(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0059735
(43) 공개일자 2025년05월07일

(51) 국제특허분류(Int. Cl.)
G06F 16/951 (2019.01) *G06F 16/955* (2019.01)
G06F 9/48 (2018.01) *G06F 9/50* (2018.01)
(52) CPC특허분류
G06F 16/951 (2019.01)
G06F 16/9566 (2019.01)
(21) 출원번호 10-2023-0143615
(22) 출원일자 2023년10월25일
심사청구일자 2023년10월25일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
박기웅
서울특별시 광진구 능동로17길 21, 304호 (화양동)
안성규
서울특별시 광진구 동일로56다길 3, B02호 (군자동)
배승진
강원특별자치도 강릉시 남부로125번길 7, 105동 1003호 (노암동, 노암3차한라아파트)
(74) 대리인
양성보

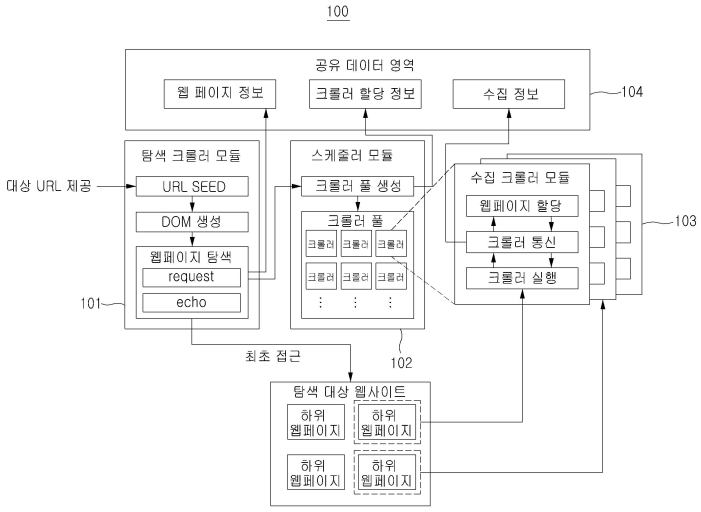
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **Swam 형태의 분산 크롤링에 대한 유동적 분산 제어기법을 통한 기밀-강인 데이터 수집 방법 및 시스템**

(57) 요 약

Swam 형태의 분산 크롤링에 대한 유동적 분산 제어기법을 통한 기밀-강인 데이터 수집 방법 및 시스템이 개시된다. 일 실시예에 따른 크롤러에 의해 수행되는 데이터 수집 방법은, 크롤링의 단일 기능을 수행하기 위해 모듈화된 다중 크롤러 모듈을 구성하는 단계; 및 상기 구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을 통해 병렬적 크롤링을 수행하는 단계를 포함할 수 있다.

대 표 도



(52) CPC특허분류
G06F 9/4881 (2013.01)
G06F 9/5077 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193909
과제번호 IITP-2021-0-01816-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 메타버스 자율트윈 핵심기술 연구
기 여 율 10/100
과제수행기관명 세종대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711179302
과제번호 00165794
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송기술국제공동연구
연구과제명 랜섬웨어 침해사고 전주기적 능동대응을 위한 다각적 수집-분석-대응 플랫폼 개발
기 여 율 30/100
과제수행기관명 세종대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711194419
과제번호 00208460
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 강인한 IoT 및 클라우드 시스템을 위한 전방위적 공격벡터 능동추출 및 사이버 면역력 강화 기술 연구
기 여 율 30/100
과제수행기관명 세종대학교 산학협력단
연구기간 2023.03.01 ~ 2024.02.29
이 발명을 지원한 국가연구개발사업
과제고유번호 2022000701
과제번호 2022-0-00701
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 국방ICT융합연구
연구과제명 국방정보통신망-상용망(5G) 연동을 위한 보안 기술개발
기 여 율 20/100
과제수행기관명 세종대학교 산학협력단
연구기간 2022.06.30 ~ 2026.04.30
이 발명을 지원한 국가연구개발사업
과제고유번호 1711195724
과제번호 00228996
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 실감콘텐츠핵심기술개발
연구과제명 우주상황인식을 위한 실-가상 연동형 국방 메타버스 기반기술 개발
기 여 율 10/100
과제수행기관명 한국전자통신연구원
연구기간 2023.04.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

데이터 수집 시스템에 의해 수행되는 데이터 수집 방법에 있어서,

크롤링의 단일 기능을 수행하기 위해 모듈화된 다중 크롤러 모듈을 구성하는 단계; 및

상기 구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을 통해 병렬적 크롤링을 수행하는 단계

를 포함하는 데이터 수집 방법.

청구항 2

제1항에 있어서,

상기 다중 크롤러 모듈은,

웹 페이지 내 크롤링에 요구되는 웹 페이지 정보를 탐색하는 탐색 크롤러 모듈, 상기 탐색된 웹 페이지 정보에 기초하여 작업 공간을 동적으로 할당하는 스케줄러 모듈 및 상기 할당된 작업 공간에 기초하여 크롤링 수행하는 수집 크롤러 모듈을 포함하는 것을 특징으로 하는 데이터 수집 방법.

청구항 3

제1항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

탐색 크롤러 모듈에서, 사용자에게 전달받은 URL 정보를 기반으로 DOM(Document Object Model) 구조를 생성하고, 리퀘스트(Request)와 에코(echo) 패킷을 전송함에 따라 콘텐츠의 크기(content_length)와 ping 명령어의 응답속도를 측정하는 단계

를 포함하는 데이터 수집 방법.

청구항 4

제3항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 측정된 응답속도에 기초하여 전체 비율 1에서 상기 ping 명령어의 응답속도에 콘텐츠의 크기를 나눈 값을 감산하여 웹 페이지 구성요소의 크기 및 크롤링을 위한 가중치를 산정하는 단계

를 포함하는 데이터 수집 방법.

청구항 5

제3항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 리퀘스트(Request)와 에코(echo) 패킷을 전송함에 따라 탐색된 웹 페이지 구성요소 및 하위 페이지를 포함하는 웹 페이지 정보를 공유 데이터 영역에 저장하는 단계

를 포함하는 데이터 수집 방법.

청구항 6

제1항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

스케줄러 모듈에서, 탐색 크롤러 모듈로부터 전달된 가중치를 기반으로 크롤러 풀을 이용하여 수집 크롤러 모듈의 개수를 연산하고, 상기 연산된 수집 크롤러 모듈의 개수에 기초하여 수집 크롤러 모듈을 생성하는 단계

를 포함하는 데이터 수집 방법.

청구항 7

제6항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 연산된 수집 크롤러 모듈의 개수에 기초하여 생성된 수집 크롤러 모듈을 작업 공간에 할당하고, 상기 생성된 수집 크롤러 모듈에게 상기 할당된 작업 공간이 가리키는 작업 정보를 전달하는 단계

를 포함하는 데이터 수집 방법.

