이진코드로 된 기계어로 변환시켜 실행 파일로 만드는 과정(linking)을 의미하는 것에 반하여, 디컴파일은 상대적으로 저수준의 추상에 있는 프로그램 코드를 고수준의 추상으로 변형하는 것으로서 구체적으로 실행 파일에서 소스 코드 파일을 알아내는 것이다. 디컴파일은 잃어버린 소스 코드를 되찾거나 컴퓨터 보안 또는 오류 검출 정정 등의 과정에서 사용된다.[23]. 디스어셈블(disassemble, 역어

21) 예컨대 아래에서 보는 서울고등법원 2016. 12. 15. 선고 2015나2074198 판결
22) 김희천, 『소프트웨어공학』, 한국방송통신대학교출판문화원, 2020년, 262~263쪽
23) 김형근, 곽덕훈, 『C 프로그래밍』, 한국방송통신대학교출판문화원, 2017년, 5~8쪽; 정광식, 원유헌, 유헌창, 『프로그래밍 언어론』, 한국방송통신대학교출판문화원, 2017년, 50~51쪽

셈블)도 이 역시 역공학의 도구 중의 하나로서, 어셈블이 어셈블리 언어로 작성된 코드를 이진코드로 변환하는 것이므로 디스어셈블은 이것의 역과정이다. 이진코드를 프로그래밍 언어로 변환한다는 점에서는 디컴파일과 같지만, 역어셈블은 어셈블리 언어를 대상으로 한다는 점에서 차이가 있다.[24].

크롤링은 인터넷상에 공개된 웹서버에 접속하여 웹 페이지 정보를 수집한다는 점에서 일반적으로 컴퓨터를 이용하여 다른 사람의 정보처리장치 또는 정보처리조직에 침입하거나 기술적인 방법으로 다른 사람의 정보처리장치가 수행하는 기능이나 전자기록에 함부로 간섭하는 일체의 행위인 해킹과는 구별된다.[25] 그러나 그와 같은 구별은 인터넷 기술의 원리 및 사용 목적에 따른 일반적인 구별일 뿐이고, 크롤링이 해킹의 과정 또는 방법의 하나인 정보통신망 침입이 되지 않는다고 점은 법률적으로 해결해야 할 과제이다. 웹 크롤링과 정보통신망 침입의 구별은 Ⅲ.에서 후술한다.

다. 웹 크롤링의 역사

웹 크롤링의 역사는 웹의 시작과 함께 검색을 위한 도구로서 등장하기 시작하였다. 1990년 개발된 최초의 인터넷 검색 엔진인 아키(Archie)[26]는 특정 FTP(File Transfer Protocol) 사이트에서 한 달에 한 번 정도 디렉토리 목록을 로컬 파일로 다운로드하는 기능을 했다.

---

24) 정광식, 원유헌, 유헌창, 『프로그래밍 언어론』, 한국방송통신대학교출판문화원, 2017년, 50~51쪽
25) 유대종, "웹검색 서비스와 ISP 책임에 관한 소고", 장작과 권리, 2007년, 122쪽
26) 아카이브(Archives)의 줄임말이다.

이후 1991년에는 일반 텍스트 문서를 색인화하는 고퍼(Gopher)가 개발되었고, 같은 해에 World Wide Web이 도입됨에 따라 고퍼사이트 중 다수가 HTML로 연결된 링크들로 변경되었다. 1993년에는 최초의 웹 검색 엔진인 월드 와이드 웹 완더러(World Wide Web Wanderer)가 크롤러를 통하여 구현되었다. 이것은 처음에는 웹의 크기를 측정하는 데 사용되었는데 나중에는 완덱스(Wandex)라는 데이터베이스에 저장된 URL을 검색하는데 사용되었다. 또 다른 초기 검색 엔진인 알리웹(Aliwe, Archie-Like Indexing for the Web)은 사용자가 수동으로 구성한 사이트 색인의 URL을 추려내기도 했다.

초창기에는 웹 크롤러가 네트워크에 과부하를 준다는 문제가 논란의 대상이 되었다. 이 문제에 대해 1994년 웹 사이트 관리자가 크롤러가 웹 사이트에 접근하여 정보를 빼내는 것을 방지하는 로봇 배제 프로토콜(Robots Exclusion Protocol)이 도입되었다. 이후 크롤러의 기능이 개선되어 1994년에는 웹 관리자가 작성한 키워드나 설명 문구가 아니라 문서의 웹 콘텐츠만 탐색할 수 있도록 하여 불필요한 검색 결과를 줄이고, 더 나은 검색 기능을 허용하였다. 이 무렵 야후(Yahoo), 알타비스타(Altavista)와 같은 상용 검색 엔진이 등장하였다. 특히 야후는 초기에는 수동으로 유지·관리되는 웹 사이트의 디렉토리 시스템이었다가 추후에 검색 엔진으로 통합되었다. 1998년 등장한 구글의 검색 엔진은 단순하고 깔끔한 인터페이스, 합리적이고 편향되지 않은 검색 결과, 스팸 검색 결과의 감소라는 특징을 갖고 있다. 비편향성과 스팸(spam) 검색 결과 배제는 구글의 페이지랭크(PageRank) 알고리즘 사용과 앵커 페이지의 검색횟수에 가중치를 둔 결과였다.

3. 웹 크롤링의 기술적 원리

웹 크롤링이 개념적으로는 매우 단순함에도 고성능의 웹 크롤러를 구현하는 것은 많은 엔지니어링 문제를 야기한다. 웹 크롤러는 합리적인 시간 내에 시드 URL에서 출발하여 모든 하이퍼링크를 따라가기 위해 초당 수천 페이지를 다운로드해야 하며 일반적으로 그 프로세스는 수십 또는 수백 대의 컴퓨터에 분산되어 실행된다(분산형 크롤러). 웹 크롤러에 사용되는 아직 크롤링되지 않은 URL 집합과 이미 크롤링한 URL 집합의 두 가지 집합의 자료구조는 일반적으로 메인 메모리에 맞지 않으므로 효율적인 디스크 기반의 자료구조를 강구해야 한다. 또한 콘텐츠 제공자에게 정중함을 유지해야 하고, 특정 웹 서버에 과부하가 걸리지 않아야 하며, 고품질 페이지에 대한 크롤링의 우선순위를 정하고 말뭉치의 최신성을 유지해야 한다.

가. 웹과 웹 크롤링

웹 크롤러가 요구되고, 작동되는 근본적인 이유는 웹의 특성에 있다. 웹은 본질적으로 중앙에서 관리되는 단일의 정보저장소가 아니라 TCP(Transmission Control Protocol), DNS(Domain Name Service), 하이퍼텍스트 등과 같은 합의된 프로토콜(예컨대 전송 프로토콜로서 HTTP, 하이퍼텍스트 마크업 언어인 HTML 등)을 통하여 유지되는 데이터의 연합정보저장소의 성격을 갖는다. 웹을 통한 콘텐츠의 수집은 업데이트된 정보를 찾기 위해 웹을 샅샅이 뒤져보는 풀 모델(pull model) 또는 콘텐츠 제공자가 관심 콘텐츠를 수집자에게 푸시(제공)

하는 푸시 모델(push model)의 방법이 있게 된다. 초창기 푸시 모델로 채택한 경우도 있었으나, 현재는 대부분의 콘텐츠가 풀 모델을 통하여 수집된다. 푸시 모델이 콘텐츠 수집의 주요 수단이 되지 못한 이유는 이렇다. 웹 서버는 매우 자동적이고 자율적으로 작동하기 때문에 본질적으로 웹을 이용하려는 콘텐츠 제공자의 진입 장벽이 낮다. 초창기 웹 프로토콜이 극히 단순했기 때문에 이 장벽은 더욱 낮았다. 그런데 새로운 콘텐츠를 제공자가 매번 푸시하는 프로토콜을 추가하게 되면 웹 프로토콜 세트가 복잡해져서 제공자가 웹을 통하여 콘텐츠를 제공하는 데에 진입 장벽이 높아진다. 반면 풀 모델에는 제공자의 입장에서 추가 프로토콜이 필요하지 않아 상대적으로 진입장벽이 낮아진다. 그래서 현재에는 대부분 풀 모델이 채택되었는데, 풀 모델에서는 콘텐츠를 수집하려는 사람이 직접 웹을 샅샅이 탐색해보아야 하는 부담을 안게 된다. 그 부담을 효율적으로 처리하려고 등장한 기술이 웹 크롤링이다.

나. 웹 크롤링의 작동 원리

웹 크롤러는 시드 URL로 알려진 초기 URL에서 시작한다. 웹 크롤러는 시드 URL에 대한 웹 페이지를 다운로드하고 다운로드한 페이지에 있는 링크를 추출한다. 검색된 웹 페이지는 저장 영역에 저장하고 인덱싱(indexing)하므로 나중에 필요할 때 다시 검색할 수 있다. 다운로드한 페이지에서 추출된 URL은 관련 문서가 이미 다운로드 되었는지 여부를 다시 확인한다. 다운로드 되지 않은 URL은 추가 다운로드를 위해 웹 크롤러에 다시 할당된다. 다운로드할 URL이 더 이상 누락되지 않을 때까지 이 프로세스가 반복된다. 단순히 정리하면 시작 시

드 URL의 결정, 이를 웹 크롤러의 시작점으로 입력, 웹 크롤러가 시드 URL에서 URL 선택, 해당 URL에 해당하는 웹 페이지 가져오기, 새 URL 링크를 찾기 위해 해당 웹 페이지의 구문 분석, 새로 발견된 모든 URL을 웹 크롤러에 추가, 모든 URL의 정보를 추출할 때까지 작업 반복의 순서로 진행된다.

그 과정을 상세히 본다. 그림 1[27]은 분산형 크롤러(distributed crawler)의 작동 원리를 보여주고 있다. 분산형 크롤러는 고속 네트워크로 연결된 서로 다른 시스템(또는 머신)에서 실행되는 여러 프로세스로 구성된다. 크롤링 프로세스는 여러 작업자 스레드(thread)[28]로 구성되며 각 작업자 스레드는 반복적인 작업 주기를 실행한다. 웹 크롤러의 주요 구성요소는 웹 크롤러에서 아직 수집하지 못한 URL을 가져오는 URL 프론티어(frontier), 지정된 웹 페이지 중 하나를 가져오는 DNS 해석기(DNS resolver), HTTP를 통해 웹 페이지 내용을 가져오는 HTTP 페처(HTTP fetcher), 가져온 웹 페이지에서 텍스트와 링크를 추출하는 파서(parser), 검색된 URL 중 중복되거나 이전에 찾았던 링크들을 삭제하는 중복 URL 제거기(Duplicate URL eliminator) 등이 있다.[29]

작업이 시작되면 웹 크롤러는 각각의 자료구조로 지정된 URL을

---

27) 『Web Crawling』(각주 12 참조) 185쪽에서 인용하였다.

28) 스레드는 프로세스 내에서 다중처리를 위하여 제안된 개념으로 실행 단위를 프로세스에서 한 단계 낮추어 규정한 것이다. 이는 하나의 프로세스 내에서 자원 소유의 단위(unit of resource ownership)와 디스패칭의 단위(unit of dispatching)를 스레드의 개수만큼 존재하게 하여 프로세스의 속도를 향상시키려는 것이다(김진욱, 이병래, 곽덕훈, 『운영체제』, 2021년, 한국방송통신대학교출판문화원, 31쪽).

29) 장준영, 임경대, 이상진, “HTML 및 URL 특징을 이용한 유해사이트 수집 시스템”, 디지털포렌식연구 제16권 제1호, 2022. 3. 55쪽

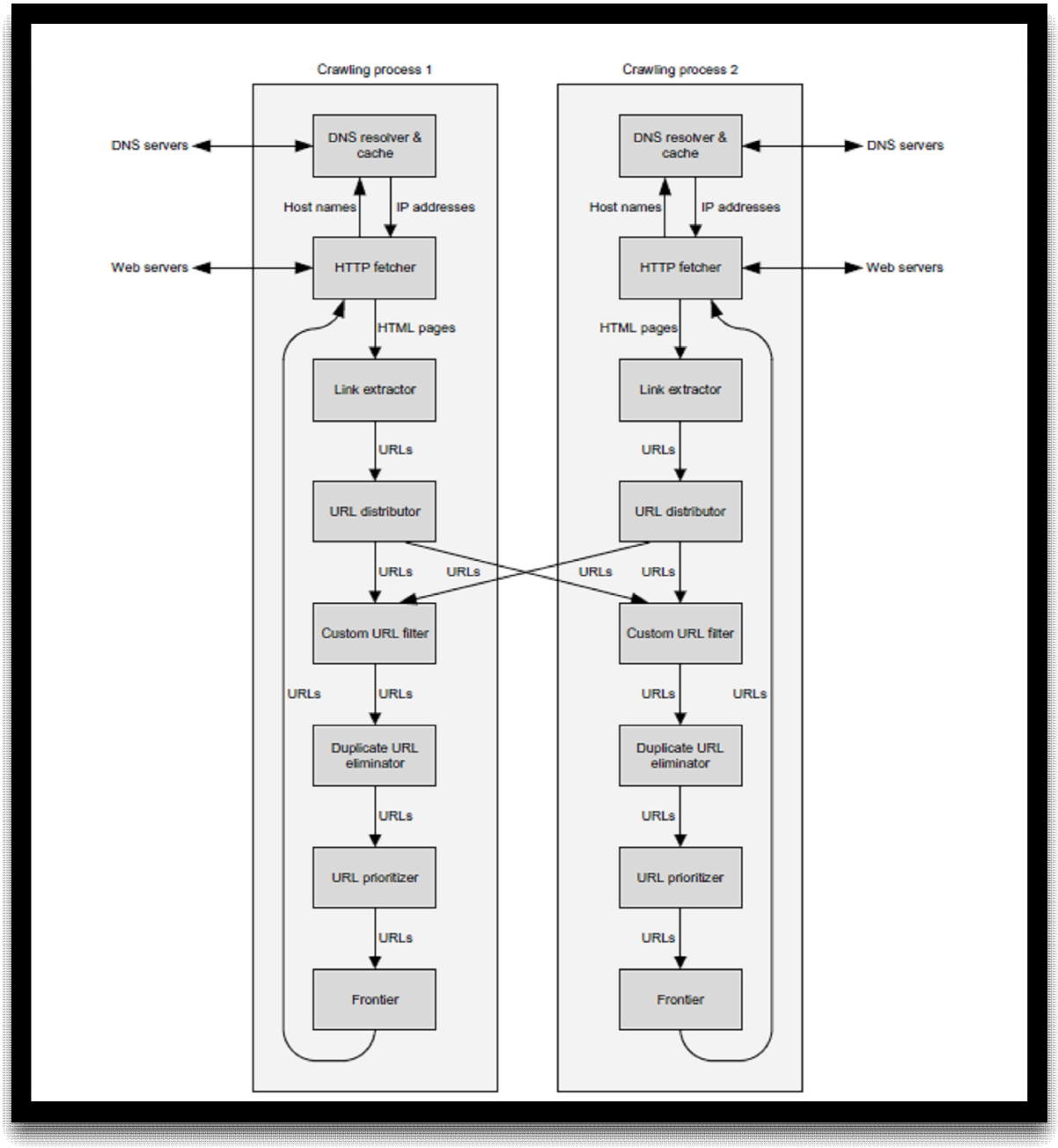

그림 1 분산형 크롤러의 작동 원리

가져와 DNS 해석기를 거쳐 우선순위에 따라 URL을 분배한다. 그 다

음 HTTP 페처를 호출한다. 페처는 먼저 DNS 하위 모듈을 호출하여 URL의 호스트 구성요소를 해당 웹 서버의 IP 주소와 확인하여(이때 이전 크롤링 작업시 저장된 캐시 값을 사용한다), 웹 서버에 연결하고 로봇 제외 프로토콜(robot exclusion protocol)을 확인한 후 웹 페이지를 다운로드한다. 다운로드한 후 그 웹 페이지를 저장소에 저장하거나 저장하지 않을 수도 있다. 어느 경우나 웹 페이지의 HTML 구문을 분석하고 그 결과를 하이퍼링크를 추출하는 링크 추출기(link extractor)로 전달한다. 링크 추출기를 통해 추출된 URL은 URL 분배기(URL distributor)에게 전달되며, URL 분배기는 각 URL을 크롤링 프로세스에 할당한다. 이 할당은 일반적으로 URL 호스트 구성요소, 해당 도메인 또는 해당 IP 주소를 해싱(hashing)하여 수행된다. 대부분의 하이퍼링크는 동일한 웹 사이트의 페이지를 참조하므로 로컬 크롤링 프로세스에 할당하는 것이 일반적이다. 그 URL은 사용자 지정 URL 필터(Custom URL filter)를 통과하여 수집의 대상이 아닌 특정 파일 확장자명을 가진 URL을 제외하도록 한 후 중복 URL 제거기로 전달되는데, 여기서 그때까지 발견된 URL 중 이전에 확인을 하였거나 다운로드한 것은 제외하고 새로운 URL만 전달한다. 마지막으로 URL 우선순위 프로그램(URL prioritizer)은 예상 페이지 중요도 또는 변경 비율과 같은 요소를 기반으로 결정된 우선순위에 따라 URL을 선택한 후 이를 다시 HTTP 페처로 보내서 앞서 한 과정을 반복하게 한다. 수집된 정보들은 설계에 따라 DBMS 또는 개별 파일형태로 저장될 수 있다.

다. 크롤링 정책(Crawling policies)

이처럼 웹 크롤러는 하나 이상의 시드 URL에서 시작하여 관련 페이지들을 다운로드하고 압축을 푸는 방식으로 웹 페이지를 다운로드하고, 이 과정을 재귀적으로 반복하여 웹 페이지를 다운로드 한다. 효율적인 웹 크롤러를 구현하기 위하여 알고리즘 차원에서 고려해야 하는 정책들은 다음과 같다. 개별 웹 크롤러의 성능 및 장·단점은 아래 정책들을 어떻게 조합하여 설계하였는지에 따라 결정된다.

1) 정중함 정책(politeness policy)

웹 크롤러에서 아직 수집하지 못한 URL을 가져오는 URL 프론티어(frontier)의 자료구조는 통상 피포(FIFO first in first out) 구조인 큐(queue)로 구현된다.[30] 이러한 구현은 웹 그래프의 너비 우선 탐색(breadth-first search, BFS)[31]을 하게 되는데, 이때 단점은 웹 페이지에 있는 하이퍼링크의 수와 구조가 일정하지 않다는 점이다. FIFO 대기열에는 동일한 웹 서버의 페이지를 참조하는 임의의 URL 집합이 포함되므로 웹 크롤러가 해당 서버에 연속으로 무수한 HTTP 요청을 발생시킨다. 웹 서버의 자원이 연속적인 요청을 처리하는 데

30) 대표적인 컴퓨터 자료구조는 스택(stack)과 큐(queue)가 있다. 스택이 한쪽 끝에서만 원소의 삽입 연산과 삭제 연산이 동시에 가능한 것과 달리 큐는 한쪽 끝에서는 원소의 삽입 연산만 가능하고 다른 쪽 끝에서는 원소의 삭제 연산만 가능하다. 가장 나중에 제출되어 작업대기 줄에 들어간 작업이 가장 먼저 처리되는 작업스케줄(Last in first out)을 컴퓨터 자료구조 측면에서 스택이라고 볼 수 있고, 반면 큐는 가장 처음에 제출되어 작업대기 줄에 들어간 작업이 가장 처음에 처리되는 작업 스케줄(first in first out)이 만들어진다(강태원, 정광식, 『자료구조』, 2017년, 한국방송통신대학교출판문화원, 73-74

31) 너비 우선 탐색은 맹목적 탐색 방법의 하나로 시작 정점을 방문한 후 시작 정점에 인접한 모든 정점들을 우선 방문하는 방법이다. 더 이상 방문하지 않은 정점이 없을 때까지 방문하지 않은 모든 정점들에 대해서도 너비 우선 검색을 적용한다(아래 웹 사이트 참조, 2022. 11. 1. 방문).
https://ko.wikipedia.org/wiki/%EB%84%88%EB%B9%84_%EC%9A%B0%EC%84%A0_%ED%83%90%EC%83%89

부족하게 되면 웹 서버 자체에 과부하가 실린다. 이처럼 짧은 간격에 많은 요청이 무작위로 쏟아지는 상황을 “무례한(impolite)” 것이라고 표현한다. 이는 자칫 웹 서버에 대한 서비스 거부 공격(DoS, denial of service attack)으로 해석될 수도 있다. 또한 웹 크롤링으로 인하여 다음과 같은 비용이 발생할 수 있다. ①웹 크롤러가 상당한 대역폭을 필요로 하고 오랜 시간 동안 높은 수준의 병렬 처리로 작동하기 때문에 그에 상응하는 네트워크 리소스가 소모된다. ②웹 크롤링의 대상이 되는 웹 서버에 대한 액세스 빈도가 너무 높은 경우 서버에 과부하가 발생한다. ③웹 크롤러의 설계가 잘못되어 웹 서버나 라우터에 충돌을 일으킬 수 있는 작업이나 웹 서버가 처리할 수 없는 페이지를 계속하여 반복하여 요청하는 경우는 그와 같은 비용이 더 커진다. ④소용량의 웹 크롤러라고 하더라도 다수의 웹 크롤러가 동시에 서버에 요청을 하여도 과부하가 발생할 수 있다.[32] 위와 같은 문제를 방지하기 위하여 웹 크롤러의 요청이 웹 서버의 작업에 과부하가 되지 않도록 하는 것을 ‘정중함 정책(politeness policy)’이라고 한다. 크롤러의 실행에 웹 사이트에 부하를 주지 않는 범위 내에서 실행할 필요가 있다. 일반적으로 1초에 1회 접속 정도라면 문제가 없다고 간주된다고 하는데,[33] 정중함 정책과 관련하여 접속의 횟수는 확립된 기준은 없어 보인다.

반면 웹 크롤러가 정중함 정책으로 웹 서버에 접근하지 못하는 동안 다른 유용한 작업을 수행하지 않고 있는 것도 컴퓨터 자원의 낭비가 된다. 이는 많은 HTTP 요청을 병렬로 실행하는 다중 스레드

32) Carlos Castillo, “Effective Web Crawling”, University of Chille 박사학위 논문, 2004년, 32쪽
33) 타쿠로 사사키, 김경록 역, 『데이터와 크롤링을 몰라도 엑셀 및 구글 스프레드시트로 쉽게 할 수 있는 웹 데이터 수집의 기술 입문편』, 2017. 8. 한빛미디어, 27~29쪽

(multi-thread) 또는 분산 크롤러(distributed crawler)에서 더욱 심각해진다. 대부분의 웹 크롤러는 동일한 서버에 대한 여러 중복 요청을 발행하지 않는데, 이는 각 URL의 호스트 구성 요소를 해싱하는 방법으로 웹 서버와 크롤링 스레드를 상호 매핑하고 이를 유지하는 것이다. 여기서 각 크롤링 스레드에는 별도의 FIFO 대기열이 있으며 그 큐에서 얻은 URL만 다운로드 한다. 보다 안전하게 웹 서버의 과부하를 줄이는 방법, 즉 정중함 정책은 해당 서버의 기능에 따라 그에 대한 요청 간격을 지정하는 것이다. 예컨대 크롤러는 서버에 대한 후속 요청을 해당 서버에서 마지막 페이지를 다운로드하는 데 걸린 시간의 10배수만큼 지연시키는 것이다. 이 방법은 웹 크롤러가 웹 서버 자원의 제한된 부분만을 사용하도록 하지만, 빠르고 응답성이 뛰어난 웹 서버보다 느리거나 제대로 연결되지 않은 웹 서버에서 다운로드되는 페이지 수가 더 적음을 의미하기 때문에 성능이 높고 효율적으로 구조화된(well-provisioned) 웹 페이지에 편향적인 크롤링 결과를 가져온다.

2) 재방문 정책(re-visit policy)

웹의 변화는 매우 동적이며, 용량에 따라 웹의 일부만을 크롤링 하는 데에도 상당한 시간이 소요될 수 있다. 웹 크롤러가 해당 웹 페이지에 대해 크롤링을 시작해 완료하는 사이에도 웹에는 생성(creations), 업데이트(updates), 삭제(deletions) 등 새로운 이벤트가 다수 발생했을 수도 있다. 이러한 이벤트를 감지하지 못하여 웹 페이지의 종전 내용에 대해 크롤링을 하는 것은 크롤링의 목적에 부합하지 않는 결과를 가져오고, 불필요한 비용을 초래한다. 이를 제거하기

위하여 신선도(freshness)와 연령(age)을 고려한 비용 함수(cost functions)를 사용한다. 신선도는 로컬 복사본이 해당 웹 사이트와 비교하여 얼마나 정확한지를 측정한 값이고, 연령은 로컬 복사본이 얼마나 오래되었는지를 측정한 값이다. 비용 함수를 기초로 동일 웹 사이트에 다시 방문하는 간격을 결정하는 데에는 균일 정책(Uniform policy)과 비례 정책(proportional policy)이 있다. 균일 정책은 변경 비율에 상관없이 크롤링된 모든 페이지를 동일한 빈도로 다시 방문하게 하는 것이고, 비례 정책은 상대적으로 더 자주 변경되는 페이지를 변경되는 빈도에 비례하여 더 자주 방문하도록 하는 것이다. 균일 정책에 따른 크롤링을 일괄 크롤링(batch crawling), 비례 정책에 따른 크롤링을 증분 크롤링(incremental crawling)으로 부르기도 한다. 일반적으로 크롤링 속도는 일괄 크롤링 순서 지정방법과 관련이 없지만, 증분 크롤링에서 페이지 재방문을 예약할 때 핵심요소가 된다. 일괄 크롤링은 크롤링 순서에는 중복되는 페이지가 포함되어 있지 않지만 이전에 크롤링된 페이지의 최신 정보를 얻기 위해 전체 크롤링 프로세스가 주기적으로 중지되고 다시 시작되는 방식이다. 이전 크롤링 사이클에서 수집된 페이지의 유용성 추정치와 같은 정보가 후속 크롤링 사이클에 제공될 수 있다. 증분 크롤링은 페이지는 크롤링 순서에 여러 번 나타날 수 있으며 크롤링은 개념적으로 절대 종료되지 않는 프로세스로 지정하는 방식이다. 대부분의 최신 상용 크롤러는 증분 크롤링을 수행하는 것으로 알려져 있다. 그 이유는 다양한 속도로 페이지를 다시 방문할 수 있기 때문에 검색에 더 유용하기 때문이다.

재방문 정책을 정할 때에는 URL 중복 테스트(URL seen test, UST 또는 duplicate URL eliminator, DUE)를 고려하여야 한다. 웹

크롤러의 자료구조가 이전에 발견되어 프론티어에 추가된 URL 집합을 확인하도록 하여 동일한 URL의 여러 인스턴스를 프론티어에 추가하는 것을 방지하는 것이 필요한데, 이를 URL 중복 테스트라고 한다. 이미 크롤러가 보았던(seen) 또는 다운로드한 URL을 큐에서 제거하는 것이다. 동일한 웹 페이지에 대해서 주기적으로 다시 크롤링을 할 때 더 이상 유효하지 않은 페이지를 계속 가리키는 URL을 삭제하는 것도 기능으로 한다.

크롤러가 방문해야 할 URL이 다수이고 주기적으로 정보를 수집하고 갱신처리하기 위해서는 방문 전략에 따른 계획된 스케줄에 의해 동작되어야 한다. 또한 새롭게 수집된 웹 페이지의 서브 링크(sub link)를 추출하게 되면 그만큼 시간이 더 소요되기 때문에 각 프로세서나 스레드에 URL 정보를 분산 배치해서 제한된 시간 안에 동작 할 수 있도록 해야 한다.[34)]

3) 선택 정책(selection policy)

정중함 정책을 제외하면 웹 크롤러는 어떤 순서로든 자유롭게 URL에 접속할 수 있는데, 웹의 크기가 매우 크고 또한 계속 확장되고 있으므로 어느 웹 페이지를 우선하여 작업을 수행할 것인지의 순서를 정하는 것이 중요하다. 그와 같은 순서를 정하는 것을 선택 정책이라고 한다. 여기에는 웹 크롤러가 원하는 페이지의 양을 모두 획득하여야 하는 점과 획득한 웹 페이지가 현재 웹 페이지의 상태에 비추어

34) 장현호, 전경식, 이후기, “분산형 병렬 크롤러 설계 및 구현”, 융합보안논문지 제19권 제3호, 2019. 9., 23쪽

최신 상태로 유지되어야 한다는 점을 고려해야 한다.

