(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년10월27일
(11) 등록번호 10-2595303
(24) 등록일자 2023년10월24일

(51) 국제특허분류(Int. Cl.)
G06F 11/34 (2006.01) *G06F 21/56* (2013.01)
H04L 9/40 (2022.01)
(52) CPC특허분류
G06F 11/3414 (2013.01)
G06F 11/3419 (2013.01)
(21) 출원번호 10-2021-0051273
(22) 출원일자 2021년04월20일
심사청구일자 2021년04월20일
(65) 공개번호 10-2022-0144666
(43) 공개일자 2022년10월27일
(56) 선행기술조사문헌
CN104391979 A*
KR1020180047353 A*
KR101978898 B1*
US20090288169 A1
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 스크립터스
서울특별시 강남구 테헤란로64길 13, 305호(대치동)
(72) 발명자
이재영
서울특별시 관악구 남부순환로241길 29, 401호 (봉천동)
(74) 대리인
이재록

전체 청구항 수 : 총 10 항 심사관 : 김계준

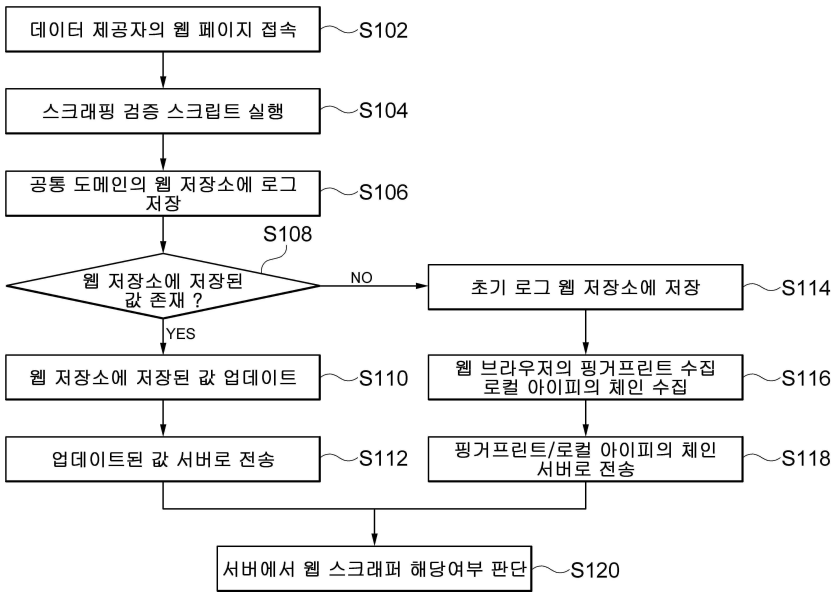
(54) 발명의 명칭 **웹 스크래핑 탐지 방법 및 이를 수행하기 위한 서버**

(57) 요 약

웹 스크래핑 탐지 방법 및 이를 수행하기 위한 서버가 제공된다. 본 발명의 일 실시예에 따른 웹 스크래핑 탐지 방법은 클라이언트의 웹 브라우저에서, 스크래핑 검증 스크립트(scraping validation script)가 삽입된 각 데이터 제공자의 웹 페이지에 접속할 때마다 상기 스크래핑 검증 스크립트가 실행되는 단계; 상기 클라이언트의 웹

(뒷면에 계속)

대 표 도



브라우저에서, 상기 스크래핑 검증 스크립트가 실행될 때마다 상기 스크래핑 검증 스크립트와 관련된 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 저장하는 단계; 상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 수집하는 단계; 및 상기 데이터 제공자의 서버에서, 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 이용하여 상기 클라이언트가 웹 스크래퍼(web scraper)인지의 여부를 판단하는 단계를 포함한다.

(52) CPC특허분류

G06F 21/566 (2013.01)

H04L 63/30 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1425146700
과제번호	S2969559
부처명	중소벤처기업부
과제관리(전문)기관명	중소기업기술정보진흥원
연구사업명	2020년도 창업성장기술개발사업 디딤돌 창업과제 2차
연구과제명	자동화된 프로그램에 의한 데이터 수집 탐지 및 차단 서비스 개발
기 여 율	1/1
과제수행기관명	주식회사 스크립터스
연구기간	2020.12.31 ~ 2021.12.30

명 세 서

청구범위

청구항 1

클라이언트의 웹 브라우저에서, 스크래핑 검증 스크립트(scraping validation script)가 삽입된 각 데이터 제공자의 웹 페이지에 접속할 때마다 상기 스크래핑 검증 스크립트가 실행되는 단계;

상기 클라이언트의 웹 브라우저에서, 상기 스크래핑 검증 스크립트가 실행될 때마다 상기 스크래핑 검증 스크립트와 관련된 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 저장하는 단계;

상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 수집하는 단계; 및

상기 데이터 제공자의 서버에서, 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 이용하여 상기 클라이언트가 웹 스크래퍼(web scraper)인지의 여부를 판단하는 단계를 포함하며,

상기 스크래핑 검증 스크립트는, 상기 공통 도메인의 스크립트로서 상기 각 데이터 제공자가 제공하는 상기 웹 페이지에 프레임(frame) 형태로 삽입되어 상기 클라이언트의 웹 브라우저에서 상기 웹 페이지에 접속할 때마다 상기 공통 도메인으로부터 로딩되어 실행되고,

상기 스크래핑 검증 스크립트의 실행과 관련된 로그는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수 및 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수 중 하나 이상을 포함하는, 웹 스크래핑 탐지 방법.

청구항 2

삭제

청구항 3

청구항 1에 있어서,

상기 클라이언트가 웹 스크래퍼인지의 여부를 판단하는 단계는, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수가 제1 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 웹 스크래핑 탐지 방법.

청구항 4

청구항 3에 있어서,

상기 클라이언트가 웹 스크래퍼인지의 여부를 판단하는 단계는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수가 제2 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 웹 스크래핑 탐지 방법.

청구항 5

청구항 1에 있어서,

상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트의 웹 브라우저에 대한 핑거프린트(fingerprint)를 수집하는

단계; 및

상기 데이터 제공자의 서버에서, 동일한 값을 갖는 상기 핑거프린트가 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는 단계를 더 포함하는, 웹 스크래핑 탐지 방법.

청구항 6

청구항 5에 있어서,

상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트가 속한 로컬 네트워크 내에서 상기 클라이언트를 식별하기 위한 로컬 아이피의 체인(chain)을 수집하는 단계; 및

상기 데이터 제공자의 서버에서, 동일한 값을 갖는 상기 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는 단계를 더 포함하는, 웹 스크래핑 탐지 방법.

청구항 7

스크래핑 검증 스크립트가 삽입된 각 데이터 제공자의 웹 페이지에 클라이언트의 웹 브라우저가 접속할 때마다 상기 스크래핑 검증 스크립트가 실행되고, 상기 스크래핑 검증 스크립트가 실행될 때마다 상기 스크래핑 검증 스크립트와 관련된 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 저장됨에 따라 상기 클라이언트가 웹 스크래퍼인지의 여부를 판단하는 데이터 제공자의 서버로서,

상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 수집하는 수집부; 및

업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 이용하여 상기 클라이언트가 웹 스크래퍼(web scraper)인지의 여부를 판단하는 판단부를 포함하며,

상기 스크래핑 검증 스크립트는, 상기 공통 도메인의 스크립트로서 상기 각 데이터 제공자가 제공하는 상기 웹 페이지에 프레임(frame) 형태로 삽입되어 상기 클라이언트의 웹 브라우저에서 상기 웹 페이지에 접속할 때마다 상기 공통 도메인으로부터 로딩되어 실행되고,

상기 스크래핑 검증 스크립트의 실행과 관련된 로그는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수 및 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수 중 하나 이상을 포함하는, 서버.