청구항 8

제1항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

수집 크롤러 모듈에서, 스케줄러 모듈에 의해 할당된 작업 공간의 웹 페이지 구성요소 및 하위 페이지 정보를 포함하는 작업 정보를 통해 크롤링을 수행하는 단계

를 포함하는 데이터 수집 방법.

청구항 9

제8항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 작업 공간에 연결되어 있는 진행 상황을 이용하여 이어서 진행할 작업이나 수집 크롤러 모듈의 작업 공간을 확인한 후, 크롤링을 수행하는 단계

를 포함하는 데이터 수집 방법.

청구항 10

제9항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 수행된 크롤링이 완료되었는지 여부를 판단하고, 상기 수행된 크롤링이 완료되지 않았을 경우 크롤링 진행 상황을 업데이트하고, 상기 수행된 크롤링이 완료되었을 경우 스케줄러 모듈이 다음 작업 공간을 할당하였는지 여부를 판단하는 단계

를 포함하는 데이터 수집 방법.

청구항 11

제10항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 스케줄러 모듈이 다음 작업 공간을 할당하지 않은 것으로 판단됨에 따라 크롤링 진행 상황을 업데이트하고 크롤러 대기열로 이동하는 단계

를 포함하는 데이터 수집 방법.

청구항 12

제11항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 스케줄러 모듈이 다음 작업 공간을 할당한 것으로 판단됨에 따라 수집 크롤러 모듈을 할당된 다음 작업 공간에 배치하는 단계

를 포함하는 데이터 수집 방법.

청구항 13

제9항에 있어서,

상기 병렬적 크롤링을 수행하는 단계는,

상기 수행된 크롤링을 통해 수집된 정보를 공유 데이터 영역에 저장하는 단계

를 포함하는 데이터 수집 방법.

청구항 14

제1항 내지 제13항 중 어느 한 항의 데이터 수집 방법을 상기 데이터 수집 시스템에 실행시키기 위해 비-일시적인 컴퓨터 판독가능한 기록 매체에 저장되는 컴퓨터 프로그램.

청구항 15

데이터 수집 시스템에 있어서,

메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로세서는,

크롤링의 단일 기능을 수행하기 위해 모듈화된 다중 크롤러 모듈을 구성하고,

상기 구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을 통해 병렬적 크롤링을 수행하는

것을 특징으로 하는 데이터 수집 시스템.

발명의 설명

기 술 분 야

[0001] 아래의 설명은 데이터 크롤링 기술에 관한 것이다.

배 경 기 술

[0003] 웹의 규모가 폭발적으로 증가하면서 웹 페이지의 양 또한 기하급수적으로 늘어나고 있다. 초기에는 단일 크롤러를 사용해 웹을 순회하며 정보를 수집하는 중앙 집중형 구조가 주를 이루었다. 그러나, 현재 인터넷의 규모와 복잡성을 고려하면, 단일 크롤러만으로는 처리해야 할 웹 페이지의 양과 다양성에 어려움을 겪게 된다. 이로 인해 대규모 웹 데이터 수집에는 많은 인력과 시간이 소요되며, 효율적인 크롤링이 어려워지고 있다.

[0004] 또한, 웹 크롤링 기술의 발전과 함께 웹 페이지를 보유한 기관이나 단체들은 자신들의 정보가 유출되는 것을 방지하려는 시도를 활발히 하고 있다. 이에 대한 대표적인 예로, 웹 크롤러에 의한 트래픽 과부하나 너무 잦은 접속을 차단하는 조치가 있다. 이러한 환경 속에서 특히 악성코드 정보와 같은 민감한 데이터를 크롤링하는 것은 더욱 어려운 일이다. 악성코드 정보는 그 자체로 중요한 정보이며, 이를 악의를 가진 사용자에 의해 심각한 문제를 초래할 수 있기 때문에, 웹 페이지 소유자들은 해당 정보에 대한 크롤링을 특히 제한하곤 한다.

[0005] 따라서, 현대의 웹 크롤링 환경에서는 다중 분산 크롤러와 같은 새로운 접근법이 필요하며, 특히 악성코드 정보와 같은 민감한 데이터 크롤링에는 더욱 신중하고 전략적인 방법이 요구된다.

[0006] 일례로, 대한민국 공개특허 10-2020-0019288(2020.02.24. 공개일)에 JCA(Java Client Application), JSA(Java Server Application), DB 서버, 웹서버로 구성되며, JCA와 JSA에 각자의 독립된 스레드를 세 개 이상 나누어 작업을 실행함으로써, 독립된 환경에서의 클라이언트가 큐와 스레드를 이용하여 빠르고 신뢰성이 향상된 웹 문서를 제공하는 다중 스레드 방식의 웹 크롤링 시스템이 개시된 바 있다.

발명의 내용

해결하려는 과제

[0008] 크롤러의 기능을 모듈화함으로써, 기존 단일 크롤링에서 수행되는 크롤링을 수행하는 연산을 최소 단위로 분할하여 다중화하고, 각 모듈에게 별도의 크롤링 타겟을 부여함으로써, 각각의 모듈이 웹 페이지에 접속하여 상호 교류를 통해 능동적으로 정보를 수집하고, 웹 페이지의 접근 차단으로 인한 크롤링 기능의 중단을 방지하기 위해 다중 노드로 구성된 서버 및 클라우드 환경을 이용한다.

[0009] 각각의 크롤링 모듈이 서로 다른 웹 페이지를 대상으로 크롤링 기능을 수행하여 웹 페이지의 차단을 회피하고 다양한 정보를 병렬적으로 수집함으로써, 빠른 크롤링이 가능하도록 하여 웹 페이지로부터의 크롤링 차단 기능을 우회하고 빠른 크롤링이 가능할 수 있도록 한다.

과제의 해결 수단

[0011] 데이터 수집 시스템에 의해 수행되는 데이터 수집 방법은, 크롤링의 단일 기능을 수행하기 위해 모듈화된 다중 크롤러 모듈을 구성하는 단계; 및 상기 구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을 통해 병렬적 크롤링을 수행하는 단계를 포함할 수 있다.

[0012] 상기 다중 크롤러 모듈은, 웹 페이지 내 크롤링에 요구되는 웹 페이지 정보를 탐색하는 탐색 크롤러 모듈, 상기 탐색된 웹 페이지 정보에 기초하여 작업 공간을 동적으로 할당하는 스케줄러 모듈 및 상기 할당된 작업 공간에 기초하여 크롤링 수행하는 수집 크롤러 모듈을 포함할 수 있다.

[0013] 상기 병렬적 크롤링을 수행하는 단계는, 탐색 크롤러 모듈에서, 사용자에게 전달받은 URL 정보를 기반으로 DOM(Document Object Model) 구조를 생성하고, 리퀘스트(Request)와 에코(echo) 패킷을 전송함에 따라 콘텐츠의 크기(content_length)와 ping 명령어의 응답속도를 측정하는 단계를 포함할 수 있다.