선택 정책을 위해서 웹 크롤러는 프런티어에 있는 URL의 우선순위를 지정할 수 있다. 예컨대 추정된 유용성의 정도에 따라 웹 페이지의 우선순위를 지정할 수 있다. 이는 구글의 PageRank에서 쓰는 방법인데 이때 유용성의 추정의 척도는 해당 웹 서버가 받는 트래픽의 양 또는 웹 사이트의 평판 등이 될 수 있다. 웹 페이지에 크롤링 우선순위를 할당하면 크롤러는 유용성에 따라 정렬된 디스크 기반 우선순위 대기열로 프론티어를 구성할 수 있다. 또 다른 방법은 우선순위를 고정된 수로 이산화(discretize)하고 각 수준에 대한 별도의 FIFO 대기열의 URL을 유지하는 것이다. URL은 이산화된 개별 우선순위에 할당되어 대응하는 큐에 삽입된다. URL을 대기열에서 빼기 위해 비어 있지 않은 우선순위가 가장 높은 대기열이 선택되거나 우선순위가 더 높은 대기열에 랜덤하게 선택된다.

4) 병렬화 정책(parallelization policy)

병렬 크롤러는 여러 프로세스를 병렬로 실행하는 크롤러이다. 이는 시간당 수집된 웹 페이지의 비율인 다운로드 율(download rate)을 최대화하기 위해 여러 개의 프로세스(process) 또는 스레드(thread) 단위로 정보를 처리를 이용하여 수집하는 방법이다. 통상 하나의 컴퓨팅 시스템에서 스레드들이 동시에 웹 페이지를 수집하는 멀티 스레드(multi-thread) 방식을 택하는데, 이를 확장하여 여러 컴퓨팅 시스템을 연결하여 동시에 웹 페이지를 수집하는 멀티 프로세스(multi-process) 방식을 취할 수도 있다.[35] 병렬 크롤러의 단점은 병

렬된 만큼 수집된 정보의 처리속도가 빨라졌지만, 그에 대응하는 DBMS의 처리속도가 따라가지 못하면 데이터의 손실이 발생할 수 있다는 점이다.[36]

라. 웹 크롤러의 유형

먼저 웹 크롤러는 다운로드하는 정보의 범위에 따라 범용 크롤러(general purpose crawling), 집중 크롤러(Focused crawling)으로 나눌 수 있다. 범용 크롤러는 특정 URL 집합과 해당 링크에서 가능한 많은 페이지를 수집하는 것이다. 이는 모든 페이지를 가져오기 때문에 속도가 느린 반면, 네트워크 트래픽이 증가할 수 있다. 집중 크롤러는 네트워크 트래픽 및 다운로드의 양을 줄일 수 있도록 특정 주제에 대한 문서만을 수집하는 것이다. 집중 크롤러는 미리 정의된 문서 집합에 적절한 페이지를 찾는 것을 목적으로 한다. 웹의 관련 영역만 크롤링하여 하드웨어 및 네트워크 자원의 낭비를 줄인다.

크롤러의 작동 방식에 따라 보면, 일반형 크롤러(general crawler)와 분산형 크롤러(distributed crawler)[37]로 나뉜다. 일반형 크롤러는 URL 목록인 시드를 기준으로 웹 문서를 수집하고 수집된 웹 문서에 포함된 URL을 다음 시드로 활용해서 다음 웹 문서를 수집하기 때문

35) 홍성학, “웹 크롤링을 이용한 스마트 가격 추적기의 구현”, 서울과학기술대학교 석사학위논문, 2015년 6~7쪽

36) 장현호, 전경식, 이후기, “분산형 병렬 크롤러 설계 및 구현”, 융합보안논문지 제19권 제3호, 2019. 9., 22쪽

37) 분산 시스템은 네트워크를 통하여 결합된 프로세서들의 집합으로 이 프로세서들은 메모리와 클럭을 공유하지 않고 자신의 로컬 메모리를 갖도록 구성된다. 분산 시스템은 서로 다른 프로세서 또는 기기들의 자원을 공유하고, 연산속도를 향상시키며, 프로세스의 신뢰성 향상 등을 목적으로 한다(김진욱, 이병래, 곽덕훈, 『운영체제』, 2021년, 한국방송통신대학교출판문화원, 225~227쪽).

에 첫 시드가 특정 웹 사이트의 전체 데이터를 수집할 수 있는 대표 시드가 아니면 해당 웹 사이트의 전체 데이터를 수집할 수 없다는 점과 방대한 웹 문서를 단일 시스템에서 크롤링하기 때문에 많은 시간이 요구된다는 단점이 있다. 반면 분산형 크롤러는 크롤링 시간을 줄이기 위해 다수의 컴퓨팅 시스템을 서버-클라이언트(sever-client) 환경으로 연결한다. 여기서 서버는 초기 시드를 클라이언트에 분배하고 클라이언트가 수집한 웹 문서를 전달받는다. 클라이언트는 서버로부터 전달받은 시드를 기준으로 크롤링하고 다음 크롤링을 위해 수집된 웹 문서에서 시드 추출 및 크롤링을 반복적으로 수행한다. 분산형 크롤러는 빠른 웹 문서 수집에 목적을 두고 있기 때문에 분산 환경에서의 네트워크 트래픽을 최소화하기 위해 각 클라이언트 간 수집한 웹 문서에 대한 정보를 실시간으로 공유하지 않는다. 그래서 각 클라이언트는 다른 클라이언트가 수집한 웹 문서에 대한 정보 없이 자신이 수행한 웹 문서 내에 포함된 모든 URL을 시드로 사용하기 때문에 다른 클라이언트에서 수집된 정보가 중복해서 크롤링될 수 있는 단점이 있다.[38] 또한 실험 결과에 비춰보면, 분산형 크롤러를 통해 단순히 머신 개수가 2배로 증가한다고 해서 소요시간이 그에 비례하여 1/2로 줄어드는 것은 아닌 것으로 확인되었다. 네트워크 상황(트래픽 또는 웹페이지 콘텐츠 양), 시스템 메모리 및 CPU 성능 등 크롤링 속도를 좌우하는 요소에는 다른 것들이 많기 때문이다.[39]

4. 웹 크롤링 방지 기술[40]

38) 김희숙, 한나, 임숙자, “빅데이터 분석 기반의 정보 검색을 위한 웹 크롤러 서비스 구현”, 디지털콘텐츠학회논문지 vol.18, no. 5, 2017. 8., 934쪽
39) 홍성학, “웹 크롤링을 이용한 스마트 가격 추적기의 구현”, 서울과학기술대학교 석사학위논문, 2015년 23쪽
40) 김선태, “웹크롤러 기반의 개인정보 침해 점검 시스템에 대한 방법론 연구”, 숭실대학

일반적으로 웹상의 자료는 웹 크롤러를 기반으로 하는 검색 엔진에 의해 검색되고 검색결과로 노출된다. 이는 통상의 웹 관리자가 의도하는 바이다. 그러나 크롤링되거나 공개되어서는 안 되는 정보, 예컨대 개인정보가 크롤링될 수도 있으므로 검색 엔진에 노출되지 않기 위한 크롤러의 접근 차단 기술도 고안되고 있다.

가. 로봇 배제 프로토콜(robot exclusion protocol, robot.txt)

robot.txt는 크롤러의 접근을 차단하거나 웹 페이지 관리자가 크롤링을 허용하는 범위와 정도에 관한 정보를 표시하는 것이다.

그림 2[41]의 왼쪽은 robot.txt가 설정되지 않는 웹 페이지에 대한 크롤링 과정이다. 웹 크롤러가 해당 웹 페이지의 모든 페이지의 모든 파일들에 대해서 크롤링이 가능하다. 그림 2의 오른쪽은 robot.txt 프로토콜을 웹 서버 설정한 모습이다. 여기서 robot.txt 프로토콜은 각 웹 사이트마다 하나만 가질 수 있고, 반드시 해당 웹 사이트 도메인의 최상위 경로에 존재해야 한다. 즉 웹 크롤러가 해당 웹 사이트에 접근하였을 때 가장 먼저 확인할 수 있는 위치에 존재하여야 하는 것이다. user-agent[42], disallow[43], crawl-delay[44] 등의 속성(attribute)을 사용하여 크롤링의 허용 범위를 정한다. 다음 표는 속성과 속성값

교 석사학위 논문, 2016., 18~27쪽

41) 김선태, "웹크롤러 기반의 개인정보 침해 점검 시스템에 대한 방법론 연구", 숭실대학교 석사학위 논문, 2016., 19쪽에서 인용하였다.

42) 자동 검색을 허용할 검색 엔진 로봇을 설정하는 속성이다.

43) 크롤링을 허용하지 않을 디렉토리 또는 파일을 정하는 속성이다.

44) 크롤러가 해당 웹 페이지에 다시 크롤링 할 수 있는 시간제한을 정하는 속성이다. 수치의 단위는 초 단위이다.

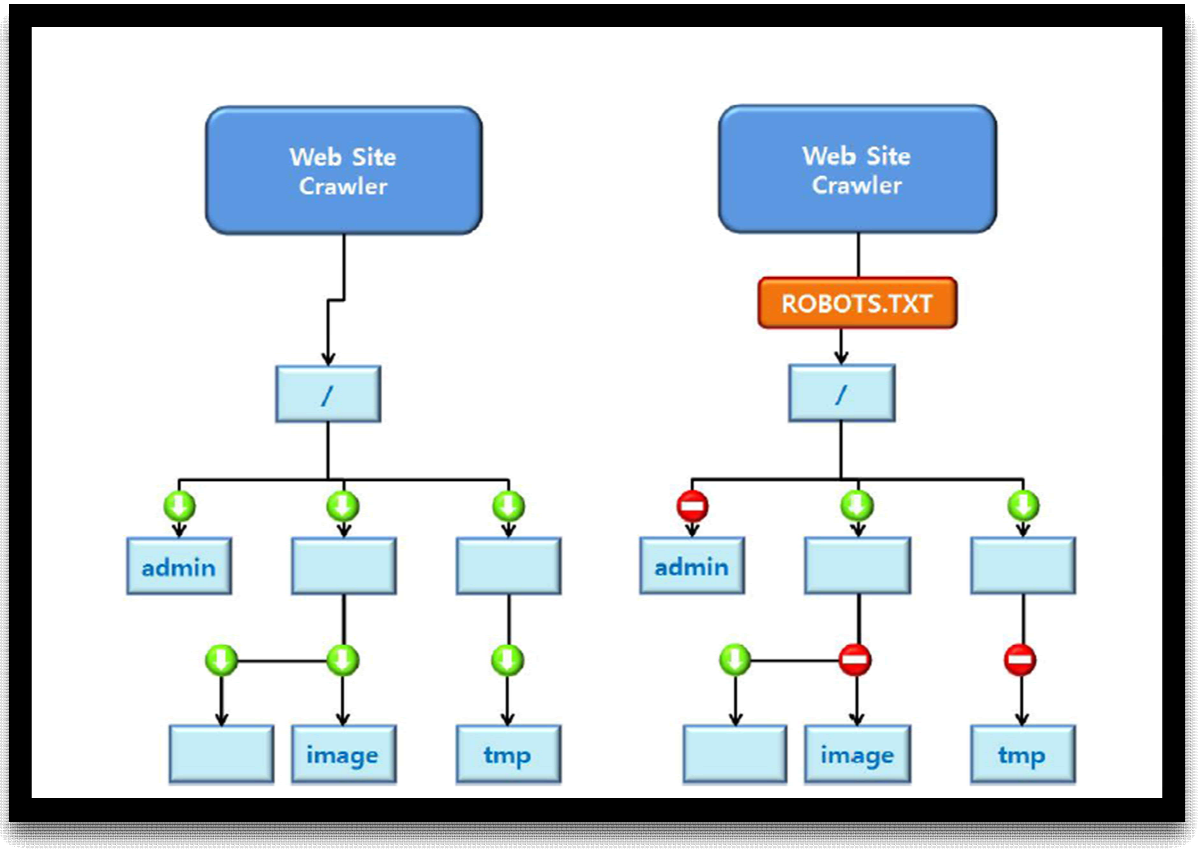

그림 2 robot.txt의 적용/미적용 알고리즘

(argument)을 가지고 크롤링의 허용범위를 정하는 예시이다.

속성-속성값	크롤링의 허용범위
user-agent : * allow : /	모든 크롤러에 대해 모든 문서에 대한 접근을 허락한다.
user-agent : * disallow : /	모든 크롤러에 대해 모든 문서에 대한 접근을 차단한다.

user-agent : googlebot disallow : /	구글 검색 엔진 로봇에 대한 접근을 차단한다.
user-agent : * disallow : /root/	모든 로봇에 대해 해당 디렉토리에 대한 접근을 차단한다.
user-agent : * disallow : /*.pdf$	모든 로봇에 대해 해당피일 유형에 대한 접근을 차단한다.
user-agent : * allow : / crawl-delay :60	모든 로봇에 대해 모든 문서에 대해 60초에 한 번씩 접근을 허락한다.
user-agent : * allow : / sitemap : https://www.google.com/sitemap.xml	모든 로봇이 사이트맵에 있는 URL에 접속하여 색인을 할 수 있다.

표 2 robot.txt의 사용예시

다만 robot.txt 프로토콜은 지침 수준의 규약이므로 크롤러의 동작을 강요할 수 없으며, robot.txt에도 불구하고 강제로 크롤링할 수 있도록 크롤러가 설계될 수도 있다. 따라서 웹 크롤러로부터 노출되지 않기를 원하는 정보는 서버 수준에서 비공개 처리하거나, 비밀번호 또는 암호화하여 보호하는 수밖에 없다.

나. 메타태그(metatag)

robot.txt는 현재 운영되고 있는 웹 사이트의 자동 검색을 차단해 주는 기능을 한다. 그러나 대형 포털사이트의 검색 엔진은 검색 결과

의 빠른 처리를 위해 크롤링한 결과 페이지를 캐시페이지로 보관하고 있다. 이 캐시페이지는 크롤링했을 당시의 페이지를 저장하여 보관하고 있는 것이므로 robot.txt의 기능으로는 자동 검색 차단을 수행할 수 없다. 따라서 이러한 부분을 해소하기 위해 메타태그를 활용한 자동검색차단을 수행한다. HTML 페이지에서 메타태그는 주로 <head> 태그와 </head> 태그 사이에 입력된다.[45] 메타태그에서 사용되는 속성은 name[46] content[47] 등이 있다. 다음은 모든 봇이 웹 페이지에 대해서 색인을 생성하지 못하도록 하는 메타태그의 예시이다.

```
<head>
<meta name="robots" content="noindex"/>
</head>
```

표 3 메타태그 예시

다. 캡차(CAPTCHA)

대형 포털사이트들은 봇의 접근으로 인한 트래픽 초과를 방지하고 해당 포털의 원활한 이용을 위하여 검색에 대한 사용 제한(rate limit)[48]을 설정하고 있다. 예컨대 1초에 검색횟수가 25회를 초과하면 검

45) <head> ..... </head> 태그는 그 사이에 있는 내용이 웹브라우저를 통해 사용자에게 보이지는 않지만 문서의 각종 정보와 문서 자체에 대한 설명을 담는 부분이다(이관용, 『HTML 웹프로그래밍』, 한국방송통신대학교출판문화원, 2020년, 11쪽).

46) 접근을 허용하지 않는 크롤러를 정하는 속성이다. robot.txt의 user-agent와 같은 역할을 한다. 「name="robots"」는 robot.txt의 「user-agnent : * 」와 같은 의미로 모든 자동 검색 봇을 차단한다는 것이다.

47) 허용하지 않는 크롤러의 활동을 정하는 속성이다. 이 속성에 사용되는 속성값은 "all"(색인 생성 또는 크롤링에 제한이 없다), "noindex"나 "none"(검색 결과에 대해 저장된 페이지와 링크표시를 하지 않는다), "noimageindex"(이미지에 대한 색인을 생성하지 않는다) 등이 있다.

48) 이는 특정 요구에 대해서 즉각적인 응답을 제한하는 것인데, 네트워크에 대용량의 트래픽을 야기할 염려가 있는 DoS 공격이나 크롤러의 접근을 제한하려는 목적이다.

색 엔진에서는 봇의 활동으로 간주하고 그 요청에 즉각적인 응답을 거부하고 CAPTCHA로 강제 이동시킨다. CAPTCHA(Completely Automated Public Turing test to tell Computers and Humans Apart)는 완전 자동화된 컴퓨터와 사람을 구별하는 기능을 한다. 사람은 구별이 가능하지만 컴퓨터는 구별하기 힘들도록 의도적으로 변조한 그림을 주고 해당 그림에 쓰여 있는 내용을 증명하게 하는 방법이 자주 사용된다. 이것은 기존의 이미지를 분석하는 것과는 다르게 변형된 이미지를 인식해야 하므로 컴퓨터 프로그램을 이용한 크롤러 기반에서는 통과하기 어렵다. CAPTCHA는 컴퓨터가 사람을 대상으로 테스트 하는 것이기 때문에 리버스 튜링 테스트(reverse Turing test)라고도 한다. 구글은 CAPTCHA를 변형한 RECAPTCHA를 사용하는데, 이것은 사용자에게 왜곡된 단어 이미지를 보여준 뒤 글자를 인식하여 입력하게 하는 방식으로 CAPTCHA와 차이점은 컴퓨터가 인식할 수 있도록 연산을 통해 데이터화된 단어와 그렇지 않은 단어 2개를 보여준 뒤, 사람은 첫 번째 단어를 인식하면 후자도 인식할 수 있다는 것을 전제로 한다.

5. 소결

웹 크롤링은 하나 이상의 시드(seed) URL에서 그와 관련된 웹 페이지를 다운로드 하고, 여기에 포함된 하이퍼링크를 추출한 후 하이퍼링크를 따라 웹 페이지를 재귀적으로 계속 다운로드하면서 웹 페이지에 있는 정보를 수집하는 프로세스이다. 효율적인 웹 크롤러를 구현하기 위해서는 웹 크롤러의 요청이 웹 서버의 작업에 과부하가 되지 않도록 크롤링 속도를 조절하는 정중함 정책(politeness policy), 크롤링

결과를 최신으로 유지하기 위한 재방문 정책(re-visit policy), 크롤링 할 웹 사이트의 순서를 정하는 선택 정책(selection policy), 크롤링 속도를 개선하기 위한 다중 스레드 방식의 병렬화 정책(parallelization policy) 등을 유념하고 설계해야 한다. 반면 웹 사이트 관라지의 입장에서 크롤링을 방지하기 위한 기술로는 로봇 배제 프로토콜(robot exclusion protocol), 메타태크(metatag), 캡차(CAPTCHA) 등이 있다.

Ⅲ. 웹 크롤링 사용의 형사법적 책임

1. 도입

가. 가치의 충돌

홍보 또는 마케팅의 수단으로 웹 사이트를 운영하는 사람이나 웹 사이트를 직접적인 영업활동의 수단으로 사용하는 사람의 경우에는 통상 대형 포털사이트나 검색 엔진에 자신의 웹 사이트가 보다 많이 보다 검색되어 그 결과로서 노출되기를 희망할 것이다. 이를 위해서는 자신의 웹 사이트에 대한 크롤링은 수인할 수밖에 없다. 반면, 경쟁자가 자신의 웹 사이트에 와서 자신의 정보를 수집해가는 것은 막으려고 할 것이다. 그런 면에서 웹 사이트 운영자의 웹 크롤링에 대한 입장은 이중적이다.

좀 더 큰 관점에서 보자. 크롤링이 대상으로 하는 웹은 기본적으로 열린 공간이다. 웹을 통해 정보를 제공하는 이상 그 정보에 대한 접근에는 원칙적으로 제한이 없다. 그 접근이 정보제공자의 의도나 이익에 반하는 것이라고 하더라도 마찬가지이다. 웹 또는 인터넷은 사람들이 자유롭게 접근하여 정보를 주고받고 여론을 형성해나가는 공공재의 성격을 갖기 때문이다. 반면 웹 사이트에 자신의 수집한 데이터를 제공하고 그것을 기초로 영업활동을 하는 사람들의 사익도 보호되어야 한다. 웹 크롤링을 통한 대규모의 웹 사이트의 데이터 수집이 가능하게 된 상황에서, 웹을 통해 공개되었다는 이유로 모든 데이터에 대한 방법적 제한 없이 그 수집이 허용된다는 것은 웹을 통한 정보교류를 활성화를 저해하는 것이된다. 그래서 웹 크롤링 사용의 형사책임은 그

와 같은 공공성과 사익을 어느 지점에서 조화시킬 것인지가 문제되는 것이다.

이하에서는 웹 크롤링의 사용 특히 데이터 수집에 대한 형사법적 책임을 검토할 것이다. 검토의 대상이 되는 구성요건은 정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 '정보통신망법'이라고 한다) 제48조 제1항[49]의 정보통신망 침입, 저작권법상 데이터베이스 무단복제에 의한 데이터베이스제작자의 권리 침해[50], 형법 제314조 제2항의 컴퓨터등장애업무방해[51]와 같은 3가지이다. 처벌구성요건의 해석은

---

49) 정보통신망법
제48조(정보통신망 침해행위 등의 금지)
① 누구든지 정당한 접근권한 없이 또는 허용된 접근권한을 넘어 정보통신망에 침입하여서는 아니 된다.
제71조(벌칙)
① 다음 각 호의 어느 하나에 해당하는 자는 5년 이하의 징역 또는 5천만원 이하의 벌금에 처한다.
9. 제48조제1항을 위반하여 정보통신망에 침입한 자

50) 저작권법
제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.
19. "데이터베이스"는 소재를 체계적으로 배열 또는 구성한 편집물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것을 말한다.
20. "데이터베이스제작자"는 데이터베이스의 제작 또는 그 소재의 갱신·검증 또는 보충(이하 "갱신등"이라 한다)에 인적 또는 물적으로 상당한 투자를 한 자를 말한다.
제93조(데이터베이스제작자의 권리)
① 데이터베이스제작자는 그의 데이터베이스의 전부 또는 상당한 부분을 복제·배포·방송 또는 전송(이하 이 조에서 "복제등"이라 한다)할 권리를 가진다.
② 데이터베이스의 개별 소재는 제1항에 따른 해당 데이터베이스의 상당한 부분으로 간주되지 아니한다. 다만, 데이터베이스의 개별 소재 또는 그 상당한 부분에 이르지 못하는 부분의 복제등이라 하더라도 반복적이거나 특정한 목적을 위하여 체계적으로 함으로써 해당 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에는 해당 데이터베이스의 상당한 부분의 복제등으로 본다.
제136조(벌칙)
② 다음 각 호의 어느 하나에 해당하는 자는 3년 이하의 징역 또는 3천만원 이하의 벌금에 처하거나 이를 병과할 수 있다.
3. 제93조에 따라 보호되는 데이터베이스제작자의 권리를 복제·배포·방송 또는 전송의 방법으로 침해한 자

51) 형법
제314조(업무방해)
① 제313조의 방법 또는 위력으로써 사람의 업무를 방해한 자는 5년 이하의 징역 또는 1천500만원 이하의 벌금에 처한다.

죄형법정주의에서 파생되는 해석의 원칙을 준수해야 한다. 또한 이론상으로 현대국가에서 형법은 시민의 자유를 제한하고 국가의 강제력을 행사하는 가장 최후의 수단이자 강력한 수단이라고 간주되므로, 형법이 최대한 겸양성을 가져야 한다는 점도 고려해야 한다. 이에 더하여 앞서 살펴본 웹 또는 인터넷의 공공성도 크롤링과 관련해서는 중요한 해석의 기준점이 될 것이다.

나. 대법원 2022. 5. 12. 선고 2021도1533 판결

대법원 2022. 5. 12. 선고 2021도1533 판결은 웹 크롤링을 통한 데이터 수집의 형사책임에 관하여 정면으로 판시한 최초의 판결이다.[52] 위 판결은 새로운 법리를 설시하면서 앞에서 언급한 3개의 구성요건의 해당성을 검토하였다. 위 판결의 주요 판시요지는 다음과 같다. [1]은 정보통신망침해와 관련하여 접근권한의 유무와 제한여부를 판단하는 기준에 관한 것이고, [2]는 저작권법상 데이터베이스제작자의 권리가 침해되었는지를 판단하는 방법에 관한 것이다.[53]

> [1] 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조 제1항은 누구든지 정당한 접근권한 없이 또는 허용된 접근권한을 넘어 정보통신망에 침입하는 것을 금지하고 있고, 이를 위반하여 정보통신망에 침입한 자에 대하여는 5년 이하의 징역 또

---

② 컴퓨터등 정보처리장치 또는 전자기록등 특수매체기록을 손괴하거나 정보처리장치에 허위의 정보 또는 부정한 명령을 입력하거나 기타 방법으로 정보처리에 장애를 발생하게 하여 사람의 업무를 방해한 자도 제1항의 형과 같다.

52) 위 판결은 웹 크롤링의 대상이 되었던 회사의 상호를 따서 일명 '야놀자 판결'로 불리기도 한다.

53) 위 판결에는 그 외에도 형법상 컴퓨터등장애업무방해와 관련된 설시도 있으나, 그것은 대법원의 기존 법리를 재확인한 것이어서 여기서는 생략한다.

는 5천만 원 이하의 벌금에 처한다(위 법 제71조 제1항 제9호). 위 규정은 이용자의 신뢰 내지 그의 이익을 보호하기 위한 규정이 아니라 정보통신망 자체의 안정성과 그 정보의 신뢰성을 보호하기 위한 것이므로, 위 규정에서 접근권한을 부여하거나 허용되는 범위를 설정하는 주체는 서비스제공자이다. 따라서 서비스제공자로부터 권한을 부여받은 이용자가 아닌 제3자가 정보통신망에 접속한 경우 그에게 접근권한이 있는지 여부는 서비스제공자가 부여한 접근권한을 기준으로 판단하여야 한다. 그리고 정보통신망에 대하여 서비스제공자가 접근권한을 제한하고 있는지 여부는 보호조치나 이용약관 등 객관적으로 드러난 여러 사정을 종합적으로 고려하여 신중하게 판단하여야 한다.

[2] 데이터베이스제작자는 그의 데이터베이스의 전부 또는 상당한 부분을 복제·배포·방송 또는 전송(이하 '복제 등'이라고 한다)할 권리를 가지고(저작권법 제93조 제1항), 데이터베이스의 개별 소재는 데이터베이스의 상당한 부분으로 간주되지 않지만, 개별 소재의 복제 등이라 하더라도 반복적이거나 특정한 목적을 위하여 체계적으로 함으로써 해당 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에는 해당 데이터베이스의 상당한 부분의 복제 등으로 본다(저작권법 제93조 제2항). 이는 지식정보사회의 진전으로 데이터베이스에 대한 수요가 급증함에 따라 창작성의 유무를 구분하지 않고 데이터베이스를 제작하거나 그 갱신·검증 또는 보충을 위하여 상당한 투자를 한 자에 대하여는

일정기간 해당 데이터베이스의 복제 등 권리를 부여하면서도, 그로 인해 정보공유를 저해하여 정보화 사회에 역행하고 경쟁을 오히려 제한하게 되는 부정적 측면을 방지하기 위하여 단순히 데이터베이스의 개별 소재의 복제 등이나 상당한 부분에 이르지 못한 부분의 복제 등만으로는 데이터베이스제작자의 권리가 침해되지 않는다고 규정한 것이다.

데이터베이스제작자의 권리가 침해되었다고 하기 위해서는 데이터베이스제작자의 허락 없이 데이터베이스의 전부 또는 상당한 부분의 복제 등이 되어야 하는데, 여기서 상당한 부분의 복제 등에 해당하는지를 판단할 때는 양적인 측면만이 아니라 질적인 측면도 함께 고려하여야 한다. 양적으로 상당한 부분인지 여부는 복제 등이 된 부분을 전체 데이터베이스의 규모와 비교하여 판단하여야 하며, 질적으로 상당한 부분인지 여부는 복제 등이 된 부분에 포함되어 있는 개별 소재 자체의 가치나 그 개별 소재의 생산에 들어간 투자가 아니라 데이터베이스제작자가 그 복제 등이 된 부분의 제작 또는 그 소재의 갱신·검증 또는 보충에 인적 또는 물적으로 상당한 투자를 하였는지를 기준으로 제반 사정에 비추어 판단하여야 한다.

또한 앞서 본 규정의 취지에 비추어 보면, 데이터베이스의 개별 소재 또는 상당한 부분에 이르지 못하는 부분의 반복적이거나 특정한 목적을 위한 체계적 복제 등에 의한 데이터베이스제작자의 권리 침해는 데이터베이스의 개별 소재 또는 상당하지 않은 부분에 대한 반복적이고 체계적인 복제 등으로 결국 상당한 부분의 복제 등을 한 것과 같은 결과를 발생하게 한 경우에 한하여 인정함이 타당하다.

