



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년03월21일
(11) 등록번호 10-1840853
(24) 등록일자 2018년03월15일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) G06F 17/27 (2006.01)
G06F 9/50 (2018.01)
(52) CPC특허분류
G06F 17/3089 (2013.01)
G06F 17/2755 (2013.01)
(21) 출원번호 10-2016-0166879
(22) 출원일자 2016년12월08일
심사청구일자 2016년12월08일
(56) 선행기술조사문헌
KR1019970002692 A

(73) 특허권자
주식회사 트레드링스
인천광역시 계양구 봉오대로 723 ,4층(작전동,
상록빌딩)
(72) 발명자
박민규
서울특별시 강남구 학동로17길 21, 101호 (논현동, 나산빌라)
구석모
서울특별시 서초구 주홍길 31-18, 201호 (반포동)
(74) 대리인
리엔목특허법인

전체 청구항 수 : 총 20 항

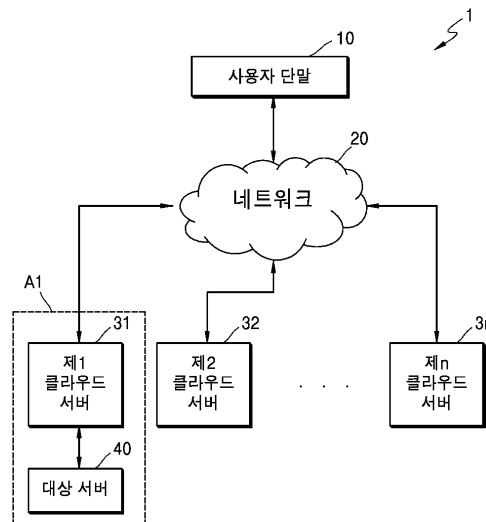
심사관 : 이복현

(54) 발명의 명칭 웹 문서 분산 수집 시스템, 방법 및 컴퓨터 프로그램

(57) 요약

본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템은, 웹 문서 수집 요청을 수신하는 사용자 인터페이스; 작업 목록에 기초하여 대상 서버로부터 웹 문서를 수집하는 복수의 가상 머신들; 및 상기 웹 문서 수집 요청에 대응하여 대상 서버와 관련된 대상 지역을 결정하고, 상기 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성하고, 상기 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 상기 복수의 가상 머신들에 상기 작업 목록을 분배하고, 상기 복수의 가상 머신들로부터 각각 분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 스케줄러;를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06F 17/30887 (2013.01)

G06F 9/5077 (2013.01)

명세서

청구범위

청구항 1

웹 문서 수집 요청을 수신하는 사용자 인터페이스;

작업 목록에 기초하여 대상 서버로부터 웹 문서를 수집하는 복수의 가상 머신들; 및

상기 웹 문서 수집 요청에 대응하여 대상 서버와 관련된 대상 지역을 결정하고, 상기 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성하고, 상기 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 상기 복수의 가상 머신들에 상기 작업 목록을 분배하고, 상기 복수의 가상 머신들로부터 각각 분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 스케줄러;를 포함하는, 웹 문서 분산 수집 시스템.

청구항 2

청구항 1에 있어서,

상기 웹 문서 수집 요청은 상기 대상 서버의 URL(uniform resource locator)을 포함하고,

상기 스케줄러는, 상기 대상 서버의 상기 URL에 대응하여 상기 대상 지역을 결정하고,

상기 대상 지역은 상기 대상 서버가 위치한 지역 또는 상기 대상 서버가 위치한 지역과 인접한 지역인, 웹 문서 분산 수집 시스템.

청구항 3

청구항 1에 있어서,

상기 스케줄러는, 상기 대상 서버로부터 최초로 웹 문서를 수집하는 경우에는 3개의 가상 머신들을 생성하고, 상기 대상 서버로부터 웹 문서를 수집한 적이 있는 경우에는 가장 최근에 생성한 가상 머신의 개수만큼의 가상 머신들을 생성하는, 웹 문서 분산 수집 시스템.

청구항 4

청구항 1에 있어서,

상기 웹 문서 수집 요청은 형태소를 포함하고,

상기 형태소는 문자 타입, 날짜 타입, 시간 타입, 또는 정수 타입인, 웹 문서 분산 수집 시스템.

청구항 5

청구항 4에 있어서,

상기 스케줄러가 수신하는 상기 대상 서버로부터 수집된 웹 문서는, 상기 복수의 가상 머신들이 각각 상기 형태소에 기초하여 파싱(parsing)한 웹 문서인, 웹 문서 분산 수집 시스템.

청구항 6

청구항 1에 있어서,

상기 스케줄러는, 상기 복수의 가상 머신들에 상기 작업 목록을 균등하게 분배하는, 웹 문서 분산 수집 시스템.

청구항 7

청구항 1에 있어서,

상기 스케줄러는, 상기 복수의 가상 머신들에 상기 작업 목록을 주기적으로 분배하는, 웹 문서 분산 수집 시스템.

템.

청구항 8

청구항 1에 있어서,

상기 스케줄러는, 상기 복수의 가상 머신들로부터 각각 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수를 포함하는 정보를 더 수신하고, 상기 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수에 기초하여 상기 복수의 가상 머신들에 상기 작업 목록을 재분배하는, 웹 문서 분산 수집 시스템.

청구항 9

청구항 1에 있어서,

상기 스케줄러는, 상기 대상 서버로부터 거부 메시지를 수신한 경우에는 상기 복수의 가상 머신들로부터 상기 거부 메시지에 포함된 가상 머신을 삭제하는, 웹 문서 분산 수집 시스템.

청구항 10

청구항 9에 있어서,

상기 스케줄러는, 상기 대상 지역에 설치된 다른 클라우드 컴퓨팅의 리소스를 이용하여 새로운 가상 머신을 생성하고, 삭제된 가상 머신에 분배된 작업 목록을 상기 새로운 가상 머신에 재분배하고, 상기 새로운 가상 머신으로부터 재분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는, 웹 문서 분산 수집 시스템.

청구항 11

사용자 인터페이스에 의해, 웹 문서 수집 요청을 수신하는 단계;

스케줄러에 의해, 상기 웹 문서 수집 요청에 대응하여 대상 서버와 관련된 대상 지역을 결정하는 단계;

상기 스케줄러에 의해, 상기 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성하는 단계;

상기 스케줄러에 의해, 상기 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 상기 복수의 가상 머신들에 상기 작업 목록을 분배하는 단계; 및

상기 스케줄러에 의해, 상기 복수의 가상 머신들로부터 각각 분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 단계;를 포함하는, 웹 문서 분산 수집 방법.

