명 세 서

청구범위

청구항 1

메인 서버에 의해 수행되는, 유해 사이트 분류 방법에 있어서,

(a) 상기 메인 서버에 미리 저장된 복수의 인터넷 웹 사이트에 대한 도메인 주소에서 숫자를 추출하고, 숫자를 추출한 도메인 주소를 통하여 해당 웹 사이트에 접속을 시도하되, 접속이 실패하는 경우, 해당 도메인 주소 내에서 추출한 숫자의 주소 내 위치를 변경하여 접속이 성공할 때까지 접속시도를 반복하며, 해당 도메인 주소의 문법 부분을 제외한 모든 문자열 사이에서 추출한 숫자의 주소 내 위치를 변경하는 모든 경우의 수에서 접속을 실패하는 경우, 상기 숫자에 기 설정된 숫자를 더하거나 뺀 뒤, 접속이 성공할 때까지 접속시도를 반복함으로써, 데이터베이스에 저장된 복수의 인터넷 웹 사이트 중 접속이 가능한 인터넷 웹 사이트를 선별하는 단계;

(b) 상기 접속이 가능한 인터넷 웹 사이트의 HTML 소스코드에서 HTML 태그, 공백 및 특수문자를 삭제하고, HTML 태그, 공백 및 특수문자를 삭제된 HTML 소스코드를 영문으로 번역한 뒤, 번역된 HTML 소스코드에서 기 설정된 문자열을 삭제하여, 전처리하는 단계;

(c) 상기 HTML 소스코드로부터, 웹사이트의 도메인 이름, 웹 사이트 내의 이미지파일의 주소, 웹 사이트 내에 개재된 링크 및 웹 사이트 내의 텍스트에 대한 HTML 소스를 포함하는 주요피처에 따라, 상기 전처리된 HTML소스코드를 분류하여 토큰화하는 단계; 및

(d) 각각의 토큰을 TF-IDF(Term Frequency-Inverse Document Frequency) 기법에 따라, 상기 토큰을 구성하는 단어 빈도를 고려하여 토큰에 포함된 모든 문서에서 기 설정된 빈도 이상 등장하는 단어에는 페널티를 주고, 해당 문서에서만 기 설정된 빈도 이상 등장하는 단어에 가중치를 주어, 각각의 토큰에 벡터 값을 부여한 후,

지도학습 방식에 따라 상기 벡터 값과 유해 사이트의HTML소스 및 정상 사이트의 HTML소스를 학습데이터로 하여 상기 (a)단계 이전에 미리 학습된 기계학습모델이, 각각의 토큰과 각각의 토큰에 부여된 벡터 값이 입력되는 경우, 입력된 토큰 및 상기 벡터 값을 분석하여 상기 웹 사이트의 유해 사이트 여부를 판단하는 단계를 포함하되,

상기 (a)단계는,

접속이 불가한 웹 사이트가 검출되는 경우, 검출된 웹사이트의 도메인이름 중 적어도 하나를 구성하는 숫자를 변경하여, 접속이 가능한 웹사이트가 발견되는지 판단하는 단계;를 포함하는 것인, 유해 사이트 분류 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 (d)단계는,

기계학습모델에 입력 데이터로 상기 벡터 값을 입력하여, 출력 데이터로 Accuracy 값 및 F1-Score를 산출하고, 산출한 Accuracy 값 및 F1-Score이 임계치 이상인 경우, 상기 벡터가 산출된 인터넷 웹 사이트를 유해 사이트로 판단하는 것인, 유해 사이트 분류 방법.

청구항 8

제7항에 있어서,

상기 기계학습모델은,

로지스틱 회귀(Logistic Regression) 모델로 구성되며, 출력데이터가 0에서 1 사이의 값으로 출력될 경우, 출력 데이터를 기초로 상기 웹 사이트가 유해 사이트에 속할 확률을 예측하는 것이고,

유해 사이트의 HTML소스와 정상사이트의 HTML 소스를 학습데이터로 하여 상기 (a)단계 전에 기 학습된 것인, 유해 사이트 분류 방법.

청구항 9

제7항에 있어서,

상기 Accuracy 값은,

입력 데이터와 출력 데이터를 비교하여 입력 데이터가 올바르게 예측된 데이터의 수를 전체 데이터의 수로 나눈 값이고,

상기 F1-Score는,

실제로 유해 사이트인 입력 데이터를 상기 기계학습모델이 유해 사이트라고 인식한 데이터의 수와, 상기 기계학습모델이 유해 사이트로 예측한 데이터 중 실제로 유해 사이트인 출력 데이터의 수를 조화평균 수식에 따라 계산하여 산출한 것인, 유해 사이트 분류 방법.

청구항 10

제1항에 있어서,

(e) 유해 사이트로 판단된 인터넷 웹 사이트의 주요 피처 및 HTML 소스코드를 상기 메인 서버의 데이터베이스에 저장하는 단계를 더 포함하는, 유해 사이트 분류 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 유해 사이트 분류 방법 및 시스템에 관한 것으로서, 보다 상세하게는, 인터넷을 통하여 접속할 수 있는 복수의 웹 사이트 중 접속이 가능한 유해 사이트의 주소를 판단하는 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 최근 인터넷 접속이 가능한 단말의 보급률이 높아짐에 따라, 청소년 및 미취학 아동 계층이 손쉽게 유해 인터넷 페이지에 접속하는 문제가 발생하고 있다.

[0003] 종래 기술의 경우, 이와 같은 문제를 방지하기 위하여, 유해사이트를 촬영한 스크린샷을 통해 분류하거나, 인력을 동원하여 수기로 유해사이트를 분류하여 왔다.

