(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0049111
(43) 공개일자 2013년05월13일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 17/40* (2006.01)
(21) 출원번호 10-2011-0114168
(22) 출원일자 2011년11월03일
심사청구일자 없음

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
이주영
대전광역시 유성구 가정로 63, 103동 1407호 (신성동, 럭키하나아파트)
길연희
대전광역시 유성구 관평동 대덕 테크노밸리아파트 909동 1502호
(뒷면에 계속)
(74) 대리인
특허법인우인

전체 청구항 수 : 총 16 항

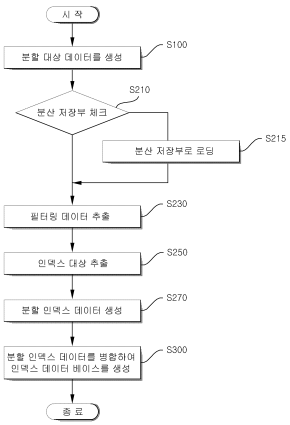
(54) 발명의 명칭 **분산 처리를 이용한 포렌식 인덱스 방법 및 장치**

(57) 요 약

본 발명은 분산 처리를 이용한 포렌식 인덱스 방법에 있어서, 인덱싱 대상 데이터를 분산처리를 위해 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성하는 단계, 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 복수의 데이터 처리부에 할당하고, 복수의 데이터 처리부에서 할당된 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 단계 및 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 단계를 포함하는 분산 처리를 이용한 포렌식 인덱스 방법을 제공한다.

본 발명에 의하면, 인덱싱 시스템의 확장성(scalability)을 높일 수 있고, 이를 통해 포렌식 분석에서의 요구사항 중 가장 큰 이슈인 처리 속도를 크게 향상 시키는 효과가 있다. 그리고 단일 플랫폼 상에서 단일 사용자용으로 사용되던 검색 시스템을 분산/병렬 플랫폼 상에서 서비스의 형태로 사용자에게 디지털 포렌식의 기능을 제공함으로써 사용자 개개인이 포렌식 툴 운용 및 관리에 노력을 기울일 필요가 없을 뿐 아니라 다수의 사용자가 동시에 접속해서 서비스를 사용할 수 있게 되므로 시스템 및 데이터의 활용성이 높아지게 된다.

대 표 도 - 도3



(72) 발명자

홍도원

대전광역시 유성구 엑스포로 448, 212동 1704호 (전민동, 엑스포아파트)

김건우

대전광역시 서구 신갈마로 95, 105동 1405호 (갈마동, 갈마아파트)

김영수

대전광역시 유성구 반석동 반석마을아파트 610동 502호

은성경

대전광역시 유성구 엑스포로 448, 404동 1106호 (전민동, 엑스포아파트)

이상수

대전광역시 유성구 하기동 송림마을아파트 509동 1504호

조수형

대전광역시 유성구 엑스포로 448, 302동 404호 (전민동, 엑스포아파트)

최우용

대전광역시 유성구 엑스포로 448, 304동 1204호 (전민동, 엑스포아파트)

조현숙

대전광역시 유성구 관평동 테크노밸리아파트 702동 601호

이 발명을 지원한 국가연구개발사업

과제고유번호	10035157
부처명	지식경제부
연구사업명	정보통신산업원천기술개발사업
연구과제명	실시간 분석을 위한 디지털 포렌식 기술 개발
주관기관	한국전자통신연구원
연구기간	2010.03.01 ~ 2014.02.28

특허청구의 범위

청구항 1

분산 처리를 이용한 포렌식 인덱스 방법에 있어서,

인덱싱 대상 데이터를 분산처리를 위해 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성하는 단계;

상기 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 복수의 데이터 처리부에 할당하고, 상기 복수의 데이터 처리부에서 할당된 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 단계; 및

상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 단계를 포함하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 2

제1 항에 있어서, 상기 분할 대상 데이터를 생성하는 단계는

상기 인덱싱 대상 데이터를 분산처리를 위하여 기설정된 분할 설정에 기초하여 상기 인덱싱 대상 데이터에 포함된 처리 파일의 개수 및 데이터 처리부의 개수로 나누어 분할 대상 데이터를 생성하거나,

상기 기설정된 분할 설정에 기초하여 상기 인덱싱 대상 데이터에 포함된 처리 파일의 개수 및 크기에 따라 분할 대상 데이터를 생성하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 생성 방법.

청구항 3

제2 항에 있어서, 분할 인덱스 데이터를 생성하는 단계는

생성된 분할 대상 데이터를 복수의 데이터 처리부에 할당하고, 상기 데이터 처리부에서 할당받은 분할 대상 데이터를 로딩(loading)하는 단계;

상기 로딩된 분할 대상 데이터를 필터링하는 단계;

상기 필터링된 분할 대상 데이터에 대하여 인덱스를 추출하는 단계; 및

상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 단계를 포함하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 4

제3 항에 있어서, 상기 분할 대상 데이터를 필터링하는 단계는

상기 분할 대상 데이터를 가상 파일 시스템(VFS)가 포함된 파일 필터를 이용하여 텍스트 데이터로 변환하는 단계; 및

상기 변환된 텍스트 데이터로부터 필터링 데이터를 추출하는 단계를 포함하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 5

제3 항에 있어서, 상기 인덱스를 추출하는 단계는

상기 필터링된 분할 대상 데이터에 대하여 N개의 음절 단위로 분할하여 인덱스 대상을 추출하고, 상기 추출된 인덱스 대상을 기설정된 패턴과의 비교를 통해 인덱스를 추출하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 6

제3 항에 있어서, 상기 분할 인덱스 데이터를 생성하는 단계는

상기 추출된 인덱스와, 상기 인덱스가 추출된 데이터를 매칭시키는 단계; 및

상기 매칭된 데이터들을 교환(Shuffling), 정렬(Sorting), 및 병합(Merging)하여 분할 인덱스 데이터로 생성하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 7

제1 항에 있어서,

업데이트를 위한 임시 인덱스를 가지는 임시 인덱스 데이터를 생성하여 상기 인덱스 데이터 베이스와 통합하여 업데이트하는 단계를 더 포함하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 8

제1 항에 있어서,

사용자로부터 검색어를 입력받아 상기 인덱스 데이터 베이스를 이용하여 검색을 수행하는 단계를 더 포함하는 분산 처리를 이용한 포렌식 인덱스 방법.

청구항 9

분산 처리를 이용한 포렌식 인덱스 장치에 있어서,

인덱싱 대상 데이터를 분산처리를 위하여 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성하는 분할 대상 데이터 관리부;

상기 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 복수의 데이터 처리부에 할당하고, 상기 복수의 데이터 처리부에서 할당된 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 분할 인덱스 데이터 생성부; 및

상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 인덱스 데이터 베이스 관리부를 포함하는 분산 처리를 이용한 포렌식 인덱스 장치.

청구항 10

제9 항에 있어서, 상기 분할 대상 데이터 관리부는

상기 인덱싱 대상 데이터를 분산처리를 위하여 기설정된 분할 설정에 기초하여 상기 인덱싱 대상 데이터에 포함된 처리 파일의 개수 및 데이터 처리부의 개수로 나누어 분할 대상 데이터를 생성하거나,

상기 기설정된 분할 설정에 기초하여 상기 인덱싱 대상 데이터에 포함된 처리 파일의 개수 및 크기에 따라 분할 대상 데이터를 생성하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 생성 장치.

청구항 11

제10 항에 있어서, 상기 분할 대상 데이터 생성부는

복수의 데이터 처리부를 포함하며,

상기 데이터 처리부는

상기 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 할당받아, 상기 할당받은 분할 대상 데이터를 로딩하는 분산 저장부;

상기 분산 저장부에 로딩된 분할 대상 데이터를 필터링하는 필터링부;

상기 필터링된 분할 대상 데이터로부터 인덱스를 추출하는 인덱스 추출부; 및

상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 분할 인덱스 데이터 생성부를 포함하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 장치.

청구항 12

제11 항에 있어서, 상기 필터링부는

상기 할당된 분할 대상 데이터를 가상 파일 시스템(VFS)가 포함된 파일 필터를 이용하여 텍스트 데이터로 변환하는 텍스트 변환부; 및

상기 텍스트 데이터로부터 필터링 데이터를 추출하는 필터링 데이터 추출부를 포함하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 장치.

청구항 13

제11 항에 있어서, 상기 인덱스 추출부는

상기 필터링된 분할 대상 데이터에 대하여 N개의 음절 단위로 분할하여 인덱스 대상를 추출하는 N-gram 분석부; 및

상기 추출된 인덱스 대상을 기설정된 패턴과의 비교를 통해 인덱스를 추출하는 패턴 비교부를 포함하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 장치.

청구항 14

제11 항에 있어서, 상기 분할 인덱스 데이터 생성부는

상기 추출된 인덱스와 상기 인덱스가 추출된 데이터를 매칭시키는 인덱스 매칭부;

상기 매칭된 데이터들을 교환(Shuffling), 정렬(Sorting) 및 병합(Merging)하여 분할 인덱스 데이터를 생성하는 것을 특징으로 하는 분산 처리를 이용한 포렌식 인덱스 장치.