청구항 12

청구항 11에 있어서,

상기 웹 문서 수집 요청은 상기 대상 서버의 URL(uniform resource locator)을 포함하고,

상기 대상 지역을 결정하는 단계는,

상기 스케줄러에 의해, 상기 대상 서버의 상기 URL에 대응하여 상기 대상 지역을 결정하는 단계이고,

상기 대상 지역은 상기 대상 서버가 위치한 지역 또는 상기 대상 서버가 위치한 지역과 인접한 지역인, 웹 문서 분산 수집 방법.

청구항 13

청구항 11에 있어서,

상기 복수의 가상 머신들을 생성하는 단계는,

상기 스케줄러에 의해, 상기 대상 서버로부터 최초로 웹 문서를 수집하는 경우에는 3개의 가상 머신들을 생성하고, 상기 대상 서버로부터 웹 문서를 수집한 적이 있는 경우에는 가장 최근에 생성한 가상 머신의 개수만큼의

가상 머신들을 생성하는 단계인, 웹 문서 분산 수집 방법.

청구항 14

청구항 11에 있어서,

상기 웹 문서 수집 요청은 형태소를 포함하고,

상기 형태소는 문자 타입, 날짜 타입, 시간 타입, 또는 정수 타입인, 웹 문서 분산 수집 방법.

청구항 15

청구항 14에 있어서,

상기 스케줄러가 수신하는 상기 대상 서버로부터 수집된 웹 문서는, 상기 복수의 가상 머신들이 각각 상기 형태소에 기초하여 파싱(parsing)한 웹 문서인, 웹 문서 분산 수집 방법.

청구항 16

청구항 11에 있어서,

상기 작업 목록을 분배하는 단계는,

상기 스케줄러에 의해, 상기 복수의 가상 머신들에 상기 작업 목록을 균등하게 분배하는 단계인, 웹 문서 분산 수집 방법.

청구항 17

청구항 11에 있어서,

상기 작업 목록을 분배하는 단계는,

상기 스케줄러에 의해, 상기 복수의 가상 머신들에 상기 작업 목록을 주기적으로 분배하는 단계인, 웹 문서 분산 수집 방법.

청구항 18

청구항 11에 있어서,

상기 스케줄러에 의해, 상기 복수의 가상 머신들로부터 각각 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수를 포함하는 정보를 수신하는 단계; 및

상기 스케줄러에 의해, 상기 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수에 기초하여 상기 복수의 가상 머신들에 상기 작업 목록을 재분배하는 단계;를 더 포함하는, 웹 문서 분산 수집 방법.

청구항 19

청구항 11에 있어서,

상기 스케줄러에 의해, 상기 대상 서버로부터 거부 메시지를 수신한 경우에는 상기 복수의 가상 머신들로부터 상기 거부 메시지에 포함된 가상 머신을 삭제하는 단계; 및

상기 스케줄러에 의해, 상기 대상 지역에 설치된 다른 클라우드 컴퓨팅의 리소스를 이용하여 새로운 가상 머신을 생성하고, 삭제된 가상 머신에 분배된 작업 목록을 상기 새로운 가상 머신에 재분배하고, 상기 새로운 가상 머신으로부터 재분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 단계;를 더 포함하는, 웹 문서 분산 수집 방법.

청구항 20

컴퓨터를 이용하여 제11항 내지 제19항 중 어느 한 항의 방법을 실행하기 위하여 기록매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명의 실시예들은 웹 문서 분산 수집 시스템, 방법 및 컴퓨터 프로그램에 관한 것이다.

배경 기술

[0002] 일반적으로, 서버에 접속하고, 접속한 서버로부터 웹 문서를 다운로드함으로써, 웹 문서를 수집할 수 있다.

[0003] 한편, 서버는 웹 문서 수집을 위한 지속적인 접속을 DDOS(distributed denial of service) 공격으로 오인할 수 있다. 이에 따라, 서버가 웹 문서 수집을 위한 접속을 차단하는 문제를 해결할 필요가 있다.

발명의 내용

해결하려는 과제

[0004] 본 발명의 실시예들은 서버에 의한 접속 차단을 방지하는 웹 문서 분산 수집 시스템, 방법 및 컴퓨터 프로그램을 제공하고자 한다.

과제의 해결 수단

[0005] 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템은, 웹 문서 수집 요청을 수신하는 사용자 인터페이스; 작업 목록에 기초하여 대상 서버로부터 웹 문서를 수집하는 복수의 가상 머신들; 및 상기 웹 문서 수집 요청에 대응하여 대상 서버와 관련된 대상 지역을 결정하고, 상기 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성하고, 상기 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 상기 복수의 가상 머신들에 상기 작업 목록을 분배하고, 상기 복수의 가상 머신들로부터 각각 분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 스케줄러;를 포함한다.

[0006] 본 실시예에 따르면, 상기 웹 문서 수집 요청은 상기 대상 서버의 URL(uniform resource locator)을 포함하고, 상기 스케줄러는, 상기 대상 서버의 상기 URL에 대응하여 상기 대상 지역을 결정하고, 상기 대상 지역은 상기 대상 서버가 위치한 지역 또는 상기 대상 서버가 위치한 지역과 인접한 지역일 수 있다.

[0007] 본 실시예에 따르면, 상기 스케줄러는, 상기 대상 서버로부터 최초로 웹 문서를 수집하는 경우에는 3개의 가상 머신들을 생성하고, 상기 대상 서버로부터 웹 문서를 수집한 적이 있는 경우에는 가장 최근에 생성한 가상 머신의 개수만큼의 가상 머신들을 생성할 수 있다.

[0008] 본 실시예에 따르면, 상기 웹 문서 수집 요청은 형태소를 포함하고, 상기 형태소는 문자 타입, 날짜 타입, 시간 타입, 또는 정수 타입일 수 있다.

[0009] 본 실시예에 따르면, 상기 스케줄러가 수신하는 상기 대상 서버로부터 수집된 웹 문서는, 상기 복수의 가상 머신들이 각각 상기 형태소에 기초하여 파싱(parsing)한 웹 문서일 수 있다.

[0010] 본 실시예에 따르면, 상기 스케줄러는, 상기 복수의 가상 머신들에 상기 작업 목록을 균등하게 분배할 수 있다.

[0011] 본 실시예에 따르면, 상기 스케줄러는, 상기 복수의 가상 머신들에 상기 작업 목록을 주기적으로 분배할 수 있다.

