



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년11월21일
(11) 등록번호 10-1327799
(24) 등록일자 2013년11월05일

(51) 국제특허분류(Int. Cl.)
H04L 12/70 (2013.01) H04L 29/06 (2006.01)
(21) 출원번호 10-2012-0069068
(22) 출원일자 2012년06월27일
심사청구일자 2012년06월27일
(56) 선행기술조사문헌
Wonchul Kang 외 2인, 'NetFlow Analysis with MapReduce', 충남대학교, 2010.04.24.

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
이연희
대전광역시 유성구 노은로 353 송림마을아파트
304동 1701호
이영석
대전광역시 유성구 테크노중앙로 32 금성백조아파트
809-602 (관평동)
(74) 대리인
김원준

전체 청구항 수 : 총 2 항

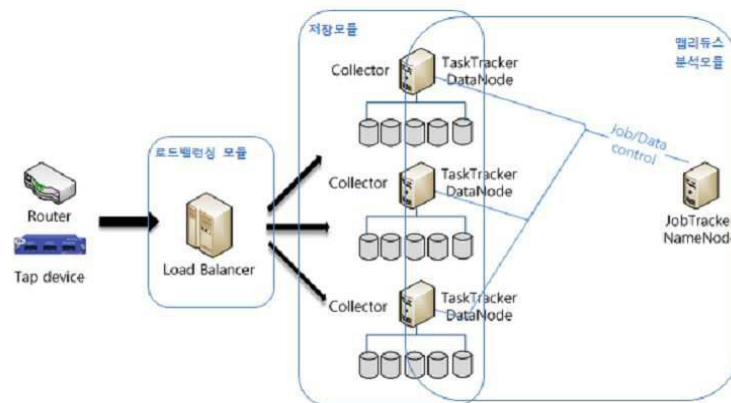
심사관 : 하은주

(54) 발명의 명칭 맵리듀스 방식을 이용한 클러스터 기반의 웹 트래픽 분석 방법 및 시스템

(57) 요약

본 발명은 분산 시스템을 이용하여 네트워크 링크에서 수집되는 대량의 패킷 트레이스로부터 웹 데이터의 분석을 효율적으로 병렬 처리할 수 있도록 고안한 맵리듀스를 기반으로 한 웹 트래픽의 분석 방법 및 시스템에 관한 것으로, 보다 상세하게는 하둡 프레임워크를 기반으로 하며, 패킷 트레이스를 네트워크 링크 또는 라우터로부터 획득하고 Http 패킷을 선별하여 각 클러스터 노드에 세션 별로 분산하여 전달하는 로드밸런싱 모듈; 상기 각 클러스터 노드에 전달된 Http 패킷을 수집하여 HDFS(하둡 분산 파일시스템)에 저장하는 저장 모듈; 및 상기 HDFS에 저장된 Http 패킷으로부터 Http 로그를 추출하여 생성하고 웹 트래픽을 분석하는 맵리듀스 모듈;을 포함하여 구성되는 것을 특징을 하는 웹 트래픽 분석 시스템과 이를 이용한 분석 방법에 관한 것이다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업	
과제고유번호	08913-05002
부처명	방송통신위원회
연구사업명	방송통신 기술개발 사업
연구과제명	초정밀 측정 및 분석 기술 연구
기 여 율	1/2
주관기관	KAIST
연구기간	2012.03.01 ~ 2013.02.28
이 발명을 지원한 국가연구개발사업	
과제고유번호	C1810-1105-0010
부처명	정보통신산업진흥원
연구사업명	대기업 연계 IT/SW창의연구과제 지원사업
연구과제명	규모확장성을 지원하는 분산 DPI 기술개발
기 여 율	1/2
주관기관	엔에이치엔(주)
연구기간	2011.11.01 ~ 2012.06.30

특허청구의 범위

청구항 1

삭제

청구항 2

(a) 네트워크 링크 또는 라우터로부터 패킷 트레이스를 획득하고 Http 헤더를 포함하는 패킷만을 선별하여 동일 세션의 Http 패킷은 동일 클러스터 노드에 전달되도록 각 클러스터 노드에 분산하여 전달하는 웹 트래픽 로드 분산 단계;