청구항 15

사용자에게 입력 인터페이스를 제공하는 단말부;

인덱싱 대상 데이터를 저장하는 데이터 저장부; 및

상기 인덱싱 대상 데이터를 기설정된 분할 설정에 따라 분할하고, 상기 분할된 분할 대상 데이터를 복수의 데이터 처리부를 이용하여 분할 인덱스 데이터를 생성하며, 상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 포렌식 인덱스 장치

를 포함하는 분산 처리를 이용한 포렌식 인덱스 시스템.

청구항 16

제15 항에 있어서, 상기 포렌식 인덱스 장치는

인덱싱 대상 데이터를 기설정된 분할 설정에 기초하여, 인덱싱 대상 데이터에 포함된 처리 파일의 개수 및 데이터 처리부의 개수로 나누어 분할 대상 데이터를 생성하거나, 인덱싱 대상 데이터에 포함된 처리 파일의 개수 및 크기에 따라 분할 대상 데이터를 생성하는 분할 대상 데이터 관리부;

상기 생성된 분할 대상 데이터를 복수의 데이터 처리부에 할당하고, 복수의 데이터 처리부에서 할당된 분할 대상 데이터를 로딩하여, 상기 로딩된 데이터를 필터링하여 인덱스를 추출하고, 상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 분할 인덱스 데이터 생성부; 및

상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 인덱스 데이터 베이스 관리부를 포함하는 분산 처리를 이용한 포렌식 인덱스 시스템.

명 세 서

기 술 분 야

[0001] 본 발명은 분산 처리를 이용한 포렌식 인덱스 방법 및 장치에 관한 것으로 보다 상세하게는 인덱싱(Indexing)하고자하는 대상 데이터를 분할하고, 분할한 데이터를 분산 처리하여 인덱싱하는 포렌식 인덱스 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 다양한 비즈니스 환경에서뿐만 아니라 일상 생활에서 생성, 보관하게 되는 디지털 데이터의 양이 급격히 증가함에 따라 디지털 포렌식의 각 단계를 수행하는데 있어서도 많은 시간을 필요로 한다.

[0003] 디지털 포렌식 중 '포렌식 검색' 과정은 대용량 데이터 중에서 증거를 신속히 찾아내기 위하여 반복적인 질의에 따른 검색을 수행해야 하기 때문에 소요되는 시간은 상당하다.

[0004] 하지만, 현재 디지털 포렌식에서 사용되고 있는 텍스트 검색 도구들은 한 번의 질의에 대해 평균 약 20MB/s의 속도록 검색을 수행한다. 이러한 도구를 이용할 경우 1TB의 데이터에서 질의어를 검색하기 위해서 14시간 이상이 걸린다는 것을 의미하며, 이는 1회 검색에 소요되는 시간으로 다수의 검색 요청 시에는 횟수에 비례한 시간이 필요하다.

[0005] 또한 디지털 포렌식 분야에서 발간되는 다양한 문헌들이나 응용 사례를 종합해보면 수사대상이 되는 하드 드라이브 디스크의 “크기”에 관련된 내용이 디지털 포렌식 분석에 있어서 가장 많이 언급되고 시급하게 여겨지는 문제임을 파악할 수 있다.

[0006] 예를 들어, 최근 디지털 포렌식 분야의 핫 이슈로 떠오르고 있는 eDiscovery 분야를 고려해보면, 기업은 기업 내에서 생성되는 전자 문서(ESD)들에 대해 3개월에서 6개월가량 저장 및 보관을 해야 하고, 관련 소송이 발생했을 때 요구되는 자료를 해당 데이터에서 찾아 기한 내에 법정에 제출해야 한다.

[0007] 그러나, 국내 대기업 그룹 내에서 하루에 생성되거나 수신되고 있는 이메일의 양은 적어도 500GB 이상 되며, 이를 3개월 동안 보관한 양은 단순하게 계산하더라도 상당함이 자명하다.

[0008] 국내 사법 환경에서 증거물에 대해 압수 후, 기소까지 걸리는 시간이 매우 단기간임을 고려해 볼 때, PC, 전자결재 시스템, 전자우편 시스템, 회계 데이터 베이스 등을 망라한 대규모의 저장매체를 수사함에 있어 효율적인 검색 방법에 대한 필요성이 증가하고 있다.

[0009] 종래 문제점을 해결할 수 있는 솔루션 중 하나는 인덱스 기반 검색 기술이다. 인덱스 검색 기술은 대상이 되는 데이터에 대해 인덱스를 생성하고 검색 시 인덱스를 참조하여 실시간으로 검색 결과를 제시하는 기술이다. 인덱스 검색 기술의 경우, 인덱스를 생성하는데 시간이 다소 소요되는 반면, 인덱스가 생성된 후에는 모든 검색이 실시간으로 이루어진다는 특징이 있다.

[0010] 그러나, 특정 매체에 저장된 데이터들에 대한 인덱스 분석 및 검색 알고리즘들이 방대하게 지속적으로 제안되어 왔고, 현재 다양한 분야에서 적용되고 있음에도 불구하고 디지털 포렌식 분야에서 연구에 대한 필요성이 다시 증가하는 것은 인덱스 검색 기술들이 디지털 포렌식 증거 검색에 적합하지 않기 때문이다.

[0011] 대부분의 인덱스 검색 기술들은 인터넷 상의 웹 데이터를 검색하기 위한 것으로 사용자의 질의어를 포함하고 있는 데이터 중 사용자의 의도와 가장 유사한 정보를 포함하는 데이터를 찾아서 되도록 많은 검색 결과를 빠르게 제공하는데 목적이 있다.

[0012] 하지만 디지털 포렌식 검색은 정해진 대용량 증거 데이터 내에서 수사관이 질의한 질의어가 포함되어 있는지 여부와 어느 위치에 포함되어 있는지를 하나의 놓침 없이 빠르고, 정확하게 제공하는데 그 목적이 있다.

[0013] 이러한 측면에서, 기존의 인덱스 검색 기술들과 인덱스 데이터베이스에 저장해야 할 정보의 내용과 양이 다르고, 처리 절차 또한 다르다. 또한 기존의 인덱스 검색 기술은 인덱스 검색 속도를 높이는데 그 초점이 맞춰져 있는 반면, 디지털 포렌식 검색은 인덱스 생성 속도에 초점이 맞추어져 있다.

[0014] 기존의 디지털 포렌식 도구에서 제공하는 인덱스 검색 기능들은 디지털 포렌식 도구들과 통합된 형태로 제공되거나 독립적으로 제공한다고 할지라도 단일 시스템에서 운용되는 형태로 제공되고 있다.

[0015] 이런 제공 형태는 대용량 데이터를 인덱싱 할 때에 시스템 하드웨어의 사양을 높인다 할 지라도 금방 다시 한계에 부딪히게 되어 성능이 빠르게 저하되는 문제점이 있다.

[0016] 따라서 디지털 포렌식에서 대용량 증거 검색이 실질적으로 가능하도록 인덱스 생성 속도 개선에 대한 요구사항이 점점 증가하고 있다.

발명의 내용

해결하려는 과제

[0017] 본 발명은 대용량 디지털 데이터를 대상으로 디지털 포렌식 검색시, 소요되는 데이터 검색 시간을 개선하기 위

해 분산 포렌식 인덱스 생성 방법 및 장치를 제공할 필요성이 있다.

과제의 해결 수단

[0018] 상술한 목적을 달성하기 위해 본 발명의 일 실시예는 분산 처리를 이용한 포렌식 인덱스 방법에 있어서, 인덱싱 대상 데이터를 분산처리를 위해 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성하는 단계; 상기 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 복수의 데이터 처리부에 할당하고, 상기 복수의 데이터 처리부에서 할당된 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 단계; 및 상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 단계를 포함하는 분산 처리를 이용한 포렌식 인덱스 방법을 제공하는 것을 특징으로 한다.

[0019] 상술한 목적을 달성하기 위해 본 발명의 다른 실시예는 분산 처리를 이용한 포렌식 인덱스 장치에 있어서, 인덱싱 대상 데이터를 분산처리를 위하여 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성하는 분할 대상 데이터 관리부; 상기 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 복수의 데이터 처리부에 할당하고, 상기 복수의 데이터 처리부에서 할당된 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 상기 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성하는 분할 인덱스 데이터 생성부; 및 상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 인덱스 데이터 베이스 관리부를 포함하는 분산 처리를 이용한 포렌식 인덱스 장치를 제공하는 것을 특징으로 한다.

[0020] 상술한 목적을 달성하기 위해 본 발명의 다른 실시예는 사용자에게 입력 인터페이스를 제공하는 단말부; 인덱싱 대상 데이터를 저장하는 데이터 저장부; 및 상기 인덱싱 대상 데이터를 기설정된 분할 설정에 따라 분할하고, 상기 분할된 분할 대상 데이터를 복수의 데이터 처리부를 이용하여 분할 인덱스 데이터를 생성하며, 상기 생성된 분할 인덱스 데이터들을 병합하여 인덱스 데이터 베이스를 생성하는 포렌식 인덱스 장치를 포함하는 분산 처리를 이용한 포렌식 인덱스 시스템을 제공하는 것을 특징으로 한다.

