스크래핑 위협과 대응 방향

2023.03.31

SCR;PTERS

스크래핑 피해 유형

1. 트래픽으로 인한 피해

✓ **국지적으로 집중되는 스크래핑 부하**
- 카드사 홈페이지 마비 – 핀테크의 이벤트
- 등본 발급 장애 – 인기 부동산 청약일

✓ **관리 비용의 증가**
- 서버 및 솔루션 구매 / 인증 및 관리 비용 등

2. 동일한 인증정보로 인한 피해 우려

✓ **공공 문서의 발급을 대행해 주는 사설업체**
- 문서 발급을 위해 인증서를 제공한 개인
- 인증서를 이용한 금융/의료 등 정보 조회

✓ **인증정보를 저장하고 있는 스크래핑 업체**
- 인증정보 암호화는 하고 있는지..
- 개인이 의도치 않은 정보수집이 가능

3. 서비스의 악의적 이용으로 인한 피해

✓ **타사의 신규 서비스, 알고 보니 우리 서비스**
- 보험 보장 분석 서비스

✓ **우리 정보를 이용한 타사의 신규 서비스**
- 2022년 이슈화 되었던 대환대출 서비스

4. 정보 탈취로 인한 피해

✓ **구인 / 숙박 업소 / 부동산 사이트 등**
- 야놀자 / 여기어때 소송 중
- LinkedIn / 하이큐랩스 소송
- 네이버 / 부동산 중개 스타트업 소송 중

✓ **타사의 서비스에 포함된 정보를 이용한 사업**
- 스크래핑으로 긁어간 후 자체 서비스

5. 어이없는 유형

✓ **하나의 웹 페이지를 동시에 10번씩 조회**
- 타 회사 스크래핑 모듈의 버그로 확인
- 버그 수정 요청 후, 정상화

✓ **개편되어 없어진 웹 페이지에 대한 요청 지속**
- 서버에서 오류(Exception) 지속
- 관리되지 않는 스크래핑 모듈

6. 개인적인 피해 유형

✓ **산장 예약, 골프 부킹, 티켓 구매가 3초면 끝?**
- 그리고 등장하는 암표 매매

✓ **한달만에 로그인 했더니 최근 로그인이 어제?**
- 찜찜함은 개인의 몫

✓ **사람보다 더 빠른 캡차 정보 입력**
- 캡차 농장(Farm)의 등장

마이데이터 사업자 - 스크래핑 금지

금융위원회 보도자료(2022.01.04)

금융위원회

□ '21.12.1일부터 API 방식의 금융 마이데이터 시범서비스 실시

○ 동 시범서비스 기간 동안 시스템 안정화, 데이터 정합성 제고, 사설인증 및 정보제공기관 확대 등 개선필요사항은 신속하게 보완

□ '22.1.5일 16시부터는 스크래핑이 금지되고 33개 마이데이터 사업자가 API 방식을 통해 금융 마이데이터 서비스를 제공할 예정

3 마이데이터 전면시행('22.1.5일) 주요내용

□ (마이데이터 사업자) '22.1.5일부터 스크래핑이 전면 금지되고 마이데이터 사업자는 모든 이용자에게 API 방식으로만 마이데이터 서비스를 제공할 수 있습니다.

2. 마이데이터 사업자 측면

❶ (안정적 사업기반 확보) 이용자가 정보전송 요청시 정보제공자에게 정보제공의무가 부여됨에 따라 마이데이터 사업자가 필요한 정보를 빠르고 안정적으로 제공받을 수 있는 환경 조성

※ 기존 스크래핑 방식 하에서는 고객정보를 보유한 정보제공자가 정보보호 등을 이유로 스크래핑 방지기술 적용시 마이데이터 서비스의 안정적 제공에 차질

2022년 1월 5일, 마이데이터 시행과 함께

스크래핑 금지

하지만 스크래핑은…

지속적인 스크래핑 요청 유입

정책적 요인

스크래핑 관련 정책

- 마이데이터 사업자만 스크래핑 금지
- 비 마이데이터 사업자는 스크래핑 가능
- 마이데이터 사업자 수: 64개(2023.02)