(b) 상기 전달된 Http 패킷을 각 클러스터 노드가 수집하여 저장하는 웹 트래픽 분산 수집·저장 단계;

(c) ① 저장된 Http 패킷을 읽어서 세션별로 Http 패킷을 조립하고, Http 헤더 필드를 추출하여 키를 생성하고 출력하는 맵 단계와 ② 맵의 결과로 생성된 키에 따라 Http 헤더의 필드를 분석하여 저장하거나 단순 저장하는 단계로 이루어진 웹 트래픽 로깅 단계; 및

(d) 저장된 Http 로그를 맵리듀스를 이용하여 분석하는 웹 트래픽 로그 분석 단계;

를 포함하는 것을 특징으로 하는 웹 트래픽의 분석 방법.

청구항 3

삭제

청구항 4

삭제

청구항 5

제 2 항에 있어서,

클라이언트로부터 서버 방향으로의 Http 패킷 또는 서버로부터 클라이언트 방향으로의 Http 패킷만을 선별하여 각 클러스터 노드에 분산하여 전달하는 것을 특징으로 하는 웹 트래픽의 분석 방법.

청구항 6

삭제

명세서

기술분야

[0001] 본 발명은 분산 시스템을 이용하여 네트워크 링크에서 수집되는 대량의 패킷 트레이스로부터 웹 데이터의 분석을 효율적으로 병렬 처리할 수 있도록 고안한 맵리듀스를 기반으로 한 웹 트래픽의 분석 방법 및 시스템에 관한 것이다.

배경기술

[0002] 고속의 링크 장비와 사용자 지향의 다양한 웹 어플리케이션의 등장으로 인터넷을 이용하는 사용자는 폭발적으로 증가하고 있고, 동영상이나 음성서비스와 같은 다량의 트래픽을 유발하는 서비스들이 성행하고 있다. 이에 따라 네트워크 링크에 유통되는 트래픽의 양의 대량화 되고 있다. 웹 트래픽(Web traffic)은 특정 웹사이트에 방문하는 사람들이 주고받는 데이터의 양으로, 인터넷 트래픽의 큰 부분을 차지한다.

[0003] 웹 트래픽의 분석은 통상 방문자 수, 방문자 당 페이지 뷰 평균 수, 평균 방문 시간, 도메인 클래스, 트래픽이

많은 시간대, 요청을 가장 많이 받은 페이지 등을 감시하는 것에 의해 이루어진다. 이와 같은 웹 트래픽의 분석은 얼마나 많은 사용자들이 방문했는지 그리고 얼마나 많은 웹 페이지를 방문했는지에 대한 정보뿐만 아니라 어떤 호스트에 웹 페이지의 요청이 집중되고 웹을 사용하는 사용자의 행위는 어떤 특징을 가지는지 등의 웹 서버와 웹의 사용자에 대한 특징을 제공한다. 웹 트래픽의 분석을 통해 웹의 운영자는 보안문제나 대역폭의 부족과 같은 현상을 유추하여 이를 통해 사이트의 구조를 설계하거나 개선하는데 많은 도움을 받을 수 있고, 비즈니스 사업자의 관점에서는 광고수익, 과금 등에 활용할 수 있다.

[0004] 이와 같이 웹 트래픽은 사이트의 성격에 관계없이 비즈니스에서 가장 관심 있는 분석 데이터이다. 웹 트래픽을 분석하는 방법으로 기존에는 웹 로그파일에서 웹에 대한 접근 정보를 추출하였으나, 근래에는 자바스크립트를 웹 페이지에 임베딩하여 페이지의 태그를 수집하는 방식을 채택하여 웹 트래픽 정보를 수집하고 분석하는 추세이다.

[0005] 패킷은 웹 서버와 클라이언트 사이에 주고받는 모든 정보를 포함하기 때문에 웹에 대한 사용자의 접근 정보를 포함한 많은 정보를 포함한다. 기존의 네트워크 트래픽 분석 시스템들은 대용량 트래픽 처리를 고려하지 않고 설계되어 있기 때문에 네트워크 트래픽 분석을 위해 고안된 시스템을 이용하여 대량으로 발생하는 패킷을 수집해서 저장하고 정확하게 웹 트래픽을 분석하는 일은 쉽지 않으며, 고 성능의 처리 서버를 구축해야하는 요구가 있다.

