(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년02월28일
(11) 등록번호 10-1238595
(24) 등록일자 2013년02월22일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 9/00* (2006.01)
(21) 출원번호 10-2006-0012089
(22) 출원일자 2006년02월08일
심사청구일자 2011년02월01일
(65) 공개번호 10-2006-0096281
(43) 공개일자 2006년09월11일
(30) 우선권주장
11/072,726 2005년03월03일 미국(US)
(56) 선행기술조사문헌
US5905980 A
US6381602 A

(73) 특허권자
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
져스키, 데니스
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
펠토넨, 카일 지.
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
샘소노브, 에브게니 에이.
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
(74) 대리인
제일특허법인

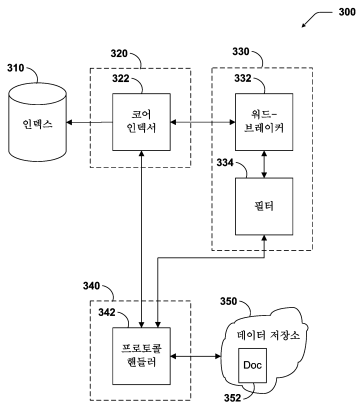
전체 청구항 수 : 총 8 항 심사관 : 권영학

(54) 발명의 명칭 **보안성 있는 풀 텍스트 인덱싱 방법 및 시스템**

(57) 요 약

프로세스의 필터링 및 워드브레이킹 부분이 제한된 보안 설정 하에서 수행되도록 프로세스의 필터링 및 워드브레이킹 부분을 분리하는 프로세스에 따라 문서가 인덱싱된다. 인덱서에 의해 문서가 요청되는 경우, 그 문서가 검색된 후 고도의 보안 프로세스로 진행된다. 다음에, 상기 문서는 인덱서로 포워딩되기 전에 제한된 보안 설정 하에서 하나 이상의 필터로 필터링되고 하나 이상의 워드브레이커에 의해 토큰화된다. 제한된 보안 설정은 보안 취약성이 필터링 및 워드브레이킹 프로세스동안 이용되는 것을 방지한다.

대 표 도 - 도3



특허청구의 범위

청구항 1

문서의 보안성있는 풀 텍스트(full-text) 인덱싱을 위한 컴퓨터 구현 방법으로서,

데이터 저장소에 저장된 사용자 데이터를 인덱서가 판독하거나 상기 사용자 데이터에 기록하는 것을 방지하는 제1 보안 설정이 적용된 제1 프로세스 하에서 실행되는 상기 인덱서로부터 문서 식별자를 수신하는 단계;

상기 인덱서에 대한 직접 호출(direct call) 시에 상기 제1 프로세스로부터 수신된 상기 문서 식별자를 상기 인덱서에 의해 송신된 문서 식별자와 비교함으로써 상기 문서 식별자를 상기 인덱서로 크로스체크(cross-checking)하는 단계;

상기 제1 프로세스로부터 수신된 상기 문서 식별자와 상기 인덱서에 의해 송신된 상기 문서 식별자가 동일하면,

프로토콜 핸들러에 의해 상기 문서 식별자에 대응하는 문서를 상기 데이터 저장소로부터 검색하고 - 상기 프로토콜 핸들러는 상기 데이터 저장소로부터 상기 프로토콜 핸들러가 데이터를 판독할 수 있도록 하는 제2 보안 설정이 적용된 제2 프로세스 하에서 실행됨 -,

제한된 보안 설정이 적용된 제3 프로세스 하에서 실행되는 워드브레이커와 필터 둘 다를 통해 상기 문서를 프로세싱하며 - 상기 제한된 보안 설정은 상기 워드브레이커와 상기 필터 모두 상기 데이터 저장소를 판독하거나 또는 상기 데이터 저장소에 기록하지 못하도록 함 -,

상기 프로세싱된 문서를 상기 인덱서로 포워딩하는 단계; 및

상기 제한된 보안 설정 하에서 실행되는 상기 제3 프로세스를 일정한 시간 간격으로 셧다운(shutting down)하는 것에 의해 보안을 침해하는 기회가 있는 윈도우(window of opportunity)를 제한하는 단계

를 포함하는 컴퓨터 구현 방법.

청구항 2

제1항에 있어서,

문서 식별자들은 상기 인덱서에 의해 송신되고 문서들은 묶음(batches)으로 검색되는 컴퓨터 구현 방법.

청구항 3

제1항에 있어서,

상기 프로토콜 핸들러는 상기 인덱서로부터 상기 문서의 요청을 직접 수신하는 컴퓨터 구현 방법.

청구항 4

제1항에 있어서,

상기 프로토콜 핸들러는, 상기 문서의 요청이 상기 제한된 보안 설정에 대응하는 프로세스를 통해 전파된 후 상기 문서의 요청을 수신하는 컴퓨터 구현 방법.

청구항 5

문서의 보안성있는 풀 텍스트 인덱싱을 위한 컴퓨터 실행가능 명령어들을 포함하는 컴퓨터 판독 가능 기록 매체로서, 상기 컴퓨터 실행가능 명령어들은 프로세서에 의해 실행될 때, 상기 프로세서가 제1항 내지 제4항 중 어느 한 항의 방법을 수행하도록 하는 컴퓨터 판독 가능 기록 매체.