[0004] 최근에는 인력을 대신하여 인터넷 크롤링을 이용한 HTML 메타 태그 분석 방식 또한 활용되고 있다.

[0005] 그러나, 종래의 방식에 따르면, 유해 사이트는 정식 사이트와 육안으로 구분이 어렵도록 구성되며 잦은 사이트 리뉴얼을 실시하기 때문에 매번 다시 대응해야 하는 문제 발생한다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 유해 사이트 분류 방법 및 시스템을 제공하는 것을 목적으로 한다.

[0007] 이를 통해, 지속적인 단속에도 불구하고 도메인 변경 후 다시 활동하는 유해 사이트의 접속 가능 주소를 빠르게 파악하고, 해당 활동 방식에 대해 적시에 대응 가능하도록 유해 사이트를 분류하는 것을 목적으로 한다.

[0008] 본 발명이 해결하려는 과제들은 이상에서 언급한 과제들로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0009] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 일 실시 예에 따르는, 메인 서버에 의해 수행되는, 유해 사이트 분류 방법은, (a) 데이터베이스에 저장된 복수의 인터넷 웹 사이트 중 접속이 가능한 인터넷 웹 사이트를 선별하는 단계; (b) 접속이 가능한 인터넷 웹 사이트의 HTML 소스코드를 추출하고, 전처리하는 단계; (c) 상기 HTML 소스코드로부터, 웹사이트의 도메인 이름, 웹 사이트 내의 이미지파일의 주소, 웹 사이트 내에 개재된 링크 및 웹 사이트 내의 텍스트에 대한 HTML 소스 중 적어도 하나를 분류하여 토큰화하는 단계; 및 (d) 각각의 토큰을 분석하여 상기 웹 사이트의 유해 사이트 여부를 판단하는 단계를 포함하되, 상기 (a)단계는, 접속이 불가한 웹 사이트가 검출되는 경우, 검출된 웹사이트의 도메인이름 중 적어도 하나를 구성하는 숫자를 변경하여, 접속이 가능한 웹사이트가 발견되는지 판단하는 단계;를 포함할 수 있다.

[0010] 또한, 상기 (a)단계는, 상기 메인 서버의 데이터 베이스에 미리 저장된 복수의 인터넷 웹 사이트에 접속한 후, 상기 인터넷 웹 사이트에 대한 도메인 주소에서 숫자를 추출하고, 숫자를 추출한 도메인 주소를 통하여 해당 웹 사이트에 접속을 시도하여 접속 가능 여부를 판단하되, 접속이 실패하는 경우, 해당 도메인 주소 내에서 추출한 숫자의 주소 내 위치를 변경하여 접속을 재시도하여 접속 가능 여부를 판단하는 것일 수 있다.

[0011] 또한, 상기 (a)단계는, 해당 도메인 주소 내에서 추출한 숫자의 주소 내 위치를 변경하는 모든 경우의 수에서 접속을 실패하는 경우, 상기 숫자에 기 설정된 숫자를 더하거나 뺀 뒤, 접속이 성공할 때까지 상기 (a)단계를 반복하는 것일 수 있다.

[0012] 또한, 상기 (b)단계는, (b-1) 상기 접속이 가능한 인터넷 웹 사이트의 HTML 소스코드에서 HTML 태그, 공백 및 특수문자를 삭제하는 단계; (b-2) HTML 태그, 공백 및 특수문자를 삭제된 HTML 소스코드를 영문으로 번역하고, 번역된 HTML 소스코드에서 기 설정된 문자열을 삭제하는 단계; 및 (b-3) 문자열이 삭제된 HTML소스코드를 주요 피처에 따라 분류하여 각각 토큰화하는 단계를 포함할 수 있다.

[0013] 또한, 상기 주요 피처는, 웹사이트의 도메인 이름, 웹 사이트 내의 이미지파일의 주소, 웹 사이트 내에 개재된 링크 및 웹 사이트 내의 텍스트에 대한 HTML 소스 중 적어도 하나 이상을 포함할 수 있다.

[0014] 또한, 상기 (d)단계는, TF-IDF(Term Frequency-Inverse Document Frequency) 기법에 따라, 상기 토큰을 구성하는 단어 빈도를 고려하여 해당 단어의 중요도를 수치화하여 벡터로 나타내는 것일 수 있다.

[0015] 또한, 상기 (d)단계는, 기계학습모델에 입력 데이터로 상기 벡터 값을 입력하여, 출력 데이터로 Accuracy 값 및 F1-Score를 산출하고, 산출한 Accuracy 값 및 F1-Score이 임계치 이상인 경우, 상기 벡터가 산출된 인터넷 웹 사이트를 유해 사이트로 판단하는 것일 수 있다.

[0016] 또한, 상기 기계학습모델은, 로지스틱 회귀(Logistic Regression) 모델로 구성되며, 출력데이터가 0에서 1 사이의 값으로 출력될 경우, 출력데이터를 기초로 상기 웹 사이트가 유해 사이트에 속할 확률을 예측하는 것이고, 유해 사이트의 HTML소스와 정상사이트의 HTML 소스를 학습데이터로 하여 상기 (a)단계 전에 기 학습된 것일 수 있다.

[0017] 또한, 상기 Accuracy 값은, 입력 데이터와 출력 데이터를 비교하여 입력 데이터가 올바르게 예측된 데이터의 수를 전체 데이터의 수로 나눈 값이고, 상기 F1-Score는, 실제로 유해 사이트인 입력 데이터를 상기 기계학습모델이 유해 사이트라고 인식한 데이터의 수와, 상기 기계학습모델이 유해 사이트로 예측한 데이터 중 실제로 유해 사이트인 출력 데이터의 수를 조화평균 수식에 따라 계산하여 산출한 것일 수 있다.