[0012] 본 실시예에 따르면, 상기 스케줄러는, 상기 복수의 가상 머신들로부터 각각 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수를 포함하는 정보를 더 수신하고, 상기 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수에 기초하여 상기 복수의 가상 머신들에 상기 작업 목록을 재분배할 수 있다.

[0013] 본 실시예에 따르면, 상기 스케줄러는, 상기 대상 서버로부터 거부 메시지를 수신한 경우에는 상기 복수의 가상 머신들로부터 상기 거부 메시지에 포함된 가상 머신을 삭제할 수 있다.

[0014] 본 실시예에 따르면, 상기 스케줄러는, 상기 대상 지역에 설치된 다른 클라우드 컴퓨팅의 리소스를 이용하여 새로운 가상 머신을 생성하고, 삭제된 가상 머신에 분배된 작업 목록을 상기 새로운 가상 머신에 재분배하고, 상기 새로운 가상 머신으로부터 재분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신할

수 있다.본 발명의 일 실시예에 따른 웹 문서 분산 수집 방법은, 사용자 인터페이스에 의해, 웹 문서 수집 요청을 수신하는 단계; 스케줄러에 의해, 상기 웹 문서 수집 요청에 대응하여 대상 서버와 관련된 대상 지역을 결정하는 단계; 상기 스케줄러에 의해, 상기 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성하는 단계; 상기 스케줄러에 의해, 상기 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 상기 복수의 가상 머신들에 상기 작업 목록을 분배하는 단계; 및 상기 스케줄러에 의해, 상기 복수의 가상 머신들로부터 각각 분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 단계;를 포함한다.

[0015] 본 실시예에 따르면, 상기 웹 문서 수집 요청은 상기 대상 서버의 URL(uniform resource locator)을 포함하고, 상기 대상 지역을 결정하는 단계는, 상기 스케줄러에 의해, 상기 대상 서버의 상기 URL에 대응하여 상기 대상 지역을 결정하는 단계이고, 상기 대상 지역은 상기 대상 서버가 위치한 지역 또는 상기 대상 서버가 위치한 지역과 인접한 지역일 수 있다.

[0016] 본 실시예에 따르면, 상기 복수의 가상 머신들을 생성하는 단계는, 상기 스케줄러에 의해, 상기 대상 서버로부터 최초로 웹 문서를 수집하는 경우에는 3개의 가상 머신들을 생성하고, 상기 대상 서버로부터 웹 문서를 수집한 적이 있는 경우에는 가장 최근에 생성한 가상 머신의 개수만큼의 가상 머신들을 생성하는 단계일 수 있다.

[0017] 본 실시예에 따르면, 상기 웹 문서 수집 요청은 형태소를 포함하고, 상기 형태소는 문자 타입, 날짜 타입, 시간 타입, 또는 정수 타입일 수 있다.

[0018] 본 실시예에 따르면, 상기 스케줄러가 수신하는 상기 대상 서버로부터 수집된 웹 문서는, 상기 복수의 가상 머신들이 각각 상기 형태소에 기초하여 파싱(parsing)한 웹 문서일 수 있다.

[0019] 본 실시예에 따르면, 상기 작업 목록을 분배하는 단계는, 상기 스케줄러에 의해, 상기 복수의 가상 머신들에 상기 작업 목록을 균등하게 분배하는 단계일 수 있다.

[0020] 본 실시예에 따르면, 상기 작업 목록을 분배하는 단계는, 상기 스케줄러에 의해, 상기 복수의 가상 머신들에 상기 작업 목록을 주기적으로 분배하는 단계일 수 있다.

[0021] 본 실시예에 따르면, 상기 스케줄러에 의해, 상기 복수의 가상 머신들로부터 각각 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수를 포함하는 정보를 수신하는 단계; 및 상기 스케줄러에 의해, 상기 가상 머신이 상기 대상 서버로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수에 기초하여 상기 복수의 가상 머신들에 상기 작업 목록을 재분배하는 단계;를 더 포함할 수 있다.

[0022] 본 실시예에 따르면, 상기 스케줄러에 의해, 상기 대상 서버로부터 거부 메시지를 수신한 경우에는 상기 복수의 가상 머신들로부터 상기 거부 메시지에 포함된 가상 머신을 삭제하는 단계; 및 상기 스케줄러에 의해, 상기 대상 지역에 설치된 다른 클라우드 컴퓨팅의 리소스를 이용하여 새로운 가상 머신을 생성하고, 삭제된 가상 머신에 분배된 작업 목록을 상기 새로운 가상 머신에 재분배하고, 상기 새로운 가상 머신으로부터 재분배된 작업 목록에 기초하여 상기 대상 서버로부터 수집된 웹 문서를 수신하는 단계;를 더 포함할 수 있다.

[0023] 본 발명의 일 실시예에 따른 컴퓨터 프로그램은, 컴퓨터를 이용하여 본 발명의 일 실시예에 따른 웹 문서 분산 수집 방법을 실행하기 위하여 기록매체에 저장된다.

[0024] 진술한 것 외의 다른 측면, 특징, 이점이 이하의 도면, 특허청구범위 및 발명의 상세한 설명으로부터 명확해 질 것이다.

발명의 효과

[0025] 본 발명의 실시예들에 따르면, 대상 서버에 의한 접속 차단을 방지하고, 웹 문서의 수집 시간을 단축시킬 수 있는 웹 문서 수집 서비스를 제공할 수 있다.

[0026] 또한, 본 발명의 실시예들에 따르면, 대상 서버와 지역적으로 가까운 리소스를 이용함으로써, 신뢰도가 보다 향상된 웹 문서 수집 서비스를 제공할 수 있다.

도면의 간단한 설명

[0027] 도 1은 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템을 개략적으로 도시한 블록도이다.

도 3은 본 발명의 일 실시예에 따른 웹 문서 분산 수집 방법을 설명하기 위한 흐름도이다.