청구항 6

문서 식별자를 구비한 문서의 보안성있는 풀 텍스트 인덱싱을 위한 시스템에 있어서,

인덱스의 엔트리들이 상기 문서의 워드들에 대응하는 상기 인덱스를 구축하도록 구성된 인덱서 - 상기 인덱서는 데이터 저장소에 저장된 사용자 데이터를 상기 인덱서가 판독하거나 상기 사용자 데이터에 기록하는 것을 방지

하는 제1 보안 설정이 적용된 제1 프로세스 하에서 실행됨 - ;

상기 인덱서에 대한 직접 호출 시에 상기 제1 프로세스로부터 수신된 상기 문서 식별자를 상기 인덱서에 의해 송신된 문서 식별자와 비교함으로써 상기 문서 식별자를 상기 인덱서로 크로스체크하고 그 다음에 상기 제1 프로세스로부터 수신된 상기 문서 식별자가 상기 인덱서에 의해 송신된 상기 문서 식별자와 동일하면 상기 인덱서로부터의 문서 요청 수신 시에 상기 데이터 저장소로부터 상기 문서를 검색하도록 구성된 프로토콜 핸들러 - 상기 프로토콜 핸들러는 상기 데이터 저장소로부터 상기 프로토콜 핸들러가 데이터를 판독할 수 있도록 하는 제2 보안 설정이 적용된 제2 프로세스 하에서 실행됨 - ; 및

제한된 보안 설정이 적용된 제3 프로세스 하에서 상기 문서를 프로세싱하고, 상기 제1 프로세스로부터 수신된 상기 문서 식별자가 상기 인덱서에 의해 송신된 상기 문서 식별자와 동일하면 상기 프로세싱된 문서를 상기 인덱서로 포워딩하도록 구성된 워드브레이커와 필터 - 상기 제3 프로세스는 일정한 시간 간격으로 셧다운되도록 구성되고, 상기 제한된 보안 설정은 상기 워드브레이커와 상기 필터 모두 상기 데이터 저장소를 판독하거나 상기 데이터 저장소에 기록하지 못하도록 함 -

를 포함하는 시스템.

청구항 7

제6항에 있어서,

상기 필터는 상기 문서를 순수 텍스트(pure text)로 변환하도록 구성된 시스템.

청구항 8

제7항에 있어서,

상기 워드브레이커는 상기 순수 텍스트를 워드들로 토큰화(tokenize)하도록 구성된 시스템.

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

[0012] 콘텐츠용의 파일 시스템 및 네트워크 간의 검색은 많은 형태로 제공되지만 거의 대부분 공통적으로 검색 엔진의 변형으로서 제공된다. 검색 엔진은 특정 키워드에 대하여 네트워크 상에서 문서를 검색하고 그 키워드가 발견된 문서의 리스트를 리턴하는 프로그램이다. 종종, 네트워크 상의 문서는 네트워크를 "크롤링(crawling)"함으로서 우선적으로 식별된다.

[0013] 크롤링 시에 문서를 검색하기 위해서, 네트워크 상에서 각 문서에 대한 동작은 그 문서를 획득하고 그 문서에 대한 기록으로 인덱스를 수집하도록(populate) 실행된다. 그러한 검색 시스템에는 보안 취약성이 존재한다. 종종, 인터넷으로부터 들어오는 문서는 이들이 악의적일 수 있거나 또는 특히 취약성 중 하나를 노출하도록 가공됨에 따라 반드시 신뢰성이 있는 것은 아니다. 검색의 임의의 부분 및 인덱싱 프로세스는 사용자의 머신을 완전히 인수하기 위해 사적 정보로부터의 범위의 서로 다른 리스크를 노출하는 보안 결점을 가질 수 있다.

발명이 이루고자 하는 기술적 과제

[0014] 본 발명의 실시예는 보안성있는 풀 텍스트 인덱싱을 위한 시스템 및 방법에 관한 것이다. 본 발명은 필터링 및 인덱싱의 워드 브레이킹 프로세스를 제한된 보안 설정(예를 들면, 읽기 전용 보안 설정)을 구비한 프로세스로 이동시킴으로써 정보 노출의 위험성을 경감한다. 이전 인덱싱 시스템에서, 악의적인 사용자는 필터 및/또는 워드 브레이커의 보안 결함을 악용하여 비밀 정보에 액세스하거나 또는 사용자의 머신을 탈취할 수 있었다. 필터 및 워드브레이커를 제한된 보안 설정을 구비한 프로세스로 이동시킴으로써, 문서를 인덱싱할 때 관여하는 다른 프로세스에 영향을 미치지 않고 보다 높은 보안 하에서 필터링 및 워드 브레이킹 프로세스가 행해질 수 있다.

[0015] 본 발명의 일 양상에 따라, 문서의 보안성있는 풀 텍스트 인덱싱에 대한 프로세스가 제공된다. 문서 식별자는 인덱서로부터 수신된다. 문서 식별자에 대응하여 문서가 검색된다. 상기 문서는 제한된 보안 설정 하에서 프로세싱되고, 프로세싱된 문서는 인덱서로 포워딩된다. 부가하여, 문서 식별자는 문서를 검색하기 전에 인덱서에 의해 크로스체크된다. 또한, 상기 제한된 보안 설정 하에서의 프로세스는 보안성을 해칠 기회가 있는 윈도우를 제한하도록 간헐적으로 셧다운된다.

[0016] 본 발명의 다른 양상에 따라, 문서의 보안성있는 풀 텍스트용 시스템이 제공되며, 이 시스템은 인덱서, 프로토콜 핸들러 및 제한된 프로세스를 포함한다. 인덱서는 인덱스를 구축하도록 배열되며, 여기에서, 인덱스에 있는 엔트리는 문서의 워드에 대응한다. 프로토콜 핸들러는 인덱서로부터 수신된 문서 요청에 따라 데이터 소스로부터 문서를 검색하도록 배열된다. 제한된 프로세스는 제한된 보안 설정 하에서 문서를 프로세싱하고 프로세싱된 문서를 인덱스로 포워딩하도록 배열된다.