[0014] 상기 병렬적 크롤링을 수행하는 단계는, 상기 측정된 응답속도에 기초하여 전체 비율 1에서 상기 ping 명령어의 응답속도에 콘텐츠의 크기를 나눈 값을 감산하여 웹 페이지 구성요소의 크기 및 크롤링을 위한 가중치를 산정하는 단계를 포함할 수 있다.

[0015] 상기 병렬적 크롤링을 수행하는 단계는, 상기 리퀘스트(Request)와 에코(echo) 패킷을 전송함에 따라 탐색된 웹 페이지 구성요소 및 하위 페이지를 포함하는 웹 페이지 정보를 공유 데이터 영역에 저장하는 단계를 포함할 수 있다.

[0016] 상기 병렬적 크롤링을 수행하는 단계는, 스케줄러 모듈에서, 탐색 크롤러 모듈로부터 전달된 가중치를 기반으로 크롤러 풀을 이용하여 수집 크롤러 모듈의 개수를 연산하고, 상기 연산된 수집 크롤러 모듈의 개수에 기초하여 수집 크롤러 모듈을 생성하는 단계를 포함할 수 있다.

[0017] 상기 병렬적 크롤링을 수행하는 단계는, 상기 연산된 수집 크롤러 모듈의 개수에 기초하여 생성된 수집 크롤러 모듈을 작업 공간에 할당하고, 상기 생성된 수집 크롤러 모듈에게 상기 할당된 작업 공간이 가리키는 작업 정보를 전달하는 단계를 포함할 수 있다.

[0018] 상기 병렬적 크롤링을 수행하는 단계는, 수집 크롤러 모듈에서, 스케줄러 모듈에 의해 할당된 작업 공간의 웹 페이지 구성요소 및 하위 페이지 정보를 포함하는 작업 정보를 통해 크롤링을 수행하는 단계를 포함할 수 있다.

[0019] 상기 병렬적 크롤링을 수행하는 단계는, 상기 작업 공간에 연결되어 있는 진행 상황을 이용하여 이어서 진행할 작업이나 수집 크롤러 모듈의 작업 공간을 확인한 후, 크롤링을 수행하는 단계를 포함할 수 있다.

[0020] 상기 병렬적 크롤링을 수행하는 단계는, 상기 수행된 크롤링이 완료되었는지 여부를 판단하고, 상기 수행된 크

롤링이 완료되지 않았을 경우 크롤링 진행 상황을 업데이트하고, 상기 수행된 크롤링이 완료되었을 경우 스케줄러 모듈이 다음 작업 공간을 할당하였는지 여부를 판단하는 단계를 포함할 수 있다.

[0021] 상기 병렬적 크롤링을 수행하는 단계는, 상기 스케줄러 모듈이 다음 작업 공간을 할당하지 않은 것으로 판단됨에 따라 크롤링 진행 상황을 업데이트하고 크롤러 대기열로 이동하는 단계를 포함할 수 있다.

[0022] 상기 병렬적 크롤링을 수행하는 단계는, 상기 스케줄러 모듈이 다음 작업 공간을 할당한 것으로 판단됨에 따라 수집 크롤러 모듈을 할당된 다음 작업 공간에 배치하는 단계를 포함할 수 있다.

[0023] 상기 병렬적 크롤링을 수행하는 단계는, 상기 수행된 크롤링을 통해 수집된 정보를 공유 데이터 영역에 저장하는 단계를 포함할 수 있다.

[0024] 데이터 수집 방법을 상기 데이터 수집 시스템에 실행시키기 위해 비-일시적인 컴퓨터 판독가능한 기록 매체에 저장되는 컴퓨터 프로그램을 포함할 수 있다.

[0025] 데이터 수집 시스템에 있어서, 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 크롤링의 단일 기능을 수행하기 위해 모듈화된 다중 크롤러 모듈을 구성하고, 상기 구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을 통해 병렬적 크롤링을 수행할 수 있다.

발명의 효과

[0027] 웹 페이지에서 정보를 수집하는 과정에서 지속적인 크롤링 기능의 수행으로 인한 트래픽 유발 등의 이유로 발생하는 크롤러의 차단을 최소화하고 능동적 병렬 방식의 크롤러를 구성함으로써, 웹 페이지 구성요소에 대한 분석 결과를 기반으로 필요 자원을 분배하여, 최적화된 크롤러 모듈의 병렬적인 실행을 통해 웹페이지 구성요소를 동시에 크롤링 할 수 있어 크롤링 과정에서 발생하는 소요 시간을 최소화할 수 있다.

도면의 간단한 설명

[0029] 도 1은 일 실시예에 있어서, 크롤러의 구조를 설명하기 위한 도면이다.

도 2는 일 실시예에 있어서, 데이터 수집 방법을 설명하기 위한 흐름도이다.

도 3은 일 실시예에 있어서, 탐색 크롤러 모듈의 동작을 설명하기 위한 흐름도이다.

도 4는 일 실시예에 있어서, 스케줄러 모듈의 동작을 설명하기 위한 도면이다.

도 5는 일 실시예에 있어서, 수집 크롤러 모듈의 동작을 설명하기 위한 도면이다.

도 6은 일 실시예에 있어서, 다수 개의 노드로 구성된 서버 또는 클라우드 환경을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0030] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0032] 도 1은 일 실시예에 있어서, 크롤러의 구조를 설명하기 위한 도면이다.

[0033] 데이터 수집 시스템(이하 '크롤러'로 기재하기로 함)(100)은 웹 페이지 내 크롤링에 요구되는 웹 페이지 정보를 탐색하는 탐색 크롤러 모듈(110), 탐색된 웹 페이지 정보에 기초하여 작업 공간을 동적으로 할당하는 스케줄러 모듈(120) 및 할당된 작업 공간에 기초하여 크롤링을 수행하는 수집 크롤러 모듈(130)로 구성될 수 있다.

[0034] 탐색 크롤러 모듈(101)은 웹 페이지의 구성 요소를 전체적으로 분석할 수 있다. 탐색 크롤러 모듈(101)은 DOM(Document Object Model)을 기반으로 웹 페이지(웹 사이트) 크롤링에 필요한 정보와 소요시간을 연산할 수 있다. 탐색 크롤러 모듈(101)은 DOM을 이용하여 웹 페이지 구성요소를 분석하고 구성요소와 관련된 링크에 리퀘스트(request)와 에코 패킷(echo packet)을 전송하여 콘텐츠의 크기(content_length)와 ping 명령어의 응답 속도를 측정할 수 있다. 이후, 탐색 크롤러 모듈(101)은 전체 비율 1에서 ping 속도에 콘텐츠의 크기를 나눈

값을 감산하여 링크에 대한 가중치를 산정할 수 있다. 이때, 산정된 가중치는 스케줄러 모듈(102)에게 전달되며, 크롤러의 총 개수에 가중치를 곱해 해당 작업을 수행하는데 필요한 크롤러의 개수가 도출될 수 있다. 탐색 크롤러 모듈(101)은 웹 페이지를 탐색함에 따라 획득된 웹 페이지 정보를 공유 데이터 영역(104)에 저장할 수 있다.