발명의 효과

[0021] 본 발명에서 따르면, 디지털 포렌식 분석에 필수적인 데이터 검색을 기존 단일 시스템에서 수행하는 것과 비교해 빠른 시간에 수행하는 효과가 있다. 또한, 확장성(Scalability)이 제공되므로 분석해야 할 데이터의 양에 따라 사용되는 자원(Resource)을 증가 혹은 감소 시킬수 있으므로 성능을 조절하는 효과가 있다.

[0022] 인덱싱 시스템의 확장성(scalability)을 높일 수 있고, 이를 통해 포렌식 분석에서의 요구사항 중 가장 큰 이슈인 처리 속도를 크게 향상시키는 효과가 있다.

[0023] 그리고 단일 플랫폼 상에서 단일 사용자용으로 사용되던 검색 시스템을 분산/병렬 플랫폼 상에서 서비스의 형태로 사용자에게 디지털 포렌식의 기능을 제공함으로써 사용자 개개인이 포렌식 도구 운용 및 관리에 노력을 기울일 필요가 없을 뿐 아니라 다수의 사용자가 동시에 접속해서 서비스를 사용할 수 있게 되므로 시스템 및 데이터의 활용성이 높아지게 된다.

도면의 간단한 설명

[0024] 도1 은 종래 기술에 따른 일반적인 인덱스 장치의 예를 도시한 참고도이다.

도2 는 종래 단일 시스템에서 운용되는 포렌식 인덱스 장치에 관한 블록도이다.

도3 은 본 발명의 일 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 방법을 도시한 순서도이다.

도4a 내지 도4b는 본 발명의 일 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 방법에서 인덱스 데이터 베이스의 저장 구조를 도시한 참고도이다.

도5a 내지 도5b는 본 발명의 일 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 방법에서 업데이트시 데이터 구조를 도시한 참고도이다.

도6은 본 발명의 다른 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 장치를 도시한 블록도이다.

도7은 본 발명의 다른 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 장치의 인덱스 추출부를 도시한 참고도이다.

도8은 본 발명의 또 다른 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 시스템을 도시한 블록도이다.

도9는 본 발명의 또 다른 실시예에 따른 하둡 시스템을 이용한 분산 처리를 이용한 포렌식 인덱스 시스템을 도시한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0025] 이하에서는 본 발명의 일부 실시예를 첨부된 도면들을 참조하여 상세히 설명한다. 아울러 본 발명을 설명함에 있어 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다.

[0026] 도1은 종래 기술에 따른 일반적인 인덱스 장치의 예를 도시한 참고도이다. 인덱스어 분석부는 데이터 상에서 인덱싱될 단어의 분석을 수행하는데, 일반적으로 형태소 중심 색인법, 단어 중심 색인법, N-gram 중심 색인법 등이 사용된다. 인덱싱될 단어는 분석 및 인덱스 데이터 베이스 처리를 거쳐 인덱스 데이터 베이스를 생성한다.

[0027] 인덱스 데이터 베이스가 생성되면, 사용자는 사용자 인터페이스를 통하여 검색을 위한 질의를 수행할 수 있게 되며, 일반적으로 질의는 검색입력 전처리부를 통해 인덱싱될 단어의 분석을 통하여 인덱싱 알고리즘에 따라 불용어 처리, N-gram 처리, 인코딩 변경 등 전처리가 수행된다.

[0028] 전처리가 수행된 검색어는 검색부에 의해 인덱스 데이터 베이스에 등록된 인덱스를 대상으로 검색이 수행되며, 그 결과가 인덱스 데이터 베이스 처리부를 거쳐 사용자에게 제공된다.

[0029] 도2는 종래 단일 시스템에서 운용되는 포렌식 인덱스 장치에 관한 블록 다이어 그램이다. 단일 시스템 기반 포렌식 인덱스 장치는 일반적인 인덱스 장치와 유사한 구조를 가지나 사용자 인터페이스와 인덱스어 분석부 사이에 가상 드라이브 생성기와 파일 필터링부를 이용하여 포렌식 이미지를 로딩하고 포렌식 수사를 위한 데이터 인덱싱을 수행할 수 있도록 한다. 다만, 대용량 데이터에 대한 인덱스를 생성하는데 시간이 상당히 소요되며, 실제로 포렌식 수사에서 인덱스 검색 기능을 활용하지 못하는 문제점이 있었다.

[0030] 도3 을 참조하여 본 발명의 일 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 방법을 설명한다.

[0031] 복수의 데이터 처리부를 가지는 인덱스 장치가 인덱싱 대상 데이터를 분산 처리를 위하여 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성한다.(S100)

[0032] 인덱싱 대상 데이터란 사용자로부터 입력받아 인덱싱을 하기 위한 데이터로서 중앙 저장 장치 혹은 외부의 데이터 저장 장치로 구성된 데이터 저장부로부터 입력받을 수 있다.

[0033] 기설정된 분할 설정이란 인덱싱 대상 데이터를 분할하여 분산 처리하기 위하여 설정된 정보들로서, 인덱싱 대상 데이터를 분할처리하기 위한 데이터 처리부의 갯수, 인덱싱 대상 데이터의 파일의 크기 및 갯수에 따른 분할 설정을 포함할 수 있다. 본 발명의 일 실시예에 따라 인덱싱 대상 데이터에 포함된 파일 개수를 포렌식 인덱스 장치에 포함된 데이터 처리부의 개수에 따라 균등하게 분할하도록 구현할 수 있으며, 또 다른 구현예로는, 인덱싱 대상 데이터에 포함된 파일들의 크기를 계산하여, 각각의 데이터 처리부를 이용하여 분산 처리할 경우 인덱싱 종료가 일정한 시간 범위 내에서 이루어지도록 파일의 크기를 고려해 인덱싱 대상 데이터를 분할할 수 있다. 일정한 시간 범위는 인덱싱 데이터 베이스를 효율적으로 수행하기 위한 시간 범위로서 사용자로부터 미리 설정받을 수 있다.

[0034] 인덱싱 대상 데이터를 분할하여 분할 대상 데이터를 생성한 경우 개별 데이터 처리부에서는 각각 할당된 분할 대상 데이터를 로딩(loading)한다.(S230) 본 발명의 일 구현예에 따라, 분할 대상 데이터는 데이터 리스트와 그에 따른 데이터를 포함할 수 있다.

[0035] 개별 데이터 처리부에 이미 분할 대상 데이터의 로딩이 이루어진 경우에는 개별 데이터 처리부로 데이터 로딩하는 과정을 스킵(Skip)할 수 있다.(S215)

[0036] 개별 데이터 처리부에 각각 할당된 분할 대상 데이터 로딩이 이루어지면 인덱스 생성을 수행하도록 각각의 데이터 처리부를 제어한다. 본 발명의 일 구현예에 따라 개별 데이터 처리부는 분산 저장부를 가질 수 있으며, 포렌식 인덱스 장치는 개별 데이터 처리부의 분산 저장부를 모아 하나의 가상 분산 저장부를 데이터 저장부로 사용하도록 구현될 수 있다. 이러한 경우 데이터 저장부에 대한 접근(Access)를 위해 분산 파일 시스템(DFS :

Distributed File System)을 사용할 수 있으며, 개별 데이터 처리부는 DFS 입출력 제어부를 포함할 수 있다.

[0037] DFS 입출력 제어부는 데이터 저장부에서 인덱싱 프로시저가 수행되는 포렌식 인덱스 장치의 분산 저장부에 있는 데이터를 접근(Access)하도록 보장함으로써 작업 성능을 향상시킬 수 있다. 즉, 인덱스 생성을 분산 처리하기 위해서 데이터를 개별 데이터 처리부에 포함된 분산 저장부로 로딩하는 경우, 데이터 처리부가 로딩된 데이터를 처리함으로써 로컬리티(locality)를 보장하는 효과를 가진다.

[0038] 따라서, 로컬리티(Locality)를 보장하고 안정적인 분산처리를 위하여 Apache Hadoop와 같은 분산 처리 프레임워크를 사용할 수 있다. Hadoop는 대용량 데이터를 분산처리하기 위해 개발된 분산 처리 프레임워크로 이를 위한 분산 파일 시스템으로 HDFS와 분산 처리 프로그래밍 모델인 MapReduce를 제공한다.

[0039] 개별 분할 인덱스 데이터 생성부(200)는 각각의 데이터 처리부에 할당된 분할 대상 데이터를 필터링한다. 본 발명의 일 실시예에 따라 분할 대상 데이터에 메타데이터가 포함되거나(S230) 포렌식 이미지 파일(DD, Encase) 처리를 위하여 가상 파일 시스템(VFS)를 포함하는 파일 필터를 이용하여 텍스트 데이터로 변환한다.(S215) 본 발명의 일 실시예에 따라, 분할 대상 데이터가 DD나 EnCase 이미지 등의 포렌식 이미지일 경우 이미지에 대한 가상 파일 시스템(VFS : Virtual File System)가 포함된 파일 필터를 이용하여, 해당 이미지에 포함된 데이터를 추출한다. 이미지에 포함된 데이터를 추출한다는 것은 NTFS, FAT 등의 파일시스템 테이블을 구성해서 이미지 상에 흩어져있는 정보를 모아 하나의 파일 형태로 추출 또는 복원하여 사용자에게 제공한다는 것을 의미한다. 추출 과정을 통하여 삭제되거나 유실된 파일에 대하여 복구가 가능하며, 이렇게 추출된 파일은 기타 일반 파일들과 마찬가지로 텍스트 데이터로 변환될 수 있다.