결론적으로 보자면 위 판결은 웹 크롤링을 통한 데이터 수집에 있어 세 가지의 구성요건, 즉 정보통신망침입, 데이터베이스제작자의 권리 침해 및 컴퓨터등장애업무방해 모두에 대해서 그 구성요건해당성을 부정하였다. 관련 부분에서 상술하겠지만 나는 위 판결의 결론에는 찬성하지만, 정보통신망침입과 관련하여 위 판결이 들고 있는 이용약관과 보호조치가 객관적 사정이라고 볼 수 있을 지에 대해서는 의문이다. 왜냐하면 이용약관과 보호조치는 모두 정보통신 서비스제공자의 임의에 달린 것이어서 '객관적'인 것이라고 할 수 없을 뿐만 아니라, 그것들에 따라 접근권한의 범위를 정한다는 것은 서비스제공자의 임의의 의사에 따라 형사책임의 범위가 달라진다는 것을 의미하기 때문이다. 이용약관과 보호조치가 접근권한의 범위를 정하는 여러 가지 기준 중의 하나는 될 수 있을 것이나, 그것들을 절대적인 것으로 접근권한의 범위를 정해서는 안 된다.

위 판결의 사실관계(공소사실), 그 진행경과는 아래와 같다. 위 판결의 사실관계와 진행경과를 상세하게 확인하는 이유는, 웹 또는 플랫폼 비즈니스에서 웹 크롤링을 하는 목적과 그 활용 국면은 물론 웹 크롤링의 사용에 법적 책임(형사 책임에 국한하지 않는다)을 확인하는데 필요한 정황과 간접사실들을 입체감 있게 알 수 있기 때문이다.[54]

1) 공소사실의 요지[55][56]

54) 이에 관하여는 이미 서문에서 밝혔다.

55) 논의의 편의성을 위하여 이 글의 주제와 관련이 있는 부분으로 한정하여 요약하였다. 나는 이하에서 나오는 대법원 또는 사실심 법원 판결례의 사실관계는 모두 관련 쟁점을 파악하는 데에 필요한 한도에서 요약하거나 간략히 정리하였다.

피고인들은 숙박업체 정보 제공 및 예약 서비스를 제공하는 회사(이하 '피고인 회사'라고 한다)의 임직원들로서 경쟁 관계에 있는 피해자 회사가 운영하는 모바일 어플리케이션인 '바로예약'(이하 '이 사건 앱'이라고 한다)이나 PC용 홈페이지에 접속하여 제휴 숙박업소 목록, 주소 정보, 가격정보 등을 확인하고 영업을 위하여 이를 내부적으로 공유하고 있었다.

① 정보통신망법위반(정보통신망침해등)

피고인들은 이 사건 앱의 프로그램 소스를 '패킷캡처' 앱[57]을 이용하여 분석하여 모바일 앱 프런트엔드(FRONT-END) API 서버(이하 '이 사건 API 서버'라고 한다)[58]의 모듈, 해당 서버의 URL 주

---

56) '야놀자 판결'의 사실관계, 즉 '야놀자 판결'의 사실관계에 대해서는 형사소송과 별도로 민사소송도 함께 진행되었다. 형사판결에서 피해자의 입장에 있는 '야놀자'가 피고인들의 회사인 '여기 어때'를 상대로 손해배상 등을 구하는 소송을 제기하였던 것이다. 재밌는 점은 형사판결과 민사판결의 결론이 서로 달랐던 점이다. 민사판결의 내용에 대해서는 Ⅳ.에서 상술한다.

57) 패킷캡처(Packet Capture, 또는 packet analyzer, packet sniffer, protocol analyzer, network analyzer라고도 한다)는 컴퓨터 네트워크상에서 이동하는 패킷 또는 트래픽을 가로채거나 저장할 수 있는 프로그램이다. 사용법도 어렵지 않고 널리 사용되고 있는 프로그램으로 구글 플레이에서 쉽게 다운받을 수 있다.

58) 데이터 센터에서 모든 데이터를 처리할 때 과부하와 그에 따른 지연 현상을 완화하기 위하여 사용자 또는 엔드포인트(endpoint) 단말기의 물리적 위치와 인접한 곳에서 컴퓨팅을 수행하는 것을 엣지(edge) 컴퓨팅이라고 한다(정재화, 『클라우드 컴퓨팅』, 한국방송통신대학교출판문화원, 2020년, 155쪽). 다시 말해 엣지 컴퓨팅에서는 사용자와 가장 가까운 서버에서 데이터를 처리하고 서비스를 제공하여 서비스의 지연을 방지하는 것이다. 엣지 컴퓨팅에서 최종 사용자와 가장 밀접한 계층으로 엣지 단말 기기로 연산 처리를 할 수 있는 능력이 있는 경미한 수준의 장치를 프런트엔드(Front-End)라고 한다(정재화, 『클라우드 컴퓨팅』, 한국방송통신대학교출판문화원, 2020년, 160쪽).

API(Application Programming Interface)는 컴퓨터나 컴퓨터 프로그램 사이의 연결로서 일종의 소프트웨어 인터페이스를 말한다. 사용자 인터페이스(user interface)가 컴퓨터와 인간을 연결하는 것을 의미하는 것에 비하여, API는 컴퓨터나 소프트웨어를 서로 연결하는 것으로 최종 사용자(사람)가 직접 사용하기 위하여 고안된 것이 아니며, 대신 소프트웨어에 이를 통합하고자 하는 컴퓨터 프로그래머가 사용하도록 고안된 것이다(위키백과, https://ko.wikipedia.org/wiki/API, 2022. 8. 3. 방문).

여기서 '모바일 앱 프런트엔드 API 서버'라고 함은 이 사건 앱의 사용자 인터페이스

소 및 위도, 경도, 반경, 입실 날짜, 퇴실 날짜 등 API 서버로 정보를 호출하는 명령구문들을 알아내었다.

피고인들은 PC를 통하여 이 사건 앱의 API 서버 URL 주소에 마치 정상적인 이 사건 앱 이용자가 이 사건 앱을 이용하는 것처럼 이 사건 API 서버로 정보를 호출하는 명령구문을 입력하는 방식으로 접근하면서, 특정 위치로부터 일정 반경 내에 있는 숙박업소 정보를 모두 불러오는 기능[59]을 특정 URL에 탑재한 프로그램(이하 '이 사건 크롤링 프로그램'이라고 한다)을 개발하였다. 피고인들은 2016. 6. 1. 경부터 같은 해 10. 3.경까지 이 사건 크롤링 프로그램을 이용하여 일 1~2회 가량 이 사건 API 서버에 접근하여 제휴 숙박업소 업체명, 주소, 방 이름 등의 정보를 무단으로 복제하였다.

그 중간에 피해자 회사가 이 사건 크롤링 프로그램 이용 등으로 인한 대량 호출 신호를 감지하여 피고인들이 이용하는 아마존 웹 서비스 클라우드 서버의 IP 주소를 차단하자, 피고인들은 위 서버의 전원을 차단하였다가 다시 켜는 방식으로 IP 주소를 변경하는 등 서버의 설정을 변경하는 방법으로 계속하여 정보를 무단으로 복제하였다.

② 저작권법위반

피고인들은 위와 같은 방법으로 이 사건 크롤링 프로그램을 이용하여 피해자 회사의 데이터베이스 중 제휴 숙박업소 업체명, 주소, 방 이름, 원래금액, 할인금액, 업체주소, 입실시간, 퇴실시간, 날짜

와 직접 연결된 서버를 의미한다. 이 사건 앱은 프런트엔드 API 서버가 직접 연결되어 있고, 최종적인 데이터베이스 서버까지는 다른 API 서버들이 순차적으로 연결되어 있었던 것으로 보인다.

59) 피고인들의 회사를 기준으로 반경 1,000㎞ 이내에 있는 숙박업소 정보를 불러오도록 하였다. 이 사건 앱은 이용자의 위치로부터 7㎞ 또는 30㎞의 범위 내의 숙박업소 검색만 가능하도록 고정되어 있었다.

와 같은 상당한 부분을 무단으로 복제하여 데이터베이스제작자의 권리를 침해하였다.

③ 컴퓨터등장애업무방해

피고인들은 위와 같은 방법으로 이 사건 크롤링 프로그램을 이용하여 정보처리장치에 부정한 명령을 입력하여 장애가 발생하게 함으로써 피해자 회사의 숙박 예약에 관한 업무를 방해하였다.

2) 제1심[60]의 판단

제1심은 공소사실 전부를 유죄로 판단하였다. 정보통신망법위반의 점정보통신망위반에 대하여 제1심은 피고인 회사는 이 사건 API 서버에 대한 접근권한이 없었다고 보았는데, 그 근거는 다음과 같다. ① 이 사건 API 서버에 저장된 정보는 피해자 회사가 상당한 비용과 시간을 들여 수집, 보충, 갱신, 가공한 것으로, 이 사건 앱을 통해 사전에 허용된 검색조건에 따라 개별적·제한적으로만 공개되었고, 피고인 회사와 같은 경쟁업체에 유출될 경우에는 피해자 회사의 경쟁력이 저하되는 등의 손해가 발생할 수 있었다. ② 피해자 회사는 이 사건 API의 모듈, URL 주소 및 이 사건 API 서버로 정보를 호출하는 명령 구문들을 외부에 공개하지 않았고, PC 접속 신호를 처리하기 위한 별도의 서버를 운영하고 있었으며, 정상적인 이용자는 이 사건 앱을 통하지 않고서는 이 사건 API 서버에 접속할 수 없었다. ③ 당시 피해자 회사의 '서비스 이용약관' '제9조(회원, 이용자의 의무)' 제2항은, 리버스엔지니어링, 디컴파일, 디스어셈블 및 기타 일체의 가공행위를

60) 서울중앙지방법원 2020. 2. 21. 선고 2019고단1777 판결

통하여 서비스를 복제, 분해 또는 모방 기타 변형하는 행위(제8호), 자동 접속 프로그램 등을 사용하는 등 정상적인 용법과 다른 방법으로 서비스를 이용하여 회사의 서버에 부하를 일으켜 회사의 정상적인 서비스를 방해하는 행위(제9호) 등을 금지하고 있었는데, 위 조항의 표제에 '이용자(서비스에 접속하여 위 약관에 따라 회사가 제공하는 서비스를 이용하는 회원 및 비회원. 위 약관 제2조 제1항 제2호 참조)'가 규정되어 있고, 위 조항에서 규정된 금지 대상 행위의 내용과 성격 등에 비추어 볼 때, 각 항에 '회원'만 명시되어 있더라도 비회원도 적용대상에 포함된다.

나아가 제1심은 피고인들이 이 사건 크롤링 프로그램을 이용하여 피해자 회사의 정보통신망에 침입하였다고 보았는데, 그 근거는 다음과 같다. ① 피고인들은 이 사건 크롤링 프로그램을 통하여 복제한 피해자 회사의 정보를 피고인 회사의 영업 전략을 수립하는 등의 용도에 사용하였다. ② 피고인 A는 피고인 B에게 "한 시간마다 크롤링할 때 우리 회사 서버를 쓰는 것은 위험하다. 우리 서버 아이피가 노출되면 피해자 회사가 해킹이라고 신고할 수 있다."라는 취지로 말하였고, 이에 피고인 B는 이 사건 크롤링 프로그램을 아마존 웹서비스 클라우드로 이전하여 설치하였는데, 이에 비추어 보면, 경쟁업체인 피고인 회사가 이 사건 API 서버에 접속하는 것이 피해자 회사의 의사에 반한다는 것은 피고인들도 알고 있었다. ③ 피해자 회사는 이 사건 크롤링 프로그램 이용으로 인한 대량 호출 신호를 감지하고 피고인 회사가 이용하는 아마존 웹서비스 클라우드 서버의 IP 주소를 수차례 차단하였는데, 이에 피고인 회사는 서버의 전원을 차단하였다가 다시 켜는 방식으로 IP 주소를 변경하여 위 차단을 회피하면서 이 부분 공소사실을 계속하였다. ④ 피고인들이 복제한 피해자 회사의 정보가 모바

일 앱을 통해서도 얻을 수 있는 것이었다고 하여, 피고인들이 권한 없이 이 사건 API 서버에 침입한 사실에는 영향을 줄 수 없다.

저작권법위반에 대한 제1심의 유죄의 근거는 다음과 같다. ① 피고인들은 6개월여 동안 264회에 걸쳐 피해자 회사의 데이터베이스를 무단으로 복제하였는데, 이와 같은 각종 정보의 대량 복제는 피해자 회사 데이터베이스의 상당한 부분을 복제한 것에 해당한다. ② 피고인들은 피고인 회사가 피해자 회사와의 경쟁에서 우위를 점하기 위해서 반복적·조직적으로 피해자 회사의 데이터베이스를 무단으로 복제하였다. 따라서 설령 위 복제가 상당한 부분에 이르지 못하는 부분의 복제라고 하더라도, 이는 '반복적이거나 특정한 목적을 위하여 체계적으로 함으로써 위 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자인 피해자 회사의 이익을 부당하게 해치는 경우'에 해당한다.

컴퓨터등장애업무방해에 대한 제1심의 유죄의 근거는 다음과 같다. ① 이 사건 앱 이용자들은 자신의 위치로부터 7㎞ 또는 30㎞의 범위 내의 숙박업소만 검색이 가능하였다. ② 피고인들은 피해자 회사의 서비스를 이용하기 위해서가 아니라 피해자 회사와의 경쟁에서 우위를 점하기 위해 이 사건 API 서버에 접속하여 전국에 있는 모든 숙박업소 정보를 요청하여 대량의 정보 호출을 발생시켰다. ③ 이 사건 서버는 '이 사건 앱을 통한 피해자 회사의 숙박 예약서비스 이용'이라는 사용목적과 다른 기능을 하게 되어 정보처리에 장애가 발생하였고, 나아가 피해자 회사의 숙박 예약에 관한 업무 방해의 결과가 초래되었거나 적어도 그러한 결과가 초래될 위험이 발생하였다.

3) 항소심[61]의 판단

항소심은 제1심 판결을 취소하고 공소사실 전부를 무죄로 판단하였다. 먼저 정보통신망법위반에 대해서는, 피고인들이 이 사건 앱을 통하지 않고 PC를 통하여 이 사건 API 서버에 접속하였다거나 이 사건 크롤링 프로그램 또는 명령어의 확장 등을 통하여 정보를 수집하였다는 사정만으로 접근권한이 없거나 접근권한을 넘어 피해자의 정보통신망에 침입하였다고 할 수 없다고 보았다. 그 근거는 다음과 같다. ① 이 사건 앱이나 API에 접속하기 위해서 회원 가입이나 비밀번호가 따로 필요하지 않고, 피고인들이 사용한 패킷캡처 프로그램은 통상 이용되는 프로그램이다. ② 피해자 회사는 이 사건 API 서버의 URL을 적극적으로 공개하지는 않았지만, 모바일 앱의 특성상 모바일 화면에 URL이 나타나지 않을 뿐 피해자 회사가 의도적으로 이를 숨기려 한 것으로 보이지 않는다. 피해자 회사는 이 사건 앱이나 API 서버 접속을 금지하는 조처를 하지 않았다. URL은 일응 인터넷 주소로서 통상 숨길 이유가 없다. 피해자가 기술적 조처를 하지 않는 한 간단한 기술적 조작으로 쉽게 URL을 파악할 수 있다. ③ 피고인들이 크롤링을 통하여 가져간 정보들은 피해자 회사가 자신의 숙박 예약 영업을 위하여 이용자들에게 공개한 정보들이다. 크롤링이나 지역 범위 검색 명령어를 확장하지 않고 이 사건 앱을 통하더라도 다소 번거롭긴 하지만 크롤링한 것과 같은 종류와 양의 정보들을 가져올 수 있다. ④ 피고인들이 사용한 API 서버의 명령구문은 이 사건 API 서버가 허용한 것이다. 검색 지역 범위 등 그 기능이 모바일 앱에서 다소

61) 서울중앙지방법원 2021. 1. 13. 선고 2020노611 판결

제한되어 있다고 하더라도 이는 이용자의 편의를 위해서 설정되어 있을 뿐으로 보이고 이용자의 필요 때문에 그 검색 범위를 넓혔다는 사정만으로 접근권한을 넘었다고 볼 수 없다. ⑤ 약관은 문언상 피해자 회사의 회원들에게 적용되는 것으로 되어 있다. 피고인들의 패킷캡처나 크롤링이 약관에서 금지하고 있는 리버스엔지니어링 등에 해당한다고 보이지 않는다. ⑥ 피해자 회사는 피고인 회사 측의 반복적인 접근에 따른 대량 호출 신호를 감지하고 피고인 회사가 이용하는 아마존 웹서비스 클라우드 서버의 IP 주소를 수차례 차단하였고, 이에 피고인 회사는 서버의 전원을 차단하였다가 다시 켜는 방식으로 IP 주소를 변경하여 위 차단을 회피한 사실은 있으나, 이러한 사정만으로는 피해자 회사가 피고인들의 접근을 일률적으로 제한한 것으로 보기 어렵다.

저작권법위반에 대해 항소심은 피고인들이 수집한 데이터가 피해자 회사 데이터베이스의 전부 또는 상당한 부분에 해당하지 않고, 피고인들의 데이터베이스 복제가 데이터베이스의 통상적인 이용과 충돌하거나 피해자의 이익을 부당하게 해치는 경우에 해당하지 않는다고 취지로 판단하였다. 또한 피고인들의 목적은 피고인 회사의 영업전략 수립을 위한 것이고, 피해자 회사는 숙박 예약 영업의 선두 주자로서 그 영업을 활성화하는 데 상당한 투자와 노력과 시간을 들였는데, 후발 주자인 피고인 회사가 피해자 회사의 노력에 의한 결과에 편승하여 무형의 이익을 얻었다고 하더라도, 후발 주자의 경쟁 시장에 대한 정보 수집을 다른 특별한 사정없이 데이터베이스제작자의 권리를 침해한 것으로 구성할 수 없다고 하였다. 그 근거는 다음과 같다. ① 피고인들이 피해자 회사의 숙박업소 데이터베이스의 50여 항목 중

수집한 세부 항목은 '업체명, 주소, 지역, 방이름(마이룸 여부), 원래 금액, 할인금액, 날짜, zone, 이용시간, 예외사항, 입실시간, 퇴실시간, 대실가격, 숙박가격, 타입, 카테고리'로서, 한 번에 적게는 '업체명, 주소, 지역'의 3개 항목에 관한 정보를 수집하고 많게는 '업체명, 방이름, 원래금액, 할인금액, 업체주소, 입실시간, 퇴실시간, 날짜' 등과 같이 8개 항목에 관한 정보를 수집하였다. ② 위 8개 항목 중 숙박업소의 업체명, 업체주소, 지역, 타입 등은 이미 상당히 알려진 정보로서 수집에 상당한 비용이나 노력이 들 것으로 보이지 않고, 할인금액, 대실가격, 숙박가격 등은 피해자 회사의 서비스 상품에 대한 가격으로 영업을 위해서는 공개할 수밖에 없는 정보이다. 피고인들은 이 데이터들을 자신들의 영업 전략에 참조할 수 있을 뿐 그대로 쓸 수 있는 것은 아닌 것으로 보인다. ③ 이와 같은 정보들은 피해자 회사가 가진 숙박업소에 대한 50여개의 데이터 항목 중 이용자에게 공개한 정보로서 모바일 앱을 통해서도 확보할 수 있었다.

컴퓨터등장애업무방해에 대해 항소심이 무죄로 본 근거는 다음과 같다. ① 이 사건 API 서버의 사용 목적은 주어진 명령구문에 대응하는 숙박업소 정보를 반환하는 것으로서 피고인들은 이 사건 API 서버 본래의 목적에 따라 숙박업소의 정보를 전송받고자 위 서버의 명령구문들에 거리 정보 1,000km 등의 정보를 입력하여 전국 숙박업소의 정보를 전송받았으므로 이는 '허위의 정보 또는 부정한 명령의 입력'에 해당하지 않는다. ② 이 사건 앱에 접속하는 데 장애가 발생한 공소사실 기재 일자들은 순차로 토요일, 토요일, 일요일, 추석, 토요일로서 평일보다 접속이 훨씬 많을 때여서, 이는 자연 이용자 증가에 따른 것이었을 가능성이 있다. ③ 피고인들이 이 사건 크롤링 프

로그램을 이용한 목적은 피해자 회사의 데이터베이스를 확보하여 경쟁사인 피해자 회사의 사업 현황을 파악하고 이를 업무에 참고하고자 함에 있었으므로, 그 접속으로 인한 장애의 발생 가능성을 미필적으로나마 인식하고 용인하였다고 보기 어렵다.

2. 정보통신망 침입

가. 정보통신망법상 정보통신망침입죄

본죄는 정당한 접근권한 없이 또는 허용된 접근권한을 넘어 정보통신망에 침입하는 행위를 처벌한다. 이는 이용자의 신뢰 내지 그의 이익을 보호하기 위한 규정이 아니라 정보통신망 자체의 안정성과 그 정보의 신뢰성을 보호하기 위한 것이라고 할 것이므로, 위 규정에서 접근권한을 부여하거나 허용되는 범위를 설정하는 주체는 서비스제공자라 할 것이고, 따라서 서비스제공자로부터 권한을 부여받은 이용자가 아닌 제3자가 정보통신망에 접속한 경우 그에게 접근권한이 있는지 여부는 서비스제공자가 부여한 접근권한을 기준으로 판단하여야 한다.[62] 정당한 접근권한 없이 또는 허용된 접근권한을 넘어 정보통

62) 대법원 2005. 11. 25. 선고 2005도870 판결, 원심은 피고인의 직속상관인 공소외 소령이 소관업무에 관한 보고를 육군본부 등에 대신 발송할 수 있도록 하는 등 업무상 필요에 의해 아이디와 비밀번호를 예하 장교와 사병들에게 공지시킴에 따라 피고인도 이를 알게 되었음을 기화로, 2회에 걸쳐 피고인의 컴퓨터로 소령의 아이디와 비밀번호를 입력하여 육군 웹 메일과 핸드오피스 시스템에 접속한 후 소령 명의로 대장인 1군사령관에게 “군사령관 보아라. 네놈이 감히 ○대위(피고인을 지칭)를 징계하려고 했던 것에 피가 거꾸로 솟구친다. 장성 하나쯤 인사 처리하는 것은 문제도 아니다. 몸조심해라.”라는 취지의 이메일을 보냄으로써 정보통신망인 소령의 육군 웹 메일 및 핸드오피스 계정에 각 침입하였다는 공소사실에 대하여, 타인의 아이디와 비밀번호를 이용하여 타인의 정보통신망에 침입하는 행위도 정당한 접근권한 없이 또는 허용된 접근권한을 초과하여 정보통신망에 침입하는 행위에 포함된다는 이유로, 위 공소사실을 유죄라고 인정하였다.
대법원은 피고인의 상고를 기각하면서, 이용자가 자신의 아이디와 비밀번호를 알려주

신망 침입에 밀접한 행위를 하면 실행의 착수가 있고, 그와 같이 침입이 완료되면 기수에 이르렀다고 볼 것이고 정보통신망에 침입한 후 시스템상 프로그램을 실행하거나 데이터를 송·수신하는 등의 행위까지 나아갈 것을 요구하지 않는다.[63]

나. 접근권한 유무의 판단기준 - '객관적으로 드러난 사정'

접근권한에 대하여 '야놀자 판결'은 정보통신망에 대하여 서비스제공자가 접근권한을 제한하고 있는지 여부는 보호조치나 이용약관 등 객관적으로 드러난 여러 사정을 종합적으로 고려하여 신중하게 판단하여야 한다고 하였다.[64]

며 사용을 승낙하여 제3자로 하여금 정보통신망을 사용하도록 한 경우라고 하더라도, 그 제3자의 사용이 이용자의 사자 내지 사실행위를 대행하는 자에 불과할 뿐 이용자의 의도에 따라 이용자의 이익을 위하여 사용되는 경우와 같이 사회통념상 이용자가 직접 사용하는 것에 불과하거나, 서비스제공자가 이용자에게 제3자로 하여금 사용할 수 있도록 승낙하는 권한을 부여하였다고 볼 수 있거나 또는 서비스제공자에게 제3자로 하여금 사용하도록 한 사정을 고지하였다면 서비스제공자도 동의하였으리라고 추인되는 경우 등을 제외하고는, 원칙적으로 그 제3자에게는 정당한 접근권한이 없다고 봄이 상당하다고 하였다.

위 판결은 정보통신망침입에 있어서는 가장 대표적인 판결로 읽혀지고 있으나, 그 결론에 대해서는 나는 약간의 의문을 갖고 있는데, 그것은 다음과 같다. 위 판결은 '서비스제공자가 제3자로 하여금 사용하도록 한 사정을 고지하였다면 서비스제공자도 동의하였으리라고 추인되는 경우'를 들고 있는데, 그것이 접근권한을 판단하는 데에 과연 적절한 것인가 하는 것이다. 위 판결의 취지는 서비스제공자로부터 정당한 접근권한을 부여받아 아이디와 비밀번호까지 설정한 이용자로부터 아이디와 비밀번호를 제공받아 그것을 사용한 제3자가 이용자로부터 허락받은 범위 밖의 일을 한 경우에는 정보통신망침입이 된다는 것이다. 그런데 만약 위 사안에서 이용자인 공소외 소령이 자신의 자신의 아이디와 비밀번호를 이용하여 상급자에게 그를 모욕하는 취지의 메일을 보냈다면 그 경우에도 정보통신망침입을 인정할 것인가? 제3자가 상급자에게 위와 같은 메일을 보내는 것과 이용자가 메일을 보내는 것 사잉에 어떠한 법적 차이가 있는가? 만약 대법원의 취지가 이용자가 자신의 아이디와 비밀번호를 이용하여 상급자 또는 타인을 모욕하는 메일을 보낸 경우까지 정보통신망침입을 인정하려는 것이라면, 인터넷상의 게시판 또는 포탈에서 타인을 모욕하는 내용의 댓글을 게시한 경우에도 모욕죄 외에 정보통신망침입죄도 성립한다는 것인데, 이것은 불합리하게 처벌의 범위를 확장한 것이라고 생각한다.

63) 이창범, 황창근, 정필운, 『이론&실무 정보통신망법』, 박영사, 2021년, 266~267쪽

64) 이 부분 판시는 '야놀자 판결'에서 최초로 설시한 것이다. 이는 정보통신망 침입을 인

1) 객관적 상황

'야놀자 판결' 이전까지 대법원은 위와 같이 접근권한을 부여하거나 접근권한의 범위를 제한할 권한은 서비스제공자에게 있으므로 서비스제공자가 부여한 접근권한을 기준으로 접근권한의 유무를 판단하여야 한다고 하였다. '야놀자 판결'은 보다 구체적인 기준과 방법론적인 접근법을 제시하였다. 보호조치나 이용약관 등 객관적으로 드러난 여러 사정을 종합하여 고려하여 신중하게 판단하여야 한다는 것이다. 즉 서비스제공자가 접근을 막기 위한 기술적 보호조치를 취하였는지 아닌지, 이용약관에 구체적으로 이용자의 사용범위나 접근권한을 밝혔는지 아닌지 등과 같은 객관적 사정을 기준으로 신중하게 파악하여야 한다는 것이다.

그렇다면 이제 접근권한을 판단하는 데에 객관적으로 외부적으로 드러난 사정을 기준으로 하되, 정보통신서비스 제공자의 주관적 의사는 배제하고 판단하여야 하는 것인가? 그렇지는 않다고 본다. 우선 후술하는 바와 같이 이용약관에 특정 행위 또는 특정 이용자의 이용을 금지하는 내용이 기재되어 있고, 보호조치가 설정되어 있다는 사정만으로 그에 기초하여 접근권한을 해석할 수는 없다고 본다. 이용약관의 내용이나 보호조치의 여부는 접근권한을 해석하는 하나의 기준이 될 뿐이지 절대적인 기준이 될 수 없다.[65] 또한 이용약관이나 보

정할 것인가의 문제는 접근권한의 범위를 어떻게 인정할 것인지가 핵심이라는 점을 보여주고 있는 것이다. 이 부분은 후술하는 미국의 CFAA를 둘러싼 논의에도 마찬가지로 나타난다.