도 4는 본 발명의 일 실시예에 따른 웹 문서 분산 수집 방법을 상세하게 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0028] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시 예를 가질 수 있는 바, 특정 실시 예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변환, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.
- [0029] 이하의 실시예에서, 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 용어들에 의해 한정되어서는 안 된다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.
- [0030] 이하의 실시예에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 다수의 표현을 포함한다. 이하의 실시예에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0031] 본 발명의 실시예들은 기능적인 블록 구성들 및 다양한 처리 단계들로 나타내어질 수 있다. 이러한 기능 블록들은 특정 기능들을 실행하는 다양한 개수의 하드웨어 또는/및 소프트웨어 구성들로 구현될 수 있다. 예를 들어, 본 발명의 실시예들은 하나 이상의 마이크로프로세서들의 제어 또는 다른 제어 장치들에 의해서 다양한 기능들을 실행할 수 있는, 메모리, 프로세싱, 로직(logic), 룩업 테이블(look-up table) 등과 같은 직접 회로 구성들을 채용할 수 있다. 본 발명의 실시예의 구성 요소들이 소프트웨어 프로그래밍 또는 소프트웨어 요소들로 실행될 수 있는 것과 유사하게, 본 발명의 실시예는 데이터 구조, 프로세스들, 루틴들 또는 다른 프로그래밍 구성들의 조합으로 구현되는 다양한 알고리즘을 포함하여, C, C++, 자바(Java), 어셈블러(assembler) 등과 같은 프로그래밍 또는 스크립팅 언어로 구현될 수 있다. 기능적인 측면들은 하나 이상의 프로세서들에서 실행되는 알고리즘으로 구현될 수 있다. 또한, 본 발명의 실시예들은 전자적인 환경 설정, 신호 처리, 및/또는 데이터 처리 등을 위하여 종래 기술을 채용할 수 있다. 매킨리즘, 요소, 수단, 구성과 같은 용어는 넓게 사용될 수 있으며, 기계적이고 물리적인 구성들로서 한정되는 것은 아니다. 상기 용어는 프로세서 등과 연계하여 소프트웨어의 일련의 처리들(routines)의 의미를 포함할 수 있다.
- [0032] 도 1은 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템을 설명하기 위한 도면이다.
- [0033] 도 1을 참조하면, 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템(1)은 사용자 단말(10), 네트워크(20), 복수의 클라우드 서버(31 내지 3n), 대상 서버(40)를 포함한다.
- [0034] 웹 문서 분산 수집 시스템(1)은 제1 클라우드 서버(31)에서 수집된 대상 서버(40)의 웹 문서가 네트워크(20)를 통해 사용자 단말(10)에 전송될 수 있는 구성을 제공할 수 있다.
- [0035] 사용자 단말(10)은 관리자의 웹 문서 수집 요청을 수신하고, 웹 문서 수집 요청에 대응하는 웹 문서 수집 결과를 관리자에게 제공한다.
- [0036] 사용자 단말(10)은 관리자의 사용자 입력을 수신하고, 최종 웹 문서 수집 결과를 수신하여 관리자에게 제공할 수 있다. 한편, 사용자 단말(10)은 관리자의 사용자 입력에 대응하여 복수의 가상 머신들을 생성하고, 복수의 가상 머신들로부터 수신한 복수의 웹 문서들 중에서 필요한 정보만을 편집하여 관리자에게 제공할 수도 있다.
- [0037] 사용자 단말(10)은 적어도 하나 이상의 프로세서를 포함할 수 있다. 사용자 단말(10)은 마이크로 프로세서나 범용 컴퓨터 시스템과 같은 다른 하드웨어 장치에 포함된 형태로 구동될 수 있다. 사용자 단말(10)의 일부는 하드웨어 장치로 구현되고, 나머지 일부는 컴퓨터 프로그램에 의하여 구현될 수 있다.
- [0038] 네트워크(20)는 유선 네트워크 또는 무선 네트워크를 포함할 수 있다. 무선 네트워크는 2G(Generation) 또는 3G 셀룰러 통신 시스템, 3GPP(3rd Generation Partnership Project), 4G 통신 시스템, LTE(Long-Term Evolution), WiMAX(World Interoperability for Microwave Access) 등이 될 수 있다.
- [0039] 사용자 단말(10)은 네트워크(20)를 통해 복수의 클라우드 서버(31 내지 3n)를 이용하여 대상 서버(40)에 접근할

수 있다. 예를 들어, 복수의 클라우드 서버(31 내지 3n)는 각각 퍼블릭 클라우드 서비스(public cloud service)를 제공할 수 있다. 즉, 복수의 클라우드 서버(31 내지 3n)는 클라우드 컴퓨팅의 리소스일 수 있다.

[0040] 대상 서버(40)는 웹 문서를 저장 및 제공할 수 있다. 대상 서버(40)는 대상 서버(40)에 접근하는 리소스 관련 정보 예컨대, 리소스의 식별 정보, 요청된 웹 문서의 식별 정보, 웹 문서 요청 수신 시간, 웹 문서 응답 전송 시간 등을 저장할 수 있다. 대상 서버(40)는 복수의 클라우드 서버(31 내지 3n) 중에서 접근이 허용된 리소스에 웹 문서를 제공할 수 있다.

[0041] 한편, 대상 서버(40)는 소정 시간 동안 소정 횟수 이상 접근한 리소스의 접근을 차단할 수 있다. 이때, 대상 서버(40)는 복수의 클라우드 서버(31 내지 3n) 중에서 접근이 차단된 리소스의 웹 문서 요청에 대응하여 거부 메시지를 전송할 수 있다.

[0042] 복수의 클라우드 서버(31 내지 3n)는 각각 대상 서버(40)에 접근할 수 있다. 복수의 클라우드 서버(31 내지 3n)는 각각 대상 서버(40)의 URL(uniform resource locator)을 이용하여 대상 서버(40)에 접근할 수 있다.

[0043] 한편, 웹 문서 분산 수집 시스템(1)은 제1 지역(A1)에 설치된 제1 클라우드 서버(31)에 의해 대상 서버(40)로부터 웹 문서를 수집할 수 있다. 이처럼, 대상 서버(40)와 지역적으로 가까운 리소스를 이용하여 대상 서버(40)의 웹 문서를 수집함으로써, 웹 문서 수집 시간을 단축시킬 수 있고, 전송 중 잃어버리는 웹 문서의 양을 줄일 수 있다.

[0044] 만약, 제1 클라우드 서버(31)가 대상 서버(40)로부터 거부 메시지를 수신한 경우, 웹 문서 분산 수집 시스템(1)은 제1 지역(A1)에 설치된 다른 클라우드 서버 또는 제1 지역(A1)과 인접한 지역에 설치된 제2 클라우드 서버(32)에 의해 대상 서버(40)로부터 웹 문서를 수집할 수 있다. 따라서, 대상 서버(40)로부터의 웹 문서 수집이 끊임 없이 지속되므로, 신뢰도가 보다 향상된 웹 문서 수집 서비스를 제공할 수 있다.

[0045] 도 2는 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템을 개략적으로 도시한 블록도이다.