[0040] 본 발명의 일 실시예에 따라 메타데이터, 포렌식 이미지 형태의 데이터 파일을 텍스트 형태로 변환하는 단계를 마치면, 할당된 데이터를 읽어서 필터링 데이터를 추출한다.(S230) 할당된 데이터는 일반 문서 뿐만이 아니라 압축 파일, 그림 파일, 동영상 파일, 음악 파일 등 다양한 포맷의 문서를 포함할 수 있으며, 포렌식 이미지가 첨부된 경우에는 가상 파일 시스템(VFS)에 따라 텍스트 데이터로 변환된 데이터를 포함할 수 있다. 필터링 데이터는 포맷이 없는 플레인 텍스트(Plain text)로 변환한 데이터를 포함한다.

[0041] 또한, 필터링부는 인덱싱 대상 데이터에 포함된 정상적으로 존재하는 파일뿐만 아니라 삭제된 파일을 추출할 수 있다. 경우에 따라, 고의로 시스템상에 숨겨져있거나, 손실, 손상등 데이터의 일부가 유실된 데이터에 대하여도 추출하는 기능을 포함할 수 있다. 즉, NTFS, FAT 등의 파일 시스템 테이블을 구성해서 이미지 상에 흩어져 있는 정보를 모아 하나의 파일 형태로 추출해 사용자에게 제공함으로써 삭제되거나 유실된 파일에 대하여도 복구가 가능하도록 구현될 수 있다.

[0042] 분할 대상 데이터를 필터링한 데이터를 분석하여 인덱스를 추출한다.(S270) 인덱스 추출이 되는 데이터를 분석하는 방법으로는 N-gram 방법, 형태소 분석 방법, 패턴 매칭 방법 등 다양한 방법이 사용될 수 있다. 본 발명의 일 실시예에 따라 바이그램 분석과 패턴 분석을 이용하여 인덱스를 추출하도록 구현될 수 있다.

[0043] 필터링된 분할 대상 데이터에 대하여 N-gram, 즉 N개의 음절 단위로 분할하여 인덱스 대상을 추출하여 인덱스 데이터 베이스를 구성할 수 있으며, 추출된 인덱스 대상을 기설정된 패턴과의 비교를 통해 인덱스를 추출함으로써 인덱스 데이터 베이스를 구성할 수 있다.

[0044] 도4a를 참조하여 바이그램 분석을 통한 인덱스 대상 추출 방법을 설명한다. 바이그램 분석을 통한 인덱스 대상을 추출하는 경우에 키(Primary row key)로서 인덱스어(Term)를 이용하며, 개별 키에 해당하는 저장 데이터들은 파일 식별자 리스트를 사용할 수 있다. 바이그램 분석부의 경우에 2개의 음절로 텍스트를 분석하여 인덱스 대상을 추출할 수 있다. 따라서, 바이그램을 사용함으로써 재현율(Recall Ratio)을 향상시킬 수 있으며, 재현율은 어떤 조건으로 검색되는 정보와 검색되어야 하는 모든 정보의 비율을 말한다.

[0045] 도4b를 참조하여 패턴 분석을 통한 인덱스 추출을 설명한다. 패턴 분석을 이용하여 인덱스 추출하는 경우에 키(Primary row key)로서 패턴(Pattern)를 이용하며, 개별 키에 해당하는 저장 데이터들은 파일 식별자 리스트를 사용할 수 있다. 파일 식별자 리스트는 개별 패턴에 대하여 식별자(FID)를 부여하고, 개별 식별자에 대하여 빈도(Frequency)및 메모리상에서 시작 및 종료 어드레스(Start/End Address)를 포함하도록 구성될 수 있다.

[0046] 패턴 분석은 주민번호, 전화번호와 같이 일정한 패턴을 가지는 데이터를 분석한 결과를 이용해 개별 문서에서 일치하는 인덱스와 인덱스에 대응하는 위치를 분석하는 것이다. 패턴 분석에는 사용자가 포렌식 분석을 위하여 미리 설정한 용어 데이터등을 이용하도록 설정할 수 있다. 미리 설정한 용어 데이터는 이메일, 전화번호, 주민

번호를 포함하는 데이터를 대상으로 할 수 있으나, 이에 한정되지 않는다.

[0047] 미리 설정한 용어 데이터와 필터링 데이터를 패턴 분석을 이용해 매칭시킴으로써 인덱스 및 인덱스의 사용빈도를 추출할 수 있으며, 대용량의 많은 포렌식 데이터를 분석할 때 성능 면에서 우수하다.

[0048] 분할 인덱스 데이터를 생성하는 단계(S270)는 바이그램 분석 및 패턴 분석을 이용하여 추출된 인덱스와, 인덱스가 추출된 데이터를 매칭시키고, 매칭된 데이터를 교환(Shuffling), 정렬(Sorting), 및 병합(Merging)하여 분할 인덱스 데이터로 생성한다. 본 발명의 일 실시예에 따라, 다수의 파일에서 추출된 인덱스와, 각 인덱스가 추출된 데이터를 매칭하여 어느 파일에서 추출되었는지 정보를 포함하는 분할 인덱스 데이터를 생성한다.

[0049] 포렌식 인덱스 장치는 각각의 데이터 처리부로부터 분할 인덱스 데이터를 수신하여 수신된 분할 인덱스 데이터를 병합하여 인덱스 데이터 베이스를 생성한다.(S300) 관리 비용을 고려할 때 개별 파일을 작은 크기의 파일을 저장하는 것보다 큰 파일을 저장하는 것이 관리 비용면에서 유리하다.

[0050] 따라서, 대용량 분산 처리시 처리 속도 향상 및 저장 공간 낭비를 막기 위하여 작은 크기의 원본 파일을 하나의 데이터베이스로 구성하는 것을 특징으로 하며, 테이블로 구성하게 됨으로써 저장할 자료 구조가 간단해진다.

[0051] 또한, 본 발명의 일 실시예에 따라 인덱스 업데이트 단계를 더 포함할 수 있다. 종래 인덱스 업데이트 방식은 인덱스 데이터 베이스가 트리(Tree) 구조 형태로 구성되어 있는 경우, 인덱스 대상이 되는 파일이 추가, 삭제되었을 경우, 해당 인덱스 데이터 베이스에 새로운 인덱스 데이터를 추가, 삭제하는 작업이 쉽지 않았다.

[0052] 인덱스 데이터가 추가되는 경우, 추가되는 단어에 따라 인덱스 데이터 베이스의 트리(Tree) 구조가 재구성되어야 하고, 삭제되는 경우에는 모든 인덱스 데이터를 재검색(Traverse)하면서 해당 삭제될 문서의 식별자를 제거하는 작업이 필요하다. 따라서, 종래의 인덱스 추가, 삭제 작업은 인덱스를 재생성하는 것과 비슷한 과정 및 시간이 필요하였다.

[0053] 따라서, 종래의 인덱스 추가, 삭제 작업의 문제점을 개성하기 위하여 임시 인덱스를 가지는 인덱스 데이터를 이용하여 인덱스 데이터 베이스와 통합하여 업데이트하는 단계를 포함할 수 있다.

[0054] 도5a와 도5b를 참조하여 인덱스 데이터 베이스(의 자료 구조와 삭제 파일 리스트를 설명한다. 인덱스 데이터 베이스는 키(Key)로서 인덱스 데이터베이스 식별자(IndexDBName)를 사용하고, 키(Key)에 대응하여 저장되는 데이터로서 인덱스를 가지는 식별자 리스트(List)를 포함할 수 있다. 또한, 삭제파일 정보를 별도의 데이터 베이스를 구성하여 삭제된 파일의 식별자(FID)를 저장하도록 구현될 수 있다.

[0055] 키(Key)가 되는 인덱스 데이터베이스 식별자는 사용자에게 보여지는 실제 인덱스 데이터이고, 실제 인덱스 데이터 베이스는 하나 이상의 인덱스 데이터를 포함하는 식별자 리스트로 구성된다.

[0056] 인덱스 데이터 베이스에 인덱스가 새로 생성되는 경우, 인덱스 생성이 완료되는 시점에 인덱스 이름을 인덱스 데이터 베이스의 키(Key)로 설정하며, 인덱스에 대응하는 값(Value)을 등록한다.

[0057] 기존 인덱스 데이터 베이스에 파일을 추가하는 경우, 추가 파일만으로 임시 인덱스를 갖는 임시 인덱스 데이터를 새로 생성하고, 인덱스 데이터 베이스에 생성된 인덱스를 등록한다. 그리고, 인덱스 데이터 베이스에 임시 인덱스 데이터의 키(Key) 및 값(Value)을 등록한다.