65) 그래서 '야놀자 판결'의 이 부분 판시에 접근권한을 '종합적으로' 고려하여 '신중하게' 판단하여야 한다고 설시하였다고 본다. 정보통신망법상 정보통신망 침입을 적용함에 있

호조치는 모두 서비스제공자가 정보통신망 침입 행위가 있기 전에 일방적으로 결정하는 것이어서 그 자체가 이미 서비스제공자의 주관적 의사가 반영된 것이다. 또한 정보통신망침입죄와 관련하여 대표적 판례로 언급되는 대법원 2005. 11. 25. 선고 2005도870 판결은 '서비스제공자에게 제3자로 하여금 사용하도록 한 사정을 고지하였다면 서비스제공자도 동의하였으리라고 추인되는 경우'라고 하여 서비스제공자의 가상적 의사를 판단기준의 하나로 들고 있기 때문이다. 위 판결은 '야놀자 판결'로 명시적으로 폐기된 바 없고, '야놀자 판결' 이후에도 여전히 정보통신망침입죄에 관한 대표적 판결로서 위상을 잃지 않았다.[66]

그러나 '야놀자 판결'이 객관적 사정을 강조한 것은 온당하다고 생각한다. 서비스제공자의 주관적 의사는 매우 불확정적이고, 이용자가 언제나 명시적·확정적으로 파악할 수 있는 것이 아니라는 점에서도 그렇다. 정보통신망의 사용이 불특정·다수의 사람들에게 일반적으로 열려있고, 인터넷이나 정보통신망의 발전이 그와 같은 개방성을 전제로 하여 왔다는 점에서도 그렇다.

2) 이용약관

여기서 이용약관은 서비스제공자와 이용자 사이에 체결된 약관을 의미한다. 대법원은 정보통신망법 제49조[67]와 관련하여 그 전제

어 적극적 태도를 경계한 것이다.

66) 그러나 위 판결에 대한 필자의 비판적 태도는 각주 61) 참조

67) 제49조(비밀 등의 보호) 누구든지 정보통신망에 의하여 처리·보관 또는 전송되는 타인의 정보를 훼손하거나 타인의 비밀을 침해·도용 또는 누설하여서는 아니 된다.

가 되는 정보의 귀속은 정보통신서비스 제공자에 의하여 그 접근권한이 부여되거나 허용된 자가 누구인지에 따라 정해져야 할 것이고, 이는 정보통신서비스 제공자가 정한 인터넷 온라인 게임 이용약관상 계정과 비밀번호 등의 관리책임 및 그 양도나 변경의 가부, 그에 필요한 절차와 방법 및 그 준수 여부, 이용약관에 따른 의무를 이행하지 않았을 경우 행해질 수 있는 조치내용, 캐릭터 및 아이템 등 게임 정보에 관한 이용약관상 소유관계 등 여러 사정을 종합적으로 고려하여야 한다[68]고 하여 '야놀자 판결' 이전에도 이용약관을 접근권한의 기준으로 하고 있었다.

그러나 이용약관을 접근권한의 판단기준으로 하는 데에는 유의하여야 할 점이 있다. 약관은 계약 당사자 중 일방의 의사에 의하여 작성되는 것이고, 그 상대방의 의사는 적극적으로 반영되지 않고 반영할 수도 없다. 또한 정보통신망서비스 제공자의 이용약관은 그것을 이용하려는 사람들에게는 일방적으로 의무를 승인 내지 동의할 것이 강제된다. 이용자가 이용약관을 승인하지 않으면 서비스 자체를 이용할 수가 없기 때문이다. 나아가 이용자는 이용약관을 승인하는 것만으로 대가 없이 무상으로 서비스를 제공받는 것이 아니다. 이용자는 서비스 이용과 관련하여 자신의 데이터를 서비스제공자에게 제공하고 있다. 이처럼 보이지 않는 교환관계가 데이터 이코노미의 기본 운영 원리가 되었다. 마지막으로 인터넷 또는 정보통신망은 개방성을 그 본질로 하고 있고 현대 사회에서는 그에 대한 접근이 사람의 기본권 또는 인권

68) 대법원 2010. 7. 22. 선고 2010도63 판결; 인터넷 온라인 게임의 이용자이자 계정 개설자 겸 명의자가 자신의 계정을 양도한 이후 그 계정을 현재 사용 중인 전전양수인이 설정해 둔 비밀번호를 변경하여 접속을 불가능하게 한 사안에서, 위 계정에 대한 구 정보통신망법상 정당한 접근권한자가 누구인지를 밝혀 같은 법 제49조의 위반 여부를 판단하였어야 함에도 그 인정사실만으로 유죄라고 판단한 원심판결에 법리오해 및 심리미진의 위법이 있다고 한 사례이다.

의 하나로까지 거론되고 있다. 그런 상황에서 서비스제공자인 기업[69]의 일방적인 의사만으로 접근권한이 결정될 수는 없음이 당연하다. 가령 이용약관에서 “오페라를 좋아하는 왼손잡이는 접근을 금지한다.”라고 정하였다고 하여 실제로 그러한 사람들의 접근이 금지된다고 볼 수 없는 것과 같다.[70]

또 이용약관은 그 문언에 기초하여 매우 엄격하게 해석되어야 하고, 해석상 논란의 여지가 있을 경우 이용자의 접근권한을 확장하는 방향으로 해석되어야 한다. 이용약관에서 접근권한의 제한을 정하더라도, 그것은 접근 그 자체를 금지해야 하고, 그 접근하려는 목적을 규제할 수는 없다. 다시 말해 이용자의 접근 목적에 따라 접근을 제한하는 약관은 그 자체로 매우 불명확하고 이용자의 이용권을 부당하게 제한할 염려가 있으므로 접근권한의 판단기준으로 사용될 수 없다고 본다.

3) 보호조치

정보통신망법 제48조는 침입을 방지하기 위한 기술적 보호조치를 침해하거나 훼손할 것을 요구하지 않는다. 대법원도 정보통신망법 제48조 제1항은 ‘정당한 접근권한 없이 또는 허용된 접근권한을 초과하여 정보통신망에 침입’하는 행위를 금지하고 있으므로, 정보통신망법은 그 보호조치에 대한 침해나 훼손이 수반되지 않더라도 부정

69) 대부분의 서비스제공자는 소위 대기업일 것이다.
70) Orin S. Kerr, “Norms of Computer Trespass”, Columbia Law Review V. 116, No 4.(https://columbialawreview.org/content/norms-of-computer-trespass/ 2022. 8. 6. 방문)

한 방법으로 타인의 식별부호(아이디와 비밀번호)를 이용하거나 보호조치에 따른 제한을 면할 수 있게 하는 부정한 명령을 입력하는 등의 방법으로 침입하는 행위도 금지한다고 보아야 한다[71]고 하였다. 구체적으로 피고인들이 서비스제공자인 SK 브로드밴드로부터 정당한 접근권한을 부여받지 않고, 장애처리용 전화기를 이용하여 SK 브로드밴드 주배전반의 통신포트에 연결한 후 피고인들의 휴대폰에 전화연결을 하는 부정한 방법으로 SK 브로드밴드의 정보통신망에 접속한 행위(그렇게 통신포트에 연결한 후 피고인들의 휴대폰에 전화연결하여 에스케이브로드밴드 가입자의 전화번호가 위 휴대폰에 착신되도록 하는 방법으로 전화번호를 수집하였다)는 정보통신망법 제48조 제1항에서 규정하는 정당한 접근권한 없이 정보통신망에 침입하는 행위에 해당한다고 할 것이고, 시스템의 정상적인 운영을 저해함이 없이 시스템에 접속하는 경우에는 '침입'에 해당하지 않는다는 피고인들의 이 부분 법리오해 주장은 이유 없다고 한 원심 판단[72]에 위법이 법리오해의 위법이 없다고 하였다.[73]

대법원은 '야놀자 판결' 이전에 이미 보호조치 여부도 접근권한의 판단기준으로 들고 있었다. 대법원은 '5급(행정) 공무원 공개경쟁 채용 제2차 시험' 합격자 발표와 관련하여 담당자는 발표 전날 합격자 명단을 사이버국가고시센터 홈페이지에 올리면서, 공식 발표 시각 이전에 위 홈페이지에 게시되지 않도록 설정기간을 미리 예약하여

71) 대법원 2005. 11. 25. 선고 2005도870 판결
72) 서울서부지방법원 2012. 4. 5. 선고 2012노10 판결
73) 대법원 2013. 10. 17. 선고 2012도4387 판결; KT 직원들이 전화·인터넷 등 아파트 전 세대의 통신회선이 집중되어 있는 아파트 통합통신장비실에 들어가 장애처리용 전화기를 SK 브로드밴드의 통신포트에 연결하여 개인용 휴대전화 등으로 전화를 걸어 고객의 발신번호가 표시되도록 하는 방법으로 SK 브로드밴드 가입 고객의 전화번호를 수집한 사안이다.

게시하였고, 웹브라우저 주소창에 공고 게시글의 주소를 입력해도 웹브라우저에 빈 화면이 나타나도록 소스 코드를 구성하였는데, 피고인은 위 시험에 응시한 여자친구의 합격 여부를 공식 발표 전에 미리 알기 위하여 위 사이트 게시판에 이미 게시된 '2016년도 외교관 후보자 선발시험 최종합격자 명단'의 첨부파일 주소[74]를 확인하고 이를 복사하여 인터넷 주소창에 붙여넣기 한 다음 그 주소의 파일 숫자 끝 번호를 계속 변경·입력하다가[75] 위 시험의 합격자 명단 파일의 파일 주소(URL)[76]를 입력하게 되었고 이에 의하여 합격자 명단 파일을 다운로드한 사안에서, 제2심 법원은 피고인이 위 합격자 명단 파일을 다운로드할 당시 사이버국가고시센터 사이트는 누구나 접속할 수 있었고, 위 합격자 명단 파일에 아무런 보호조치가 되어 있지 않았으며, 피고인은 누구나 자유롭게 이용할 수 있는 이 사건 사이트에 접속하여 우연히 알아낸 위 합격자 명단 파일의 주소를 웹브라우저 주소창에 입력하여 파일을 다운로드하였을 뿐이고, 웹브라우저 주소창에 직접 어떤 주소를 입력하는 행위가 금지되어 있다고 볼 수 없다는 이유로 정보통신망법 위반 혐의를 무죄로 판단한 원심을 그대로 인용하였다.[77]

---

74) http://www.gosi.go.kr/cmm/fms/FileDown.do?atchField=FILE_000000000121782&fileSn=1

75) 121782의 끝 번호 2 대신 6을 입력하였다.

76) http://www.gosi.go.kr/cmm/fms/FileDown.do?atchField=FILE_000000000121786&fileSn=1

77) 본문의 내용은 원심(서울서부지방법원 2017. 7. 20. 선고 2017노345 판결)의 내용이다. 위 판결에 대하여 대법원은 구체적인 내용의 설시 없이 상고기각(대법원 2017. 10. 12. 선고 2017도12758 판결)

위 판결은 여러모로 흥미로운 점이 많은데, '야놀자 판결'과의 관계에서 보자면, '야놀자 판결'에서 데이터를 수집한 방법이 크롤러라는 컴퓨터 프로그램에 의한 것이라면, 위 판결은 데이터를 수집하기 위하여 이미 알려진 URL의 일부 숫자를 일일이 바꿔가며 입력한 것이다. 전자가 자동화 방법이라면 후자는 수기 또는 사람이 일일이 웹브라우저에 가능한 URL을 입력한 것이다(나는 그것을 '손크롤링'이라고 부른다). 이 글을 쓰게

보호조치 역시 그것을 할 것인지 여부, 한다면 보호되는 정보와 그렇지 않은 정보의 구별, 기술적으로 서버의 계층구조에서 어느 단계에서 보호조치를 취할 것이지 등 보호조치와 관련된 모든 사항은 서비스제공자가 결정한다. 따라서 보호조치와 관련된 사정만으로 접근권한을 일의적으로 결정할 수는 없다. 위 대법원 2017. 10. 12. 선고 2017도12758 판결의 원심은 무죄 이유 중의 하나로 보호조치가 없는 사정을 들었지만, 대법원 2013. 10. 17. 선고 2012도4387 판결에서는 보호조치가 없었음에도 외부인이 접근할 수 없는 아파트 통합통신장비실에 들어가 접속한 사안에서 유죄를 인정하기도 하였다.

다. 미국의 CFAA

우리나라의 정보통신망법과 입법형식이 가장 유사하다고 평가되는 것은 미국의 컴퓨터 사기 및 남용 방지법(Computer Fraud and Abuse Act, CFAA, 이하 'CFAA'라고 한다)이다. CFAA는 미국에서 1986년 제정되었는데, 애초에는 연방정부의 정보통신망에 권한 없이 또는 권한을 넘어 침입하는 것을 처벌하기 위한 것이었으나, 수차례 개정을 거쳐 적용범위가 넓어졌다. 우리나라 정보통신망법 제48조와 가장 유사한 것은 §1030(a)(2) 이다.[78]

된 동기는 다음과 같은 호기심이었다. 사람이 일일이 손으로 하는 일이 무죄라면 그것을 컴퓨터 프로그램과 같은 자동화 도구를 사용하는 것도 무죄가 되어야 하는 아닌가? 만약 손으로 하는 일이 무죄이지만 그것을 자동화 도구를 사용하는 것은 유죄라면, 양자의 질적 차이는 무엇인가? 양자의 가치적·법적 평가가 달라지는 이유는 무엇인가? 자동화에도 여러 단계가 있을 터인데, 과연 자동화의 어느 지점에서 유죄와 무죄가 갈리는 것인가? 이는 크롤링뿐만 아니라 뒤에서 자동화된 도구를 사용하는 것의 형사책임과도 관련 있다. '야놀자 판결'은 위와 같은 호기심에 하나의 대답을 제시하였는데 그것은 저작권법위반에 관련된 두 번째 판시사항이다.

78) 번역은 조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등

18 U.S.C. § 1030(a)(2)[79]

누구든지 고의로 권한 없이 컴퓨터에 접근하거나 권한의 범위를 넘어 접근하여 다음과 같은 정보를 취득한 자는 본조 (c)에 규정된 것과 같이 처벌한다.

(A) 금융기관 또는 카드발급자가 보유하는 금융기록에 있는 정보, 또는 소비자신용보고회사가 보유하는 소비자에 대한 정보

(B) 연방정부기관이 보유하는 정보

(C) 보호되는 컴퓨터에 있는 정보

18 U.S.C. § 1030(a)(3)[80]

누구든지 고의로 권한 없이 연방정부기관이 배타적으로 사용하고 일반에 공개되지 아니한 컴퓨터에 접근한 자는 본조 (c)에 규정된 것과 같이 처벌한다. 다만 그 컴퓨터가 연방정부기

에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 128~132쪽에서 인용하였으나 그 중 일부는 필자가 수정하였다.

79) 원문은 다음과 같다.

(a) Whoever ... (2) intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains (A) information contained in a financial record of a financial institution, or of a card issuer as defined in section 1602(n) 1 of title 15, or contained in a file of a consumer reporting agency on a consumer, as such terms are defined in the Fair Credit Reporting Act (15 U.S.C. 1681 et seq.); (B) information from any department or agency of the United States; or (C) information from any protected computer; ... shall be punished as provided in subsection (c) of this section.

80) (a) Whoever ... (3) intentionally, without authorization to access any nonpublic computer of a department or agency of the United States, accesses such a computer of that department or agency that is exclusively for the use of the Government of the United States or, in the case of a computer not exclusively for such use, is used by or for the Government of the United States and such conduct affects that use by or for the Government of the United States; ... shall be punished as provided in subsection (c) of this section.

관이 배타적으로 사용하는 것이 아니어도, 일반에 공개되지 아니한 것이고 연방정부기관에 의하여 사용되거나 연방정부기관을 위하여 사용되는 것이라면, 그 사용에 영향을 주는 경우에 한하여 위와 같다.

1) 보호법익

CFAA는 무단침입죄(criminal tresspass)나 범죄목적침입죄(burglary)에 그 연원을 두고 있는데, 무단침입죄 등이 부동산에 대한 권리를 보호하는 것과 같이 CFAA는 컴퓨터에 대한 권리를 보호하는 것이다.[81] 그러나 대법원은 정보통신망법 제48조 제1항이 누구든지 정당한 접근권한 없이 또는 허용된 접근권한을 넘어 정보통신망에 침입하는 것을 금지하고 있는 것은 이용자의 신뢰 내지 그의 이익을 보호하기 위한 것이 아니라 정보통신망 자체의 안정성과 그 정보의 신뢰성을 보호하기 위한 것[82]이라고 하여, CFAA 와는 그 보호법익을 달리 보고 있다. CFAA의 입법취지를 우리식으로 표현하면 개인의 부동산에 대한 타인의 무단 침입을 방지하기 위한 것으로 그 개인적 법익에 대한 범죄인 반면, 정보통신망법 제48조 제1항은 정보통신망 자체의 안정성과 그 정보의 신뢰성이라서 일종의 사회적 법익에 대한 범죄로 보고 있는 것이다. 따라서 양자 사이에는 보호법익에 차이가 있어서 직접적 비교의 대상이 될 수는 없을 것이다. 그러나 그 입법형식에 있어서 CFAA 가 정보통신망법과 가장 유사한 것으로 알려져 있어서[83] 그 해석에 있어서 일응 참고의 대상이 될 수 있다고 본다.

---

81) 조성훈, “정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-”, 법조 Vol. 687, 2013. 12., 134-135쪽

82) 이는 대법원의 일관된 태도이다. 예컨대 대법원 2022. 5. 12. 선고 2021도1533 판결

2) 접근(access)

CFAA는 권한 없는 접근을 금지하면서도 '권한 없는'의 의미를 정의하지는 않고 있다. 먼저 여기서 '접근'은 좁게 해석하면 일반에 공개되지 않은 정보를 취득한 경우 또는 패스워드나 기술적 수단에 의해 보호되는 정보를 취득한 경우만을 가르킨다고 보는 견해와 넓게 해석하여 컴퓨터로 하여금 명령을 수행하게 하였다면 일응 접근에 해당한다고 보는 견해의 대립이 있을 수 있으나, 후자에 의하더라도 '권한'의 범위를 축소하여 부당한 구성요건의 포섭을 줄일수 있다는 점에서 양자의 입장은 실질적으로 다르지 않기 때문에 악성프로그램, 서비스거부공격 등의 다양한 행위태양을 포섭하기 위해서는 '접근'의 범위를 넓게 해석하는 것이 타당하다.[84]

3) 권한(authority)

'권한'은 어떻게 해석되는가. CFAA가 '권한 없는 접근'(access without authorization) 의 의미를 정의하고 있지 않고 있다. CFAA의 핵심은 권한이 무엇인지, 권한이 없는 경우 또는 권한이 있더라도 그 범위를 넘어서는 것이 어떤 것인지를 어떻게 해석할 것인지에 달려있다. 앞서 본 바와 같이 접근의 해석도 권한의 정도에 따라 탄력적

83) 조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 127쪽; 정현순, "크롤링에 의한 데이터 수집이 정보통신망 침입에 해당하는지 여부-대상판결: 대법원 2022. 5. 12. 선고 2021도1533 판결", 사법61호, 사법발전재단, 2022., 352-353쪽

84) 조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 143-144쪽

으로 해석될 수 있으므로 결국 해석의 중심은 권한이 될 것이니, CFAA 의 핵심은 권한을 어떻게 해석할 것인지에 달려있다.

미국에서의 해석론으로는 먼저 코드 기반 규제(code-based restriction)를 위반한 경우에는 권한 없는 접근에 해당한다고 본다.[85] 코드 기반 규제란 패스워드에 의하여만 접근을 허용하거나 사용권한을 각자 다르게 규정한 계정을 부여하는 등 프로그램을 기반으로 접근권한을 정하여 주는 방법인데, 이는 다른 사람의 사용자명과 패스워드를 도용하는 방법(false identification), 권한부여 프로그램을 오작동하게 함으로써 허용될 수 없는 권한을 부여하도록 하는 방법에 의하여 위반할 수 있다.[86] 여기서 '코드'는 컴퓨터 시스템의 설계도라고 할 수 있는 프로그램 코드상 접근이 허용되었는지 여부를 따지는 것으로 보인다. 코드기반규제를 기반으로 접근이 제한되어 있다면 이는 물리적으로 불가능한 것이어서, 정상적인 시스템의 접근 또는 컴퓨터 시스템 설계자나 관리자가 상정하고 있는 방법으로는 접근 자체가 불가능한 것이어서 이를 우회하기 위한 기술적인 별도의 조치가 있어야 할 것이다.

다음으로 권한부여에 있어 가장 느슨한 방식인 정보서비스 이용약관(terms of service)을 위반한 경우에는 권한 없는 접근에 해당하지 않는다고 본다.[87] 예컨대 소셜 네트워크 서비스의 이용 약관을

85) Patricia L. Bellia, Defending Cyberproperty, 79 N.Y.U. L. REV. 2253 (2004); Orin S. Kerr, Cybercrime's Scope: Interpreting Access and Authorization in Computer Misuse Statutes, 78 N.Y.U. L. REV. 1648-1660 (2003)(조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 144쪽 각주 71에서 재인용함)

86) 조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 144-145쪽

위반하여 자신의 나이와 같은 인적사항을 허위로 꾸민 행위는 접근 없는 권한으로 볼 수 없다는 것인데, 만약 이러한 경우까지 접근 없는 권한을 인정한다면 '막연하기 때문에 무효(void for vagueness)' 원칙에 반하게 되고 나아가 CFAA 규정을 극도로 모호하게 만들게 되고 또한 서비스 운영자에게 형사처벌의 대상을 정할 수 있는 실질적인 권한을 부여하게 되기 때문이다.[88] 결국 해석의 문제는 권한부여에 있어 엄격한 기준인 코드기반규제와 느슨한 방식인 정보서비스 이용약관 사이에 있는 행위 태양 중에서 어디까지 CFAA 에 포섭할 수 있을지에 달려있다.

4) CFAA 관련 사례

웹 크롤링에 대하여 CFAA의 적용 여부가 문제되었던 미국의 사례들은 최근 20년간 미국의 법원의 태도는 4단계로 진행되었다고 평가된다.[89] 1단계는 2000년대 초반부터 2000년대 후반까지로 이 시기에는 대체로 CFAA를 넓게 해석하여, 코드 기반 규제인지 이용약관 규제인지를 불문하고 웹 사이트 운영자가 제한한 접근권한을 위반한

87) Patricia L. Bellia, Defending Cyberproperty, 79 N.Y.U. L. REV. 2253 (2004); Orin S. Kerr, Cybercrime's Scope: Interpreting Access and Authorization in Computer Misuse Statutes, 78 N.Y.U. L. REV. 1648-1660 (2003)(조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 144쪽 각주 71에서 재인용함)

88) United States v. Drew, 259 F.R.D. (C.D. Cal. 2009)(조성훈, "정보통신망 침입에 대한 연구 - 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로-", 법조 Vol. 687, 2013. 12., 145쪽에서 재인용함)

89) Andrew Sellars, 'Twenty Years of Web Scraping and the Computer Fraud and Abuse Act', 24 boston University Journal of Science & Technology Law 379-381 (2018); Jennie E. Christensen, The Demise of The CFAA In Data Scraping Cases, 34 Notre Dame Journal of Law, Ethics & Public Policy (2020), 5-8 (정현순, "크롤링에 의한 데이터 수집이 정보통신망 침입에 해당하는지 여부-대상판결: 대법원 2022. 5. 12. 선고 2021도1533 판결", 사법61호, 사법발전재단, 2022., 355쪽 각주27에서 재인용, 이하의 내용은 355-364쪽의 내용을 재정리한 것이다)

경우에는 CFAA 위반을 인정하였다. 대표적인 판결로는 Southwest 항공이 실제 또는 잠재적인 고객을 대상으로 항공노선의 운임 정보를 자신의 웹 사이트에서 제공하고 있었고, 그 이용 약관에서 크롤링을 금지한다고 명시하고 있었는데, Outtask가 위 웹 사이트의 정보를 크롤링한 사안으로, 법원은 Southwest 의 이용 약관과 Southwest 가 직접 Outtask 에 접근권한이 없음을 알렸다는 것을 근거로 위 크롤링은 CFAA 위반에 해당한다 하였다.[90]

2단계는 2009년부터 2013년까지는 이 시기에는 가급적 CFAA를 좁게 해석하려고 시도하였는데, 이 기간 동안 '단지 계약에 기초한 제한(contract based controls)'보다는 '기술적 제한(technical controls)'이 있는지와 '단순한 이용 제한(mere use restrictions'보다는 '접근 제한(access restrictions)'이 있는지를 중점을 두었다. 이 시기의 대표적인 판결은 LVRC 에 근무하는 Brekka 가 자신의 개인 사업을 위해 LVRC 의 자료들을 자신의 개인 컴퓨터로 이메일로 전송한 사안에서, 고용주가 종업원에게 어떤 제한을 부과하면서 회사 컴퓨터를 이용하도록 승인한 경우, 법원은 그 종업원이 그러한 제한을 위반하였더라도 여전히 그 회사 컴퓨터를 사용할 수 있는 권한을 가지고 있으며, 접근권한이 없다고 보기 위해서는 그 사람이 어떠한 목적이든 컴퓨터를 이용할 수 있다는 허락을 받지 못하였거나, 고용주가 그 허락을 취소하였으나 그 종업원이 계속하여 그 컴퓨터를 사용한 경우여야 한다고 하였다. Brekka 는 위 둘 중 어느 경우에도 해당하지 않으므로 CFAA에 해당하지 않는다고 한 것[91]이다.

---

90) Southwest Airline Co. v. FareChase, Inc., 318 F.Supp.2d 435 (N.D.Tex. 2004)
91) LVRC Holdings LCC v. Brekka, 581 F.3d 1127 (9th Cir. 2009)

3단계는 2013년부터 2018년까지로 그 전 시기의 법원이 접근(access)과 이용(use) 사이의 구별에 집중한 것과 달리, 위 LVRC 사건에서 컴퓨터 소유자가 컴퓨터에 대한 접근 허락을 취소하였음에도 계속 그 컴퓨터를 이용하는 행위를 접근 권한이 없는 경우로 본 점에 집중하여 이를 취소 이론(revocation theory)이라고 부르며 접근 권한 유무의 판단 기준으로 삼았다. 대표적인 사안으로는 Craigslist가 이용자들이 광고를 올리거나 검색할 수 있는 서비스를 제공하는데, 3Taps가 공개된 Craigslist의 사이트의 게시물들을 크롤링하여 자신이 운영하는 사이트에 게시한 사건이다. Craigslit는 먼저 3Taps에게 크롤링의 중단 및 자신의 사이트에 대한 접근을 금지하는 내용의 통지를 한 다음, 3Taps와 관련된 IP 주소들의 접근을 차단하는 조치를 하였는데, 3Taps는 다른 IP 주소를 사용하거나 프록시(Proxy) 서버를 사용하는 방법으로 기술적 조치를 우회하여 계속하여 위 사이트를 크롤링하였다. 법원은 접근권한을 부여하거나 취소할 수 있는 권한이 시스템 관리자에게 있으므로 공개 사이트라 하더라도 Craigslist는 3Taps에 대해서만 접근권한 부여를 취소할 수 있고, 3Taps에게 웹 사이트 이용 정책만이 아닌 명백한 통지와 IP 차단을 통하여 웹 사이트 접근 제한을 알렸으며, 특정 목적으로 웹 사이트 이용을 제한하는 것과 선별적으로 해당 사이트 접근을 완전히 제한하는 것은 다르다는 것을 근거로 3Taps의 CFAA 위반을 인정하였다.[92)]

4단계는 2018년 이후로서 누구나 이용할 수 있는 웹 사이트에 대한 웹 크롤링은 공익적 관점, 수정헌법 제1조, 웹 크롤링과 웹 브라우징의 기술적 유사성, CFAA가 크롤링과 같은 행위를 규제하기 위하

92) Craigslit, Inc. v. #TAPS, Inc., 964 F. Supp. 2d 1178 (N.D.Cal. 2013)

여 입법된 것이 아니라는 점을 근거로 CFAA 위반 여부를 엄격히 판단하고 있다. 대표적 사건으로는 Van Buren v. United States 사건과 HiQ Labs v. LinkedIn 사건이 있다. 국내에서는 특히 후자가 소위 '링크드인 사건'으로 많이 알려져 있다.