[0046] 도 2를 참조하면, 본 발명의 일 실시예에 따른 웹 문서 분산 수집 시스템은 사용자 인터페이스(110), 스케줄러(130), 및 클라우드 서버(150)를 포함한다.

[0047] 사용자 인터페이스(110)는 웹 문서 수집 요청을 수신한다. 관리자는 사용자 인터페이스(110)를 통해 관리 포털(management portal)에 접속하고, 관리 포털 상에서 대상 서버(40)의 형태소를 입력할 수 있다. 이때, 사용자 인터페이스(110)는 대상 서버(40)의 형태소를 포함하는 웹 문서 수집 요청을 수신할 수 있다.

[0048] 웹 문서 수집 요청은 대상 서버(40)의 형태소(morpheme)를 포함할 수 있다. 형태소는 수집 대상을 특정하는 정보로서, 문자 타입(character type), 날짜 타입(data type), 시간 타입(time type), 또는 정수 타입(integer type) 일 수 있다. 이때, 대상 서버(40)의 형태소는 대상 서버(40)의 URL을 포함할 수 있다.

[0049] 대상 서버(40)가 항공사 서버인 경우, 문자 타입의 형태소는 URL, 출발지, 출발 공항 코드, 도착지, 도착 공항 코드, 항공사 코드, 항공기 타입, 항공기 번호, 항공기 등록 번호, 상태 등을 포함할 수 있다. 이때, 날짜 타입의 형태소는 출발일, 도착일 등을 포함할 수 있다. 이때, 시간 타입의 형태소는 출발 예정 시각, 실제 출발 시각, 도착 예정 시각, 실제 도착 시각 등을 포함할 수 있다. 이때, 정수 타입의 형태소는 항공사 프리픽스(prefix) 코드, 연식 등을 포함할 수 있다.

[0050] 사용자 인터페이스(110)는 관리 포털 상에서 에러 메시지를 표시함으로써, 관리자에게 형태소 분석 결과가 유효하지 않음을 알려줄 수 있다.

[0051] 스케줄러(130)는 웹 문서 수집 요청에 대응하여 대상 서버(40)와 관련된 대상 지역을 결정하고, 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성하고, 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 복수의 가상 머신들에 작업 목록을 분배하고, 복수의 가상 머신들로부터 각가 분배된 작업 목록에 기초하여 대상 서버(40)로부터 수집된 웹 문서를 수신한다.

[0052] 스케줄러(130)는 형태소 분석기(131), 가상 머신 매니저(133), 작업 부하 분산기(135), 및 데이터베이스(137)를 포함할 수 있다.

[0053] 형태소 분석기(131)는 사용자 인터페이스(110)로부터 웹 문서 수집 요청을 수신할 수 있다. 형태소 분석기(131)는 웹 문서 수집 요청에 포함된 대상 서버(40)의 형태소를 분석할 수 있다. 한편, 형태소 분석기(131)는 형태소 분석 결과가 유효하지 않은 경우, 사용자 인터페이스(110)에 에러 메시지를 전송할 수 있다.

- [0054] 형태소 분석기(131)는 웹 문서 수집 요청으로부터 대상 서버(40)의 URL을 추출하고, 대상 서버(40)의 URL에 대응하여 대상 지역을 결정할 수 있다. 이때, 형태소 분석기(131)는 대상 서버(40)의 URL을 DNS(domain name server, 50)에 전송하고, DNS(50)로부터 대상 서버(40)가 위치한 지역을 나타내는 정보를 수신할 수 있다.
- [0055] 형태소 분석기(131)는 대상 서버(40)가 위치한 지역 또는 대상 서버(40)가 위치한 지역과 인접한 지역을 대상 지역으로 결정할 수 있다. 이때, 형태소 분석기(131)는 대상 지역을 나타내는 대상 지역 코드를 생성하고, 가상 머신 매니저(133)에 대상 지역 코드를 전달할 수 있다.
- [0056] 형태소 분석기(131)는 형태소 분석 결과가 유효한 경우, 작업 부하 분산기(135)에 형태소 분석 결과를 전달할 수 있다.
- [0057] 가상 머신 매니저(133)는 대상 지역 코드에 기초하여, 대상 지역에 설치된 클라우드 서버(150)를 이용하여 복수의 가상 머신들을 생성할 수 있다. 이때, 복수의 가상 머신들은 각기 다른 IP(internet protocol) 주소를 가질 수 있다.
- [0058] 가상 머신 매니저(133)는 대상 서버(40)로부터 최초로 웹 문서를 수집하는 경우에는 3개의 가상 머신들을 생성하고, 대상 서버(40)로부터 웹 문서를 수집한 적이 있는 경우에는 가장 최근에 생성한 가상 머신의 개수만큼의 가상 머신들을 생성할 수 있다.
- [0059] 예를 들어, 가상 머신 매니저(133)는 네트워크(20)를 통해, 대상 지역에 설치된 클라우드 서버(150)를 이용하여 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3) 및 스토리지(S)를 생성할 수 있다. 이때, 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3)은 각각 제1 가상 머신 상태 정보 내지 제3 가상 머신 상태 정보를 생성하여 가상 머신 매니저(133)에 전송할 수 있다.
- [0060] 한편, 스케줄러(130)는 가상 머신으로부터 종료 메시지를 수신한 경우에는, 복수의 가상 머신들에 작업 목록을 재분배하고, 복수의 가상 머신들로부터 재분배된 작업 목록에 기초하여 대상 서버(40)로부터 수집된 웹 문서를 수신할 수 있다.
- [0061] 한편, 스케줄러(130)는 대상 서버(40)로부터 거부 메시지를 수신한 경우에는, 복수의 가상 머신들로부터 거부 메시지에 포함된 가상 머신을 삭제할 수 있다. 이때, 스케줄러(130)는 대상 지역에 설치된 다른 클라우드 컴퓨팅의 리소스를 이용하여 새로운 가상 머신을 생성하고, 삭제된 가상 머신에 분배된 작업 목록을 새로운 가상 머신에 재분배하고, 새로운 가상 머신으로부터 재분배된 작업 목록에 기초하여 대상 서버(40)로부터 수집된 웹 문서를 수신할 수 있다.
- [0062] 예를 들어, 가상 머신 매니저(133)는 거부 메시지에 기초하여 제1 가상 머신(VM1)을 삭제하고, 제4 가상 머신(VM4)을 생성할 수 있다. 제4 가상 머신(VM4)은 제1 가상 머신(VM1)을 포함하는 클라우드 서버로부터 생성될 수도 있고, 대상 지역에 설치된 클라우드 서버로서 제1 가상 머신(VM1)을 포함하는 클라우드 서버와 다른 클라우드 서버로부터 생성될 수도 있다. 이때, 제4 가상 머신(VM4)은 제1 가상 머신(VM1)의 IP 주소와 다른 IP 주소를 가질 수 있다.
- [0063] 이후에, 가상 머신 매니저(133)는 제1 가상 머신(VM1)과 동일한 IP 주소를 가지는 제5 가상 머신이 생성된 경우에는, 제5 가상 머신을 삭제함으로써, 대상 서버(40)가 클라우드 서버(150)의 접근을 DDOS 공격으로 오인하는 문제를 해결할 수 있다.
- [0064] 가상 머신 매니저(133)는 복수의 가상 머신들을 가리키는 가상 머신 목록을 작업 부하 분산기(135)에 전달할 수 있다. 예를 들어, 가상 머신 매니저(133)는 상태 정보를 전송한 가상 머신을 가리키는 가상 머신 목록을 생성할 수 있다. 즉, 가상 머신 목록은 제1 가상 머신 상태 정보 내지 제3 가상 머신 상태 정보를 각각 전송한 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3)을 가리킬 수 있다.
- [0065] 작업 부하 분산기(135)는 웹 문서 수집 요청에 기초하여 작업 목록을 생성할 수 있다. 작업 목록은 하나 이상의 작업 대상을 가리키는 목록을 의미할 수 있으며, 대상 서버(40)의 형태소를 포함할 수 있다.
- [0066] 작업 부하 분산기(135)는 가상 머신 매니저(133)로부터 수신한 가상 머신 목록에 기초하여, 복수의 가상 머신들에 작업 목록을 분배할 수 있다. 예를 들어, 작업 부하 분산기(135)는 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3)에 작업 목록을 분배할 수 있다.
- [0067] 작업 부하 분산기(135)는 가상 머신 매니저(133)로부터 수신한 가상 머신 목록에 기초하여, 복수의 가상 머신들에 작업 목록을 균등하게 분배할 수 있다.