청구항 8

삭제

청구항 9

청구항 7에 있어서,

상기 판단부는, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수가 제1 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 서버.

청구항 10

청구항 9에 있어서,

상기 판단부는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수가 제2 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 서버.

청구항 11

청구항 7에 있어서,

상기 수집부는, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트의 웹 브라우저에 대한 핑거프린트(fingerprint)를 수집하고,

상기 판단부는, 동일한 값을 갖는 상기 핑거프린트가 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 서버.

청구항 12

청구항 7에 있어서,

상기 수집부는, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트가 속한 로컬 네트워크 내에서 상기 클라이언트를 식별하기 위한 로컬 아이피의 체인(chain)을 수집하고,

상기 판단부는, 동일한 값을 갖는 상기 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 서버.

발명의 설명

기 술 분 야

[0001] 본 발명의 실시예들은 웹 스크래핑 탐지 기술과 관련된다.

배 경 기 술

[0003] 웹 스크래핑(Web Scraping)이란 프로그램에 의해 자동으로 웹 시스템에 접속하여 필요한 자료를 추출하여 가져오는 기술이나 프로그램을 의미한다. 일반적으로, 웹 페이지에는 개인정보나 고객 데이터 등과 같이 제3자가 수집해서는 안되는 민감한 데이터가 포함될 수 있다. 그러나, 최근 들어 자동화된 툴을 이용하여 개인정보와 같은 민감한 정보를 수집하거나, 방문 실적에 따른 광고 수익을 노리는 자들이 웹 페이지에 접속하지 않으면서 데이터만 수집해 가는 경우가 종종 발생한다. 공공기관이나, 은행, 카드사, 증권사 등과 같은 금융기관에서는 이러한 웹 스크래핑으로 인한 시스템 안정성 문제가 지속적으로 제기되고 있다. 특히, 이러한 웹 스크래핑이 특정 시간에 집중되는 경우 공공기관이나 금융기관 등과 같은 데이터 제공자가 제공하는 서비스가 느려지거나 데이터 제공자의 서버가 다운되기도 한다. 이와 같이, 웹 스크래핑으로 인해 데이터 제공자의 시스템 운영비용 및 불안정성이 증가하고 있으며 이에 따라 웹 스크래핑의 탐지에 대한 요구가 발생하고 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국공개특허공보 제10-2015-0085716호(2015.07.24)

발명의 내용

해결하려는 과제

[0006] 본 발명의 실시예들은 공통 도메인과 웹 브라우저의 웹 페이지 접속 패턴(또는 행동 패턴)을 이용하여 웹 스크래핑을 탐지하기 위한 것이다.

과제의 해결 수단

[0008] 예시적인 실시예에 따르면, 클라이언트의 웹 브라우저에서, 스크래핑 검증 스크립트(scraping validation script)가 삽입된 각 데이터 제공자의 웹 페이지에 접속할 때마다 상기 스크래핑 검증 스크립트가 실행되는 단계; 상기 클라이언트의 웹 브라우저에서, 상기 스크래핑 검증 스크립트가 실행될 때마다 상기 스크래핑 검증 스크립트와 관련된 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 저장하는 단계; 상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 수집하는 단계; 및 상기 데이터 제공자의 서버에서, 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 이용하여 상기 클라이언트가 웹 스크래퍼(web scraper)인지의 여부를 판단하는 단계를 포함하는, 웹 스크래핑 탐지 방법이 제공된다.

[0009] 상기 스크래핑 검증 스크립트의 실행과 관련된 로그는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수 및 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수 중 하나 이상을 포함할 수 있다.

[0010] 상기 클라이언트가 웹 스크래퍼인지의 여부를 판단하는 단계는, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수가 제1 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단할 수 있다.

[0011] 상기 클라이언트가 웹 스크래퍼인지의 여부를 판단하는 단계는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수가 제2 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단할 수 있다.

[0012] 상기 웹 스크래핑 탐지 방법은, 상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트의 웹 브라우저에 대한 핑거프린트(fingerprint)를 수집하는 단계; 및 상기 데이터 제공자의 서버에서, 동일한 값을 갖는 상기 핑거프린트가 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는 단계를 더 포함할 수 있다.

[0013] 상기 웹 스크래핑 탐지 방법은, 상기 데이터 제공자의 서버에서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트가 속한 로컬 네트워크 내에서 상기 클라이언트를 식별하기 위한 로컬 아이피의 체인(chain)을 수집하는 단계; 및 상기 데이터 제공자의 서버에서, 동일한 값을 갖는 상기 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는 단계를 더 포함할 수 있다.

[0014] 다른 예시적인 실시예에 따르면, 스크래핑 검증 스크립트가 삽입된 각 데이터 제공자의 웹 페이지에 클라이언트의 웹 브라우저가 접속할 때마다 상기 스크래핑 검증 스크립트가 실행되고, 상기 스크래핑 검증 스크립트가 실행될 때마다 상기 스크래핑 검증 스크립트와 관련된 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 저장됨에 따라 상기 클라이언트가 웹 스크래퍼인지의 여부를 판단하는 데이터 제공자의 서버로서, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 수집하는 수집부; 및 업데이트된 상기 스크래핑 검증 스크립트의 실행과 관련된 로그를 이용하여 상기 클라이언트가 웹 스크래퍼(web scraper)인지의 여부를 판단하는 판단부를 포함하는, 서버가 제공된다.

[0015] 상기 스크래핑 검증 스크립트의 실행과 관련된 로그는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수 및 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수 중 하나 이상을 포함할 수 있다.

[0016] 상기 판단부는, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트가 실행된 횟수가 제1 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단할 수 있다.

[0017] 상기 판단부는, 상기 스크래핑 검증 스크립트가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트가 실행된 총 횟수가 제2 임계치를 초과하는 경우 상기 클라이언트를 웹 스크래퍼로 판단할 수 있다.

[0018] 상기 수집부는, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트의 웹 브라우저에 대한 핑거프린트(fingerprint)를 수집하고, 상기 판단부는, 동일한 값을 갖는 상기 핑거프린트가 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단할 수 있다.

[0019] 상기 수집부는, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로

저장될 때마다 상기 클라이언트가 속한 로컬 네트워크 내에서 상기 클라이언트를 식별하기 위한 로컬 아이피의 체인(chain)을 수집하고, 상기 판단부는, 동일한 값을 갖는 상기 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단할 수 있다.

발명의 효과

[0021] 본 발명의 실시예들에 따르면, 공통 도메인의 스크래핑 검증 스크립트를 각 데이터 제공자의 웹 페이지에 삽입한 후 상기 스크래핑 검증 스크립트를 통해 클라이언트의 웹 페이지 접속 패턴을 모니터링하여 클라이언트가 웹 스크래퍼에 해당하는지의 여부를 효율적으로 탐지할 수 있다.

[0022] 또한, 본 발명의 실시예들에 따르면, 공격자가 공통 도메인의 웹 저장소를 초기화(삭제)하는 것에 대비하여 동일한 값의 웹 브라우저 핑거프린트 또는 로컬 아이피의 체인이 설정된 횟수 이상 반복 수집되는 경우 클라이언트를 웹 스크래퍼로 판단함으로써, 웹 스크래퍼의 탐지 확률을 보다 향상시킬 수 있다.