기술적 요인

익명성

- 스크래핑은 정확한 출처 확인이 어려움
- 내가 해도, 상대는 내가 한 줄 모른다?
- 스크래핑 데이터 ≠ 마이데이터 API 데이터

경제적 요인

데이터 수집 비용

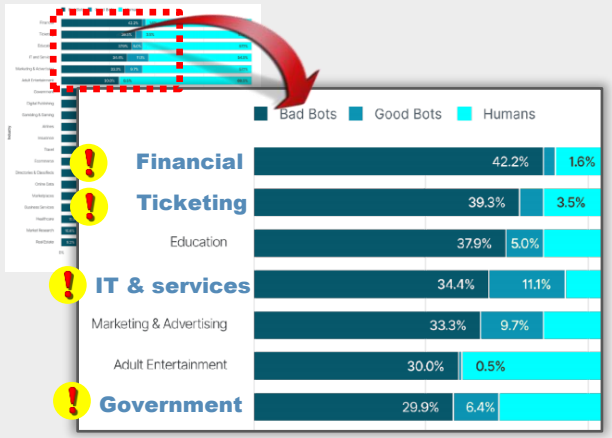
- 무료 스크래핑 vs 유료 마이데이터 API
- API 라는 이름의 스크래핑 서비스
- 스크래핑을 이용한 핀테크 기업의 증가

마이데이터 사업자만 금지 | **익명성 & 데이터 확보** | **이미 보유중인 무료 기술**

스크래핑의 지속적인 유입

스크래핑 관련 통계 및 해외 동향

Scraping / OWASP / JPMorgan



<출처: "Bad Bot Report 2019" by https://www.globaldots.com>

스크래핑이 전체 웹 트래픽의 30~42%

OWASP Open Web Application Security Project

OAT-011 Scraping

Definition

OWASP Automated Threat (OAT) Identity Number

OAT-011

Threat Event Name

Scraping

Summary Defining Characteristics

Collect application content and/or other data for use elsewhere.