[0018] 또한, (e) 유해 사이트로 판단된 인터넷 웹 사이트의 주요 피처 및 HTML 소스코드를 상기 메인 서버의 데이터베이스에 저장하는 단계를 더 포함할 수 있다.

발명의 효과

[0019] 본 발명은 유해 사이트 분류 방법과 시스템을 제공함으로써, 잦은 주기로 도메인 변경 또는 사이트를 리뉴얼한 이후 반복적으로 활동하는 유해 사이트를 분류하여 최신 접속 주소를 파악할 수 있다.

[0020] 또한, 파악한 주소에 대하여 차단, 제제 및 징계 등의 다양한 조치를 신속하게 제공하여, 웹 사이트에 방문하는 사용자가 유해 사이트에 노출되어 겪을 수 있는 피해를 최소화할 수 있다.

[0021] 또한, 기존에 인력으로 수행되어 왔던 유해 사이트 분류 작업을 대체하여, 작업자가 분류 작업 중 지속적으로 유해 콘텐츠에 노출되어 정신적인 피로감을 겪는 문제를 해결할 수 있다.

도면의 간단한 설명

[0022] 도1은 본 발명의 일 실시 예에 따르는, 유해 사이트 분류 시스템에 대한 구조도 이다.

도2는 본 발명의 일 실시 예에 따르는, 메인 서버의 내부구성을 나타내는 블록도 이다.

도3은 본 발명의 일 실시 예에 따르는, 접속부의 내부구성을 나타내는 블록도 이다.

도4는 본 발명의 일 실시 예에 따르는, 분석부의 내부구성을 나타내는 블록도 이다.

도5는 본 발명의 일 실시 예에 따르는, 주요피처를 나타내는 도표이다.

도6은 본 발명의 일 실시 예에 따르는, TF-IDF 수식 이다.

도7은 본 발명의 일 실시 예에 따르는, 유해 사이트의 접속 가능한 주소를 파악하는 동작에 관한 순서도 이다.

도8은 본 발명의 일 실시 예에 따르는, 전처리 동작에 관한 순서도 이다.

발명을 실시하기 위한 구체적인 내용

[0023] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0024] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0025] 본 명세서에 있어서 '부(部)'란, 하드웨어에 의해 실현되는 유닛(unit), 소프트웨어에 의해 실현되는 유닛, 양방을 이용하여 실현되는 유닛을 포함한다. 또한, 1 개의 유닛이 2 개 이상의 하드웨어를 이용하여 실현되어도 되고, 2 개 이상의 유닛이 1 개의 하드웨어에 의해 실현되어도 된다. 한편, '~부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니며, '~부'는 어드레싱 할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 '~부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로코드, 회로, 데이터, 데이터 베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 '~부'들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '~부'들로 결합되거나 추가적인 구성요소들과 '~부'들로 더 분리될 수 있다. 뿐만 아니라, 구성요소들 및 '~부'들은 디바이스 또는 보안 멀티미디어카드 내의 하나 또는 그 이상의 CPU들을 재생시키도록 구현될 수도 있다.

[0026] 이하에서 언급되는 "단말"은 네트워크를 통해 서버나 타 단말에 접속할 수 있는 컴퓨터나 휴대용 단말기로 구현될 수 있다. 여기서, 컴퓨터는 예를 들어, 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크톱(desktop), 랩톱(laptop), VR HMD(예를 들어, HTC VIVE, Oculus Rift, GearVR, DayDream, PSVR 등)등을 포함할 수 있다. 여기서, VR HMD 는 PC용 (예를 들어, HTC VIVE, Oculus Rift, FOVE, Deepon 등)과 모바일용(예를 들어, GearVR,

DayDream, 폭풍마경, 구글 카드보드 등) 그리고 콘솔용(PSVR)과 독립적으로 구현되는 Stand Alone 모델(예를 들어, Deepon, PICO 등) 등을 모두 포함한다. 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, 스마트폰(smart phone), 태블릿 PC, 웨어러블 디바이스뿐만 아니라, 블루투스(BLE, Bluetooth Low Energy), NFC, RFID, 초음파(Ultrasonic), 적외선, 와이파이(WiFi), 라이파이(LiFi) 등의 통신 모듈을 탑재한 각종 디바이스를 포함할 수 있다. 또한, "네트워크"는 단말들 및 서버들과 같은 각각의 노드 상호 간에 정보 교환이 가능한 연결 구조를 의미하는 것으로, 근거리 통신망(LAN: Local Area Network), 광역 통신망(WAN: Wide Area Network), 인터넷 (WWW: World Wide Web), 유무선 데이터 통신망, 전화망, 유무선 텔레비전 통신망 등을 포함한다. 무선 데이터 통신망의 일례에는 3G, 4G, 5G, 3GPP(3rd Generation Partnership Project), LTE(Long Term Evolution), WIMAX(World Interoperability for Microwave Access), 와이파이(Wi-Fi), 블루투스 통신, 적외선 통신, 초음파 통신, 가시광 통신(VLC: Visible Light Communication), 라이파이(LiFi) 등이 포함되나 이에 한정되지는 않는다.

[0027] 본 발명은 유해 사이트 분류 방법 및 시스템을 제공함으로써, 도메인 변경 후 다시 활동하는 유해 사이트의 접속 가능 주소를 빠르게 파악하고, 해당 활동 방식에 대해 적시에 대응 가능하도록 유해 사이트를 분류하는 기술이다.