[0006] 맵리듀스(MapReduce)는 대용량의 데이터를 안정적으로 처리하기 위한 프로그래밍 모델로서 구글(Google)에 의해 제안되었다. 이 방식을 이용하면 저가의 컴퓨팅 장비를 이용한 분산 프로그래밍이 가능하며, 이러한 맵리듀스를 오픈소스로 구현한 것이 바로 하둡(Hadoop)이다.

[0007] 맵리듀스는 맵과 리듀스 단계의 처리과정을 통해 대량의 데이터를 처리하는 방식으로 각 단계는 입력과 출력으로 키와 값의 쌍을 가진다. 맵 단계에서는 주로 대량의 데이터로부터 유입되는 레코드 형태의 Key-Value의 쌍으로부터 처리하려고 하는 Key에 해당하는 값과 Value에 해당하는 값을 추출하는 일을 수행하고, 리듀스 단계에서는 맵 단계에서 추출한 Key에 그룹핑되어 전달되는 일련의 Value들을 순회하며 원하는 계산을 수행하여 결과를 얻어내는 일을 수행한다. 맵리듀스 프로그래밍에서 이 두 함수, 즉 맵과 리듀스만 정의하면 분산된 컴퓨팅 자원을 이용하여 대용량의 데이터를 분산 처리할 수 있다.

[0008] 맵리듀스를 이용한 대용량 데이터의 처리는 다양한 분야에서 개발되어 사용되어지나, 이를 이용한 패킷 트레이스로부터의 웹 트래픽을 분석하는 방법과 시스템에 대해서는 아직 개발된 바 없다.

발명의 내용

해결하려는 과제

[0009] 상기와 같은 종래기술의 문제점을 해소하기 위한 본 발명의 목적은 분산 시스템을 이용하여 네트워크 링크에서 수집되는 대량의 패킷 트레이스로부터 웹 데이터의 분석을 효율적으로 병렬 처리할 수 있는 웹 트래픽의 분석 방법 및 시스템을 제공하는 것이다.

과제의 해결 수단

[0010] 전술한 목적을 달성하기 위한 본 발명은 하둡 프레임워크를 기반으로 한 웹 트래픽 분석 시스템에 관한 것으로, 보다 상세하게는 하둡 프레임워크를 기반으로 하며, 패킷 트레이스를 네트워크 링크 또는 라우터로부터 획득하고 Http 패킷을 선별하여 각 클러스터 노드에 세션 별로 분산하여 전달하는 로드밸런싱 모듈; 상기 각 클러스터 노드에 전달된 Http 패킷을 수집하여 HDFS(하둡 분산 파일시스템)에 저장하는 저장 모듈; 및 상기 HDFS에 저장된 Http 패킷으로부터 Http 로그를 추출하여 생성하고 웹 트래픽을 분석하는 맵리듀스 모듈;을 포함하여 구성되는 것을 특징을 하는 웹 트래픽 분석 시스템에 관한 것이다.

[0011] 본 발명은 또한 (a) 네트워크 링크 또는 라우터로부터 패킷 트레이스를 획득하고 Http 패킷을 선별하여 각 클러스터 노드에 분산하여 전달하는 웹 트래픽 로드 분산 단계; (b) 상기 전달된 Http 패킷을 각 클러스터 노드가 수집하여 저장하는 웹 트래픽 분산 수집·저장 단계; (c) 상기 저장된 Http 패킷으로부터 맵리듀스 방식으로 Http 로그를 추출하여 저장하는 웹 트래픽 로깅 단계; 및 (d) 저장된 Http 로그를 맵리듀스를 이용하여 분석하

는 웹 트래픽 로그 분석 단계;를 포함하는 것을 특징으로 하는 웹 트래픽의 분석 방법에 관한 것이다.

발명의 효과

[0012] 이상과 같이 본 발명에 의하면, 인터넷 트래픽 데이터 중 패킷 트레이스를 이용하여 웹 트래픽을 여러 대의 컴퓨팅 노드에 분산하여 측정, 분석할 수 있다. 특히 패킷 트레이스로부터 로그 데이터를 추출하여 분산하는 과정을 여러 대의 컴퓨터에 분산하여 병렬로 처리하도록 하여 별도의 웹 로깅 시스템이 없는 경우에도 빠른 분석 성능을 보장할 수 있으며 고가의 서버 장비를 필요로 하지 않기 때문에 비용 절감효과를 가져올 수 있다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일 실시예에 의한 웹 트래픽 분석 시스템을 보여주는 구성도.