[0058] 검색 시에는 인덱스 데이터 베이스의 값(Value)이 존재하는 인덱스 데이터 베이스를 모두 검색하고 결과를 통합해서 사용자에게 제공한다.

[0059] 본 발명의 일 실시예에 따라 검색어를 입력받아 인덱스 데이터 베이스를 이용하여 검색을 수행하는 단계를 더 포함할 수 있다. 바이그램 분석을 통하여 인덱스 데이터 베이스 저장 구조를 단순화하고 패턴 검색을 통하여 정규식으로 표현된 패턴을 찾아 해당 패턴에 대한 정보를 테이블형태의 인덱스 데이터 베이스로 저장하였는바, 이러한 인덱스 데이터 베이스를 포렌식 검색을 수행하여 사용자에게 결과를 제공할 수 있다.

[0060] 본 발명의 일 구현예에 따라 검색어에 대한 전처리를 수행하고, 인덱스 데이터 베이스를 검색하여, 검색 결과에 대한 정확도를 높이기 위하여 후처리를 수행할 수 있다.

[0061] 검색어에 대한 전처리란, 검색어로 의미가 없는 불용어(Stopword)를 제거하거나 인코딩을 변경하는 것을 의미하며, 불용어는 검색시 검색 용어로 사용하지 않는 단어로 예를 들어, 관사, 전치사, 조사, 접속사등으로 검색 인덱스로서는 의미가 없는 단어를 의미한다.

[0062] 검색결과에 대한 후처리란 예를 들어, 바이그램 인덱스를 대상으로 검색된 디지털 자료를 필터링하여 검색결과를 출력하는 것을 의미한다. 예를 들어 출력되는 검색 결과에는 검색어가 포함된 각 문서 이름과 그 문서 내에서의 빈도수가 포함될 수 있다.

[0063] 본 발명의 일 실시예에 따라 인덱스 갱신 상태를 유지하면서 인덱스 검색을 수행하는 방법을 설명한다.

[0064] 포렌식 인덱스 장치가 인덱스 데이터를 검색하기 위해서 사용자로부터 검색 키워드에 대하여 필터링을 수행한다. 본 발명의 일 실시예에 따라 N-gram 중 바이그램 분석을 수행할 수 있다. 인덱스 데이터 베이스에서 검색하고, 검색된 결과들(FID)의 교집합을 구한 값을 해당 문자열(String)에 대한 1차 결과로 한다. 복수의 문자열(String)가 존재할 경우, 이러한 과정을 반복하고, 완료시 해당 결과들에 대해 입력된 AND, OR, NOT 등의 연산(Operation)을 적용한다.

[0065] 다음 단계로, 1차 결과에 대해 파일 내용을 얻어 문자열(String) 검색을 수행하고 문자열(String)가 포함된 파일을 2차 결과로 한다. 바이그램 분석과 문자열(String) 검색을 함께 수행하면, 바이그램 분석에 의한 재현율을 높이고, 검색 결과에 포함된 불필요한 정보(Garbage)를 제거함으로써 검색 결과에 대한 정확도를 높이는 효과가 있다.

[0066] 2차 결과에 포함된 파일 식별자가 삭제 파일 리스트에 포함되었는지 여부를 판단해 포함되지 않은 경우, 해당 결과를 최종 검색 결과에 포함한다.

[0067] 도면 5a에 도식화된 인덱스 데이터 베이스를 참조해서 검색 대상이 되는 인덱스 데이터가 하나 이상의 인덱스 데이터로 구성이 되었다면 해당 DB에 대해 동일한 과정을 거쳐 검색을 수행하고 해당 결과를 최종 결과에 포함한다.

[0068] 최종 결과는 문자열(String)를 포함하고 있는 파일의 파일 식별자, URI 또는 URL 형태의 파일 위치, 파일 경로 등을 포함할 수 있다. 특히 클라이언트가 모바일 디바이스와 같은 계산력(Computing power)이 떨어지는 디바이스를 이용하는 경우, URL을 이용해 검색 결과를 제공함으로써 문서 필터링 등의 대부분의 처리는 서버 측에서 담당하게 함으로써 클라이언트의 부하를 줄이면서 사용자는 원하는 결과를 언제, 어디서나 획득할 수 있다.

[0069] 도6 을 참조하여 본 발명의 일 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 장치를 도시한다. 본 발명의 포렌식 인덱스 장치는 분할 대상 데이터 관리부(100), 분할 인덱스 데이터 생성부(200), 인덱스 데이터 생성부(300)를 포함할 수 있다.

[0070] 분할 대상 데이터 관리부(100)는 인덱싱 대상 데이터를 분산 처리를 위하여 기설정된 분할 설정에 따라 분할하여 분할 대상 데이터를 생성한다. 인덱싱 대상 데이터는 중앙 저장 장치 혹은 외부의 데이터 저장 장치로부터 입력받을 수 있다. 본 발명의 구현 예에 따라 인덱싱 대상 데이터의 파일 개수를 데이터 처리부의 개수로 균등하게 분할하거나, 파일의 크기 및 개수를 고려하여 데이터를 분할할 수 있다.

[0071] 분할 인덱스 데이터 생성부(200)는 복수의 데이터 처리부를 포함하며, 개별 데이터 처리부는 생성된 분할 대상 데이터를 기설정된 분할 설정에 따라 할당받아, 할당받은 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성한다.

[0072] 개별 데이터 처리부에 이미 분할 대상 데이터 로딩이 이루어져 있거나, 분산 저장부에 분할된 인덱싱 대상 데이터가 이미 할당된 경우에는 데이터 처리부로 데이터 로딩하는 과정을 스킵(Skip)하도록 제어할 수 있다. 즉, 개별 데이터 처리부가 분산 저장부를 통하여 데이터를 제어함으로써 로컬리티(Locality)를 보장하는 효과를 가진다.

[0073] 분할 인덱스 데이터 생성부(200)는 할당된 분할 대상 데이터를 필터링하여 인덱스를 추출하고, 추출된 인덱스를 포함하는 분할 인덱스 데이터를 생성한다. 분할 인덱스 데이터 생성부(200)는 분산 저장부(210),필터링부(230), 인덱스 추출부(250), 분할 인덱스 데이터 생성부(270)를 포함할 수 있다.

[0074] 포렌식 인덱스 장치는 분산 저장부(210)를 모아 가상 분산 저장부를 데이터 저장부로 사용하도록 구현할 수 있다. 본 발명의 일 실시예에 따라 Apache Hadooop와 같은 분산 처리 프레임워크를 사용할 수 있으며, 분할 대상

데이터 관리부와 데이터 처리부는 DFS 입출력 제어부를 더 포함할 수 있다.

[0075] 필터링부(230)는 데이터 처리부에 할당된 분할 대상 데이터를 필터링하는 것으로, 할당된 분할 대상 데이터를 텍스트 데이터로 변환한다. 할당된 데이터는 일반 문서, 압축 파일, 그림 파일등 다양한 포맷의 문서를 포함하며, DD나 Encase와 같은 포렌식 이미지가 포함된 경우에는 가상 파일 시스템(VFS)가 포함된 필터를 이용하여 포맷이 없는 플레인 텍스트(Plain text)로 데이터로 변환하여 필터링한다. 즉, NTFS, FAT등의 파일 시스템 테이블을 구성해서 이미지 상에 흩어져있는 정보를 모아 하나의 파일 형태로 제공하면, 제공된 파일을 플레인 텍스트(Plain test)로 변환시킨다.

[0076] 도7을 참조하여 인덱스 추출부(250)을 설명한다. 인덱스 추출부(250)는 필터링된 분할 대상 데이터를 분석하여 인덱스를 추출한다. 인덱스 추출이 되는 데이터를 분석하는 방법으로 N-gram 방법, 형태소 분석 방법, 패턴 매칭 방법등을 사용할 수 있으며, 본 발명의 일 실시예에 따라 바이그램 분석부(251), 패턴 매칭부(253)를 포함할 수 있다. 바이그램 분석 및 패턴 매칭 방법을 살피면 바이그램 분석부(251)는 필터링된 데이터를 2개의 음절로 텍스트를 분석하여, 패턴 매칭부(253)은 사용자로부터 기설정된 용어 데이터와 패턴 매칭되는지 여부를 판단하여 인덱스를 추출한다.

[0077] 분할 인덱스 데이터 생성부(270)는 인덱스 추출부에서 추출된 인덱스와 인덱스가 추출된 데이터를 매칭하여 어느 파일에서 인덱스가 추출되었는지 정보를 포함하는 데이터를 교환, 정렬 및 병합하여 분할 인덱스 데이터를 생성한다.

[0078] 인덱스 데이터 생성부(300)는 각각의 분할 인덱스 데이터 생성부(200)로부터 분할 인덱스 데이터를 수신하여 수신된 분할 인덱스 데이터를 병합하여 인덱스 데이터 베이스를 생성한다.