도면의 간단한 설명

[0024] 도 1은 일반적인 사용자가 웹 브라우저를 통해 데이터 제공자의 웹 페이지에 접속하는 예시를 나타낸 도면

도 2는 자동화 프로그램의 웹 브라우저가 웹 스크래핑을 위해 데이터 제공자의 웹 페이지에 접속하는 예시를 나타낸 도면

도 3은 본 발명의 일 실시예에 따른 웹 스크래핑 탐지 시스템의 상세 구성을 나타낸 도면

도 4는 본 발명의 일 실시예에 따른 웹 스크래핑 탐지 방법을 설명하기 위한 흐름도

도 5는 본 발명의 일 실시예에 따른 단위 시간 값을 계산하는 예시

도 6은 본 발명의 일 실시예에 따른 서버의 상세 구성을 나타낸 블록도

도 7은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

발명을 실시하기 위한 구체적인 내용

[0025] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0026] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

[0028] 도 1은 일반적인 사용자가 웹 브라우저를 통해 데이터 제공자의 웹 페이지에 접속하는 예시를 나타낸 도면이다.

[0029] 도 1을 참조하면, 사용자는 사용자 단말(50)의 웹 브라우저를 통해 각 데이터 제공자가 제공하는 웹 페이지(60)에 접속할 수 있다. 본 실시예들에 있어서, 데이터 제공자는 사용자가 요청한 데이터나 서비스를 제공하는 은행, 카드사, 보험사, 공공기관 등이 될 수 있다. 사용자는 자신이 원하는 데이터 또는 서비스를 제공 받기 위해 데이터 제공자가 제공하는 웹 페이지(60) 중 하나 이상을 선택하여 순차적으로 접속할 수 있다. 예를 들어, 사용자는 A 은행의 웹 페이지(60)에 접속하여 A 은행과 관련된 데이터 또는 서비스를 제공 받을 수 있으며, 이후 B 카드사의 웹 페이지(60)에 접속하여 B 카드사와 관련된 데이터 또는 서비스를 제공 받을 수 있다.

[0030] 반면, 후술할 바와 같이, 웹 스크래핑을 수행하는 공격자 단말(70)의 웹 페이지(60) 접속 패턴은 도 1에 도시된 일반적인 사용자의 웹 페이지 접속 패턴과 상이하다.

[0032] 도 2는 자동화 프로그램의 웹 브라우저가 웹 스크래핑을 위해 데이터 제공자의 웹 페이지에 접속하는 예시를 나타낸 도면이다.

[0033] 도 2를 참조하면, 공격자 단말(70)의 자동화 프로그램은 웹 브라우저를 통해 각 데이터 제공자가 제공하는 웹 페이지(60)에 접속할 수 있다. 본 실시예들에 있어서, 자동화 프로그램은 웹 스크래핑을 위해 자동으로 웹 페이지(60)에 접속하여 필요한 자료를 추출하여 가져오는 소프트웨어 또는 애플리케이션으로서, 예를 들어 스크래퍼(scraper), 봇(bot) 등이 될 수 있다. 상기 자동화 프로그램은 웹 스크래핑을 수행하고자 하는 공격자 단말(70)에 장착될 수 있다. 자동화 프로그램은 HTTP request에 포함되는 파라미터 키, 파라미터 값 및 URL을 미리 가지고 있으며, 이를 통해 각 웹 페이지(60)에 자동으로 접속할 수 있다.

[0034] 이때, 자동화 프로그램은 도 1에서와 달리 각 데이터 제공자가 제공하는 모든 웹 페이지(60)에 순차적 또는 동시 접속하는 패턴을 보인다. 즉, 일반적인 사용자가 자신이 필요한 웹 페이지(60)에만 선택적으로 접속하는 것과 달리, 자동화 프로그램은 각 데이터 제공자가 제공하는 웹 페이지(60)에 모두 접속하는 패턴을 보인다.

[0035] 이하에서는, 이러한 자동화 프로그램의 웹 페이지(60) 접속 패턴을 고려하여 웹 스크래퍼를 효율적으로 탐지하는 웹 스크래핑 탐지 시스템(100)에 대해 살펴보기로 한다.

[0037] 도 3은 본 발명의 일 실시예에 따른 웹 스크래핑 탐지 시스템의 상세 구성을 나타낸 도면이다.

[0038] 도 3에 도시된 바와 같이, 웹 스크래핑 탐지 시스템(100)은 클라이언트(102), 및 웹 페이지(104)를 제공하는 데이터 제공자를 포함한다.

[0039] 클라이언트(102)는 사용자나 공격자가 소지하는 단말로서, 예를 들어 스마트폰, 데스크탑, 태블릿 PC, PDA(Personal Digital Assistant), 노트북 등이 될 수 있다.

[0040] 데이터 제공자는 데이터나 서비스를 제공하는 은행, 카드사, 보험사, 공공기관 등이 될 수 있으며, 상기 데이터나 서비스 제공을 위한 웹 페이지(104)를 제공할 수 있다.

[0041] 본 실시예들에 있어서, 각 데이터 제공자가 제공하는 웹 페이지(104)에는 기 정의된 스크래핑 검증 스크립트(scraping validation script)(106)가 삽입될 수 있다. 스크래핑 검증 스크립트(106)는 기 설정된 공통 도메인(common domain)의 스크립트로서, 각 데이터 제공자가 제공하는 웹 페이지(104)에 프레임(frame) 형태(예를 들어, iframe)로 삽입되어 클라이언트(102)가 상기 웹 페이지(104)에 접속할 때마다 상기 공통 도메인으로부터 로딩되어 실행될 수 있다. 여기서, 스크립트는 소스 코드를 컴파일(compile)하지 않고도 실행할 수 있는 프로그래밍 언어로서, 예를 들어 자바스크립트 등이 될 수 있다.

[0043] 상기 공통 도메인 및 스크래핑 검증 스크립트(106)의 예시는 아래와 같다.

[0045] * 공통 도메인

[0046] https://common.com

[0048] * 스크래핑 검증 스크립트(106)

[0049] validation.js

[0051] 상기 스크래핑 검증 스크립트(106)는 예를 들어, 아래와 같은 형태로 각 웹 페이지(104)에 삽입될 수 있다.

[0053] * 스크래핑 검증 스크립트의 웹 페이지 삽입 형태

[0054] <script src="https://common.com/validation.js"></script>

[0056] 이 경우, 클라이언트(102)의 웹 브라우저에서 각 데이터 제공자의 웹 페이지(104)에 접속할 때마다 상기 스크래핑 검증 스크립트(106)가 실행된다. 예를 들어, 클라이언트(102)의 웹 브라우저에서 A 은행의 웹 페이지(104), B 은행의 웹 페이지(104),… A 카드사의 웹 페이지(104), B 카드사의 웹 페이지(104),… A 보험사의 웹 페이지(104), B 보험사의 웹 페이지(104)…에 접속할 때마다 상기 스크래핑 검증 스크립트(106)가 반복적으로 실행될 수 있다.

[0057] 또한, 상기 스크래핑 검증 스크립트(106)가 실행될 때마다 상기 스크래핑 검증 스크립트(106)와 관련된 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 저장될 수 있다.