발명의 구성 및 작용

[0017] 본 발명은, 본 명세서의 일부를 형성하고, 본 발명을 실행하기 위한 특정 예시적 실시예를 도해적으로 도시하는 첨부 도면을 참조하여 이후에 보다 상세히 설명된다. 본 발명은, 그러나, 다수의 서로 다른 형태로 구체화될

수 있으며, 본 명세서에서 설명되는 실시예에 한정되는 것으로 해석되어서는 안되며, 오히려, 이들 실시예는 본 발명의 범위를 당업자에게 철저하고 완벽하게, 그리고 충분히 전달하도록 제공된다. 그들 중에서 특히, 본 발명은 방법 또는 장치로서 구체화될 수 있다. 따라서, 본 발명은 전적으로 하드웨어 실시예 형태를 취하거나, 전적으로 소프트웨어 실시예 형태를 취하거나 또는 소프트웨어와 하드웨어 양상을 결합한 실시예의 형태를 취할 수 있다. 따라서, 후술하는 상세한 설명은 한정하는 의미로 취해지는 것은 아니다.

[0018] 예시적 오퍼레이팅 환경

[0019] 도 1을 참조하면, 본 발명을 구현하기 위한 하나의 예시적 시스템은 컴퓨팅 장치(100) 같은 컴퓨팅 장치를 포함한다. 컴퓨팅 장치(100)는 클라이언트, 서버, 모바일 장치, 또는 임의의 기타 컴퓨팅 장치로서 구성될 수 있다. 매우 기본적 구성에서, 컴퓨팅 디바이스(100)는 전형적으로 적어도 하나의 프로세싱 유닛(102) 및 시스템 메모리(104)를 포함한다. 컴퓨팅 장치의 유형 및 정확한 구성에 따라, 시스템 메모리(104)는 (RAM같은) 휘발성, (ROM, 플래시 메모리 등과 같은) 비휘발성 또는 그 둘의 몇몇 조합일 수 있다. 시스템 메모리(104)는 통상적으로 오퍼레이팅 시스템(105), 하나 이상의 애플리케이션(106)을 포함하고, 프로그램 데이터(107)를 포함할 수 있다. 일 실시예에서, 애플리케이션(106)은 본 발명의 기능성을 구현하기 위한 검색 및 인덱싱 애플리케이션(120)을 포함한다. 이러한 기본 구성은 도 1에 대시(dashed) 라인(108) 내의 컴포넌트로 도시되어 있다.

[0020] 컴퓨팅 장치(100)는 부가의 특징 또는 기능을 구비할 수 있다. 예를 들면, 컴퓨팅 장치(100)는 또한, 예를 들면, 자기 디스크, 광학 디스크, 또는 테이프 같은 부가의 (분리형 및/또는 비분리형) 데이터 저장 장치를 포함할 수 있다. 그러한 부가이 저장 장치는 도 1에 분리형 저장 장치(109) 및 비분리형 저장 장치(110)로 도시되어 있다. 컴퓨터 저장 매체는, 컴퓨터 판독가능 명령, 데이터 구조, 프로그램 모듈, 또는 다른 데이터와 같은 정보를 저장하기 위한 기술 및 임의의 방법으로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 포함할 수 있다. 시스템 메모리(104), 분리형 저장 장치(109) 및 비분리형 저장 장치(110)는 컴퓨터 저장 매체의 모든 예이다. 컴퓨터 저장 매체는, RAM, ROM, EEPROM, 플래시 메모리 또는 다른 메모리 기술, CD-ROM, DVD(digital versatile disks) 또는 다른 광학 저장 장치, 자기 카세트, 자기 테이프, 자기 디스크 저장 장치 또는 다른 자기 저장 장치, 또는 원하는 정보를 저장하고 컴퓨팅 장치(100)에 의해 액세스될 수 있는 임의의 기타 매체를 포함하지만, 이에 한정되는 것은 아니다. 임의의 그러한 컴퓨터 저장 매체는 장치(100)의 일부일 수 있다. 컴퓨팅 장치(100)는 또한 키보드, 마우스, 펜, 음성 입력 장치, 터치 입력 장치 등과 같은 입력 장치(들)(112)을 구비할 수 있다. 디스플레이, 스피커, 프린터 등과 같은 출력 장치(들) 또한 포함될 수 있다.

[0021] 컴퓨팅 장치(100)는 또한 이 장치가 네트워크를 통해 다른 컴퓨팅 장치(118)와 통신가능하게 하는 통신 접속부(116)를 포함한다. 통신 매체는 통상적으로 컴퓨터 판독가능 명령, 데이터 구조, 프로그램 모듈, 또는 반송파 또는 다른 전송 메카니즘 같은 변조된 데이터 신호에 있는 다른 데이터에 의해 구체화될 수 있으며, 임의의 정보 전달 매체를 포함한다. "변조된 데이터 신호"라는 용어는 신호 내에 정보를 인코딩하도록 설정되거나 변환된 특성을 하나 또는 그 이상을 갖는 신호를 의미한다. 예를 들면, 통신 매체는 유선 네트워크 또는 직접 유선 접속 등의 유선 매체와, 음향, RF, 적외선 및 기타 무선 매체 등의 무선 매체를 포함하지만, 이에 한정되지 않는다. 본 명세서에서 사용되는 컴퓨터 판독가능 매체라는 용어는 저장 매체 및 통신 매체 모두를 포함한다.

[0022] 보안성 있는 풀 텍스트 인덱싱을 위한 예시적 실시예

[0023] 본 발명은 네트워크 상에서 문서를 보안성 있게 풀 텍스트 인덱싱하는 것에 관한 것이다. 후술하는 설명 및 청구 범위에서, "문서"라는 용어는 검색 질의 또는 네트워크의 크롤의 결과로서, 네트워크 문서, 파일, 폴더, 웹 페이지, 이메일 첨부물 같은 리턴될 수 있는 임의의 가능한 자원 및 기타 자원을 말한다.

[0024] 인터넷으로부터 들어오는 문서가 악의적이거나 또는 특히 검색 및 인덱싱 시스템에서 일부 취약성을 노출하도록 악용될 수 있음에 따라 그 문서가 신뢰성이 있는지에 대한 의문점이 생긴다. 예를 들면, 이메일이 임의의 사용자 인터랙션없이 수신될 수 있음에 따라 이메일 프로세싱시 특히 위험성이 있다.