[0079] 본 발명의 일 실시예에 따라 검색어를 입력받아 인덱스 데이터 베이스를 이용하여 검색부를 더 포함할 수 있다. 본 발명의 일 구현예에 따라 검색어에 대한 전처리를 수행하고, 인덱스 데이터 베이스를 검색하여, 검색 결과에 대한 정확도를 높이기 위하여 후처리를 수행할 수 있다.

[0080] 검색어에 대한 전처리란, 검색어로 의미가 없는 불용어(Stopword)를 제거하거나 인코딩을 변경하는 것을 의미하며, 검색결과에 대한 후처리란 예를 들어, 바이그램 인덱스를 대상으로 검색된 디지털 자료를 필터링하여 검색결과를 출력하는 것을 의미한다.

[0081] 도8 를 참조하여 본 발명의 일 실시예에 따른 분산 처리를 이용한 포렌식 인덱스 시스템을 설명한다. 본 발명의 포렌식 인덱스 시스템은 단말부(400), 데이터 저장부(500), 포렌식 인덱스 장치를 포함할 수 있다.

[0082] 단말부(400)는 사용자로부터 인덱스 생성, 검색 등의 입력을 받을 수 있거나 포렌식 인덱스 생성 및 검색을 위한 인터페이스를 제공하고, 입력된 명령을 포렌식 인덱스 장치에 전달한다.

[0083] 본 발명의 일 구현예에 따라 단말부는 포렌식 이미지를 핸들링하기 위한 VFS(Virtual File System)부를 포함할 수 있다. VFS부가 포함된 경우, 사용자가 선택한 포렌식 이미지의 파일 시스템(File System)을 마운트(Mount)해 사용자에게 제시하고, 전체 파일 또는 사용자가 선택한 파일만을 인덱싱하도록 추출해 데이터 저장부에 저장하도록 할 수 있다. 또한, VFS부는 정상 파일뿐 아니라 삭제 파일, 유실 데이터 및 은닉 데이터 등을 추출할 수 있는 기능을 포함할 수 있다.

[0084] 본 발명의 일 구현예에 따라 단말부는 포렌식 인덱스 장치를 외부에서 제어할 수 있도록 웹 분석 기능을 포함할 수 있으며, 이러한 기능을 위하여 포렌식 인덱스 장치에 접속하는 대상의 인증 및 권한을 확인할 수 있는 접근 제어부를 포함할 수 있다.

[0085] 본 발명의 일 구현예에 따라 단말부는 일반 PC 어플리케이션 및 웹 기반 또는 모바일 기반 응용의 형태로 제공될 수 있다. 웹 기반 또는 모바일 디바이스에서 클라이언트가 운용되는 경우, 인덱스 생성 및 검색을 위한 복잡한 작업이 포렌식 인덱스 장치에서 수행되고 요청에 대한 결과만을 단말부로 전달함으로써 클라이언트가 운용되는 디바이스의 성능에 의존하지 않고 포렌식 인덱스 검색을 수행할 수 있다는 장점이 있다.

[0086] 데이터 저장부(500)는 인덱싱 대상 데이터를 저장하며, 중앙 저장 장치를 포함할 수 있다. 데이터 저장부로부터 분산 저장부로의 접근(Access)을 위하여 분산 파일 시스템(DFS)이 사용될 수 있으며, 분산 저장부는 데이터 처리부에서 수행되는 물리 시스템의 로컬 저장 장치를 모아 하나의 가상 분산 저장부로 구현될 수 있다.

[0087] 포렌식 인덱스 장치는 정해진 프로토콜을 기반으로 인덱스 생성 및 검색 요청을 받고 인덱스 생성을 위해 데이터 처리부를 제어하거나 검색부를 제어할 수 있다.

[0088] 인덱스 생성을 위해 복수의 데이터 처리부를 이용하여 인덱스 생성을 위한 명령을 처리하고 그 결과를 단말부로 전송할 수 있다.

[0089] 인덱스 검색의 경우, 데이터 저장부에 저장되어 있는 인덱스 데이터 베이스에 대한 검색 쿼리를 수행하여, 수행된 검색 결과를 단말부에 전달할 수 있다.

[0090] 본 발명의 일 구현예에 따라 포렌식 인덱스 장치는 사용자 인증부를 더 포함할 수 있으며, 인증/인가부는 통신부를 통하여 단말부가 작업을 요청하는 경우 해당 작업이 인증(Authentication)된 사람에 의한 것인지 여부 및 접근하고자 하는 데이터가 인가(Authorization)된 것인지를 판단한다. 인증된 사용자가 아니거나 사용자에게 접근 권한이 없는 데이터을 접근하고자 하는 경우에는 작업 요청을 거부할 수 있다.

[0091] 도9 를 참조하여 본 발명의 일 구현예로서, 하둡 프레임워크를 기반으로 한 분산 처리를 이용한 포렌식 인덱스 방법을 설명한다.

[0092] 사용자로부터 인덱싱 요청을 받은 클라이언트는 인덱싱 대상 데이터를 데이터 저장부에 저장한다. 본 발명의 일 실시예에 따라 데이터 저장부는 NAS(Network Attached Storage)와 같은 고속 저장 디바이스가 사용될 수 있으며, 데이터 Access를 위하여 NFS(Network File System)등의 파일 시스템을 이용할 수 있다.

[0093] 포렌식 인덱스 장치는 분산 처리를 위해서 데이터 저장부()에 저장된 데이터를 분할하여 데이터 처리부로 로딩하면, 각 데이터 처리부는 할당된 데이터를 분산 파일 시스템으로 업로드한다.

[0094] 이 때 하둡 파일 시스템에서 저장할 수 있는 객체(파일, 디렉토리등)의 개수가 메모리 크기에 의해 결정되고, 각 파일을 관리하는데 비용이 들게 되므로 작은 크기의 파일을 저장하는 것보다 큰 파일을 저장하는 것이 관리 비용면에서 유리하다.

[0095] 따라서, 대용량 분산처리시 처리 속도 향상 및 저장 공간 낭비를 막기 위하여 작은 크기의 원본 파일을 하나의 데이터베이스 테이블로 구성하는 것을 특징으로 한다. 테이블로 구성하게 됨으로써 저장할 자료 구조가 간단해지고, 파일 검색시에도 파일 식별자(File Identifier: FID)를 키(Primary Key)로 해서 원본 파일을 쉽게 가져올 수 있다.

[0096] 원본 파일 테이블은 파일의 식별자(FID)를 키(Key)로 설정하고 파일 내용을 저장값(Value)로 갖는다. 데이터 처리부가 데이터 저장부에 직접 접근 가능한 경우 데이터 로딩 단계는 생략될 수 있다. 데이터를 데이터 저장부에 저장하는 과정에서 FID와 원본 파일 경로를 매칭시킨 파일 정보 테이블을 함께 생성한다. 파일 정보 테이블의 구성에 따라 사용자에게 검색 결과를 제공할 때 FID 대신 원본 파일 경로를 제공할 수 있으며, Key를 FID로 사용하고 Value로 원본 파일 경로를 가진다.

[0097] 또한, 원본 파일 테이블로부터 FID를 Key로 하고 파일 내용을 Value로 하는 Key-Value 쌍으로 읽어 필터링, 인덱스어 분석, 분석된 인덱스어를 교환(Shuffling), 정렬(Sorting), 병합(Merging)하는 일련의 과정을 Mapping과 Reducing으로 구성된 Hadoop의 프로그래밍 모델로 구성할 수 있다.

[0098] Map/Reduce 과정을 통해 생성된 결과는 인덱스어를 Key로 하고 인덱스어가 포함된 파일들의 식별자 리스트 즉, FID 리스트를 Value로 하는 Key-Value 쌍이다. Key-Value쌍은 인덱스 테이블에 저장되며, 패턴 매칭 인덱스의 경우, Key는 정규식 패턴이 되고, Value는 각 FID당 빈도수와 빈도수에 따른 파일에서의 위치주소(시작주소와 끝주소)를 포함하는 정보로 value를 구성하는 쌍이 된다.

[0099] 원본 파일을 테이블로 구성하면 Hadoop에서 처리하는 기본 단위(블록, Default: 64MB)보다 작은 크기의 파일을 HDFS에 저장할 때 발생할 수 있는 저장공간 부족 문제와 MapReduce 수행 시, 작은 파일이 저장된 블록 당 하나의 mapper 프로세스가 수행될 때 발생하는 스케줄링, 프로세스 fork, 프로세스간 통신, 메모리 사용 등의 비용 증가 문제를 해결할 수 있다.

[0100] 또한, 파일 검색 결과에서 사용자가 파일 내용을 검토하고자 할 때 테이블에서 Key값으로 바로 파일 내용을 찾을 수 있기 때문에 사용자 요청에 빠르게 대처할 수 있다.

[0101] 기존의 단일 시스템에서 수행되는 인덱스 기술들은 인덱스 DB를 Tree구조로 구성해 파일에 저장함으로써, 검색

속도를 향상시키는 방법을 취해왔으나, 대용량 데이터에 대한 인덱싱의 경우, 인덱스 DB의 크기가 늘어남에 따라 Scalability와 Reliability를 보장해 줄 필요가 있다.