[0028] 도1을 참조하면, 본 발명의 일 실시 예에 따르는 유해 사이트 분류 시스템은, 메인 서버(100) 및 유해 사이트 서버(200)로 구성될 수 있다.

[0029] 메인 서버(100)는 통신망을 통해 인터넷에 접속하여 유해 사이트 서버(200)로부터 제공되는 각 종 유해 사이트에 접속할 수 있으며, 이를 위해, 통신망과 유무선으로 연결되어 통신을 수행하는 것일 수 있다.

[0030] 또한, 도2를 참조하면, 본 발명의 일 실시 예에 따르는 메인 서버(100)는 유해 사이트 분류 방법을 수행하는 프로그램(또는 애플리케이션)이 저장된 메모리와 위 프로그램을 실행하는 프로세서 및 복수의 유해 사이트 접속에 필요한 데이터를 저장하는 데이터 베이스(110)를 포함하여 구성될 수 있다.

[0031] 여기서 프로세서는 메모리에 저장된 프로그램의 실행에 따라 다양한 기능을 수행할 수 있는데, 각 기능에 따라 프로세서에 포함되는 세부 구성요소들을 접속부(120), 분석부(130), 분류부(140) 및 저장부(150)로 나타낼 수 있다.

[0032] 상술한 각각의 세부 구성요소에 관한 설명은 추후, 본 발명의 일 실시 예에 따르는 유해 사이트 분류 방법에 대한 설명과 함께 설명하도록 한다.

[0033] 한편, 유해 사이트 서버(200)는 통신망을 통해 인터넷에 접속하여 유해 사이트 각 종 유해 사이트를 제공하는 서버로서, 유해 사이트의 접속을 지원하기 위해, DNS(Domain Name System)서버를 경유하는 것일 수 있다.

[0034] 여기서, DNS서버는 스마트폰이나 노트북 등의 사용자 단말로부터 웹 사이트 또는 웹 컨텐츠를 서비스하는 서버에 이르기까지 인터넷상의 모든 컴퓨터가 IP 주소를 통하여 통신하는 대신, 웹 브라우저를 열고 웹 사이트로 이동할 때, example.com과 같은 도메인 이름을 입력해도 원하는 웹 사이트로 갈 수 있도록 배포된 서비스로서, www.example.com과 같이 사람이 읽을 수 있는 이름을 192.0.2.1과 같은 숫자 IP 주소로 변환하여 사용자 단말을 어떤 서버에 연결할 것인지를 제어하는 것을 의미한다.

[0035] 이러한 DNS서버의 경우, 종래기술에 해당하므로, 본 명세서에서는 자세한 설명은 생략하도록 한다.

[0036] 이하에서, 본 발명의 일 실시 예에 따르는, 메인 서버(100)에 의해 수행되는, 유해 사이트 분류 방법에 대하여 설명하도록 한다.

[0037] 먼저, 메인 서버(100)는, 유해 사이트 서버(200)에 의해 제공되는 유해 사이트에 접속한다.

[0038] 이때, 유해 사이트에 대한 최초 접속에 필요한 IP주소 및 도메인 주소 등의 데이터는 메인 서버(100)의 데이터 베이스(110)에 미리 저장된 것일 수 있다.

[0039] 이를 위해, 메인 서버(100)는 데이터 베이스(110)에 미리 저장된 복수의 인터넷 웹 사이트, 즉, 유해 사이트에 접속을 시도하며, 해당 인터넷 웹 사이트에 대한 도메인 주소에서 숫자를 추출한다.

[0040] 이는, 최근 운영되고 있는 대다수의 유해 사이트가, 차단 또는 징계 등의 처벌을 회피하기 위한 수단으로서, 기설정된 주기마다 해당 사이트를 리뉴얼하며, 사이트의 접속을 위한 도메인 주소 내에 특정 숫자를 포함시켜, 매 리뉴얼 마다 해당 숫자를 변경하는 방식을 활용하는 것에 대처하기 위함이다.

[0041] 이를 위해, 본 발명의 일 실시 예에 따르는 메인 서버(100)는, 숫자를 추출한 도메인 주소를 통하여 해당 웹 사이트에 접속을 시도하여 접속 가능 여부를 판단한다.

[0042] 해당 주소를 통한 접속이 실패하는 경우, 메인 서버(100)는 해당 도메인 주소 내에서 추출한 숫자의 주소 내 위치를 변경하여 접속을 재시도하여 접속 가능 여부를 판단한다.

[0043] 예를 들어, 메인 서버(100)가 www.example1.com과 같은 도메인 주소로 접속을 시도하였고, 해당 주소로 접속이 실패하였다고 가정하도록 한다.

[0044] 이 경우, 메인 서버(100)는 www.example1.com의 주소에서 숫자 1을 추출하고, 추출한 숫자의 위치를 변경하여 다시 접속을 시도한다.

[0045] 이러한 숫자의 변경은 www.e1xample.com 또는 www.exam1ple.com과 같은 실시 예로 구성될 수 있으며, 본 발명의 추가 실시 예에 따르면, 도메인 주소 내에서 숫자가 변경될 수 있는 위치는, 도메인 주소의 문법 부분, 예를 들면, www(World Wide Web)이나, com(Company), net(Network) 및 마침표(.)는 제외될 수 있으며, 도메인 주소의 특징적 부분에 해당하는 example과 같은 문자열 사이만이 그 대상이 되는 것일 수도 있다.

[0046] 또한, 메인 서버(100)는, 해당 도메인 주소 내에서 추출한 숫자의 주소 내 위치를 변경하는 모든 경우의 수에서 접속을 실패하는 경우, 숫자에 기 설정된 숫자를 더하거나 뺀 뒤, 접속이 성공할 때까지 반복하는 것일 수 있다.