[0058] 일반적으로, 클라이언트(102)의 웹 브라우저마다 도메인별로 웹 저장소(108)가 나누어져 있다. 예를 들어, A 은행 도메인의 웹 저장소(108a), B 은행 도메인의 웹 저장소(108b),… 공통 도메인의 웹 저장소(108c)는 각각 별

개로 구비된다. 이때, 웹 브라우저의 보안 정책상 A 은행 도메인의 웹 저장소(108a)에는 클라이언트(102)가 B 은행의 웹 페이지(104)에 접속한 기록이 저장되지 않으며, B 은행 도메인의 웹 저장소(108b)에는 클라이언트(102)가 A 은행의 웹 페이지(104)에 접속한 기록이 저장되지 않는다. 이에 따라, 본 실시예들에서는 각 웹 페이지(104)에 공통 도메인의 스크래핑 검증 스크립트(106)를 삽입함으로써, 클라이언트(102)가 각 웹 페이지(104)에 접속할 때마다 상기 공통 도메인으로부터 스크래핑 검증 스크립트(106)가 로딩되어 실행될 수 있도록 하였다. 이 경우, 상기 스크래핑 검증 스크립트(106)가 실행될 때마다 상기 스크래핑 검증 스크립트(106)와 관련된 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 저장되며, 이에 따라 클라이언트(102)가 접속한 모든 웹 페이지(104)에 대한 기록(예를 들어, 모든 은행 웹페이지, 모든 카드사 웹 페이지, 모든 보험사 웹 페이지, 모든 공공기관의 웹 페이지에 대한 기록)이 남게 된다.

[0059] 또한, 데이터 제공자의 서버(미도시)는 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그를 이용하여 클라이언트(102)가 웹 스크래퍼(web scraper)인지의 여부를 판단할 수 있다. 서버의 웹 스크래핑 탐지 방법은 도 4 및 도 5를 참조하여 구체적으로 후술하기로 한다.

[0061] 도 4는 본 발명의 일 실시예에 따른 웹 스크래핑 탐지 방법을 설명하기 위한 흐름도이다. 도시된 흐름도에서는 상기 방법을 복수 개의 단계로 나누어 기재하였으나, 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 또는 도시되지 않은 하나 이상의 단계가 부가되어 수행될 수 있다.

[0062] 먼저, 클라이언트(102)는 웹 브라우저를 이용하여 데이터 제공자가 제공하는 웹 페이지(104)에 접속한다(S102).

[0063] 다음으로, 클라이언트(102)가 웹 브라우저를 통해 웹 페이지(104)에 접속함에 따라 스크래핑 검증 스크립트(106)가 실행된다(S104). 상술한 바와 같이, 각 웹 페이지(104)에는 공통 도메인의 스크래핑 검증 스크립트(106)가 삽입되며, 클라이언트(102)가 각 웹 페이지(104)에 접속할 때마다 상기 공통 도메인으로부터 스크래핑 검증 스크립트(106)가 로딩되어 실행될 수 있다.

[0064] 다음으로, 스크래핑 검증 스크립트(106)가 실행될 때마다 상기 스크래핑 검증 스크립트(106)와 관련된 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 저장된다(S106). 이때, 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그는 예를 들어, 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트(106)가 실행된 횟수, 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트(106)가 실행된 총 횟수, 마지막 단위 시간 값 등을 포함할 수 있다. 상기 공통 도메인의 웹 저장소(108c)는 W3C(World Wide Web Consortium) 표준 기술인 WebCrypto 를 이용하여 상기 로그를 암호화된 형태로 보관 가능하다. 또한, 상기 공통 도메인의 웹 저장소(108c)에 저장된 값들은 공격자에 의해 삭제(초기화)는 가능하나, 복제나 변조는 불가능하다.

[0065] 만약, 공통 도메인의 웹 저장소(108c)에 기 저장된 값이 존재하는 경우, 공통 도메인의 웹 저장소(108c)에 저장된 값이 업데이트된다(S108, S110).

[0066] 예를 들어, 설정된 단위 시간이 10초이고 스크래핑 검증 스크립트(106)가 최초로 실행된 시각이 오전 10시 1분 10초라 가정할 때, 오전 10시 1분 12초에 스크래핑 검증 스크립트(106)가 실행되면 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그는 아래와 같이 업데이트될 수 있다.

[0068] 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 : 오전 10시 1분 10초

[0069] 단위 시간 동안 스크래핑 검증 스크립트(106)가 실행된 횟수 : 2번

[0070] 스크래핑 검증 스크립트(106)가 실행된 총 횟수 : 2번

[0071] 마지막 단위 시간 값 : 오전 10시 1분 10초

[0073] 이후, 오전 10시 1분 13초에 스크래핑 검증 스크립트(106)가 다시 실행되면 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그는 아래와 같이 업데이트될 수 있다.

[0075] 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 : 오전 10시 1분 10초

[0076] 단위 시간 동안 스크래핑 검증 스크립트(106)가 실행된 횟수 : 3번

[0077] 스크래핑 검증 스크립트(106)가 실행된 총 횟수 : 3번

[0078] 마지막 단위 시간 값 : 오전 10시 1분 10초

[0080] 또한, 오전 10시 1분 21초에 스크래핑 검증 스크립트(106)가 다시 실행되면 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그는 아래와 같이 업데이트될 수 있다.

[0082] 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 : 오전 10시 1분 10초

[0083] 단위 시간 동안 스크래핑 검증 스크립트(106)가 실행된 횟수 : 1번

[0084] 스크래핑 검증 스크립트(106)가 실행된 총 횟수 : 4번

[0085] 마지막 단위 시간 값 : 오전 10시 1분 20초

[0087] 이와 같이, 상기 스크래핑 검증 스크립트(106)가 실행될 때마다 상기 스크래핑 검증 스크립트(106)와 관련된 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 저장되고, 업데이트될 수 있다.

[0089] 다음으로, 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 데이터 제공자의 서버로 전송된다(S112). 이때, 서버로 데이터를 전송하는 통신 구간에서는 SSL(Secure Socket Layer) 및 난독화 기술을 이용한 메시지 암호화가 가능하다.

[0090] 다음으로, 서버는 업데이트된 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그를 이용하여 클라이언트(102)가 웹 스크래퍼(web scraper)인지의 여부를 판단한다(S120).

[0091] 일 예시로서, 서버는 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트(106)가 실행된 횟수가 제1 임계치를 초과하는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다. 예를 들어, 서버는 10초 동안 상기 스크래핑 검증 스크립트(106)가 실행된 횟수가 5번을 초과하는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다.

[0092] 다른 예시로서, 서버는 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트(106)가 실행된 총 횟수가 제2 임계치를 초과하는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다. 공격자는 위 웹 스크래핑의 탐지를 피하기 위해 상기 단위 시간에 대한 정보를 해킹하여 단위 시간 동안 제1 임계치를 초과하지 않는 범위에서 웹 스크래핑을 반복 수행할 수 있다. 이에 따라, 본 발명의 실시예들에서는 설정된 단위 시간 동안 스크래핑 검증 스크립트(106)가 실행된 횟수뿐 아니라 스크래핑 검증 스크립트(106)가 실행된 총 횟수(즉, 누적 횟수)를 기 정해진 임계치와 비교하여 웹 스크래핑을 보다 높은 확률로 탐지할 수 있다. 예를 들어, 서버는 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트(106)가 실행된 총 횟수가 100번을 초과하는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다.

[0093] 또한, S108 단계에서의 판단 결과 공통 도메인의 웹 저장소(108c)에 기 저장된 값이 존재하지 않는 경우 공통 도메인의 웹 저장소(108c)에 초기 로그 값이 저장된다(S114). 즉, 스크래핑 검증 스크립트(106)가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트(106)가 실행된 횟수 및 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트(106)가 실행된 총 횟수, 마지막 단위 시간 값 등이 아래와 같이 공통 도메인의 웹 저장소(108c)에 저장될 수 있다.