[0035] 스케줄러 모듈(102)은 필요한 크롤러의 개수만큼 작업 공간(배열)을 동적 할당하고, 할당된 작업 공간이 가리키는 작업 정보(크롤러 할당 정보)를 공유 데이터(102)에 정의할 수 있다. 스케줄러 모듈(102)은 소요 시간을 종합하여 크롤링 모듈을 할당할 수 있다.

[0036] 수집 크롤러 모듈(103)은 실제 크롤링 기능을 담당할 수 있다. 수집 크롤러 모듈(130)은 작업 공간의 빈 공간을 찾아 들어가게 되고, 작업 공간에 있는 공유 데이터 영역(104)에 저장된 크롤링 진행상황을 불러와 이어서 진행할 작업이나 자신의 작업 구역을 확인한 후 크롤링 코드를 실행할 수 있다. 수집 크롤러 모듈(103)은 크롤링을 실행함에 따라 수집된 정보를 공유 데이터 영역(104)에 저장할 수 있다.

[0037] 보다 상세하게는, 크롤러(100)는 사용자로부터 크롤링하기 위한 특정 웹 페이지에 대해 URL 정보(URL SEED)를 전달받아 웹 페이지 구성요소 및 하위 웹 페이지를 탐색하기 위한 탐색 크롤러 모듈(101), 탐색 크롤러 모듈(101)을 통해 수집된 웹 페이지 정보를 기반으로 수집 크롤러 모듈을 생성하기 위한 스케줄러 모듈(102), 스케줄러 모듈에 의해 할당받은 웹 페이지 구성요소의 정보를 크롤링하는 수집 크롤러 모듈(103), 탐색 크롤러 모듈(101), 스케줄러 모듈(102), 수집 크롤러 모듈(103)의 정보를 통합해서 관리하기 위한 공유 데이터 영역(104)으로 구성될 수 있다.

[0038] 탐색 크롤러 모듈(101)은 사용자에게 전달받은 URL 정보(즉, URL SEED)를 기반으로 하여 DOM 구조를 생성하고, 해당 웹 페이지에 리퀘스트와 에코 패킷을 전송하여 콘텐츠의 크기와 ping 명령어의 응답속도를 측정할 수 있다. 탐색 크롤러 모듈(101)은 측정된 응답속도에 기초하여 웹 페이지 구성요소의 크기 및 크롤링에 필요한 가중치를 연산할 수 있다. 탐색 크롤러 모듈(101)은 가중치 연산 결과를 스케줄러 모듈(102)에게 전달할 수 있다.

[0039] 스케줄러 모듈(102)은 탐색 크롤러 모듈(101)로부터 전달된 가중치를 기반으로 수집 크롤러 모듈(103)을 생성하기 위한 크롤러 풀을 활용함으로써, 크롤러 풀에서 가중치에 해당하는 크롤러의 개수를 할당할 수 있다.

[0040] 수집 크롤러 모듈(103)은 스케줄러 모듈(102)에 의해 웹 페이지 구성요소 및 하위 페이지 정보를 할당받아 크롤링 기능을 수행할 수 있다. 스케줄러 모듈(102)에 의해 생성된 크롤러 모듈(103)은 웹페이지 구성요소 및 하위 페이지의 개수에 따라 다수 개로 생성되어 병렬적으로 크롤링 연산을 수행할 수 있다. 수집 크롤러 모듈(103)의 각각은 통신 기능을 포함함으로써, 각 모듈 간의 크롤링 완료 여부 파악이 가능하므로, 수집 크롤러 모듈(103)에 의해 수집된 웹 페이지 정보는 공유 데이터 영역(104)로 전송될 수 있다. 정보 수집이 완료된 수집 크롤러 모듈(103)은 스케줄러 모듈(102)에게 완료 신호를 전송한 뒤 삭제되며 스케줄러 모듈(102)은 크롤러 풀에 새로운 크롤러 영역을 할당한다.

[0041] 도 2는 일 실시예에 있어서, 데이터 수집 방법을 설명하기 위한 흐름도이다.

[0042] 단계(210)에서 크롤러는 크롤링의 단일 기능을 수행하기 위해 모듈화된 다중 크롤러 모듈을 구성할 수 있다. 크롤러는 웹 페이지 내 크롤링에 요구되는 웹 페이지 정보를 탐색하는 탐색 크롤러 모듈, 탐색된 웹 페이지 정보에 기초하여 작업 공간을 동적으로 할당하는 스케줄러 모듈 및 할당된 작업 공간에 기초하여 크롤링 수행하는 수집 크롤러 모듈을 포함할 수 있다.

[0043] 단계(220)에서 크롤러는 구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을 통해 병렬적 크롤링을 수행할 수 있다. 크롤러는 탐색 크롤러 모듈을 통해 사용자에게 전달받은 URL 정보를 기반으로 DOM(Document Object Model) 구조를 생성하고, 리퀘스트(Request)와 에코(echo) 패킷을 전송함에 따라 콘텐츠의 크기(content_length)와 ping 명령어의 응답속도를 측정할 수 있다. 크롤러는 측정된 응답속도에 기초하여 전체 비율 1에서 ping 명령어의 응답속도에 콘텐츠의 크기를 나눈 값을 감산하여 웹 페이지 구성요소의 크기 및 크롤링을 위한 가중치를 산정할 수 있다. 크롤러는 리퀘스트(Request)와 에코(echo) 패킷을 전송함에 따라 탐색된 웹 페이지 구성요소 및 하위 페이지를 포함하는 웹 페이지 정보를 공유 데이터 영역에 저장할 수 있다. 또한, 크롤러는 스케줄러 모듈을 통해 탐색 크롤러 모듈로부터 전달된 가중치를 기반으로 크롤러 풀을 이용하여 수집 크롤러 모듈의 개수를 연산하고, 연산된 수집 크롤러 모듈의 개수에 기초하여 수집 크롤러 모듈을 생성할 수 있다. 크롤러는 연산된 수집 크롤러 모듈의 개수에 기초하여 생성된 수집 크롤러 모듈을 작업 공간에 할당하고, 생성된 수집 크롤러 모듈에게 할당된 작업 공간이 가리키는 작업 정보를 전달할 수 있다. 또한, 크