[0047] 예를 들어, 앞서 설명한 예와 같이, 메인 서버(100)가 www.example1.com과 같은 도메인 주소로 접속을 시도하였고, 숫자 1을 도메인 주소의 변경하는 모든 경우의 수에서 접속이 실패하였다고 가정하도록 한다.

[0048] 이러한 경우, 메인 서버(100)는 기존 www.example1.com의 주소에서, 숫자 1에 기 설정된 숫자(예를 들어, 1로 가정)를 더 하거나 뺀 주소인, www.example2.com와 www.example0.com 로 접속을 재시도 할 수 있다.

[0049] 또한, 변경된 숫자의 주소에서도 숫자 부분인 0과 2를 추출하여 도메인 주소 내의 위치를 변경한 접속 시도 또한 이어서 수행될 수 있다.

[0050] 이러한 접속 시도 과정은, 본 발명의 일 실시 예에 따르는 메인 서버(100)의 접속부(120)에 의해 수행될 수 있다.

[0051] 도3을 참조하면, 본 발명의 일 실시 예에 따르는 접속부(120)는, 상술한 바와 같이 데이터 베이스(110)에서 접속을 시도할 복수의 웹 사이트에 대한 도메인 주소를 조회하고, 접속 요청을 생성하여 접속을 시도할 수 있다.

[0052] 이때, 접속이 성공하는 경우, 해당 웹 사이트에 대한 도메인 주소를 분석부(130)로 전달하며, 접속이 실패하는 경우에는, 접속을 실패한 도메인 주소를 접속 주소 예측부에 송신할 수 있다.

[0053] 본 발명의 일 실시 예에 따르면, 접속 주소 예측부는 상술한 바와 같은 방식으로 접속이 실패한 도메인 주소 내의 숫자의 위치와 숫자의 크기를 변경하는 역할을 수행한다.

[0054] 다음으로, 메인 서버(100)는 웹 사이트의 HTML 소스코드를 추출하고, 전처리하여 토큰화를 수행한다.

[0055] 본 발명의 일 실시 예에 따르는, HTML 소스코드는 인터넷 웹 페이지를 제공하기 위해 웹 브라우저에서 동작하는 언어를 텍스트로 표현한 것으로서, 인터넷 웹 페이지의 제공 서버로부터 제공되거나 크롤링 등의 수단을 통하여 추출되는 것일 수 있다.

[0056] 이러한 HTML 소스코드에는, HTML 태그, 코딩에 필요한 문법 상 사용되는 공백 및 특수문자가 포함된다.

[0057] 메인 서버(100)는 접속이 가능한 인터넷 웹 사이트의 HTML 소스코드에서 HTML 태그, 공백 및 특수문자를 삭제하고, 태그, 공백 및 특수문자를 삭제된 HTML 소스코드를 영문으로 번역하고, 번역된 HTML 소스코드에서 기 설정된 문자열을 삭제하여 전처리 과정을 수행한다.

[0058] 메인 서버(100)의 분류부(140)는 전처리가 수행된 HTML 소스코드를, 주요 피처에 따라 분류하여 각각 토큰화(Tokenization)한다.

[0059] 본 발명에서 토큰화는, 해당 데이터를 사용하고자 하는 용도에 맞게 분할하여 각각의 토큰(Token)으로 생성하는 것이며, 이는 종래 기술에 해당하므로, 본 명세서에서는 자세히 설명하지 않는다.

[0060] 도5를 참조하면, 본 발명의 일 실시 예에 따르는 주요 피처는, 웹사이트의 도메인 이름, 웹 사이트 내의 이미지 파일의 주소, 웹 사이트 내에 개재된 링크 및 웹 사이트 내의 텍스트에 대한 HTML 소스 중 적어도 하나 이상을

포함하는 것으로서, 메인 서버(100)는 상술한 주요 피처 별로, HTML 소스코드를 토큰화한다.

[0061] 다음으로, 메인 서버(100)는 TF-IDF(Term Frequency - Inverse Document Frequency)기법에 따라, 각 토큰을 구성하는 단어 빈도를 고려하여 해당 단어의 중요도를 수치화하여 벡터로 나타냄으로써, 각각의 토큰을 벡터화한다.

[0062] 본 발명의 일 실시 예에 따르면, 메인 서버(100)는 TF-IDF 기법에 따라, 각각의 토큰에서 자주 등장하는 단어에 높은 가중치를 주되, 해당 토큰에 포함된 문서에 전반적으로 자주 등장하는 단어에 대하여 패널티와 가중치를 주는 방식으로 벡터값을 부여한다.

[0063] 이때, 토큰에 포함된 모든 문서에서 자주 등장하는 단어에는 페널티를 주고, 해당 문서에서만, 자주 등장하는 단어에 높은 가중치를 주는 방식을 활용함으로써, 패널티 혹은 가중치를 받은 단어가 실질적으로 중요한 단어인지 검사할 수 있다.

[0064] 이를 위해, 도6에 도시된 바와 같은 TF-IDF 수식이 활용될 수 있다.

[0065] 도6을 참조하면, 도시된 수식과 같이 문서, 단어, 문서의 총 개수를 변수로 하여, 메인 서버(100)는 특정 토큰에 포함된 문서에서 특정 단어가 몇 번 나타났는지 count할 수 있다.

[0066] 이때, 수식에서 n은 고정된 값이기 때문에, df(t)가 증가할수록 log(n/df(t))는 감소한다. 여기서 df(t)는 특정 단어 t를 포함하는 문서의 개수를 의미하므로, 특정 단어 t를 포함하는 문서가 많다는 것은 t가 보편적으로 사용되는 단어라는 뜻이고, 이는 t가 실질적으로 중요한 단어가 아니라는 뜻일 수 있다. 따라서 log(n/df(t)) 값이 작아지며 페널티가 적용될 수 있다.