[0095] 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 : 오전 10시 1분 10초

[0096] 단위 시간 동안 스크래핑 검증 스크립트(106)가 실행된 횟수 : 1번

[0097] 스크래핑 검증 스크립트(106)가 실행된 총 횟수 : 1번

[0098] 마지막 단위 시간 값 : 오전 10시 1분 10초

[0100] 도 5는 본 발명의 일 실시예에 따른 단위 시간 값을 계산하는 예시를 나타낸 도면이다.

[0101] 도 5를 참조하면, A 시점 ~ B 시점 사이는 단위 시간 값이 A로, B ~ C 시점 사이는 단위 시간 값이 B로 계산될 수 있다. 여기서, A는 예를 들어 오전 10시 1분 10초, B는 오전 10시 1분 20초일 수 있으며, 단위 시간은 10초일 수 있다.

[0102] 이 경우, 오전 10시 1분 13초의 단위 시간 값은 오전 10시 1분 10초가 되며, 오전 10시 1분 21초의 단위 시간 값은 오전 10시 1분 20초가 된다.

[0104] 다시 도 4로 돌아오면, 서버는 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 최초로 저장되는 경우 클라이언트(102)의 웹 브라우저에 대한 핑거프린트(fingerprint)를 수집한

다(S116). 여기서, 웹 브라우저에 대한 핑거프린트는 클라이언트(102)를 식별하기 위한 식별값으로서, 예를 들어 웹 브라우저의 종류, 버전(version), 언어(language), 운영체제(OS), 쿠키 사용여부(브라우저 설정값), 화면 해상도, 보유 글꼴 및 색상 등이 될 수 있다.

[0105] 상술한 바와 같이, 서버는 공통 도메인의 웹 저장소(108c)에 저장된 스크래핑 검증 스크립트(106)의 실행과 관련된 로그를 이용하여 클라이언트(102)가 웹 스크래퍼인지의 여부를 판단할 수 있다. 그러나, 공격자는 위 웹 스크래핑의 탐지를 피하기 위해 공통 도메인의 웹 저장소(108c)에 저장된 로그를 지속적으로 삭제하여 초기화할 수 있다. 이를 위해, 서버는 공통 도메인의 웹 저장소(108c)에 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 최초로 저장될 때마다(즉, 공통 도메인의 웹 저장소(108c)가 초기화된 후 새롭게 로그가 저장될 때마다) 클라이언트(102)의 웹 브라우저에 대한 핑거프린트(fingerprint)를 수집할 수 있다.

[0106] 이후, 서버는 상기 핑거프린트를 수신하고(S118), 동일한 값을 갖는 핑거프린트가 설정된 횟수 이상 수집되는 경우 클라이언트(102)를 웹 스크래퍼로 판단한다(S120). 예를 들어, 서버는 동일한 값을 갖는 핑거프린트가 5회 이상 수집되는 경우 공통 도메인의 웹 저장소(108c)가 공격자에 의해 반복적으로 초기화된 것으로 판단하고 상기 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다.

[0107] 또한, S116 단계에서 서버는 공통 도메인의 웹 저장소(108c)에 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 최초로 저장되는 경우 클라이언트(102)가 속한 로컬 네트워크 내에서 상기 클라이언트(102)를 식별하기 위한 로컬 아이피의 체인(chain)을 추가적으로 수집할 수 있다. 상술한 웹 브라우저의 핑거프린트는 서로 다른 클라이언트(102)라 하더라도 동일한 값(즉, 중복된 값)을 가질 수 있으므로, 서버는 웹 브라우저의 핑거프린트뿐 아니라 클라이언트(102)를 식별하기 위한 로컬 아이피의 체인까지 추가적으로 수집하여 클라이언트(102)가 웹 스크래퍼인지의 여부를 보다 높은 확률로 탐지할 수 있다. 공격자 단말의 외부 아이피는 프록시 서버를 통해 손쉽게 변경이 가능하지만 공격자 단말의 로컬 아이피의 체인은 변경이 어렵기 때문에, 웹 브라우저의 핑거프린트뿐 아니라 클라이언트(102)가 속한 로컬 네트워크 내에서의 로컬 아이피의 체인까지 수집할 경우 클라이언트(102)가 웹 스크래퍼인지의 여부를 보다 높은 확률로 탐지할 수 있다.

[0108] 여기서, 로컬 아이피의 체인은 클라이언트(102)의 사설 아이피 주소(예를 들어, 172.30.x.x 형태의 아이피 , 192.168.x.x 형태의 아이피 주소)뿐 아니라 클라이언트(102)가 속한 로컬 네트워크 내에서 상기 클라이언트(102)와 연결된 네트워크 장치들(예를 들어, router, hub 등)의 아이피 주소들을 포함하는 의미로 사용된다.

[0109] 이러한 로컬 아이피의 체인에 대한 예시는 아래와 같다.

[0111] * 로컬 아이피의 체인 예시

[0112] 클라이언트의 사설 아이피 주소(내부 아이피 주소) - 허브(hub)의 아이피 주소 - 라우터(router)의 아이피 주소

[0114] 이후, 서버는 상기 로컬 아이피의 체인을 수신하고(S118), 동일한 값을 갖는 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다(S120). 예를 들어, 서버는 동일한 값을 갖는 로컬 아이피의 체인이 5회 이상 수집되는 경우 공통 도메인의 웹 저장소(108c)가 공격자에 의해 반복적으로 초기화된 것으로 판단하고 상기 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다.

[0115] 즉, 서버는 S110 ~ S112 단계를 통해 공통 도메인의 웹 저장소(108c)에 업데이트된 로그를 기반으로 클라이언트(102)가 웹 스크래퍼에 해당되는지의 여부를 판단할 수 있으며, 이와 별개로 공격자에 의한 공통 도메인의 웹 저장소(108c)의 초기화에 대비하여 S114 ~ S118 단계를 통해 동일한 값의 웹 브라우저 핑거프린트/로컬 아이피의 체인의 반복 수신여부에 따라 클라이언트(102)가 웹 스크래퍼에 해당되는지의 여부를 판단할 수 있다.

[0117] 도 6은 본 발명의 일 실시예에 따른 서버의 상세 구성을 나타낸 블록도이다.

[0118] 도 6에 도시된 바와 같이, 본 발명의 일 실시예에 따른 서버(104a)는 수집부(602) 및 판단부(604)를 포함한다.

[0119] 수집부(602)는 공통 도메인의 웹 저장소(108c)에 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 업데이트될 때마다 업데이트된 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그를 수집한다. 상기 스크래핑 검증 스크립트(106)의 실행과 관련된 로그는 예를 들어, 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각, 설정된 단위 시간 동안 상기 스크래핑 검증 스크립트(106)가 실행된 횟수, 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 이후 상기 스크래핑 검증 스크립트(106)가 실행된 총 횟수, 마지막 단위 시간 값 등을 포함할 수 있다.

[0120] 또한, 수집부(602)는 공통 도메인의 웹 저장소(108c)에 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가

최초로 저장될 때마다 클라이언트(102)의 웹 브라우저에 대한 핑거프린트를 수집할 수 있다.

[0121] 또한, 수집부(602)는 공통 도메인의 웹 저장소(108c)에 스크래핑 검증 스크립트(106)의 실행과 관련된 로그가 최초로 저장될 때마다 클라이언트(102)가 속한 로컬 네트워크 내에서 클라이언트(102)를 식별하기 위한 로컬 아이피의 체인을 수집할 수 있다.

[0122] 판단부(604)는 클라이언트(102)가 웹 스크래퍼인지의 여부를 판단한다.