롤러는 수집 크롤러 모듈을 통해 스케줄러 모듈에 의해 할당된 작업 공간의 웹 페이지 구성요소 및 하위 페이지 정보를 포함하는 작업 정보를 통해 크롤링을 수행할 수 있다. 크롤러는 작업 공간에 연결되어 있는 진행 상황을 이용하여 이어서 진행할 작업이나 수집 크롤러 모듈의 작업 공간을 확인한 후, 크롤링을 수행할 수 있다. 크롤러는 수행된 크롤링이 완료되었는지 여부를 판단하고, 수행된 크롤링이 완료되지 않았을 경우 크롤링 진행 상황을 업데이트하고, 수행된 크롤링이 완료되었을 경우 스케줄러 모듈이 다음 작업 공간을 할당하였는지 여부를 판단할 수 있다. 크롤러는 스케줄러 모듈이 다음 작업 공간을 할당하지 않은 것으로 판단됨에 따라 크롤링 진행 상황을 업데이트하고 크롤러 대기열로 이동할 수 있다. 크롤러는 스케줄러 모듈이 다음 작업 공간을 할당한 것으로 판단됨에 따라 수집 크롤러 모듈을 할당된 다음 작업 공간에 배치할 수 있다. 크롤러는 수행된 크롤링을 통해 수집된 정보를 공유 데이터 영역에 저장할 수 있다.

[0044] 도 3은 일 실시예에 있어서, 탐색 크롤러 모듈의 동작을 설명하기 위한 흐름도이다.

[0045] 탐색 크롤러 모듈은 최초 웹 페이지를 탐색할 수 있다. 탐색 크롤러 모듈은 URL SEED를 통한 웹 페이지를 탐색할 수 있다(310). 탐색 크롤러 모듈은 탐색된 웹 페이지 내에 새로운 링크가 존재하는지 여부를 판단할 수 있다(320). 탐색 크롤러 모듈은 웹 페이지 구성요소 및 하위 웹 페이지에 접속이 가능한 새로운 링크가 존재한다면 해당 웹 페이지 구성요소를 대상으로 DOM을 제작할 수 있다(330). 탐색 크롤러 모듈은 웹 페이지 구성요소를 대상으로 DOM을 추가하여 구조를 분석한 뒤 예상 가중치를 계산하기 위해 리퀘스트와 에코 패킷을 전송함에 따라 콘텐츠의 크기와 ping 명령어의 응답속도를 측정하고, 측정된 응답속도에 기초하여 웹 페이지 구성요소의 크기 및 크롤링에 필요한 가중치를 계산할 수 있다(340). 탐색 크롤러 모듈은 가중치 계산이 완료됨에 따라 스케줄러 모듈에게 계산된 가중치 값을 전송할 수 있다(350). 이때, 스케줄러 모듈은 전송받은 가중치 값에 기초하여 크롤러 풀에서 수집 크롤러 모듈을 할당할 수 있다. 이와 같은 과정으로 웹 페이지의 구조를 미리 파악하고 필요 자원을 계산하여 다른 크롤링 기법보다 빠르게 연속적으로 수집할 수 있게 된다.

[0046] 도 4는 일 실시예에 있어서, 스케줄러 모듈의 동작을 설명하기 위한 도면이다.

[0047] 스케줄러 모듈은 작업 공간의 공유 데이터에 크롤러의 행위를 저장할 수 있다(410). 스케줄러 모듈은 탐색 크롤러 모듈에서 전달받은 가중치 값을 이용하여 작업 공간을 동적으로 할당할 수 있다(420). 스케줄러 모듈은 크롤러 풀에서 가용할 수 있는 수집 크롤러 모듈의 개수를 연산하고, 연산된 수집 크롤러 모듈의 개수에 기초하여 수집 크롤러 모듈을 작업 공간에 동적으로 할당할 수 있다. 스케줄러 모듈은 크롤러 풀에 예약된 크롤러 모듈을 이용하여 새로운 수집 크롤러 모듈을 생성하고, 수집 크롤러 모듈의 생성이 완료되는 과정에서 탐색 크롤러 모듈에서 수집된 웹 페이지 구성요소 및 하위 페이지 정보를 할당할 수 있다. 이때, 스케줄러 모듈은 수집 크롤러 모듈 생성 과정에서 사용자의 요구 및 시스템 가용자원에 따라 각 웹 페이지 구성요소 및 하위 웹 페이지 개수와 동일한 수집 크롤러 모듈을 생성하거나, 웹 페이지 구성요소보다 적은 개수의 수집 크롤러 모듈을 생성할 수 있다. 스케줄러 모듈은 작업 공간에 남는 자리가 있는지 판단할 수 있다(430). 스케줄러 모듈은 작업 공간에 남는 자리가 없는 것으로 판단될 경우, 일정시간 대기할 수 있다. 스케줄러 모듈은 일정시간 대기 후, 다시 작업 공간에 남는 자리가 있는지 판단할 수 있다. 스케줄러 모듈은 작업 공간에 남는 자리가 있는 것으로 판단될 경우, 작업이 완료된 수집 크롤러 모듈이 존재하는지 판단할 수 있다(440). 스케줄러 모듈은 작업이 완료된 수집 크롤러 모듈이 존재하는 것으로 판단되는 경우, 남는 자리가 있는 작업 공간에 수집 크롤러 모듈을 배치할 수 있다(450). 스케줄러 모듈은 작업이 완료된 수집 크롤러 모듈이 존재하지 않는 것으로 판단되는 경우, 프로세스를 종료할 수 있다.

[0048] 도 5는 일 실시예에 있어서, 수집 크롤러 모듈의 동작을 설명하기 위한 도면이다.

[0049] 수집 크롤러 모듈은 스케줄러 모듈에 의해 생성될 수 있다. 수집 크롤러 모듈은 스케줄러 모듈 내의 작업 공간 할당을 대기할 수 있다(510). 수집 크롤러 모듈은 생성 과정에서 크롤링 대상 웹 페이지 구성요소 및 하위 페이지에 대한 정보를 할당받을 수 있다. 수집 크롤러 모듈은 할당받은 정보를 토대로 크롤링을 수행할 수 있다. 수집 크롤러 모듈은 작업 공간에 연결되어 있는 진행 상황과 코드를 가져와 이어서 크롤링을 실행할 수 있다(520). 수집 크롤러 모듈은 크롤링을 완료하였는지 여부를 판단할 수 있다(530). 수집 크롤러 모듈은 크롤링을 완료하지 않았을 경우, 진행상황을 업데이트할 수 있다(540). 수집 크롤러 모듈은 크롤링을 완료하였을 경우, 스케줄러 모듈이 다음 작업 공간을 할당하였는지 여부를 판단할 수 있다(550). 수집 크롤러 모듈은 스케줄러 모듈이 다음 작업을 할당한 것으로 판단됨에 따라 할당된 다음 작업 공간에 수집 크롤러 모듈을 배치할 수 있다(570). 수집 크롤러 모듈은 스케줄러 모듈이 다음 작업 공간을 할당하지 않은 것으로 판단됨에 따라 크롤링 진행 상황을 업데이트하고 크롤러 대기열로 이동할 수 있다(560). 이때, 수집 크롤러 모듈은 크롤링 수행 과정에서 실시간으로 웹 페이지 구성요소 및 웹 페이지의 크롤링 범위 및 현재 크롤링 진행 과정에 대한 정보를

공유 데이터 영역에 전송할 수 있다. 또한, 수집 크롤러 모듈은 수집 크롤링 모듈에 포함되어 있는 통신 모듈을 이용하여 다른 수집 크롤링 모듈 또는 공유 데이터 영역에 필요한 데이터를 요청하거나 전송할 수 있다.