[0025] 입력되는 문서를 인덱싱하기 위해, 콘텐츠 필터링(서로 다른 포맷으로부터 평문(plain text) 추출) 및 워드 브레이킹(breaking)이 행해진다. 필터는 상당히 복잡할 수 있고 에러가 있을 가능성이 있는 것으로 알려져 있다. 예를 들면, 사용자의 데이터에 액세스하는 프로세스에 필터링이 행해지고 버퍼 오버런(overrun)(또는 다른 보안 결함)이 일부 필터에 이용되면, 또 다른 위협이 가능하다. 필터와 관계된 보안 침해의 위협은 사적 정보 노출 범위에 있어 사용자 머신을 점거할 수 있다. 워드브레이커는 필터보다는 덜 복잡한 것으로 알려져 있지만, 그럼에도 불구하고 위험을 포함하고 있다.

[0026] 도 2는 종래의 풀 텍스트 검색 및 인덱싱 시스템을 도시한다. 시스템(200)은 인덱스(210), 검색 엔진(220), 및

데이터 저장소(230)를 포함한다. 검색 엔진(220)은 또한 코어 인덱서(222), 워드브레이커(224), 필터(226), 및 프로토콜 핸들러(228)를 포함한다. 동작시에, 검색 엔진(220)은 네트워크(230)로부터 문서(예를 들면, 참조부호 232)를 검색하고, 그 문서를 프로세싱하여 그 문서를 인덱스(210)에 인덱싱한다.

[0027] 프로토콜 핸들러(228)는 특정 데이터 저장소로부터 문서를 획득하도록 구성된 소프트웨어 모듈이다. 일 실시예에서, 검색 엔진(220)에 의해 액세스되는 데이터 저장소의 각 유형에 대하여 서로 다른 프로토콜 핸들러가 포함될 수 있다. 일 실시예에서, 프로토콜 핸들러(228)는 다수의 데이터 저장소를 통해 다양한 문서 유형에 대해 요청을 핸들링하는 다수의 프로토콜 핸들러를 포함할 수 있다. 데이터 저장소는 로컬 및 공유 파일 시스템, 인터넷, 근거리 통신망, 원거리 통신망, 이메일 저장 시스템, 및 검색 엔진이 액세스가능한 문서의 다른 저장 위치를 포함할 수 있다. 일 실시예에서, 데이터 저장소(230)는 특정 애플리케이션과 관계된 저장 파일(예를 들면, 이메일 애플리케이션을 위한 저장 위치)에 대응한다.

[0028] 필터(226)는 문서를 순수한 텍스트 등가물로 변환하도록 배열된 소프트웨어 모듈이다. 일 실시예에서, 필터(226)는 도시된 신호 필터보다는 문서 콘텐츠를 필터링하는 다수의 필터를 포함할 수 있다. 하나 이상의 이들 필터는 "IFilter" 또는 "IFilter interface"라 칭해질 수 있다. IFilter interface는 텍스트 및 특성(소위 속성이라 불림)에 대하여 문서를 스캐닝한다. 이것은 이들 문서로부터 텍스트의 청크(chunks)를 추출하고, 내장된 포맷팅을 필터링 아웃하며 그 텍스트의 위치에 관한 정보를 유지한다. IFilter는 또한 전체 문서 또는 문서의 잘 정의된(well-defined) 부분의 속성인 값의 청크를 추출한다. IFilter는 문서 인덱서 및 애플리케이션 독립 뷰어 같은 고레벨 애플리케이션을 구축하는 기초를 제공한다.

[0029] 워드브레이커(224)는 필터의 순수한 텍스트 출력을 취하고 그 텍스트를 워드 또는 그 텍스트의 언어에 의존하는 다른 유닛으로 토큰화하도록 구성된 소프트웨어 모듈이다. 그 결과의 워드 또는 유닛은 그 특정 언어에 대한 워드-경계 규칙에 의존한다. 예를 들면, 영어의 변형은 워드 경계로서 여백(whitespace)을 우선적으로 고려한다. 워드브레이커는 종종 풀 텍스트 인덱스용 뿐만 아니라, 검색 스트링이 토큰화되고 이들 용어가 인덱스(210)로 포워딩되어 매치(matches)를 찾을 때의 질의 시간에 사용된다. 또 다른 실시예에서, 워드브레이커(224)는 문서 콘텐츠 내의 워드를 토큰화하는 다수의 워드브레이커 알고리즘을 포함할 수 있다. 예를 들면, 다수의 워드브레이커가 포함되어 다수의 언어로 되어 있는 문서의 데이터 저장을 핸들링할 수 있다.

[0030] 코어 인덱서(222)는 워드브레이커(224)로부터 출력된 유닛 또는 워드로부터 풀 텍스트 인덱스(예를 들면, 인덱스(210))를 구축하도록 배열된 소프트웨어 모듈이다. 인덱스(210)의 구축 버전으로, 검색 엔진(220)은 인덱스(210)에 있는 엔트리에 정합되는 검색 용어에 대응하는 문서를 검색하도록 사용될 수 있다. 많은 코어 인덱서 유형 및 설계는 공지되어 있다. 사용되는 코어 인덱서의 특정 유형은 설명된 본 발명에 한정되는 것이 아니다. 따라서, 본발명은 본 명세서에서 코어 인덱서(222)의 구조를 상세하게 설명하지는 않는다. 상이한 많은 코어 인덱서 구성이 본 발명의 사상 또는 범위를 벗어나지 않고 사용될 수 있다.