- [0068] 한편, 작업 부하 분산기(135)는 복수의 가상 머신들에 작업 목록을 주기적으로 분배할 수 있다. 이때, 작업 부하 분산기(135)는 가상 머신이 대상 서버(40)에 접근한 횟수 및 가상 머신이 대상 서버(40)로부터 웹 문서를 수신한 횟수 중 적어도 하나에 기초하여 복수의 가상 머신들에 작업 목록을 불균등하게 분배할 수 있다. 예를 들어, 작업 부하 분산기(135)는 가상 머신이 대상 서버(40)로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수에 기초하여 복수의 가상 머신들에 작업 목록을 재분배할 수 있다. 작업 부하 분산기(135)는 대상 서버(40)로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수가 많은 가상 머신들과 대상 서버(40)로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수가 적은 가상 머신들에 작업 목록을 불균등하게 분배할 수 있다.
- [0069] 작업 부하 분산기(135)는 가상 머신으로부터 수신한 거부 메시지를 가상 머신 매니저(133)에 전달할 수 있다.
- [0070] 가상 머신은 작업 목록에 기초하여 대상 서버(40)로부터 웹 문서를 수집한다. 가상 머신은 작업 목록에 포함된 대상 서버(40)의 URL에 기초하여 대상 서버(40)에 접근할 수 있다. 이때, 가상 머신은 대상 서버(40)의 형태소가 포함된 웹 문서 요청을 대상 서버(40)에 전송할 수 있다.
- [0071] 가상 머신은 대상 서버(40)로부터 웹 문서 요청에 대응하는 웹 문서 응답을 수신할 수 있다. 웹 문서 응답은 대상 서버(40)의 형태소가 포함된 웹 문서를 포함할 수 있다.
- [0072] 가상 머신은 웹 문서 응답에 포함된 웹 문서를 스토리지(S)에 저장할 수 있다. 한편, 가상 머신은 해쉬 함수를 이용하여 웹 문서의 해쉬 값을 생성할 수 있다. 가상 머신은 웹 문서의 해쉬 값을 이용하여 대상 서버(40)로부터 수신한 웹 문서가 스토리지(S)에 저장된 웹 문서와 동일한지 여부를 판단할 수 있다. 이에 따라, 가상 머신은 웹 문서의 해쉬 값에 기초하여 동일한 웹 문서의 반복 수집을 방지할 수 있다.
- [0073] 이때, 가상 머신은 대상 서버(40)의 형태소에 기초하여 웹 문서 응답에 포함된 웹 문서를 파싱(parsing)한 웹 문서를 스토리지(S)에 저장할 수 있다. 이에 따라, 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3)은 분배된 작업 목록에 따라 대상 서버(40)의 웹 문서를 분산 수집할 수 있다. 이처럼, IP 주소가 각기 다른 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3)의 웹 문서 분산 수집은, IP 주소가 하나인 가상 머신의 동일한 작업량의 웹 문서 분산 수집에 비하여, 대상 서버(40)에 의하여 DDOS 공격으로 오인받을 가능성이 현저히 낮아질 수 있다.
- [0074] 한편, 가상 머신은 웹 문서 관련 정보를 생성할 수 있다. 웹 문서 관련 정보는 가상 머신이 대상 서버(40)에 접근한 횟수 및 가상 머신이 대상 서버(40)로부터 웹 문서를 수신한 횟수를 포함할 수 있다. 가상 머신이 대상 서버(40)에 접근한 횟수는, 가상 머신이 대상 서버(40)에 웹 문서 요청을 전송한 횟수일 수 있다. 가상 머신이 대상 서버(40)로부터 웹 문서를 수신한 횟수는, 가상 머신이 대상 서버(40)로부터 웹 문서 응답을 수신한 횟수일 수 있다. 가상 머신이 대상 서버(40)로부터 웹 문서를 수신한 횟수는 가상 머신이 대상 서버(40)로부터 각기 다른 해쉬 값을 가지는 웹 문서를 수신한 횟수일 수 있다.
- [0075] 가상 머신은 웹 문서 수집이 종료되면 작업 부하 분산기(135)에 종료 메시지를 전송할 수 있다. 가상 머신은 작업 목록에 기초하여 웹 문서 수집이 종료되었는지 여부를 판단할 수 있다.
- [0076] 가상 머신은 웹 문서 및 웹 문서 관련 정보를 스케줄러(130)에 전송할 수 있다. 이때, 스케줄러(130)가 가상 머신으로부터 수신하는, 대상 서버(40)로부터 수집된 웹 문서는, 가상 머신이 형태소에 기초하여 파싱한 웹 문서일 수 있다. 스케줄러(130)의 데이터베이스(137)는 웹 문서 및 웹 문서 관련 정보를 저장할 수 있다.
- [0077] 가상 머신은 대상 서버(40)로부터 웹 문서 요청에 대응하는 거부 메시지를 수신할 수 있다. 가상 머신은 대상 서버(40)로부터 수신한 거부 메시지를 작업 부하 분산기(135)에 전송할 수 있다.
- [0078] 도 3은 본 발명의 일 실시예에 따른 웹 문서 분산 수집 방법을 설명하기 위한 흐름도이다.
- [0079] 도 3을 참조하면, 사용자 인터페이스(110)는 웹 문서 수집 요청을 수신한다(S10). 웹 문서 수집 요청은 문자 타입, 날짜 타입, 시간 타입, 또는 정수 타입의 형태소를 포함할 수 있다. 웹 문서 수집 요청은 대상 서버(40)의 URL을 포함할 수 있다.
- [0080] 이어서, 스케줄러(130)의 형태소 분석기(131)는 웹 문서 수집 요청에 대응하여 대상 서버(40)와 관련된 대상 지역을 결정한다(S30). 스케줄러(130)는 대상 서버(40)의 URL에 대응하여 대상 지역을 결정할 수 있다. 이때, 대상 지역은 대상 서버(40)가 위치한 지역 또는 대상 서버(40)가 위치한 지역과 인접한 지역일 수 있다.
- [0081] 이어서, 스케줄러(130)의 가상 머신 매니저(133)는 대상 지역에 설치된 클라우드 컴퓨팅의 리소스를 이용하여 복수의 가상 머신들을 생성한다(S50). 스케줄러(130)는 대상 서버(40)로부터 최초로 웹 문서를 수집하는 경우에는 3개의 가상 머신들을 생성하고, 대상 서버(40)로부터 웹 문서를 수집한 적이 있는 경우에는 가장 최근에 생