[0050] 도 6은 일 실시예에 있어서, 다수 개의 노드로 구성된 서버 또는 클라우드 환경을 설명하기 위한 도면이다.

[0051] 크롤링 환경은 다수 개의 노드로 구성된 서버 또는 클라우드 환경에서 구동될 수 있다. 각 노드는 크롤링 모듈을 포함하며 각각 다른 네트워크 IP 등을 사용하는 환경으로 구성될 수 있다.

[0052] 각 노드에는 각 크롤링 모듈을 구동하는데 필요한 개별 공유 데이터 영역이 구성되어 있으며, 각 노드에서 생성된 최종 데이터는 별도로 구성된 최상위 공유 데이터 영역으로 전송되어, 최종적으로 사용자는 최상위 공유 데이터 영역을 통해 크롤링 결과 및 크롤링 데이터를 확인할 수 있다.

[0053] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0054] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

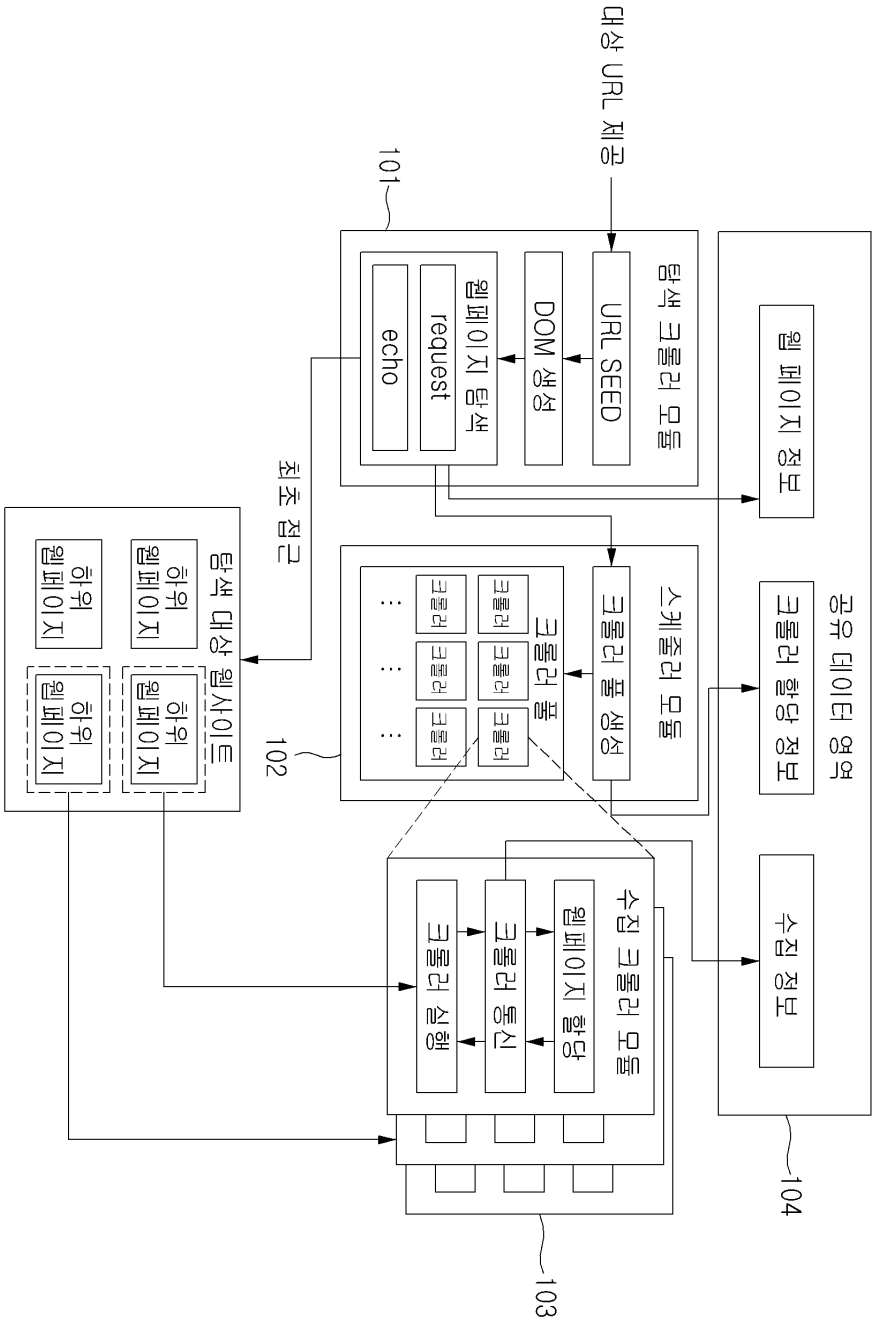
[0055] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0056] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

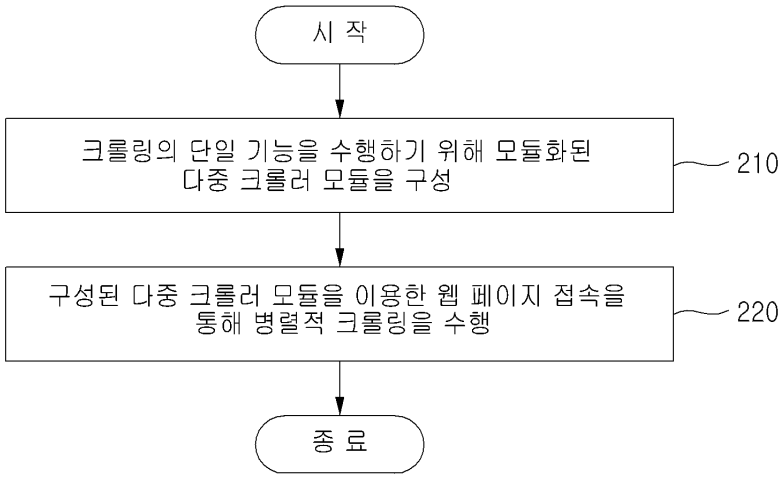
[0057] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