① Van Buren v. United States[93)]

경찰관인 Van Buren이 개인적인 목적으로 순찰차에 설치된 컴퓨터를 이용하여 자동차번호관련 기록을 열람한 사안(그와 같은 열람은 경찰의 내부 지침상 금지되어 있다)에서, 미국연방대법원은 CFAA § 1030(a)(2)의 허용된 접근권한의 범위를 넘은 접근으로 볼 수 없다고 판단하였다. 연방대법원은 상황 또는 목적에 따라 CFAA 위반 여부를 판단하게 되면 불합리하게 형사처벌의 범위를 확장할 수 있으므로 CFAA 위반 여부를 판단하는 기준은 컴퓨터의 해당 영역에 접근할 수 있는 권한의 유무가 되어야 한다고 하였다. 근로계약의 사용자는 통상 컴퓨터나 전자 기기들은 오직 사업 목적을 위해서만 사용되어야 한다고 하므로 사용자의 태도를 기준으로 하면 근로자가 사적인 이메일을 보내는 것 까지 CFAA 위반이 될 수 있는데, 그와 같은 해석론을 따르면 인터넷 웹 사이트의 관리자가 상황에 기초한 접근제한(circumstance-based access restrictions)을 하였는데 이용자가 그 제한을 넘어 사용하는 경우도 CFAA 위반이 될 수 있다는 것이다. 특히 연방대법원은 위 판결에서 CFAA 위반 여부는 문이 열려 있는지 닫혀 있는지의 문제(gates-up-or-down inquiry)인데 이는 어떤 사람이 컴퓨터 시스템이나 시스템 내의 특정 영역에 접근할 수 있는지가 기준이 되므로 Van Buren이 정당한 자격의 경찰관으로서 그

93) Van Buren v. United States, 141 S. Ct. 1648 (2021)

의 직무상 순찰차에 설치된 컴퓨터로 자동차번호관련 기록을 열람할 권한이 있었다면 Van Buren에게 그 목적과 무관하게 그 컴퓨터는 열린 문이었다는 것이다. 그와 같은 기록 열람이 경찰서 내부 지침에 위반된 접근이었다는 사정은 CFAA의 포섭 여부에 기준으로 삼지 않겠다는 태도이다. 즉 CFAA는 온라인에서 약속을 어긴 것을 범죄로 보거나 계약 규정에 위반하는 것을 범죄로 보는 것이 아니고, CFAA에서 중요한 것은 문(gates)이다.[94] 위와 같은 연방대법원의 태도를 보면 결국 접근권한의 문제를 가능성으로 환원해서 보는 것 같은 인상을 준다. 즉 행위자의 주관적 의도나 목적은 물론 컴퓨터 시스템 관리자와 행위자 사이의 계약 또는 약정의 내용도 차치하고 그가 갖고 있는 지위에 비추어 그와 같은 접근 자체가 허용되어 있는지 여부만을 문제삼고 있는 것이다. 만약 그와 같은 접근 자체가 컴퓨터 시스템 불가능한 것이었는데, 이를 우회하기 위한 다른 별도의 조작이나 행동(예컨대 제3자의 아이디와 패스워드의 사용 또는 악성 프로그램의 사용)을 하였다면 CFAA에 해당한다고 볼 수 있다.

② HiQ Labs v. LinkedIn[95]

구인구직 플랫폼인 링크드인(LinkedIn)이 공개적으로 게시하고 있는 회원들의 이름, 직업, 경력, 자격 등의 정보를 경쟁사인 하이큐 랩스(HiQ Labs)가 크롤링하여 가져가는 것에 대하여 링크드인이 하이큐 랩스의 IP 접속을 차단하자, 하이큐 랩스가 링크드인의 그와 같은 접속차단의 금지명령(preliminary injunction)을 구한 사안이다.

94) Orin S. Kerr, The Wupreme Court Reins In the CFAA in Van Buren , LAWFARE, 2021, 6. 9. article(정현순, “크롤링에 의한 데이터 수집이 정보통신망 침입에 해당하는지 여부-대상판결: 대법원 2022. 5. 12. 선고 2021도1533 판결”, 사법61호, 사법발전재단, 2022., 354쪽 각주 23에서 재인용

95) HiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985 (9th Cir. 2021)

앞의 사례와는 절차의 성격이 다른데, 우리 법제로 설명하자면 앞의 사례는 형사재판의 본안에서 유무죄가 문제가 된 것이었고, 이 사례는 일종의 가처분신청사건에서 경쟁법적 문제가 쟁점이 된 것이다. 미국 제9연방항소법원은 하이큐 랩스의 청구를 인용하여 링크드인의 크롤링 금지 조치에 대하여 금지명령을 발령하는 것이 타당하다는 취지로 판단하였는데 그 근거는 다음과 같다. CFAA의 권한 없는 접근은 접근이 일반적으로 가능하지 않고, 허가가 일상적으로 요구되는 상황을 전제한다. 여기서 권한은 적극적 개념이고 접근이 특별히 인정되거나 허가된 사람들에게만 허용됨을 의미한다. CFAA의 입법취지는 해킹과 같은 컴퓨터에 대한 의도적 침입을 방지하기 위한 것이므로 '성과도용법(misappropriation statue)'[96]으로 해석되면 안 된다. Van Buren 판결에서 말한 문이 열려 있는지 닫혀 있는지의 문제(gates-up-or-down inquiry)는 이용자가 권한을 가지고 있다면 컴퓨터 시스템에 접근할 수 있고 권한을 가진 이용자는 코드에 기초하든 계약이나 정책에 기초한 것이든 접근에 대한 제한들로부터 영향을 받지 않는다는 의미이다. 웹 크롤링은 정보를 수집하는 일반적인 방법이고, 인터넷을 통한 정보의 자유로운 흐름을 극대화하여 공중의 이익에 기여하는 반면, 대량의 이용자 정보를 축적하고 있는 회사들로 하여금 누가 데이터를 크롤링 할 수 있는지를 결정할 수 있게 한다면 정보의 이용 방법에 대한 과도한 통제권을 부여하는 것이 될 것이다. 크롤링 방지조치를 금지하는 명령이 내려지더라도 링크드인이 악의적 행위자들에 대한 기술적 수단(가령 robot.txt)을 실시할 수 있다.

---

96) 이는 우리 법제상 부정경쟁방지 및 영업비밀보호에 관한 법률 제2조 제1호 (파)목의 '그 밖에 타인의 상당한 투자나 노력으로 만들어진 성과 등을 공정한 상거래 관행이나 경쟁질서에 반하는 방법으로 자신의 영업을 위하여 무단으로 사용함으로써 타인의 경제적 이익을 침해하는 행위'를 의미한다. 자세한 것은 Ⅳ.에서 서술한다.

라. 유럽연합의 사이버범죄협약(Convention on Cybercrime)

유럽연합은 2001년 사이버범죄협약[97]을 제정하였는데, 독일, 프랑스와 같은 유럽의 주요 국가들과 미국, 일본, 캐나다 등 65개국이 가입하였다.[98] 사이버범죄협약 제2조는 컴퓨터 시스템에 대한 불법접속을 금지하고 있다.[99]

> 제2조(불법접속)
> 각 당사국은 권한 없이 고의로 컴퓨터 시스템의 일부 또는 전부에 접속하는 행위를 국내법상 범죄로 하는데 필요한 입법 및 그 밖의 조치를 취해야 한다. 당사국은 컴퓨터 데이터를 획득할 목적이나 기타 부정한 목적으로 또는 다른 컴퓨터 시스템과 연결되어 있는 컴퓨터 시스템과 관련하여 보호조치를 침해할 것을 그 요건으로 할 수 있다.[100]

위 협약 제2조는 권한 없이 컴퓨터 시스템의 일부 또는 전부에 고의적으로 접속하는 행위, 컴퓨터 데이터를 획득하거나 기타 부정한 목

---

97) 2001. 11. 23. 헝가리 부다페스트에서 위 협약의 서명식이 개최되었기 때문에 '부다페스트 협약'으로도 불린다.
98) 현재 우라니라는 아직 가입하지 않았다.
99) 아래의 번역은 기본적으로 유민종, "사이버범죄협약과 국내 법제의 양립 가능성 연구", 서울대학교 석사학위논문, 2019., 58쪽의 것을 기본으로 하여 필자가 일부 수정하였다.
100) 원문은 아래와 같다.
Article 2 – Illegal access
Each Party shall adopt such legislative and other measures as man be necessary to establish as criminal offences under its domestic law, when committed intentionally, the access to the whole or any part of a computer system without right. A Party may require that the offence be committed by infringing security measures, with the intent of obtaining computer data or other dishonest intent, or in relation to a computer system that is connected to another computer system.

적 또는 다른 컴퓨터 시스템과 관련하여 그와 연결된 컴퓨터 시스템의 보호조치를 침해 하는 것을 처벌하도록 하고 있다. 권한 없이 컴퓨터 시스템에 접속하는 행위에는 별도의 목적을 필요로 하지 않고, 보호조치를 침해하는 것은 컴퓨터 데이터 획득 등의 목적을 필요로 하고 있다. 이 규정은 컴퓨터 시스템 및 그 데이터 자체의 기밀성, 무결성 및 효용성을 해킹으로부터 보호하는 데에 목적이 있다.[101] 위 협약에 첨부된 주석서(Explanatory Report)에 따르면 공중에게 자유롭게 공개된 접근을 허용하는 컴퓨터 시스템에 접근하는 행위는 정당한 것으로서 범죄화하지 않으며, 일상적으로 적용되는 통신프로토콜이나 프로그램에 제공되는 기본 도구를 이용하는 것은 권한 없는 것에 해당하지 않는다고 한다.[102]

마. 웹 크롤링과 정보통신망 침입

'야놀자 판결'은 피해자 회사에 의하여 피고인들의 이 사건 크롤링 프로그램에 의한 이 사건 API 서버로의 접근이 제한되었다고 보기 어렵다고 하였다. 구체적인 근거는 다음과 같다. 구체적으로 ①은 보호조치에 관한 것이고, ②, ③은 이용약관에 관한 것이다. ① 이 사건 API 서버의 URL이나 명령구문은 피해자 회사가 적극적으로 공개하지는 않았지만 누구라도 간단한 기술조작이나 통상 사용되는 소위 '패킷 캡쳐 프로그램' 등을 통해 쉽게 알아낼 수 있는 정보이다. 일반 이용자들은 이 사건 앱을 통해 API 서버에 회원 가입 후 또는 회원 가입

101) 유민종, "사이버범죄협약과 국내 법제의 양립 가능성 연구", 서울대학교 석사학위논문, 2019., 58쪽

102) 정현순, "크롤링에 의한 데이터 수집이 정보통신망 침입에 해당하는지 여부-대상판결: 대법원 2022. 5. 12. 선고 2021도1533 판결", 사법61호, 사법발전재단, 2022., 365쪽

없이 자유롭게 접근할 수 있었고, 이 사건 앱이나 API 서버로의 접근을 막는 별도의 보호조치는 없었다.[103] ② 피해자 회사의 이 사건 앱 서비스 이용약관에서 '이용자는 회사를 이용함으로써 얻은 정보를 회사의 사전 승낙 없이 복제, 송신, 출판, 배포, 방송 등 기타 방법에 의하여 영리 목적으로 이용하거나 제3자에게 이용하게 하여서는 안 된다'고 정하고 있으나, 이는 이 사건 앱 또는 API 서버로부터 취득한 정보의 이용을 제한하는 내용일 뿐, 이에 대한 접근을 제한하는 내용으로 볼 수 없다. ③ 위 이용약관에서 회원에 대하여 '자동접속프로그램 등을 사용하여 회사의 서버에 부하를 일으켜 회사의 정상적인 서비스를 방해하는 행위'를 금지하고 있기는 하지만, 위 약관 규정을 회원가입을 하지 않은 이용자들에게 적용할 수 있는 근거를 찾기 어렵고, 규정의 내용 또한 접근권한 자체를 제한하는 것으로 볼 수 없어 위와 같은 약관상의 규정만으로 API 서버에 대한 접근권한이 객관적으로 제한되었다고 보기 어렵다.

'야놀자 판결'에 대해서 정보통신망침입은 인정하지 않으면서도 피해자 회사가 대량 호출 신호를 감지하고 피고인들이 사용하는 서버의 IP 주소를 차단한 이후에도 피고인들이 서버의 전원을 껐다가 다시

---

103) 제1심 법원도 인정하였듯이 피해자 회사는 이 사건 API 서버에 SSL(secure socket layer)을 사용하지 않았다.
SSL은 웹브라우저와 웹서버 간에 주고받는 데이터의 안전성과 보안성을 위하여 사용하는 통신 프로토콜이다. 구체적으로 웹브라우저가 웹서버에 SSL을 적용한 웹 페이지를 요청하면 웹서버는 웹브라우저에 공개키와 인증서를 발행하고, 웹브라우저는 증명서가 신뢰할 수 있는지 확인한 후 대칭키 방식의 암호통신을 위한 세션키를 생성한다. 웹브라우저에서 데이터는 세션키로, 세션키는 공개키로 암호화하여 암호화된 세션키와 데이터를 웹서버에 전송하면, 웹서버는 암호화된 세션키를 비밀키로 복호화하고 암호화된 데이터는 세션키로 복호화하는 방식으로 통신한다(손진곤, 길준민, 정보통신망, 한국방송통신대학교출판문화원, 2021년, 303쪽).
이에 대하여 제1심법원은 피해자 회사가 SSL을 사용하지 않은 이유는 구버전 안드로이드를 사용하는 일부 이용자들이 이 사건 앱을 사용할 수 없게 되기 때문이라고 하였다.

켜는 방식으로 IP 주소를 변경하여 이 사건 API 서버에 접속한 것은 정보통신망의 안정성을 해하는 것이므로 정보통신망침입이 인정된다는 견해도 있다.[104] 그러나 그와 같은 견해는 받아들일 수 없다. 위와 같은 피해자 회사의 IP 주소 차단으로 피해자 회사가 의도한 정보통신망의 안정성은 이미 달성된 것이다. 피고인들은 그와 같이 차단된 IP 주소를 통한 접속을 한 것이 아니고, 단순히 자신들이 이용하는 서버의 전원을 다시 켰을 뿐이다. 통상 IP 주소는 동적으로 할당(dynamic allocation)되고 정보통신망은 이것을 전제하고 설계되므로, 동적으로 할당되어 변경된 IP 주소를 통하여 접속한 행위는 정보통신망의 안정성과는 무관하다. 또한 피해자 회사의 위 IP 주소 차단이 피고인들의 접근권한을 제한했다고 볼 수도 없다. 앞서 본 바와 같이 접근권한의 제한은 객관적이고 명백한 사정에 따라 판단되어야 한다. 피해자 회사의 입장에서 차단된 IP 주소를 통한 접근을 금지하였다고 볼 수는 있을지언정 다른 IP 주소를 통한 접근까지 배제하였다고 볼 수 없다.[105]

3. 데이터베이스제작자의 권리 침해

가. 저작권법상 데이터베이스와 데이터베이스제작자의 권리

1) 데이터베이스

104) 최상진, “경쟁사의 무단 크롤링에 대한 법적 대응방안에 관한 연구”, Law & Technology 제17권 제1호, 2021년, 34쪽

105) 이 부분은 '야놀자 판결'에서 언급되지 않았다. 다만 원심 판결에서는 위와 같은 IP 주소의 변경으로 정보통신망침입이 된다고 할 수 없다고 하였다.

저작권법은 데이터베이스에 대해서 정의하고 있다. 데이터베이스는 '소재를 체계적으로 배열 또는 구성한 편집물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것[106]을 말한다. 종래에는 별도로 데이터베이스를 보호의 대상으로 삼지 않았고 대신 편집저작물의 하나로서 취급하였다. 따라서 구 저작권법(2003. 5. 27. 법률 제6881호로 개정되기 전의 것) 제6조 제1항의 편집저작물에 '논문, 수치, 도형 기타 자료의 집합물로서 이를 정보처리장치를 이용하여 검색할 수 있도록 체계적으로 구성한 편집물'을 포함하는 대신 그 소재의 선택 또는 배열이 창작성이 있는 것에 한하여 편집저작물의 한 형태로서 보호하고 있었다. 2003년 저자저작권법에 의하여 보호되는 데이터베이스에 창작성을 요건으로 하였으나, 2003년 저작권법 개정으로 별도의 데이터베이스를 정의하는 규정을 두었고 창작성이 없는 데이터베이스도 저작권법에 의하여 보호되도록 하였다. 이는 창작성이 없는 데이터베이스에 대한 배타적 권리를 부여하는 것이 비록 사회 전체의 정보공유를 저해하여 정보화 사회에 역행한다는 부정적 측면을 제기할 수 있으나, 제작 및 갱신 등에 상당한 투자가 있을 경우 그 투자 노력에 대해서도 법적으로 보호할 가치가 있고, 이는 데이터베이스 제작을 활성화시킨다는 점에서 궁극적으로 지식정보사회의 요구에 부합하는 조치라는 필요가 있다는 것이 그 개정의 이유였다.[107] 이는 창작성이 없는 데이터베이스를 불법행위법으로 보호하는 미국, 일본과 달리 EU 데이터베이스 지침을 따라 데이터베이스제작자의 '독자적 권리(sui generis right)'를 규정하였다는 점에서 특징이 있다.[108]

---

106) 저작권법 제2조 제19호

107) 문화관광위원회, 저작권법중개정법률안 심사보고서, 2003. 2. 7.(한지영, 데이터베이스의 법적 보호에 관한 연구, 서울대학교 박사학위논문, 2005, 23쪽에서 재인용함)

108) 김종호, "빅데이터의 재산법상 보호 가능성에 관한 법적 고찰", 법이론실무연구 제9

저작권법의 규정에 따른 데이터베이스는 다음의 요건을 갖추어야 한다.[109]첫째, 편집물이어야 한다. 즉 저작물이나 부호·부호·문자·음·영상 그 밖의 형태의 자료(이하 "소재"라 한다)의 집합물[110] 즉 편집물이어야 한다. 편집물을 이루는 소재는 저작물일 수도 있고 저작물이 아닐 수도 있다. 둘째 소재를 체계적으로 배열 또는 구성한 것이어야 한다. 단순히 소재를 모아 놓은 것만으로는 부족하고 소재의 배열 또는 구성에 있어서 체계성이 인정되어야 하는 것이다. 창작성은 요구되지 않으므로 누가 하더라도 동일하게 할 수 밖에 없는 방법으로 배열 또는 구성하여도 데이터베이스에 해당한다. 셋째, 개별적으로 소재에 접근하거나 소재를 검색할 수 있도록 되어 있어야 한다. 원하는 정보를 검색하기 위하여 데이터베이스 전체를 처음부터 끝까지 다 살펴보아야 할 필요 없이 손쉽게 그 정보를 찾아낼 수 있도록 구성되어 있어야 한다. 둘째, 셋째 요건은 데이터베이스의 주된 가치가 보호의 근거가 되는 자료 검색의 편리성을 위하여 요구되는 것이다. 다만 데이터베이스의 제작·갱신·검증·보충 또는 운영에 이용되는 컴퓨터프로그램과 무선 또는 유선통신을 기술적으로 가능하게 하기 위하여 제작되거나 갱신·검증 또는 보충 등이 되는 데이터베이스는 보호의 대상에서 제외된다.[111] 데이터베이스의 제작·갱신·검증·보충 또는 운영에 이용되는 컴퓨터프로그램은 데이터베이스와 결합되어 이용되지만, 데이터베이스와는 별도의 저작물로 보호된다.[112] 데이터베이스로서 창작성이 있는 것은 데이터베이스인 동시에 편집저작물에 해당하므로

권 제4호, 2021년, 404쪽
109) 오승종, 『저작권법 강의』, 박영사, 2018년, 553쪽
110) 저작권법 제2조 제17호
111) 저작권법 제92조
112) 허희성, 『신 저작권법 축조개설(하)』, 명문프리컴, 2011, 474쪽

저작권법 제6조에 의한 저작물로서의 보호와 저작권법 제4장에 의한 데이터베이스제작자의 보호를 중첩적으로 받을 수 있음은 당연하다.[113]

2) 데이터베이스제작자의 권리

데이터베이스제작자는 데이터베이스의 제작 또는 그 소재의 갱신·검증 또는 보충에 인적 또는 물적으로 상당한 투자를 한 자를 말한다.[114] 여기서 상당한 투자에 대한 구체적인 기준은 없으나 특정한 종류의 데이터베이스와 그것을 구성하는 정보의 사회·경제적 중요성, 데이터 수집·조직의 용이성, 그 보호가 시장에 미치는 영향, 그리고 그 데이터베이스를 구성하는 개별적인 정보에 대한 접근의 중요성 등을 고려하여야 하며,[115] 양적인 상당성뿐만 아니라 질적인 상당성도 감안하여 데이터베이스의 내용 중 핵심정보를 지니고 있는지, 가장 전략적이고 최신의 정보인지, 데이터베이스 제작과정에서 문제된 부분을 수집, 검증, 표현하기 위하여 한 투자의 정도등이 주요 기준이 될 것이다.[116]

데이터베이스제작자는 그의 데이터베이스의 전부 또는 상당한 부분을 복제·배포·방송 또는 전송(이하 '복제 등'이라고 한다)할 권리를 가진다.[117] 데이터베이스의 개별 소재는 데이터베이스의 상당한

113) 염호준, "유럽연합의 데이터베이스 보호에 관한 지침과 최근의 동향", 세계의 언론법 제 2006년 상권(통권 제19호), 16쪽
114) 저작권법 제2조 제20호
115) 오승종, 『저작권법 강의』, 556쪽
116) 이해완, 『저작권법(제4판)』, 박영사, 2019, 1015쪽
117) 저작권법 제93조 제1항

부분으로 보지 않으나, 데이터베이스의 개별 소재 또는 상당한 부분에 이르지 못하는 부분의 복제 등의 행위라고 하더라도 그러한 행위를 반복적 또는 특정한 목적을 위하여 체계적으로 함으로써 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에는 해당 데이터베이스의 상당한 부분의 복제 등을 하는 행위로 본다.[118] 여기서 데이터베이스의 통상적인 이용과 충돌한다는 것은 그 데이터베이스와 시장에서 경쟁하는 관계에 놓인다거나 현재적 또는 잠재적 시장에 영향을 미칠 정도에 이른 경우를 의미하며 데이터베이스자작자의 이익을 부당하게 해치는 경우는 그 데이터베이스제작자가 통상적으로 이용허락을 함으로써 얻을 수 있는 이익을 부당하게 상실하게 되는 경우를 의미한다.[119]

또한 데이터베이스제작자에 대한 보호는 데이터베이스의 구성부분이 되는 소재 그 자체에는 미치지 않으므로 데이터베이스를 구성하는 개별 소재 자체를 데이터베이스제작자의 허락 없이 복제·배포·방송 또는 전송하는 것은 데이터베이스제작자의 권리를 침해하는 것이 아니다. 예컨대 인명편 전화번호부에서 소수 특정인의 전화번호 몇 개를 찾아 유인물로 만들어 배포하는 것은 데이터베이스인 전화번호부 제작자의 복제권이나 배포권을 침해하는 행위가 아니다.[120] 그러나 데이터베이스의 개별 소재 또는 그 상당한 부분에 이르지 못하는 부분의 복제라 하더라도 반복적이거나 특정한 목적을 위하여 체계적으로 함으로써 해당 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에는 해당 데이터베이

118) 저작권법 제93조 제2항
119) 오승종, 『저작권법 강의』, 박영사, 558쪽
120) 오승종, 『저작권법 강의』, 박영사, 557쪽

스의 상당한 부분의 복제로 보게 되어(같은 법 제93조 제2항 단서), 같은 법에 따라 보호되는 데이터베이스제작자의 권리를 침해하는 행위로 평가될 수 있다.[121]

나. 데이터베이스제작자 권리 침해의 판단기준

저작권법의 규정을 정리하면, 데이터베이스제작자의 권리 침해가 인정되는 것은 ①데이터베이스의 전부를 복제·배포·방송 또는 전송한 경우, ②데이터베이스의 전부가 아니더라도 그 상당한 부분을 복제·배포·방송 또는 전송한 경우, ③데이터베이스의 개별 소재를 복제·배포·방송 또는 전송하였더라도 그것이 반복적이거나 특정한 목적으로 체계적으로 이루어졌고, 그로 인하여 해당 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 권리를 부당히 침해한 경우와 같이 세 가지로 정리된다. ①은 데이터베이스 전부를 대상으로 한 것이므로 해석상 논란의 여지가 별로 없다.[122] 문제는 ②에서 '상당한 부분'의 해석과 ③에서 '반복성/체계성'과 '통상적인 이용/부당한 침해'의 해석이다.

'야놀자 판결'은 위 부분에 대해서 다음과 같이 판시하였다. 저작권법이 데이터베이스제작자의 권리를 인정하고 데이터베이스의 전부를

121) 다만 데이터베이스제작자의 권리는 저작권법상 저작권자의 권리가 제한되는 일반적인 경우(제23조, 제28 내지 37조)에 더하여 교육·학술 또는 연구를 위하여 이용하는 경우(다만, 영리를 목적으로 하는 경우는 제외), 시사보도를 위하여 이용하는 경우에는 누구든지 데이터베이스의 전부 또는 그 상당한 부분을 복제·배포·방송·또는 전송할 수 있다. 그러나 데이터베이스의 통상적인 이용과 저촉되는 경우에는 그러하지 아니하다(저작권법 제94조 제2항).

122) 데이터베이스 전부를 복제하였다면, 이 경우는 기술상 크롤링이라고 할 수는 없고 미러링에 해당한다고 봄이 옳은 기술의 적용이라고 본다.

대상으로 복제·배포·방송 또는 전송이 있는 경우 형사처벌의 대상으로 하면서도, 위 ②, ③의 경우에는 형사처벌에 제한을 가하고 있는 취지에 대해서, “지식정보사회의 진전으로 데이터베이스에 대한 수요가 급증함에 따라 창작성의 유무를 구분하지 않고 데이터베이스를 제작하거나 그 갱신·검증 또는 보충을 위하여 상당한 투자를 한 자에 대하여는 일정기간 해당 데이터베이스의 복제 등 권리를 부여하면서도, 그로 인해 정보공유를 저해하여 정보화 사회에 역행하고 경쟁을 오히려 제한하게 되는 부정적 측면을 방지하기 위하여 단순히 데이터베이스의 개별 소재의 복제 등이나 상당한 부분에 이르지 못한 부분의 복제 등만으로는 데이터베이스제작자의 권리가 침해되지 않는다고 규정”하였다고 보았다. 나아가 “데이터베이스제작자의 권리가 침해되었다고 하기 위해서는 데이터베이스제작자의 허락 없이 데이터베이스의 전부 또는 상당한 부분의 복제 등이 되어야 하는데, 여기서 상당한 부분의 복제 등에 해당하는지를 판단할 때는 양적인 측면만이 아니라 질적인 측면도 함께 고려하여야 한다. 양적으로 상당한 부분인지 여부는 복제 등이 된 부분을 전체 데이터베이스의 규모와 비교하여 판단하여야 하며, 질적으로 상당한 부분인지 여부는 복제 등이 된 부분에 포함되어 있는 개별 소재 자체의 가치나 그 개별 소재의 생산에 들어간 투자가 아니라 데이터베이스제작자가 그 복제 등이 된 부분의 제작 또는 그 소재의 갱신·검증 또는 보충에 인적 또는 물적으로 상당한 투자를 하였는지를 기준으로 제반 사정에 비추어 판단하여야 한다”고 하면서, “데이터베이스의 개별 소재 또는 상당한 부분에 이르지 못하는 부분의 반복적이거나 특정한 목적을 위한 체계적 복제 등에 의한 데이터베이스제작자의 권리 침해는 데이터베이스의 개별 소재 또는 상당하지 않은 부분에 대한 반복적이고 체계적인 복제 등으로 결국 상당한 부분

의 복제 등을 한 것과 같은 결과를 발생하게 한 경우에 한하여 인정" 되어야 한다고 하였다. 이하에서는 우선 '야놀자 판결' 이전 데이터베이스제작자의 권리 침해와 관련된 민·형사상 사례들을 살펴본다.