<출처: Open Web Application Security Project>

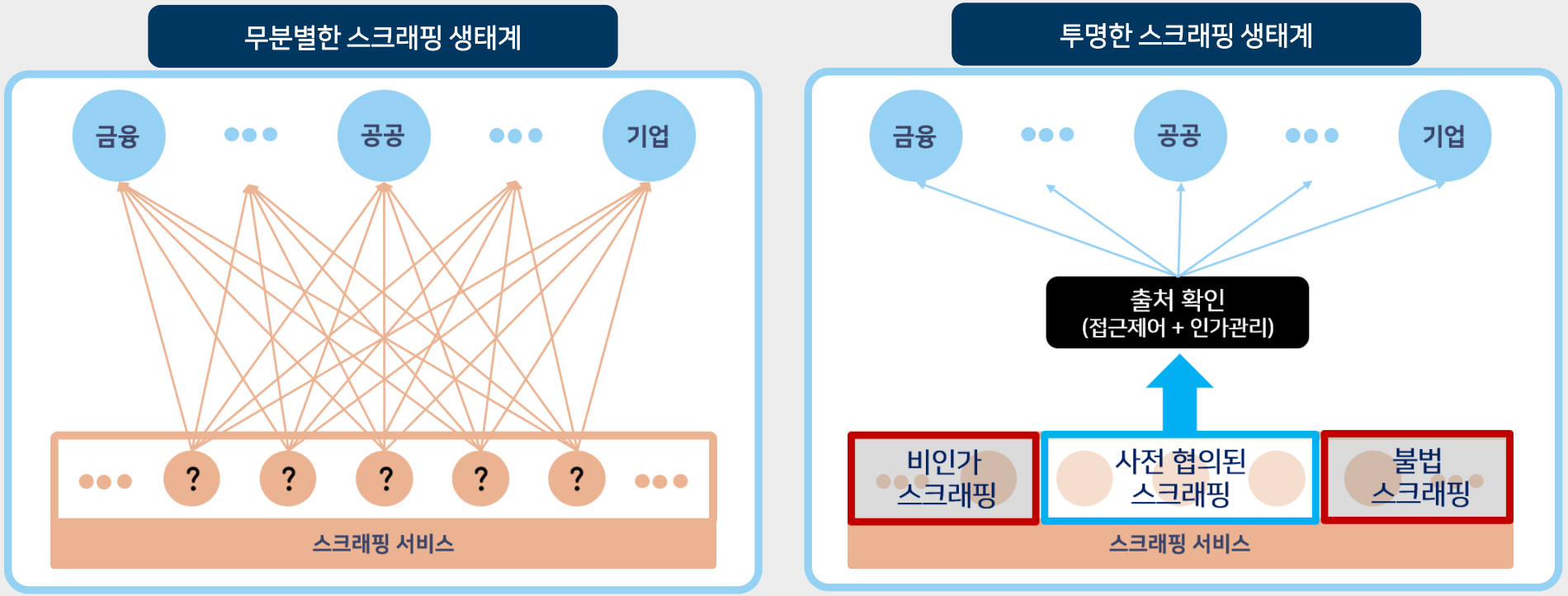
스크래핑을 자동화된 위협으로 지정

<출처: ItNews, Feb 14 2020>

해외 금융기관의 스크래핑 대응

대응 방향

익명성을 제거하고, 상호 협의된 상태에서만 가능한 스크래핑



스크래핑 관리는 데이터 경쟁력 확보 및 마이데이터 활성화

출처를 알면 …

1. 트래픽으로 인한 피해

✓ **국지적으로 집중되는 스크래핑 부하**
- 카드사 홈페이지 마비 – 핀테크의 이벤트
- 등본 발급 장애 – 인기 부동산 청약일

✓ **관리 비용의 증가**
- 서버 및 솔루션 구매 / 인증 및 관리 비용 등

✓ **선별적으로 스크래핑 허용 / 부하관리**

2. 동일한 인증정보로 인한 피해 우려

✓ **공공 문서의 발급을 대행해 주는 사설업체**
- 문서 발급을 위해 인증서를 제공한 개인
- 인증서를 이용한 금융/의료 등 정보 조회

✓ **인증정보를 저장하고 있는 스크래핑 업체**
- 인증정보 암호화는 하고 있는지..
- 개인이 의도치 않은 정보수집이 가능

✓ **기관 A는 허용했지만, 금융 A는 허용 안함**

3. 서비스의 악의적 이용으로 인한 피해

✓ **타사의 신규 서비스, 알고 보니 우리 서비스**
- 보험 보장 분석 서비스

✓ **우리 정보를 이용한 타사의 신규 서비스**
- 2022년 이슈화 되었던 대환대출 서비스

✓ **협의된 적이 없는 요청은 처리 불가**

4. 정보 탈취로 인한 피해

✓ **구인 / 숙박 업소 / 부동산 사이트 등**
- 야놀자 / 여기어때 소송 중
- LinkedIn / 하이큐랩스 소송
- 네이버 / 부동산 중개 스타트업 소송 중

✓ **타사의 서비스에 포함된 정보를 이용한 사업**
- 스크래핑으로 긁어간 후 자체 서비스

✓ **협의된 적이 없으면 긁어가지 못함**

5. 어이없는 유형

✓ **하나의 웹 페이지를 동시에 10번씩 조회**
- 타 회사 스크래핑 모듈의 버그로 확인
- 버그 수정 요청 후, 정상화

✓ **개편되어 없어진 웹 페이지에 대한 요청 지속**
- 서버에서 오류(Exception) 지속
- 관리되지 않는 스크래핑 모듈

✓ **잘 관리되는 스크래핑 모듈과 협력**

6. 개인적인 피해 유형

✓ **산장 예약, 골프 부킹, 티켓 구매가 3초면 끝?**
- 그리고 등장하는 암표 매매

✓ **한달만에 로그인 했더니 최근 로그인이 어제?**
- 찜찜함은 개인의 몫

✓ **사람보다 더 빠른 캡차 정보 입력**
- 캡차 농장(Farm)의 등장

감사합니다.

SCR;PTERS

Contact | 이 재 영
kwangbul@scripters.kr