도 2는 본 발명의 웹 트래픽의 분석 방법에 대한 흐름도.

도 3은 웹 트래픽 로깅 단계에서 웹 트래픽 로그 추출의 흐름도.

도 4는 맵리듀스 기반의 웹 트래픽 분석 모듈의 구성의 일 실시예와 이에 따른 데이터의 흐름도.

발명을 실시하기 위한 구체적인 내용

[0014] 이하 첨부된 도면을 참조하여 본 발명을 보다 상세히 설명한다. 그러나 첨부된 도면은 본 발명의 기술적 사상의 내용과 범위를 쉽게 설명하기 위한 예시일 뿐, 이에 의해 본 발명의 기술적 범위가 한정되거나 변경되는 것은 아니다. 또한 이러한 예시에 기초하여 본 발명의 기술적 사상의 범위 안에서 다양한 변형과 변경이 가능함은 당업자에게는 당연할 것이다.

[0015] 본 발명은 분산 시스템을 이용하여 네트워크 링크 또는 라우터에서 수집되는 대량의 패킷 트레이스로부터 웹 데이터를 효율적으로 병렬 처리할 수 있도록 고안한 웹 트래픽의 분석 방법 및 시스템에 관한 것으로, 도 1은 본 발명의 일 실시예에 의한 웹 트래픽 분석 시스템을 보여주는 구성도이다. 도 1을 참조하면, 본 발명의 웹 트래픽 분석 시스템은 하둡 프레임워크를 기반으로 하여, 패킷 트레이스를 네트워크 링크 또는 라우터로부터 획득하고 Http 패킷을 선별하여 각 클러스터 노드에 세션 별로 분산하여 전달하는 로드밸런싱 모듈; 상기 각 클러스터 노드에 전달된 Http 패킷을 수집하여 HDFS(하둡 분산 파일시스템)에 저장하는 저장 모듈; 및 상기 HDFS에 저장된 Http 패킷으로부터 Http 로그를 추출하여 생성하고 웹 트래픽을 분석하는 맵리듀스 모듈;을 포함하여 구성된다.

[0016] 본 발명의 분석 시스템에서 상기 로드밸런싱 모듈, 저장 모듈 및 맵리듀스 모듈은 개별의 모듈로 설명하였으나, 이는 클러스터 환경에서 각 노드가 담당하는 기능을 기준으로 분류한 것으로 실질적으로는 하나의 클러스터 노드가 저장 모듈 및 맵리듀스 모듈의 기능을 동시에 수행할 수 있다. 예를 들면, 하나의 클러스터 노드가 저장 모듈과 맵리듀스 모듈로 동시에 작용할 수도 있으며, 저장 모듈 또는 맵리듀스 모듈로서만의 기능을 수행할 수도 있다.

[0017] 본 발명은 또한 하둡 프레임워크를 기반으로 한 웹 트래픽의 분석 방법에 관한 것으로, 도 2는 본 발명의 웹 트래픽의 분석 방법에 대한 흐름도를 도시한 것이다. 도 2에 도시한 바와 같이, 본 발명의 웹 트래픽 분석 방법은 (a) 네트워크 링크 또는 라우터로부터 패킷 트레이스를 획득하고 Http 패킷을 선별하여 각 클러스터 노드에 분산하여 전달하는 웹 트래픽 로드 분산 단계(100); (b) 상기 전달된 Http 패킷을 각 클러스터 노드가 수집하여 저장하는 웹 트래픽 분산 수집·저장 단계(200); (c) 상기 저장된 Http 패킷으로부터 맵리듀스 방식으로 Http 로그를 추출하여 저장하는 웹 트래픽 로깅 단계(300); 및 (d) 저장된 Http 로그를 맵리듀스를 이용하여 분석하는 웹 트래픽 로그 분석 단계(400);를 포함한다.