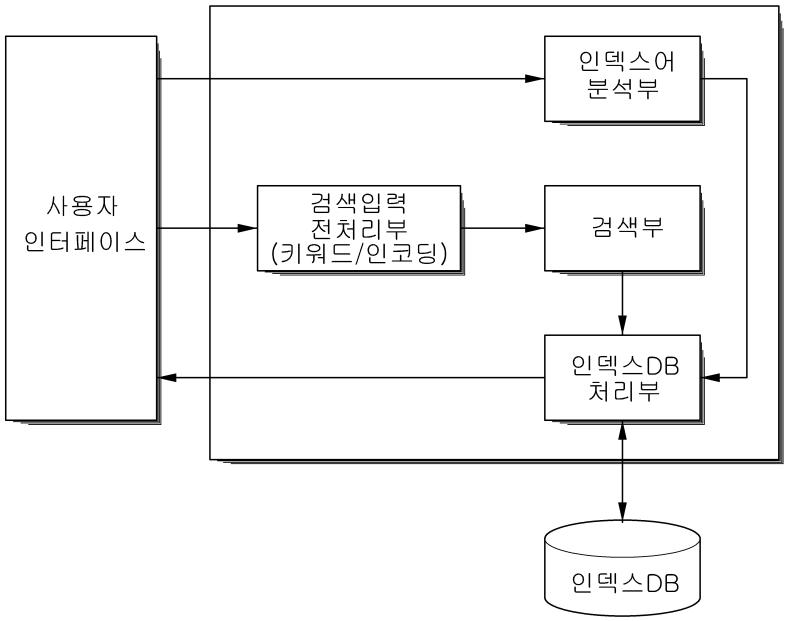
[0102] 따라서 대규모의 데이터를 저장, 관리할 필요가 있으며, 본 발명에서는 이러한 요구를 충족시킬 수 있는 NoSql 데이터 베이스에 인덱스 DB를 저장하는 것을 특징으로 할 수 있다.

[0103] 본 발명에 의한 실시예들은 컴퓨터 프로그램으로 작성 가능하다. 이 컴퓨터 프로그램을 구성하는 코드들 및 코드 세그먼트들은 당해 분야의 컴퓨터 프로그래머에 의하여 용이하게 추론될 수 있다. 또한, 해당 컴퓨터 프로그램은 컴퓨터가 읽을 수 있는 정보저장매체(Computer Readable Media)에 저장되고, 컴퓨터에 의하여 읽혀지고 실행됨으로써 실시예를 구현한다. 정보저장매체는 자기 기록매체, 광 기록매체 및 캐리어 웨이브 매체를 포함한다.

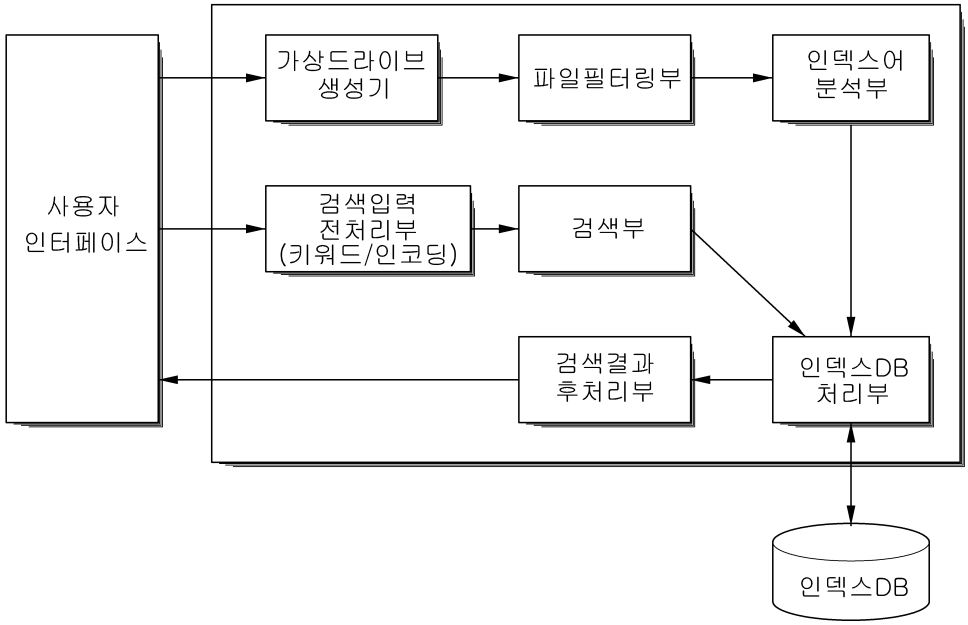
[0104] 이제까지 본 발명에 대하여 바람직한 실시예를 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 본 발명을 구현할 수 있음을 이해할 것이다. 그러므로, 상기 개시된 실시예 들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 한다.

도면

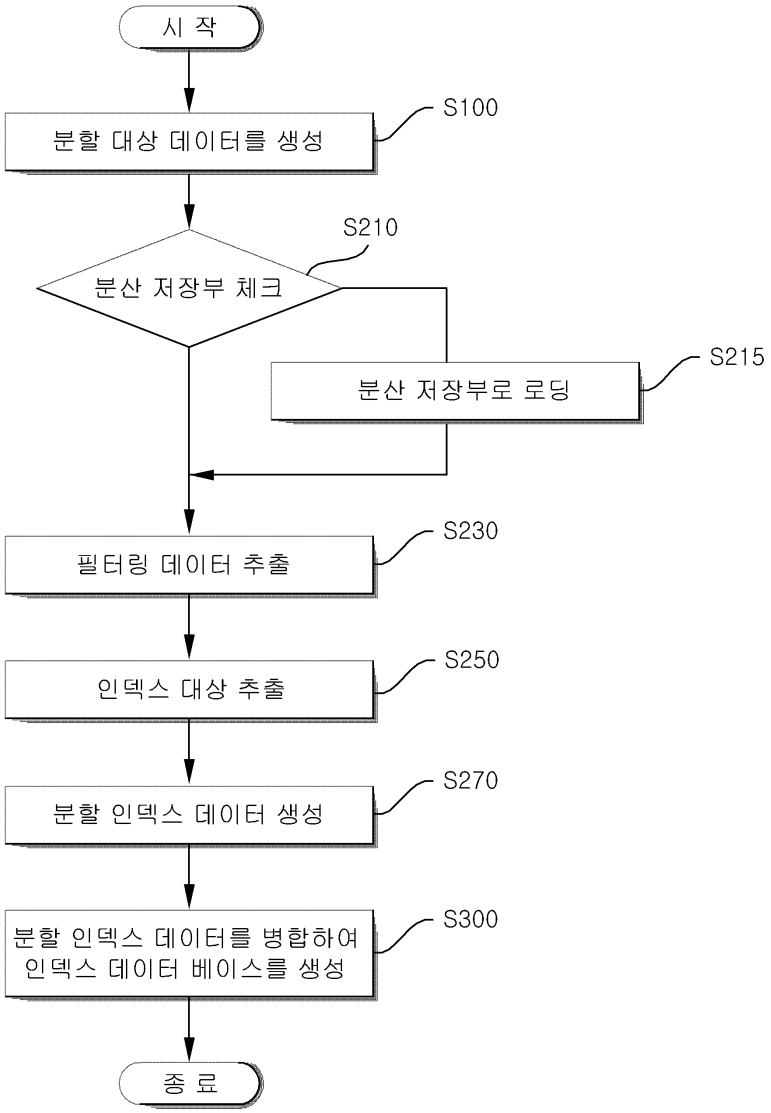
도면1



도면2


사용자
인터페이스
가상드라이브
생성기
파일필터링부
인덱스어
분석부
검색입력
전처리부
(키워드/인코딩)
검색부
검색결과
후처리부
인덱스DB
처리부
인덱스DB

도면3



도면4a

Term (Primary Row Key)	FIDs
word 1	1, 2, 4, 8
word 2	2, 3, 5
word 3	6, 9

도면4b

Pattern (Primary Row Key)	FID	Frequency	Start/End Address
Pattern 1	1	2	0x0001~0x0003, 0x0010~
	2	1	0x0012~0x0013
	3	2	0x00C2~0x00D1
Pattern 2	3	1	…
	4	5	…