[0031] 동작시에, 코어 인덱서(222)는 데이터 저장소(230)로부터 프로토콜 핸들러(228)로 검색될 문서(예를 들면, 232)의 식별자를 제공한다. 예를 들면, 데이터 저장소(230)가 파일 시스템이면, 식별자는 파일명 및 경로에 대응할 수 있다. 데이터 저장소(230)가 네트워크인 경우에, 인덱서(222)는 문서의 URL를 프로토콜 핸들러(228)로 제공할 수 있다. 프로토콜 핸들러(228)는 데이터 저장소(230)로부터 문서를 검색하고 그 문서를 필터(226)로 전파한다. 필터(226)는 문서를 순순한 텍스트로 변환하고 그 순수한 텍스트를 워드브레이커(224)로 출력한다. 워드브레이커(224)는 순순한 텍스트를 개별 워드(또는 유닛)으로 토큰화하고 그 워드를 코어 인덱서(222)로 출력한다. 코어 인덱서(222)는 수신된 워드를 사용하여 인덱스(210)를 구축한다.

[0032] 도시된 종래의 시스템에서, 코어 인덱서(222), 워드브레이커(224), 필터(226), 및 프로토콜 핸들러(228)의 동작은 특정 컴퓨팅 장치 상의 동일한 프로세스의 일부 또는 동일한 보안성 우선 설정을 공유하는 다수의 프로세스의 일부이다. 일 실시예에서, 모든 프로세스는 로컬 보안 컨텍스트에서 실행한다. 그러나, 워드브레이커(224) 및 필터(226)는, 워드브레이커(224)와 필터(226)가 사용되는 것에 의존하여 많은 상이한 저자 중 하나를 구비할 수 있다. 이들 컴포넌트는 데이터 저장소(230)의 문서와 관계될 수 있는 콘텐츠 및 언어의 다양성 때문에 다양한 소스에 의해 기록될 수 있다. 저자의 다양성으로 인해 수많은 보안 결함이 발생된다. 예를 들면, 버퍼 오버런은 악의적인 문서가 인덱싱 프로세스를 "점거"할 수 있는 워드브레이커(224) 또는 필터(226)의 문맥에서 발생할 수 있다. 이 문제는, 프로토콜 핸들러(228)가 문서를 액세스하기 위해, 상기 프로세스가 임의의 우선 순위 레벨(예를 들면, 판독 및 기록)로 실행될 필요가 있다는 사실과 관계가 있다. 보호하지 않으면, 그러한 보안 침해는 비밀 정보를 유포시키게 되거나 사용자 컴퓨팅 장치를 탈취하게 된다.

[0033] 도 3은 본 발명에 따른 풀 텍스트 검색 및 인덱싱 시스템을 도시한다. 검색 및 인덱싱 시스템(300)은 이전 시

스템의 보안 결함을 해결하는 시스템을 설명한다. 시스템(300)은, 인덱스(310), 코어 인덱서(322), 워드브레이커(332), 필터(334), 프로토콜 핸들러(342), 및 데이터 저장소(350)를 포함한다는 점에서 도 2의 시스템(200)과 유사하다. 시스템(300)의 소프트웨어 모듈은 도 2에 도시된 시스템(200)의 소프트웨어 모듈과 유사하게 배열된다. 그러나, 본 발명은 소프트웨어 모듈을 단일 프로세스로서 실행하기 보다는 세개의 프로세스(320, 330, 340)로 분리한다.서로 다른 보안 설정을 서로 다른 세개의 프로세스에 적용함으로써, 워드브레이커(332) 및 필터(334)와 관계된 이전의 보안 결점이 경감될 수 있다.

[0034] 일 실시예에서, 프로토콜 핸들러(342)는 프로세스(340)에 따라 실행되며 프로토콜 핸들러(342)가 데이터 저장소(350)로부터 판독할 수 있는 적용된 보안 설정을 구비한다. 프로토콜 핸들러(342)용의 상기 보안 설정은 코어 인덱서(322) 및 워드브레이커(332) 및 필터(334)에 적용된 제한된 보안 설정과는 분리된다.

[0035] 코어 인덱서(322)는 데이터 소스(350)에 저장된 사용자 데이터를 코어 인덱서가 판독하거나 그 사용자 데이터에 기록하는 것을 방지하는 제한된 보안 설정을 구비한 프로세스(320)에 따라 실행된다. 대신에, 코어 인덱서(322)는 특정 위치(인덱스(310))에 대하여 기록 액세스 우선 순위로 한정된다.

[0036] 워드브레이커(332) 및 필터(334)는 또한 또 다른 제한된 보안 설정 하에서 소프트웨어 모듈을 실행하는 프로세스(330)로 분리된다. 일 실시예에서, 워드브레이커(332) 및 필터(334)는 최상의 제한된 보안 설정 하에 있으며, 여기에서는 어떠한 소프트웨어 모듈도 사용자 데이터에 대응하는 위치를 포함하여 임의의 메모리 위치로의 판독 액세스 또는 기록 액세스도 허여되지 않는다. 워드브레이커(332) 및 필터(334) 프로세스가 제한된 보안 설정 하에서 실행되기 때문에, 정보 노출 또는 컴퓨팅 장치의 하이잭킹의 위험성이 경감된다. 본 발명에 따라, 시스템(300)의 아키텍처 또한 보안 침해의 위험성을 추가로 감소시키는 부가의 보안 측정이 취해질 수 있게 한다. 부가된 보안에 대한 이들 추가의 단계가 시스템(300)의 아키텍처를 이용하는 보안성있는 풀 텍스트 인덱싱에 대한 프로세스와 관련하여 하기(도 4)에 설명된다.

[0037] 도 4는 본 발명에 따라 도 3의 시스템에 대응하는 문서의 풀 텍스트 인덱싱에 대한 예시적 프로세스를 도시한다. 프로세스(400)는 블록(402)에서 시작하며, 여기에서, 도 3에 도시된 시스템(300)은 문서가 인덱싱을 위해 검색될 준비가 되어 있는 상태에 있게 된다. 프로세싱은 블록(404)에서 계속된다.