[0067] 다음으로, 본 발명의 일 실시 예에 따르는 메인 서버(100)는 다양한 수단을 통하여 추출한 벡터값을 이용하여 웹 사이트의 유해 사이트 여부를 판단할 수 있으며, 이 중 바람직한 실시 예로서, 기계학습모델이 활용될 수 있다.

[0068] 해당 실시 예에서, 기계학습모델은, 로지스틱 회귀(Logistic Regression) 모델로 구성되며, 출력데이터가 0에서 1 사이의 값으로 출력될 경우, 출력데이터를 기초로 웹 사이트가 유해 사이트에 속할 확률을 예측하는 것일 수 있다.

[0069] 이때, 기계학습모델은, 유해 사이트의 HTML소스와 정상사이트의 HTML 소스를 학습데이터로 하여, 지도학습 방식으로, 메인 서버(100)가 특정 웹 사이트에 접속하는 동작 이전에 미리 학습된 것일 수 있다.

[0070] 본 발명의 일 실시 예에 따르는 기계학습모델은 기계학습모델에 입력 데이터로 벡터가 입력되는 경우, 출력 데이터로 Accuracy 값 및 F1-Score를 산출하고, 산출한 Accuracy 값 및 F1-Score이 임계치 이상인 경우, 상기 벡터가 산출된 인터넷 웹 사이트를 유해 사이트로 판단하는 것일 수 있다.

[0071] 이때, Accuracy 값은, 입력 데이터와 출력 데이터를 비교하여 입력 데이터가 올바르게 예측된 데이터의 수를 전체 데이터의 수로 나눈 값이며, F1-Score는, 실제로 유해 사이트인 입력 데이터를 기계학습모델이 유해 사이트라고 인식한 데이터의 수와, 기계학습모델이 유해 사이트로 예측한 데이터 중 실제로 유해 사이트인 출력 데이터의 수를 조화평균 수식에 따라 계산하여 산출한 것일 수 있다.

[0072] 따라서, 본 발명의 일 실시 예에 따르는 메인 서버(100)는 Accuracy 값과 F1-Score를 기반으로, 기계학습모델이 얼마나 정확하게 유해 서버를 판단하는지 파악할 수 있다.

[0073] 상술한 바와 같은 과정을 통하여 유해 사이트로 판단된 웹 사이트에 대하여, 메인 서버(100)의 저장부(150)는 유해 사이트로 판단된 인터넷 웹 사이트의 주요 피처 및 HTML 소스코드를 상기 메인 서버(100)의 데이터 베이스(110)에 저장한다.

[0074] 본 발명의 추가 실시 예에 따르면, 데이터 베이스(110)에 저장된 유해 사이트의 주요 피처 및 HTML 소스코드는 저장부(150)에 의해, 메인 서버(100)가 유해 사이트 분류를 위하여 데이터 베이스(110) 내의 특정 웹 사이트에 접속할 때, 유해 사이트로 분류되지 않았거나 분류되기 이전인 웹 사이트보다 더 낮은 우선순위로 접속될 수 있다.

[0075] 따라서 메인 서버(100)는 이미 유해 사이트로 분류된 웹 사이트를 더 나중에 접속 시도함으로써, 더 많은 수의 유해 사이트로 분류되지 않았거나 분류되기 이전인 웹 사이트에 대한 검증을 수행할 수 있다.

[0076] 이하에서, 도7 내지 도8을 참조하여 본 발명의 일 실시 예에 따르는 유해 사이트의 접속 가능한 주소를 파악하

는 과정 및 전처리 과정에 관하여 다시 한번 설명하도록 한다.

[0077] 도7을 참조하면, 본 발명의 일 실시 예에 따르는 유해 사이트의 접속 가능한 주소를 파악하는 과정은 먼저, 메인 서버(100)의 데이터 베이스(110)에 미리 저장된 복수의 인터넷 웹 사이트에 접속한 후, 접속이 실패하는 경우, 인터넷 웹 사이트에 대한 도메인 주소에서 숫자를 추출(S101)하며 시작된다.

[0078] 다음으로, 메인 서버(100)는 숫자를 추출한 도메인 주소를 통하여 해당 웹 사이트에 접속을 시도하여 접속 가능 여부를 판단(S102)한다.

[0079] 이후, 메인 서버(100)는 접속이 실패하는 경우, 해당 도메인 주소 내에서 추출한 숫자의 주소 내 위치를 변경하여 접속을 재시도하여 접속 가능 여부를 판단(S103)한다.

[0080] 다음으로, 도8을 참조하면, 본 발명의 일 실시 예에 따르는 전처리 과정은 먼저, 접속이 가능한 인터넷 웹 사이트의 HTML 소스코드에서 HTML 태그, 공백 및 특수문자를 삭제(S201)하고, HTML 태그, 공백 및 특수문자를 삭제된 HTML 소스코드를 영문으로 번역하고, 번역된 HTML 소스코드에서 기 설정된 문자열을 삭제(S202)하여, 문자열이 삭제된 HTML소스코드를 주요 피처에 따라 분류하여 각각 토큰화(S203)함으로써 수행될 수 있다.