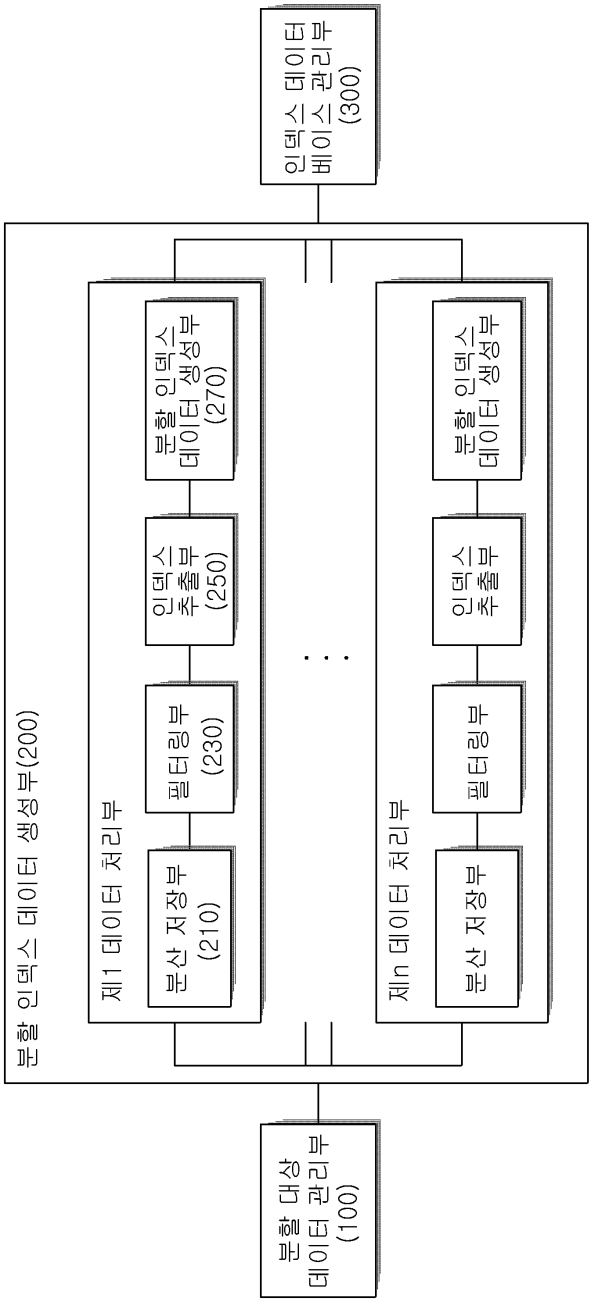
도면5a

IndexDBName	List
A. idx	A. idx
B. idx	B. idx
C. idx	C. idx X. idx
D. idx	C. idx D. idx
E. idx	E. idx E'. idx
E'. idx	E'. idx

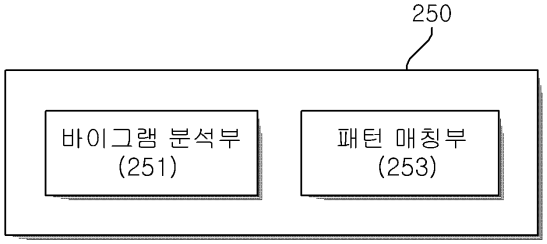
도면5b

FileID (FID)
3
9
. .

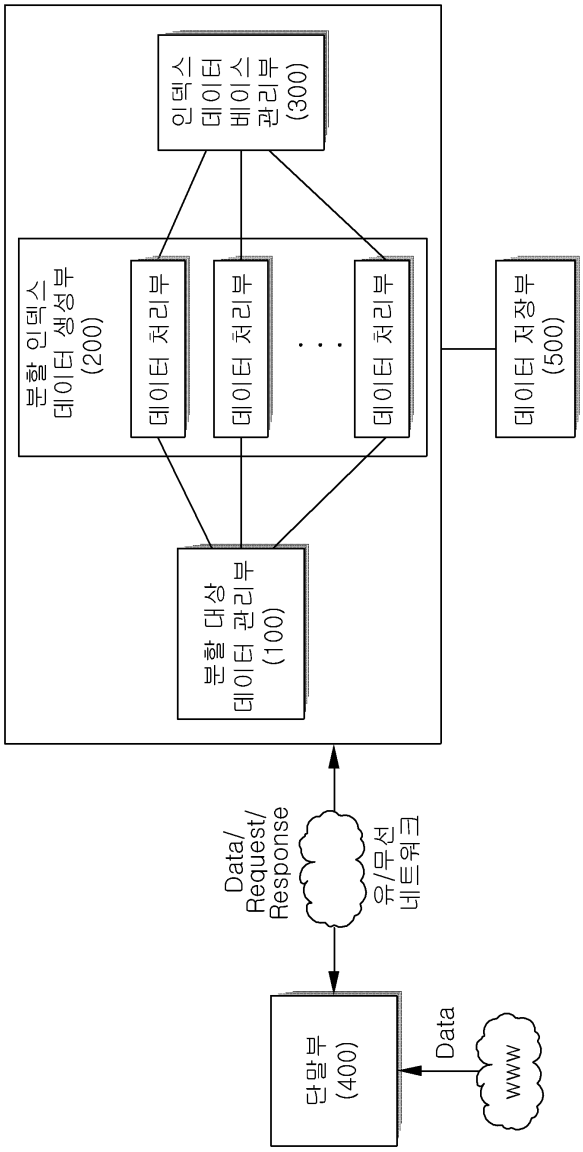
도면6


분할 대상 데이터 관리부 (100)
분할 인덱스 데이터 생성부(200)
제1 데이터 처리부
분산 저장부 (210)
필터링부 (230)
인덱스 추출부 (250)
분할 인덱스 데이터 생성부 (270)
제n 데이터 처리부
분산 저장부
필터링부
인덱스 추출부
분할 인덱스 데이터 생성부
인덱스 데이터 베이스 관리부 (300)

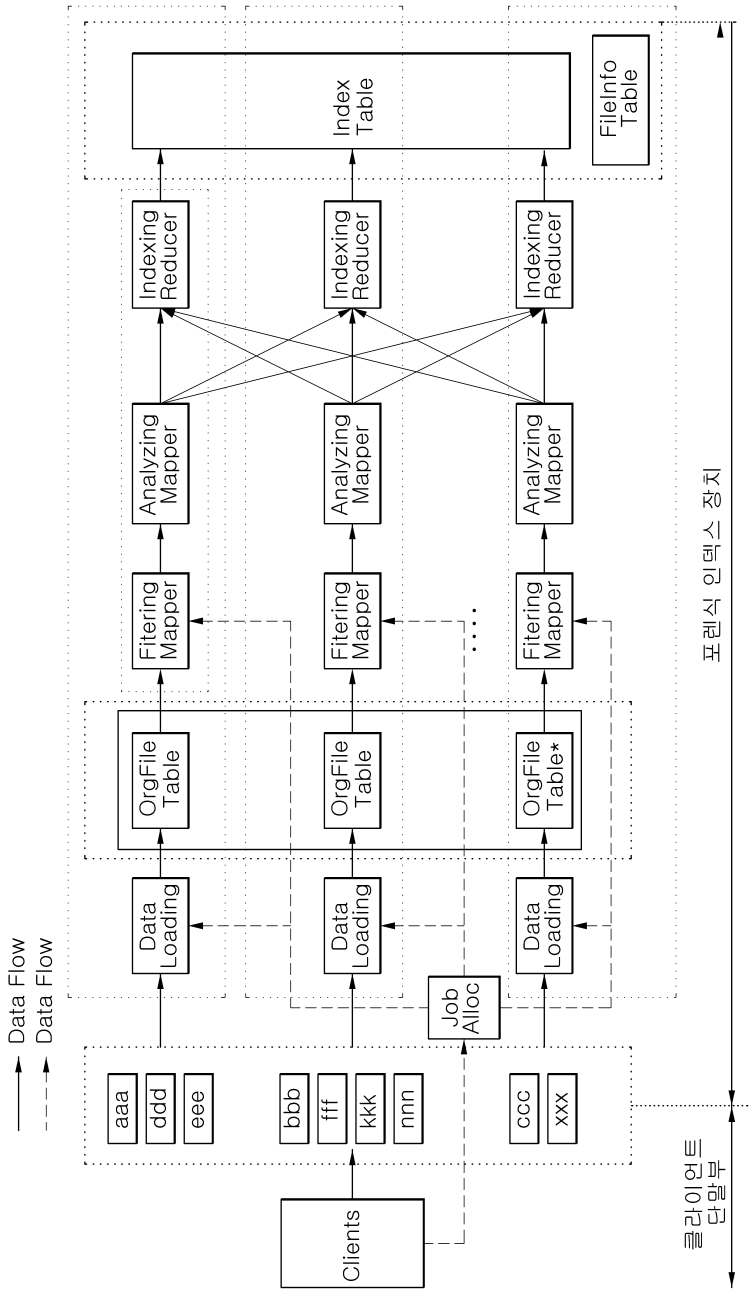
도면7


250
바이그램 분석부 (251)
패턴 매칭부 (253)

도면8


인덱스 데이터베이스 관리부 (300)
분할 인덱스 데이터 생성부 (200)
데이터 처리부
데이터 처리부
데이터 처리부
데이터 저장부 (500)
분할 대상 데이터 관리부 (100)
Data/ Request/ Response
유/무선 네트워크
단말부 (400)
Data
www

도면9


Data Flow
Data Flow
Clients
aaa
ddd
eee
bbb
fff
kkk
nnn
ccc
xxx
Job Alloc
Data Loading
Data Loading
Data Loading
OrgFile Table
OrgFile Table
OrgFile Table*
Fitering Mapper
Fitering Mapper
Fitering Mapper
Analyzing Mapper
Analyzing Mapper
Analyzing Mapper
Indexing Reducer
Indexing Reducer
Indexing Reducer
Index Table
FileInfo Table
포렌식 인덱스 장치
클라이언트 단말부