[0123] 구체적으로, 판단부(604)는 공통 도메인의 웹 저장소(108c)에 업데이트된 스크래핑 검증 스크립트(106)의 실행과 관련된 로그를 이용하여 클라이언트(102)가 웹 스크래퍼인지의 여부를 판단할 수 있다. 일 예시로서, 판단부(604)는 설정된 단위 시간 동안 스크래핑 검증 스크립트(106)가 실행된 횟수가 제1 임계치를 초과하는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다. 또한, 판단부(604)는 상기 스크래핑 검증 스크립트(106)가 최초로 실행된 시각 이후 스크래핑 검증 스크립트(106)가 실행된 총 횟수가 제2 임계치를 초과하는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수도 있다.

[0124] 또한, 판단부(604)는 동일한 값을 갖는 핑거프린트 또는 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 클라이언트(102)를 웹 스크래퍼로 판단할 수 있다.

[0125] 이와 같이, 본 발명의 실시예들에 따르면, 공통 도메인의 스크래핑 검증 스크립트(106)를 각 데이터 제공자의 웹 페이지(104)에 삽입한 후 상기 스크래핑 검증 스크립트(106)를 통해 클라이언트(102)의 웹 페이지(104) 접속 패턴을 모니터링하여 클라이언트(102)가 웹 스크래퍼에 해당하는지의 여부를 효율적으로 탐지할 수 있다.

[0126] 또한, 본 발명의 실시예들에 따르면, 공격자가 공통 도메인의 웹 저장소(108c)를 초기화(삭제)하는 것에 대비하여 동일한 값의 웹 브라우저 핑거프린트 또는 로컬 아이피의 체인이 설정된 횟수 이상 반복 수집되는 경우 클라이언트(102)를 웹 스크래퍼로 판단함으로써, 웹 스크래퍼의 탐지 확률을 보다 향상시킬 수 있다.

[0128] 또한, 공격자가 웹 스크래핑 관점에서 고려 가능한 공격 포인트 및 이에 대한 본 실시예들에서의 방어 로직을 정리하면 아래와 같다.

[0130] 1. 웹 저장소의 로그 매번 삭제(초기화)

[0131] 일반적인 사용자가 소지하는 사용자 단말의 웹 브라우저에 대한 웹 저장소는 매번 초기화되지 않는다. 만약, 공격자가 자신이 사용하는 웹 브라우저에 대한 웹 저장소의 로그를 매번 삭제하는 경우, 서버(104a)는 동일한 값을 갖는 핑거프린트 또는 로컬 아이피의 체인을 반복적으로 수신하게 되며 이에 따라 웹 스크래핑이 일어난 것으로 판단할 수 있다.

[0133] 2. 웹 저장소의 로그를 조작

[0134] 상술한 바와 같이, 상기 공통 도메인의 웹 저장소에 저장된 값들은 공격자에 의해 삭제(초기화)는 가능하나, 복제나 변조는 불가능하다.

[0136] 3. 서버로 전송되는 값을 드롭(drop)하거나 호스트 파일(host file)을 조작하여 스크래핑 검증 스크립트가 처음부터 실행되지 않도록 조작

[0137] 상술한 바와 같이, 본 실시예들에 있어서 서버는 은행, 카드사, 보험사, 공공기관 등과 같은 데이터 제공사의 서버이다. 따라서, 스크래핑 검증 스크립트가 실행되지 않아 서버에서 웹 저장소에 업데이트된 로그나 핑거프린트/로컬 아이피의 체인을 수집하지 않는 경우, 서버는 데이터 제공자가 제공하는 데이터 또는 서비스를 제공하지 않도록 함으로써 해당 업무(예를 들어, 은행 업무, 카드 업무, 보험 업무 등) 자체를 볼 수 없도록 할 수 있다.

[0139] 도 7은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술되지 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0140] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 서버(104a), 또는 웹 스크래핑 탐지 시스템(100)에 포함되는 하나 이상의 컴포넌트일 수 있다.

[0141] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있

다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0142] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0143] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0144] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.

[0146] 이상에서 대표적인 실시예를 통하여 본 발명에 대하여 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 전술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