1) 데이터베이스제작자의 권리 침해 여부에 관한 민사사례

① 서울고등법원 2016. 12. 15. 선고 2015나2074198 판결[123]

원고는 이용자들이 특정한 주제어에 관한 게시물을 자유롭게 작성하여 게시하거나 이미 게시된 내용을 자유롭게 수정하는 방식으로 웹 사이트(리그베다위키, rigvedawiki)를 운영하였는데 피고가 원고 웹 사이트에 집적된 자료 전부를 미러링(mirroring) 방식으로 복제하여 피고 운영 웹 사이트('엔하위키 미러')에 게재한 사안이다. 원고는 데이터베이스제작자의 권리 침해를 주장했는데, 제1심 법원은 원고 사이트에 집적된 20만 건 이상에 이르는 게시물 대부분은 각 이용자가 작성하거나 이를 수정하여 온 것으로 보이는 점 등을 들어 원고가 데이터베이스제작자에 해당한다고 볼 수 없다고 판시하였다. 그러나 항소심은 1심과 달리 (i) 원고가 개인적으로 운영하던 상식 사전 사이트 데이터를 12,000~13,000개 항목으로 정리하고, 그 중 100여개를 선별하여 구성의 체계성, 개별 소재의 접근성, 검색 기능 등을 테스트한 후 나머지 데이터 10,000여 개를 엔하위키 게시판에 모두 업데이트하였고, 그 후에도 이용자들의 요구에 따라 목차 구조와 페이지 작성 양식 등을 만들거나 '최근 변경내역' 등을 도입하여 개별 자료에의 접근성을 높인 점, (ii) 검색 엔진을 변경하면서 사이트 환경에 맞추어

123) 대법원 2017. 4. 13. 선고 2017다204315 판결(심리불속행)로 확정되었다. 일명 '리그베다위키' 사건이다.

검색 기능 등을 추가 개발하는 등 접근 및 검색 가능성을 높여 플랫폼을 구축한 점, (iii) 원고 명의의 서버 4대를 운영하면서 약 16,000명의 가입자와 25만 개의 위키 문서가 있는 원고 사이트를 유지·관리한 점 등을 종합하여 원고의 데이터베이스제작자로서의 지위를 인정하였고, 피고가 원고 사이트를 미러링하는 방법으로 원고 사이트의 미러 사이트를 개설·운영하였으므로 원고의 복제권·전송권 침해가 인정된다고 하였다.

② 서울고등법원 2017. 4. 6. 선고 2016나2019365 판결[124)]

피고가 경쟁업체인 원고 운영 웹 사이트('잡코리아')에 게시된 '채용정보 전부'를 원고의 동의 없이 수집 프로그램을 통해 크롤링하여 피고의 웹 사이트('사람인') 서버에 저장한 후 구인업체의 동의를 받는 절차를 거쳐 피고 웹 사이트에 게재한 사안이다. 제1심은 원고 웹 사이트의 HTML 소스에 창작성이 없어 저작물에 해당하지 않으므로, 저작권 침해는 인정하지 않았으나, 부정경쟁방지법 제2조 제1호 (차)목의 부정경쟁행위에 해당한다고 판단하였다. 그러나 항소심은 원고 웹 사이트는 여러 구인업체의 채용정보를 체계적으로 배열하여 수록함으로써 이용자가 원고 웹 사이트로부터 각종 채용정보를 각 분류별로 자신이 원하는 기준에 따라 모아서 열람하거나 검색할 수 있도록 한 데이터베이스에 해당하며, 원고는 원고 웹 사이트를 제작 및 소재의 갱신·검증 또는 보충을 위하여 인적 또는 물적으로 상당한 투자를 한 자로서 원고 사이트에 대한 데이터베이스제작자에 해당한다고 하고, 피고가 별도의 마케팅 비용 등의 지출 없이 피고의 영업에 이용할 목적으로 반복적, 체계적으로 원고 데이터베이스의 채용정보 부분

124) 대법원 2017. 8. 24. 선고 2017다224395 판결(심리불속행)로 확정되었다. 일명 '잡코리아' 사건이다.

을 복제함으로써 데이터베이스제작자인 원고의 이익을 부당하게 해쳤다"고 판단하였다. 또한 원고의 데이터베이스제작자의 권리 침해를 인정하는 이상, 원고가 선택적으로 구하는 원고 웹 사이트의 HTML 소스에 대한 전송권, 복제권, 2차적 저작물작성권 침해 주장이나 부정경쟁방지법 제2조 제1호 (차)목에 관한 주장은 모두 따로 판단하지 않는다고 하였다.

③ 서울고등법원 2016. 5. 12. 선고 2015나2004441 판결[125)]

원고는 인터넷을 통한 부동산경매에 관한 정보제공서비스를 제공하는데, 피고들이 원고의 동의 없이 원고 인터넷 홈페이지에 게시된 각 경매사건의 매각금액에 관한 정보를 무단 복제하여 자신들이 운영하는 인터넷 홈페이지에 게시한 사안이다. 원고가 전국 경매법원에서 경매정보를 수집하는 조사원들에게 보수를 지급하고 그들로부터 경매정보를 제공받는 방식으로 자료를 모아 위 데이터베이스에 추가하여 옴으로써 원고 인터넷 홈페이지의 경매정보의 제작 및 소재의 갱신에 인적 또는 물적으로 상당한 투자를 하였으므로 데이터베이스제작자에 해당한다고 보았다. 또한 피고들은 원고의 데이터베이스의 개별 소재에 해당하는 각 경매사건의 매각대금에 관한 정보들의 복제물을 반복적으로 영업목적을 위하여 체계적으로 자신들의 인터넷 홈페이지에 게시함으로써 원고의 데이터베이스제자갖의 권리를 침해하였다고 보았다. eklaks 피고들도 조사원들에게 일정한 보수를 지급하고 경매정보를 제공받아왔던점, 한 달에 법원에서 약 17,000건 내지 20,000건 정도의 경매사건이 진행되는데, 6개월 동안 피고들이 운영하는 각 인터넷 홈페이지 경매정보에 원고가 의도적으로 수정·입력한

125) 대법원 2016. 9. 9. 선고 2016다223227 판결(심리불속행)로 확정되었다.

매각대금과 동일한 매각대금이 반영된 건은 피고별로 29건에서 57건에 불과한 점을 근거로 피고들에게 고의·과실이 인정되지 않는다고 하여 손해배상책임을 부정하였다.

④ '야놀자 판결'의 사안과 비교

위 세 사건에서 법원은 정보의 분류, 검색 프로그램 개발, 서버의 관리 등을 투자의 상당성 판단에서 중요한 고려요소로 보고 있으며, 피고가 원고의 정보 모두를 복제하였기 때문에 특별히 '상당한 복제'인지 여부를 고려할 필요는 없었던 것으로 보인다. '야놀자 판결'의 사안은 다음과 같은 점에서 위 두 사건과 사실관계를 달리한다.[126] (i) 웹 사이트 자체가 아닌 프런드엔드(front-end)의 API 서버에의 접근이 문제가 되었는데 그 API 서버 접근에 대한 특별한 보호조치가 없었다. (ii) 크롤링을 통해 수집된 정보가 데이터베이스를 구성하는 50여개 항목 중 3 내지 8개로 양적 상당성을 인정하기 어려운 상황이었다. (iii) 그 정보들이 대부분 이용자들에게 공개된 것이거나 쉽게 수집할 수 있는 것이어서 질적 상당성 역시 인정하기 어려운 측면이 있었다. (iv) 약관상 크롤링 제한의 내용이 모호했고 피고인들이 약관의 적용을 받는 회원이 아니었다.

2) 데이터베이스제작자의 권리 침해 여부에 관한 형사사례[127]

126) 물론 가장 큰 차이는 위 두 사건은 손해배상을 구하는 민사절차에서 이루어진 것이고, '야놀자 판결'은 직접 크롤링에 가담한 피고인 회사의 임직원들에 대한 형사책임을 묻는 형사절차에서 이루어졌다는 점이다.

127) 그 행위태양이 크롤링을 통한 것인 경우에 한정한다. 모두 하급심의 판결이고 상소하지 않아 그대로 확정되었다. 앞서 언급한 바와 같이 크롤링을 통한 정보수집에 대한 형사책임을 밝힌 대법원의 판결은 '야놀자 판결'이 최초이다.

① 서울남부지방법원 2021. 9. 8. 선고 2021고단588 판결(확정)

직업소개사업을 하는 피고인들이 수집한 정보를 판매하기 위한 목적으로, 관련 웹 사이트[128]에 접속하여[129] '인재검색 서칭 서비스' 상품을 결제한 뒤 ATS(Advanced Target Search) 프로그램[130]을 이용하여 '잡코리아'의 이력서 페이지에 약 69,000회 반복적으로 접속하여 구직자들의 사진, 이름, 연락처, 주소 일부를 제외한 나머지 정보(성별, 나이, 학력, 교육, 경력, 자격증 등)을 수집, 복제하였다. 잡코리아는 구직자정보의 경우 직무, 경력, 근무지역, 학력, 나이, 거주지역 등으로 분류하여 유형별로 '인재검색'이 가능하지만, 구직자정보를 모두 블라인드 처리하여 사이트 이용자가 '인재검색 서칭 서비스' 상품에 가입하여야 구직자의 사진, 이름, 연락처, 주소 일부를 제외한 나머지 정보(성별, 나이, 학력, 교육, 경력, 자격증 등)을 제공한다.

법원은 저작권법 제136조 제2항 제3호, 제93조(데이터베이스 제작자 권리 침해의 점), 정보통신망법 제71조 제1항 제9호, 제48조 제1항, 형법 제30조(정보통신망 침입의 점), 을 모두 유죄로 인정하였다. 구체적으로 저작권법위반에 대하여는, 피고인들은 영리를 목적으로 자동 정보 수집 프로그램인 이 사건 프로그램을 이용하여 판시 각 정보를 반복적·체계적으로 수집하여 대량으로 복제하였고, 이는 피해자 회사들의 데이터베이스의 통상적인 이용과 충돌하고 피해자 회사

128) '잡코리아'와 '사람인'의 각 웹 사이트였다.
129) 타인 또는 허무인의 계정을 통하여 접속한 것으로 보이는데, 타인인지 허무인 인지는 알 수 없다.
130) 웹 사이트 화면에 보이는 텍스트를 전체 선택(Ctrl+A)하여 복사(Ctrl+C), 붙여넣기(Ctrl+V)를 한 뒤 일정한 형태의 텍스트를 추출하여 파일로 저장하는 구동을 반복적으로 실행하는 프로그램이다.

들의 이익을 부당하게 해치는 결과를 초래하는 것으로 보이는바, 피고인들이 복제한 판시 각 정보가 피해자 회사들의 데이터베이스 그 자체는 아니고, 데이터베이스를 구성하는 소재에 해당하는 것이라 하더라도, 피고인들의 행위는 피해자 회사들의 데이터베이스의 상당한 부분의 복제에 해당하고, 이로써 저작권법에 따라 보호되는 피해자 회사들의 데이터베이스제작자로서의 권리가 침해되었다고 판단된다. 한편, '크롤링'이란 일반적으로 웹 사이트, 하이퍼링크, 데이터, 정보 자원을 자동화된 방법으로 수집, 분류, 저장하는 것을 의미하는 것으로, 법률상의 용어는 아닌바,[131] 크롤링에 해당하는지 여부에 따라 피고인들의 복제행위의 위법성이나 데이터베이스제작자에 대한 권리 침해 여부가 달라지지 않는다. 이에 더하여 법원은 이 사건 프로그램은 일반적인 크롤러 프로그램에 비하여 속도가 느리고 기능이 단순한 편인 것으로 보이기는 하나, 그 구동 과정 대부분은 자동적으로 이루어지는 것[132]으로 피고인 A는 이 사건 프로그램을 이용하여 피해자 회사들의 각 인터넷 사이트에서 피고인들이 원하는 정보를 선택적으로 추출한 후 배열을 달리하여 텍스트 파일로 저장하였다는 점과 피고인들이 복제한 정보에는 기간별로 가격이 다른 유료 서비스 상품을 구매한 경우에만 열람이 허용되는 정보(경력 중 직장명 등)가 포함되어 있었고, 피고인들은 복제한 정보를 구글드라이브에 업로드하고 헤드헌터들

131) 법률상의 개념이 아니라고 하여 판결에서 무작정 배척하는 것은 일반적인 관점에서 바람직하지 않은 경우가 더 많아고 생각한다. 실제로 위 판결에서도 크롤링이 법률상의 개념이 아니라고 하여 그에 관한 심리가 미진하였다고는 보이지 않는다. 좀 더 '세련된' 설시도 가능했을 것이다.

132) 이와 같은 설시는 자동화된 도구를 사용할 경우 데이터베이스제작자의 권리침해에 대해서는 유죄로 판단할 여지가 커진다는 것을 전제로 하고 있다. 이는 앞서 각주 8)에서 언급한 나의 호기심과 관련된다. 위 판결과 '야놀자 판결'이 그와 같은 호기심에 하나의 답을 제시하고 있기는 하지만, 특히 **자동화된 알고리즘을 사용한 정보통신망의 이용**에 대해서 형법적으로 어떻게 평가해야 하는가에 대한 문제에 대해서는 이제 겨우 첫발을 떼었다고 할 정도로 풀어내야할 쟁점이 많아 보인다.

에게 유상 제공하여 헤드헌터들이 보유한 데이터베이스와 통합 검색이 가능하도록 하는 영리사업을 계획하고 이를 위하여 판시와 같이 각 정보를 복제하였다는 점을 유죄의 근거로 들기도 하였다.

정보통신망침입에 대해서는, 계정에 대하여 허용한 접근권한을 넘어 피해자 회사들의 정보통신망인 각 인터넷 사이트에 침입하였다고 인정되고, 피고인 A가 피고인 B 등의 계정과 비밀번호를 이용하였다 하더라도 위 죄의 성립에는 영향이 없다고 하였다. 그 근거는 다음과 같다. (ⅰ) 피해자 회사들은 인터넷 사이트 회원을 서치펌회원 또는 기업회원, 개인회원 등으로 구분하여 관리하고 있고, 헤드헌터 등 이용자들이 피해자 회사들의 인터넷 사이트에서 구직자 정보를 검색·열람하기 위해서는 서치펌회원 또는 기업회원으로 가입 또는 이용계약을 체결하고 아이디를 부여받아야 하는 것으로 보이고, 피해자 회사들은 구직자 정보 중 핵심적인 부분은 유료 서비스 상품을 구매한 경우에만 열람할 수 있도록 회원의 서비스 이용에 제한을 두기도 하였다. (ⅱ) 피고인이 동의한 것으로 보이는 피해자 회사들의 서치펌회원 이용약관 또는 기업회원 약관에 의하면, 회원은 유·무료로 개인회원이 등록한 이력서를 검색할 수 있으나, 그 서비스를 직원채용 및 채용중개 또는 구인 구직 이외의 목적으로 사용해서는 안 되고, 서비스를 이용하여 얻은 정보를 피해자 회사의 사전 동의 없이 복제 등 방법으로 사용하거나 타인에게 제공할 수 없으며, 사이트를 통해 열람한 이력서 정보를 피해자 회사 및 당사자의 허락 없이 재배포할 수 없고, 인터넷 사이트의 정보 및 서비스를 이용한 영리 행위를 해서는 안 되는바, 이는 피해자 회사들의 영업상의 핵심 이익 등을 보호하기 위한 것으로 보인다. (ⅲ) 피고인들은 이 사건 범행 당시 피해자 회사들의 인터넷

사이트에서 복제한 정보(유료 서비스 상품을 구매한 경우에만 열람할 수 있는 정보 포함)를 헤드헌터들에게 유료로 제공하는 사업을 계획하였는바, 그와 같은 목적으로 피고인들이 판시와 같이 피해자 회사 운영의 인터넷 사이트와 이력서 페이지에 반복적으로 접속한 것을 단순히 직원채용 및 채용중개 또는 구인 구직 활동을 위한 허용된 목적 범위 내의 서비스 이용이라고 볼 수는 없고, 피해자 회사들의 인터넷 사이트에서 정보를 복제하여 제3자에게 판매 또는 유상 제공하기 위한 것이었다고 보이는데, 이는 피해자 회사들이 금지하고 있는 정보의 무단 제3자 제공 또는 재배포 및 정보를 이용한 영리 행위를 목적으로 하는 것이다.

② 대전지방법원 2018. 11. 30. 선고 2018고정909 판결 (확정)

자동정보 수집 프로그램(크롤링 프로그램)인 '수집로봇'을 이용하여 피해자가 반려견 분양사이트와 반려묘 분양 사이트에서 각 반려동물의 사진, 특징, 분양자의 인적사항(성명, 거주지역, 휴대전화번호)이 기재된 데이터베이스를 복제하여 그 데이터베이스를 자신이 운영하는 반려동물 분양사이트에 게시한 사안이다. 법원은 별도의 설시 없이 저작권법 제136조 제2항 제3호, 제93조 제1항(데이터베이스제작자 권리 침해의 점), 개인정보보호법 제72조 제2호, 제59조 제1호(개인정보 부정취득의 점)를 적용하여 유죄로 인정하였다.

③ 대전지방법원 2018. 9. 28. 선고 2018고단2373 판결 (확정)

피고인이 파이썬을 이용하여 패스워드 검증절차 없이 자신이

다니던 대학교 사이버교육시스템에 접속하여 재학생 및 교직원의 개인정보를 수집할 수 있는 크롤링 프로그램을 제작한 후, 자신의 휴대전화로 도서관 와이파이를 통해 위 크롤링 프로그램을 실행하여 재학생 및 교직원의 학번(사번), 이름, 단과대학정보(교직원부서), 전화번호, 이메일, 주소가 저장된 개인정보를 복제한 사안이다.[133] 법원은 별도의 설시 없이 정보통신망법 제71조 제1항 제1호, 제48조 제1항(정보통신망 침해의 점), 제71조 제1항 제11호, 제49조(타인의 비밀 침해의 점)[134]를 적용하여 유죄로 인정하였다.

다. 데이터베이스제작자의 권리 침해와 웹 크롤링

'야놀자 판결'에 의하면 데이터베이스제작자의 권리 침해가 인정되는 경우는 다음과 같다. ① 데이터베이스의 전부 또는 상당한 부분의 복제가 있는 경우인데, 여기서 전부 복제는 판단의 어려움이 없다. 문제는 '상당한 부분'의 복제는 어떠한 정도에 이르러야 하는가 인데, 전체 데이터베이스와 복제된 부분의 규모를 비교하는 양적인 측면과 데이터베이스제작자가 그 복제된 부분의 제작 또는 그 소재의 갱신·검증 또는 보충에 상당한 투자를 하였는지를 확인하는 질적인 측면을 모두 고려하여야 한다. ③ 상당한 부분의 복제에 이르지 못하였어도 반복적이거나 특정한 목적을 위하여 체계적으로 복제하여 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에는 '상당한 부분의 복제'에 이르렀다고 간주되나, 이때에도 데이터베이스의 개별 소재 또는 상당하지 않은 부분에 대한 반복적이고 체계적인 복제로 결국 상당한 부분의 복제 등을 한 것과 같은

133) 피고인은 위와 같이 복제한 개인정보를 자신이 만든 웹 사이트에 게재하였다.
134) 이 사안은 저작권법위반으로 기소되지는 않았다.

결과를 발생하게 한 경우에 한하여 인정하여야 한다. 결국 데이터베이스제작자의 권리 침해의 가장 핵심 요건은 ‘상당한 부분’의 해석에 달려 있게 된다. '야놀자 판결'은 데이터베이스제작자의 권리 침해를 인정하지 않은 항소심의 판단을 그대로 인용하였다. ①피고인들이 피해자 회사의 API 서버로부터 수집한 정보들은 피해자 회사의 숙박업소 관련 데이터베이스의 일부에 해당한다. ②위 정보들은 이미 상당히 알려진 정보로서 그 수집에 상당한 비용이나 노력이 들었을 것으로 보이지 않거나 이미 공개되어 있어 이 사건 앱을 통해서도 확보할 수 있었던 것이고, 데이터베이스의 갱신 등에 관한 자료가 없다. ③이러한 피고인들의 데이터베이스 복제가 피해자 회사의 해당 데이터베이스의 통상적인 이용과 충돌하거나 피해자의 이익을 부당하게 해치는 경우에 해당한다고 보기 어렵다.

4 컴퓨터등장애업무방해

가. 컴퓨터등장애업무방해죄

1) 허위의 정보 또는 부정한 명령

형법 제314조 제2항은 컴퓨터와 같은 정보처리장치에 허위의 정보 또는 부정한 명령을 입력하여 정보처리에 장애를 일으켜 업무를 방해한 자를 처벌하는 규정이다. 여기서 ‘허위의 정보’는 객관적으로 진실에 반하는 내용의 정보를 의미하고, ‘부정한 명령’은 객관적으로 정당하지 않은 명령이나 사무처리 과정에서 주어서는 안 되는 명령을 의미하는데, 구체적으로 관리자가 정보처리장치를 운영하는 본래의 목

적과 상이한 명령을 입력하는 것을 의미한다. 그리하여 정보처리장치를 관리 운영할 권한이 없는 자가 그 정보처리장치에 입력되어 있던 관리자의 아이디와 비밀번호를 무단으로 변경하는 행위는 정보처리장치에 부정한 명령을 입력하여 정당한 아이디와 비밀번호로 정보처리장치에 접속할 수 없게 만드는 행위로서 정보처리에 장애를 현실적으로 발생시킬 뿐 아니라 이로 인하여 업무방해의 위험을 초래할 수 있으므로, 컴퓨터등장애업무방해죄를 구성하고,[135] 포털사이트 운영회사의 통계집계시스템 서버에 허위의 클릭정보를 전송하여 검색순위 결정 과정에서 위와 같이 전송된 허위의 클릭정보가 실제로 통계에 반영됨으로써 정보처리에 장애가 현실적으로 발생하였다면, 그로 인하여 실제로 검색순위의 변동을 초래하지는 않았다 하더라도 '컴퓨터 등 장애 업무방해죄'가 성립한다.[136] 또한 피고인이 악성프로그램이 설치된 피해 컴퓨터 사용자들이 실제로 인터넷 포털사이트 '네이버' 검색창에 해당 검색어로 검색하거나 검색 결과에서 해당 스폰서링크를 클릭하지 않았음에도 악성프로그램을 이용하여 그와 같이 검색하고 클릭한 것처럼 네이버의 관련 시스템 서버에 허위의 신호를 발송하는 방법으로 정보처리에 장애를 발생하게 하였다고 하여 컴퓨터등장애업무방해로 기소된 사안에서, 피고인의 행위는 객관적으로 진실에 반하는 내용의 정보인 '허위의 정보'를 입력한 것에 해당하고, 그 결과 네이버의 관련 시스템 서버에서 실제적으로 검색어가 입력되거나 특정 스폰서링크가 클릭된 것으로 인식하여 그에 따른 정보처리가 이루어졌으므로 이는 네이버의 관련 시스템 등 정보처리장치가 그 사용목적에 부합하는 기능을 하지 못하거나 사용목적과 다른 기능을 함으로써 정보처리의 장애가 현실적으로 발생하였고, 이로 인하여 네이버의 검색어

135) 대법원 2006. 3. 10. 선고 2005도382 판결
136) 대법원 2009. 4. 9. 선고 2008도11978 판결

제공서비스 등의 업무나 네이버의 스폰서링크 광고주들의 광고 업무가 방해되었다고 본다.[137] 킹크랩 프로그램을 이용한 댓글 순위 조작 작업이 허위의 정보나 부정한 명령을 입력하여 정보처리에 장애를 발생하게 함으로써 피해자 회사들의 댓글 순위 산정 업무를 방해한 것에 해당한다고 판단하였다.[138]

2) 정보처리에 장애 발생

'정보처리에 장애'를 발생하게 하여 '업무를 방해'하여야 하는데, 정보처리의 장애란 가해행위 결과 정보처리장치가 그 사용목적에 부합하는 기능을 하지 못하거나 사용목적과 다른 기능을 하는 것을 말한다.[139] 컴퓨터 등 정보처리장치의 일반적인 기능(자료의 저장, 연산, 검색)을 불가능하게 하는 것은 물론 업무자가 정보처리장치에 의해서 달성하려는 구체적인 정보처리를 불가능하게 하는 것을 포함한다. 정보처리에 장애를 발생하게 하여 업무방해의 결과를 초래할 위험이 발생한 이상, 나아가 업무방해의 결과가 실제로 발생하지 않더라도 위 죄가 성립한다.[140] 따라서 포털사이트 운영회사의 통계집계시스템 서버에 허위의 클릭정보를 전송하여 검색순위 결정 과정에서 위와 같이 전송된 허위의 클릭정보가 실제로 통계에 반영됨으로써 정보처리에 장애가 현실적으로 발생하였다면, 그로 인하여 실제로 검색순위의 변동을 초래하지는 않았다 하더라도 '컴퓨터 등 장애 업무방해죄'가 성립한다.[141]

---

137) 대법원 2013. 3. 28. 선고 2010도14607 판결
138) 대법원 2020. 2. 13. 선고 2019도12194 판결
139) 대법원 2009. 4. 9. 선고 2008도11978 판결
140) 위 죄는 추상적 위험범에 해당하기 때문에 업무방해라는 구체적인 결과의 발생을 구성요건으로 하지 않는다.

나. 컴퓨터등장애업무방해와 웹 크롤링

1) '부정한 명령'인지 여부

웹 크롤링을 통한 데이터 수집이 컴퓨터등장애업무방해가 성립할 것인지를 일의적으로 말할 수는 없다. '야놀자 판결'을 대상으로 살펴본다. 공소사실에서 부정한 명령은 피고인들이 이 사건 API 서버에 접속하여 위도, 경도, 반경에 관하여 피고인 회사의 위치를 기준으로 반경 1,000㎞로 하여 그 안에 있는 모든 숙박업소 정보를 요청하는 것이라고 한다. 결국 2가지 사정의 해석 문제, 즉 첫째는 앱을 통하지 않고 PC를 통하여 접속하였다는 것, 둘째는 반경을 피해자 회사가 설정해 놓은 범위를 넘어 1,000㎞로 하였다는 것을 컴퓨터등장애업무방해와 관련하여 형법적으로 어떻게 평가할 것인지로 귀결된다.

전술한 바와 같이 이 사건 API 서버에 이 사건 크롤링 프로그램을 통하여 명령구문을 입력한 것이 명령입력 권한이 없는 행위라고 할 수 없다고 본다. 왜냐하면 피해자 회사는 이 사건 API 서버에 명령 입력의 범위를 제한하지 아니하였기 때문이다. 이 사건 앱에서만 검색범위의 제한이 일부 설정되어 있었다고 하더라도 이 사건 API 서버는 기본적으로 주어진 명령구문에 대응하는 숙박업소 정보를 반환하는 것에 있으므로 반경을 넓혀서 검색한 것만으로 그것이 부정하다고 할 수 없다. 물론 피해자 회사가 검색범위를 3㎞로 제한해 놓은 의도를 '야놀자 판결' 및 관련 하급심 판결의 내용만으로는 정확히 알

141) 대법원 2009. 4. 9. 선고 2008도11978 판결

수 없다. 그와 같은 제한이 이 사건 앱의 접속 시간과 접속 횟수를 늘리기 위한 피해자 회사의 경영적 판단일 수도 있고, 단순히 이용자에게 그의 위치를 기준으로 적절한 정보를 제공해 주기 위한 이용자 편의를 위한 것일 수도 있고, 아니면 그 밖의 다른 이유가 개재되어 있었을 수도 있다. 그러나 그와 같은 피해자 회사의 일방적 의사에 따라 설정된 검색범위의 제한을 우회하는 이용을 부정한 것이라고 한다면, 정보통신망 관리자의 주관적 의사에 의하여 형사처벌이 불합리하게 확대되는 결과가 될 것이다.

따라서 '부정한 명령'의 판단도 정보통신망 침입의 접근권한과 같이 객관적 사정을 중심으로 판단해야 한다. 본죄의 성립을 인정한 주요 선례들을 보아도, 권한 없는 자가 아이디와 비밀번호를 무단으로 변경하거나, 허위의 클릭정보를 전송하거나, 권한 없는 자가 함부로 컴퓨터에 비밀번호를 설정한 행위 등으로 명백하게 허위의 정보를 전송하거나 정보처리장치의 관리 권한 없으면서도 아이디와 비밀번호를 무단으로 입력하거나 비밀번호를 설정한 사안들이어서, 그와 같은 것은 객관적 사정에 비추어 보아도 부정한 것이라고 볼 수 있다.

2) '정보처리에 장애 발생' 여부

'야놀자 판결'에서 이는 사실인정의 문제로 귀결된 것으로 보인다. 항소심은 접속장애가 처음 발생기 전까지 초당 평균 접속횟수가 더 많았던 일자들에서는 접속 장애가 발생하지 않았고, 그 이후에도 초당 접속횟수가 더 많았던 일자들에서도 접속 장애가 발생하지 않았다는 점, 접속장애가 발생한 일자들은 토요일, 일요일, 추석 등으로 자

연 이용자 증가에 따른 것이었을 가능성을 배제할 수 없는 점, 접속 장애가 일어난 일자들의 초당 최다 접속횟수는 30명대인데, 이는 접속 장애가 일어나지 않은 날의 초당 접속 횟수보다 적고, 피해자 회사의 서버는 초당 50명의 동시 접속자를 수용할 수 있게 설계되기도 한 점 등을 들었다.