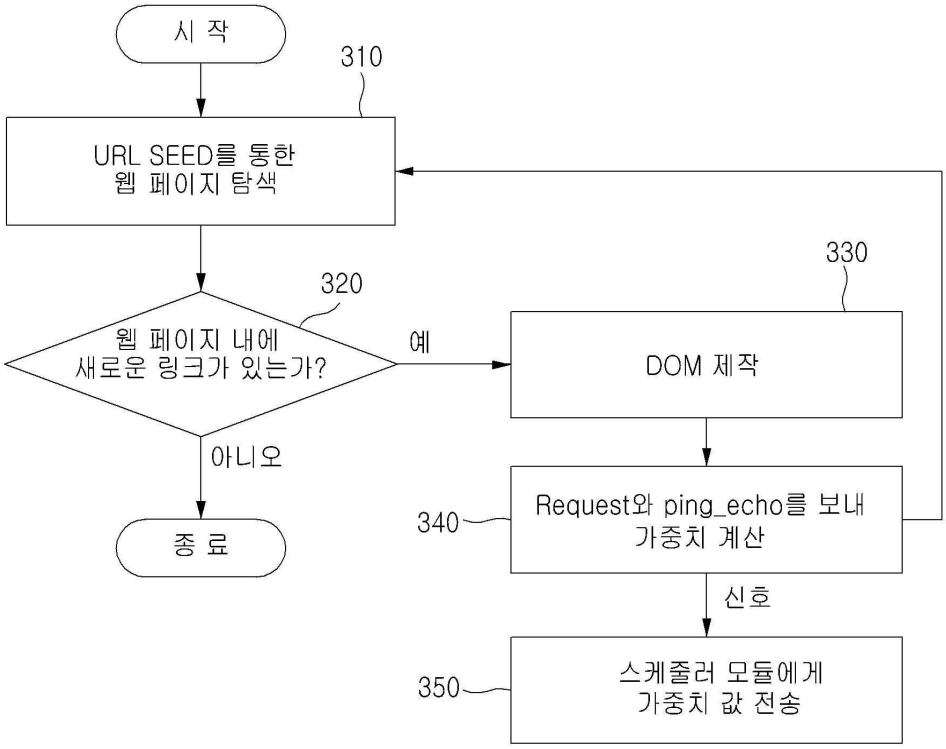
도면1


100
대상 URL 제공
101
탐색 크롤러 모듈
URL SEED
DOM 생성
웹페이지 탐색
request
echo
최초 접근
스케줄러 모듈
크롤러 풀 생성
크롤러 풀
크롤러
탐색 대상 웹사이트
하위 웹페이지
102
수집 크롤러 모듈
웹페이지 할당
크롤러 통신
크롤러 실행
103
공유 데이터 영역
웹 페이지 정보
크롤러 할당 정보
수집 정보
104

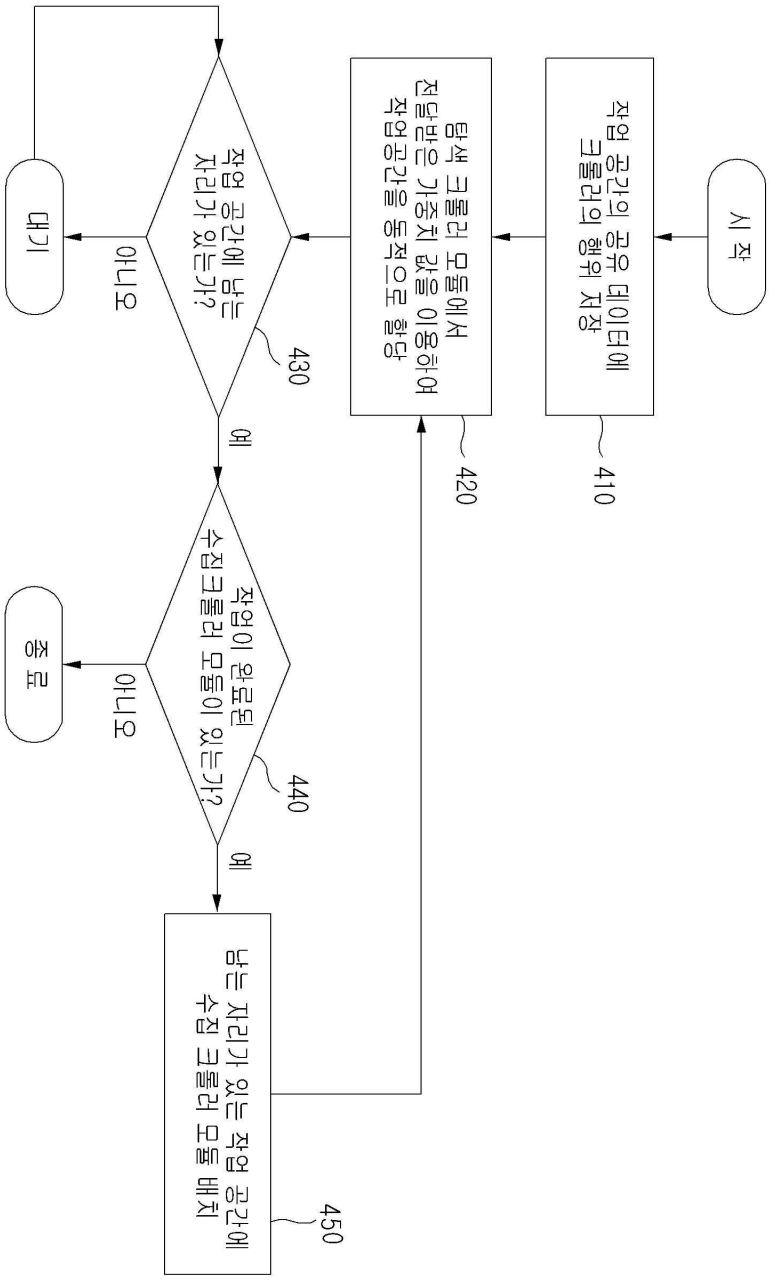
도면2


시 작
크롤링의 단일 기능을 수행하기 위해 모듈화된
다중 크롤러 모듈을 구성
210
구성된 다중 크롤러 모듈을 이용한 웹 페이지 접속을
통해 병렬적 크롤링을 수행
220
종 료