성한 가상 머신의 개수만큼의 가상 머신들을 생성할 수 있다.

- [0082] 이어서, 스케줄러(130)의 작업 부하 분산기(135)는 웹 문서 수집 요청에 기초하여 작업 목록을 생성하고, 복수의 가상 머신들에 작업 목록을 분배한다(S70). 이때, 스케줄러(130)는 복수의 가상 머신들에 작업 목록을 균등하게 분배할 수 있다. 한편, 스케줄러(130)는 복수의 가상 머신들에 작업 목록을 주기적으로 분배할 수 있다.
- [0083] 스케줄러(130)의 데이터베이스(137)는 복수의 가상 머신들로부터 각각 분배된 작업 목록에 기초하여 대상 서버(40)로부터 수집된 웹 문서를 수신한다(S90).
- [0084] 도 4는 본 발명의 일 실시예에 따른 웹 문서 분산 수집 방법을 상세하게 설명하기 위한 흐름도이다.
- [0085] 도 4를 참조하면, 스케줄러(130)는 제1 가상 머신(VM1)을 생성하면(S101), 제1 가상 머신(VM1)으로부터 제1 가상 머신 상태 정보를 수신한다(S103).
- [0086] 도시하지 않았으나, 스케줄러(130)는 제2 가상 머신(VM2)을 생성하면 제2 가상 머신(VM2)으로부터 제2 가상 머신 상태 정보를 수신하고, 제3 가상 머신(VM3)을 생성하면 제3 가상 머신(VM3)으로부터 제3 가상 머신 상태 정보를 수신할 수 있다.
- [0087] 스케줄러(130)는 제1 가상 머신(VM1) 내지 제3 가상 머신(VM3)에 작업 목록을 분배할 수 있다. 그 결과, 스케줄러(130)는 제1 가상 머신(VM1)에 제1 작업 목록을 전송한다(S105). 도시하지 않았으나, 스케줄러(130)는 제2 가상 머신(VM2) 및 제3 가상 머신(VM3)에 각각 제2 작업 목록 및 제2 작업 목록을 전송할 수 있다. 스케줄러(130)가 대상 서버(40)로부터 최초로 웹 문서를 수집하는 경우에는, 제1 작업 목록 내지 제3 작업 목록 각각의 작업량이 균등할 수 있다.
- [0088] 제1 가상 머신(VM1)은 제1 작업 목록에 기초하여 대상 서버(40)에 웹 문서 요청을 전송하고(S107), 대상 서버(40)로부터 웹 문서 요청에 대응하는 웹 문서 응답을 수신한다(S109).
- [0089] 웹 문서 응답이 거부 메시지인 경우, 제1 가상 머신(VM1)은 스케줄러(130)에 거부 메시지를 전송한다(S121).
- [0090] 스케줄러(130)는 제1 가상 머신(VM1)으로부터 거부 메시지를 수신한 경우에는, 제1 가상 머신(VM1)을 삭제한다(S123). 즉, 스케줄러(130)는 대상 서버(40)로부터 거부 메시지를 수신한 경우에는 복수의 가상 머신들로부터 거부 메시지에 포함된 가상 머신을 삭제할 수 있다.
- [0091] 이어서, 스케줄러(130)는 제4 가상 머신(VM4)을 생성한다(S125). 즉, 스케줄러(130)는 대상 지역에 설치된 다른 클라우드 컴퓨팅의 리소스를 이용하여 새로운 가상 머신을 생성할 수 있다.
- [0092] 이어서, 스케줄러(130)는 제4 가상 머신(VM4)으로부터 제4 가상 머신 상태 정보를 수신하고(S127), 제4 가상 머신(VM4)에 제4 작업 목록을 전송한다(S129). 이때, 스케줄러(130)는 삭제된 가상 머신에 분배된 작업 목록을 새로운 가상 머신에 재분배할 수 있다. 즉, 스케줄러(130)는 삭제된 제1 가상 머신(VM1)의 제1 작업 목록을 제4 작업 목록으로 재분배할 수 있다. 도시하지 않았으나, 제4 가상 머신(VM4)은 재분배된 작업 목록에 기초하여 대상 서버(40)로부터 웹 문서를 수집할 수 있다.
- [0093] 한편, 웹 문서 응답에 웹 문서가 포함된 경우, 제1 가상 머신(VM1)은 웹 문서 관련 정보를 생성하여 저장한다(S113).
- [0094] 제1 가상 머신(VM1)은 작업 목록에 포함된 작업들을 수행하기 위하여 대상 서버(40)에 복수 회의 웹 문서 요청을 전송할 수 있고, 대상 서버(40)로부터 복수 회의 웹 문서 응답을 수신할 수 있다. 제1 가상 머신(VM1)은 웹 문서 요청을 전송 횟수 및 웹 문서 응답을 수신한 횟수를 카운트한 결과에 기초하여 웹 문서 관련 정보를 생성할 수 있다. 이때, 웹 문서 관련 정보는 제1 가상 머신(VM1)이 대상 서버(40)에 접근한 횟수 및 대상 서버(40)로부터 웹 문서를 수신한 횟수를 포함할 수 있다.
- [0095] 이어서, 제1 가상 머신(VM1)은 스케줄러(130)에 웹 문서 관련 정보를 전송한다(S115). 스케줄러(130)는 웹 문서 관련 정보에 기초하여 제1 가상 머신(VM1)의 삭제 여부 또는 제1 가상 머신(VM1)에 분배하기 위한 제1 작업 목록의 재생성 여부를 판단할 수 있다.
- [0096] 한편, 제1 가상 머신(VM1)은 웹 문서 응답에 포함된 웹 문서를 파싱하여 스토리지(S)에 저장한다(S117).
- [0097] 제1 가상 머신(VM1)은 수집된 웹 문서를 스케줄러(130)에 전송할 수 있다(S119). 이때, 스케줄러(130)가 수신하는 대상 서버(40)로부터 수집된 웹 문서는, 제1 가상 머신(VM1)이 대상 서버(40)의 형태소에 기초하여 파싱한 웹 문서일 수 있다.