5. 소결

웹 크롤링을 통한 데이터 수집의 형사책임을 정보통신망 침입(①), 데이터베이스제작자의 권리 침해(②), 컴퓨터등장애업무방해(③) 각각의 구성요건을 중심으로 살펴보았다. 위 각 구성요건은 웹 크롤링의 순서상 문제가 되는 국면이 다르다. ①은 크롤링 대상 서버에 접근하는 행위 자체의, ②는 크롤링 대상 자료의 성격과 크롤링의 방법의, ③은 크롤링의 결과 정보통신망에 장애가 발생하였는지의 각각의 국면에 대해서 형법적 평가를 하는 것이다. ①에 있어서는 가장 중요한 것이 접근권한이다. 크롤링 대상 웹 서버에 대한 접근권한의 유무가 ①의 구성요건해당성 여부를 결정짓는다. 접근권한은 해당 웹 사이트의 접근이 일반적으로 허용되어 있거나 접근에 특별한 제약이 없는 이상 이용약관과 보호조치와 같은 객관적 사정에 따라 엄격하게 판단되어야 한다. 이용약관과 보호조치가 진정으로 객관적인 사정인가에는 의문이 있다. 이용약관과 보호조치 모두 웹 서버 관리자가 일방적으로 설정하는 것이고, 그 웹 서버에 접근하려고 하는 자가 이용약관의 내용을 숙지해야 한다거나, 보호조치 여부를 매번 확인해야 할 의무까지 인정할 수는 없다고 생각하기 때문이다. 그래서 위 두 가지를 가지고 접근권한을 판단할 때에는 약관을 작성 또는 보호조치를 설정한 서버

관리자의 진정한 의사를 찾으려고 해서는 안 된다. 그것들을 통해서 외부로 드러난 객관화된 관리자의 의사를 찾아야 하고, 어떤 경우에도 인터넷이나 웹의 공공성을 전제로 해야 한다. 접근권한을 제한적으로 해석하거나 부여되었던 접근권한을 서버 관리자의 일방적인 의사로 철회하였다고 인정하는 데에는 매우 신중한 태도를 취해야 한다.

②는 웹 크롤링을 통해 크롤링 대상 웹 서버와 연결된 데이터베이스 전부가 아닌 일부분이 수집되었을 때(만약 데이터베이스 전부를 수집해갔다면, 그것은 이미 웹 크롤링이 아닌 미러링으로 볼 것이어서, 이 글의 범위를 벗어난다), 그것을 데이터베이스제작자의 권리 침해로 인정하기 위해서는 어떤 요건이 필요한가에 관한 것이다. 데이터베이스제작자의 권리 침해를 인정하기 위해서는 데이터베이스의 '상당한 부분'의 복제가 인정되어야 하고, '상당한 부분'의 복제는 반복적이거나 특정한 목적을 위하여 체계적으로 복제하여 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에 인정될 수 있다. '상당한 부분'의 판단은 양적인 측면과 질적인 측면이 모두 검토되어야 한다. '야놀자 판결' 중 이 부분에 관한 설시에 나는 특별한 관심을 갖고 있는데, 왜냐하면 이것이 서문에서 밝힌 웹 크롤링을 주제로 선택하게 된 나의 호기심과 직접적으로 관련이 되기 때문이다. 사람이 일일이 손으로 할 때에는 아무런 책임을 지지 않는 것이 컴퓨터 프로그램 등 자동화 도구를 사용할 때에는 어떻게 달라질 수 있는지에 대한 하나의 답을 제시하고 있다. '야놀자 판결'의 설시 중 '반복적이고 체계적'이라는 부분이 자동화 도구의 사용을 포섭할 수 있다. 즉 반복적이고 체계적인 자동화 도구를 사용하게 되면 법적 책임을 질 수도 있다는 것이다.[142)]

③의 핵심은 '부정한 명령'과 '정보처리에 장애 발생' 여부이다. '부정한 명령'은 해당 웹 사이트의 통상적인 이용과의 관계에서 보아야 할 것이나, 웹 크롤러의 활동이 통상적인 이용과 모습을 달리한다는 이유만으로 쉽사리 '부정한 명령'이라고 보아서는 안 된다. 앞서 본 접근권한과 같이 엄격하게 볼 필요가 있다. '정보처리에 장애 발생' 여부는 일의적으로 판단하기 어렵다. 정보처리에 장애가 발생하였다고 인정할 수 있는 기준은 무엇인지, 웹 크롤러의 활동과 정보처리 장애 사이에 인과관계는 어떻게 판단할 것인지 등의 문제는 다분히 사실인정의 문제로 보인다. 이는 웹 크롤링의 정중함 정책(politeness policy)과 관련될 것이다.

마지막으로 첨언할 것은 웹 크롤러의 유형별로 법적 책임을 달리 볼 수 있을 것인가라는 질문에 대한 나의 답변이다. 학문이 세계에 대한 추상화(abstraction) 과정이라고 한다면, 이 글의 주제인 웹 크롤링에 대해서도 추상화 작업이 수행되어야 한다. 추상화의 가장 손쉬운 방법은 유형화(categorization)이다. 그러니 웹 크롤링을 유형화하고 각 유형별로 법적 책임을 검토하는 것이 가장 손쉬운 추상화 작업이 될 것이다. 앞서 **「Ⅱ. 웹 크롤링(Web Crawling)」**에서 웹 크롤러의 유형을 다운로드하는 정보의 범위에 따라 범용(general purpose)과 집중(focused)으로, 작동 방식에 따라 일반형(general)과 분산형(distributed)으로 구별하였다. 일응 크롤러가 집중하되고, 분산화될수록 웹 서버에 일으키는 부하가 커지고 수집되는 정보의 양도 많아질 것이니 법적으로 문제가 될 소지도 커지게 된다. 그러나 본문에서

142) 물론 이 부분에 대한 정당성 및 그와 같은 법적 책임이 저작권법 이외의 다른 법 영역에서도 그대로 적용될 수 있는지는 추가 연구가 필요하다.

웹 크롤러의 위와 같은 유형별로 형사 책임을 논하지는 않았다. 그 이유는 이렇다. 웹 크롤링은 시드 URL에서 하이퍼링크를 따라가며 해당 웹 사이트를 다운하는 것이 기본 개념이지만, 그 개념상 단순성과는 달리 실제에서 사용되고 있는 모습은 매우 다양하여 이들을 법적 책임을 구별짓기에 적절한 방법으로 구별하는 기준을 찾을 수 없었다. 섣부른 유형화는 실제 크롤러들의 설계 모습과 운용되는 상황과는 유리된 논의가 될 수 있는 위험성도 있다. 또한 앞서 본 바와 같이 크롤러의 설계는 개별 크롤러의 목적과 용도에 따라 크롤링 정책들의 다양한 조합들에 의하여 이루어질 터인데, 그 조합들의 경우의 수를 적정한 수로 나눌 수 있는 기준이 없다. 그래서 웹 크롤러의 유형별 법적 책임의 구별 작업은 보류하였다. 다만 개별 사안에서 해당 웹 크롤링의 법적 책임을 따지는 데에는 해당 크롤러가 채택한 크롤링 정책을 고려하는 것이 유의미할 것임은 두말할 나위가 없다. 그와 같은 접근은 법적 책임을 논할 때에 법적인 도그마에 매몰되지 않고 기술적 특성을 함께 따져볼 수 있는 넓은 시야를 제공할 것이다.

Ⅳ. 웹 크롤링 사용의 기타 법적 책임

1. 도입

여기에서는 웹 크롤링을 통한 데이터 수집의 법적 책임을 형사 책임을 제외한 나머지 영역에서 검토한다. 민사적으로 웹 크롤링의 사용에 대해서 불법행위로 손해배상책임을 지는가가 문제되는데, 이에 관해서는 부정경쟁방지 및 영업비밀보호에 관한 법률(이하 '부정경쟁방지법'이라고 한다)이 부정경쟁행위로서 불법행위 태양을 구체화하고 있으므로 위 법을 중심으로 살펴본다. 그 외에는 경쟁법적으로 행정처분의 대상이 될 수 있는지가 문제되는데, 이에 관해서는 독점규제 및 공정거래에 관한 법률(이하 '공정거래법'이라고 한다)이 정한 시장지배적 사업자의 남용행위 또는 불공정거래행위에 해당하는지가 직접적으로 관련된다.[143]

2. 부정경쟁방지 및 영업비밀보호에 관한 법률상 책임

부정경쟁방지법 제2조 제1호 (카)목과 (파)목[144]에 대해서 본다.

---

143) 독점규제 및 공정거래에 관한 법률상 시장지배적 사업자의 남용행위 또는 불공정거래행위를 이유로 민사상 불법행위 책임을 물을 여지도 충분히 있다. 이에 관한 논의는 주로 경쟁법 연구자들 사이에서 진행되고 있는데, 단순히 위 법이 금지하는 행위를 하였다는 이유만으로 민사상 불법행위 책임까지 곧바로 인정되는 것은 아니고, 또한 손해의 산정 역시 쉽지 않은 매우 착잡한 영역으로 남아 있는 것으로 보인다. 그 부분에 대한 연구는 이 글의 범위를 넘는다.

144) 부정경쟁방지 및 영업비밀보호에 관한 법률]
제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.
1. "부정경쟁행위"란 다음 각 목의 어느 하나에 해당하는 행위를 말한다.
카. 데이터(「데이터 산업진흥 및 이용촉진에 관한 기본법」 제2조 제1호에 따른 데이터 중 업(業)으로서 특정인 또는 특정 다수에게 제공되는 것으로, 전자적 방법으로 상당량 축적·관리되고 있으며, 비밀로서 관리되고 있지 아니

가. 부정경쟁행위로서 성과물의 부정차용

부정경쟁방지법 제2조 제1호 (파)목은 부정경쟁행위의 태양 중 하나로서 '그 밖에 타인의 상당한 투자나 노력으로 만들어진 성과 등을 공정한 상거래 관행이나 경쟁질서에 반하는 방법으로 자신의 영업을 위하여 무단으로 사용함으로써 타인의 경제적 이익을 침해하는 행위'를 포함하고 있다. 이 규정은 다른 부정경쟁행위인 출처혼동이나 품질오인 등의 해당여부와 관계없이 타인의 투자나 노력의 성과물을 부정차용(misappropriation)하는 행위를 포괄적으로 부정경쟁행위에 포함시키고 있는데, 이는 부정경쟁행위에 관한 보충적 일반조항으로서의 성격을 갖는다.[145] 부정경쟁방지법은 부정경쟁행위 중 상당수에 대해서 형사처벌도 규정하고 있으나, 이 조항에 대해서는 처벌대상에서 제외하고 있다. 이는 형법의 명확성 원칙상, 구성요건이 모호하고 불합리하게 처벌 대상의 범위를 확대할 위험이 있기 때문인 것으로 보인

한 기술상 또는 영업상의 정보를 말한다. 이하 같다)를 부정하게 사용하는 행위로서 다음의 어느 하나에 해당하는 행위

1) 접근권한이 없는 자가 절취·기망·부정접속 또는 그 밖의 부정한 수단으로 데이터를 취득하거나 그 취득한 데이터를 사용·공개하는 행위

2) 데이터 보유자와의 계약관계 등에 따라 데이터에 접근권한이 있는 자가 부정한 이익을 얻거나 데이터 보유자에게 손해를 입힐 목적으로 그 데이터를 사용·공개하거나 제3자에게 제공하는 행위

3) 1) 또는 2)가 개입된 사실을 알고 데이터를 취득하거나 그 취득한 데이터를 사용·공개하는 행위

파. 그 밖에 타인의 상당한 투자나 노력으로 만들어진 성과 등을 공정한 상거래 관행이나 경쟁질서에 반하는 방법으로 자신의 영업을 위하여 무단으로 사용함으로써 타인의 경제적 이익을 침해하는 행위

제18조(벌칙)

③ 다음 각 호의 어느 하나에 해당하는 자는 3년 이하의 징역 또는 3천만원 이하의 벌금에 처한다.

1. 제2조 제1호(아목, 차목, 카목1)부터 3)까지, 타목 및 파목은 제외한다)에 따른 부정경쟁행위를 한 자

145) 정상조, 박준석, 『지식재산권법(제5판)』, 홍문사, 2020년, 692쪽

다. 위 조항을 과도하게 널은 범위에서 적용하게 되면 새로운 투자나 노력을 위축시키는 부작용을 초래할 수 있으므로 위 일반조항의 적용에는 신중을 기해야 한다.[146]

대법원은 부정경쟁방지법 제2조 제1호 (파)목이 제정되기 이전에도 인터넷 포털사이트에 접속한 인터넷 사용자들의 모니터에서 프로그램을 이용한 광고행위를 하여 그 포털사이트의 광고이익을 가로채갔던 사안에서, "경쟁자가 상당한 노력과 투자에 의하여 구축한 성과물을 상도덕이나 공정한 경쟁질서에 반하여 자신의 영업을 위하여 무단으로 이용함으로써 경쟁자의 노력과 투자에 편승하여 부당하게 이익을 얻고 경쟁자의 법률상 보호할 가치가 있는 이익을 침해하는 행위는 부정한 경쟁행위로서 민법상 불법행위에 해당한다"고 하면서, 광고프로그램업자의 위와 같은 행위가 그 포털사이트의 투자나 노력의 결과 형성된 신용과 고객흡인력을 부정차용하는 불법행위에 해당한다고 판단하여, 성과물의 부정차용이 민법상 불법행위에 해당할 수 있음을 밝힌바 있다.[147][148] 이후 부정경쟁방지법 제2조 제1호 (파)목 제정된 이후에도 위 조항에 대한 구체적 해석론을 제시하였다.[149] 그 판결에 따르면, 위 (파)목은 새로이 등장하는 경제적 가치를 지닌 무형의 성과를 보호하고 입법자가 부정경쟁행위의 모든 행위를 규정하지 못한 점을 보완하여 법원이 새로운 유형의 부정경쟁행위를 좀 더

146) 정상조, 박준석, 『지식재산권법(제5판)』, 홍문사, 2020년, 693-694쪽
147) 대법원 2010. 8. 25.자 2008마1514 결정
148) 위 대법원 결정 이후 2013. 7. 30. 법률 제11963호로 개정된 부정경쟁방지 및 영업비밀보호에 관한 법률은 그 취지를 반영하여 "그 밖에 타인의 상당한 투자나 노력으로 만들어진 성과 등을 공정한 상거래 관행이나 경쟁질서에 반하는 방법으로 자신의 영업을 위하여 무단으로 사용함으로써 타인의 경제적 이익을 침해하는 행위"를 부정경쟁행위의 하나로 추가하였다.
149) 대법원 2020. 3. 26. 선고 2016다276467 판결

명확하게 판단할 수 있도록 함으로써, 변화하는 거래관념을 적시에 반영하여 부정경쟁행위를 규율하기 위한 보충적 일반조항인데, 위 (파)목은 그 보호대상인 '성과 등'의 유형에 제한을 두고 있지 않으므로, 유형물뿐만 아니라 무형물도 이에 포함되고, 종래 지식재산권법에 따라 보호받기 어려웠던 새로운 형태의 결과물도 포함될 수 있고, 성과 등을 판단할 때에는 위와 같은 결과물이 갖게 된 명성이나 경제적 가치, 결과물에 화체된 고객흡인력, 해당 사업 분야에서 결과물이 차지하는 비중과 경쟁력 등을 종합적으로 고려해야 한다고 한다. 다만 그러한 성과 등이 상당한 투자나 노력으로 만들어진 것인지는 권리자가 투입한 투자나 노력의 내용과 정도를 그 성과 등이 속한 산업분야의 관행이나 실태에 비추어 구체적·개별적으로 판단하되, 성과 등을 무단으로 사용함으로써 침해된 경제적 이익이 누구나 자유롭게 이용할 수 있는 이른바 공공영역(public domain)에 속하지 않아야 한다. 또한 행위태양에 있어서도 (파)목이 정하는 '공정한 상거래 관행이나 경쟁질서에 반하는 방법으로 자신의 영업을 위하여 무단으로 사용'한 경우에 해당하기 위해서는 권리자와 침해자가 경쟁관계에 있거나 가까운 장래에 경쟁관계에 놓일 가능성이 있는지, 권리자가 주장하는 성과 등이 포함된 산업분야의 상거래 관행이나 경쟁질서의 내용과 그 내용이 공정한지, 위와 같은 성과 등이 침해자의 상품이나 서비스에 의해 시장에서 대체될 수 있는지, 수요자나 거래자들에게 성과 등이 어느 정도 알려졌는지, 수요자나 거래자들의 혼동가능성이 있는지 등을 판단기준으로 제시하고 있다.

나. 부정경쟁행위로서 데이터 부정사용

2021. 12. 7. 법률 제18548호로 개정된 부정경쟁방지법 제2조 제1호 (카)목은 부정경쟁행위의 태양으로서 데이터 부정사용을 추가하였다. 이는 국내 데이터 산업발전과 기업성장을 위한 노력의 일환으로 기업들이 안심하고 데이터를 거래·유통할 수 있는 환경을 구축하기 위한 목적[150]으로 이루어졌는데, 거래 목적으로 생성한 데이터를 접근권한 없이 부정취득하거나 사용한 행위를 부정경쟁행위의 유형으로 신설하였다. 앞서 본 바와 같이 종래에도 데이터의 부정사용에 대해서는 부정경쟁방지법 제2조 제1호 (파)목[151]의 일반조항에 따라 손해배상은 물론 금지를 청구할 수 있었으나, 개정법은 이를 명문으로 인정하였다. 여기서 보호대상이 되는 데이터는 특정대상에게 제공하기 위한 목적으로 생성한 것으로, 아이디, 비밀번호설정으로 접근을 제한하는 등 전자적으로 관리되어야 하며, 상당량 축적되어 경제적 가치를 가지고 있고, 비밀로서 관리되고 있지 아니한 기술상 또는 영업상 정보에 대한 것이어야 한다.

위 데이터 부정사용에 대해서는 후속 연구를 통해 구체적인 적용범위, 요건 및 효과 등에 대해서 밝혀야 할 것이나, (파)목이 (카)목에 대해 특별법 관계에 있으므로 웹 크롤링 사안에 있어서는 (파)목보다 우선적으로 적용이 검토될 것이다. 사실 '야놀자 판결' 이후 관련 업계와 행정 당국에서는 크롤링을 통한 데이터의 수집이 전면적으로 허용되는 것이 아닌가 하는 염려가 있었는데,[152] 부정경쟁방지법 담

150) 특허청 공식 블로그(https://bolg.naver.com/kipoworld2/222704963601: 2022. 11. 8. 방문)
151) 이 사건은 위 (카)목이 제정되기 전의 것이어서 (카)목은 적용되지 않는다.
152) 언론은 '야놀자 판결'에 대해 자극적인 제목으로 기사를 전하였다. 예컨대 한국경제 2022. 5. 15.자 「"크롤링 무죄" 판결에 속 타는 대형 플랫폼」(https://www.hankyung.com/society/article/2022051568601); 중앙일보 2022. 5. 13.자 「웹 돌아다니며 정보 수집…데이터 '크롤링' 날개 다나」

당부처인 특허청에서 그와 같은 분위기를 반영하여 위와 같은 개정을 주도하였던 것으로 보인다. 위 개정법은 데이터에 대한 공공성 보다는 데이터를 지식재산 또는 그와 유사한 재산에 대한 보호의 객체로만 취급하였다는 점에서 아쉬움이 남는다.

다. 서울고등법원 2022. 8. 25. 선고 2021나2034740 판결[153]

'야놀자 판결'과 같은 사안에서 피해자인 '야놀자'(원고)가 '어기어때'(피고)를 상대로 손해배상을 청구한 사안이다. '야놀자'는 자신이 상당한 투자와 노력, 시간을 들여 취득한 자신의 성과인 제휴 숙박업소 정보를 무단으로 복제하여 '여기어때'의 숙박업소 정보 제공 업무에 이용함으로써 마케팅 비용 절감, 서비스 제공을 통한 수수료 상당의 이익을 취득하고 자신의 경제적 이익을 침해하였는데, 이는 부정경쟁방지법 제2조 제1호 (파)목의 부정경쟁행위에 해당한다고 주장하였다. 이에 대하여 '여기어때'는 자신이 사용한 정보검색기술인 크롤링은 위법한 기술이 아니고, 자신은 '야놀자'가 허용한 방법에 따라 '야놀자'가 공개한 정보를 검색하였을 뿐이라고 주장하였다.

법원은 제휴 숙박업소 정보가 원고의 상당한 투자나 노력으로 만들어진 성과에 해당한다고 보았는데, 그 주된 근거는 다음과 같다. ① 온라인으로 숙박업소를 광고하고 고객이 예약에 나아가는 경우 온라인 예약절차를 제공하는 서비스에서 숙박업소 관련 정보의 양과 질은 위 사업의 핵심 요소이고, 특히 경쟁업체가 제휴계약을 체결한 숙박업

(https://www.joongang.co.kr/article/25070824) 등이다. 그러한 언론보도를 통해 보면, 기대보다는 염려가 대부분이었던 것으로 보인다.

153) 위 판결에 대해서는 원·피고 모두 상소하지 않아 그대로 확정되었다.

소의 업체명, 해당 숙박업소에서 제공되는 서비스 내용 등의 정보는 영업 전략에 이용되는 주요 정보이다. ②원고는 국내에서 전국 숙박업소를 대상으로 이러한 온라인 숙박예약 서비스를 제공하는 영업을 개척한 선두주자이다. 원고는 2005년경부터 이와 같은 정보를 수집하고 이를 데이터베이스로 구축하였다. ③원고의 영업사원들이 수해 동안 전국의 숙박업소들을 직접 방문하여 숙박시설 등을 평가하고 반복적으로 영업하는 등의 방법으로 제휴 숙박업소를 발굴하고 관리해 왔다. ④원고는 숙박업소 정보를 수집하여 분류한 후에도 지속적으로 갱신·검증·보충 작업을 수행하였다. ⑤원고는 숙박업소 정보 수집과 분류, 갱신·검증·보충 업무를 위하여 상당한 비용을 투자하였다. ⑥원고는 제휴 숙박업소 정보를 공개하고 있으나, 서비스 이용약관[154]에서 그 정보의 사용을 제한하고 있으므로 그것이 공개하고 있다는 사정만으로 이용자 누구나 어떤 목적과 방식으로든 자유롭게 복제하여 사용할 수 있는 공공역역에 속한 것이라고 볼 수 없다.

나아가 법원은 피고가 제휴 숙박업소 정보를 공정한 상거래 관행이나 경쟁질서에 반하는 방법으로 자신의 영업을 위하여 무단으로 사용하여 원고의 경제적 이익을 침해하였다고 보았는데, 그 주된 근거는 다음과 같다. ①원고와 피고는 모두 숙박업소 정보를 제공하는 인터넷 웹 사이트를 운영하는 회사로서, 상호 경쟁자이이고, 제휴 숙박업소 정보는 원·피고의 영업에 핵심 요소이다. ③피고는 원고가 당초 접속을 예정한 방법을 우회하여 PC를 통해 모바일앱 이용자가 모바일앱을 이용하는 것처럼 어플리케이션 API 서버에 접속하고, 원고가 설정한

154) 제19조 제6호 이용자는 회사를 이용함으로써 얻은 정보를 회사의 사전 승낙 없이 복제, 송신, 출판, 배포, 방송 등 기타 방법에 의하여 영리 목적으로 이용하거나 제3자에게 이용하게 하여서는 안 됩니다.

검색범위 명령어를 변경하여 반경 1,000㎞ 즉, 전국에 있는 모든 숙박업소 정보를 요청하도록 설계된 크롤링 프로그램을 개발하였고, 위 프로그램을 매일 1~3회 또는 1시간마다 또는 특정일의 특정시간대에 지속적·반복적으로 실행함으로써 정상적인 모바일앱 접속 방식으로 정보를 수집하는 경우와 비교할 수 없는 대량의 동시간대 전국 숙박업소의 정보를 실시간 전송받아 복제하였다. ④피고의 크롤링은 원고와의 경쟁에서 우위를 차지하기 위한 목적이었고, 9개월간 조직적·지속적으로 이루어졌다. ⑤ 원고의 API 서버는 '모바일앱을 통한 원고의 숙박 예약 서비스 이용'이라는 원고가 제공한 이용 목적과 다른 기능을 수행하게 되었고, 원고의 서비스를 제공하는 어플리케이션에 비정상적인 대량의 접속이 이루어져 원고의 서버에 서비스 제공과 무관한 대량의 트래픽이 발생하였다. 원고는 반복적인 접근에 따른 대량 호출 신호를 감지하고 피고가 이용하는 아마존 웹서비스 클라우드 서버의 IP 주소를 수차례 차단하였는데, 피고는 서버의 전원을 차단하였다가 다시 켜는 방식으로 IP 주소를 변경하면서 위와 같은 행위를 계속하였다. ⑥피고는 장기간 반복적으로 대량 복제한 제휴 숙박업소 정보를 분석하여, 원고의 영업 전략 및 현황을 파악하고, 이를 피고의 영업 전략을 수립하는 데 활용하였다.

라. 데이터 부정사용 또는 성과물의 부정차용과 웹 크롤링

앞서 말한 바와 같이 웹 크롤링에 대해서는 (파)목의 성과물의 부정차용보다 특별법적 관계에 있는 (카)목의 데이터 부정사용이 우선적으로 검토될 것이다. 다만 (카)목에 해당되지 않는데 (파)목의 일반조항에 해당되는 웹 크롤링 유형은 쉽게 상정하기 어렵다. 그만큼

(카)목이 요건이 포괄적이고 추상적으로 되어 있다. 데이터의 부정사용을 ①'접근권한 없는 자'의 '절취·기망·부정접속 또는 그 밖의 부정한 수단'에 의한 '데이터 취득·사용·공개', ②접근권한이 있더라도 '부정한 이익을 얻거나 데이터 보유자에게 손해를 가할 목적'으로 그 데이터를 '사용·공개·제3자 제공' 및 ③그와 같은 사실을 알고 데이터를 취득·사용·공개하는 행위로 정하고 있다. '부정'이라는 추상적인 규정 때문에 모든 웹 크롤링을 통한 데이터 수집이 부정경쟁방지법의 데이터 부정사용에 해당될 수도 있다. 나는 기본적으로 재판에 있어서 민사와 형사의 결론은 얼마든지 달라질 수 있고, 굳이 그러한 것을 법질서의 통일이라는 관점을 들이대면서 방지하려는 노력은 불필요한 것이자 유해한 것이라고 생각한다. 처벌규성요건의 해석과 적용은 엄격하게 하여야 하나 민사상 책임의 인정은 그 보다 완화되어도 별 문제가 되지 않고, 오히려 그러한 법 적용이 모든 민사사건이 형사화되어 전체적으로 불필요한 분쟁이 증가하는 것을 막을 수도 있다고 본다. 그러나 그렇다고 하여 (카)목과 같이 민사상 불법행위책임의 인정범위를 사실상 무한히 넓힐 수 있는 여지를 주는 것은 바람직하지 않다. 특히 위 (카)목이 제정된 것이 웹 크롤링을 통한 데이터 수집을 애초에 막으려는 의도인 것이고 거기에는 웹 크롤링을 통한 데이터 수집 자체를 금지 또는 제재해야 할 대상이라고 전제하고 있는 것으로 보인다. 특히 제정된 (카)목은 데이터 '취득'까지도 부정경쟁행위로 규정하고 있는데, 경쟁업체의 공개된 정보를 취득하여 자신의 영업전략 수립에 참고하는 것이 어떻게 불법행위를 구성할 수 있는지 납득하기 어렵다. 위 (카)목의 해석은 계속해서 많은 사례의 축적과 연구로 정립되어야 할 것이나, 앞서 본 논의들 즉 '접근권한'에 있어서는 정보통신망 침입에서의 논의를, '부정한 수단'에 있어서는 데이터베이스제작

자의 권리 침해에서와 같은 논의를 차용하여 엄격하게 해석해야 한다고 생각한다.