[0018] 상기 웹 트래픽 로드 분산(100) 단계에서는 libpcap이나 PF_RING, PF_Q와 같은 패킷 수집 라이브러리를 이용하

여 네트워크 링크 또는 라우터로부터 패킷트레이스를 획득한 후, 이 중 Http 프로토콜을 사용하는 웹 트래픽에 해당하는 패킷만을 선별하여 각 클러스터 노드에 분산하여 전달한다. 이때, 이후 (c) 단계에서의 Http 로그 추출의 편의를 위하여 각 Http 패킷의 세션을 식별하여 동일한 세션에 속하는 Http 패킷은 동일한 클러스터 노드(데이터 노드)에 전달하도록 처리하는 것이 바람직하다. 이때 세션 별로 각 노드에 Http 패킷을 전달하기 전에, 같은 노드로 할당되는 다른 세션의 Http 패킷과의 구분을 위해서 같은 세션에 속하는 Http 패킷을 그룹핑하여 전달하는 것이 보다 바람직하다. 세션 식별을 위해서는 통상 하나의 플로우라고 인정할 수 있는 5tuple 값(출발지 IP, 출발지 port, 도착지 IP, 도착지 port, 프로토콜 종류)과 SYN 및 FIN 플래그를 세션키로 사용하여 같은 5tuple 키를 가지는 플로우에 대해서 TCP 세션의 시작과 종료를 감지하여 세션을 분류할 수 있도록 구성될 수 있다.

[0019] 또한 (c) 웹 트래픽 로깅 단계의 성능을 향상하기 위하여 상기 웹 트래픽 로드 분산 단계에서는 Http 헤더를 포함하지 않는 패킷은 폐기하고, Http 헤더를 포함하는 패킷만을 선별하여 각 클러스터 노드에 분산하여 전달할 수 있다. Http 헤더를 포함하는 패킷의 선별은 하나의 세션에 속하는 Http 패킷들 중 Http 헤더필드가 시작되는 패킷으로부터 통상 Http 헤더 필드를 포함한다고 보여지는 수 만큼의 패킷만을 선별하여 전달하는 것에 의해 달성될 수 있다. 여기서 Http 헤더 필드가 시작되는 패킷의 식별은 Http 메소드, 즉 'GET', 'POST', 'HEAD', 'PUT' 등의 문자열을 매칭하여 봄으로써 해결할 수 있다.

[0020] 웹 트래픽의 분석 특성에 따라서는 보다 효율적인 분석을 위해서 웹 트래픽 로드 분산 단계에서 클라이언트로부터 서버 방향으로의 Http 패킷 또는 서버로부터 클라이언트 방향으로의 Http 패킷만을 선별하여 각 클러스터 노드에 분산하여 전달하도록 할 수도 있다.

[0021] 상기 웹 트래픽 분산 수집·저장 단계에서는 웹 트래픽 로드 분산 단계에서 각 클러스터 노드로 전달된 Http 패킷을 ibpcap이나 PF_RING, PF_Q와 같은 패킷 수집 라이브러리를 이용하여 획득한다. 이때, 상기 웹 트래픽 로드 분산 단계에서 Http 패킷이 세션별로 그룹핑 되지 않은 경우에는 본 단계에서 Http 패킷을 세션별로 그룹핑하여 저장할 수 있다. 하둡 프레임워크를 기반으로 하는 경우 상기 Http 패킷은 HDFS에 저장된다.

[0022] 상기 웹 트래픽 로깅 단계는 저장된 Http 패킷으로부터 맵리듀스 방식으로 Http 로그를 추출하여 저장하는 단계이다. 보다 상세하게는 (A) 저장된 Http 패킷을 읽어서 세션별로 Http 패킷을 조립하고, Http 헤더 필드를 추출하여 키를 생성하고 출력하는 맵 단계; (B) 맵의 결과로 생성된 키에 따라 Http 헤더의 필드를 분석하여 저장하거나 단순 저장하는 단계;에 의해 수행된다. 웹 트래픽 로드 분산 단계에서 Http 패킷이 세션별로 선별되고 더 나아가 그룹핑되어있는 경우에는 상기 (A) 단계의 Http 패킷의 조립이 더욱 효율적으로 이루어 질 수 있다.