[0081] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독 가능 매체는 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0082] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0083] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0084] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0085] 100: 메인 서버 110: 데이터 베이스

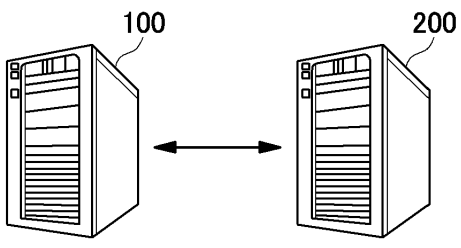
120: 접속부 130: 분석부

140: 분류부 150: 저장부

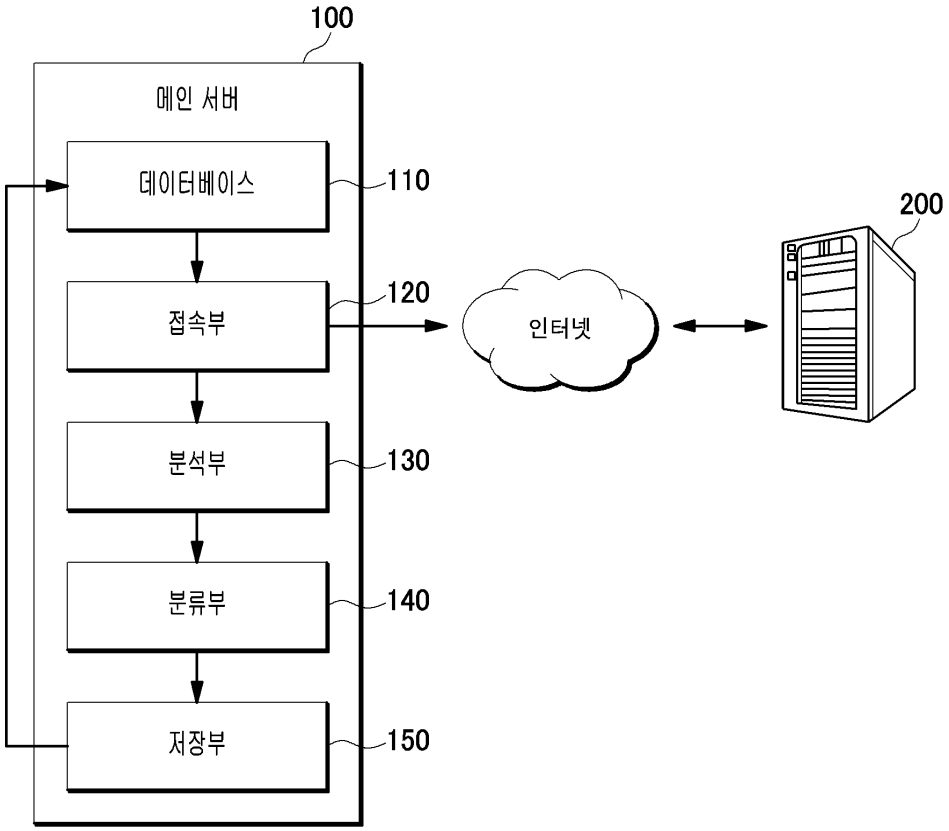
200: 유해 사이트 서버

도면

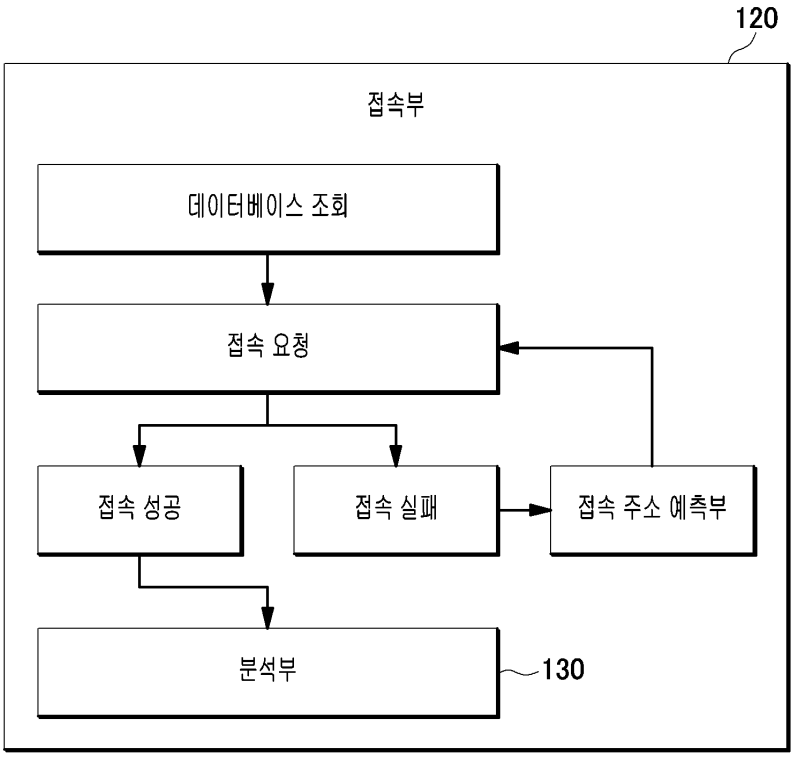
도면1



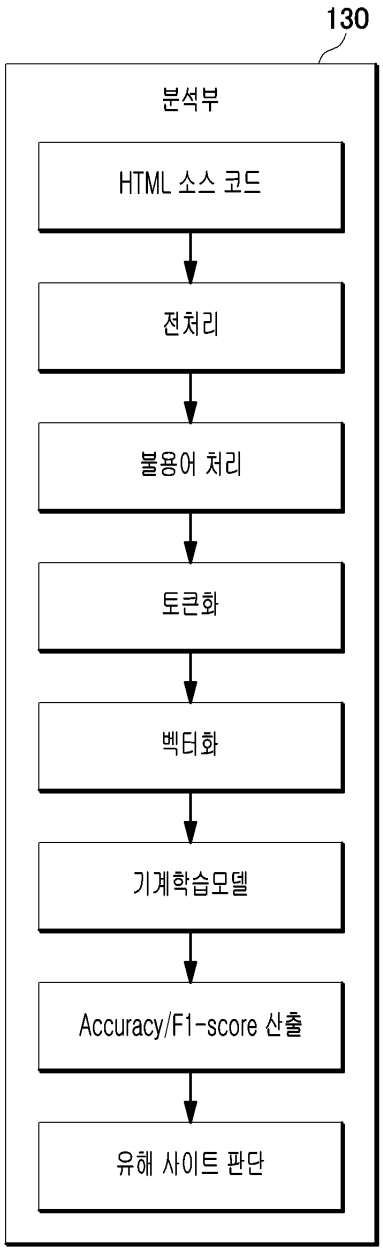
도면2


100
메인 서버
데이터베이스
110
접속부
120
분석부
130
분류부
140
저장부
150
인터넷
200

도면3


120
접속부
데이터베이스 조회
접속 요청
접속 성공
접속 실패
접속 주소 예측부
분석부
130

도면4


130
분석부
HTML 소스 코드
전처리
불용어 처리
토큰화
벡터화
기계학습모델