[0148] 10 : 컴퓨팅 환경

12 : 컴퓨팅 장치

14 : 프로세서

16 : 컴퓨터 판독 가능 저장 매체

18 : 통신 버스

20 : 프로그램

22 : 입출력 인터페이스

24 : 입출력 장치

26 : 네트워크 통신 인터페이스

50 : 사용자 단말

60 : 데이터 제공자의 웹 페이지

70 : 공격자 단말

100 : 웹 스크래핑 탐지 시스템

102 : 클라이언트

104 : 데이터 제공자의 웹 페이지

104a : 데이터 제공자의 서버

106 : 웹 스크래핑 검증 스크립트

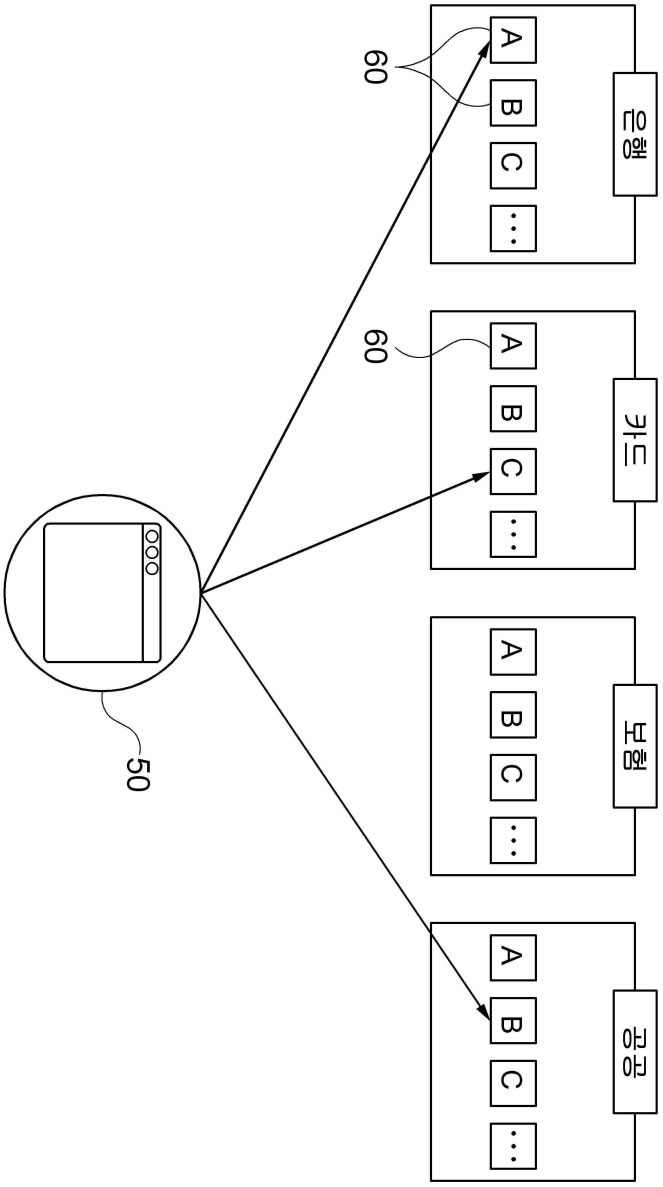
108 : 웹 저장소

602 : 수집부

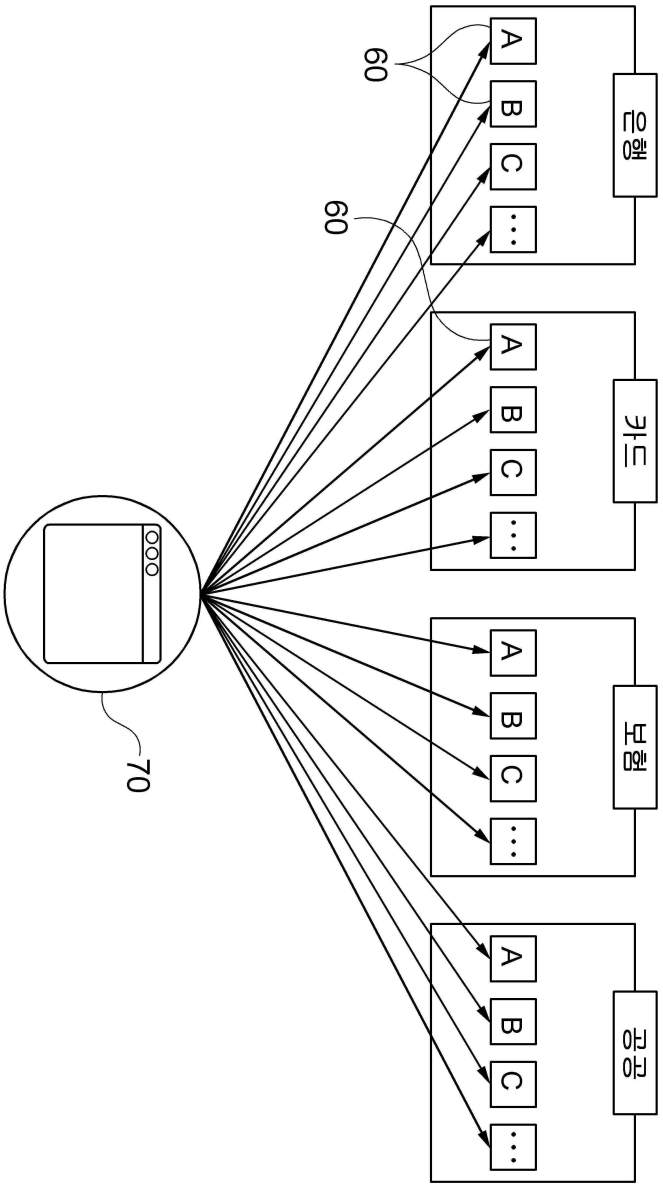
604 : 판단부

도면

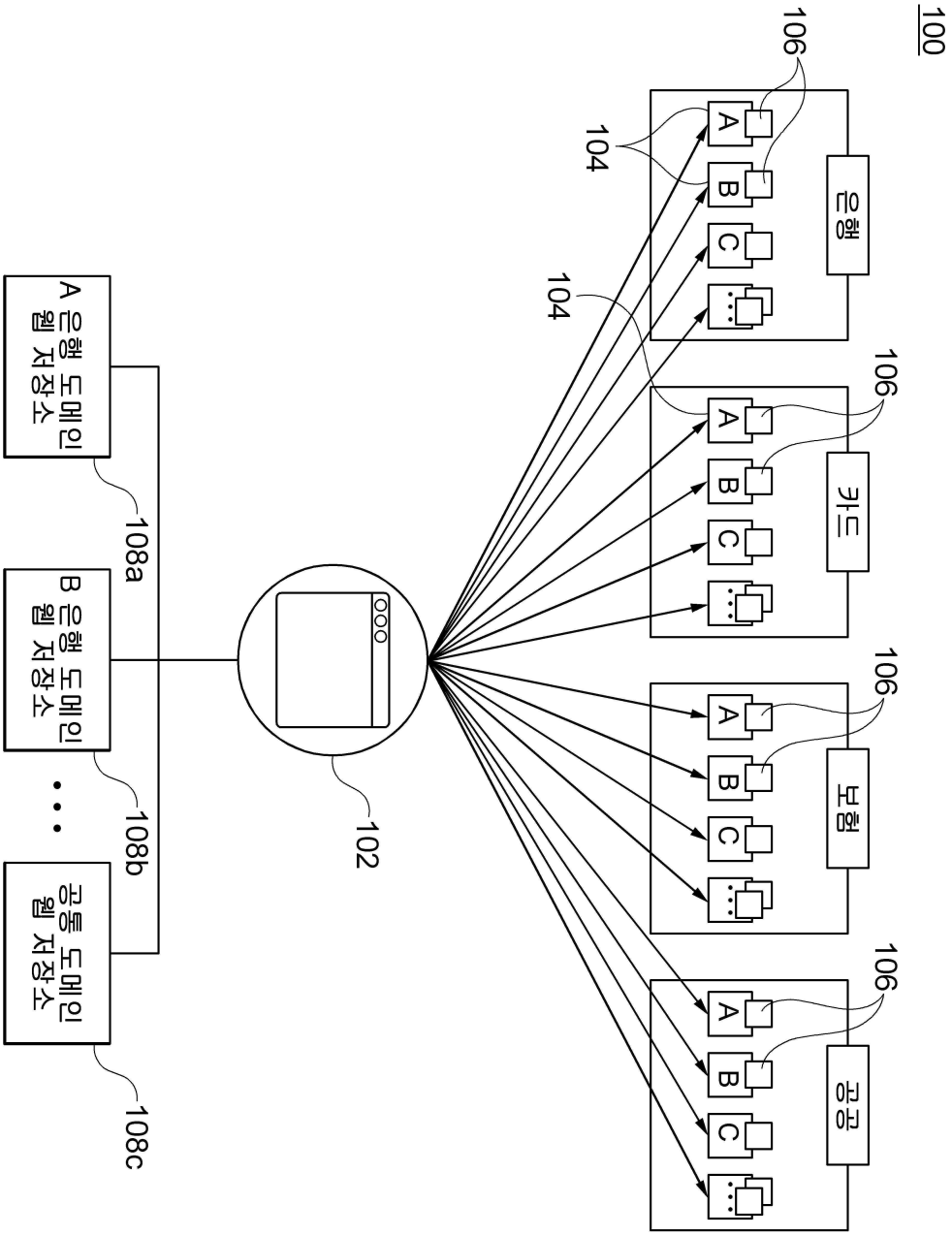
도면1



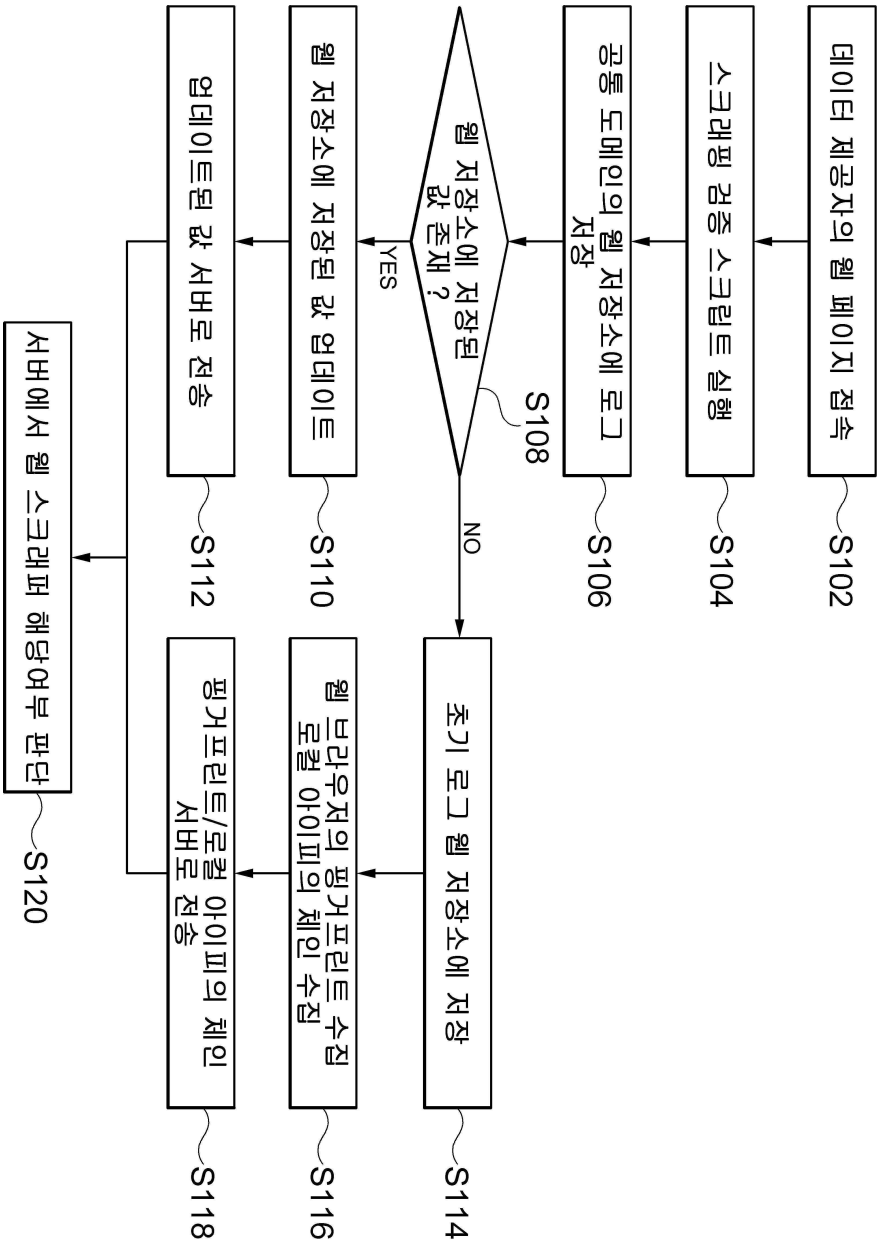
도면2



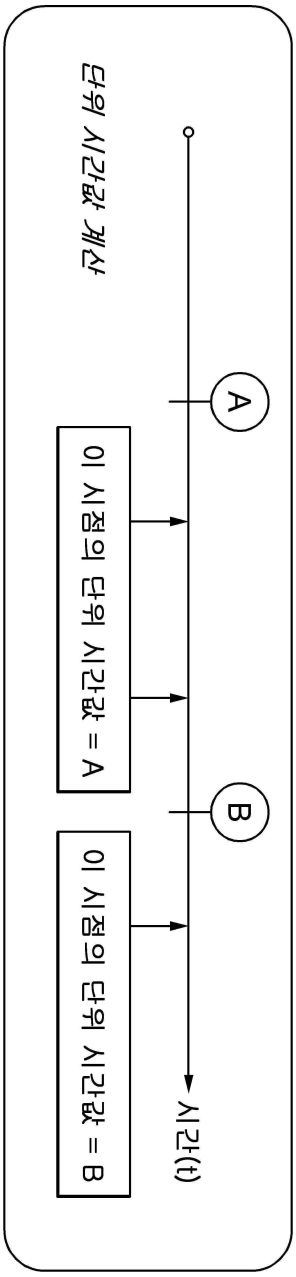
도면3


100
106
104
104
106
106
106
A
B
C
...
은행
카드
보험
공공
102
A 은행 도메인 월 저장소
B 은행 도메인 월 저장소
공통 도메인 월 저장소
108a
108b
108c

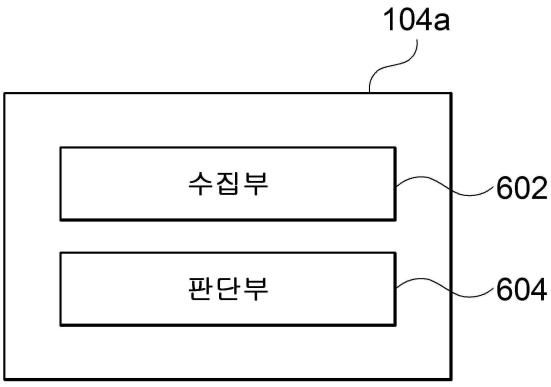
도면4


데이터 제공자의 웹 페이지 접속
S102
스크래핑 검증 스크립트 실행
S104
공통 도메인의 웹 저장소에 로그 저장