[0098] 이제까지 본 발명에 대하여 바람직한 실시예를 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 본 발명을 구현할 수 있음을 이해할 것이다.

[0099] 그러므로 상기 개시된 실시예는 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 특허청구범위에 의해 청구된 발명 및 청구된 발명과 균등한 발명들은 본 발명에 포함된 것으로 해석되어야 한다.

부호의 설명

[0100] 1: 웹 문서 분산 수집 시스템

10: 사용자 단말

20: 네트워크

31: 제1 클라우드 서버

32: 제2 클라우드 서버

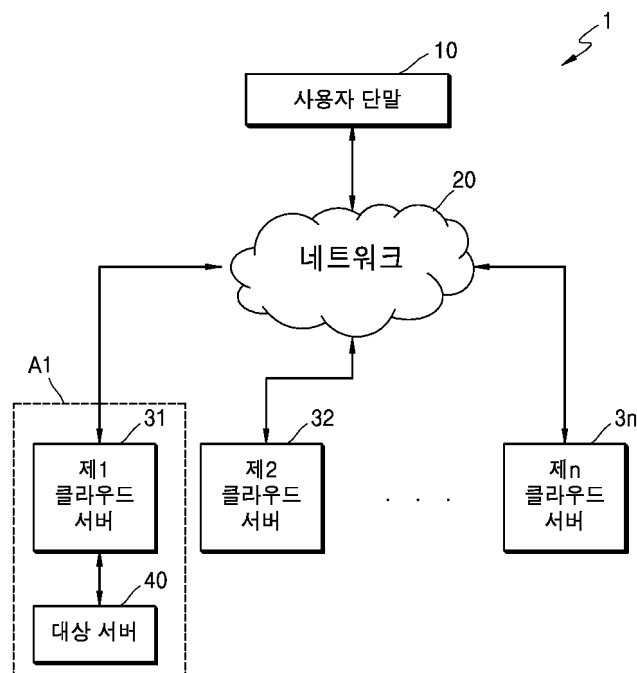
3n: 제n 클라우드 서버

40: 대상 서버

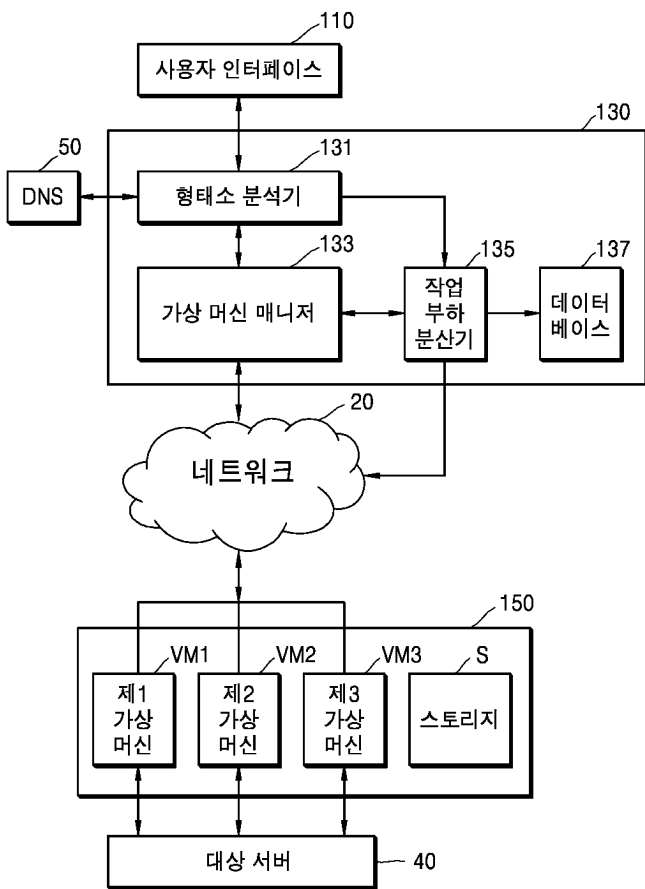
A1: 제1 지역

도면

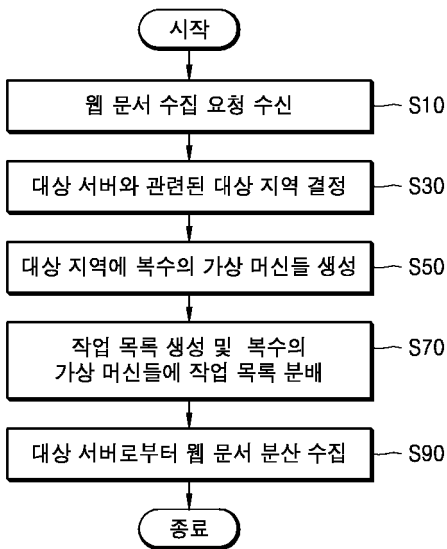
도면1



도면2



도면3



도면4