[0038] 블록(404)에서, 코어 인덱서(322)는 프로세스(330)를 통해 문서 요청을 프로토콜 핸들러(342)로 전송한다. 전술한 바와 같이, 문서 요청은 일부 문서 ID(예를 들면, URL, 파일 경로 등)에 따라 문서를 식별한다. 상기 요청이 코어 인덱서(322)로 전송되면, 프로세싱은 결정 블록(406)에서 계속된다.

[0039] 결정 블록(406)에서, 문서 ID는 코어 인덱서(322)가 요청한 실제 문서로서 프로토콜 핸들러(322)에 의해 검증된다. 프로토콜 핸들러(322)는 호출(call)을 코어 인덱서로 전송하여 문서 ID를 직접 크로스체크한다. 이론적으로, 악의적 공격자가 프로세스(330)를 하이재킹할 수 있다면, 공격자는 프로세스(330)로부터 프로토콜 핸들러(342)로 직접 문서 요청을 열거함으로써 사용자 데이터를 또 다른 위치로 흡수할 수 있는 가능성이 있다. 문서 ID가 코어 인덱서(322)에 의해 요청된 문서에 대응하는지를 체크함으로써, 코어 인덱서(322)는 문서를 검색하기 전에 그 문서를 검증할 수 있다. 크로스체크가 수행된 임의의 논-매치(non-matches)는 수신된 문서 요청이 허위라는 것을 프로토콜 핸들러(342)에 환기시킨다. 문서 ID가 코어 인덱서(322)로부터 입력되는 것으로 검증되는 경우, 프로세싱은 블록(410)을 진행한다. 그러나, 문서 ID가 크로스체크 동안 코어 인덱서(322)로부터 입력되는 것으로 검증되지 않는 경우에는 프로세싱은 블록(408)로 진행한다.

[0040] 블록(408)에서, 프로토콜 핸들러(342)에 의한 문서의 페치(fetch)는 방지된다. 또한, 제한된 프로세스(330)가 하이잭킹되었을 가능성이 있기 때문에 그 프로세스를 중단해야 한다는 것을 지시하는 플래그가 설정될 수 있다. 다음에, 프로세싱은 제한된 프로세스(330)의 셧다운(shut down)이 수행될 수 있는 결정 블록(416)으로 진행한다.

[0041] 블록(410)에서, 문서 ID가 검증되었기 때문에, 프로토콜 핸들러(342)는 데이터 저장소(350)로부터 문서를 페치한다. 문서가 검색되면, 그 문서는 제한된 프로세스(330)로 포워딩되고, 프로세싱은 블록(412)에서 계속된다.

[0042] 블록(412)에서, 문서는 필터(334) 및 워드브레이커(332)에 의해 제한된 보안 설정 하에서 프로세싱된다. 전술한 바와 같이, 제한된 보안 설정은 보안 단점에 기인한 프로세스(330)를 침해할 수 있는 악의적 공격자에 의한 사용자 데이터로의 액세스를 경감한다. 프로세싱은 블록(414)에서 계속된다.

[0043] 블록(414)에서, 현재 토큰화된 워드에 대응하는 프로세싱된 문서가 코어 인덱서(322)로 포워딩된다. 다음에, 코어 인덱서(322)는 수신된 데이터로부터 인덱스(310)를 구축할 수 있다. 일 실시예에서, 인덱스(310)는 풀 텍

스트 인덱스에 대응한다. 다음에, 프로세싱은 결정 블록(416)에서 계속된다.

[0044] 결정 블록(416)에서, 제한된 프로세스(330)가 일시적으로 셧다운되거나 중지되어야 하는지가 결정된다. 일 실시예에서, 제한된 프로세스(330)는 경과 시간 기간(예를 들면, 수 초)에 대응하여 간헐적으로 중지된다. 또 다른 실시예에서, 제한된 프로세스(330)는 (예를 들면, 문서 크로스체크가 실패한 경우, 블록(406 및 408) 참조) 제한된 프로세스(300)가 하이잭킹되었다는 의혹이 있기 때문에 일시적으로 셧 다운된다. 제한된 프로세스(330)가 셧다운 되어야 한다고 결정되면, 프로세싱은 제한된 프로세스(330)가 일시적으로 중지되는 블록(418)으로 진행된다. 일 실시예에서, 제한된 프로세스(330)는 소정 시간 기간 동안 일시적으로 중지된다. 또 다른 실시예에서, 제한된 프로세스(330)는 상기 프로세스로부터 하이잭커를 쫓아내기에 충분한 듀레이션 동안 중지된다. 제한된 프로세스(300)를 셧다운하게 되면, 악의적 공격자가 프로세스를 제어하는 시간을 제한하게 되고, 그 결과 보안 침해의 기간을 제한할 수 있다. 제한된 프로세스가 중지되고 재시작되면, 또는 일시적 중지가 필요하지 않으면, 프로세싱은 프로세스(400)가 다른 문서 요청으로 계속되거나 다른 프로세싱으로 진행할 수 있는 블록(420)으로 진행한다.

[0045] 추가의 실시예에서, 프로세스(400)에 따른 문서 요청 및 문서 검색은 일괄적으로 수행된다. 달리 말하면, 코어 인덱서(322)는 문서 ID 묶음을 프로토콜 핸들러(342)로 포워딩한다. 프로토콜 핸들러(342)는 문서의 묶음을 검색하고 이들을 프로세싱을 위해 제한된 프로세스(330)로 포워딩한다. 다음에, 코어 인덱서가 제한된 프로세스(330)로부터 프로세싱된 문서의 출력을 수신함에 따라 문서 묶음이 코어 인덱서(322)에 의해 인덱싱된다.

[0046] 또 다른 실시예에서, 제한된 프로세스(330)를 통해 전파하는 것 대신에 문서 요청이 프로토콜 핸들러(342)로 직접 포워딩될 수 있다. 그러한 실시예에서, 블록(406 및 408)의 크로스체크 단계는, 프로토콜 핸들러(342)가 소스에 기인한 요청의 유효성을 즉시 확인할 수 있기 때문에 필요하지 않다.