S106
웹 저장소에 저장된 값 존재 ?
S108
YES
NO
웹 저장소에 저장된 값 업데이트
S110
업데이트된 값 서버로 전송
S112
초기 로그 웹 저장소에 저장
S114
웹 브라우저의 핑거프린트 수집
로컬 아이피의 체인 수집
S116
핑거프린트/로컬 아이피의 체인 서버로 전송
S118
서버에서 웹 스크래퍼 해당여부 판단
S120

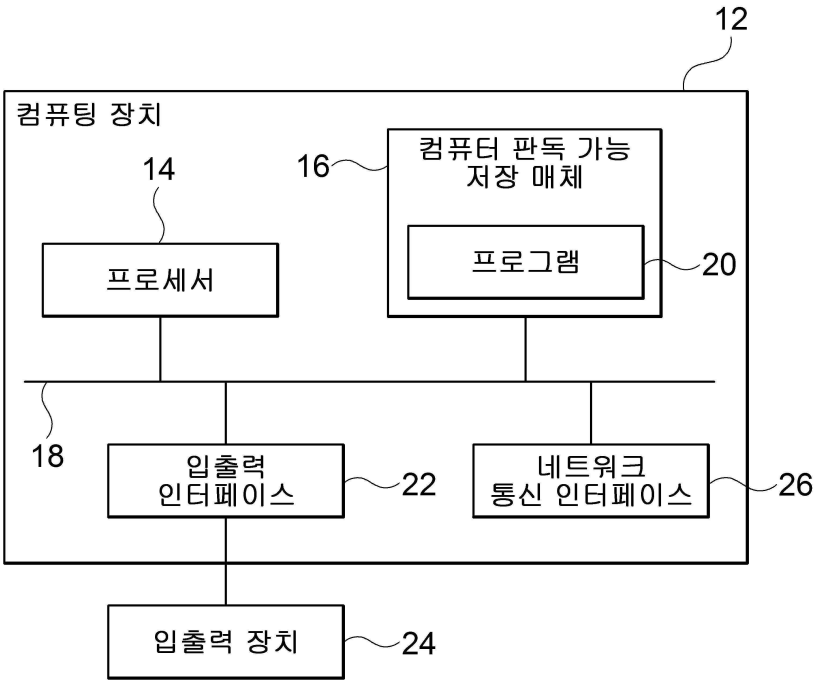
도면5



도면6



도면7



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 12

【변경전】

청구항 8에 있어서,

상기 수집부는, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트가 속한 로컬 네트워크 내에서 상기 클라이언트를 식별하기 위한 로컬 아이피의 체인(chain)을 수집하고,

상기 판단부는, 동일한 값을 갖는 상기 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 서버.

【변경후】

청구항 7에 있어서,

상기 수집부는, 상기 공통 도메인의 웹 저장소에 상기 스크래핑 검증 스크립트의 실행과 관련된 로그가 최초로 저장될 때마다 상기 클라이언트가 속한 로컬 네트워크 내에서 상기 클라이언트를 식별하기 위한 로컬 아이피의 체인(chain)을 수집하고,

상기 판단부는, 동일한 값을 갖는 상기 로컬 아이피의 체인이 설정된 횟수 이상 수집되는 경우 상기 클라이언트를 웹 스크래퍼로 판단하는, 서버.