[0023] 도 3은 웹 트래픽 로깅 단계에서 Http 로그를 생성하기 위한 맵과 리듀스 단계를 보여주는 흐름도이다. 맵 단계(301)에서는 저장된 Http 패킷 트레이스를 읽어 세션을 식별하고 하나의 세션에 포함된 n개의 패킷의 TCP 페이로드를 조립하여 Http 헤더 필드들을 추출한다. 이 때 Key로 생성되는 값은 분석하고자 하는 항목에 따라 설정이 가능하다. Value는 Http 헤더의 필드값을 사용할 수 있다. 예를 들어 클라이언트로부터 웹서버로 향하는 세션의 경우 이 때 생성되는 Key 값을 출발지 IP와 timestamp으로 생성하게 되면 리듀스 단계를 거쳐 출발지 IP를 기준으로 그룹핑되고 다시 그룹 내에서 timestamp로 정렬이 되므로 특정사용자에 대한 웹 페이지의 접근 순서에 따른 Http 헤더 필드 값들을 얻을 수 있다. 리듀스 단계(302)에서는 맵리듀스의 특성에 따라 Key 값에 의해 Http 로그가 정렬이되고 결과는 HDFS에 저장(303)되어 Http의 로그 생성이 완성된다.

[0024] 상기 웹 트래픽 로그 분석 단계는 웹 트래픽 로깅 단계(300)에서 생성한 Http 로그 정보를 읽어 맵리듀스를 이용하여 웹 트래픽을 분석하는 단계이다. 여기에서의 분석은 분석하고자 하는 항목에 따라 다양하게 구현되어 수행될 수 있다.

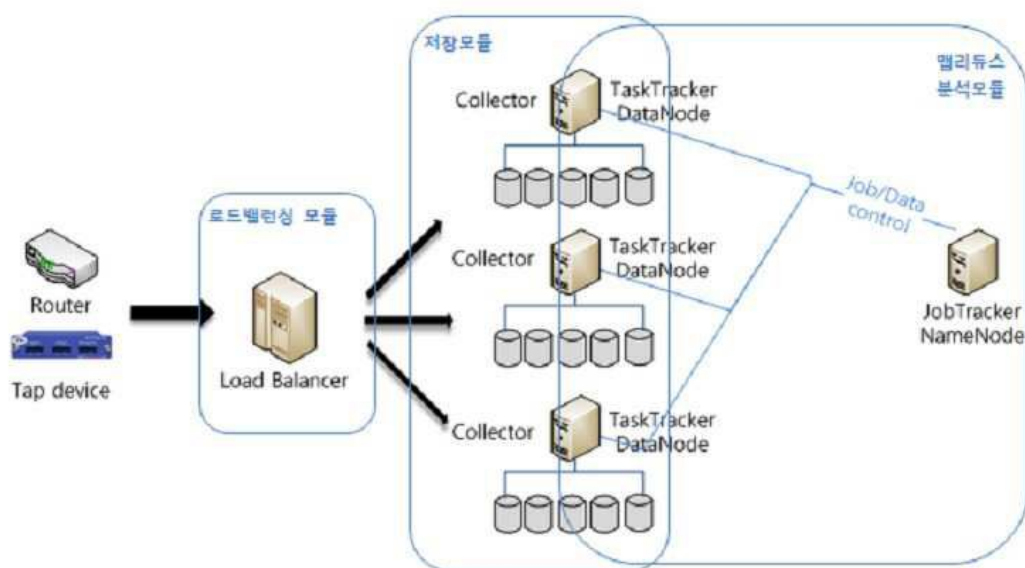
[0025] 도 4는 맵리듀스 기반의 웹 트래픽 분석 모듈의 구성의 일실시예와 이에 따른 데이터의 흐름을 보여준다. 웹 트래픽 로깅 단계(300)에서 Http 로깅에 대한 맵리듀스 잡을 거쳐 Http 패킷 레코드로부터 Http 로그가 생성이 되면, 이 로그를 읽어 URL별 user 수와 view 수를 계산하는 맵리듀스 Job과 사용자별 접속 URL을 구하는 맵리듀스 Job이 실행이 되고, 추가적인 분석을 위해 Hive와 같은 데이터웨어하우징 솔루션을 이용하기위한 데이터의 로딩이 수행된다.