Accuracy/F1-score 산출
유해 사이트 판단

도면5

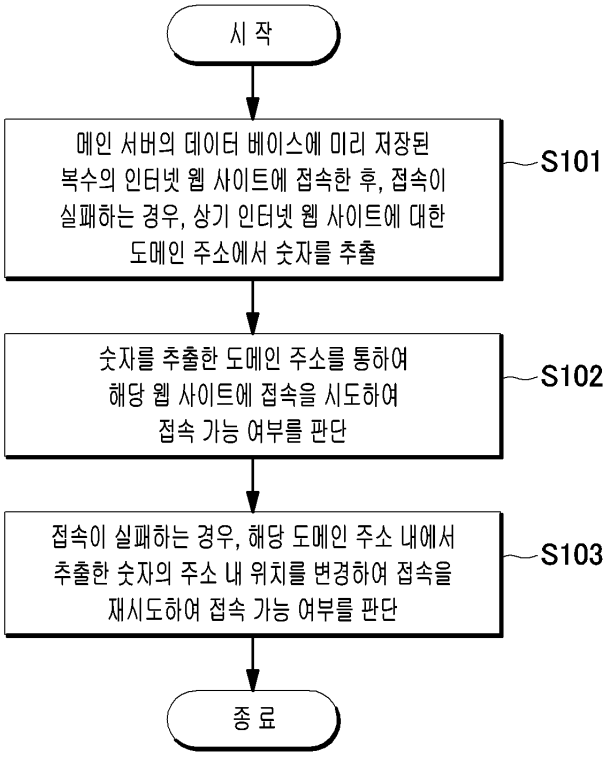
주요 피처
도메인 이름
웹 사이트 내의 이미지파일의 주소
웹 사이트 내에 개재된 링크
웹 사이트 내의 텍스트에 대한 HTML 소스

도면6

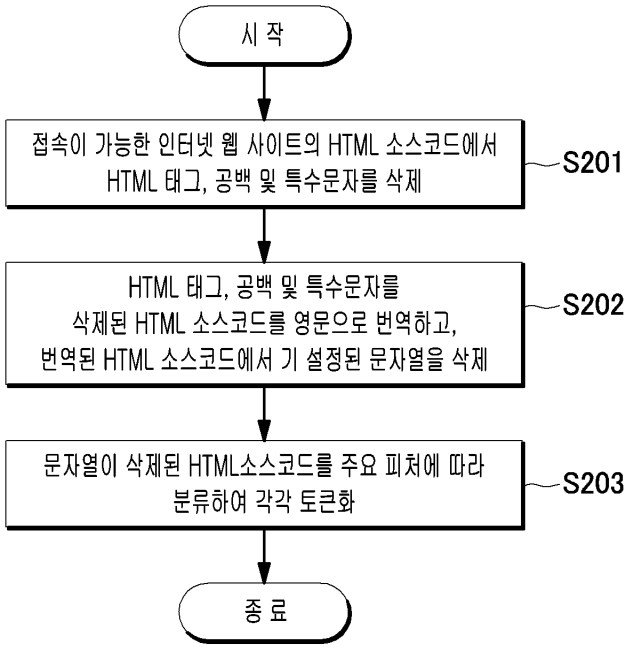
$$idf(d,\ t) = \log\left(\frac{n}{1 + df(t)}\right)$$

<TF-IDF 수식>
d:문서, t:단어, n:문서의 총 개수

도면7



도면8



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 7

【변경전】

제6항에 있어서,

상기 (d)단계는,

기계학습모델에 입력 데이터로 상기 벡터 값을 입력하여, 출력 데이터로 Accuracy 값 및 F1-Score를 산출하고, 산출한 Accuracy 값 및 F1-Score이 임계치 이상인 경우, 상기 벡터가 산출된 인터넷 웹 사이트를 유해 사이트로 판단하는 것인, 유해 사이트 분류 방법.

【변경후】

제1항에 있어서,

상기 (d)단계는,

기계학습모델에 입력 데이터로 상기 벡터 값을 입력하여, 출력 데이터로 Accuracy 값 및 F1-Score를 산출하고, 산출한 Accuracy 값 및 F1-Score이 임계치 이상인 경우, 상기 벡터가 산출된 인터넷 웹 사이트를 유해 사이트로 판단하는 것인, 유해 사이트 분류 방법.