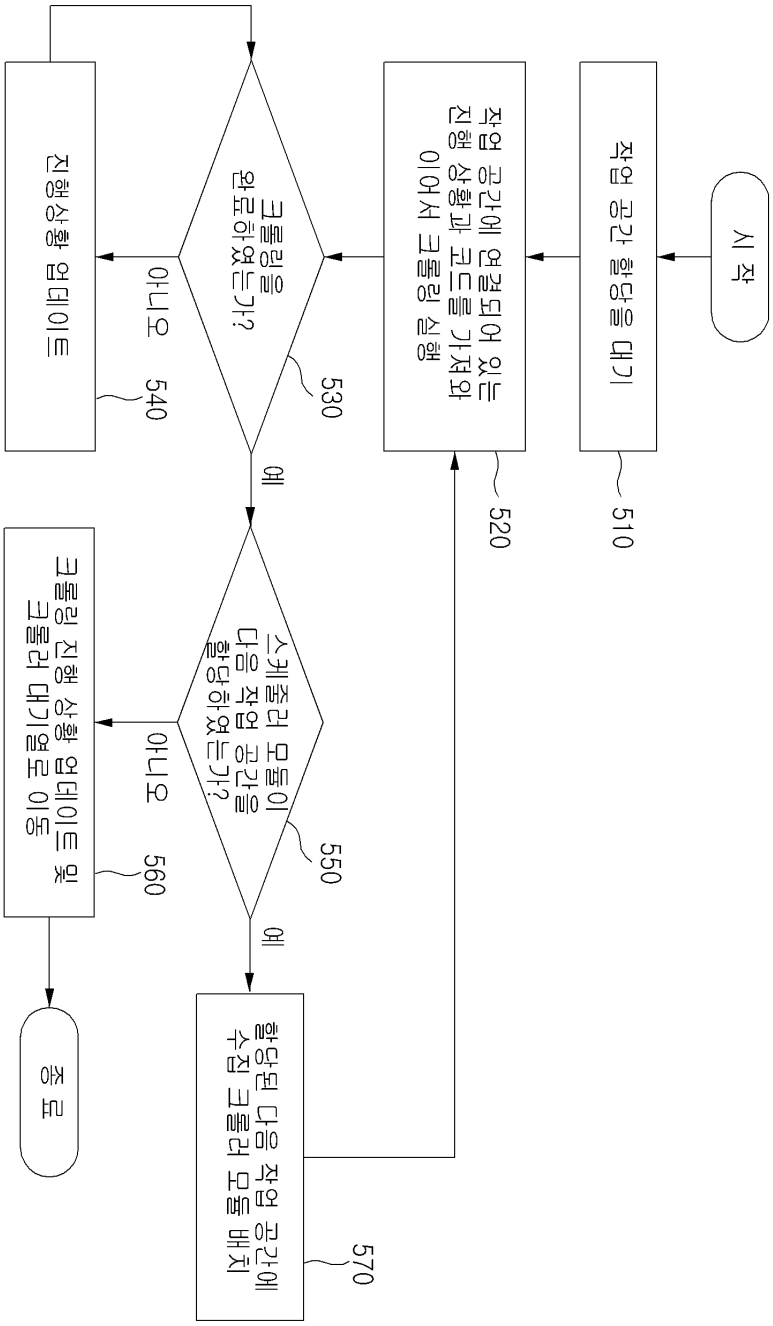
도면3


시 작
310
URL SEED를 통한
웹 페이지 탐색
320
웹 페이지 내에
새로운 링크가 있는가?
예
330
DOM 제작
아니오
종 료
340
Request와 ping_echo를 보내
가중치 계산
신호
350
스케줄러 모듈에게
가중치 값 전송

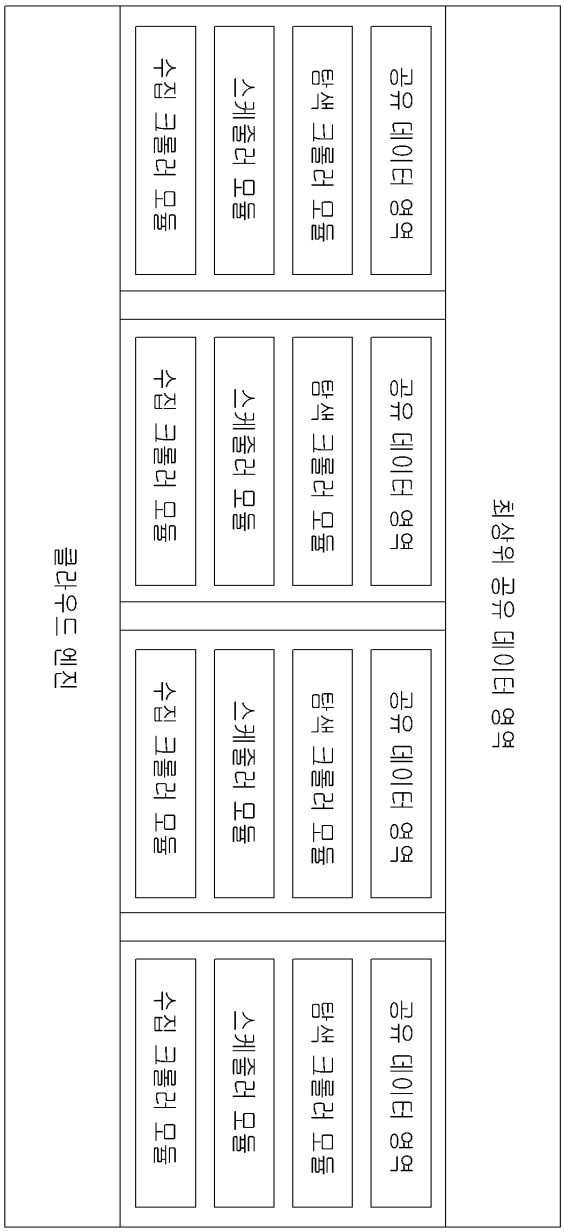
도면4


시 작
작업 공간의 공유 데이터에 크롤러의 행위 저장
410
탐색 크롤러 모듈에서 전달받은 가중치 값을 이용하여 작업공간을 동적으로 할당
420
작업 공간에 남는 자리가 있는가?
430
아니오
대기
예
작업이 완료된 수집크롤러 모듈이 있는가?
440
아니오
종료
예
남는 자리가 있는 작업 공간에 수집 크롤러 모듈 배치
450

도면5


시 작
작업 공간 할당을 대기
510
작업 공간에 연결되어 있는
진행 상황과 코드를 가져와
이어서 크롤링 실행
520
크롤링을
완료하였는가?
530
아니오
진행상황 업데이트
540
예
스케줄러 모듈이
다음 작업 공간을
할당하였는가?
550
아니오
크롤링 진행 상황 업데이트 및
크롤러 대기열로 이동
560
예
할당된 다음 작업 공간에
수집 크롤러 모듈 배치
570
종 료

도면6


최상위 공유 데이터 영역
공유 데이터 영역
탐색 크롤러 모듈
스케줄러 모듈
수집 크롤러 모듈
공유 데이터 영역
탐색 크롤러 모듈
스케줄러 모듈
수집 크롤러 모듈
공유 데이터 영역
탐색 크롤러 모듈
스케줄러 모듈
수집 크롤러 모듈
클라우드 엔진