발명의 효과

[0047] 전술한 바와 같이, 본 발명은 필터링 및 워드브레이킹 프로세스로부터 사용자 데이터로의 직접 액세스를 제거하고 또 다른 프로세스에 위임함으로써 정보 노출의 위험성을 경감한다. 또한, 문서 식별자를 크로스체크하여 요청된 문서가 사실상 인덱싱되고 있다는 것을 보장한다. 또한, 필터링 및 워드 브레이킹 프로세스는 주기적으로 중단할 수 있고, 따라서, (버퍼 오버런 또는 다른 메카니즘을 통해) 제한된 보안 설정이더라도 하이잭킹되는 경우에, 사용자 데이터로의 시간 노출이 프로세스의 유효 기간에 의해 제한된다.

[0048] 전술한 사양, 예 및 데이터는 본 발명의 구성의 용도 및 제조에 완전한 설명을 제공한다. 본 발명의 많은 실시예가 본 발명의 사상 및 범위를 벗어나지 않고 행해질 수 있으므로, 본 발명은 첨부된 청구범위에 귀속된다.

도면의 간단한 설명

[0001] 도 1은 본 발명의 예시적 일 실시예에 사용될 수 있는 예시적 컴퓨팅 장치를 도시한 도면.

[0002] 도 2는 종래의 풀 텍스트 검색 및 인덱싱 시스템을 도시한 도면.

[0003] 도 3은 본 발명에 따른 풀 텍스트 검색 및 인덱싱 시스템을 도시한 도면.

[0004] 도 4는 본 발명에 따라, 도 3의 시스템에 대응하는 문서의 풀 텍스트 인덱싱을 위한 예시적 프로세스를 도시한 도면.