[0026] 보다 구체적으로 URL별 user 수와 view 수를 계산하는 맵리듀스 Job의 경우를 예로 들면, 웹 트래픽 로깅 단계(300)에서 추출한 Http 로그를 읽어 맵 단계에서 URL을 주 정렬키로, 출발지 IP를 보조 정렬 키로하고 URL을 처리할 리듀서를 할당하기 위한 파티션 함수의 해쉬값 항목으로 사용한다. 리듀스 단계에서는 주 정렬키와 보조 정렬 키 값에 의해 정렬된 레코드를 읽으면서 더하여 주 정렬키인 URL에 대한 View 수를 구하고, 이전 레코드와의 출발지 IP값이 상이한지 여부를 판별하여 URL에 대한 유일한 user의 수를 구하여 결과를 얻는 것이 가능하다.

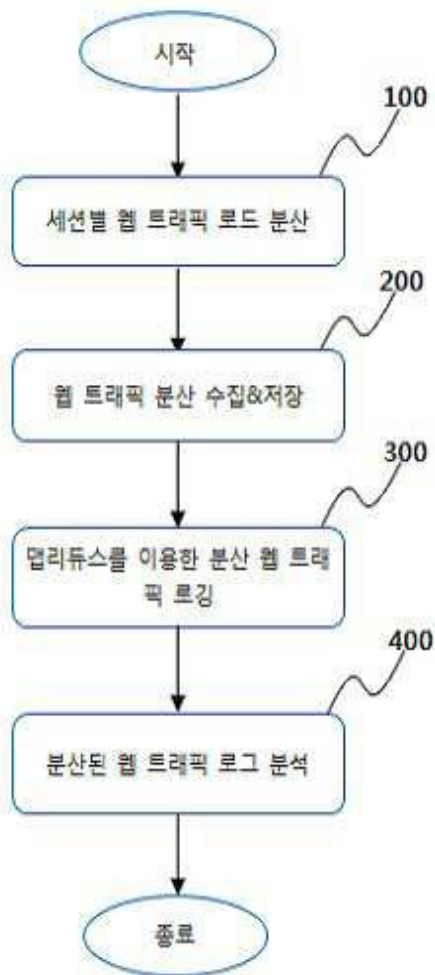
[0027] 사용자별 접속 URL의 경우 웹 트래픽 로깅 단계(300)에서 추출한 Http 로그를 읽어 맵 단계에서 출발지 IP를 주 정렬 키로, 타임스탬프 값을 보조 정렬 키로하고 주 정렬 키인 출발지 IP를 처리할 리듀서를 할당하기 위한 파티션 함수의 해쉬값 항목으로 사용한다. 리듀스 단계에서는 주 정렬 키와 보조 정렬 키 값에 의해 정렬된 레코드를 읽어서 출발지 IP에 대한 접속 URL을 리스트 형태로 출력하는 것이 가능하다.

도면

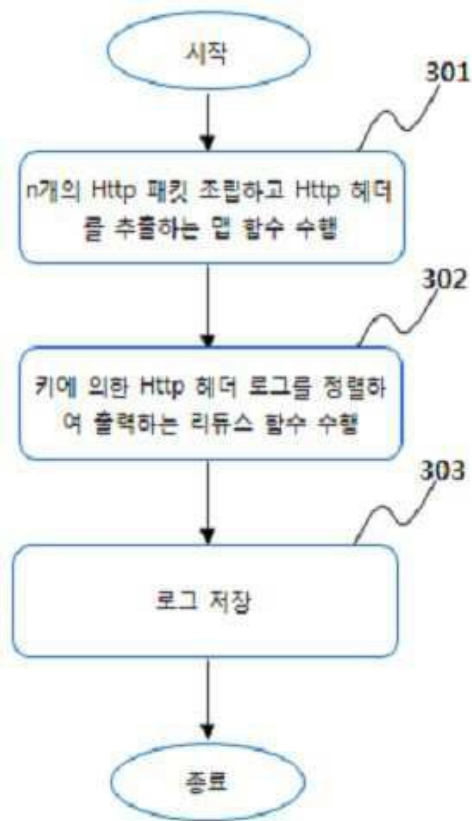
도면1



도면2



도면3



도면4