3. 독점규제 및 공정거래에 관한 법률상 책임

가. 웹 크롤링에 대한 경쟁법의 평가

웹 크롤링에 대해서 경쟁법 학계에서 보는 관점은 크게 두 가지로 나뉜다. 공공성의 관점과 투입 자원의 관점이다. 공공성의 관점에서 보자면, 웹 크롤링은 인터넷에 존재하는 수많은 정보에 쉽고 빠르게 접근할 수 있는 필수적인 행위로서 경쟁자의 공개된 정보를 수월하게 취득하여 이용한다는 점에서 경쟁에 기여하는 공익적 성격이 있다.[155] 또한 웹 사이트의 개방성으로 인하여 원칙적으로 누구나 접근 가능하고 사람이 직접 웹 브라우징을 통하여 데이터에 접근하는 것과 웹 크롤링을 통하는 것 사이에 질적인 차이를 인정할 수 없다. 특히 웹 크롤링은 빅데이터나 이를 기반으로 하는 인공지능 분야에도 주목을 받는 데이터 수집의 기술로, 이에 대한 과도한 제한은 관련 산업 발전에 장애가 될 수 있다는 우려도 있다.[156] 반면 투입 자원의 관점에서 보자면, 데이터의 공공성만을 강조하게 된다면 사업자가 서비스를 제공하기 위여 많은 양의 데이터를 수집하고 이를 분석·처리해야 하는 데 인적·물적 자원이 투입하게 될 유인이 저해되고 경쟁질서에 왜곡이 초래되는 결과가 된다. 소위 '무임승차(free rider)'의 문제이다. 아무런 제한 없이 데이터 공유가 되면 시장에는 새로운 가치는 제

155) 유대종, "웹검색 서비스와 ISP책임에 관한 소고", 창작과 권리 46권, 세창출판사, 2007. 3., 107쪽
156) 강정희, "웹 크롤링의 데이터 수집행위와 경쟁법적 논의", 489쪽

공하지 않으면서 혁신의 동기를 약화시켜 시장의 경쟁기능이 왜곡되고 시장의 효율성이 저해되어 궁극적으로 소비자 후생에도 부정적인 영향을 미치게 된다.

웹 크롤링의 경쟁법상 책임은 위와 같은 관점을 전제로 논의되는데, 독점규제 및 공정거래에 관한 법률(이하 '공정거래법'이라고 한다)의 적용여부가 문제된다. 만약 위 법을 적용한다면 공정거래위원회를 통한 행정적 규제의 대상이 될 것이다. 동시에 경우에 따라서는 위 법 위반행위에 대해서 민사상 손해배상도 소구할 수 있을 것이다.

나. 불공정거래행위 또는 시장지배적지위 남용과 웹 크롤링

여기서는 웹 크롤링을 통하여 정보를 수집하려는 행위뿐만 아니라 웹 크롤링에 대응하여 정보 유출을 막으려는 행위를 구별하여 본다. 전자는 불공정거래행위로서 부당하게 웹 크롤링의 대상이 되는 사업자의 사업활동을 방해하는지가 문제된다(공정거래법 제45조 제1항 제8호).[157] 후자는 웹 크롤링을 제한하는 행위가 시장지배적지위 남용으로서 정보를 수집하려는 사업자의 사업활동을 부당하게 방해하는지(동법 제5조 제1항 제3호), 새로운 경쟁사업자의 참가를 부당하게 방해하는지(동조항 제4호), 부당하게 경쟁사업자를 배제하기 위하여 거래하거나 소비자의 이익을 현저히 해칠 우려가 있는지(동조항 제5호)[158] 등이 문제된다.[159] 경쟁법적 검토에 있어서 흥미로운 점은

157) 공정거래법
제45조(불공정거래행위의 금지) ① 사업자는 다음 각 호의 어느 하나에 해당하는 행위로서 공정한 거래를 해칠 우려가 있는 행위(이하 "불공정거래행위"라 한다)를 하거나, 계열회사 또는 다른 사업자로 하여금 이를 하도록 하여서는 아니 된다.
8. 부당하게 다른 사업자의 사업활동을 방해하는 행위

웹 크롤링에 대응하여 행위도 공정거래법에 저촉될 수 있다는 점이다.

웹 크롤링을 통하여 정보를 수집하는 것이 수집의 대상이 되는 사업자의 사업활동을 방해하는 행위로서 불공정거래행위에 포섭될 수 있는가. 생각건대 웹 크롤링 기술을 활용하여 웹 사이트 소유자가 공개한 데이터를 수집한 것만으로는 그와 같은 불공정거래행위에 포섭되기 어려울 것이다. 데이터를 수집한 것만으로 그 상대방 사업자의 사업에 직접적인 장애가 발생한 것을 인정하기 어렵기 때문이다. 그러나 상대방 사업자가 웹 사이트에 로봇 배제 프로토콜을 비롯한 기술적 보호수단으로 웹 크롤링을 제한하였는데, 그것을 우회하거나 크롤링 기술을 넘어서 위 보호수단을 무력화하였을 때 또는 웹 크롤러가 웹 페이지의 정보를 수집하는 데 단시간에 많은 데이터를 크롤링하여 대상 사이트 운영자의 웹 사용용량이 넘쳐서 비용이 발생하거나 웹 사이트의 호스트 서버에 과부하가 결려 속도가 느려지거나 접속 자체가 불가능하게 된다면[160] 이는 상대방 사업자의 사업활동에 장애가 있다고 할 것이어서 위 불공정거래행위에 포섭될 여지가 크다.[161]

---

158) 공정거래법
제5조(시장지배적지위의 남용금지) ① 시장지배적사업자는 다음 각 호의 어느 하나에 해당하는 행위(이하 "남용행위"라 한다)를 해서는 아니 된다.
3. 다른 사업자의 사업활동을 부당하게 방해하는 행위
4. 새로운 경쟁사업자의 참가를 부당하게 방해하는 행위
5. 부당하게 경쟁사업자를 배제하기 위하여 거래하거나 소비자의 이익을 현저히 해칠 우려가 있는 행위

159) 위와 같은 양자의 구별은 강정희, "웹 크롤링의 데이터 수집행위와 경쟁법적 논의-대법원 2022. 5. 12. 선고 2021도1533 판결을 중심으로", 사법 61호, 사법발전재단, 2022., 490쪽-492쪽; 황태희, "웹 크롤링의 경쟁법적 고찰", 경쟁법연구 43권, 한국경쟁법학회, 2021. 5., 184-194쪽에 따른 것이다.

160) 결국 해당 크롤러가 앞서 본 정중함 정책(politeness policy)을 어느 정도 반영하여 구현되었는지가 문제가 될 것이다.

161) 황태희, "웹 크롤링의 경쟁법적 고찰", 경쟁법연구 43권, 한국경쟁법학회, 2021. 5., 194-195쪽

나아가 웹 크롤링을 통한 정보수집을 제한하는 행위가 시장지배적 지위 남용으로 포섭될 수 있는가.[162] 이는 웹 크롤링을 통한 데이터 수집의 대상이 되는 사업자가 오히려 크롤링 자체를 금지하거나 제한함으로써 크롤링을 통해 정보를 수집하고자 하는 경쟁 사업자의 사업활동을 방해한다고 볼 수 있는지가 핵심이다. 이 부분은 경쟁법계에서도 계속 논의 중인데 아직 일반적으로 받아들여지는 기준은 없으며, 이에 관한 국내의 사례도 없다. 이 부분과 관련하여 거의 유일하게 참고할 수 있는 사례는 앞서 살펴본 HiQ Labs v. LinkedIn 사건이다. 미국 제9연방항소법원은 하이큐 랩스의 청구를 인용하여 링크드인의 웹 크롤링 금지 조치에 대해 금지명령(preliminary injunction)을 발령하는 것이 타당하다고 하였다. 위 사건은 여러모로 흥미로운데, 우선은 웹 크롤링을 통한 정보 수집에 관한 직접적 사례라는 점, 그것도 형사책임이 아니라 경쟁법적 관점에서의 사례라는 점, 게다가 웹 크롤링을 직접 수행한 사람에 대한 책임을 묻는 것이 아니라 오히려 웹 크롤링을 금지한 사람에 대한 책임을 추궁할 수 있는 여지를 남겼다는 점에서 그렇다. 그러나 위 사건은 하이큐 랩스의 예비적 금지명령(preliminary injunction) 신청을 인용한 1심 법원의 결정을 인용한 것뿐이고, 하이큐 랩스가 자신의 서비스 제공을 위해서는 링크드인의 공개 프로필 데이터를 사용하는 것 외 다른 방법이 없고 예비적 구제가 없을 경우 회복할 수 없는 피해를 입을 가능성이 큰지를 중심적으로 판단 기준으로 한 것이어서, 위 사건만으로 링크드인의 크롤링 제한 행위가 잠재적 경쟁자인 하이큐 랩스를 배제함으로써 경쟁제한을 우려를 야기하는지 여부에 대한 본안 판단 수준의 확정적인 것이 아니

162) 시장지배적지위를 인정하는 문제에도 데이터 및 웹의 개방성 등이 반영되어야 할 것이나, 거기에는 관련시장을 획정해야 하는 등 선결되어야 하는 문제가 많다. 그와 같은 논의는 이 글의 주제를 벗어나므로 생략한다.

라는 점에서 위 사건의 의미를 과대하거나 그 결론을 일반화하여 경쟁법 일반에 적용하는 것을 경계하는 견해도 있다.[163]

4. 소결

타인의 투자나 노력의 성과물을 부정차용하는 행위에 대해 민사상 불법행위를 이유로 손해배상책임을 부담할 수 있다는 점은 대법원 2010. 8. 25.자 2008마1514 결정부터 인정되고 있었다. 부정경쟁방지법은 위 대법원 결정의 취지를 반영하여 제2조 제1호 (파)목을 부정경쟁행위로 규정을 신설하였다. 그런데 '야놀자 판결' 이후 특허청의 주도로 다시금 위 법이 개정되어 (카)목으로 데이터 부정사용을 신설하였으므로 향후에는 웹 크롤링을 통한 데이터 수집에는 (카)목이 우선적으로 검토될 것이다. (카)목의 데이터 부정사용에 관한 내용이 포괄적이고 규정이 다소 추상적인데, 정보통신망 침입과 데이터베이스제작자의 권리 침해에서의 논의를 차용하여 엄격하게 해석되어야 한다. 한편 주로 웹 크롤링에 대한 경쟁법적 검토는 웹 크롤링으로 데이터를 수집하는 행위는 물론 웹 크롤링을 금지하는 행위도 규제의 대상이 될 수 있다. 전자는 공정거래법상 불공정거래행위로서 부당하게 웹 크롤링의 대상이 되는 사업자의 사업활동을 방해하는지가, 후자는 웹 크롤링을 제한하는 행위가 시장지배적지위 남용이 되는지 등이 문제된다. 이 부분은 행정적 규제의 대상이 될 수 있는지의 문제를 전제로 한 것이다.

163) 강정희, "웹 크롤링의 데이터 수집행위와 경쟁법적 논의-대법원 2022. 5. 12. 선고 2021도1533 판결을 중심으로", 사법 61호, 사법발전재단, 2022., 491쪽

Ⅴ. 결론

이제까지 웹 크롤링의 기본 원리와 그 사용의 법적 책임에 대해 살펴보았다. 웹 크롤러의 구현에서 가장 중요한 것은 크롤링 정책(crawling policies)의 반영 여부와 그 정도이다. 크롤링 정책 중 크롤링의 법적 책임을 논하는 데에 핵심이 되는 것은 정중함 정책(politeness policy)이다. 크롤링 대상이 되는 웹 서버에 실질적인 네트워크 장애를 발생하게 하였다면 형사책임든 민사책임이든 법적인 책임을 피해가기 어렵기 때문이다. 크롤링 사용의 법적 책임을 논하는 데에는 그 외에도 웹 크롤링이 데이터 수집과 관련되어 문제가 되고 있는 상황, 웹의 개방성 그리고 데이터의 중요성이 고려되어야 한다.

대부분의 웹 사이트들은 웹 크롤링을 허용하고 있다. 이는 검색 엔진의 크롤러들이 자신의 사이트를 발견하여 크롤링하여 검색 엔진에 검색될 수 있도록 하기 위한 것이다. 그런데 경쟁사의 크롤러들이 웹 사이트에 들어와 정보를 다운로드해가는 상황이 발생하고, 그것이 일반적인 영업 전략을 세우기 위한 시장조사 또는 자료조사의 기본적인 방법으로 인식되고 있어서 관련 사업자들 사이에 분쟁의 소재가 되고 있다. 크롤링의 대상이 되는 웹 사이트의 관리자 입장에서는 robot.txt의 사용이나 여타의 방법으로 웹 크롤링을 완벽하게 차단할 수 없을 뿐만 아니라, 설사 완벽하게 크롤러들을 차단할 수 있는 기술적 방안이 마련된다고 하더라도 그것은 웹 사이트를 운영하는 사업자의 입장에서는 검색 엔진의 검색 대상에서 자신의 웹 사이트를 배제하게 되어 이용자의 접근을 막는 결과가 되기 때문에 쉽사리 적용할 수도 없을 것이다. 여기가 웹 크롤링이 데이터 수집과 관련되어 문제가 제기

되는 지점이다.

형사책임에 관해서는 ‘야놀자 판결’의 논의를 큰 틀로 이용하여 논의했다. 종래 웹 크롤링에 대한 형사책임에 관한 논의가 전무하던 상황[164]에서 ‘야놀자 판결’은 최초로 웹 크롤링을 통한 데이터 수집의 형사책임을 밝힌 선구적 사건임에 틀림없다. 특히 ‘야놀자 판결’의 사실관계는 민사소송으로도 이어져 부정경쟁방지법 및 민사상 불법행위 책임에 대해서도 논의를 할 수 있는 사례를 제공하였다. 마지막으로 미국의 사례지만 HiQ Labs v. LinkedIn 사건도 이 글의 주제에는 매우 유용하다.

특히 형사책임에 관하여 ‘야놀자 판결’을 토대로 해서 크롤링 대상 서버에 접근하는 측면에서 정보통신망 침입(①), 크롤링 대상 자료의 성격과 크롤링의 방법의 측면에서 데이터베이스제작자의 권리 침해(②), 크롤링의 결과로서 정보통신망에 장애가 발생하였는지의 측면에서 컴퓨터등장애업무방해(③) 각각의 구성요건을 중심으로 살펴보았다. 크롤링 대상 웹 서버에 대한 접근권한의 유무가 ①의 구성요건해당성이 결정되기 때문에 여기서는 접근권한의 범위가 가장 중요하다. 전술한 바와 같이 접근권한은 해당 웹 사이트의 접근이 일반적으로 허용되어 있거나 접근에 특별한 제약이 없는 이상 이용약관과 보호조치와 같은 객관적 사정에 따라 엄격하게 판단되어야 한다. 다만 나는 이용약관과 보호조치 모두 웹 서버 관리자가 일방적으로 설정하는 것이고, 그 웹 서버에 접근하려고 하는 자가 이용약관의 내용을 숙지해

164) 크롤링에 관한 법학 분야의 연구에 대해서는 권세진, 이정훈, 이창무, “데이터 경제 시대에 있어서 웹 크롤링의 법적 인식에 관한 연구”, 한국산업보안연구, 제11권 제3호, 2021년, 85쪽 참조

야 한다거나, 보호조치 여부를 매번 확인해야 할 의무까지 인정할 수는 없다고 생각하지 않는다. 그래서 위 두 가지를 가지고 접근권한을 판단할 때에는 약관을 작성 또는 보호조치를 설정한 서버 관리자의 진정한 의사를 찾으려고 해서는 안 된다. 그것들을 통해서 외부로 드러난 객관화된 관리자의 의사를 찾아야 하고, 어떤 경우에도 인터넷이나 웹의 공공성을 전제로 해야 한다. 접근권한을 제한적으로 해석하거나 부여되었던 접근권한을 서버 관리자의 일방적인 의사로 철회하였다고 인정하는 데에는 매우 신중한 태도를 취해야 한다. 그렇게 된다면 크롤링에 대해서 정보통신망침입을 인정하게 되는 경우는 매우 줄어들게 될 것이다.

②는 웹 크롤링을 통해 크롤링 대상 웹 서버와 연결된 데이터베이스 전부가 아닌 일부분이 수집된 경우, 그것을 데이터베이스제작자의 권리 침해로 인정하기 위한 요건에 관한 것이다. '야놀자 판결'에 따르면 데이터베이스제작자의 권리 침해를 인정하기 위해서는 데이터베이스의 '상당한 부분'의 복제가 인정되어야 하고, '상당한 부분'의 복제는 반복적이거나 특정한 목적을 위하여 체계적으로 복제하여 데이터베이스의 통상적인 이용과 충돌하거나 데이터베이스제작자의 이익을 부당하게 해치는 경우에 인정될 수 있다. '상당한 부분'의 판단은 양적인 측면과 질적인 측면이 모두 검토되어야 한다. '야놀자 판결' 중 이 부분에 관한 설시에 나는 특별한 관심을 갖고 있는데, 왜냐하면 이것이 서문에서 밝힌 웹 크롤링을 주제로 선택하게 된 나의 호기심과 직접적으로 관련이 되기 때문이다. 사람이 일일이 손으로 할 때에는 아무런 책임을 지지 않는 것이 컴퓨터 프로그램 등 자동화 도구를 사용할 때에는 어떻게 달라질 수 있는지에 대한 하나의 답을 제시하고 있다.

‘야놀자 판결’의 설시 중 ‘반복적이고 체계적’이라는 부분이 자동화 도구의 사용을 포섭할 수 있다. 즉 반복적이고 체계적인 자동화 도구를 사용하게 되면 법적 책임을 질 수도 있다는 것이다. 이에 관해서는 추가적인 연구가 필요하다.

③은 ‘야놀자 판결’ 이전부터 대법원이 설시하던 법리로도 충분히 판단할 수 있는 부분이다. 실제로 ‘야놀자 판결’ 중 이 부분에 관한 설시는 기존의 법리를 그대로 인용하고 있을 뿐이다. 여기서 중요한 것은 ’부정한 명령‘과 ’정보처리에 장애 발생‘ 여부이다. ’부정한 명령‘은 해당 웹 사이트의 통상적인 이용과의 관계에서 보아야 할 것이나, 웹 크롤러의 활동이 통상적인 이용과 모습을 달리한다는 이유만으로 쉽사리 ’부정한 명령‘이라고 보아서는 안 된다. 앞서 본 접근권한과 같이 엄격하게 볼 필요가 있다. ’정보처리에 장애 발생‘ 여부는 일의적으로 판단하기 어렵다. 정보처리에 장애가 발생하였다고 인정할 수 있는 기준은 무엇인지, 웹 크롤러의 활동과 정보처리 장애 사이에 인과관계는 어떻게 판단할 것인지 등의 문제는 다분히 사실인정의 문제로 보이고. 이는 웹 크롤링의 정중함 정책(politeness policy)과 관련된다.

내가 이 글을 통해 일관되게 주장하는 것은 웹이 본질적으로 개방적이라는 점이다. 상호 정보의 전달을 위한 공개된 가상의 공간이 웹이기 때문이다. 그래서 웹에서의 교환되는 정보는 기본적으로 공개된 것이고, 비공개를 원할 때에는 그 정보를 웹에 개재하는 측에서 그러한 비공개 또는 공개 범위의 제한과 같은 조치를 취해야 할 것이지, 웹에 접근하는 사람에게 비공개 정보에 대한 어떠한 의무를 부담하게 하는 것은 부당하다. 비공개 정보를 웹에 개재하면서 비공개 또는 접

근 범위를 제한한다는 표시를 하는 것만으로는 이용자에게 아무런 영향을 줄 수 없다. 그것이 약관에 기재하였다고 하더라도 마찬가지이다.

빅데이터와 인공지능의 역할이 부상하고 있는 인더스트리 4.0에서 대량의 데이터를 수집하고 처리·분석하는 일은 대부분의 기업 활동에서 중요해졌고, 그러한 능력이 기업의 경쟁력을 좌우하고 있다.[165] 사업의 초기에는 거대한 양의 데이터로부터 가치를 추출해 낼 수 있을 만큼의 규모에 누가 먼저 도달하는지, 그리고 그 데이터로부터 가치를 누가 먼저 끌어내는지가 중요하다. 이미 거대한 양의 데이터를 보유하고 있는 사업자가 데이터의 독점하고자 한다면 이는 새로운 사업자의 시장참여를 방해하는 행위가 되어 경쟁을 저해하게 된다.

정보통신기술과 관련된 기술 즉 자동화 도구가 앞으로도 무수히 나올 것이다. 기술을 가치중립적이므로 그 어떤 기술도 그 자체만으로 법적 평가를 할 수 없다. 웹 크롤링도 마찬가지이다. 비유컨대 칼이 살인의 도구로 사용되었다고 하여 칼 자체를 금지할 수 없다. 어떤 도구도 도구 그 자체로는 법률적으로 의미가 없다. 도구가 사용되는 맥락과 상황, 사용자의 의도 또는 목적 그리고 사용의 결과와 같은 것이 법률적 평가의 대상이 될 뿐이다.

---

165) 이를 데이터 이코노미(data economy), 또는 데이터 주도 경제(data-driven economy)라고 하는데, 데이터를 기반으로 한 경제 또는 데이터가 사회와 경제를 움직이는 핵심적인 동력이 되는 새로운 시대를 의미하는 것이라고 한다(서울대 법과경제연구센터, 『데이터이코노미』, 한스미디어, 2017년, 4쪽, 245쪽)

- 참고문헌 -

【단행본】

서울대 법과경제연구센터, 『데이터이코노미』, 한스미디어, 2017.

오승종, 『저작권법강의(제2판)』, 박영사, 2018.

이창범, 황창근, 정필운, 『이론&실무 정보통신망법』, 박영사, 2021.

이해완, 『저작권법(제4판)』, 박영사, 2019.

정상조, 박준석, 『지식재산권법(제5판)』, 홍문사, 2020.

타쿠로 사사키, 김경록 역, 『데이터와 크롤링을 몰라도 엑셀 및 구글 스프레드시트로 쉽게 할 수 있는 웹 데이터 수집의 기술 입문편』, 한빛미디어, 2017.

허희성, 『신 저작권법 축조개설(하)』, 명문프리컴, 2011.

【한국방송통신대학교출판문화원 단행본】

강태원, 정광식, 『자료구조』, 2017.

김진욱, 이병래, 곽덕훈, 『운영체제』, 2021.

김진욱, 류대현, 김희천, 『컴퓨터보안 』, 2020.

김형근, 곽덕훈, 『C프로그래밍』, 2017.

김형근, 손진곤, 『컴퓨터구조』, 2021.

김희천, 『소프트웨어공학』, 2020.

손진곤, 길준민, 『정보통신망』, 2021.

이관용, 『HTML웹프로그래밍』, 2020.

정광식, 원유헌, 유헌창, 『프로그래밍언어론』, 2017.
정재화, 『클라우드컴퓨팅』, 2020.
최대우, 장영재, 이석호, 『데이터과학입문』, 2020.

【논문】

강정희, “웹 크롤링의 데이터 수집행위와 경쟁법적 논의-대법원 2022. 5. 12. 선고 2021도1533 판결을 중심으로”, 사법 61호, 사법발전재단, 2022.
권세진, 이정훈, 이창무, “데이터 경제 시대에 있어서 웹 크롤링(crawling)의 법적 인식에 관한 연구”, 한국산업보안연구 제11권 제3호, 2021.
김선태, “웹크롤러 기반의 개인정보 침해 점검 시스템에 대한 방법론 연구”, 숭실대학교 석사학위논문, 2016.
김종호, “빅데이터의 재산법상 보호 가능성에 관한 법적 고찰”, 법이론실무연구 제9권 제4호, 2021.
김현숙, “크롤링을 이용한 공개데이터 수집·활용의 법적 쟁점에 대한 비판적 검토”, 강원법학 제61권, 2020.
김희숙, 한나, 임숙자, “빅데이터 분석 기반의 정보 검색을 위한 웹 크롤러 서비스 구현”, 디지털콘텐츠학회논문지 Vol. 18, No. 5, 2017.
염호준, “유럽연합의 데이터베이스 보호에 관한 지침과 최근의 동향”, 세계의 언론법제 2006년 상권(통권 제19호)
유대종, “웹검색 서비스와 ISP 책임에 관한 소고”, 장작과 권리, 2007.
유민종, “사이버범죄협약과 국내 법제의 양립 가능성 연구”, 서울대학

교 석사학위논문, 2019.
이주호, “크롤링에 의한 민감 정보 침해에 대응하는 로그인 강화 연구”, 숭실대학교 석사학위논문, 2020.
장준영, 임경대, 이상진, “HTML 및 URL 특징을 이용한 유해사이트 수집 시스템”, 디지털포렌식연구 제16권 제1호, 2022.
장현호, 정경식, 이후기, “분산형 병렬 크롤러 설계 및 구현”, 융합보안논문지 제19권 제3호, 2019.
정현순, “크롤링에 의한 데이터 수집이 정보통신망 침입에 해당하는지 여부-대상판결: 대법원 2022. 5. 12. 선고 2021도1533 판결”, 사법61호, 사법발전재단, 2022.
조성훈, “정보통신망 침입에 대한 연구-정보통신망 이용촉진 및 정보보호 등에 관한 법률 제48조를 중심으로”, 법조 Vol. 687, 2013. 12.
최상진, “경쟁사의 무단 크롤링에 대한 법적 대응방안에 관한 연구”, Law & Technology 제17권 제1호, 2021.
한지영, 데이터베이스의 법적 보호에 관한 연구, 서울대학교 박사학위논문, 2005.
황태희, “웹 크롤링의 경쟁법적 고찰”, 경쟁법연구 43권, 2021.
홍성학, “웹 크롤링을 이용한 스마트 가격 추적기의 구현”, 서울과학기술대학교 석사학위논문, 2015.

【인터넷 자료】

Christopher Olston, Marc Najork, 『Web Crawling』, Foundations and Trends® in Information Retrieval, Vol. 4, No. 3, 2010.
(http://infolab.stanford.edu/~olston/publications/crawling_survey.pdf: 2022. 8. 5. 방문)

Marc Najork, "Web Crawler Architecture"
(https://marc.najork.org/pdfs/eds2009a.pdf: 2022. 8. 5. 방문)

Md. Abu Kausar, V.S. Dhaka, Sanjeev Kumar Singh, "Web Crawler: A Review", International Journal of Computer Applications Volume 63, No. 2, 2013.
(https://research.ijcaonline.org/volume63/number2/pxc3885125.pdf: 2022. 11. 1. 방문)

Carlos Castillo, "Effective Web Crawling", University of Chille 박사 학위 논문, 2004.
(https://chato.cl/papers/crawling_thesis/effective_web_crawling.pdf: 2022. 11. 1. 방문)

Orin S. Kerr. "Norms of Computer Trespass", Columbia Law Review V. 116, No. 4

(https://columbialawreview.org/content/norms-of-computer-trespass/ 2022. 8. 6. 방문)

[Astract]

A Study on Web Crawling - Focused on its working mechanism and legal responsibility

Taekyun Kim
Master of Science in Digital Forensics
Department of Mathematical Information Science
The Graduate School of Convergence Science and Technology
Seoul National University

Web Crawling means a way of collecting web data using a program which is called Crawler or Spider. A Crawler, given seed URLs, downloads the web pages associated with these URLs, extracts hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks. Web Crawling is now broadly being used as one of the most effective and useful method in data collecting wherever data becomes a key factor. Especially in decision-making process in marketing or business strategy, Web Crawling is now sine qua non with the emerging of big data or artificial intelligence.

While Web Crawling is getting more and more important, few studies on legal responsibilities of using it can be found. This article focuses on the working mechanism and legal responsibilities of using Web Crawler. Recently the Supreme Court has found, in the case where the defendants, the employees of a company running a accommodation information offering service accessed the competiter's mobile application server and copied the database such as a list of accommodation through their crawling computer program, that the defendants are not guilty. The Case showed new legal principles, ① whether the service provider has set any limitation on access right to network should be determined by object things such as protective actions or terms and conditions of use, ② significant copy of database should be determined in the aspects of quantity and quality both. And The Case sentenced not guilty to the charges(invasion of network, piracy of database and obstruction of business). The Case was the first judgement of

the Supreme Court on crawling data collection.

In this article, The Case will be examined thoroughly in the 3 aspects(invasion of network, piracy of database, obstruction of business) including Unfair Competition Prevention Act and competition laws. The conclusion of this article is as follows. Legal responsibilities for using web crawlers should be judged by the standards of the scope of access to networks, whether the rights of database producers are violated and whether there is any obstruction of business, which must be based on the circumstances where web crawlers are used, the intetion of the user and the result of the use.

keywords: web crawling, web crawler, data, data-collecting, invasion of network, infringement of database producers' right, obstruction of business, act of unfair competition, unfair business practice, abuse of market-dominant position

Student Number: 2021-24496