[0005] <도면의 주요 부분에 대한 부호의 설명>

[0006] 310: 인덱스

[0007] 322: 코어 인덱서

[0008] 332: 워드브레이커

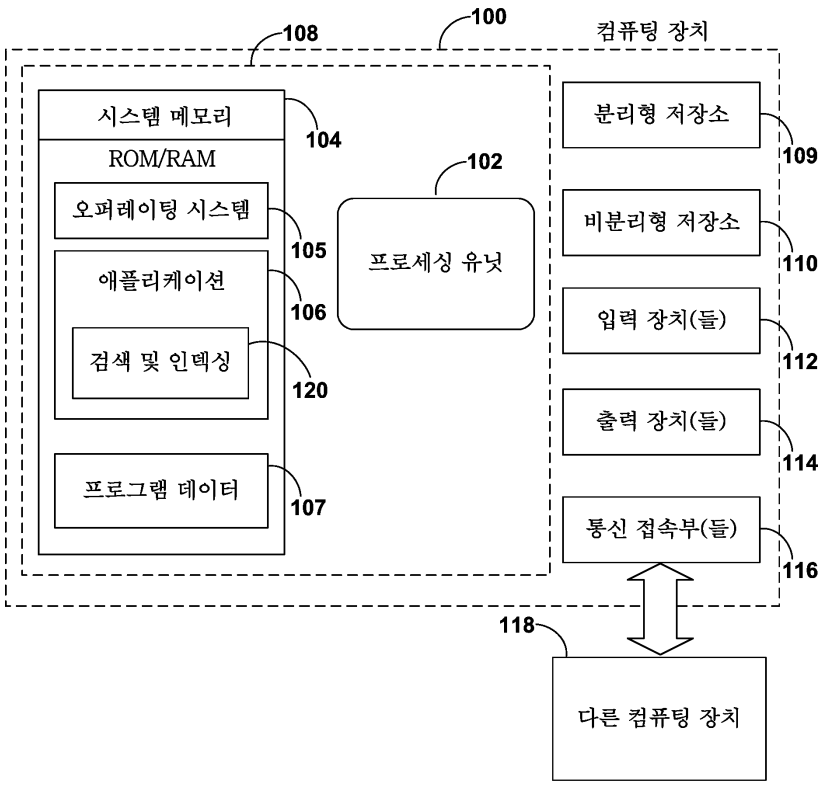
[0009] 334: 필터

[0010] 342: 프로토콜 핸들러

[0011] 350: 데이터 저장소

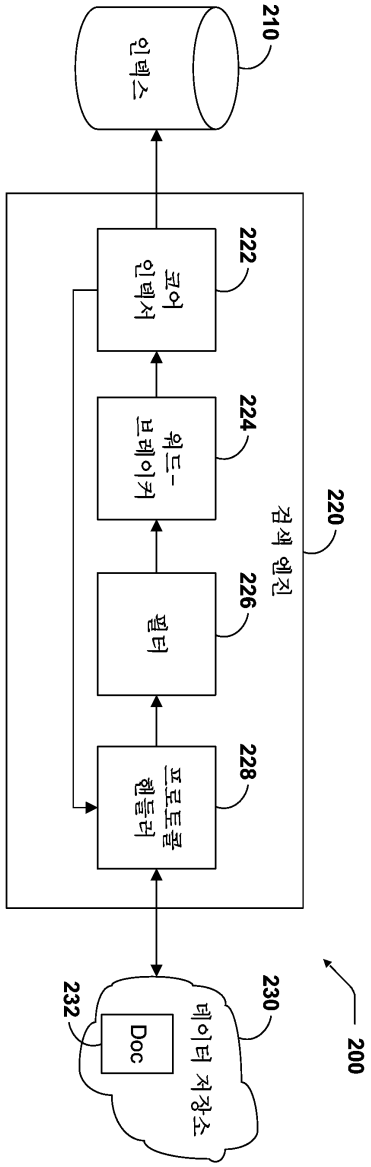
도면

도면1

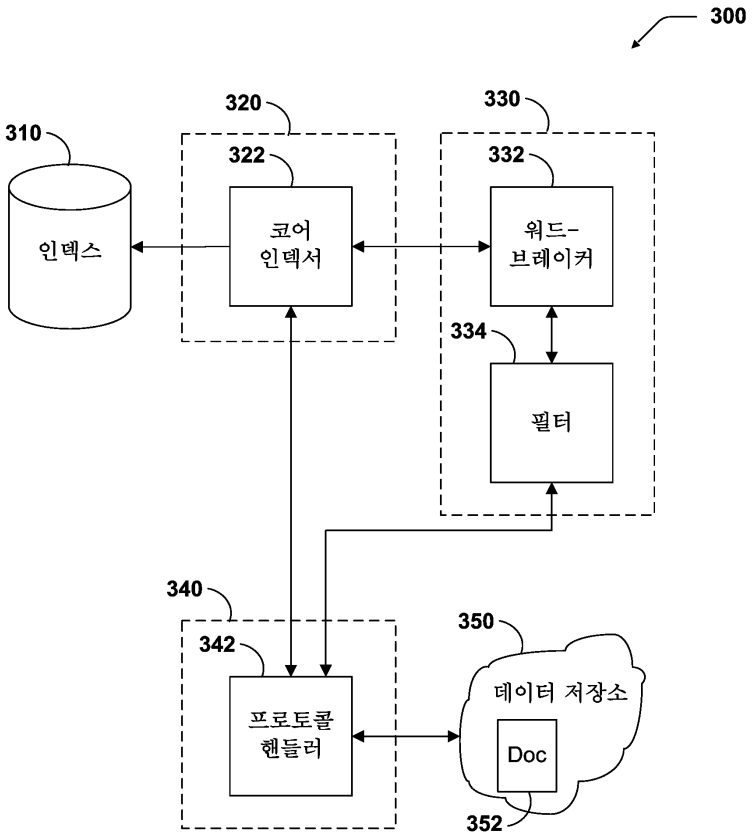

100
컴퓨팅 장치
108
시스템 메모리
104
ROM/RAM
오퍼레이팅 시스템
105
애플리케이션
106
검색 및 인덱싱
120
프로그램 데이터
107
102
프로세싱 유닛
분리형 저장소
109
비분리형 저장소
110
입력 장치(들)
112
출력 장치(들)
114
통신 접속부(들)
116
118
다른 컴퓨팅 장치

도면2

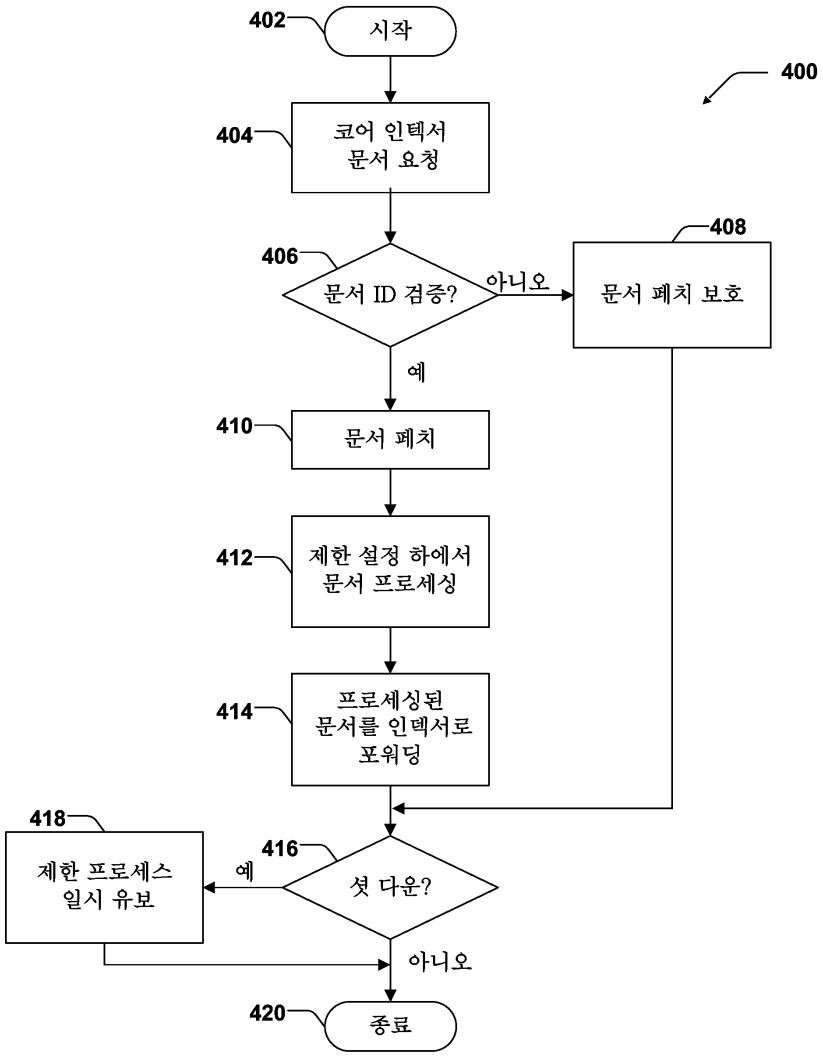
(종래 기술)


210
인덱스
222
코어 인덱서
224
워드-브레이커
220
검색 엔진
226
필터
228
프로토콜 핸들러
200
230
데이터 저장소
232
Doc

도면3


300
310
320
322
330
332
인덱스
코어
인덱서
워드-
브레이커
334
필터
340
342
350
프로토콜
핸들러
데이터 저장소
Doc
352

도면4


400
402
시작
404
코어 인텍서
문서 요청
406
문서 ID 검증?
아니오
408
문서 폐치 보호
예
410
문서 폐치
412
제한 설정 하에서
문서 프로세싱
414
프로세싱된
문서를 인덱서로
포워딩
416
셧 다운?
예
418
제한 프로세스
일시 유보
아니오
420
종료